

Conceptualisation et exploitation d'un graphe de pangénome partitionné comme représentation compacte de la diversité du répertoire génique des espèces procaryotes

Guillaume Gautreau

▶ **To cite this version:**

Guillaume Gautreau. Conceptualisation et exploitation d'un graphe de pangénome partitionné comme représentation compacte de la diversité du répertoire génique des espèces procaryotes. Génomique, Transcriptomique et Protéomique [q-bio.GN]. Université Paris-Saclay, 2020. Français. ⟨NNT : 2020UPASE001⟩. ⟨tel-02933311⟩

HAL Id: tel-02933311

https://theses.hal.science/tel-02933311v1

Submitted on 8 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Thèse de doctorat

NNT : 2020UPASE001

université PARIS-SACLAY

Conceptualisation et exploitation d'un graphe de pangénome partitionné comme représentation compacte de la diversité du répertoire génique des espèces procaryotes

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°577 : Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Sciences de la vie et de la santé
Unité de recherche : Université Paris-Saclay, Univ Évry, CNRS, CEA, Génomique métabolique, 91057, Évry-Courcouronnes, France.
Référent : Université d'Évry Val d'Essonne

Thèse présentée et soutenue à Évry-Courcouronnes, le 27 février 2020, par

Guillaume GAUTREAU

Composition du Jury

Sophie SCHBATH Directrice de recherche INRAE, Unité Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE)	Présidente
Philippe LOPEZ Professeur Sorbonne Université, Université Pierre et Marie Curie, Muséum national d'Histoire naturelle	Rapporteur
Éric RIVALS Directeur de recherche CNRS, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM)	Rapporteur
Laurent JACOB Chargé de recherche CNRS, Laboratoire Biométrie et Biologie Évolutive	Examinateur
Claire LEMAITRE Chargée de recherche INRIA, INRIA Rennes Bretagne Atlantique	Examinatrice
Claudine MÉDIGUE Directrice de recherche CNRS, Genoscope CEA	Directrice de thèse
David VALLENET Chercheur CEA, Genoscope CEA	Co-Directeur de thèse

À l'espoir que le réchauffement du climat soit contenu en dessous de 1.5°C sans étendre la misère (MASSON-DELMOTTE et al., 2018).

À la mémoire des français morts, blessés et mutilés lors des événements de 2018-2019 (CHAUVIN et al., 2019 ; LARTIZIEN et al., 2019).

À ma chérie Mélanie.

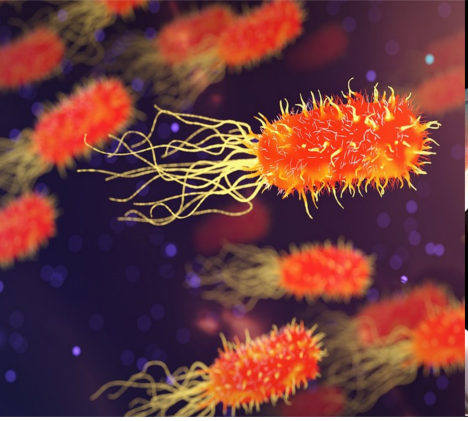




Table des Matières

Remerciements 11

Introduction, contexte et enjeux 15

I Du génome au pangénome

1 Biologie et évolution des procaryotes 21

1.1 Bases moléculaires de l'évolution 22

1.2 Structure des génomes procaryotes 24

1.2.1 Structure d'un chromosome procaryote 24

1.2.2 Les gènes procaryotes 24

1.3 Classification des procaryotes 25

1.3.1 Critères morphologiques 25

1.3.2 Critères physiologiques 26

1.3.3 Critères de pathogénicité 27

1.3.4 Critères de sérogroupage 27

1.3.5 Critères génomiques 27

1.4 Les trois modes d'évolution des procaryotes 28

1.4.1 L'évolution verticale 28

1.4.2 Réduction de génomes 30

1.4.3 Les transferts horizontaux 30

2 Aux origines de la pangénomique 33

2.1 Les génomes procaryotes, clones ou chimères ? 34

2.1.1 Les recombinaisons étaient admises mais sous-estimées 34

2.1.2 Réévaluation de la fréquence des recombinaisons homologues 35

2.1.3 Réévaluation de l'influence des transferts horizontaux de gènes 35
2.2 Qu'est-ce qu'une espèce procaryote ? 36
2.2.1 Vue historique de la taxonomie procaryote 36
2.2.2 Quels critères univoques pour délimiter les espèces ? 37
2.2.3 Vers une taxonomie repensée avec des critères génomiques ? 38
2.3 Les premières analyses de pangénomes 39
2.3.1 Le premier pangénome de l'espèce *Escherichia coli* 40
2.3.2 Le premier pangénome de l'espèce *Buchnera aphidicola* 41
2.3.3 Le supra-génome et l'hypothèse du génome distribué 41
2.3.4 Concevoir un vaccin par l'étude du pangénome de l'espèce *S. agalactiae* 42
2.3.5 La naissance « officielle » de la pangénomique 42

3 Modélisation des pangénomes 45
3.1 Modélisation par décroissance exponentielle 45
3.2 Modélisation par la loi de Heaps 49
3.3 Modélisation d'un pangénome par modèles de mélange binomiaux 51
3.3.1 Modèle binomial utilisant des classes de fréquence fixées *a priori* 51
3.3.2 Modèle binomial déterminant les classes de fréquence des gènes 54
3.3.3 Critique des modèles à pangénomes finis 57
3.4 La trichotomie : cœur, coquille et nuage 57
3.5 Le modèle infiniment plein de gènes (IMG) corrobore la trichotomie 59
3.5.1 Le modèle généalogique infiniment plein de gènes (IMG) 59
3.5.2 Modèle d'évolution totalement neutre 63
3.5.3 Impact d'une approche trichotomique sur les prédictions du modèle IMG . . 63
3.5.4 Impact de trois types d'arbres différents sur les prédictions du modèle IMG . . 65
3.6 Rejet de l'hypothèse d'une évolution neutre 66
3.7 Conclusion sur les modèles proposés 68

4 Le pangénome est-il adaptatif ou non ? 71
4.1 L'apport de la génétique des populations 71
4.1.1 Notions de base en génétique des populations 71
4.1.2 Estimation de la diversité pangénomique à travers sa fluidité 72
4.1.3 Corrélation entre la fluidité des génomes et la diversité en mutations synonymes 72
4.1.4 Critique de la génétique des populations pour expliquer le pangénome 73
4.2 L'acquisition de nouveaux gènes est délétère en soi 74
4.3 Pourquoi donc les procaryotes ont-ils des pangénomes ? 74
4.3.1 L'hypothèse du gène égoïste 75
4.3.2 Les transferts horizontaux seraient bénéfiques 75
4.3.3 Des gènes avantageux entraîneraient un balayage génétique 76
4.3.4 Les migrations et les transferts horizontaux de gènes diversifient le pangénome 76

4.3.5 Conclusion et réserves sur le caractère adaptatif des pangénomes 77

5 Les outils utilisés en pangénomique 79

5.1 Approches basées sur les familles de gènes 80

5.1.1 EDGAR 81
5.1.2 Panseq 82
5.1.3 The Prokaryotic-genome Analysis Tool (PGAT) 82
5.1.4 The Pan-genome Analysis Pipeline (PGAP) 83
5.1.5 PanOCT 84
5.1.6 GET_HOMOLOGUES 84
5.1.7 Integrated Toolkit for Exploration of Microbial Pan-genomes (ITEP) 85
5.1.8 Roary 85
5.1.9 Piggy 86
5.1.10 The Bacterial Pan Genome Analysis (BPGA) 86
5.1.11 Large Scale Blast Score Ratio (LS-BSR) 86
5.1.12 micropan 87

5.2 Approches à l'échelle de la séquence 87

5.3 Les approches d'indexation de génomes basées sur une référence 88

5.3.1 Referentially Compressed Search Index (RCSI) 88
5.3.2 Multiple Genome Index (MuGI) 88
5.3.3 The Journaled String Tree (JST) 88
5.3.4 BWBBLE 89

5.4 Les graphes de séquence 91

5.4.1 GenomeMapper 91
5.4.2 Variation Graph (VG) 91
5.4.3 HISAT2 92
5.4.4 PanCake 92

5.5 Les graphes de De Bruijn 92

5.5.1 Cortex 93
5.5.2 SplitMEM 93
5.5.3 Bloom Filter Trie (BFT) 94
5.5.4 TwoPaCo 95
5.5.5 DBGWAS 96

II La méthode PPanGGOLiN

6 Motivation de l'approche PPanGGOLiN 99

6.1 Améliorer le partitionnement des pangénomes 99

6.1.1 Les problèmes causés par la dichotomie génome cœur/génome accessoire 99
6.1.2 Causes de la perte des gènes cœurs 100
6.1.3 Assouplir la dichotomie entre le génome cœur et le génome accessoire . . . 102

6.1.4 Détecter le génome coquille et le génome nuage 105
6.2 Limites des approches tenant compte des contextes génomiques 107
6.2.1 Périmètre de l'usage des graphes de pangénome à l'échelle de la séquence 107
6.2.2 Intérêt de considérer la colocalisation des gènes 108
6.3 PPanGGOLiN, le chaînon manquant de la pangénomique 110

7 Construire un graphe de pangénome partitionné 111
7.1 Vue d'ensemble de la méthode PPanGGOLiN 111
7.2 Les familles de gènes, les points de repère entre les génomes 113
7.3 La construction d'un graphe de pangénome 116
7.4 Le partitionnement du graphe de pangénome 117
7.4.1 Modéliser la matrice du pangénome par un modèle de mélange de Bernoulli 117
7.4.2 Partitionnement spatial au travers du graphe de pangénome 120
7.4.3 Estimation des paramètres du modèle et partitionnement par NEM 121
7.4.4 Partitionnement par morceaux 125
7.4.5 Estimation du coefficient de régularité spatiale (β) 127
7.5 L'emmêlement du graphe de pangénome 129
7.5.1 Cas des éléments transposables 129
7.5.2 Cas des familles multigéniques 131
7.6 L'écaillage du graphe de PPanGGOLiN 134
7.7 L'échantillonnage des génomes 135

8 Description technique de la méthode PPanGGOLiN 139
8.1 Informations sur l'implémentation effectuée 139
8.1.1 Informations sur le langage et les dépendances utilisés 139
8.1.2 Disponibilité de l'application 140
8.1.3 Contributions 140
8.2 Fonctionnement de l'application PPanGGOLiN 140
8.2.1 Interface en ligne de commande de la méthode PPanGGOLiN 140
8.2.2 Fichiers d'entrée 142
8.2.3 Fichiers de sortie 143
8.2.4 Fichiers de visualisation 144
8.2.5 Fichiers temporaires 148
8.3 Performances de l'application PPanGGOLiN 148
8.3.1 Temps de calcul 148
8.3.2 Usage mémoire 149

9 Résultats de la méthode appliquée sur 439 espèces 151
9.1 Analyse des pangénomes de GenBank 151
9.1.1 À quoi ressemble un graphe de pangénome partitionné ? 151
9.1.2 Résultats du partitionnement de PPanGGOLiN 157
9.1.3 Comparaison du génome persistant avec le génome cœur assoupli 159
9.1.4 Structure et dynamique du génome coquille et nuage 162
9.1.5 Liens entre la phylogénie et les schémas de présence du génome coquille . 167
9.2 Application sur des génomes issus de métagénomes 173
9.2.1 Rappels sur la notion de métagénomique 173
9.2.2 Protocole de la comparaison entre les MAGs et les isolats de GenBank 173
9.2.3 Résultat de la comparaison des génomes persistants 175
9.2.4 Quels résultats aurait-on obtenu avec le génome cœur assoupli ? 176
9.2.5 Quelles différences du point de vue du génome coquille ? 177
9.2.6 Conclusion sur le partitionnement du pangénome des MAGs 178

III Applications basées sur la méthode PPanGGOLiN

10 Vers la constitution d'une banque de pangénome 181
10.1 L'idée d'une banque de graphes de pangénomes partitionnés 181
10.2 Limites des ressources en ligne existantes 182
10.3 Obtenir les génomes de GenBank 183
10.4 Filtrer les erreurs dans la banque de données GenBank 183
10.4.1 Filtrer les erreurs d'assignation taxonomique 183
10.4.2 Filtrer les redondances 185
10.4.3 Filtrer les assemblages fragmentés 185
10.5 Une interface pour explorer les graphes de pangénomes partitionnés 186

11 Détecter les régions de plasticité génomique 189
11.1 Limites des approches actuelles de détection des RGPs 189
11.1.1 Limites des approches de détection de GIs par approche compositionnelle 189
11.1.2 Limites des approches de détection de GIs par la génomique comparée . . 190
11.2 panRGP : utiliser un pangénome partitionné pour détecter les GIs 191
11.2.1 Motivation de l'approche panRGP 191
11.2.2 Étape d'attribution des scores 191
11.2.3 Étape d'identification des RGPs 192
11.2.4 Regroupement des RGPs appartenant au même spot 194
11.3 Résultats de la méthode panRGP 196
11.3.1 Exemple d'une RGP extraite de la souche d'*A. baumannii* AYE 196
11.3.2 Comparaison avec d'autres outils 196

11.4 Conclusion sur la méthode panRGP **198**

Conclusion, bilan et perspectives 199

11.5 Bilan des contributions scientifiques **199**

11.5.1 PPanGGOLiN 199
11.5.2 panRGP 201
11.5.3 MicroScope 201
11.5.4 PanGBank 201
11.5.5 Réserves sur l'approche pangénomique proposée 202

11.6 Perspectives à approfondir ou à envisager **202**

11.6.1 Association des chemins à des données externes 203
11.6.2 Faire le lien entre PPanGGOLiN et les approches à l'échelle de la séquence 204
11.6.3 Autres perspectives 205

11.7 Conclusion générale **205**

A Annexes 207

A.1 Soutenance de thèse de doctorat **207**

A.2 Présentation orale de PPanGGOLiN **208**

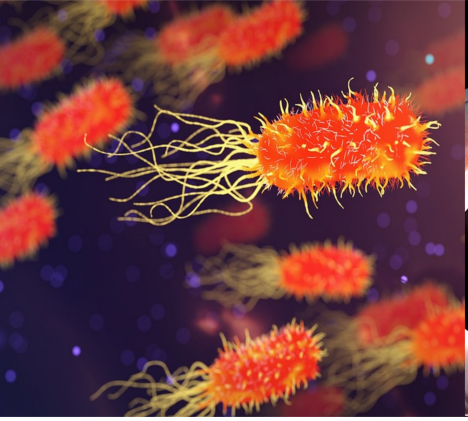
A.3 Poster PPanGGOLiN **209**

A.4 Poster PanGBank **210**

A.5 Poster panRGP **211**

A.6 Poster MicroScope **212**

Bibliographie 212





Remerciements

SACHEZ QU'ICI CHAQUE MOT EST PESÉ ET DÉPLOIE AUSSI PRÉCISÉMENT QU'IL M'EST POSSIBLE L'ESTIME QUE JE VOUS PORTE. SI VOUS EN ÊTES ABSENT, NUL BESOIN DE CROIRE À DE L'ANIMOSITÉ, PRÉFÉREZ-Y LE RASOIR D'HANLON.

Au début de ma thèse, je pensais travailler avec Stéphane qui avait obtenu un financement pour m'encadrer sur une thèse portant sur la méthode de comparaison quantitative de métagénomes. Pour y parvenir, David proposa une idée qui lui tenait à cœur, celle d'utiliser comme référence, des graphes de pangénomes. Cette structure de données devait être la brique première de mon travail, qu'on estimait rapide à publier, avant de s'en servir pour traiter de la métagénomique quantitative. Que nenni, les graphes de pangénomes occuperont la majeure partie de ce travail de thèse. Notamment avec la méthode statistique pour partitionner ce graphe que m'ont proposée Catherine et Christophe que je tiens à remercier. Même si je sais que de multiples facteurs sont intervenus dans ta décision, Stéphane, j'imagine que le fait que je n'ai pas traité le sujet que tu m'avais proposé, a dû participer à motiver ton départ du Genoscope. Je souhaite donc m'excuser Stéphane pour ces moments où tu voyais mon travail s'éloigner de ton dessein initial. Malgré cela, tu restas patient, à l'écoute, compréhensif et impliqué dans le suivi de mon travail. Merci pour cela.

Cher David, tu forces le respect, malgré la quantité de travail considérable que te demande la plateforme Microscope et les projets du LABGeM, la bibliographie, la commission CNRS, les nuits de maintenance et de « debugging », les nombreuses réunions qui débordent, les mails par dizaines, les questions qui te viennent de toute part à tout moment, les trois thèses que tu as encadrées en même temps que la mienne, tu es toujours resté un co-directeur de thèse très impliqué. On dira que c'est ton rôle de chercheur, que c'est normal mais moi ce que je vois c'est quelqu'un qui mérite mon profond respect doublé de mes plus vifs remerciements. J'ai beaucoup appris avec toi. Bien sûr parfois, on s'est mal compris, on s'est opposé mais au final cette dialectique fut séminale car je suis fier, et je n'ai pas l'auto-compliment facile, de ce que nous avons fait avec PPanGGOLiN.

Enfin, ma directrice de thèse, Claudine, je dois te le dire, j'aurais aimé qu'on discute un peu plus de mon travail de thèse ensemble. Et tout en disant ça, je ne t'en fais pas grief car je ne t'en ai pas fait la demande. Au fond, je savais que tu suivais mes recherches de loin, que tu étais toujours là pour moi si nécessaire, que tu t'assurais que je puisse opérer dans de bonnes conditions, bref, que tu

participais à toutes ces choses invisibles mais finalement indispensables pour le bon fonctionnement d'un laboratoire de recherche, s'assurer que les moyens soient là, que les objectifs soient réalisables dans de bonnes conditions de travail. Pour tout ce travail souterrain, et dont je peine à mesurer l'étendue, mais qui a fait que concrètement j'ai eu tout ce dont j'avais besoin pour mener à bien ma thèse, je t'exprime ma spéciale reconnaissance.

Adelme, je suis très heureux que tu sois venu nous rejoindre au LABGeM. J'apprécie ta sincérité, ta capacité d'écoute. Dans les moments d'incertitude doute tu m'as bien aidé. Tu savais trouver les mots pour écarter les apories, lever les doutes, apporter les solutions, provoquer le dépassement dialectique entre moi et David. Ton arrivée fut heureuse, je te recommande au bonheur, je te vois devenir un grand chercheur.

Comme je suis heureux de t'avoir rencontré Rémi, en 3 ans, tu resteras pour moi un grand personnage de ma vie. Nos discussions à refaire le monde à perte de nuit, l'impression de pouvoir tout te dire, de l'humour qui respecte, toi l'athlète, d'une excellence technique (merci pour l'interface de PanGBank !), me voilà incapable face à l'ineffable que tu m'évoques. Tu es de ces personnes qui rendent la vie belle et font oublier le reste. Quel merci serait à la hauteur ? Félicitations à toi et Alice pour l'heureuse nouvelle que vous m'avez apprise il y a peu.

Si je devais nommer le scientifique qui m'épate le plus au Genoscope, je crois sans doute que je répondrais Amin(e). Véritable savant, tu es de tout, partout, curieux de tout, toujours prêt à de nouveaux projets, toujours de nouvelles idées et une connaissance à la fois profonde et large. D'une assurance et d'une vivacité d'esprit qui m'ont souvent laissé interdit, tu es drôle et puissant, Hercule et Bacchus de concert pour la Science. J'apprécierais d'avoir l'honneur de participer à tes travaux enivrés de savoir.

Pourfendeur à tout crin des « zozos », Mathieu, tu es le personnage le plus totalitaire qu'il m'ait été donné de rencontrer, tu as un désir de contrôle et d'attention fascinant et non paradoxalement (pour peu qu'on ait connaissance de l'essai sur le don de Marcel Mauss), tu es quelqu'un d'une générosité, d'une serviabilité, d'une attention, d'une gentillesse, d'une dévotion exacerbées. D'une (France) culture remarquable, ayant une voix sur tout, travailleur tel Stakhanov, portant un culte à l'autorité, tu es l'homme de la caricature, de toutes les moqueries et raisonnements insupportables. Je crois qu'il faut donc que je sois remercié pour avoir supporté cet histrion durant deux ans. Merci à moi :)

Mais la thèse déborde aussi sur la vie personnelle. Mélanie, ma chérie, tu as souffert, de ces soirs où je rentrais tard, où tu m'attendais en contemplant notre chat Carou se suffire de bolduc. Et moi de rentrer, trop fatigué pour discuter de ces choses qui font la vie. Merci pour ta patience face au poids des silences, des tâtonnements et des déceptions. Merci beaucoup d'avoir relu ma thèse. Tu le sais bien toi aussi, la recherche est faite de passion, de grandes et de petites victoires mais toujours de doutes, il faut être de bonne constitution pour soutenir tout cela. Malgré des moments difficiles, nous nous aimons, je t'aime.

Et pour tous ceux qui ne pourront avoir un paragraphe dédié, sachez que vous avez tous eu une importance pour moi, Valentin j'ai été très heureux de t'accompagner dans ton stage de M2, Mathieu Gachet merci pour ton aide avec Snakemake et PanGBank ; David Roche pour les compliments esthétiques sur les graphes ; Mylène pour ta joie de vivre, et de m'avoir appris le mot japonais

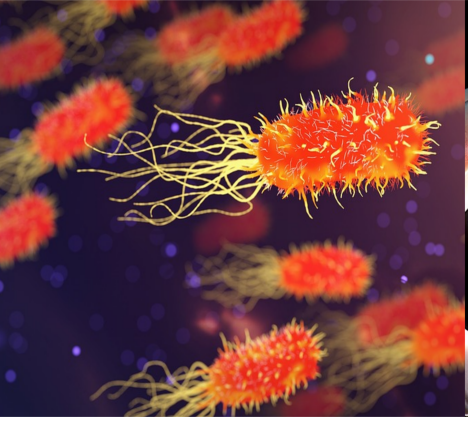
« kawaii », qui te sied d'ailleurs ; Johan pour les discussions, bon courage pour ta thèse ; Jordan, Marine, Stéphanie et Zoé pour la gourmandise, miam à vous ; Kévin, félicitations pour ta thèse, j'espère te revoir ; qu'on est modeste en tant que bioinformaticien face à quelqu'un comme Guilhem qui soigne au quotidien en plus de son travail de recherche, que d'estime je te porte ; Jonathan pour tous tes conseils au début, ce me fut fort utile. ; Laura pour ta gentillesse, ta joie de vivre ; Alexandra pour tes conseils scientifiques.

Mes remerciements à tous de m'avoir laisser parler lorsque sous la forme d'un monologue, j'ai pu partager mes passions pour quelques disciplines parfois trop peu enseignées par l'école française, à savoir les sciences politiques, la philosophie, le droit, les sciences humaines et sociales, les chats et les sciences économiques. Bien sûr j'aurais aimé vous en dire davantage mais je me suis souvent silencié pour ne pas vous assommer de mes névroses et puis aussi car je préfère me taire plutôt que de dire des choses imprécises. Je présente mes excuses à ceux qui ont reçu de longs mails de commentaires de ma part sur des sujets divers. Je prépare un essai résumant quelques travaux de recherches personnelles sur ces sujets. J'offre avec plaisir un exemplaire à toute personne citée dans ces remerciements.

Je rends hommage à Eduardo Rocha, à Pierre Peterlongo, à Nicolas Pons d'avoir participé à mes comités de thèse, pour les avis, les conseils, les critiques et les suggestions.

Merci à l'école doctorale « Structure et dynamique des systèmes vivants », à Pierre Capy, à Florence Gonnet, bien à vous.

Au Genoscope, j'ai fait de belles rencontres, le mot qui me revient est « impressionnant », mon esprit est impressionné, pareil à la plaque d'iodure d'argent du daguerréotype.





Introduction, contexte et enjeux

Insaisissables mais vastement présents partout dans nos environnements, dans tous les milieux et d'une capacité d'adaptation sans égale, les microorganismes fascinent, inquiètent et intéressent du microbiologiste au médecin en passant par les industriels et même désormais jusqu'aux informaticiens. Les microbiologistes et les médecins, cela est entendu, imaginent concurremment diverses stratégies, tout à la fois pour les éliminer lorsqu'ils sont pathogènes tout en leur permettant pour la majorité de prospérer harmonieusement dans divers environnements dont nos intestins. De même pour les industriels, car nous imaginons sans mal leur importance dans la fermentation et la maturation d'aliments comme les fromages, dans le retraitement des eaux et bien entendu dans le contrôle de l'hygiène agroalimentaire. Pensons aussi aux espoirs qu'offrent la décontamination des sols ou bien la production de biocarburants de nouvelle génération. Mais les informaticiens, eux, en quoi sont-ils concernés ? Et bien ils le sont car la microbiologie, comme la quasi-totalité des disciplines scientifiques désormais, s'appuie sur des méthodes informatiques, des algorithmes, des équations, des calculs pour analyser les données obtenues à partir des échantillons microbiens. Ces problèmes, ces questions, ces énigmes, sont spécialement l'objet de la discipline scientifique qu'est la bioinformatique.

L'étude de ces organismes microscopiques, et particulièrement des bactéries, pose sans cesse des difficultés au bioinformaticien ; à commencer par la diversité de leur génome. Car bien qu'ils soient généralement plus petits que ceux des organismes eucaryotes, leurs chromosomes connaissent une évolution très rapide et sont maintenant connus pour être particulièrement mobiles et plastiques. La compréhension des phénomènes qui en résultent, mais aussi, qui participent de cette variété génomique est un enjeu majeur du fait des très nombreuses applications, notamment médicales, que cela engage. Que ce soit la prédiction de la résistance aux antibiotiques, à l'identification des facteurs épidémiologiques de virulence, à la défense contre les menaces bioterroristes, tous ces sujets sont aujourd'hui cruciaux. Enfin, au-delà de ces applications, du fait de leur rapidité de croissance et les facilités techniques et éthiques à les modifier, les cellules du règne bactérien demeurent et demeureront assurément les modèles d'étude incontournables pour la compréhension d'un bon nombre des mécanismes fondamentaux de la biologie.

Jusqu'au début des années 2000, l'accès aux génomes complets d'une bactérie était particulièrement laborieux et onéreux de sorte que le travail de séquençage, c'est-à-dire de lecture de l'information génomique, se limitait le plus souvent à la caractérisation d'un unique génome de référence par espèce. Au mieux, cela était éventuellement complété par quelques portions supplémentaires d'autres génomes de l'espèce. Ce travail a permis d'établir les premiers liens de parenté entre les différentes espèces sur des bases moléculaires a contrario des approches observationnelles pratiquées notamment au microscope depuis des décennies. On le conçoit, avec l'arrivée de ces génomes, de nombreuses questions se sont posées aux bioinformaticiens. Où sont les gènes dans ces grands textes formés de millions de A-C-T-G sans ponctuation ? Et après les avoir identifiés que veulent-ils dire ? Ces gènes sont-ils les mêmes entre toutes les bactéries ? Comment les comparer ? Ainsi naquit la génomique comparée comme discipline de recherche.

Mais ce n'est pas tout. Depuis près de 15 ans, le travail de recherche des microbiologistes a été considérablement modifié par la diminution considérable des coûts de séquençage permettant un accès facilité à ces appareillages dit « à haut débit » dans les laboratoires de microbiologie. Il en résulte une explosion du nombre de projets ayant pour objet de générer des collections de génomes. On constate par conséquent qu'un nombre de plus en plus important de génomes par espèce est maintenant accessible publiquement dans les banques de référence. Ce nombre grimpe même jusqu'à plusieurs dizaines de milliers de génomes pour certaines espèces très étudiées comme *Escherichia coli* et le phénomène, loin d'approcher une phase de stabilisation, semble s'accélérer chaque année. Les informations qui avaient déjà été établies sur les premières références étudiées pour chaque espèce ont ainsi pu être confrontées aux nouveaux génomes. D'autre part, l'effort de recherche s'est aussi porté sur la caractérisation de l'origine et sur la dynamique des parties plus spécifiques de ces nouveaux génomes afin de comprendre les mécanismes du surgissement de nouvelles particularités biologiques comme l'adaptation à de nouveaux environnements ou l'émergence d'épidémies.

Pourtant, malgré la création de plateformes de recherche dédiées, à l'image de la plateforme MicroScope (VALLENET et al., 2019) du *Laboratoire d'Analyses Bioinformatiques pour la Genomique et le Métabolisme* (LABGeM, Genoscope) au sein de duquel ce doctorat s'est effectué, et l'amélioration notable des capacités de parallélisation des calculs bioinformatiques, il reste que le traitement de ces données massives pose de plus en plus de difficultés. Très concrètement, il n'est donc plus possible de continuer à analyser ces génomes avec les méthodes de génomique comparée habituelles. Dès lors, de nouvelles approches sont requises pour organiser et hiérarchiser ces informations afin de les rendre biologiquement exploitables, c'est-à-dire, à même de fournir des hypothèses explicatives des phénomènes biologiques.

On pourrait se poser la question de l'intérêt de collectionner une telle quantité de données. On constate en effet que l'information génomique est notoirement redondante, de qualité très variable et ce malgré l'amélioration des méthodes de séquençage. On pourrait se dire qu'après tout, Albert Einstein n'a pas eu besoin de millions de clichés astronomiques pour construire sa théorie de la relativité générale. Pourtant, ce qu'on observe avec les procaryotes c'est l'existence d'une biodiversité génomique considérable qui nécessite d'être répertoriée pour être étudiée. En effet, comprendre comment les gènes apparaissent ou disparaissent, les fonctions associées,

les mécanismes moléculaires en jeu, là est la clé de compréhension d'un très grand nombre de phénomènes biologiques y compris en interaction avec des animaux et des plantes. Certes, cette quantité de données est importante, mais plutôt que de la considérer comme une contrainte, il faut le voir comme un moyen de penser de nouvelles solutions de stockage et de représentation de données afin de mieux en extraire l'information.

On pourrait aussi objecter qu'à cause du fait que la communauté scientifique s'intéresse majoritairement à quelques espèces d'intérêt mais également des difficultés de culture en laboratoire inhérentes à certaines d'entre elles, il s'ensuit que la quantité de génomes dans les bases de données est très inégalement répartie. Indéniablement, le premier point est exact mais on découvre aussi via des approches génomiques de nombreuses espèces jusqu'alors inconnues. En effet, nous constatons l'émergence d'une nouvelle approche méthodologique permettant d'accéder au génome d'organismes difficiles à isoler et à cultiver : la métagénomique. Ici le but n'est plus d'isoler une souche puis d'en obtenir le génome, mais bien de séquencer l'ensemble de l'information génomique d'un échantillon pour pouvoir reconstituer informatiquement les séquences qui y étaient présentes. Cette nouveauté technique révolutionne la génomique et donne accès là encore à des milliers de nouveaux gènes et génomes reconstruits à partir d'espèces non cultivées.

C'est avec ces problèmes à l'esprit qu'au milieu de la décennie 2000, émerge une approche originale pour organiser l'information génique présente dans une espèce. À l'inverse d'une comparaison « génome à génome » ou bien de « génome à génome de référence », le concept de **pangénome** agrège l'ensemble de l'information génique retrouvée dans une espèce. Ainsi, il est possible de modéliser la diversité présente dans une espèce à travers l'étude de l'occurrence des gènes dans le pangénome. L'objectif est d'estimer la variabilité génomique contenue dans ces génomes, la modéliser, la stocker, la représenter. Dans ce pangénome, on distinguera d'une part le génome cœur, c'est-à-dire les gènes systématiquement présents dans les souches de l'espèce et d'autre part les gènes accessoires plus ou moins conservés dans les génomes, donc susceptibles d'expliquer des particularités phénotypiques.

Cependant, en procédant à ce type d'analyse pangénomique, on constate empiriquement que le concept de génome cœur est limitant car des gènes, bien que fonctionnellement indispensables, peuvent être exceptionnellement absents des génomes, soit pour des raisons techniques, soit pour des raisons biologiques. Dès lors, il peut apparaître une diminution du génome cœur à mesure que l'on analyse de nouveaux génomes. Une autre limitation de cette distinction entre le génome cœur et le génome accessoire est de ne pas rendre compte de la grande différence d'occurrence des gènes dans le génome accessoire, attendu que certains gènes sont particulièrement rares et que d'autres possèdent une fréquence de présence intermédiaire. Pour pourvoir à cette objection, il a notamment été proposé une division du pangénome en trois parties : le génome cœur, le génome coquille (gènes souvent présents) et le génome nuage (gènes rarement présents). Pour autant, même avec cette délimitation conceptuelle, il reste difficile d'établir des seuils de fréquence adaptés à toutes les espèces pour le partitionnement du pangénome. De plus, on pourrait aussi penser que la présence et l'absence de gènes désignent des schémas spécifiques qui pourraient être employés pour assigner les gènes à ces parties.

Néanmoins, quand bien même le pangénome serait correctement partitionné, il reste que les études pangénomiques laissent généralement une partie importante de l'information génomique inexploitée. En effet, en considérant les gènes comme des éléments isolés, ces approches ne tiennent pas compte de l'organisation chromosomique des gènes dans les différents génomes de l'espèce. Cette information est donc perdue. Pourtant, il est déjà établi que l'emplacement des gènes sur le génome est capital, par exemple pour détecter les régions susceptibles d'être échangées entre les génomes au sein des îlots génomiques.

Le travail de doctorat dont vous entreprenez de lire le manuscrit de thèse, a visé au développement d'une méthode combinant à la fois le voisinage chromosomique des gènes et leur schéma de présence, afin de partitionner à l'aide d'une approche statistique le pangénome des espèces procaryotes. On verra notamment que cette approche permet de détecter de façon convaincante les régions et non plus seulement les gènes variables dans les génomes. De plus, cette approche basée sur des schémas de présence permet également d'analyser les génomes obtenus à partir du concept précédemment évoqué de métagénome.

Dans de ce manuscrit de thèse, la partie I page 21, passera en revue la littérature pertinente pour la pangénomique. Cela commencera par un chapitre rappelant les fondamentaux de la génomique et de l'évolution des procaryotes (chapitre 1), suivi de trois autres décrivant l'histoire du concept de pangénome (chapitre 2) ainsi que les propositions de modélisation de celui-ci (chapitre 3) et des réflexions sur les conséquences biologiques résultant de son existence (chapitre 4), avant de parcourir l'abondante littérature d'outils, de méthodes et d'approches utilisés pour le construire et le représenter (chapitre 5).

Enfin viendra le cœur de mon travail, la méthode PPanGGOLiN abordée à la partie II page 99, vous en comprendrez les motivations à travers le chapitre 6 avant que vous en soit exposés les détails statistiques et techniques ainsi que les différentes pistes explorées au cours de son développement (chapitre 7 et 8). Enfin, vous pourrez découvrir les résultats obtenus en appliquant cette méthode sur 439 espèces procaryotes à travers le chapitre 9.

La partie III page 181, vous montrera que PPanGGOLiN jettent les bases de tout un écosystème de fonctionnalités abordées à travers deux chapitres. Tout d'abord une banque de données de pangénomes dont la communauté manque aujourd'hui, appelée PanGBank (chapitre 10) avant de présenter une méthode appelée panRGP permettant de détecter les régions de plasticité présentes dans les génomes (chapitre 11).

Comme il se doit, un retour critique sera offert sur le travail présenté pour en signaler certaines limites et proposer au lecteur des améliorations possibles. Enfin, après avoir fait le bilan des contributions, nous terminerons par la présentation de quelques perspectives en vue d'étendre l'écosystème PPanGGOLiN comme une approche de détection de chemins associés à des variables externes. Mais tout d'abord, si vous le permettez, commençons ensemble par un retour sur les fondamentaux de la microbiologie vue sous le prisme de l'évolution des génomes procaryotes [1].

1. Avant d'entamer la lecture de ce manuscrit, je conseille également de visionner au préalable un enregistrement de la soutenance de thèse (dont le lien est disponible en annexe à la section A.1 page 207).

Partie

Du génome au pangénome

1 Biologie et évolution des procaryotes 21
1.1 Bases moléculaires de l'évolution
1.2 Structure des génomes procaryotes
1.3 Classification des procaryotes
1.4 Les trois modes d'évolution des procaryotes

2 Aux origines de la pangénomique 33
2.1 Les génomes procaryotes, clones ou chimères ?
2.2 Qu'est-ce qu'une espèce procaryote ?
2.3 Les premières analyses de pangénomes

3 Modélisation des pangénomes 45
3.1 Modélisation par décroissance exponentielle
3.2 Modélisation par la loi de Heaps
3.3 Modélisation d'un pangénome par modèles de mélange binomiaux
3.4 La trichotomie : cœur, coquille et nuage
3.5 Le modèle infiniment plein de gènes (IMG) corrobore la trichotomie
3.6 Rejet de l'hypothèse d'une évolution neutre
3.7 Conclusion sur les modèles proposés

4 Le pangénome est-il adaptatif ou non ? 71
4.1 L'apport de la génétique des populations
4.2 L'acquisition de nouveaux gènes est délétère en soi
4.3 Pourquoi donc les procaryotes ont-ils des pangénomes ?

5 Les outils utilisés en pangénomique 79
5.1 Approches basées sur les familles de gènes
5.2 Approches à l'échelle de la séquence
5.3 Les approches d'indexation de génomes basées sur une référence
5.4 Les graphes de séquence
5.5 Les graphes de De Bruijn



Chapitre 1 : Biologie et évolution des procaryotes

Même si on trouve dans la pensée philosophique des théories sur l'existence d'animaux invisibles à l'œil nu dès l'Antiquité [1], on peut faire remonter la naissance de la microbiologie à l'observation des premiers microorganismes par le néerlandais Antonie van Leeuwenhoek en 1676. Ce dernier, drapier de son état, conçoit des lentilles pour scruter la qualité des mailles des étoffes, quand en 1676, il observa pour la première fois, à l'aide d'une de ses inventions (illustré sur la figure 1.1) d'étranges animaux qu'il nomma en latin *animalculum* et qu'il entreprit de dessiner (exemple sur la figure 1.2).



FIGURE 1.1 – Photo personnelle prise au musée des Sciences de l'université d'Utrecht du premier microscope ayant permis d'observer des *animalculum*. Pour avoir utilisé une réplique, il fatigue les yeux !

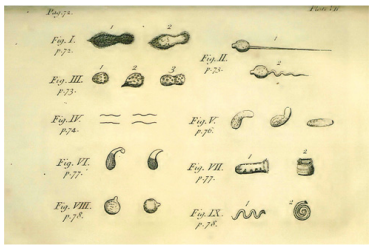


FIGURE 1.2 – Croquis de Antonie van Leeuwenhoek (Source : Université de Delft). « Fig IV » est probablement le premier dessin d'une bactérie connue (potentiellement de la famille des Spirillaceae).

Quatre siècles plus tard, la microbiologie et en particulier l'étude des procaryotes, c'est-à-dire des organismes unicellulaires sans noyaux, a bien évolué et des milliers de souches bactériennes ont été recensées et caractérisées notamment du point de vue de leur morphologie, de leur mode de vie et de leur génome. Au-delà, la connaissance des procaryotes a notamment servi de socle aux travaux ayant permis de découvrir les fondamentaux moléculaires du vivant. On peut penser en particulier aux expériences de Jacques Monod et François Jacob sur l'opéron lactose de la bactérie *Escherichia coli*, dont les conclusions sont notamment à l'origine du dogme central de la biologie

1. Notamment par le philosophe romain Marcus Terentius Varro (dit Varron), le philosophe turc Muhammad Shams al-Din bin Hamzah (dit Akshamsaddin) et dans la religion indienne jaïnique.

moléculaire [1]. Jacques Monod et Francois Jacob ayant même été jusqu'à reprendre à leur compte l'aphorisme suivant : « Tout ce qui est vrai pour la bactérie *Escherichia coli* est vrai pour l'éléphant » (MONOD et al., 1961).

De par leur génome à la fois généralement plus petit que les eucaryotes mais très diversifié, l'étude des procaryotes a pleinement profité du développement de la génomique. Preuve de l'intérêt que leur porte la communauté scientifique, à ce jour, sur les 94 946 espèces dont l'ADN a été séquencé [2], 59 415 sont des espèces procaryotes (58 153 bactéries et 1 262 archées). Il en est de même en termes de nombre de génomes puisque, toujours à la même date, les génomes procaryotes représentaient 212 278 entrées sur un total de 287 407 génomes répertoriés.

Dans ce chapitre, nous passerons en revue des notions importantes de la biologie des procaryotes qu'il faut avoir à l'esprit pour bien saisir l'intérêt de l'approche pangénomique pour l'étude de leur évolution.

1.1 Bases moléculaires de l'évolution

À sa naissance, notamment sur la base des expériences de Gregor Mendel sur le pois, la génétique n'avait pas de support moléculaire connu pour la transmission de l'information du vivant. La théorie de l'évolution se basait donc sur des inférences recoupées qui dessinaient une théorie crédible expliquant la manière dont les espèces accumulent des variations, apparaissent et disparaissent mais sans être en mesure d'invoquer une mécanistique génétique et physiologique appuyant cette théorie [3]. Pour pallier au défaut d'un support informationnel à sa théorie, Charles Darwin propose d'hypothétiques « gemmules » (aussi appelées pangènes) dans sa théorie dite de la pangenèse. Ces gemmules présentes dans le corps des individus stockeraient à la fois l'information propre de l'espèce mais aussi les caractères acquis par l'individu. Au moment de la reproduction, ces pangènes, en capacité de migrer, se concentreraient dans les gamètes ce qui expliquerait la transmission de ces deux types de traits aux descendants. Cette théorie fut réfutée par Francis Galton et August Weismann tandis que les travaux conjoints notamment de Theodor Boveri, Walter Sutton, Eleanor Carothers et Thomas Morgan nous apprîmes au début du XXème siècle que seules les variations représentées dans les gamètes sont transmissibles aux descendants à travers les chromosomes.

Pourtant la nature moléculaire du chromosome resta encore inconnue. Pour aller plus loin, il fallut le concours du monde bactérien. En 1928, Frederick Griffith constate qu'en injectant à des souris une dose de pneumocoques (*Streptococcus pneumoniae*) de souche R (« rough », c'est-à-dire sans capsule) connus comme non-pathogènes, couplée à une infime dose inactive (bactéries tuées par chauffage) de la souche S (« smooth », souche pathogène possédant une capsule la protégeant du

1. D'après CRICK, 1970 : « The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid ».

2. D'après la banque de séquences NCBI RefSeq consultée le 09 septembre 2019

3. D'ailleurs des historiens s'accordent pour dire que bien que les manuscrits de Mendel étaient présents dans l'immense bibliothèque de Darwin, celui-ci n'en aurait probablement pas compris l'importance du fait du style d'écriture mathématique de Mendel (GALTON, 2009). Comme disait Darwin : « Mathematics in biology was like a scalpel in a carpenter's shop, there was no use for it ».

système immunitaire de l'hôte), les souris contractent une pneumonie, et ce bien que la bactérie pathogène soit morte. De plus, on observe des pneumocoques de souches R et S en grand nombre dans leur sang. Griffith expliquera ce phénomène par le fait que l'information nécessaire à la formation de la capsule protectrice qui était présente dans la bactérie morte a été transférée dans la bactérie vivante, devenant ainsi pathogène. Ainsi, Griffith montra que des échanges d'information pouvaient s'effectuer horizontalement entre des souches bactériennes, ce qu'il appela un « transforming principle ». Reste une question, quel est le support moléculaire médiant cette transformation ? Deux candidats sont en lice, la macromolécule d'ADN (Acide DésoxyriboNucléique) et les protéines. Pour réponse à cette question, Oswald Avery, Colin MacLeod and Maclyn McCarty modifient l'expérience en incubant cette fois les bactéries de la souche S avec, soit différents extraits de protéines de la souche R (l'ADN étant détruit par l'ajout d'une enzyme appelée nucléase), soit l'ADN pur uniquement (les protéines étant détruites par l'ajout d'enzymes appelées protéases). Ils constateront en laboratoire l'apparition de colonies bactériennes ayant le phénotype de la souche S uniquement dans le cas de l'ajout de l'ADN seul (AVERY et al., 1944). En 1952, Hershey et Chase confirmeront le rôle de l'ADN dans la transmission de l'information génétique en réalisant des expériences sur des phages (HERSHEY et al., 1952). Au moment de cette découverte, il était déjà connu que l'ADN était un polymère constitué de quatre nucléotides (Adénine, Thymine, Cytosine, Guanine [1]), quand, un an plus tard, sa structure chimique précise fut établie par le travail conjoint de Francis Crick, James Watson, Maurice Wilkins et Rosalind Franklin [2] (WATSON et al., 1953). L'ADN, molécule désormais iconique, se caractérise par une structure chimique associant deux brins appareillés de façon complémentaire assurant par là même une redondance de toute l'information génétique. Cet attribut est capital pour bien comprendre la grande fiabilité que présente le vivant, dans la conservation de son information génétique face au nombre de circonstances susceptibles d'en altérer la chimie (UV, radioactivité, molécules mutagènes, radicaux libres). De plus, ces deux brins appariés s'enroulent pour former une double-hélice, assurant ainsi une grande compacité à cette structure. Cette compacité est renforcée par la grande flexibilité du polymère permettant un très grand niveau de condensation de la molécule jusqu'à former des structures filamenteuses, les fameux chromosomes.

Malgré sa grande fiabilité et de nombreux mécanismes de réparation, il est inévitable que, sur un grand nombre de cellules, des mutations (substitutions, insertions ou délétions) apparaissent. C'est en particulier le cas au moment de la division cellulaire lorsque des erreurs sont introduites durant une recopie des génomes. Ces dernières peuvent avoir des conséquences neutres, bénéfiques ou délétères à l'organisme, que ce soit à court terme dans un environnement donné mais aussi à plus long terme pour l'espèce. Ces errements jalonnent ainsi l'histoire des destins vivants, en s'accumulant ils bricolent les génomes de sorte que ce mouvement, dans sa diaprure, constitue la matrice de l'évolution.

1. Les mots Adénine, Thymine et Guanine proviennent de l'échantillon duquel ces molécules furent isolées. « Aden » ($\alpha\delta\eta\nu$) signifiant gland en grec, « thymine » provenant de « thymus » et « guanine » provenant de « guano ». La molécule de cytosine bien qu'elle aussi soit extraite du thymus porte ce nom car elle est très présente dans le noyau des cellules.

2. Cette dernière n'ayant pas obtenu le prix Nobel car elle décéda d'un cancer vraisemblablement provoqué par l'irradiation aux rayons X qu'elle manipulait pour obtenir les clichés cristallographiques à l'origine de la découverte de la structure des molécules, et notamment de l'ADN.

1.2 Structure des génomes procaryotes

Il est difficile de décrire exhaustivement la grande diversité de structures des génomes procaryotes attendu qu'une certaine multiplicité de formes peut y être rapportée. Toutefois, certaines caractéristiques semblent relativement répandues.

1.2.1 Structure d'un chromosome procaryote

Comme précédemment indiqué, le ou les chromosomes bactériens sont constitués de deux brins d'ADN complémentaires et appareillés. Pourtant, vu à l'échelle de la cellule grâce à un microscope, le chromosome bactérien est une masse compacte ressemblant à un « sac de nœuds » appelée chromatine. De surcroît, si nous démêlions cette chromatine, nous constaterions que dans le cas des procaryotes, ce double brin d'ADN forme en réalité une molécule circulaire. Il faut toutefois noter qu'il existe des exceptions où le chromosome est linéaire comme dans l'espèce *Borrelia burgdorferi* par exemple.

Pour information, les chromosomes possèdent un nombre de paires de nucléotides (aussi appelés paires de bases ou bases, abrégé : pb ou b) s'échelonnant d'environ 160Kb (pour la bactérie *Carsonella ruddii* possédant $\approx$ 224 gènes, KATSIR et al., 2018) jusqu'à 15Mb (pour la bactérie *Sorangium cellulosum* possédant presque $\approx$ 9367 gènes, SCHNEIKER et al., 2007). Quand on examine la distribution de la taille de tous les génomes procaryotes, il existe deux pics de densité, un à environ 2Mb et un autre plus petit à environ 5Mb (KOONIN et al., 2008). S'ajoutent facultativement à ce ou ces chromosomes de plus petites molécules d'ADN circulaires ou non (généralement de 1Kb à 200Kb) et en nombre variable, appelées plasmides. Quand ces plasmides sont de taille importante, on parle de mégaplasmide. Les plasmides comme tous les chromosomes sont des molécules possédant la capacité d'être répliquées et possèdent une origine de réplication (souvent abrégée « ori »). On les regroupe ainsi généralement sous le terme de réplicons. Plus rarement utilisé, le terme chromide désigne un réplicon dont la composition fréquentielle (fréquence en paires de bases Guanine-Cytosine, abrégé GC) est proche de celle du chromosome principal et dont la présence est obligatoire pour la croissance de la bactérie.

1.2.2 Les gènes procaryotes

Le gène est l'unité d'information fondamentale du chromosome. Concrètement, il correspond à une portion d'ADN en mesure d'être transcrite sous forme d'ARN (Acide RiboNucléoique) afin que son information génétique puisse induire une fonction biologique. Généralement la transcription est suivie d'une étape de traduction où l'ARN est traduit en protéine par une macromolécule appelée ribosome selon un code de conversion de l'information appelé code génétique. C'est ensuite la protéine ainsi obtenue qui assurera une fonction biologique. Dans ce cas, le gène est dit codant et l'ARN est dit messager. Dans le cas où l'ARN n'est pas traduit en protéine, c'est l'ARN directement qui aura une fonction biologique.

L'expression des gènes conduisant à des fonctions biologiques, elle doit être régulée et coordonnée afin de permettre la survie de l'être vivant. Plusieurs mécanismes participent de la régulation de l'expression des gènes. Tout d'abord pour être exprimé un gène doit être accessible à un ensemble de molécules constituant l'ARN polymérase qui se fixe en amont d'un gène en reconnaissant de

courtes séquences très particulières appelées « promoteurs ». Les promoteurs sont des séquences ayant une plus ou moins bonne adéquation avec l'ARN polymérase, ce qui en conséquence détermine la probabilité de fixation d'une ARN polymérase et donc la fréquence de la transcription des gènes.

Il faut aussi noter que les gènes sont souvent transcrits ensemble sur la même molécule d'ARN afin notamment d'obtenir dans le même temps plusieurs protéines agissant de façon coordonnée. On appelle l'ensemble de ces gènes colocalisés et coexprimés : un opéron. Les opérons possèdent très souvent un emplacement, proche du promoteur commun à tous ces gènes, appelé opérateur chargé d'agir comme un interrupteur d'expression pour tous les gènes de l'opéron. En effet, des molécules peuvent venir se fixer sur cet opérateur afin d'en activer ou d'en réprimer l'expression.

Dans un chromosome procaryote, un gène possède une taille moyenne de 1Kb tandis que les distances intergéniques (contenant notamment les promoteurs et les opérateurs) sont d'environ 100b, néanmoins entre les gènes situés dans les opérons, cette distance est proche de zéro ou même négative car les gènes peuvent parfois être chevauchants (KOONIN et al., 2008).

1.3 Classification des procaryotes

Depuis l'exceptionnel travail conduit par Aristote notamment dans son livre « Histoire des animaux », l'une des tâches principales de la biologie est de catégoriser les formes vivantes. Avant l'apparition de la génomique, le typage et l'identification des souches bactériennes étaient généralement effectués sur la base de quatre types de critères.

1.3.1 Critères morphologiques

Deux des critères évidents de classement des procaryotes sont leur forme et leur éventuel mode de regroupement (on peut voir quelques exemples sur la figure 1.3 page suivante). D'un point de vue morphologique, la présence ou l'absence de flagelle est également un critère important du fait de son lien avec le type de mobilité de la cellule.

Il est également très courant de classifier les procaryotes en fonction de la nature de leur paroi. En particulier, l'une des classifications les plus célèbres se base sur la coloration proposée par Hans Christian Gram. Cette classification distingue les cellules selon la nature de leur paroi. Le protocole de Gram consiste dans l'ajout de plusieurs colorants sur des cellules procaryotes, suivi de leur observation au microscope. Si les cellules possèdent une paroi épaisse faite de peptidoglycanes, elles seront colorées en violet. À l'inverse, les cellules possédant une fine paroi de peptidoglycanes seront, elles, colorées en rose. Certains procaryotes ne peuvent cependant pas être colorés car ils ne possèdent pas de paroi, comme les Mollicutes par exemple.

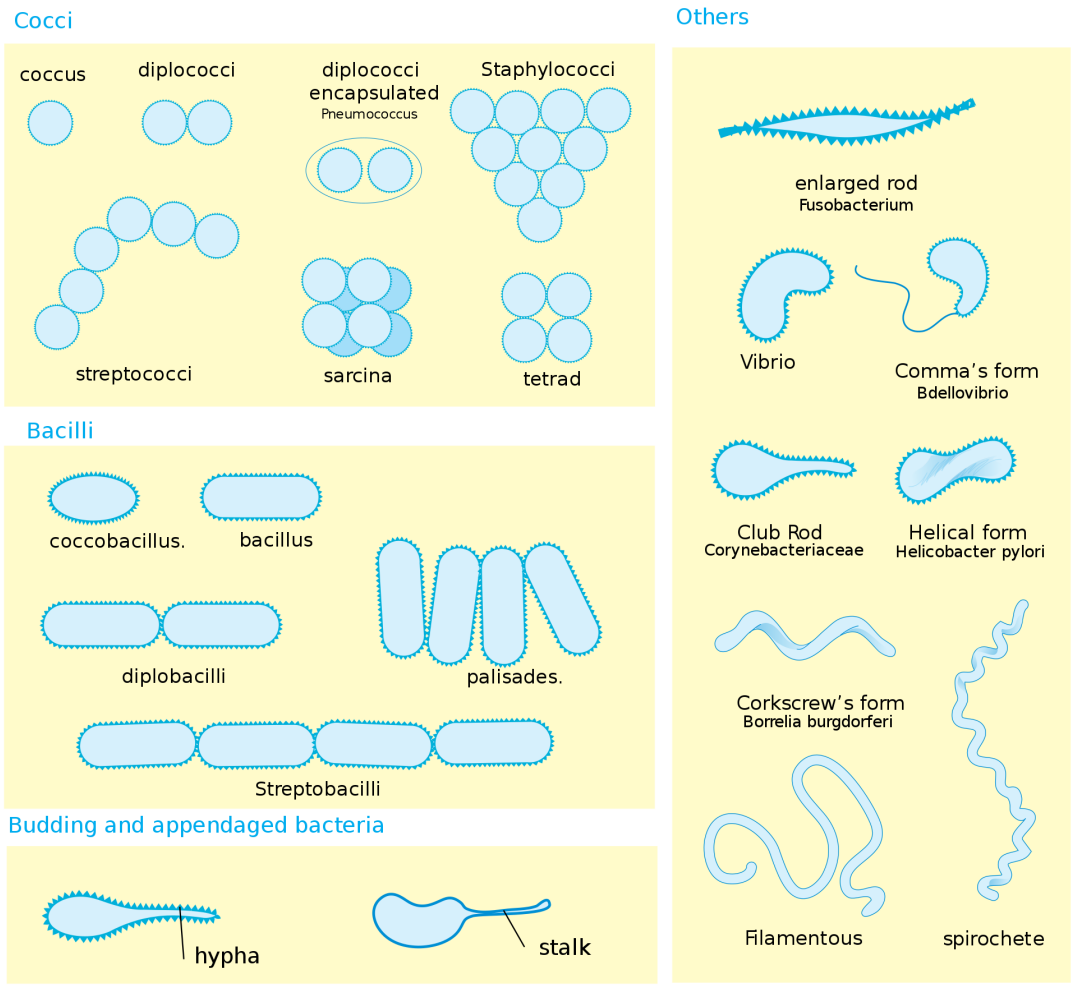


FIGURE 1.3 – Figure reproduite depuis WIKIPEDIA, 2019b illustrant une partie de la multiplicité de formes que peuvent revêtir les organismes procaryotes.

1.3.2 Critères physiologiques

Les procaryotes peuvent également être classifiés en fonction de leurs types métaboliques et physiologiques :

— Leur mode de vie aérobie ou anaérobie.
— Les différentes sources d'énergie qu'ils utilisent.
— Les sources de carbone et d'azote qu'ils emploient.
— Le type de support qu'ils préfèrent (support solide ou liquide).

Toutes ces caractéristiques définissent des milieux de culture. Enfin, les procaryotes peuvent également être catégorisés par leur capacité à produire certaines molécules comme des produits de fermentation ou des métabolites secondaires.

1.3.3 Critères de pathogénicité

Pour des justifications cliniques, les espèces procaryotes et plus particulièrement les espèces bactériennes sont également séparées sur la base des maladies qu'elles sont susceptibles de causer. Bien que d'une grande utilité pour les soignants, ce critère présente certaines limites, à commencer par le fait que seul un faible nombre d'espèces procaryotes sont potentiellement pathogènes. De plus, ce critère de regroupement ne recoupe pas toujours les autres critères de classification comme ceux mentionnés aux sections précédentes et suivantes. On peut citer l'exemple de *Bacillus cereus* et *Bacillus anthracis* qui sont catégorisées dans deux espèces différentes bien qu'elles soient très proches sur la base d'autres critères. En effet, ces deux espèces sont très proches, notamment sur le plan génétique, mais elles ont été distinguées car *B. anthracis* possède, pour sa part, deux plasmides (*pX*01 and *pX*02) rendant la bactérie capable de développer la maladie du charbon (« anthrax » en anglais [1]) chez son hôte (HELGASON et al., 2000).

1.3.4 Critères de sérogroupage

Dans le cas des bactéries possédant une capsule c'est-à-dire une enveloppe (le plus souvent composée de polysaccharides) entourant la paroi bactérienne, les souches d'une espèce sont généralement typées par les molécules qui la composent. En effet, les capsules ont une forte utilité clinique et épidémiologique car elles sont en partie responsables de leur immunogénicité, de leur virulence et de leur résistance aux composés antimicrobiens. La diversité des types de capsule (sérotype capsulaire) est donc une information importante pour classifier la diversité bactérienne, notamment au sein d'une même espèce (diversité intraspécifique). Le sérotypage s'effectue par l'observation *in vitro* grâce à un microscope d'une éventuelle agglutination des cellules procaryotes lors de l'ajout de sérum extrait d'un individu préalablement infecté. Si ce sérum contient des anticorps spécifiques du sérotype de la cellule observée, ceux-ci se colleront aux cellules provoquant ainsi une agglutination des cellules. On en déduit ainsi que la cellule ayant infecté le patient, dont on a extrait le sérum, a donc le même sérotype que les autres cellules étudiées. À l'inverse, si ce sérum ne contient pas d'anticorps spécifiques du sérotype, il n'y aura pas d'agglutination observable indiquant donc que les cellules procaryotes sont d'un sérotype différent. Ainsi, il est possible de classer les cellules procaryotes par les anticorps qu'elles génèrent du fait de la réaction immunitaire de leur hôte.

1.3.5 Critères génomiques

Il est également possible de regrouper les procaryotes par une approche génomique. Historiquement, celle-ci est basée sur la comparaison de portions d'ADN à des loci conservés (souvent situés au niveau de gènes de ménage [2]) entre deux génomes. Ces portions sont similaires mais pas nécessairement identiques. Concrètement, cela nécessite de recopier les séquences d'ADN cibles dans le but d'amplifier leur signal via une méthode appelée PCR (« Polymerase Chain Reaction », MULLIS et al., 1986). Une fois ces loci amplifiés, ces portions d'ADN sont lues via une méthode

1. À noter que cette maladie ne doit pas être confondue avec l'anthrax en français qui désigne une staphylococcie cutanée.
2. Gènes souvent hautement exprimés qui assurent les fonctions indispensables à la vie des cellules.

de séquençage (comme par exemple la méthode Sanger, SANGER et al., 1977). Ainsi, en comparant leur similarité par des algorithmes d'alignement de séquences (comme par exemple celui des auteurs de NEEDLEMAN et al., 1970), il est possible d'extrapoler ces comparaisons de locus à locus à la similarité entre les génomes bactériens en entier. Une fois les similarités entre plusieurs souches d'une même espèce mesurées, il devient possible d'inférer l'évolution de l'espèce sous forme d'un arbre phylogénétique. De même, sur la base de ces résultats, les microbiologistes ont introduit différentes catégories afin d'organiser les génomes procaryotes comme par exemple les phylogroupes ou les ribotypes (basés sur la similarité de séquence pour l'ARN ribosomique 16S). Aujourd'hui, il est devenu très simple de comparer les génomes entiers de différentes souches procaryotes, ce qui a amené au développement de nouvelles métriques sur lesquelles nous reviendrons.

1.4 Les trois modes d'évolution des procaryotes

On distingue généralement trois voies évolutives pour les génomes procaryotes (comme illustré sur la figure 1.4 page suivante).

1. La voie verticale par mutations, duplications et réarrangements aléatoires.
2. La voie verticale par perte aléatoire de portions d'ADN.
3. La voie horizontale par acquisition de portions d'ADN depuis l'environnement.

Tous les êtres vivants semblent être en mesure de suivre ces trois voies évolutives, néanmoins pour beaucoup d'organismes vivants, la voie (1) semble largement dominante de sorte que les deux autres et en particulier la voie (3) sont d'importance mineure. Ainsi, la particularité des procaryotes est que l'évolution par réduction de génome et par transferts horizontaux semblent être des modes d'évolution non négligeables et même potentiellement prépondérants.

1.4.1 L'évolution verticale

Bien sûr, comme les autres formes de vie, l'évolution des procaryotes peut résulter de duplications de portions d'ADN, de réarrangements chromosomiques et de mutations locales (substitutions, insertions, délétions). Le point commun de toutes ces modifications est d'apparaître aléatoirement et de pouvoir être transmises aux cellules filles lors de la division cellulaire.

Mutations locales

Pour le cas des mutations locales, ces modifications dans le génome des cellules apparaissent à cause de deux phénomènes :

— À la faveur d'erreurs de recopie lors de la réplication du chromosome, ce qui cause principalement des substitutions.

— En cas de cassure de la molécule d'ADN (notamment due à une irradiation). Dans ce cas, les enzymes d'entretien de l'ADN reconstituent rapidement la contiguïté de la molécule d'ADN (via la voie de réparation « Non-Homologous End Joining ») mais avec un risque important de mutations (en particulier d'insertions et de délétions).

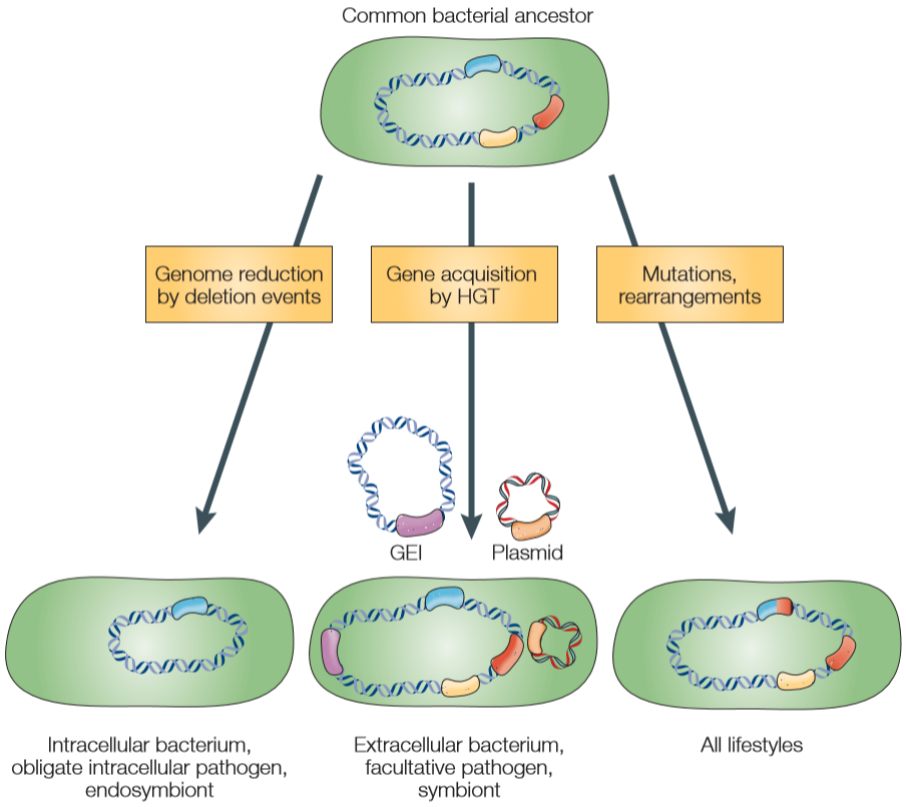


FIGURE 1.4 – Figure reproduite depuis (DOBRINDT et al., 2004) illustrant les trois types de chemins évolutifs suivis par les espèces procaryotes depuis un ancêtre commun.

Dans le cas des substitutions, si celles-ci ne sont pas létales, elles sont susceptibles d'être maintenues dans la population, ce que nous appelons des variants SNPs (Single Nuleotide Polymorphisms). Pour ce qui est des insertions et délétions, elles sont parfois en mesure d'avoir plus d'impact en décalant le cadre de lecture lors de la traduction, ce qui peut neutraliser la fonction de la totalité d'une protéine.

Variations structurales

Les réarrangements des chromosomes se produisent de façon privilégiée au niveau de régions présentant une homologie de séquence. Néanmoins, comme les recombinaisons sont des évènements ayant un impact évolutif fort, il y a un important risque que ces évènements soient délétères.

Plus fréquemment, les variations structurales des chromosomes adviennent par la duplication d'une portion d'ADN dans les génomes. Cela survient essentiellement à travers deux mécanismes :

— Par rétrotransposition : lors de la rétrotransposition d'un élément transposable, il peut y avoir transcription d'une partie de la région chromosomique à proximité de l'élément transposable. Lorsque l'ARN est rétrotranscrit au sein du génome, les segments rétrotranscrits peuvent alors être recombinés avec une autre région du génome, provoquant ainsi une duplication.

— Par glissement de la réplication (Slipped Strand Mispairing : SSM) : lors de la réplication, l'ADN polymérase commence à copier l'ADN. À un moment donné au cours du processus de réplication, la polymérase peut alors se dissocier par erreur de l'ADN. Lorsque la polymérase

se rattache au brin d'ADN, si elle se réaligne sur le brin en réplication à une position incorrecte, elle va potentiellement recopier un même segment plusieurs fois produisant ainsi une duplication. Le glissement de la réplication est notamment facilité par la présence de séquences répétées.

Ainsi par l'accumulation de ces modifications locales et structurales transmises à leur descendance, les organismes évoluent d'une manière pouvant être décrite par un arbre de descendance. C'est la raison pour laquelle on évoque la notion d'évolution verticale pour qualifier ce type d'évolution.

1.4.2 Réduction de génomes

L'évolution des génomes procaryotes peut aussi prendre la forme d'une suppression de gènes non essentiels lorsqu'un organisme s'adapte à un nouvel environnement nécessitant moins de fonctions biologiques. Cela se produit généralement lorsque des organismes adoptent un mode de vie parasitaire ou symbiotique, notamment lorsque des nutriments sont fournis par l'hôte. Ainsi, les organismes sont avantagés s'ils perdent les gènes nécessaires à la production de ces nutriments. D'ailleurs, il est souvent possible d'identifier par génomique comparée les gènes perdus depuis une espèce originelle ayant un mode de vie « en liberté » avec leurs variantes parasitaires. On peut notamment penser à l'exemple des deux espèces pathogènes : *Mycobacterium tuberculosis* et *Mycobacterium leprae*. L'espèce *M. tuberculosis* est extracellulaire et possède ≈4000 gènes tandis que l'espèce *M. leprae* n'en possède plus que ≈1600 car la bactérie est passée à l'état intercellulaire obligatoire (MALHOTRA et al., 2017). Ainsi *M. leprae* détourne une partie du métabolisme des cellules eucaryotes hôtes et peut ainsi survivre malgré la perte de ≈2400 gènes, de sorte que l'espèce présente ainsi un génome considérablement réduit.

Là encore l'évolution sera considérée comme verticale car les gènes perdus ne pourront pas être retrouvés par les descendants. Pour augmenter à niveau le répertoire génomique, il faudra alors que la cellule subisse des duplications comme évoqué précédemment ou bien qu'elle acquiert de nouveaux gènes depuis l'extérieur.

1.4.3 Les transferts horizontaux

L'évolution des génomes procaryotes peut résulter comme nous l'avons vu avec l'expérience de Griffith par l'acquisition de nouveaux gènes issus du transfert de segments d'ADN obtenus directement depuis l'environnement. Comme ces segments n'ont pas pour origine un héritage provenant d'une cellule mère, on parle alors de transferts horizontaux. Toutefois, une fois ces éléments intégrés, ils pourront être hérités par les cellules filles. Ces transferts peuvent correspondre à des portions d'ADN venant s'intégrer dans le chromosome principal ou bien des éléments périphériques comme des plasmides. Il existe trois mécanismes conduisant à des transferts horizontaux : la transformation, la transduction et la conjugaison.

La transformation

Une transformation génétique se définit par l'intégration d'un fragment d'ADN étranger dans une cellule de façon passive. Le segment d'ADN présent dans l'environnement traverse ainsi directement la paroi puis la membrane cytoplasmique des cellules avant de s'intégrer dans le chromosome procaryote. Ces segments peuvent également correspondre à des plasmides. Ce phénomène se déroule en particulier chez des cellules procaryotes dites compétentes [sic], c'est-à-dire à même de permettre à de l'ADN exogène d'entrer dans la cellule. Les cellules peuvent être naturellement compétentes mais il est aussi possible d'augmenter la fréquence des transformations en rendant compétentes des cellules en laboratoire par des chocs électriques ou par traitement avec des composés chimiques.

Enfin, il est connu que certaines bactéries peuvent notamment devenir compétentes en cas de stress environnemental ou de manque de nutriments comme par exemple chez l'espèce *Bacilius subtilis* (CLAVERYS et al., 2006). À l'inverse, cela peut aussi advenir en cas de haut niveau de disponibilité en nutriments, par exemple chez l'espèce *Acinetobacter baylyi* (NIELSEN et al., 1997). Ainsi cela permet à la bactérie d'acquérir de nouveaux gènes susceptibles d'étendre si nécessaire les capacités de la bactérie afin d'optimiser sa survie.

La transduction

La transduction génétique est un phénomène consistant dans le transfert d'un segment d'ADN entre deux cellules par le biais d'un vecteur viral. Ainsi, si une cellule procaryote est infectée par un virus de procaryote (appelé phage), le génome de ce virus peut s'intégrer dans le génome de la cellule. Or, si ce génome de virus contient certes le nécessaire pour la réplication virale, il peut aussi contenir des portions d'ADN provenant de la cellule précédemment infectée ayant synthétisé ce virus. Ainsi, du matériel génétique peut alors être transféré de cellule à cellule.

Une fois infectée par un phage, une cellule procaryote peut réagir selon deux alternatives :

— Une réponse lysogénique : les gènes viraux responsables de la réplication virale ne sont pas exprimés. L'ADN viral est alors qualifié de prophage et il est plus ou moins inactif.

— Une réponse lytique : généralement sous l'effet d'un stress (altération de l'ADN par rayonnement ultra-violet, pression, manque de nutriments), la réponse lysogénique peut devenir lytique. La réponse lytique consiste en la production d'un grand nombre de phages suivi de l'éclatement de la bactérie infectée, libérant ainsi ces phages dans l'environnement.

À l'occasion de la phase lytique, la cellule va reproduire le virus et son génome. Ainsi, peuvent alors s'intégrer de nouvelles portions d'ADN non virales provenant de la bactérie au génome viral néosynthétisé.

La bactérie peut également se défendre contre l'infection virale par des mécanismes assimilables à un système immunitaire. Le mécanisme le plus célèbre est le système CRISPR-Cas9 (« Clustered Regularly Interspaced Short Palindromic Repeats - CRISPR associated protein 9 »). Le CRISPR est un annuaire présent dans certains génomes procaryotes constitué de courtes portions d'ADN issues de séquences virales. Cet annuaire est exprimé par plusieurs enzymes afin d'aboutir à de courtes séquences d'ARN (« crRNA »). Par la suite, ces séquences s'associent à une nucléase appelée Cas9 et vont lui servir de guide. En cas d'intégration d'ADN viral suivant une infection, le complexe

CRISPR-Cas9 va alors découper de façon spécifique l'ADN viral s'hybridant à la courte séquence d'ARN. Cela protège ainsi les organismes procaryotes des infections par des phages. Les fragments d'ADN viraux ainsi découpés peuvent alors être réutilisés par la cellule pour étendre l'annuaire CRISPR en vue d'une prochaine infection.

La conjugaison

Contrairement aux précédents phénomènes s'effectuant sans contact direct, la conjugaison permet un échange actif de matériel génétique entre deux cellules procaryotes par le biais d'un tube appelé « pilus conjugatif ». Dans le schéma classique (étudié chez l'espèce *Escherichia coli*), pour aboutir à une conjugaison, une cellule donneuse doit posséder un pilus conjugatif et un plasmide F contenant un facteur de fertilité « F-factor » nécessaire à la synthèse de ce pilus. La cellule donneuse va alors déployer son pilus jusqu'à l'insérer dans une cellule acceptrice présente dans l'environnement formant ainsi un pont entre les deux cytoplasmes. Il est à noter qu'un système empêche que deux cellules donneuses puissent se relier entre elles, ainsi une cellule acceptrice doit être dépourvue de pilus. Grâce à deux complexes enzymatiques appelés relaxosome et transférosome, le plasmide F du donneur est alors clivé et un seul brin d'ADN de ce plasmide est ensuite transféré via le pilus à la cellule acceptrice. Après cela, les deux cellules synthétisent le brin complémentaire des deux plasmides. Il en résulte que la cellule appelée acceptrice peut maintenant synthétiser un pilus et devenir une cellule donneuse.

Il est à noter que le plasmide F est un épisome, c'est-à-dire un plasmide à même de se recombiner avec le chromosome principal. D'autres segments d'ADN que les gènes strictement nécessaires à la synthèse du pilus peuvent alors se retrouver sur le plasmide F permettant ainsi des échanges horizontaux de gènes entre cellules. Ces segments peuvent par exemple correspondre à des gènes de résistance aux antibiotiques ou à des gènes permettant de métaboliser de nouvelles molécules. Bien que toutes les espèces n'en soient pas capables, au vu de ses caractéristiques, la conjugaison peut ainsi être vue comme une protosexualité procaryote.

Maintenant que nous avons passé en revue les principaux mécanismes biologiques conduisant à la variation des génomes procaryotes, nous allons ainsi pouvoir introduire le concept de pangénome.



Chapitre 2 : Aux origines de la pangénomique

Loin de la microbiologie, le terme « pan-génome » [1] est introduit par le Pr. François Sigaux [2] en 2000 dans un article prospectif publié dans le bulletin de l'Académie Nationale de Médecine (SIGAUX, 2000). À cette époque, bien que le génome humain complet n'était pas encore publié, il proposa via un partenariat entre le Centre National de Séquençage (Genoscope) et la Ligue contre le Cancer, de constituer en France une base de données publique, nommée « Carte d'Identité des Tumeurs, » afin de recenser les génomes et transcriptomes des différents types de cancer. À travers ce programme, entre 2003 et 2015, 14 183 tumeurs d'une vingtaine de pathologies cancéreuses seront répertoriées dans cette base de données (LIGUE CONTRE LE CANCER, 2019). Dans la littérature scientifique actuelle, au lieu du vocable de « pangénome », se substitue désormais celui de « pancancer » pour désigner ce type d'étude.

Pour ce qui est de la microbiologie, on attribue généralement l'origine du concept de pangénome à ces deux articles de 2005 : MEDINI et al., 2005 ; TETTELIN et al., 2005. Depuis lors, on définit un pangénome de la façon suivante :

Définition 2.0.1 Un pangénome désigne l'union de toutes les séquences non identiques (usuellement les gènes) présentes dans un ensemble de génomes (provenant le plus souvent de la même espèce).

1. Le préfixe « pan » vient étymologiquement du mot grec $\pi\alpha\nu$ signifiant « tout ». Dans tout ce document, le trait d'union sera ôté du mot composé « pan-génome » puisque cette notion est maintenant ancrée dans la communauté scientifique et nécessite donc d'être consacrée par le mot non composé « pangénome », de même, par exemple, que « métagénome ». De plus, selon Wikipédia, il n'est pas nécessaire d'utiliser un trait d'union après le préfixe modificateur « pan » en français (WIKIPEDIA, 2019c).

2. Pour l'anecdote, ce médecin fut le président du jury de financement CEA/IRTELIS ayant validé le contrat doctoral à l'origine de ce travail de thèse.

Hormis l'article de SIGAUX, 2000, le mot « pangénome » n'est évoqué qu'une seule fois avant 2005 dans un article de ADDO et al., 2003 avec un sens très éloigné [1]. Pourtant, sans utiliser le terme idoine, plusieurs études précurseuses menées bien avant 2005 ont montré que la diversité en gènes des procaryotes était bien plus importante que les estimations historiques sur la variabilité des espèces. Avant de les aborder plus en détail, il convient de rappeler ces considérations historiques.

2.1 Les génomes procaryotes, clones ou chimères ?

L'état de l'art des années 1980 supposait que la diversité intraspécifique était plutôt faible. Ainsi, selon LEVIN, 1981, la diversité de l'espèce *Escherichia coli* serait faible car l'existence d'un ancêtre commun positivement sélectionné du fait d'une mutation bénéfique stabiliserait non seulement la mutation avantageuse, mais également la lignée à partir de laquelle elle est apparue. En effet, contrairement aux eucaryotes, les procaryotes se reproduisant par scissiparité, les recombinaisons sont facultatives donc réputées comme peu fréquentes. Par conséquent, la reproduction des procaryotes a longtemps été considérée comme essentiellement clonale.

2.1.1 Les recombinaisons étaient admises mais sous-estimées

Il reste qu'à cette époque, les facteurs de recombinaison étaient pourtant déjà bien étudiés et identifiés. Une revue de la littérature (CAMPBELL, 1981) classifiait ainsi les gènes des génomes procaryotes en deux catégories :

— **Les gènes cœurs** (aussi qualifiés d' « euchromosomals ») dont la recombinaison au sein de la même espèce ou vers d'autres espèces est rarement sélectivement avantageuse. En effet, ces gènes étant respectivement co-adaptés, tout changement serait vraisemblablement délétère. C'est pourquoi, ces gènes conserveraient leur ordre général dans les génomes et constitueraient l'identité même des espèces à travers l'évolution.

— **Les gènes ou éléments accessoires** [2] ayant une origine extrachromosomale comme les prophages, les plasmides, les éléments conjugatifs ou les transposons. Ces gènes ne doivent normalement pas être des gènes absolument nécessaires à la reproduction de l'organisme. Enfin, ils sont d'ordinaire capables de se répliquer de manière autonome ou de se sur-répliquer lors de la réplication de l'ADN chromosomique.

Cette distinction témoigne de la perception qu'avaient les généticiens de l'époque de la variabilité des génomes et du répertoire génique des génomes procaryotes. Pourtant, le rôle évolutif des éléments accessoires a longtemps été considéré comme négligeable. En effet, étant donné que les échanges génétiques furent jadis considérés comme rares, on leur prêtait alors peu d'impact évolutif en comparaison de la sélection naturelle fixant les variants ayant des avantages sélectifs (COHAN, 1994a ; b). Dès lors, même si l'idée que les génomes se recombinent est depuis très longtemps acceptée, les modèles, comme celui de LEVIN, 1981, prédisaient naguère que la sélection naturelle s'opposerait à une explosion de la diversité bactérienne de lignées proches.

1. Il traite du dépistage de la réaction des cellules humaines face à tous les génomes du VIH (« pan-genome screening ») sans procéder à aucune analyse de séquences génomiques.

2. Le terme « accessoire » est issu du monde eucaryote avec la publication de MUNTZING, 1945 pour désigner les chromosomes dont le nombre peut varier dans les génomes en fonction du sexe (les gonosomes). Cela ne doit bien évidemment pas être confondu avec les anomalies chromosomiques de type trisomie.

2.1.2 Réévaluation de la fréquence des recombinaisons homologues

Avec l'utilisation croissante de nouvelles méthodes expérimentales, en particulier l'électrophorèse d'enzymes multiloci (MultiLocus Enzyme Electrophoresis : MLEE, SELANDER et al., 1986) [1], les généticiens ont peu à peu découvert une variabilité bien plus importante, même entre lignées proches. Cela sera confirmé par le typage des souches par séquences d'ADN de multiloci (MultiLocus Sequence Typing : MLST, MAIDEN et al., 1998). Cette méthode a montré en particulier que la fréquence des recombinaisons homologues était bien supérieure aux précédentes estimations (FEIL et al., 2001 ; J. M. SMITH et al., 2000).

2.1.3 Réévaluation de l'influence des transferts horizontaux de gènes

Dans le même temps, l'importance des transferts horizontaux de gènes (Horizontal Gene Transfer : HGT) entre les génomes a elle aussi été réévaluée. L'opinion dominante considérait jusqu'ici ce phénomène comme négligeable dans la compréhension générale de l'évolution (KOONIN et al., 2001). Un des premiers articles remettant en cause cet état de l'art fut l'article de MÉDIGUE et al., 1991. Cet article étudie l'adéquation entre les codons présents dans 780 gènes de l'espèce *E. coli* et l'effectif des ARN de transfert (ARNt) exprimés dans l'espèce. Comme cela était connu (GOUY et al., 1982), les codons des gènes se divisent via une analyse factorielle des correspondances en deux classes, une première pour les gènes à haute expression (protéines ribosomales par exemple) ayant des codons très adaptés au contingent d'ARNt et une seconde pour les gènes peu exprimés ayant des codons moins adaptés. Néanmoins l'article montre l'existence d'une troisième classe contenant un nombre significatif de gènes (≈10 à 15%). Ces gènes, bien plus riches en nucléotides AT, présentent des codons très divergents par rapport aux deux autres classes, en particulier de nombreux codons quasi-absents des deux autres classes. Dès lors, les auteurs de l'article conjecturent que ces gènes ont été acquis récemment par des transferts horizontaux depuis des espèces plus riches en AT et n'ayant pas eu le temps de s'adapter à leur nouvel hôte.

Par la suite, les premiers génomes entiers séquencés ont finalement étendu cette hypothèse en révélant que jusqu'à 18% des gènes possèdent des taux de bases GC exotiques sur *E. coli* (LAWRENCE et al., 1998) et même jusqu'à 20% pour l'espèce *Salmonella enterica* (LAN et al., 1996). D'autres études conduites sur différentes espèces confirment également ces résultats et sont répertoriées sur la figure 2.1 page suivante. Il faut aussi noter la quasi-absence de transferts horizontaux détectés chez certaines espèces ayant plutôt de petits génomes comme *Mycoplasma genitalium*, *Rickettsia prowazekii* ou bien *Borrelia burgdorferi*.

Plus récemment, les auteurs de TREANGEN et al., 2011 ont montré que le phénomène de transfert horizontal constituerait un phénomène tellement considérable pour l'évolution des procaryotes qu'il serait le premier facteur, devant la duplication, conduisant à l'extension du répertoire génique des procaryotes. On mesure bien l'erreur en quoi a consisté d'avoir ignoré l'importance de ces transferts aussi longtemps. De plus, ces séquences transférées horizontalement peuvent très concrètement avoir de grandes conséquences. D'un point de vue clinique par exemple, l'incorporation

1. Cette méthode consiste à faire migrer les enzymes extraites d'une souche sur gel par électrophorèse afin de générer des empreintes de poids moléculaires. La comparaison des empreintes permet ainsi d'identifier des variations entre génomes.

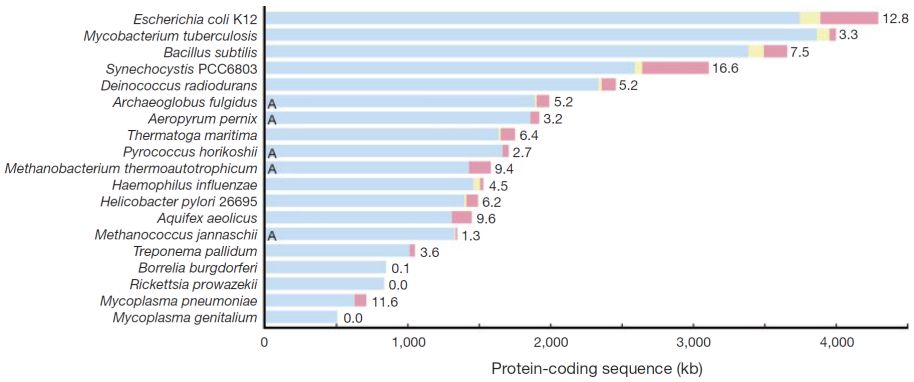


FIGURE 2.1 – Figure reproduite depuis OCHMAN et al., 2000 illustrant pour différentes espèces les quantités d'ADN euchromosomal en bleu, les éléments mobiles (transposons et prophages) en jaune et enfin l'ADN acquis par transferts horizontaux en rouge (avec le pourcentage correspondant sur la droite). Les génomes d'archées sont indiqués par la lettre A.

d'ADN peut conférer des caractéristiques de virulence pouvant rapidement modifier une souche commensale en une souche pathogène (HACKER et al., 1997 ; MCDANIEL et al., 1997). Il en est de même pour ce qui est de l'acquisition de gènes de résistance aux thérapies antibiotiques (FALKOW et al., 1971).

À long terme, l'impact cumulé des recombinaisons et de ces transferts horizontaux d'origine évolutive parfois très lointaine peut tendre à brouiller le signal phylogénétique de telle sorte que plusieurs articles suggèrent que les relations entre les lignées procaryotes devraient être représentées non plus par des arbres comme c'est classiquement le cas depuis Ernst Haeckel et Charles Darwin mais de préférence par des réseaux (BAPTESTE et al., 2009 ; DOOLITTLE, 1999). Cependant, loin de remettre totalement en cause le principe de l'évolution verticale darwinienne, le mode horizontal souligne davantage l'importance de prendre en compte les différents rouages participant de l'évolution des génomes procaryotes. Toujours est-il qu'en voyant la notion d'héritage vertical s'estomper, il est légitime de se demander si le concept d'espèce procaryote a encore un sens.

2.2 Qu'est-ce qu'une espèce procaryote ?

Autant le dire d'emblée, la systématique procaryote n'a pas encore produit de consensus pour définir l'unité de base, la brique première de la diversité biologique. Faire la diachronie du concept d'espèce nécessiterait assurément une thèse en soi. Ici, nous aborderons juste les interrogations que pose le concept d'espèce, en particulier afin de préciser le sens qu'il peut avoir dans le contexte procaryote.

2.2.1 Vue historique de la taxonomie procaryote

Même si le débat scientifique n'est pas éteint, il existe à travers la célèbre définition d'Ernst Mayr un relatif consensus pour ce qui est de la notion d'espèce animale. Cette définition s'exprime ainsi : « Species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups » (MAYR, 1942). Bien que le phénomène de

conjugaison pratiqué par certaines bactéries soit parfois considéré comme une forme de sexualité, il reste qu'il est conceptuellement difficile d'imaginer transposer *ceteris paribus* cette définition aux procaryotes en général. Certains considèrent alors que le concept d'espèce n'aurait aucun sens pour ce qui est des procaryotes (HOLLRICHER, 2007).

Depuis 1923, l'une des références régulièrement mise à jour les plus reconnues pour la taxonomie bactérienne d'abord, puis procaryote avec l'ajout des archées est le « Manual of Determinative Bacteriology » de Bergey (BERGEY et al., 1923). Ce manuel, longtemps *vade-mecum* indispensable du microbiologiste, fut initié par la société américaine de bactériologie [1] et recense la parenté des organismes procaryotes selon des critères principalement phénotypiques, c'est-à-dire reposant sur des considérants observationnels, biochimiques et cliniques. Il est à noter cependant que ces critères ont évolué dans les dernières éditions du manuel, en particulier, par la prise en compte des informations génomiques. En parallèle, il existe également un code international « International Code of Nomenclature of Prokaryotes » (ICNP) chargé de standardiser la terminologie taxonomique (ICNP, 2019).

2.2.2 Quels critères univoques pour délimiter les espèces ?

Pour autant, et malgré des manuels de référence, il n'y a rien de plus équivoque que des phénotypes, ce qui a conduit immanquablement à multiplier les démarcations possibles pour quasi chaque espèce. Afin de clarifier les différents critères de la systématique bactérienne, s'est réuni à l'Institut Pasteur en 1987 un comité international chargé de réconcilier les approches taxonomiques. L'objectif du comité était de proposer un critère formel donc univoque pour pouvoir affirmer que deux organismes relèvent ou non de la même espèce. Le comité a discuté de retenir un critère d'identité entre deux organismes sur le gène de la petite sous-unité de l'ARN ribosomique (ARNr 16S). En effet, ce dernier avait été historiquement utilisé par Carl Woese pour produire le premier arbre phylogénétique universel du vivant (WOESE et al., 1977). Néanmoins ce critère est jugé trop variable en fonction des branches de la phylogénie. Finalement, le comité tranchera en faveur du critère suivant : deux organismes doivent avoir au moins 70% de leur génome qui s'hybride à 5°C au plus en-dessous de la température de fusion de ces ADN pour être considérés comme appartenant à une même espèce (WAYNE et al., 1987).

Pourtant, les difficultés techniques à conduire des expériences d'hybridation ADN à ADN (DNA-DNA Hybridation : DDH) amènent à un regain d'intérêt pour l'utilisation de l'ARNr 16S comme gène marqueur pour l'identification des espèces. De surcroît, les auteurs de STACKBRANDT et al., 1994 montrent qu'un DDH de 70% est équivalent à ≈97% d'identité sur le gène entier de l'ARNr 16S (≈1500b). Une décennie plus tard, les auteurs de STACKEBRANDT et al., 2006 avertiront que ce seuil doit être réévalué à ≈98.7%.

Nonobstant ces équivalences, l'article de MARTINEZ-MURCIA et al., 1992 avait déjà exposé les limites que concentre l'utilisation d'un seul gène marqueur, par exemple sur l'espèce *Aeromonas trota* où 30% de DDH correspond à une identité pour l'ARNr 16S de 99.9%. Cela peut spécialement être le cas pour les génomes en réduction. Enfin, l'article de WARD et al., 1998 critique cette approche en signalant que le seuil de 70% de DDH n'est pas assez prédictif du phénotype.

1. Désormais appelée société américaine de microbiologie.

Avec le développement du séquençage de génomes entiers, l'article de KONSTANTINIDIS et al., 2005 propose de pallier à ces critiques en utilisant l'ensemble des gènes de chaque génome comme potentiels marqueurs phylogénétiques. Pour cela, les séquences de l'ensemble des gènes d'un génome A sont alignées bioinformatiquement sur celles de l'ensemble des gènes d'un génome B. La moyenne des pourcentages d'identité des alignements est ensuite calculée en utilisant les meilleures correspondances pour chaque gène du génome A dans le génome B. Cette moyenne d'identité nucléique (Average Nucléotide Identity : ANI) est ainsi utilisée comme critère pour définir les espèces procaryotes [1]. Sur la base de cette métrique, les auteurs de l'article considèrent que deux organismes dépendent de la même espèce si leur valeur d'ANI est supérieure à 94%, ceci étant corrélé avec une valeur de DDH également supérieure à 70% ($R^2 = 0.93$).

2.2.3 Vers une taxonomie repensée avec des critères génomiques ?

Aujourd'hui, le nombre de génomes accumulés dans les principales bases de données s'évalue à plusieurs centaines de milliers de génomes, de sorte qu'il n'est aujourd'hui plus computationnellement possible d'aligner tous les gènes pour des centaines de milliers de génomes contre tous les autres sans de sérieuses optimisations. Parmi celles-ci, on peut citer MASH (ONDOV et al., 2016) et FastANI (JAIN et al., 2018). Ces deux outils fonctionnent en extrayant des mots de taille k appelés k-mers dans les séquences génomiques avant de les transformer via des fonctions mathématiques en valeurs de hachage (aussi appelées « hashes »). En sélectionnant les plus petites d'entre elles, on forme ainsi un ensemble de hashes que nous pouvons considérer comme une empreinte (aussi appelée « sketch ») pour chaque séquence. La comparaison avec d'autres empreintes est ensuite très efficace puisqu'il suffit d'une simple mesure de distance de Jacquard entre les 2 ensembles de hashes. Tandis que l'outil MASH génère des empreintes pour un génome entier, l'outil fastANI applique cette méthode sur des fenêtres à l'intérieur des génomes. Ainsi, fastANI peut approcher un alignement et ainsi fournir des résultats équivalents à la méthode originelle. Pour ce qui est de l'outil MASH, nous calculons une distance entre les génomes entiers, ce qui est bien plus rapide mais aussi moins précis que l'ANI. Ainsi, on considère qu'une distance MASH de 0.06 correspond à une valeur approximée d'ANI de 94%.

Ces outils rapides sont précieux. L'outil fastANI par exemple a été employé dans l'article de JAIN et al., 2018 pour montrer en comparant $\approx 90\,000$ génomes, l'existence d'une délinéation claire plutôt qu'un continuum entre les espèces. Cela tendrait à montrer que la notion d'espèce a un sens en tant que groupe de génomes proches. Cela reste néanmoins controversé. De plus, il faut avoir conscience que l'échelle de l'espèce est ici très différente des échelles utilisées pour les animaux. Par exemple, les humains possèdent entre eux une valeur d'ANI assurément supérieure à 99.9% (AUTON et al., 2015). Si une valeur seuil d'ANI aussi stricte était appliquée aux procaryotes, le nombre d'espèces serait exorbitant. Sans aller vers ce genre d'implications proprement vertigineuses [2] au moins pourrait-on déjà s'assurer de la cohérence des rangs taxonomiques au sein du monde procaryote. C'est tout l'objet du projet GTDB (Genome Taxonomy DataBase, PARKS et al., 2018).

1. Il faut noter que même si rien ne contraint foncièrement l'ANI à être symétrique, en pratique, les résultats obtenus dans les 2 sens sont extrêmement proches.

2. On estime déjà avec les règles actuelles le nombre d'espèces bactériennes à $\approx$ 1000 milliards dont moins de 100 000 seraient au moins partiellement séquencées (LOCEY et al., 2016).

Ce projet se propose de refaire la taxonomie du NCBI sur la base d'une phylogénie obtenue à partir de 120 gènes marqueurs habituellement présents chez tous les procaryotes. Il en résulte une taxonomie modifiée à hauteur de ≈58% par rapport à l'existant (voir figure 2.2). On observe toutefois que le rang espèce ressort plutôt bien conservé de ce remaniement en comparaison des rangs supérieurs.

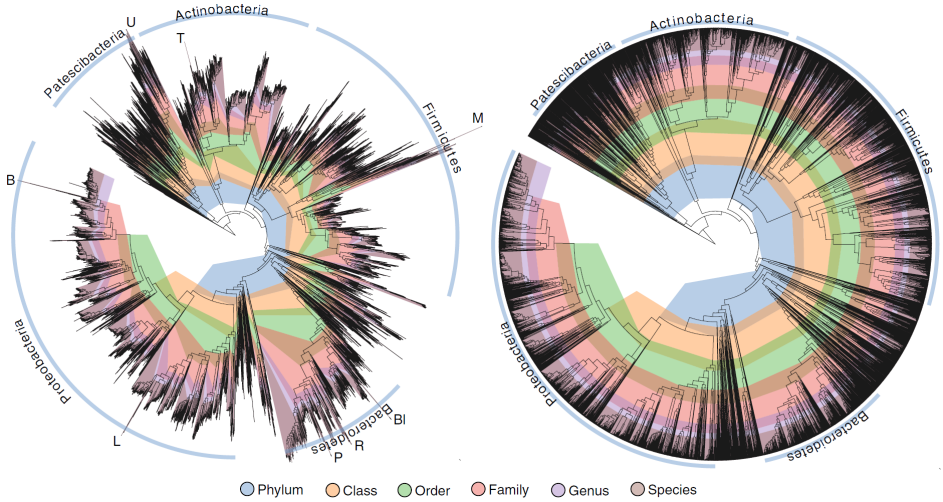


FIGURE 2.2 – Figure reproduite depuis PARKS et al., 2018 illustrant les rangs taxonomiques du NCBI projetés sur l'arbre calculé par les auteurs (sur la gauche) en comparaison de l'actualisation des rangs taxonomiques que les auteurs proposent (sur la droite).

2.3 Les premières analyses de pangénomes

Bien que les résultats provenant des analyses MLST et de l'étude de la composition des codons indiquent une partie variable pouvant atteindre jusqu'à ≈20% des génomes pour certaines espèces procaryotes, il est nécessaire d'y apporter confirmation par la comparaison directe de génomes. Déjà en 1998, une étude de AKOPYANTS et al., 1998 confirmait – en détectant de nombreux fragments d'ADN présents dans certaines souches bien qu'absents de la référence de *Helicobacter pylori* [1] – que la variabilité dans l'espèce est bien supérieure aux attentes. Une approche encore plus robuste consisterait dans le séquençage de plusieurs souches proches afin d'en comparer les génomes. Ces besoins coïncident avec la considérable explosion des projets de séquençage de nouvelles espèces au début des années 2000. Pourtant, les re-séquençages de génomes d'espèces possédant déjà une référence demeuraient anecdotiques, loin derrière, et cela se conçoit fort bien, la stimulante quête pour les génomes d'espèces inédites. Bien sûr, il est toujours possible de comparer des génomes éloignés et d'estimer le moment de l'apparition de leurs gènes spécifiques par rapport aux gènes ancestraux. Ce fut notamment l'objet de nombreux travaux en génomique comparée (JORDAN et al., 2001 ; SNEL et al., 2002). Il reste, si l'on me pardonne le truisme, que pour pouvoir étudier la variabilité entre génomes proches, il est préférable d'obtenir la séquence de génomes proches.

1. À l'aide d'une PCR (Polymerase Chain Reaction, MULLIS et al., 1986) modifiée pour l'étude en question.

2.3.1 Le premier pangénome de l'espèce *Escherichia coli*

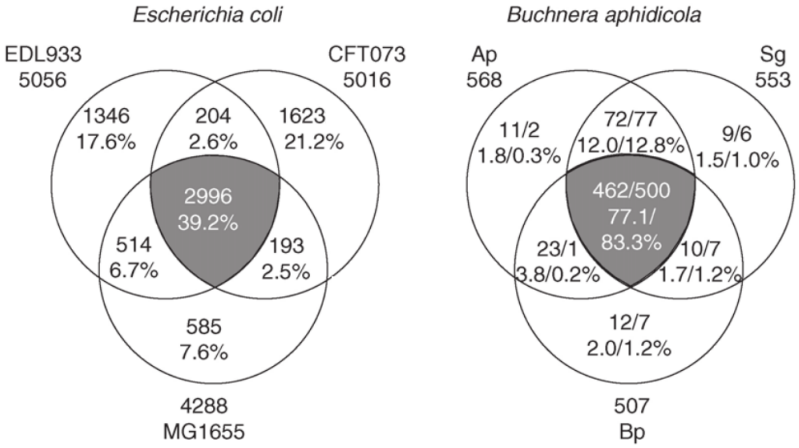


FIGURE 2.3 – Diagramme de Venn reproduit et modifié (car il y avait une erreur dans la figure originale, ce que j'ai signalé à l'auteur correspondant) depuis LAWRENCE et al., 2005 comparant le contenu en gènes codants et non redondants partagés entre les trois génomes de l'espèce *Escherichia coli* à gauche et de l'espèce *Buchnera aphidicola* à droite. Dans le cas de l'espèce *Buchnera aphidicola*, les métriques à gauche de la barre oblique indiquent le nombre de gènes après filtrage des pseudogènes tandis que les métriques à sa droite sont brutes.

Quand, à la fin de l'année 2002, les auteurs de WELCH et al., 2002 purent analyser trois génomes d'*E. coli* entiers (la souche de référence $K12-MG1655$ ainsi que 2 souches pathogènes, la souche uropathogène $CFT073$ et la souche entérohémorragique $EDL933$), ils ont constaté par comparaison des différences frappantes. En effet, après union de tous les gènes codants et non redondants de ces trois génomes, ils observent avec étonnement que seulement 39.2% (soit 2 996 gènes) sont communs à tous ces génomes, c'est-à-dire des **gènes cœurs**. Cela indique, par contraste, que pas moins de 60.8% des gènes totaux sont variables entre ces souches (voir partie gauche de la figure 2.3). Ces chiffres excédent les précédentes estimations évoquées dans l'article de LAWRENCE et al., 1998 et la première comparaison déjà effectuée seulement entre les souches $K12-MG1655$ et $EDL933$ (PERNA et al., 2001). De plus, les auteurs observent que ces gènes se localisent sous la forme d'îlots génomiques possédant un biais dans la fréquence de leurs codons relativement à ceux des gènes cœurs. Ils attribuent donc une origine extrachromosomale à ces gènes qui auraient dès lors été acquis par événements de transfert horizontal. Ils en infèrent que les génomes d'*E. coli* seraient des mosaïques composées d'une alternance d'îlots entrecoupant une colonne vertébrale de gènes cœurs. Pour qualifier ces îlots, les auteurs reprennent ici le terme de régions « accessoires » utilisé par CAMPBELL, 1981. D'un point de vue fonctionnel maintenant, ces gènes accessoires semblent, pour la souche uropathogène, être à la source de ses capacités à infecter les voies urinaires (par exemple via des gènes d'adhésion) et sanguines mais aussi de permettre à la souche de résister au système immunitaire de l'hôte. L'acquisition de ces gènes ne semble toutefois pas affecter la capacité de la bactérie à vivre dans l'intestin. Pour le cas de la souche entérohémorragique, les gènes accessoires acquis lui conféreraient la faculté de coloniser

des régions spécifiques de l'intestin. De surcroît, ces gènes modifieraient le mode d'interaction avec le tissu intestinal par rapport aux souches commensales. Ces deux caractéristiques seraient responsables de symptômes gastro-intestinaux cliniquement très irréguliers.

2.3.2 Le premier pangénome de l'espèce *Buchnera aphidicola*

Quelques mois plus tard, une étude (VAN HAM et al., 2003) du même type est menée sur trois génomes de l'espèce *Buchnera aphidicola* qui correspondent aux souches *Ap*, *Sg* et *Bp*, respectivement, symbiontes intracellulaires des espèces de pucerons *Acyrthosiphon pisum*, *Schizaphis graminum* et *Baizongia pistaciae*. Pour ces cas ci, la variabilité en gènes apparaît plus faible que pour *E. coli* (voir partie droite de la figure 2.3 page précédente). Les gènes cœurs constituent ici 77.1% de l'union des gènes codants et non redondants entre ces trois génomes. Pour autant et contrairement là encore à *E. coli*, cette diversité de gènes n'est ici pas attribuée par les auteurs à l'acquisition de nouveaux gènes. En effet, ces derniers montrent que ce phénomène n'aurait pas eu lieu dans l'espèce depuis 50 à 200 millions d'années. À l'inverse, cette variabilité serait causée par quelques pertes de gènes conduisant à une légère réduction génomique, en particulier pour la souche *Bp* qui aurait perdu autour de 10% de gènes par rapport aux deux autres génomes (voir figure 2.3 page ci-contre).

Ce cas illustre en quoi les fluctuations du contenu en gènes ne sont pas seulement attribuables aux gains de gènes mais qu'elles peuvent également résulter de leur perte. Par conséquent, contrairement à la définition de CAMPBELL, 1981 utilisée jusqu'ici, désormais sera utilisée une définition moderne des gènes accessoires, c'est-à-dire, qui inclut les gènes perdus dans certaines souches d'un ensemble de génomes quoique présents dans les autres, et ce malgré une origine euchromosomale.

2.3.3 Le supra-génome et l'hypothèse du génome distribué

Comme évoqué dans l'incipit de la partie I page 21, on attribue généralement le concept de pangénome à MEDINI et al., 2005 ; TETTELIN et al., 2005, toutefois, celui-ci a été théorisé 2 ans auparavant en 2003 sous un autre nom par les auteurs de ERDOS et al., 2003 au travers de l'hypothèse du génome distribué (« Distributed Genome Hypothesis » ou DGH). Cette dernière postule l'existence au niveau de la population, d'un « supra-génome » bien plus grand que les génomes des bactéries individuelles. Ainsi, chaque souche de bactéries contiendrait un sous-ensemble unique des gènes du supra-génome. Bien qu'ils utilisent le terme supra-génome, les auteurs décrivent là sans ambiguïté, ce que nous appelons désormais un pangénome. N'ayant accès qu'à un seul génome séquencé pour l'espèce qu'ils étudient (*Haemophilus influenzae*), les auteurs de ERDOS et al., 2003 vont corroborer l'hypothèse DGH en comparant des fragments d'ADN de plusieurs souches de l'espèce avec la référence [1]. Sur la base de ces expériences, ils concluent qu'au moins 11% de l'ADN présent dans l'espèce *Haemophilus influenzae* n'est pas présent dans le génome de référence.

1. Les auteurs brisent l'ADN des génomes de plusieurs souches différentes de *H. influenzae* en de nombreux morceaux qu'ils mélangent et qu'ils insèrent dans des vecteurs plasmiques (un seul morceau s'insère dans chaque plasmide). Ces vecteurs plasmidiques sont ensuite clonés dans *Escherichia coli* (un seul vecteur s'insère dans chaque clone). En séquençant ces clones, ils observent que 11% des clones possèdent de l'ADN absent de la référence connue.

2.3.4 Concevoir un vaccin par l'étude du pangénome de l'espèce *S. agalactiae*

En 2005, *Streptococcus agalactiae* (aussi appelé « Group B *Streptococcus* » ou GBS) est parfois décrit comme la principale cause d'infections néonatales potentiellement mortelles (GIBBS et al., 2004) [1]. Bien que des antibiotiques permettent de traiter l'infection, une vaccination de la mère (pouvant transmettre ses anticorps à l'enfant) est envisagée en traitement préventif. Un vaccin est déjà disponible mais ne fonctionne que sur les souches américaines. En particulier ce vaccin est inefficace sur les souches japonaises car celles-ci sont d'un sérotype différent (LACHENAUER et al., 1999). Afin de rechercher un vaccin à efficacité universelle sur toutes les souches de l'espèce, les auteurs de MAIONE et al., 2005 parviennent à obtenir le génome de huit souches de *S. agalactiae*. Sur la base de ces génomes, les auteurs identifient [2] une seule protéine vaccinale présente dans toutes les souches de l'espèce tandis qu'ils en trouvent trois présentes dans certaines souches seulement. En combinant ces quatre protéines, ils parviennent ainsi à obtenir le premier vaccin à efficacité universelle pour *S. agalactiae*.

En recherchant ces quatre protéines, ils découvrent que 1 811 gènes codants et non redondants sont présents dans la totalité des huit souches (≈80% de chaque génome) alors qu'ils sont 765 à être présents uniquement dans certaines souches. Trois des quatre protéines immunogènes indispensables à l'efficacité universelle de leur vaccin étaient présentes dans cet ensemble de 765 gènes accessoires et n'auraient pas pu être retrouvées si les auteurs n'avaient utilisé qu'une seule souche.

2.3.5 La naissance « officielle » de la pangénomique

À l'été 2005, l'auteur de TETZ, 2005 n'apporte pas de résultats mais expose son opinion en proposant le concept de pangénome. Le concept qu'il expose est vu d'une façon très extensive et dépasse l'échelle de l'espèce. Ici le pangénome est étendu à tout le monde vivant, de sorte qu'il désigne l'ensemble de tous les gènes présents sur Terre. À titre d'exemple, l'auteur indique que ce concept peut notamment être utilisé en médecine pour la découverte de médicaments ayant des modes d'action nouveaux. Mais c'est finalement au deuxième semestre de 2005 que trois articles (LAWRENCE et al., 2005 ; MEDINI et al., 2005 ; TETTELIN et al., 2005) vont jeter les bases historiques de la notion de pangénome en microbiologie.

Dans le premier article, LAWRENCE et al., 2005 passe en revue les études menées jusqu'alors. C'est dans cet article qu'est utilisé pour la première fois le terme « pangénome » dans le contexte de la microbiologie pour désigner l'ensemble des gènes présents dans un clade. Cet article confirme que la vision historique d'un génome présentant une certaine stabilité pour une espèce donnée est renversée par la découverte que des génomes même étroitement apparentés présentent des contenus en gène parfois très différents.

1. Environ 76.8 cas/an pour 100 000 enfants de moins de 1 an sont répertoriés aux USA en 2004 d'après CENTERS FOR DISEASE CONTROL AND PREVENTION, 2005.

2. Pour cela, les protéines candidates identifiées comme potentiellement exprimées à la surface de *S. agalactiae* (589 gènes candidats) sont clonées dans *E. coli* afin d'exprimer la protéine massivement (312 clonages réussis). Chaque protéine est ensuite injectée à un groupe de souris femelles afin qu'elles développent une réaction immunitaire. La bactérie est ensuite administrée à tous les groupes de souris. Les groupes de souris survivant à l'infection permettent ainsi d'identifier les protéines vaccinales, c'est-à-dire celles capables d'induire une réaction immunitaire protégeant contre l'infection.

Dans le deuxième article, les auteurs de TETTELIN et al., 2005 réutilisent les huit génomes de MAIONE et al., 2005 afin d'approfondir nos connaissances sur la diversité génétique intraspécifique de l'espèce *S. agalactiae*. Plus particulièrement, leur but était d'une part, de comprendre l'origine des grandes différences de pathogénicité au sein de l'espèce *S. agalactiae* et d'autre part de découvrir de nouvelles cibles antimicrobiennes. Pour y parvenir, plutôt que d'effectuer de multiples comparaisons de génome à génome (soit $\binom{8}{2} = \frac{8!}{2!(8-2)!} = 28$ comparaisons) comme c'était généralement le cas en génomique comparée, les auteurs généralisent l'approche holistique caractéristique de la pangénomique et déjà utilisé par WELCH et al., 2002 et VAN HAM et al., 2003 basée sur l'identification d'éléments non redondants. Pour cela, les auteurs agrègent l'ensemble des gènes retrouvés dans l'espèce en familles de gènes homologues. En comptant la présence ou l'absence des gènes appartenant à ces familles dans les génomes, ils vont ainsi tenter de modéliser la diversité pangénomique présente dans les espèces. À cet effet, les auteurs introduisent les concepts de **pangénome ouvert** et de **pangénome fermé**, qualifiant respectivement la tendance haute et basse, de la croissance du pangénome lors de l'ajout successif de nouveaux génomes (voir figure 2.4 page suivante).

Dans le troisième article, les auteurs de MEDINI et al., 2005 constatent qu'entre 1995, année où un premier génome bactérien est complètement séquencé (*Haemophilus influenzae* Rd par FLEISCHMANN et al., 1995) et l'année 2005, seulement un génome par espèce est généralement séquencé. À titre illustratif, ils font remarquer que 211 espèces bactériennes ne possédaient qu'un seul génome en décembre 2005, soit 83.3% des 239 génomes bactériens séquencés (Tableau 1 de MEDINI et al., 2005). Néanmoins, comme de plus en plus d'espèces comme *S. agalactiae*, *Buchnera aphidicola* ou *E. coli* commencent à être représentées par plusieurs génomes entiers dans les bases de données, les auteurs notent que cela autorise des études plus approfondies de la diversité génomique entre les génomes d'une même espèce. En particulier, ils indiquent que cela est susceptible de nous informer sur la manière dont ces espèces évoluent. En effet, les familles de gènes homologues communes aux différentes souches contiendraient le métabolisme essentiel au mode de vie de base d'une cellule tandis que le génome accessoire conférerait des avantages sélectifs non essentiels, comme des gènes permettant l'adaptation à des niches écologiques, de résister à des antibiotiques ou enfin à la capacité de coloniser de nouveaux hôtes.

Cette nouvelle approche de la génomique comparée, va amener de nombreuses questions notamment celles-ci : le pangénome est-il fini ou infini ? Quelle est la vitesse de croissance du pangénome lorsqu'on séquence de plus en plus de souches et la stabilité du génome cœur ? Combien de souches doivent être séquencées pour observer la totalité (ou presque) du pangénome d'une espèce ? Quelles informations la notion de pangénome apporte-t-elle aux théories sur l'évolution des génomes ?

Pour y répondre, des analyses supplémentaires et des modèles *ad hoc* ou réadaptés depuis d'autres disciplines scientifiques vont être nécessaires. Synthétiser les principales modélisations existantes dans la littérature, pour répondre à ces différentes questions va être l'objet du chapitre qui suit.

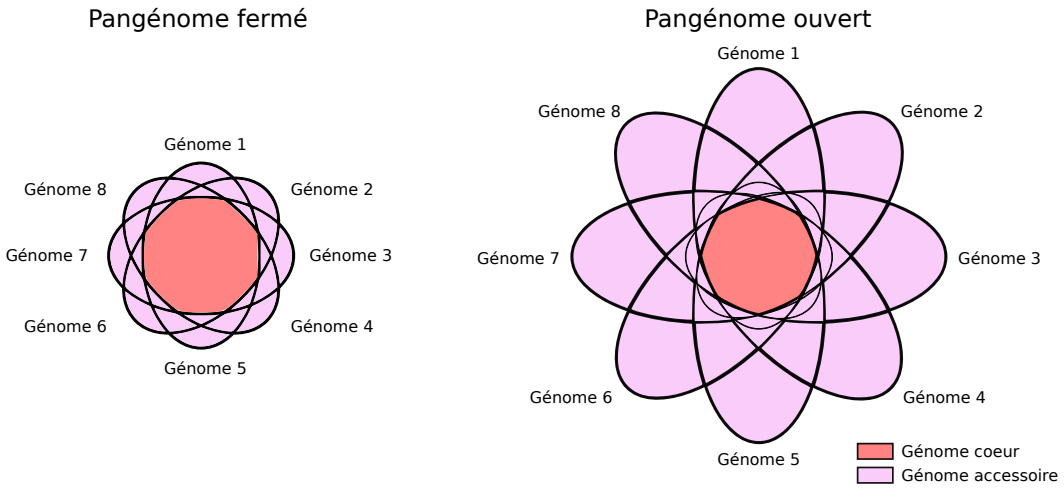
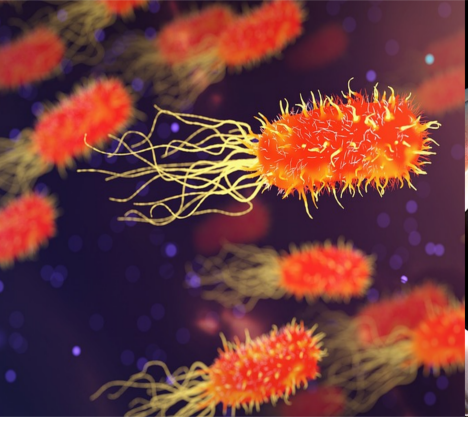


FIGURE 2.4 – Diagrammes de Venn représentant les intersections entre les ensembles de gènes que constituent les génomes. La totalité d'un diagramme (l'union des ensembles) représente le **pangénome** tandis que le centre rouge identifie l'ensemble des gènes communs à tous les génomes : **le génome cœur**. La corolle autour du génome cœur représente quant à elle, l'ensemble de gènes plus ou moins partagés entre les génomes : **le génome accessoire**. Le diagramme de gauche illustre un **pangénome fermé** où les génomes présentent peu de gènes spécifiques alors que celui de droite correspond à un **pangénome ouvert** où au contraire les génomes possèdent un nombre important de gènes spécifiques.





Chapitre 3 : Modélisation des pangénomes

3.1 Modélisation par décroissance exponentielle

Pour estimer la diversité de l'espèce *S. agalactiae*, les auteurs de TETTELIN et al., 2005 ont entrepris de modéliser la tendance à la croissance des pangénomes. Ils s'intéressent en particulier à l'apparition de nouveaux gènes et à la décroissance du génome cœur, lors de l'ajout de nouveaux génomes. Puisque ces métriques dépendent nécessairement du nombre de souches connues dans l'espèce, les auteurs vont extrapoler, via une modélisation mathématique, les métriques observées sur les huit génomes de MAIONE et al., 2005 à un nombre infini de génomes. Ainsi via ce modèle, les auteurs espèrent pouvoir prédire :

1. L'évolution du nombre de nouvelles familles de gènes découvertes à chaque ajout d'un nouveau génome dans l'espèce, soit le génome spécifique.
2. L'évolution du nombre total de familles de gènes présentes au moins une fois dans tous les génomes de l'espèce, soit la taille du pangénome. Cette métrique peut être aisément déduite par la somme cumulée de l'item précédent.
3. L'évolution du nombre de familles de gènes présentes dans absolument tous les génomes de l'espèce, soit le génome cœur.

Pour modéliser ces différentes évolutions sur la base des huit génomes, les auteurs vont tout d'abord mesurer l'évolution de la taille du génome cœur et la taille du génome spécifique pour chaque ajout successif de génome. Il semble logique de s'attendre à ce que le génome cœur se réduise à chaque ajout (ou au moins reste stable) ; de même pour le nombre de gènes spécifiques apportés par le dernier ajout de génome. Néanmoins, ces intuitions nécessitent d'être appuyées par un modèle.

Pour calculer le nombre d'arrangements, c'est-à-dire le nombre de successions distinctes d'ajouts de génome (en interdisant les doubles ajouts de génome), il est possible d'utiliser la formule suivante : $A_n^k = \frac{n!}{(n-k)!}$ où k désigne la taille des successions et n le nombre d'éléments (WIKIPEDIA, 2019a). Dans notre cas nous avons 8 génomes et nous souhaitons faire des successions de longueur 8 donc $A_8^8 = \frac{8!}{(8-8)!} = 40320$. Néanmoins ce calcul, bien que correct, n'est pas réellement utile en pratique car deux successions différentes peuvent posséder des sous-séquences d'ajouts

produisant les mêmes mesures. De sorte qu'un nombre de mesures inférieur pourrait permettre d'obtenir les mêmes résultats. En remplacement, pour calculer le nombre total d'échantillons de génomes à mesurer (N) pour toutes les combinaisons de successions possibles, les auteurs utilisent plutôt la formule suivante :

$$N = \frac{8!}{(n-1)!(8-n)!}$$

où 8 correspond au nombre de génomes, n aux différentes tailles d'échantillons de génomes à mesurer (n allant de 1 à 8). En appliquant cette formule, nous obtenons un total de 8 échantillons de 1 génome à mesurer, 56 échantillons de 2 génomes à mesurer *et cetera*... Finalement nous avons donc au total $8+56+168+280+280+168+56+8 = 1025$ échantillons à mesurer. Pour chaque échantillon à mesurer, seront calculés le génome cœur (voir figure 3.1) et le nombre de gènes spécifiques ajoutés (voir figure 3.2).

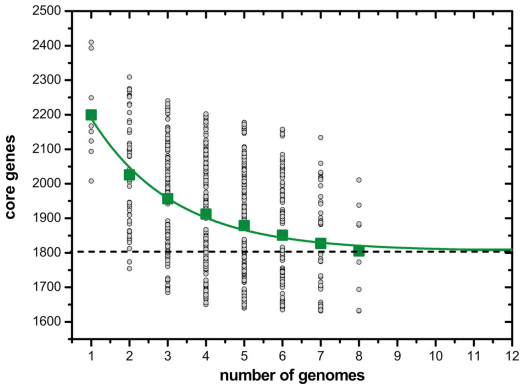


FIGURE 3.1 – Figure reproduite depuis TETTELIN et al., 2005 illustrant la courbe de raréfaction du nombre de familles de gènes dans le génome cœur en fonction du nombre de génomes de l'espèce *S. agalactiae* ajoutés successivement. Sur cette figure comme sur la figure 3.2 ci-contre, chaque point correspond à un échantillon de taille n mesuré. La courbe verte correspond à la modélisation de F_c par décroissance exponentielle.

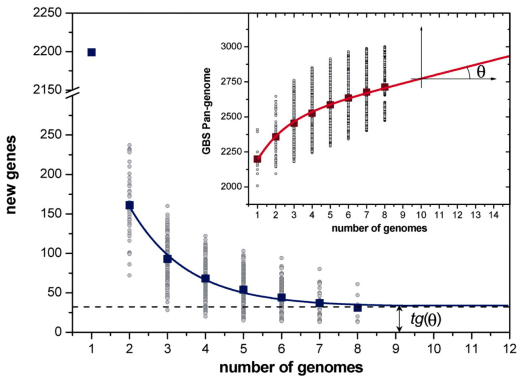


FIGURE 3.2 – Figure reproduite depuis TETTELIN et al., 2005 illustrant la courbe de raréfaction du nombre de gènes spécifiques découverts en fonction du nombre de génomes de l'espèce *S. agalactiae* ajoutés successivement. La courbe bleue correspond à la modélisation de F_s par décroissance exponentielle. La courbe rouge est obtenue par la somme cumulative des points bleus correspondant à l'évolution de la taille du pangénome.

Afin d'extrapoler les métriques à un nombre supérieur de génomes, les auteurs modélisent leur tendance par deux fonctions paramétrées pour s'adapter aux mesures. Dès lors, une fonction va exprimer la taille du génome cœur F_c et une autre va exprimer le nombre de gènes spécifiques F_s. Comme nous l'avons vu, il est raisonnable de penser que les métriques F_s et F_c diminuent lors de l'ajout successif de génomes. C'est pourquoi les auteurs postulent que ces deux fonctions suivent une tendance exponentiellement décroissante relativement à la taille des échantillons n, de sorte

qu'ils les expriment ainsi :

$$F_c = \kappa_c e^{-\frac{n}{\tau_c}} + \Omega \tag{3.1}$$

$$F_s = \kappa_s e^{-\frac{n}{\tau_s}} + tg(\theta) \tag{3.2}$$

où κ_c, κ_s, τ_c, τ_s, Ω, $tg(\theta)$ sont des paramètres s'adaptant aux mesures et $tg(\theta)$ correspond au taux de croissance extrapolé du pangénome, de sorte que si $P(n)$ est la fonction exprimant la taille du pangénome (soit la somme cumulée des F_s) en fonction du nombre maximum de génomes disponibles, alors, $\lim_{n\to+\infty} P(n) \approx tg(\theta)n$. En effet $\lim_{n\to+\infty} \kappa_s e - \frac{n}{\tau_s} = 0$.

Pour estimer ces paramètres, les auteurs les ajustent aux moyennes [1] des mesures effectuées pour chaque taille d'échantillons n par la méthode des moindres carrés. Ainsi, ils estiment d'une part que : $\kappa_s = 476 \pm 62$, $\tau_s = 1.51 \pm 0.15$ et $tg(\theta) = 33 \pm 3.5$ avec un coefficient de corrélation de Pearson entre les valeurs réelles et prédites de $R^2 = 0.99$ et d'autre part que : $\kappa_c = 610 \pm 38$, $\tau_c = 2.16 \pm 0.28$ et $\Omega = 1806 \pm 16$ avec un coefficient $R^2 = 0.995$. Sur la base de cette modélisation, ils en concluent donc que le nombre de gènes dans le génome cœur est égal à $\Omega = 1806 \pm 16$ et qu'il restera constamment autour de Ω peu importe le nombre de génomes ajoutés. De même, ils prédisent que la taille du pangénome augmentera de $tg(\theta) = 33 \pm 3.5$ nouveaux gènes à chaque ajout de génome au-delà des 8 génomes, indépendamment du nombre de génomes ajoutés. Pour les auteurs si $tg(\theta) > 0$ alors le pangénome est ouvert et inversement si $tg(\theta) = 0$ le pangénome est fermé. Sur la base de leurs résultats, puisque $tg(\theta) = 33 \pm 3.5$, ils en arrivent donc à la conclusion que le pangénome de *S. agalactiae* est ouvert, ou autrement dit, qu'il est infini.

Trois ans plus tard, grâce aux progrès des méthodes de séquençage [2], de très nombreux nouveaux génomes sont disponibles permettant d'obtenir le pangénome de plusieurs espèces. Parmi ces nouveaux pangénomes disponibles, plusieurs d'entre eux ont été analysés en suivant plus ou moins la méthode ici présentée (résultats rassemblés dans le tableau 3.1).

Clade	Nombre de génomes	Taille du pangénome	Taille du génome cœur	Pangénome ouvert ou fermé ?	Publication
Streptococcus agalactiae	8	≈2646	≈1 806	ouvert	TETTELIN et al., 2005
Streptococcus pyogenes	11	≈2 500	≈1 376	ouvert	LEFÉBURE et al., 2007
Neisseria meningitidis	6	≈3000	≈1 337	ouvert	SCHOEN et al., 2008
Escherichia coli + genre *Shigella*	32	≈11 872	≈2 241	ouvert	WILLENBROCK et al., 2007
Escherichia coli	17	≈13 000	≈2 344	ouvert	RASKO et al., 2008
genre *Prochlorococcus*	12	≈5 736	≈1 273	ouvert	KETTLER et al., 2007

TABLE 3.1 – Caractéristiques de six pangénomes constitués entre 2005 et 2008.

1. Les auteurs suggéreront 3 ans plus tard qu'il eût été préférable d'utiliser la médiane à la place de la moyenne pour limiter l'influence éventuelle des extrema (TETTELIN et al., 2008).

2. En particulier l'apparition des séquenceurs de 2ème génération (à haut débit) de type : Roche-454, Solexa-Illumina et ABI-SOLiD dont la vitesse de séquençage est supérieure de plusieurs ordres de grandeur aux anciennes technologies basées sur la méthode Sanger (SANGER et al., 1977).

L'un des inconvénients du modèle de TETTELIN et al., 2005 est que celui-ci repose sur deux hypothèses. La première est qu'une fois stabilisée la croissance du pangénome serait linéaire. La deuxième est que le génome cœur pourrait être stabilisé. Or, les résultats sur un ensemble de génomes plus important comme celui de WILLENBROCK et al., 2007 semblent réfuter empiriquement ces hypothèses. En effet, dans cette étude pangénomique se basant sur 32 génomes de l'espèce *E. coli* ainsi que du genre *Shigella* (génomes très proches d'*E. coli* mais ayant perdu de nombreux gènes), les auteurs semblent en désaccord avec les résultats de la modélisation par décroissance exponentielle du génome cœur (figure 3.3, courbe bleue). Pour répondre à cela, ils proposent une curieuse modification de la modélisation du cœur pour mieux s'adapter à leurs données (figure 3.3, courbe verte) en remplaçant le calcul de F_c dans l'équation 3.1 page précédente, par :

$$F_c = \kappa_c e^{-\frac{\sqrt[4]{n}}{\tau_c}} + \Omega \tag{3.3}$$

Il est possible que cette modification *ad hoc* sans justification de la part des auteurs présente un risque de sur-ajustement sur les données. De plus, on peut se demander si cet ajout n'est pas un pis-aller à une inexorable décroissance du génome cœur vers zéro. Pour ce qui est des résultats de la modélisation par décroissance exponentielle des nouveaux gènes (voir figure 3.4), la courbe se stabilise sur une valeur de croissance linéaire qui semble éloignée des données. Pourtant, les auteurs de l'article ne semblent pas conscients du problème en affirmant « the decay function [...] was found to be appropriate », alors que la densité de points semble bien en-dessous des valeurs prédites par la courbe modèle. En particulier, les points denses (cyans, bleus et violets) semblent très proches de zéro nouveau gène découvert.

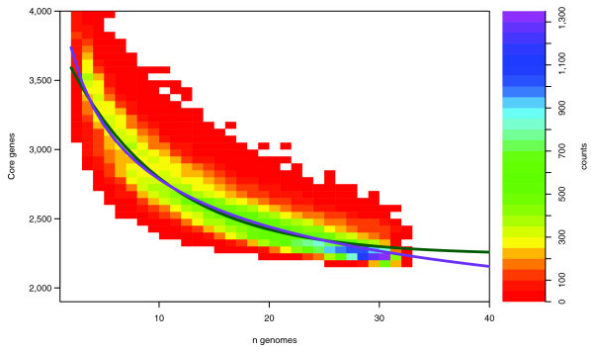


FIGURE 3.3 – Figure reproduite depuis WILLENBROCK et al., 2007 illustrant les courbes de raréfaction du nombre de gènes cœurs provenant de 32 génomes de l'espèce *E. coli* et du genre *Shigella*, en surplomb d'un diagramme de densité des mesures résultant de l'échantillonnage.

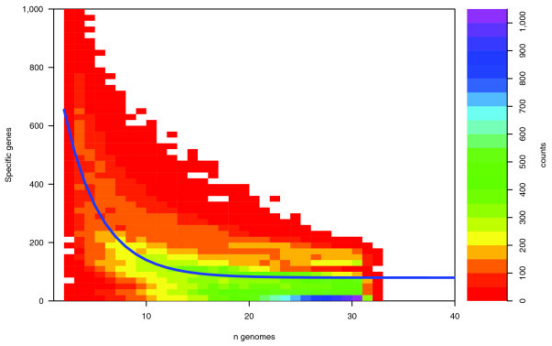


FIGURE 3.4 – Courbe de raréfaction du nombre de gènes spécifiques provenant du même article et du même ensemble de génomes que la figure 3.3 ci-contre, en surplomb d'un diagramme de densité des mesures résultant de l'échantillonnage.

Pour conclure à la vue de ses résultats, on peut empiriquement se demander si un modèle imposant une stabilisation de la tendance du pangénome est une bonne option.

3.2 Modélisation par la loi de Heaps

Pour pallier aux limites du modèle précédemment présenté, les auteurs de cet article TETTELIN et al., 2008 proposent d'utiliser la loi de Heaps (HEAPS, 1978) [1], déjà éprouvée dans le domaine de la linguistique, pour modéliser la tendance des pangénomes. Plus précisément, cette loi exprime le nombre de mots distincts empiriquement observés lors de la lecture d'un ensemble de documents comme s'accroissant de moins en moins vite au fur à mesure que le corpus déjà lu s'agrandit. En effet, si nous nous attendons à ce que des déterminants comme « le », « les » ou « une » soient présents dans tous les livres en langue française [2], à l'inverse, on s'attend aussi, à ce que des mots rares (comme les néologismes ou encore les hapax) nécessitent de lire un grand nombre de textes pour être observés. Bien sûr cela suppose un choix aléatoire de documents car à l'évidence on trouvera bien plus fréquemment le mot « emphytéotique » dans la bibliothèque d'un notaire que dans les magazines posés sur le guéridon d'un cabinet médical. S'il était encore utile de le préciser, dans l'usage réinterprété que proposent les auteurs, il faut remplacer les mots par les gènes et les documents par les génomes. Une fois réinterprétée, la loi de Heaps énonce donc que le nombre de gènes dans un pangénome F en fonction du nombre de génomes (N) peut être modélisé par la fonction suivante :

$$F = p(N) = \kappa N^{\gamma}$$

où la fonction p est une loi de Heaps, soit un polynôme paramétré par κ et γ. Le paramètre κ est une constante de proportionnalité tandis que γ paramètre la tendance de la fonction.

- $\gamma > 1$: donnera une croissance supralinéaire (courbes bleues sur la figure 3.5 page suivante).
- $\gamma = 1$: donnera une croissance linéaire (droites rouges sur la figure 3.5 page suivante).
- $0 < \gamma < 1$: donnera une croissance sublinéaire (courbes vertes sur la figure 3.5 page suivante).
- $\gamma = 0$: donnera une tendance stable (droites oranges sur la figure 3.5 page suivante).

Ainsi, le nombre de gènes spécifiques découverts à chaque ajout Δ_F n'est pas nécessairement attendu comme stable à terme comme dans l'article de TETTELIN et al., 2005 mais est donné par la dérivée p' de p :

$$\Delta_F = p'(N) = \kappa\gamma N^{\gamma-1} = \kappa(1-\alpha)N^{-\alpha}$$

En effet, les auteurs remplacent $\gamma - 1$ par $-\alpha$, avec ($\alpha = 1 - \gamma$), puis en réutilisant certaines des données disponibles au moment de la publication, ils vont estimer α pour neuf espèces bactériennes en adaptant les paramètres aux mesures des courbes de raréfaction (voir figure 3.6 page 51).

Dans l'article de TETTELIN et al., 2005 évoqué à la sous-section 3.1 page 45, un pangénome était considéré comme fermé si son taux de croissance $tg(\theta) \approx 0$ et ouvert si $tg(\theta) > 0$. Dans cet article, les auteurs redéfissent la manière de considérer un pangénome comme ouvert ou fermé. Il considère désormais que pour $\alpha > 1$, la taille du pangénome s'approche d'une constante à mesure que l'on ajoute davantage de génomes, c'est-à-dire que le pangénome est fermé. Inversement,

1. Loi originellement proposée par Gustav Herdan (HERDAN, 1960), confirmant ainsi la loi aphoristique de Stiegler. Cette dernière énonçant qu'« une découverte scientifique ne porte jamais le nom de son auteur ».
2. Peut-être à l'exception des lipogrammes, comme le roman « La Disparition » de George Pérec.

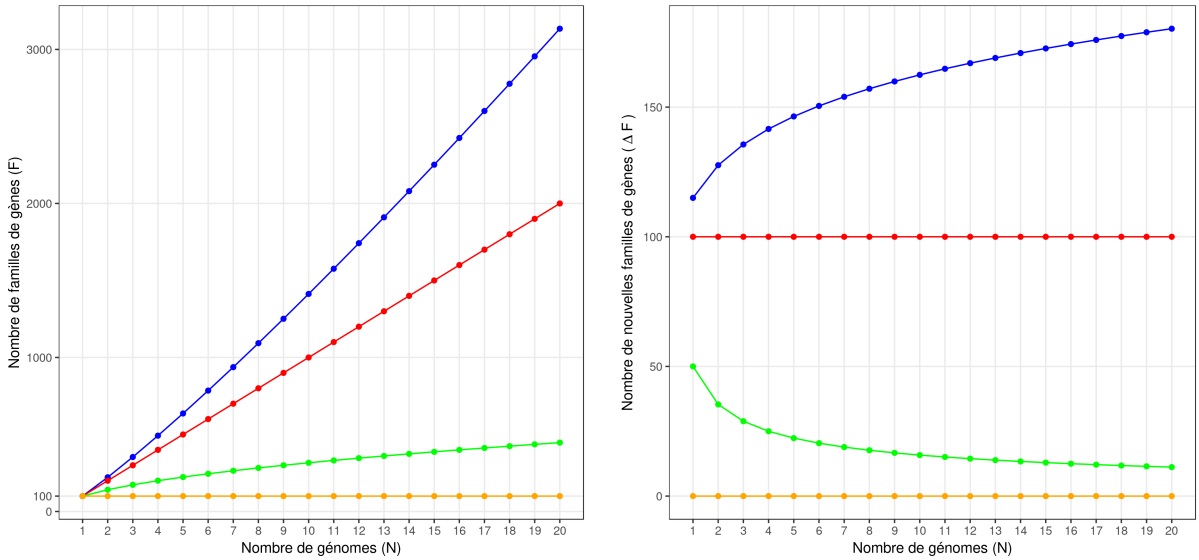


FIGURE 3.5 – Exemple de quatre lois de Heaps avec quatre paramètres de γ différents. $\kappa = 100$ pour les quatre courbes mais $\gamma = 1.15$ ($\alpha = -0.15$) pour la courbe bleue, $\gamma = 1$ ($\alpha = 0$) pour la droite rouge, $\gamma = 0.5$ ($\alpha = 0.5$) pour la courbe verte, $\gamma = 0$ ($\alpha = 1$) pour la droite orange. Les tracés de gauche montrent le nombre cumulé de familles de gènes F tandis que les tracés de droite montrent l'ajout successif Δ_F de nouvelles familles de gènes.

pour $\alpha < 1$, la taille du pangénome est une fonction croissante et illimitée du nombre de génomes considérés, c'est-à-dire que le pangénome est ouvert. Enfin, si $\alpha = 1$ la taille du pangénome suit une tendance logarithmique, c'est-à-dire qu'elle croît très lentement, mais qu'elle reste techniquement sans limite. Par conséquent, ils considèrent que si le paramètre $\alpha = 1$ (ou même $\alpha \approx 1$), comme le pangénome est croissant (même faiblement), il peut être considéré comme ouvert. Dès lors, sur la base des valeurs de α, les espèces suivantes sont considérées comme ayant des pangénomes ouverts (droites vertes sur la figure 3.6 page ci-contre) : *Bacillus cereus*, *Prochlorococcus marinus*, *Streptococcus pneumoniae*, *Escherichia coli* et *Streptococcus agalactiae* ; alors que les espèces suivantes sont considérées comme ayant des pangénomes fermés (droites rouges sur la figure 3.6 page suivante) : *Staphylococcus aureus*, *Streptococcus pyogenes*, *Ureaplasma urealyticum* et *Bacillus anthracis*.

Pour conclure, la loi de Heaps semble tout à fait pertinente pour modéliser la tendance d'un pangénome. D'une part car elle semble moins rigide que la modélisation par décroissance exponentielle et d'autre part, elle est remarquablement parcimonieuse car seulement deux paramètres permettent d'en modéliser la croissance. Un autre avantage est que ce modèle peut être aussi bien utilisé pour modéliser l'évolution du nombre de gènes spécifiques, la croissance du pangénome et enfin la décroissance (ou la stabilité) du génome cœur. Dans ce dernier cas, γ sera négatif ou nul.

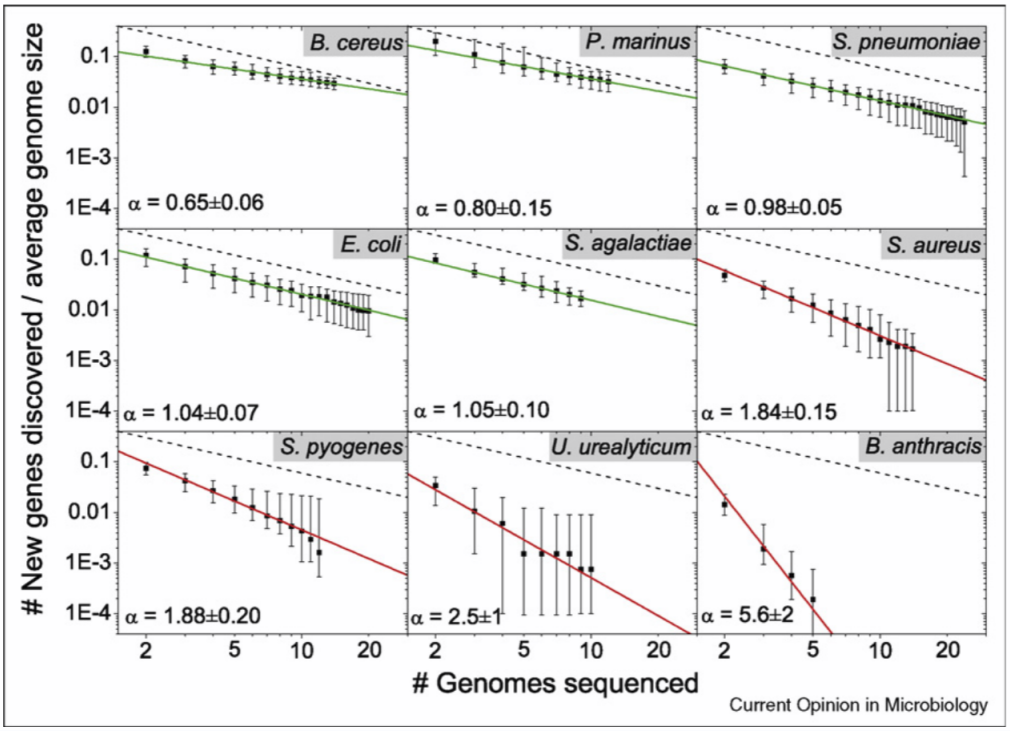


FIGURE 3.6 – Figure reproduite depuis TETTELIN et al., 2008 illustrant des courbes de raréfaction du nombre de gènes spécifiques (les 2 axes sont à l'échelle logarithmique) pour neuf espèces bactériennes. Les neuf droites représentent des fonctions d'équation $\Delta_F = p'(N) \propto N^{-\alpha}$ modélisant le nombre de nouveaux gènes découverts Δ_F à chaque ajout successif de génome. Les 2 axes sont à l'échelle logarithmique et la droite en pointillés indique la fonction $\Delta_F = N^{-1}$.

3.3 Modélisation d'un pangénome par modèles de mélange binomiaux

3.3.1 Modèle binomial utilisant des classes de fréquence fixées *a priori*

Les auteurs de HOGG et al., 2007 étudient la variabilité à l'intérieur de l'espèce *Haemophilus influenzae* et souhaitent modéliser le pangénome de l'espèce sur la base des treize souches disponibles. Plutôt que d'utiliser des approches comme celles précédemment évoquées, les auteurs proposent une méthode alternative. En effet, selon eux la modélisation de TETTELIN et al., 2005 dont ils avaient connaissance, se base sur le postulat que le pangénome est infini ; hypothèse qu'ils questionnent. Dans leur article, ils préféreront faire le choix de l'hypothèse inverse en considérant le pangénome comme fini. En effet, le pangénome est ici représenté par un modèle génératif, une sorte « d'usine » construisant des génomes selon un ensemble de règles probabilistes. Ce pangénome fini contient un nombre N de gènes distincts qui sont modélisés comme des variables aléatoires de Bernoulli avec des probabilités de réussite correspondant à la fréquence de chaque gène g dans la population. Un génome provenant d'une souche procaryote s est ainsi fabriqué par la réalisation de N expériences de Bernoulli. Pour chaque génome, celui-ci possédera un gène g si l'expérience correspondante a réussi alors que g en sera absent dans le cas inverse. Ce modèle assez simple suppose donc que chaque gène g présent dans un génome, le soit indépendamment des autres. Cette hypothèse est à l'évidence fausse en microbiologie car les îlots génomiques, par exemple,

contiennent des ensembles de gènes acquis en même temps par un événement de transfert horizontal d'ADN (LANGILLE et al., 2010). Nonobstant ces critiques, qu'ils anticipent, les auteurs estiment cette hypothèse comme raisonnable pour simplifier leur modèle.

Tout d'abord, il faut dire que le nombre et la fréquence de ces gènes hypothétiques dans la population sont inconnus. Dès lors, les auteurs supposent que ces gènes peuvent être classifiés en K classes discrètes de gènes. Chaque classe k est associée à une fréquence de gènes dans la population μ_k qui sera la même pour tous les gènes appartenant à la classe k. Chacun des gènes g est affecté à une classe k selon une distribution de probabilité donnée par le vecteur π, où π_k est la probabilité qu'un gène soit affecté à la classe k ; π_k correspondant à la proportion de gènes du pangénome ayant une fréquence μ_k dans la population. L'affectation d'un gène à une classe k est indépendante de toutes les autres affectations de gènes.

Posons Z la variable cachée dans laquelle chaque z_n indique la classe du gène g. Posons également X la variable indiquant les gènes réellement observés, où $x_{g,s}$ correspond à la présence ou à l'absence du gène g dans le génome s. Pour résumer, le modèle ainsi conçu nécessite les paramètres suivants : N, K, un coefficient de mélange π_k et une probabilité de réalisation de Bernoulli μ_k pour chaque classe. Le nombre de classes de gènes K et leur probabilité de Bernoulli associée μ_k sont fixés à l'avance. Les auteurs proposent de rendre compte des différentes classes de fréquences des gènes dans la population à travers $K = 7$ classes ayant les probabilités μ_k associées suivantes : $\mu = \{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$. La classe $k = 7$ ayant une fréquence de gènes $\mu_k = 1$ dans la population représente les gènes du génome cœur ; les autres classes expriment les différentes fréquences du génome accessoire. Sachant ces paramètres instinctivement fixés, les variables N et π sont déterminées via une méthode de maximisation de vraisemblance.

Supposons que $|S|$ génomes sont observables et qu'un gène g appartenant à une classe k est observé n fois dans les $|S|$ génomes, un gène ne pouvant être présent qu'une seule fois par génome. La probabilité de cette observation peut donc être calculée comme une probabilité binomiale.

$$P(x = n \mid z = k, \mu_k) = \frac{|S|!}{n!(|S| - n)!} \mu_k^n (1 - \mu_k)^{|S|-n}$$

Néanmoins comme nous ne connaissons pas la classe des gènes, alors il nous faut un modèle de mélange :

$$P(x = n \mid \pi, \mu) = \sum_{k=1}^{K} P(x = n \mid z = k, \mu_k) \cdot P(z = k \mid \pi_k) = \sum_{k=1}^{K} \pi_k \frac{|S|!}{n!(|S| - n)!} \mu_k^n (1 - \mu_k)^{|S|-n}$$

Nous pouvons ainsi estimer la probabilité d'un gène g d'être présent dans n génomes observés sachant le modèle. Étendons maintenant ce modèle à tous les gènes du pangénome. Considérons $c = c_0, \ldots, c_n, \ldots, c_S$, où c_n est le nombre de gènes observés dans exactement n génomes, ce vecteur c s'appelle le profil du pangénome. La probabilité de toutes ces observations est alors donnée par la

distribution multinomiale suivante :

$$P(c \mid N, \pi, \mu) = \frac{N!}{c_0! \dots c_n! \dots c_{|S|}!} \prod_{n=0}^{|S|} p(x = n \mid \pi, \mu)^{C_n}$$
$$= \frac{N!}{c_0! \dots c_n! \dots c_{|S|}!} \prod_{n=0}^{|S|} \Big(\sum_{k=1}^{K} \pi_k \frac{|S|!}{n!(|S|-n)} \mu_k^n (1-\mu_k)^{|S|-n} \Big)^{C_n}$$

Les paramètres optimaux $\hat{N}$ et $\hat{\pi}$ peuvent alors être déterminés en maximisant la log-vraisemblance des observations c :

$$\ell(c \mid N, \pi, \mu) = \log(N!) - \sum_{n=0}^{|S|} \log(C_n!) + \sum_{n=0}^{|S|} c_n \log\Big(\sum_{k=1}^{K} \pi_k \frac{|S|!}{n!(|S|-n)} \mu_k^n (1-\mu_k)^{|S|-n} \Big)$$

Pour rappel les auteurs possèdent treize génomes de l'espèce *Haemophilus influenzae* mais ils souhaitent également vérifier les résultats de leur modèle avec un ensemble plus petit de génomes. À cet effet, ils sélectionnent un sous-ensemble de seulement huit génomes. Après estimation des paramètres de leur modèle sur la base de ce sous-ensemble, celui-ci nous indique que la taille du pangénome est comprise dans cet intervalle : $2975 \leqslant \hat{N} \leqslant 3681$. Les résultats ne sont pas identiques sur la base des treize génomes où le modèle estime plutôt les valeurs suivantes : $4425 \leqslant \hat{N} \leqslant 6052$. Pour ce qui est des valeurs de π_k, les voici illustrées dans la figure 3.7. Sur cette figure, on observe que la classe ayant $\mu_k = 1$, c'est-à-dire le génome cœur, possède 1 423 gènes pour les huit génomes et 1 437 pour les treize génomes. Le génome cœur est donc plutôt stable quand le nombre de génomes augmente. On observe également que la majorité des nouveaux gènes ajoutés par le passage de huit à treize génomes sont préférentiellement associés à la classe $\mu_k = 0.01$, autrement dit ce sont des gènes rares.

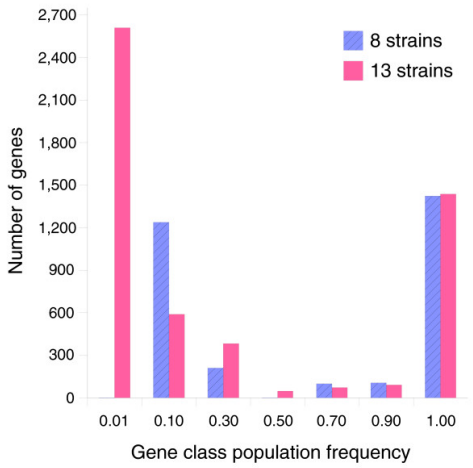


FIGURE 3.7 – Figure reproduite depuis HOGG et al., 2007 illustrant le nombre de gènes dans chacune des $K = 7$ classes de fréquences μ_k utilisées par les auteurs pour les deux ensembles de génomes (huit ou treize souches de l'espèce *Haemophilus influenzae*).

Deux mois après la publication de ce modèle, celui-ci est appliqué sur dix-sept génomes de l'espèce *Streptococcus pneumoniae* par les auteurs de HILLER et al., 2007. Leurs résultats montrent un pangénome fini de ≈5 117 gènes où ≈1,380 appartiendraient au génome cœur et ≈2,100 seraient des gènes spécifiques d'un seul génome seulement. Avec ce modèle, les auteurs estiment avoir découvert 95% du pangénome fini hypothétique de *S. pneumoniae* avec ces dix-sept souches seulement.

3.3.2 Modèle binomial déterminant les classes de fréquence des gènes

Les auteurs de SNIPEN et al., 2009 reprennent les idées de HOGG et al., 2007 en considérant le pangénome comme fini mais ils trouvent cependant que ce modèle possède de lourdes restrictions comme le fait de figer la fréquence μ_k des classes. En remplacement, ils proposent une autre modélisation inspirée des modèles « capture-recapture » déjà utilisés en écologie (EVANS et al., 1994) pour l'estimation de la taille des populations animales [1]. Ici, le pangénome est représenté comme une matrice binaire de présences/absences $M = m_{i,j}$ où chaque ligne correspond à une famille de gènes homologues i et chaque colonne à un génome j. Si une famille de gènes i possède au moins un gène dans le génome j alors $m_{i,j} = 1$ sinon $m_{i,j} = 0$ (voir figure 3.8 page suivante). Ils définissent également $y = \{y_1, \ldots, y_j \ldots, y_N\}$ le vecteur listant les nombres de familles de gènes ayant des gènes observés dans exactement j génomes ($1 \leqslant j \leqslant N$, N étant le nombre total de génomes). Ce vecteur y est également appelé le profil du pangénome ou aussi parfois le spectre de fréquence des gènes (voir figure 3.8 page ci-contre). La taille observée du pangénome n est donc calculable par $n = \sum_{j=1}^{N} y_j$. Au-delà de la taille observée, les auteurs nomment η la « vraie » taille du pangénome théorique, celui-ci contenant un nombre fini de familles de gènes. En effet, la taille de ce pangénome théorique fini devrait prendre en compte y_0 le nombre de familles de gènes présentes dans 0 génome, c'est-à-dire, le nombre de familles de gènes non observées jusqu'ici. Ainsi $\eta = n + y_0$. Sur la base de ce formalisme, les auteurs souhaitent déduire y_0 depuis la partie connue du vecteur y.

À partir de maintenant, on considérera désormais y sous sa forme étendue par l'ajout des familles de gènes non observées (comme c'est le cas sur la figure 3.8 page suivante), soit $y = y_0, y_1, \ldots, y_j, \ldots, y_N$. Par hypothèse η étant un nombre fini, on peut considérer y comme un vecteur de valeurs distribuées de façon multinomiale sous réserve d'indépendance entre les familles de gènes. On peut ainsi écrire $y \sim Multinomial(\theta, \eta)$ où les probabilités binomiales sont données par $\theta = (\theta_0, \theta_1, \ldots, \theta_j, \ldots, \theta_j, \theta_N)$ correspondant à la probabilité d'une famille de gènes i, d'être détectée dans exactement j génomes. Pour estimer ces paramètres θ, les auteurs vont utiliser, comme HOGG et al., 2007, le modèle de mélange binomial à K classes suivant :

$$\theta_j = \sum_{k=1}^{K} \pi_k \frac{N!}{j!(N-j)!} \rho_k^j (1-\rho_k)^{N-j} \tag{3.4}$$

1. Quand il est impossible de recenser exhaustivement une population, les écologues capturent, puis marquent une partie représentative de la population avant de relâcher ces animaux. Par la suite, un échantillon d'animaux est capturé afin de compter le nombre d'individus marqués ou non marqués. Une estimation de la taille de la population totale peut ainsi être obtenue en divisant le nombre total d'individus initialement marqués par la proportion d'individus marqués à la recapture.

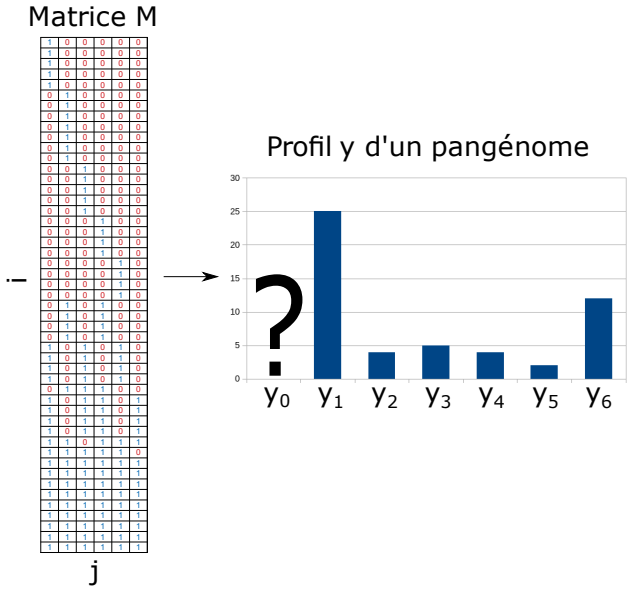


FIGURE 3.8 – Exemple illustratif d'une matrice binaire de présences/absences M comprenant $n = 52$ familles de gènes observées pour $N = 6$ génomes. Le profil de ce pangénome y donnant la distribution de la fréquence d'apparition des familles de gènes i est également représenté sous forme d'histogramme sur la droite.

où les classes k ayant une proportion de mélange π_k doivent être interprétées comme des classes de familles de gènes homologues ayant une probabilité de succès de Bernoulli ρ_k d'être détectée dans un génome j. Si ρ_k est faible, ces familles sont rarement observées dans les génomes et inversement. La classe 1 pour laquelle $\rho_1 = 1$ est la classe des familles de gènes du génome cœur.

En l'absence de données pour y_0, les auteurs sont dans l'embarras pour estimer les paramètres de leur modèle de mélange. Pour surmonter ce problème, ils proposent d'estimer les paramètres de leur modèle en maximisant sa log-vraisemblance hors y_0.

En considérant le pangénome n ayant un profil restreint y_j pour $j > 0$ seulement, alors le vecteur est aussi multinomial avec une probabilité que l'on peut exprimer par $\frac{\theta_j}{(1-\theta_0)}$ pour chaque élément du profil. Dès lors, la log-vraisemblance hors zéro du modèle de mélange binomial devient :

$$\ell(\pi, \rho \mid K) = \sum_{j=1}^{N} y_j \log\left(\frac{\theta_j}{(1-\theta_0}\right) + C \tag{3.5}$$

$$= \sum_{j=1}^{N} y_j \log\left(\frac{\sum_{k=1}^{K} \pi_k \frac{N!}{j!(N-j)!} \rho_k^j (1-\rho_k)^{N-j}}{1 - \sum_{k=1}^{K} \pi_k (1-\rho_k)^N}\right) + C \tag{3.6}$$

où la log-vraisemblance hors zéro de l'équation 3.6 est déduite de la 3.5 car θ_j dépend de π et ρ d'après l'équation 3.4 page ci-contre, et où C est une constante indépendante. Les paramètres maximisant la log vraisemblance hors 0 pour un K donné sont donc : $\hat{\pi}_K$, $\hat{\rho}_K$. Ces paramètres permettent de déduire une estimation des θ_j, incluant θ_0. Ayant obtenu θ_0, on peut alors en déduire une estimation de $\hat{y}_0$.

Dans HOGG et al., 2007, K était arbitrairement fixé à 7. Ici K va être déterminé en maximisant un critère d'information bayésien (Bayesian Information Criterion : BIC, SCHWARZ, 1978) sachant que pour chaque valeur de K possible, le BIC peut être calculé ainsi :

$$BIC(K) = -2\ell(\pi, \rho \mid K)) + (2K - 2)\log(n)$$

Le calcul du BIC pour plusieurs valeurs de K permet d'estimer le meilleur compromis entre le gain de vraisemblance apporté par un K élevé et la parcimonie du modèle. Le but étant de trouver le K maximisant le BIC.

Species	Genomes	Sample core	Median size	Sample pan	Components	Mixture core	Chao pan	Mixture pan	Coverage
Campylobacter jejuni	5	847	1697	3221	3	470	6986	6587	
Coxiella burnetii	5	963	1856	3308	3	692	4698	4612	
Acinetobacter baumannii	6	1555	3428	6421	3	900	12681	15023	
Buchnera aphidicola	6	159	550	1051	3	113	2194	2597	
Helicobacter pylori	6	917	1522	2650	3	850	4212	4178	
Rhodopseudomonas palustris	6	1820	4606	10542	3	1651	17879	17654	
Streptococcus pneumoniae	6	1213	2010	3327	3	1113	5225	5190	
Yersinia pestis	7	2385	3816	5798	4	1172	8174	8622	
Francisella tularensis	7	805	1514	2524	4	154	3183	3463	
Bacillus cereus	8	2421	5388	11276	4	1038	25920	33780	
Clostridium botulinum	8	824	3390	9606	3	692	15760	15366	
Prochlorococcus marinus	12	726	1885	7567	4	678	13759	13472	
Streptococcus pyogenes	13	1073	1797	3433	5	760	5452	6081	
Salmonella enterica	14	2425	4406	9868	5	1902	22040	24580	
Staphylococcus aureus	14	1485	2532	4646	6	613	6858	8368	
Escherichia coli	22	2446	4523	12599	6	1760	26614	42640	

TABLE 3.2 – Tableau reproduit depuis SNIPEN et al., 2009 résumant les métriques issues de la modélisation des pangénomes de seize espèces. La colonne « Components » correspond aux valeurs de K maximisant le BIC. Les colonnes « Sample core » et « Sample pan » indiquent respectivement le génome cœur et le pangénome observés tandis que les colonnes « Mixture core » et « Mixture pan » fournissent les prédictions théoriques du modèle. Les auteurs se comparent à la métrique donnée par la méthode de Chao utilisée en écologie (CHAO, 1987) pour estimer la taille théorique d'une population. Celle-ci se calcule trivialement avec la formule suivante : $\eta_{\text{Chao}} = n\frac{y_1^2}{2y_2}$.

Pour rappel, dans ce formalisme, la taille estimée du pangénome théorique est $\hat{\eta} = n + \hat{y}_0$ et la taille estimée du génome cœur théorique est égale à $\hat{\eta}\hat{\pi}_1$. Les auteurs ont appliqué leur modèle sur les génomes de seize espèces (voir les résultats sur le tableau 3.2). Sur la base de ces résultats, les auteurs notent que contrairement à ce qui était d'usage avec la dichotomie classique : génome cœur et génome accessoire, aucune espèce ne semble avoir un nombre optimal de classes $K = 2$. Les valeurs de K s'échelonnant entre 3 et 6. Cela peut bien sûr être expliqué par l'échantillonnage mais comme cela semble le cas pour toutes les espèces, il est aussi possible d'imaginer une origine biologique à ce partitionnement. À noter également que la couverture de η par n ou autrement dit la part du pangénome théorique déjà observée ne dépend pas nécessairement du nombre de génomes inclus. Par exemple, il est intéressant d'observer que pour *Escherichia coli*, bien que ce soit l'espèce avec le plus de génomes utilisés, elle est celle qui présente le plus faible niveau de couverture. Enfin, on constate des génomes cœurs théoriques relativement petits en comparaison des génomes cœurs observés, en particulier pour les espèces suivantes où ils correspondent à environ 50% seulement des valeurs observées : *Campylobacter jejuni*, *Yersinia pestis*, *Bacillius cereus*, *Staphylococus aureus* et même jusqu'à moins de 80% pour *Francisella tularensis*.

Six ans après la publication de cet article, le modèle ici décrit fut implémenté sous la forme d'un paquet R appelé « micropan » (SNIPEN et al., 2015) permettant à qui le souhaite de facilement modéliser le pangénome de son choix avec un modèle mélange binomial. Ce paquet R sera évoqué derechef à la sous-section 5.1.12 page 87 et à la sous-section6.1.4 page 105, car c'est l'outil le plus comparable à la méthode de partitionnement qui sera développée dans ce manuscrit de thèse.

3.3.3 Critique des modèles à pangénomes finis

La principale critique, selon moi, que l'on peut adresser aux deux modèles précédemment décrits consiste à s'interroger sur la stabilité des prédictions qu'ils proposent lorsque de nouveaux génomes sont ajoutés. En effet, ces modèles se basent sur l'idée d'un pangénome fini, c'est-à-dire supposent qu'il existerait une population finie de gènes à même de s'intégrer dans les génomes d'une espèce donnée. S'il est vrai que pour certaines espèces bactériennes vivant isolées de leurs congénères, comme celles ayant un mode de vie intracellulaire par exemple, il semble raisonnable d'affirmer qu'elles ont un pangénome quasi-fini, ce constat ne me semble pas universalisable à toutes les espèces procaryotes. En effet, ces dernières colonisent des milieux parfois très ouverts où se déploie une faramineuse diversité d'organismes vivants et donc d'acides nucléiques, de sorte qu'on peine à imaginer les caractéristiques biologiques spécifiques que pourraient avoir ces gènes en nombre fini, eux seuls susceptibles de s'intégrer dans les génomes d'une espèce donnée. Du reste, si dans l'absolu, le nombre de molécules sur Terre est fini, de sorte qu'à un temps t, le pangénome doit nécessairement l'être lui aussi, pour autant, celui-ci est si colossal que le considérer comme fini serait superfétatoire. Pour illustrer ce problème, prenons en exemple l'espèce *S. pneumoniae* déjà abordée précédemment. Pour ce cas, le modèle de HOGG et al., 2007 prédit que 95% du pangénome fini de l'espèce comprenant ≈5 117 gènes serait découvert avec seulement dix-sept souches. Le modèle de SNIPEN et al., 2009 semble plutôt en accord avec ces résultats en affirmant que η=5 100 et que ≈64% du pangénome serait observé avec six génomes. Pourtant des études plus récentes menées sur le camp de réfugiés de *Meala* en Thaïlande [1] indiqueraient un pangénome pour cette même espèce d'une taille de ≈12 184 gènes (TONDER et al., 2019). On conçoit dès lors les limites que soulève la notion de population de gènes finis, sauf à avoir séquencé tous les êtres vivants sur Terre.

3.4 La trichotomie : cœur, coquille et nuage

La grande majorité des études en pangénomique se contentent de partitionner le pangénome en deux, le génome cœur et le génome accessoire. Subvertissant cette habitude, les auteurs de HOGG et al., 2007 déterminaient arbitrairement que le pangénome pouvait se partitionner en $K = 7$ parties. En raffinant ce modèle, les auteurs de SNIPEN et al., 2009 suggèrent sur la base du critère BIC que le nombre de parties (K) dans un pangénome est variable et qu'en tout cas, il semble être supérieur à deux. En parallèle de ce travail, les auteurs de KOONIN et al., 2008 ; MAKAROVA et al., 2007 se sont pour leur part intéressés non pas au pangénome d'une espèce en particulier comme nous l'avons généralement vu jusqu'ici mais au pangénome total à l'échelle du règne, les archées pour

1. Accueillant ≈50 000 réfugiés principalement de la minorité Karènes persécutée par la junte militaire en Birmanie.

le premier et un mélange d'archées et de bactéries pour le second. En étudiant la répartition des familles de gènes orthologues dans les génomes procaryotes, les auteurs montrent que le nombre d'organismes ayant des gènes dans ces familles se distribue sous une forme de lettre **U** asymétrique où la hampe de gauche culminerait plus haut que celle sur la droite de la lettre. Cette distribution peut ainsi se décomposer en au moins trois fonctions exponentielles s'adaptant aux données, visible sous forme de droite à l'échelle logarithmique (voir à la gauche de la figure 3.9) :

1. Le **génome cœur** : représentant la petite portion de familles de gènes orthologues universelles, c'est-à-dire le noyau de gènes communs aux deux règnes bactérien et archéen, soit ≈70 familles de gènes orthologues procaryotes et ≈230 familles pour les archées uniquement.
2. Le **génome coquille** [1] : représentant les familles de gènes orthologues modérément présentes dans les génomes, soit ici ≈5700 familles de gènes orthologues procaryotes et ≈2 200 familles d'archées.
3. Le **génome nuage** [2] : représentant les familles de gènes orthologues rares, c'est-à-dire partagés par un très petit nombre de génomes, soit ici ≈24 000 familles de gènes orthologues procaryotes et ≈5 200 familles d'archées.

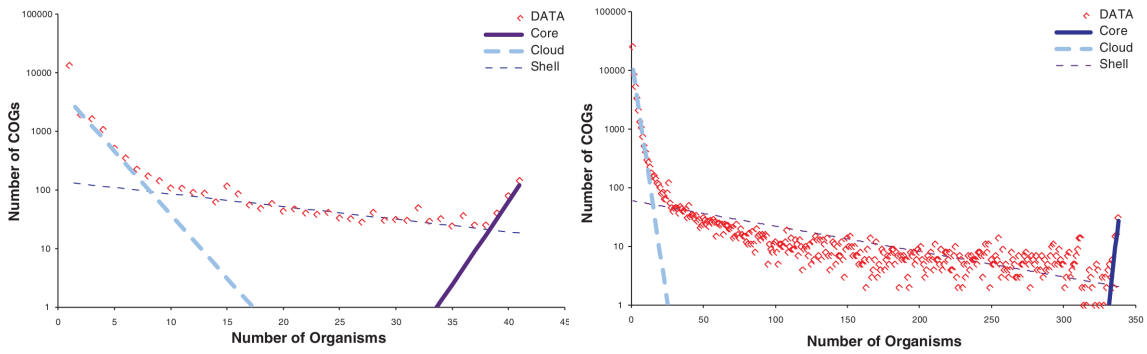


FIGURE 3.9 – Nombre d'organismes ayant des gènes dans 2 ensembles de familles de gènes orthologues. Dans la figure de gauche, les familles de gènes orthologues sont obtenues à partir de 41 génomes d'archées provenant de la base de données arCOG (MAKAROVA et al., 2007). Dans la figure de droite, elles sont obtenues à partir de 338 génomes procaryotes provenant de la base de données EggNOG (JENSEN et al., 2008). Les deux figures sont reproduites depuis le chapitre 3 du livre de KOONIN, 2011.

En outre, l'étude des auteurs de LAPIERRE et al., 2009 conduite elle aussi sur le pangénome de l'ensemble des génomes bactériens (573 génomes) disponibles dans les bases de données en 2009 confirme un profil de pangénome en forme de **U** asymétrique se partitionnant lui aussi en trois parties dont la partie génome nuage serait potentiellement infinie. Enfin, les auteurs montrent également un génome étendu (familles de gènes présentes dans ⩾99% des génomes) de 250 familles de gènes homologues.

1. Le terme génome coquille ayant été introduit par MAKAROVA et al., 1999 pour décrire les gènes variables entre les quatre premières espèces d'archées déjà séquencées en 1999 : *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Archaeoglobus fulgidus* et *Pyrococcus horikoshii*

2. Le terme provenant du lyrisme de LAPIERRE et al., 2009 : « This constant rain of genetic material on genomes from a cloud of frequently transferred genes enhances the chance of survival of species by introducing variability in the population ».

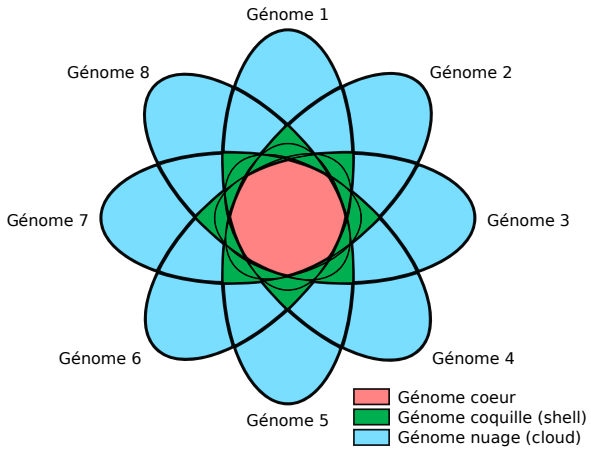


FIGURE 3.10 – Diagramme de Venn illustrant un pangénome ouvert partitionné en trois parties. En comparaison de la figure 2.4 page 44, le génome cœur battant de gènes essentiels à la vie des procaryotes est toujours présent mais celui-ci est maintenant revêtu d'un génome coquille, le tout confortablement emmitouflé dans un généreux génome nuage azurin libérant sa constante mousson de nouveaux gènes [1].

Sur la base de ces évolutions paradigmatiques, il est possible d'actualiser le partitionnement illustré sur la figure 2.4 page 44, on obtient alors la figure 3.10.

3.5 Le modèle infiniment plein de gènes (IMG) corrobore la trichotomie

3.5.1 Le modèle généalogique infiniment plein de gènes (IMG)

Contrairement aux approches abordées jusqu'ici, les auteurs de BAUMDICKER et al., 2012 proposent de modéliser la diversité du pangenome via un modèle de gains et de pertes de gènes à travers la généalogie d'une population. D'après les auteurs, ce modèle serait capable, à partir du spectre de la fréquence observée des familles de gènes, de prédire la taille moyenne des génomes, la taille du génome cœur ainsi que la taille du pangénome. Ce modèle s'appelle le modèle infiniment plein de gènes (Infinitely Many Genes : IMG) car les auteurs considèrent contrairement à HOGG et al., 2007 et SNIPEN et al., 2009 que le nombre de gènes dans le pangénome est potentiellement infini car le « réservoir » permettant d'obtenir des gènes par transfert horizontal ou bien par mutation est illimité [2].

Dans leur modèle, ils supposent que les génomes se composent de deux parties : les gènes nécessaires à la survie, soit le génome cœur ici noté c, et les gènes neutres qui peuvent être présents ou absents sans aucun avantage ou inconvénient pour l'individu. Ces gènes neutres seraient alors nécessairement dans le génome accessoire. Pour modéliser l'évolution du génome accessoire, les auteurs supposent que les nouveaux gènes sont acquis (par mutation ou bien à partir d'une source externe) avec une probabilité u et que les gènes existants sont perdus avec une probabilité v pour chaque génération. À l'échelle d'une population importante et constante N_e, les auteurs introduisent

1. Pour filer le lyrisme de LAPIERRE et al., 2009.
2. Les auteurs s'inspirent du modèle « Infinitely Many Sites », très utilisé en génétique des populations (HEIN et al., 2005).

les variables $\theta = 2N_e u$ et $\rho = 2N_e v$. Sur $2N_e$ générations le long d'une même lignée, θ correspond au nombre moyen de gènes acquis et ρ correspond au taux de perte. Autrement dit, si une lignée possède x gènes, elle gagne θ nouveaux gènes et perd ρx gènes en moyenne en $2Ne$ générations. Par conséquent, la taille d'équilibre du génome accessoire serait $x = \frac{\theta}{\rho}$ où les gains seraient compensés par les pertes de gènes. Pour affirmer cela, les auteurs se basent sur l'article de HUSON et al., 2004 qui avait montré que la taille du génome accessoire d'un individu à l'équilibre suit une loi de Poisson de paramètre $\frac{\theta}{\rho}$.

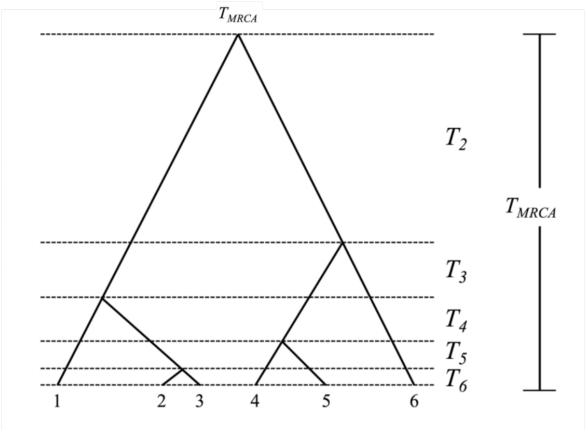


FIGURE 3.11 – Figure reproduite et adaptée depuis VILLANEA, 2019 illustrant un arbre coalescent de six organismes. En commençant au présent, le processus de fusion de lignées démarre à la base de l'arbre, et à mesure que les lignées se fusionnent remonte jusqu'à l'ancêtre commun le plus récent au sommet (Most Recent Common Ancestor : MRCA). Chaque processus de fusion de lignées prend un temps T_i de plus en plus grand à mesure que l'on remonte dans l'arbre.

En supposant une évolution neutre, les auteurs modélisent sur la base de la théorie coalescente de KINGMAN, 1982, la généalogie théorique sous-jacente reliant les individus par un arbre aléatoire appelé le « coalescent ». Pour une population d'organisme de taille N_e dont provient un échantillon de taille n, cette généalogie correspond à un arbre ultramétrique [1] aléatoire (exemple sur la figure 3.11). À partir de l'échantillon originel de taille n, pour construire l'arbre, les lignées ancestrales de deux organismes choisis au hasard sont fusionnées. Le temps T_i nécessaire pour cette fusion suit une distribution exponentielle $\lambda_n e^{-\lambda_n t}$ dont le paramètre $\lambda_n = \binom{n}{2}$. Puis sur la base des $n-1$ lignées restantes, le processus est recommencé avec un temps T_{n-1} suivant un paramètre cette fois-ci de $\lambda_{n-1} = \binom{n-1}{2}$. Et ainsi de suite pour les $n-2$ lignées restantes avec un paramètre $\lambda_{n-2} = \binom{n-2}{2}$. Le processus est arrêté quand l'ancêtre commun le plus récent à tous les organismes n est atteint. On note que comme les coefficients λ décroissent car $\binom{n}{2} > \binom{n-1}{2} > \binom{n-2}{2} > \cdots > \binom{2}{2}$, alors le changement de forme de la distribution exponentielle induit a pour conséquence que le temps de coalescence $T_n < T_{n-1} < T_{n-2} < \cdots < T_2$ est de plus en plus grand à mesure qu'on remonte dans l'arbre. Le processus est arrêté quand l'ancêtre commun le plus récent est atteint. Sur cet arbre, une branche de taille 1 correspond à N_e générations. Dès lors, en reprenant les égalités précédemment introduites, on peut dire que le long des lignées de cet arbre, les gains de gènes se produisent avec

1. Cela signifie que toutes les distances de la racine aux feuilles de l'arbre sont égales.

une probabilité $u = \frac{\theta}{2}$ et les pertes avec une probabilité $v = \frac{\rho}{2}$ toutes les N_e générations. Dans leur notation, les auteurs utilisent $\mathbb{E}_{\theta,\rho,c}[.]$ pour désigner la moyenne des trois sources d'aléas (probabilité de gains, de perte, de la généalogie) alors qu'il écrivent $\mathbb{E}_{\theta,\rho,c}[.|\tau]$ lorsqu'ils n'ont que deux source d'aléas car le vrai arbre généalogique τ est connu.

Pour exprimer les métriques du pangénome, la notation G^n indique le nombre de familles de gènes total dans le pangénome tandis que G^n_k correspond aux nombres de familles de gènes trouvés dans exactement k génomes sur n. On notera que $G^n = \sum_{a=1}^{n} G^n_a$. Sur la base de ce formalisme, les auteurs montrent les relations suivantes [1] :

— La taille du pangénome est égale à :

$$\mathbb{E}_{\theta,\rho,c}[G^n] = c + \theta \sum_{a=0}^{n-1} \frac{1}{a+\rho} \tag{3.7}$$

— Le nombre de gènes présents dans exactement k génomes peut être calculé :

— Si $1 \leqslant k < n$ par :

$$\mathbb{E}_{\theta,\rho}[G^n_k] = \frac{\theta}{k} \frac{n \ldots (n-k+1)}{k(n-1+\rho)\ldots(n-k+\rho)} \tag{3.8}$$

— Et si $k = n$ (c'est-à-dire, le génome cœur) par :

$$\mathbb{E}_{\theta,\rho,c}[G^n_n] = c + \theta \sum_{a=0}^{n-1} \frac{1}{a+\rho} \tag{3.9}$$

— Enfin, le nombre de gènes nouveaux apportés par le $n^{\text{ème}}$ génome appelé S_n est donné par :

$$\mathbb{E}_{\theta,\rho}[S_n] = \frac{1}{n}\mathbb{E}_{\theta,\rho}[G^n_1] = \frac{\theta}{n-1+\rho} \tag{3.10}$$

Toutes ces métriques sont ainsi obtenues à partir des paramètres θ et ρ, qu'il faut, dès lors, estimer. Pour cela, les auteurs ont besoin de deux types d'informations : d'une part le profil réel d'un ensemble de génomes (modélisé avec les mêmes équations 3.7, 3.8, 3.9 et 3.10, en remplaçant G par g) mais aussi d'un arbre réel décrivant la généalogie des génomes. Cet arbre τ doit également être ultramétrique pour permettre une comparaison avec l'arbre théorique.

Les auteurs conseillent ainsi l'utilisation du logiciel ClonalFrame (DIDELOT et al., 2007) sur les gènes conservés uniquement pour obtenir cet arbre ultramétrique. Sur la base de ce vrai arbre τ, on peut ainsi obtenir les vrais temps entre chaque fusion de lignées $(t_2, \ldots, t_i, \ldots, t_n)$ et ainsi injecter ces valeurs dans le modèle, de sorte que $T_2 = t_2, \ldots, T_i = t_i, \ldots, T_n = t_n$. Rappelons qu'à l'équilibre le nombre de gènes présents dans un génome accessoire suit une loi de Poisson de paramètre $\frac{\theta}{\rho}$. Dès lors, les auteurs vont estimer les paramètres θ de leur modèle en maximisant la log vraisemblance du modèle exprimée avec la formule suivante :

1. Voir le matériel supplémentaire de l'article de BAUMDICKER et al., 2012 pour en obtenir les démonstrations.

$$\ell(\theta,\rho \mid g_1^n,\ldots,g_{n-1}^n,\tau) = b + \sum_{k=1}^{n-1} \gamma_k^{(n)}(\theta,\rho,\tau) + g_k^n \log(\gamma_k^{(n)}(\theta,\rho,\tau))$$

où $\gamma_k^{(n)}(\theta,\rho,\tau)$ est le résultat de la loi de Poisson paramétrée par $\frac{\theta}{\rho}$ et τ et b une constante indépendante de θ et ρ.

Pour ce qui est du génome cœur, il est obtenu par :

$$\hat{c} = g_n^n - \gamma_n^{(n)}(\theta,\rho,\tau)$$

Enfin, ils estiment la taille de la population procaryote via cette formule :

$$\hat{N_e} = \frac{1}{n-1} \sum_{i=2}^{n} \binom{i}{2} T_i$$

De plus, une fois les paramètres maximisant la vraisemblance $\hat{\theta}$ et $\hat{\rho}$ estimées, les auteurs proposent un test statistique permettant de vérifier l'hypothèse de neutralité sous-jacente à leur modèle.

$$\chi^2 = \sum_{k=1}^{n-1} \frac{(G_k^n - \mathbb{E}_{\theta,\rho}[G_k^n])^2}{\mathbb{E}_{\theta,\rho}[G_k^n]}$$

Les auteurs vont ensuite appliquer ce modèle sur deux ensembles de génomes de cyanobactéries : onze génomes du genre *Prochlorococcus* d'une part et onze génomes du genre *Synechococcus* d'autre part. Pour commencer, les auteurs vont utiliser leur test de neutralité sur ces deux ensembles. Le résultat du test sur les génomes du genre *Prochlorococcus* suggère une évolution plutôt neutre, le modèle peut donc y être appliqué (voir les prédictions du modèle pour le genre *Prochlorococcus* sur le tableau 3.3 page suivante). En effet, l'hypothèse nulle de neutralité n'est pas à rejeter (valeur $p = 0.63$). En revanche, c'est précisément l'inverse qui est constaté sur le genre *Synechococcus* où l'hypothèse nulle de neutralité semble à rejeter (valeur $p = 0.01$), de sorte que les conditions d'application du modèle ne semblent pas réunies pour cette espèce. Le rejet de l'hypothèse de neutralité peut notamment être attribué aux possibles avantages ou désavantages sélectifs qu'apporteraient les gènes accessoires pour l'évolution du genre *Synechococcus*. On peut regretter cependant qu'en 2010, date de publication de l'article, les auteurs n'aient pas utilisé leur modèle sur davantage de données pour avoir une meilleure idée des conséquences de ce modèle sur différentes espèces. Sur le genre *Prochlorococcus*, le modèle IMG indique une taille de pangénome qui semble se placer entre les modèles de HOGG et al., 2007 et le modèle de TETTELIN et al., 2008. Il faudrait probablement davantage de génomes, pour pouvoir vérifier ces prédictions.

Métriques	Valeur Observée	Modèle IMG avec arbre fixé	Modèle de TETTELIN et al., 2008	Modèle de HOGG et al., 2007
Nombre de gènes par génome	2 019	2 033	NA	2 032
Taille du pangénome pour 11 génomes	5 025	5 245	5 041	5 023
Taille du pangénome pour 1000 génomes	NA	15 225	28 051	9 421
Taille du pangénome pour $\hat{N}_e$ génomes ($\hat{N}_e \approx 1^{11}$)	NA	57 792	15 337 650	9 421

TABLE 3.3 – Prédiction de la taille du pangénome pour onze génomes du genre *Prochlorococcus* avec un nombre croissant de génomes selon trois des principaux modèles évoqués jusqu'ici. Il faut rappeler, même si cela n'est pas parfaitement comparable, que le modèle de SNIPEN et al., 2009 sur douze génomes de *Prochlorococcus marinus* donnait un pangénome observé de 7 567 et un pangénome théorique de 13 472 (voir le tableau 3.2 page 56). La valeur de $\hat{N}_e$ est celle que fournit le modèle IMG.

Pour conclure, sur ce modèle, il faut garder en tête l'hypothèse forte sous-jacente à ce formalisme qui édicte que le génome cœur est absolument essentiel à la survie de l'organisme d'une part mais surtout, ce qui est plus discutable, que la totalité du génome accessoire a un impact évolutif parfaitement neutre. Comme nous l'avons vu, ce ne semble pas être le cas pour le genre *Synechococcus*.

3.5.2 Modèle d'évolution totalement neutre

Quelques mois plus tard, l'article de HAEGEMAN et al., 2012 est venu suggérer qu'un modèle d'évolution totalement neutre même pour le génome cœur, contrairement au modèle IMG, était également susceptible de générer des profils de pangénomes en forme de **U** pour six espèces : *Bacillius anthracis*, *Escherichia coli*, *Staphylococcus aureus*, *Streptococcus pneumonia*, *Streptococcus pyogenes*, *Neisseria meningitidis*. Ce modèle prend de plus en compte les possibles variations du nombre d'individus au cours des générations contrairement au modèle IMG. Ces résultats tendent à confirmer la pertinence des modèles à évolution neutre comme le modèle IMG. Néanmoins les auteurs notent qu'ils n'expliquent pas l'asymétrie du profil en **U**. Peut-être que la modélisation de plusieurs parties dans le pangénome ayant des taux de gains et de pertes différents pourrait améliorer la capacité prédictive de ce type de modèle ?

3.5.3 Impact d'une approche trichotomique sur les prédictions du modèle IMG

À travers leur article, les auteurs de COLLINS et al., 2012 vont appliquer le modèle IMG de BAUMDICKER et al., 2012 à 172 génomes complets appartenant à sept espèces de la classe des Bacilli. De plus, les auteurs souhaitent tester si des évolutions du modèle IMG pourraient en améliorer les prédictions. Ainsi, les auteurs proposent d'évaluer l'impact sur les prédictions du modèle IMG d'un partitionnement du pangénome en trois parties au lieu de deux. Ces modifications, faisant bien sûr écho aux articles de KOONIN et al., 2008 ; LAPIERRE et al., 2009 ; MAKAROVA et al., 2007 (évoquées dans la sous-section 3.4 page 57), permettraient de comparer la différence entre le modèle IMG originel où les familles de gènes sont soit dans le génome cœur (c'est-à-dire les gènes essentiels), soit dans le génome accessoire (c'est-à-dire les gènes à impact évolutif neutre) avec un modèle identique hormis que le génome accessoire est ici divisé en deux parties : les

familles de gènes à acquisitions et pertes de gènes rapides (la partie à évolution rapide) *versus* celles à acquisitions et pertes de gènes lentes (la partie à évolution lente). De la sorte, chacune de ces deux parties du génome accessoire possède ainsi ses propres paramètres θ et ρ.

Pour estimer la qualité des prédictions du modèle les auteurs vont mesurer si la courbe de raréfaction du pangénome – c'est-à-dire la courbe de croissance du pangénome lors de l'ajout de génomes – prédite par le modèle possède la même forme que celle empiriquement observée. Pour cela, les auteurs calculent l'écart quadratique (Root Mean Square : RMS) entre la courbe prédite par le modèle IMG et la moyenne des mesures à chaque ajout successif de génomes à la courbe de raréfaction.

Clade	RMS avec deux parties	RMS avec trois parties
Staphylococcus aureus	1.418	**0.088**
Streptococcus pyogenes	1.021	**0.027**
Streptococcus pneumoniae	1.318	**0.034**
Bacillus cereus	1.506	**0.040**
Genre Listeria	0.538	**0.003**
Genre Staphylococcus	1.428	**0.035**
Genre Streptococcus	1.538	**0.119**
Famille des Bacillaceae	1.834	**0.020**
Famille des Lactobacillaceae	1.076	**0.078**
Classe des Bacilli	0.863	**0.065**

TABLE 3.4 – Tableau illustrant, pour dix clades différents, l'écart quadratique (RMS) entre les courbes de raréfaction observées et celles prédites par un modèle IMG à deux parties et le même modèle à trois parties. Les valeurs minimisant le RMS sont indiquées en gras.

Les résultats obtenus (présentés dans le tableau 3.4) montrent un RMS dont la médiane est 38 fois plus faible avec trois parties au lieu de deux, soit un très net avantage prédictif pour cette modification du modèle IMG. Cela corrobore les études précédentes allant dans le sens de l'abandon de la dichotomie génome cœur/génome accessoire. Par ailleurs, les auteurs constatent que si à l'échelle du pangénome, les familles de la partie à évolution rapide sont majoritaires par rapport aux familles de la partie à évolution lente, ce n'est pas le cas à l'intérieur des génomes où les familles de la partie lente semblent majoritaires. Bien que nous ayons ici un modèle neutre, le fait d'avoir des différences entre les parties rapides et lentes suggère d'interpréter ce phénomène comme sélectif. Les gènes de la partie lente seraient bénéfiques pour certains des génomes probablement associés à des milieux écologiques spécifiques. Par conséquent, ces gènes auraient alors une grande probabilité d'être conservés plus longtemps dans le clade.

Par ailleurs, les auteurs suggèrent qu'il y a beaucoup d'informations dans les schémas de présence/absence des gènes qui ne sont pas visibles à travers le profil du pangénome et qui pourraient être exploitées pour améliorer la modélisation. Cette remarque m'invite à dire au lecteur qu'exploiter ces informations sera une partie des travaux développés dans la partie II page 99, de ce travail de thèse. Enfin les auteurs nous alertent sur l'importance des arbres qui sont fournis au modèle IMG, ces derniers ayant des impacts importants sur ses prédictions comme nous allons le voir.

3.5.4 Impact de trois types d'arbres différents sur les prédictions du modèle IMG

En plus du nombre de parties utilisées, les auteurs de COLLINS et al., 2012 font la remarque que la forme de l'arbre semble également cruciale dans le modèle IMG. Ainsi, comme précédemment, ils souhaitent calculer l'impact de plusieurs formes d'arbre sur la courbe de raréfaction du pangénome prédite par le modèle. Trois types d'arbres sont donc comparés dans cet article (voir également la figure 3.12) :

(a) Les arbres coalescents (soit le modèle originel de BAUMDICKER et al., 2012) : sur un arbre coalescent, plus le temps passe depuis l'ancêtre commun, plus les branches sont courtes, ce qui correspond à un pangénome dont la taille augmenterait logarithmiquement avec le nombre de génomes.

(b) Les arbres-étoiles : sur un arbre en étoile, la taille du pangénome augmente linéairement avec le nombre de génomes car la longueur de chaque nouvelle branche est constante peu importe le nombre de génomes déjà ajoutés, de sorte que le même nombre de gènes s'ajoute en moyenne sur chacune des branches de l'étoile. Les auteurs montrent que ce type d'arbre correspond implicitement au modèle de TETTELIN et al., 2005. En effet, les auteurs démontrent que le modèle IMG se simplifie dans le cas d'un arbre-étoile par les équations 3.2 et 3.1 page 47.

(c) Les arbres phylogénétiques de longueur de branche fixe prédéfinie : pour un arbre phylogénétique, la forme de l'arbre sera déterminante pour établir ou non une meilleure capacité prédictive. Ici les auteurs génèrent leur arbre à l'aide de l'outil PhyML (GUINDON et al., 2010) en se basant sur 55 gènes en copie unique dans les génomes qu'ils alignent avec l'outil MUSCLE (EDGAR, 2004).

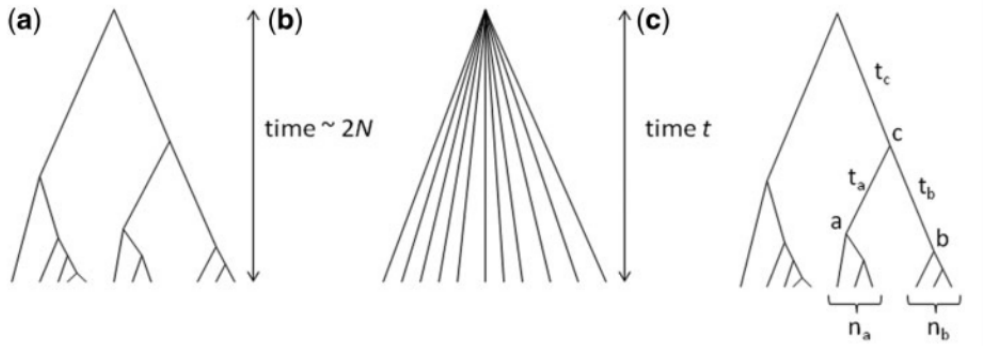


FIGURE 3.12 – Figure reproduite depuis COLLINS et al., 2012 illustrant la topologie ainsi que l'échelle de temps des trois types d'arbres comparés : (a) un arbre coalescent, (b) un arbre-étoile (c) un arbre phylogénétique.

Ainsi que nous l'avons vu précédemment, les auteurs ayant constaté que c'est le modèle à 3 parties qui semble produire les meilleures prédictions, la comparaison des trois arbres dans le tableau 3.5 page suivante, se fait pour les modèles IMG à trois parties uniquement. En analysant les valeurs RMS de ce tableau, les auteurs observent que la médiane du RMS avec un arbre coalescent est quinze fois plus faible qu'avec un arbre étoile, soit de très faibles capacités prédictives pour le modèle IMG basé sur un arbre-étoile. Puisque les auteurs avaient montré l'équivalence entre

Clade	RMS - arbre coalescent	RMS - arbre-étoile	RMS - arbre phylogénétique
Staphylococcus aureus	0.088	0.244	**0.023**
Streptococcus pyogenes	**0.027**	0.322	0.066
Streptococcus pneumoniae	**0.034**	0.375	0.071
Bacillus cereus	0.040	0.426	**0.019**
Genre Listeria	**0.003**	0.179	0.008
Genre Staphylococcus	**0.035**	0.344	0.310
Genre Streptococcus	**0.119**	2.247	0.123
Famille des Bacillaceae	**0.020**	1.785	0.030
Famille des Lactobacillaceae	**0.078**	1.928	**0.078**
Classe des Bacilli	0.065	2.497	**0.040**

TABLE 3.5 – Tableau illustrant l'écart quadratique (RMS) entre les courbes de raréfaction observées et celles prédites par un modèle IMG à 3 parties basé soit sur un arbre coalescent, soit sur un arbre-étoile ou enfin sur un arbre phylogénétique.

le modèle basé sur un arbre-étoile et le modèle TETTELIN et al., 2005, ils en déduisent que cela réfute les modèles de décroissance exponentielle en tant que modélisation correcte d'un pangénome. Pour ce qui est des résultats entre l'arbre phylogénétique et l'arbre coalescent, ceux-ci sont alternativement un peu mieux ou un peu moins bons de sorte que le modèle est globalement indifférent aux différences entre ces deux arbres. Pour tenter de les départager, les auteurs vont comparer leur capacité à prédire le profil observé des pangénomes. Dans ce cas aussi, les résultats sont en ballottage mais les auteurs tendent plutôt à donner l'avantage aux modèles coalescents en général et aux modèles à arbres phylogénétiques dans certains cas. Ces cas correspondent aux profils n'étant pas parfaitement dessinés en **U** mais étant plutôt en **W** comme c'est le cas pour l'ordre des Bacilicae dans l'article [1]. Les espèces ayant ces profils en **W** étant appelées « espèces non homogènes » dans l'article de MOLDOVAN et al., 2018.

En conclusion, les auteurs ont constaté que les variantes proposées du modèle IMG font de bonnes prédictions des courbes de raréfaction et de la forme du profil de fréquence des familles de gènes dans plusieurs ensembles de génomes bactériens. Les auteurs attribuent les avantages de leur modèle au fait qu'il soit basé sur des règles génétiques et donc qu'il serait plus utile que les approches correspondant simplement à des fonctions d'ajustement sur les données ou bien que les modèles statistiques s'appuyant sur des distributions n'étant associées à aucun mécanisme évolutif.

3.6 Rejet de l'hypothèse d'une évolution neutre

Les auteurs de LOBKOVSKY et al., 2013 sont relativement sceptiques par rapport aux modèles d'évolution neutre proposés par BAUMDICKER et al., 2012 ; COLLINS et al., 2012 ; HAEGEMAN et al., 2012. Les auteurs du modèle IMG eux-mêmes l'avaient d'ailleurs déjà rejeté pour le genre *Synechococcus*. Dans l'article de LOBKOVSKY et al., 2013, les auteurs introduisent un modèle autorisant les caractéristiques évolutives des nouveaux gènes à être non plus seulement neutres mais également avantageuses ou désavantageuses. Comme les auteurs du modèle IMG, ils vont étudier le modèle qu'ils proposent en comparant le profil théorique prédit par leur modèle aux vrais profils de pangénomes observés. Comme c'était le cas pour le modèle IMG, ces profils théoriques sont obtenus à partir d'un arbre phylogénétique couplé à un modèle de gains et de pertes de gènes (appelé

1. Voir la figure 6 de l'article de COLLINS et al., 2012 pour plus d'information.

modèle de taux de remplacement des gènes). En l'occurrence, les auteurs vont ici tester quatre modèles de ce type. Pour cela, ils étudieront le profil de pangénomes issus de 392 combinaisons de 10 génomes uniquement mais ayant une vaste gamme de proximité génétique allant de génomes de la même espèce à des génomes taxonomiquement très éloignés (par exemple entre bactéries et archées). Les auteurs admettent cependant ne pas pouvoir étudier plus de 10 génomes à la fois car leur modèle serait trop long à exécuter sur davantage de données.

Voici les quatre modèles étudiés :

A Modèle neutre où tous les gènes possèdent le même taux de remplacement (correspond au modèle IMG originel de BAUMDICKER et al., 2012).

B Modèle ayant un taux de remplacement suivant une distribution Gamma [1] où le paramètre d'échelle θ gouverne le taux d'évolution global tandis que le paramètre de forme β indique la force de la sélection.

C Modèle neutre où les gènes variables sont séparés en 2 classes ayant des taux de remplacement différents (correspond au modèle IMG modifié de COLLINS et al., 2012).

D Le modèle proposé par les auteurs dans lequel l'adaptation d'un organisme à son environnement correspond à la somme des contributions de chaque gène.

Le modèle qu'introduisent les auteurs est relativement complexe à décrire bien que relativement parcimonieux (il nécessite seulement 2 paramètres : f et v évoqués ci-dessous). Par souci de concision, je vais donc le décrire de façon résumée et j'invite le lecteur à consulter l'article LOBKOVSKY et al., 2013 pour plus de détails. Ce modèle (modèle D) considère une population de N génomes contenant chacun M gènes. Chaque gène possède une contribution additive fixe f à la capacité d'adaptation des organismes à leur environnement. La sélection et la dérive génétique sont modélisées selon un processus de Moran (MORAN, 1962). Dans ce processus itératif, la descendance de chaque individu est mutée à chaque génération, c'est-à-dire que chaque gène est remplacé par un autre selon une probabilité V. L'effet évolutif de ces nouveaux gènes est arbitrairement considéré comme suivant une distribution exponentielle – ainsi que le propose GILLESPIE, 1984 – car cette distribution semble permettre de modéliser un large éventail de contributions évolutives. Sur la base de ce modèle, les auteurs vont pouvoir étudier comment la contribution de chaque gène affecte le taux de remplacement et donc le profil du pangénome.

En analysant les résultats obtenus (exemple sur la figure 3.13 page suivante), les faibles capacités prédictives du modèle d'évolution neutre à une classe (le modèle A) sautent aux yeux des auteurs. En particulier ce modèle sous-estime notablement la taille du génome cœur. Pour ce qui est des trois autres modèles, ces derniers semblent plutôt bien prédire les profils empiriquement observés. Cependant le modèle D proposé par les auteurs a en moyenne donné un ajustement un peu meilleur que les modèles B et C. Pour conclure, les auteurs affirment donc que prendre en compte la sélection apporte une contribution substantielle pour prédire le profil de pangénome. C'est particulièrement le cas face au modèle d'évolution neutre à une seule classe de gènes de BAUMDICKER et al., 2012. Pour les auteurs, leur travail démontre qu'un modèle neutre sous-estime largement le génome cœur et que prendre en compte la sélection semble nécessaire pour correctement rendre compte du profil des pangénomes. On pourrait cependant leur objecter que par rapport au modèle d'évolution neutre

1. À ne pas confondre avec le paramètre γ utilisé dans la loi de Heaps.

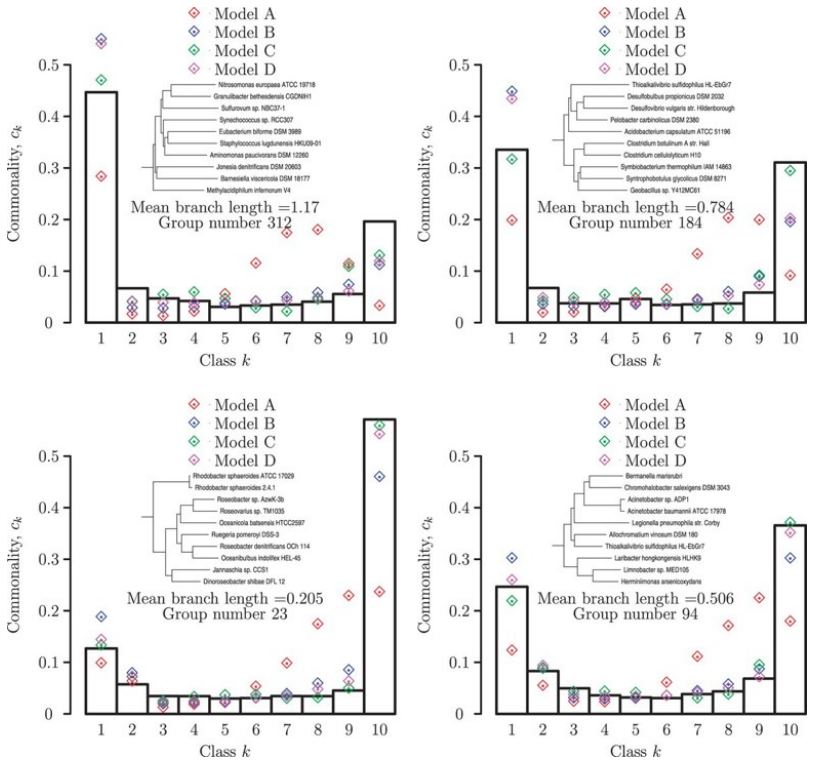


FIGURE 3.13 – Figure reproduite depuis LOBKOVSKY et al., 2013 illustrant les résultats obtenus par les 4 modèles testés par les auteurs en comparaison des profils de 4 pangénomes empiriquement observés. Ces 4 profils sont des exemples représentatifs des différents types de profils obtenus parmi les 392 groupes de 10 génomes étudiés.

à deux classes leur modèle ne semble pas apporter un gain considérable. Néanmoins on peut aussi dire que le fait de proposer qu'il y ait 2 classes de gènes dans le génome accessoire ayant une évolution considérée comme neutre et avec des dynamiques différentes, peut déjà être vu comme un ersatz de modèle sélectif.

3.7 Conclusion sur les modèles proposés

Nous l'avons vu, il existe une grande diversité de prismes et de méthodes permettant d'appréhender la biodiversité des pangénomes. On peut par exemple approcher cette dernière par la forme des courbes de raréfaction des différents compartiments du pangénome ou bien par la fréquence des différents gènes dans les espèces, c'est-à-dire le profil des pangénomes. En fonction du type de données prises en compte, on modélisera le pangénome par l'ajustement de fonctions sur des courbes, par des modèles statistiques ou bien par des modèles basés sur des hypothèses évolutives. Comme nous l'avons dit et discuté à travers ce chapitre, ces modèles sont basés sur de nombreuses hypothèses simplificatrices qu'il faut bien avoir à l'esprit quand on analyse leurs prédictions. Enfin, aucun des modèles présentés ne semble exploiter la totalité de l'information potentiellement accessible comme la matrice de présence/absence ou le contexte génomique des gènes par exemple.

Ces modélisations permettent de donner des éléments de réponses aux questions mentionnées au chapitre précédent. À la question : le pangénome est-il fini ou infini ? Je pense qu'il est raisonnable de répondre, comme j'ai eu l'occasion de le discuter, qu'un pangénome fini semble une hypothèse peu envisageable sur la base des prédictions des modèles basés sur cette hypothèse, puisqu'ils semblent en contradiction avec certaines observations. Pour ce qui est des questions sur la vitesse de croissance des pangénomes lorsqu'on séquence de plus en plus de souches et sur la stabilité du génome cœur, les modèles basés sur les courbes de raréfaction permettent de répondre que leur croissance semble suivre une loi de Heaps. Néanmoins, la forme de cette loi dépendra des espèces étudiées, certaines ayant des croissances plus rapides que d'autres. Enfin, quelles informations la notion de pangénome apporte-t-elle aux théories sur l'évolution des génomes ? Dans ce chapitre, les approches basées sur l'évolution neutre se sont ici opposées aux approches considérant un caractère adaptatif aux gènes. Nous avons notamment terminé ce chapitre en abordant le modèle de LOBKOVSKY et al., 2013 qui affirme que les gènes s'ajoutant aux pangénomes auraient un caractère adaptatif. Cette affirmation si elle était vérifiée aurait plusieurs conséquences biologiques, elle reste cependant très discutée et je vous propose d'aborder cette discussion à travers un nouveau chapitre.



Chapitre 4 : Le pangénome est-il adaptatif ou non ?

L'acquisition de nouveaux gènes accessoires peut avoir des effets bénéfiques, neutres ou délétères sur les capacités d'adaptation des organismes. Cependant la prévalence relative de ces différents types d'effets reste une question ouverte importante en microbiologie évolutive. Deux écoles s'affrontent : celle en faveur d'une acquisition de nouveaux gènes majoritairement neutres ou délétères (GOGARTEN et al., 2005) et une autre penchant pour le fait que les nouveaux gènes soient majoritairement bénéfiques (MCINERNEY et al., 2017). Chacune d'elles tente d'expliquer pourquoi les procaryotes ont des pangénomes sur la base de théories sur le caractère adaptatif des gènes transférés horizontalement. Nous verrons dans ce chapitre les bases de leur argumentaire respectif.

4.1 L'apport de la génétique des populations

4.1.1 Notions de base en génétique des populations

Si un variant purement neutre apparaît dans une population, il aura alors une fréquence initiale de $1/N$. Plaçons-nous dans l'hypothèse d'un modèle d'évolution neutre. La génétique des populations nous apprend que si la probabilité d'apparition d'un variant est u, alors pour une population effective à long terme de taille N, il aura une probabilité de se fixer par pure dérive génétique de $Nu \cdot 1/N = \frac{\not{N}u}{\not{N}}$, soit si on simplifie la formule, une probabilité u indépendante de la taille effective de la population N. Ainsi, les nouveaux variants obtenus pourraient rester très longtemps dans la population avant de pouvoir se fixer. Plus la population effective est grande et plus un nombre important de variants polymorphiques sera retrouvé. Un moyen simple d'estimer la taille de la population est donc d'estimer les variants synonymes qui y sont représentés.

Même si le lien n'est pas évident à première vue, je prie le lecteur d'accepter sans transition une rapide explication de ce qu'est la fluidité d'un pangénome afin de pouvoir relier ce concept avec la génomique des populations.

4.1.2 Estimation de la diversité pangénomique à travers sa fluidité

Les auteurs de KISLYUK et al., 2011 ont introduit une mesure qu'ils présentent comme robuste pour rendre compte de la diversité en gènes de chaque génome parmi un groupe de génomes (habituellement l'espèce). Cette métrique s'appelle la fluidité des génomes. Dès lors, les génomes ayant une petite fluidité correspondront aux espèces clonales, c'est-à-dire très stables tandis qu'à l'inverse les espèces à haute fluidité seront les espèces s'échangeant beaucoup de gènes et ayant donc une proportion de gènes variables très importante entre 2 génomes de la même espèce.

La fluidité des génomes (ϕ) se calcule ainsi :

$$\phi = \frac{2}{N(N-1)} \sum_{\substack{k=1 \\ l=1 \\ k<l}}^{N} \frac{U_k + U_l}{M_k + M_l}$$

où U_k et U_l correspondent au nombre de familles ayant des gènes dans k mais pas dans l pour le premier et inversement pour le second et où M_k et M_l indiquent le nombre total de familles présentes respectivement dans k et dans l.

On le voit la fluidité correspond donc à la moyenne du rapport des familles de gènes uniques sur la somme des familles de gènes totales, pour chaque paire de génomes présents dans une espèce. Cette valeur est donc égale à $\approx$0 pour des espèces clonales et peut monter jusqu'à $\approx$0.35 pour des espèces aux génomes très fluides.

4.1.3 Corrélation entre la fluidité des génomes et la diversité en mutations synonymes

En étudiant le pangénome de 90 espèces, les auteurs de ANDREANI et al., 2017 montrent qu'il existe une corrélation entre les espèces ayant un pangénome fluide et les espèces ayant une forte diversité en mutations synonymes (voir figure 4.1 page suivante). Nous l'avons dit, en génétique des populations, une forte diversité en mutations synonymes s'interprète comme le signe d'une population effective importante. Dès lors, les auteurs suggèrent que cette corrélation peut être expliquée par le fait que si une population effective est plus grande, alors de nombreux variants polymorphiques sont ajoutés à la population sur la base de la théorie de l'évolution neutre (KIMURA, 1983) évoquée précédemment à la sous-section 4.1.1 page précédente. Ils en déduisent donc que la diversité en mutation synonyme survient en l'absence de sélection et comme cela est en corrélation avec la fluidité du génome, alors le génome accessoire est probablement neutre également.

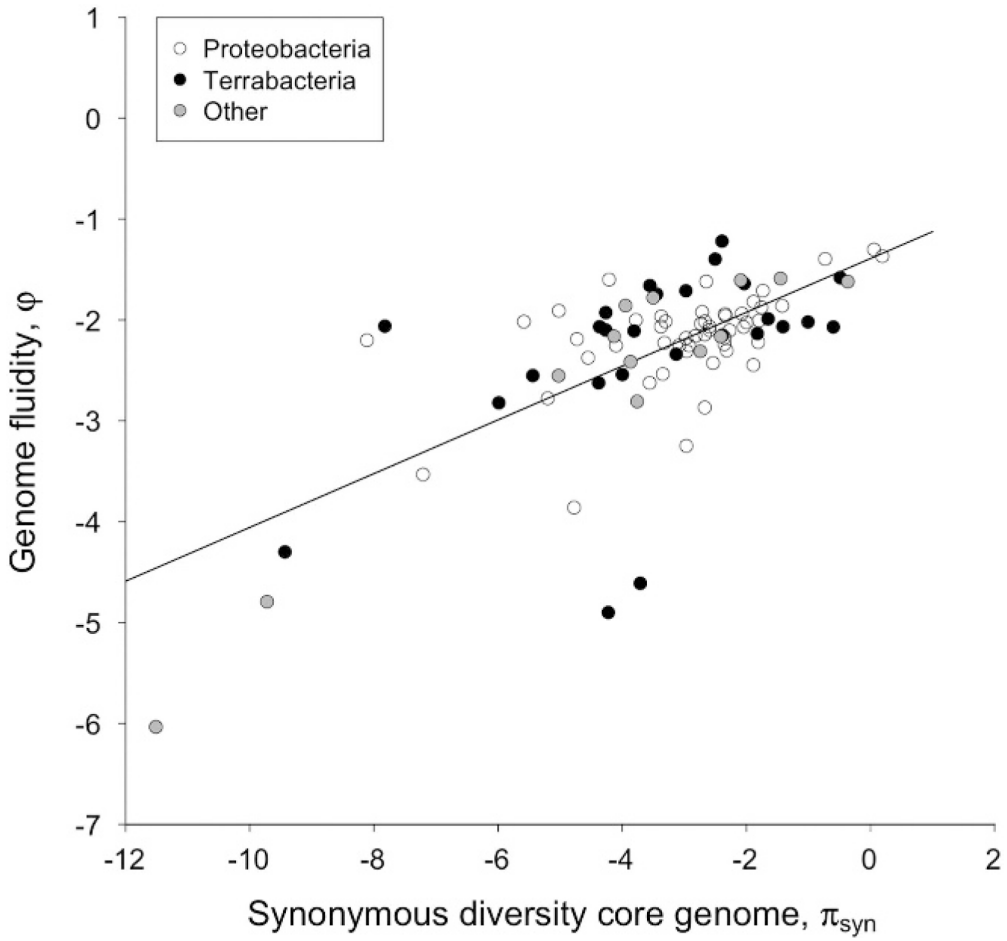


FIGURE 4.1 – Figure reproduite depuis ANDREANI et al., 2017 illustrant la corrélation entre la fluidité et la diversité en codons synonymes dans le génome cœur de 90 espèces. Les axes sont transformés logarithmiquement. La droite indique une corrélation linéaire d'équation $ln(\phi) = -1.39(0.12) + 0.27(0.03) \cdot ln(\pi_{syn})$ avec une part de variance expliquée de $R^2 = 0.45$ (valeur-p<0.001).

4.1.4 Critique de la génétique des populations pour expliquer le pangénome

L'interprétation des auteurs de ANDREANI et al., 2017 les conduisant à tenir leur corrélation comme validant le modèle neutre de Kimura est à modérer pour ce qui est du pangénome. D'une part, il faut rappeler qu'à l'origine, ce modèle a été introduit par Kimura principalement pour modéliser l'apparition de variants nucléotidiques. (KIMURA, 1983). On ne peut donc pas affirmer en l'état que cette théorie est aussi applicable pour des gènes. D'autre part, les gènes susceptibles d'apparaître chez un individu ne proviennent pas uniquement du réservoir de gènes de la population propre à l'espèce : ces derniers peuvent également être obtenus par des transferts horizontaux depuis d'autres populations. Ainsi la population effective ne devrait pas être seulement considérée en termes d'individus mais également en termes de gènes disponibles dans l'environnement, ce qui a pour conséquence que le formalisme issu de la génomique de population n'est pas nécessairement applicable *ceteris paribus*.

4.2 L'acquisition de nouveaux gènes est délétère en soi

Comme nous l'avons vu le modèle d'évolution neutre a été introduit pour modéliser l'apparition de variants nucléotidiques. Mais dans le cas qui nous intéresse, c'est-à-dire, celui de l'apparition d'un nouveau gène, on peut imaginer que sa seule présence dans un génome et indépendamment de son effet, implique un impact intrinsèquement négatif sur l'organisme, contrairement à un variant nucléotidique. En effet comme le rappellent les auteurs de MCINERNEY et al., 2017, ce nouveau gène augmentera la taille du génome devant être répliqué. De plus, si ce gène est fonctionnel, il sera transcrit, et de même, s'il est codant, il sera potentiellement traduit. Ainsi, avant même de prendre en compte son apport, celui-ci implique donc nécessairement un « surcoût » métabolique non négligeable pour l'organisme. Or, si ce gène n'exprime qu'une fonction ayant un caractère neutre pour l'organisme, alors à cause de son impact négatif propre, le gène devrait tendre à être éliminé. De plus, si la maintenance de ces gènes avait un coût nul, alors on pourrait penser que la taille des génomes procaryotes augmenterait indéfiniment. Or ce n'est évidemment pas ce que l'on observe, les génomes procaryotes ne dépassant que très rarement une taille de 10 Mb. Il a d'ailleurs été montré que les génomes bactériens ont une forte tendance à supprimer l'ADN inutile des génomes, confirmant ainsi l'hypothèse que les gènes neutres seraient rapidement éliminés (KUO et al., 2009). Enfin, les auteurs de SELA et al., 2016 ont montré qu'il y avait une corrélation négative entre le ratio dN/dS (ratio entre le taux de mutations non synonymes et de mutations synonymes) et la taille des génomes. Autrement dit, selon eux, la pression de sélection augmenterait sur les grands génomes. Cela est étonnant car on pourrait au contraire penser que plus la sélection est forte et plus les génomes sont petits. De plus, cette conclusion tenant à relier le ratio dN/dS à une pression de sélection reste elle aussi à modérer sachant que cette théorie – certes couramment admise entre individus d'espèces différentes – ne semble pas s'appliquer aux génomes très proches phylogénétiquement (KRYAZHIMSKIY et al., 2008). Au final, la conclusion que l'on pourrait en tirer est que les gènes acquis dans ces génomes plus grands auraient un caractère adaptatif positif plus fort susceptible de lutter contre cette forte sélection purificatrice.

4.3 Pourquoi donc les procaryotes ont-ils des pangénomes ?

Dans un cadre où une forte pression de sélection empêche la fixation de gènes ayant un caractère évolutif neutre ou légèrement délétère, on s'attendrait alors à ce que très peu de gènes acquis aléatoirement par transferts horizontaux soient en mesure d'être maintenus dans la population. En effet, ces derniers seraient le plus souvent incapables d'être en mesure de compenser leur impact négatif par un avantage sélectif supérieur à leur coût de maintenance dans les génomes. Pour être conservés, ces gènes vont donc devoir posséder *a minima* des avantages sélectifs en mesure de compenser leur coût d'entretien propre. On s'attend donc à ce que très peu de ces gènes puissent être conservés dans les génomes. De plus, les conséquences même du mécanisme d'obtention de nouveaux gènes par le mode du transfert horizontal de gènes semblent avoir un impact négatif supplémentaire (BALTRUS, 2013).

On pourrait donc penser que comme ces transferts sont majoritairement délétères, les cellules pourraient tenter en réponse de diminuer leur capacité à acquérir de nouveaux gènes. Or ce n'est pas ce que l'on observe. Au contraire, les procaryotes font beaucoup de transferts et c'est même la principale source de l'extension de leur répertoire génique (ROCHA, 2008).

4.3.1 L'hypothèse du gène égoïste

On pourrait répondre avec la théorie du gène égoïste de DAWKINS, 1976 que les organismes ne sont pas en mesure de contrôler les gènes qu'ils présentent dans leur génome. Ainsi, les gènes possèdent plus ou moins leur capacité évolutive propre et vont tenter de se reproduire le plus possible. Avec cette théorie, les gènes non autoréplicatifs seront davantage répliqués s'ils apportent un avantage à leur hôte lui permettant de croître plus rapidement, à l'inverse pour le cas des gènes réplicatifs comme les transposons, les plasmides, les prophages, leur reproduction pourrait se faire potentiellement au détriment de leur hôte, un peu à l'image *mutatis mutandis* des « cancers » pour les eucaryotes. Cela expliquerait donc pourquoi des gènes, ayant un caractère évolutif délétère, seraient tout de même insérés en grande proportion dans les génomes, et donc l'existence d'un pangénome diversifié.

Selon les auteurs de MCINERNEY et al., 2017, cela leur semble improbable car les éléments génétiques égoïstes dans les génomes semblent minoritaires. Pour cela ils invoquent une étude montrant que l'essentiel des gènes transférés correspondent à des gènes de surface des cellules, des protéines régulatrices et des gènes de « processus cellulaires » (impliquant notamment des fonctions d'entretien de l'ADN et de pathogénicité), donc des gènes a priori très peu égoïstes (NAKAMURA et al., 2004). Il est vrai que dans cette étude, ces trois catégories de gènes additionnés (respectivement $\approx$13.8%, $\approx$11.0% et $\approx$10.0%) comprennent plus de gènes que les éléments potentiellement égoïstes (transposases, prophages, plasmides) qui totalisent $\approx$28.3%. Pour autant, cela reste tout même non négligeable. Enfin, dans une autre étude les auteurs de TOUCHON et al., 2009 ont montré que ces éléments égoïstes étaient largement dominants dans le génome accessoire d'*E. coli*. De sorte que cette hypothèse semble plausible pour expliquer une part de la diversité des pangénomes.

4.3.2 Les transferts horizontaux seraient bénéfiques

Les auteurs de MCINERNEY et al., 2017 proposent une autre explication à l'existence d'un pangénome. Pour eux, les gènes acquis par transferts horizontaux auraient, dans la majorité des cas, un caractère adaptatif positif suffisant pour contrebalancer la sélection. Pour supporter cette hypothèse, les auteurs se basent notamment sur les travaux des auteurs de KARCAGI et al., 2016 qui se demandent si la réduction des génomes est vraiment bénéfique. Pour vérifier leur hypothèse, ces derniers ont réduit expérimentalement le génome d'*Escherichia coli* en supprimant les régions ayant pour origine un transfert horizontal de gènes (jusqu'à 20% du génome). À travers cette expérience, les auteurs ont observé que la réduction forcée du génome a diminué l'adaptabilité des organismes même en conditions non stressantes. Ils en concluent que la réduction du génome des bactéries n'est pas bénéfique et qu'au contraire cela provoque une utilisation moins efficace de l'énergie de la cellule et une diminution de la tolérance au stress.

Les auteurs de Vos et al., 2017 reprochent à cette étude d'avoir consisté dans la délétion de nombreux gènes, y compris d'après eux des gènes du génome cœur sans pour autant apporter de précisions. Ils considèrent donc cette étude comme non concluante. En effet, cette étude tendrait à remettre en cause un certain consensus sur la caractère bénéfique des réductions de génomes (Lee et al., 2012 ; Wolf et al., 2013). Néanmoins, les auteurs de Mcinerney et al., 2017 en tirent des arguments pour considérer que les gènes transférés horizontalement pourraient être avantageux dans la majorité des cas.

4.3.3 Des gènes avantageux entraîneraient un balayage génétique

Partons de l'hypothèse que les gènes transférés horizontalement sont avantageux. Dans ce cas, l'introduction des gènes avantageux dans un contexte sélectif amènerait à un balayage sélectif autour du gène. Ainsi, on assisterait à une purification de la population et donc à une chute importante de la biodiversité génétique des génomes (comme illustré sur la partie (a) de la figure 4.2 page suivante) et ce même dans des génomes présentant un fort taux de recombinaison (Bendall et al., 2016 ; Charlesworth, 2009). Dès lors, cette baisse de la diversité aurait probablement tendance à fortement limiter la taille des pangénomes. Comment alors continuer à soutenir la thèse que les gènes transférés horizontalement seraient bénéfiques ?

4.3.4 Les migrations et les transferts horizontaux de gènes diversifient le pangénome

Une autre hypothèse, pour expliquer l'existence des pangénomes malgré la purification sélective, est proposée par les auteurs Mcinerney et al., 2017. Ces derniers invoquent un modèle rendant théoriquement compatible l'augmentation de la biodiversité avec l'existence de variants adaptatifs obtenus par transfert. En effet, le modèle de Niehus et al., 2015 permettrait de concilier le fait que les transferts horizontaux soient bénéfiques avec l'existence d'une grande diversité pangénomique dans le cas où ces adaptations entraînent des spécialisations en de nombreuses niches écologiques.

Tout d'abord, les auteurs de Niehus et al., 2015 confirment que les transferts horizontaux avantageux seraient rares s'il n'y avait pas de migration. En effet, d'après leur modèle, le transfert d'un gène bénéfique au sein d'un organisme dans une communauté sous sélection serait limité par des balayages sélectifs susceptibles de propager l'allèle verticalement. En effet, la transmission verticale propagerait plus rapidement ce trait par la croissance de l'individu avantagé que ne le ferait le transfert horizontal de ce trait aux autres souches (comme illustré sur la partie (a) de la figure 4.2 page ci-contre). On assisterait donc, comme nous l'avons dit, à une purification de la population.

Quand il se produit des migrations dans une population, les auteurs de Niehus et al., 2015 ont montré que ce phénomène de purification semble contrecarré. En effet, les migrations entraîneraient selon leur modèle de grandes quantités de transferts horizontaux ayant dès lors la possibilité de pouvoir transférer le trait génétique avantageux plus rapidement que la transmission verticale. Ainsi cela permettrait donc de maintenir les fonds génétiques des individus présents dans la population microbienne tout en diffusant le trait génétique avantageux (comme illustré sur la partie (b) de la figure 4.2 page suivante). De la sorte, la biodiversité serait ainsi conservée.

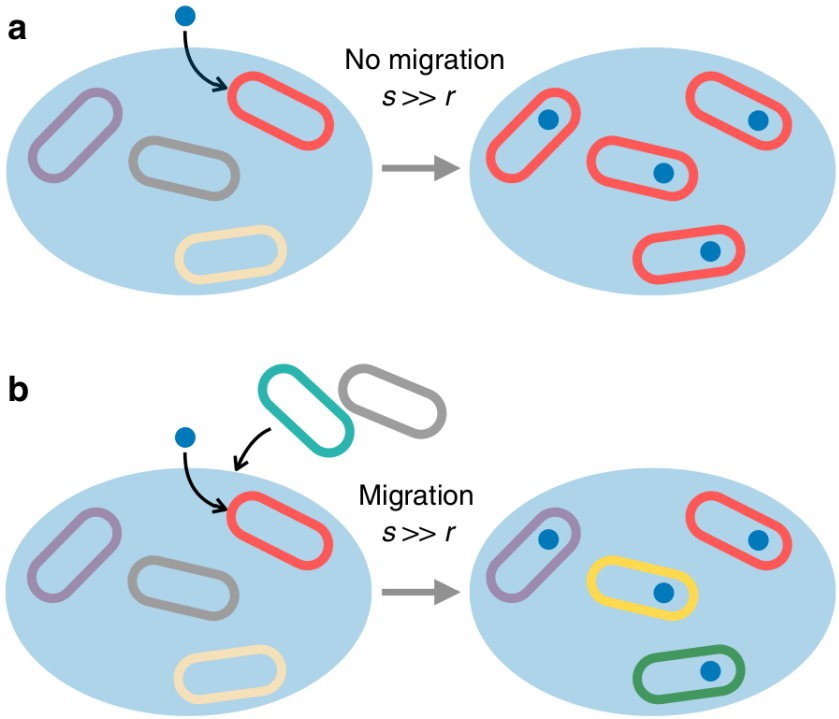


FIGURE 4.2 – Figure reproduite depuis NIEHUS et al., 2015 illustrant la sélection (taux de sélection s) de gènes transférés (taux de transfert r) dans plusieurs modèles écologiques. (a) Absence de migration : un trait positif (point bleu) transféré dans un génome entraîne un balayage sélectif du génome dont lequel il apparaît pour la première fois (cellule rouge) ce qui élimine la biodiversité dans la population (si $s >> r$). (b) Présence d'une migration : l'immigration du nouveau génotype dans une niche entraîne une augmentation des transferts horizontaux. Si un trait positif transféré horizontalement est présent, il sera partagé avec tous les génomes de la population, donc dans des fonds génétiques diversifiés.

Les auteurs de MCINERNEY et al., 2017 en concluent que cela expliquerait les raisons pour lesquelles les génomes ayant des niches écologiques très restreintes (comme par exemple le pathogène intracellulaire obligatoire *C. trachomatis*) ont de très petits pangénomes très proches de la taille du génome cœur tandis que les espèces vivant dans des niches écologiques très diversifiées – comme les bactéries marines ou bien du sol – présentent, pour leur part, des pangénomes très ouverts.

4.3.5 Conclusion et réserves sur le caractère adaptatif des pangénomes

D'après les auteurs de MCINERNEY et al., 2017, le pangénome serait le résultat d'une évolution adaptative positive et non d'une évolution neutre (ou délétère). On pourrait objecter aux auteurs de MCINERNEY et al., 2017 un fait qu'ils ne semblent pas avoir pris en compte : le profil de pangénomes se distribue sous la forme d'un **U** ce qui semble indiquer qu'une grande partie des gènes rares (hampe gauche du **U**), donc potentiellement d'acquisition récente (le génome nuage), ne percolent pas dans le génome coquille (la base du **U**). Pourquoi donc ces nouveaux gènes seraient-ils éliminés ou ne se propageraient pas davantage s'ils sont si avantageux pour les organismes ?

Plusieurs études de génomique comparée dont LUKJANCENKO et al., 2010 ont déjà montré que, dans les génomes étroitement apparentés, on constate que de nombreux gènes obtenus par transfert sont relativement transitoires et rapidement supplantés par d'autres gènes eux aussi obtenus par transfert, augmentant ainsi à nouveau la taille du génome nuage à l'échelle de la population. Dès lors cela tendrait plutôt à réfuter l'idée que les gènes acquis horizontalement seraient quasi systématiquement bénéfiques. Il reste que le raisonnement de MCINERNEY et al., 2017 même s'il est contestable sur bien des points (ce qui a provoqué une vive réaction de la part de VOS et al., 2017 notamment) semble plausible sur bien des points.

Pour terminer de répondre à la question posée par le titre de ce chapitre, il serait donc souhaitable que les futurs travaux théoriques et empiriques prennent davantage en compte tous les paramètres de l'évolution comme les transferts de gènes, la sélection, la dérive génétique, les migrations, la taille des populations et la taille des pangénomes pour fonder leurs conclusions.

Ce chapitre clos l'état de l'art orienté vers les aspects biologiques des pangénomes. Nous l'avons vu à travers les trois précédents chapitres, il existe un grand nombre de points de vue théoriques, de modélisations, d'interprétations reliés au concept de pangénome. Cependant si l'on revient à une approche plus bioinformatique, ces modèles, méthodes et théories sont souvent testés sur quelques génomes dont on tire des conclusions pour l'ensemble des procaryotes. De plus, les méthodes sont souvent difficiles à réemployer sur de nouveaux génomes, de sorte que nous puissions regretter que peu de biologistes les utilisent effectivement dans leurs études de pangénome. Dès lors, la masse des études appliquées en pangénomique en reste aux concepts de base comme la simple dichotomie entre génome cœur et génome accessoire pour rendre compte de la diversité des espèces qu'ils étudient. Enfin, ce que les adeptes de la pangénomique semblent particulièrement apprécier en pratique, si l'on en juge le nombre de citations qu'ils obtiennent, ce sont les outils de constitution de pangénome. Passons donc en revue les différents outils bioinformatiques dédiés aux pangénomes qui existent dans la littérature.



Chapitre 5 : Les outils utilisés en pangénomique

L'essor notable de la pangénomique durant la décennie 2010 (voir figure 5.1) a été renforcé par l'apparition de très nombreuses méthodes pour détecter et indexer les différentes parties d'un pangénome.

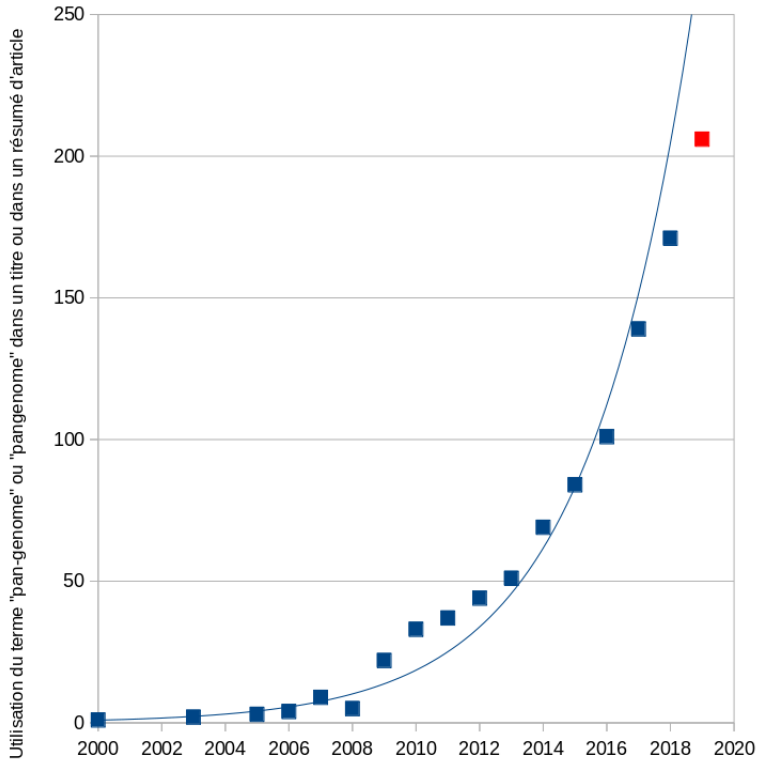


FIGURE 5.1 – Figure illustrant la croissance de l'usage du terme « pan-genome » et « pangenome » dans les titres et résumés référencés dans la base de données PubMed au 13 décembre 2019. Le point rouge représente l'année en cours (2019). On constate une croissance de l'intérêt pour cette thématique encore relativement exponentielle (l'ajustement avec une courbe de tendance exponentielle donne un $R^2 = 0.96$).

À travers ces développements, on peut constater une scission paradigmatique de la pangénomique en deux catégories d'outils selon l'approche choisie pour capturer la diversité d'un pangénome. On retrouve tout d'abord des outils dans la continuité des travaux décrits jusqu'ici, où le pangénome correspond à l'ensemble des familles de gènes homologues d'un clade. Néanmoins, plus récemment ont émergé en pangénomique des approches à l'échelle de la séquence nucléotidique plutôt qu'à l'échelle du gène. Dans ces cas de figure, les pangénomes sont principalement organisés sous la forme d'un graphe dont les nœuds décrivent des portions de séquence partagées par un certain nombre de génomes et où les arêtes relient ces nœuds en fonction des différentes colocalisations de ces éléments dans les génomes. Pour commencer ce chapitre, nous allons tous d'abord nous intéresser à la première catégorie d'approches se basant sur les familles de gènes, puis nous aborderons les développements effectués pour représenter les pangénomes à l'échelle de la séquence.

Cet inventaire tentera de lister les approches les plus utilisées, les plus originales ou les plus discutées mais n'a pas vocation à être parfaitement exhaustif ; des outils associés au concept de pangénome apparaissant presque toutes les semaines.

5.1 Approches basées sur les familles de gènes

Si l'on fait preuve d'abstraction, la méthode générale de la pangénomique basée sur les familles de gènes est semblable à celle utilisée depuis l'article de Welch et al., 2002. Ainsi, ces approches se basent sur un ensemble de génomes annotés et combinent systématiquement ces trois étapes successives :

1. Constitution d'un ensemble rassemblant toutes les familles de gènes homologues (parfois les orthologues seulement) d'une population de génomes (généralement une espèce) par alignement de séquences nucléotidiques ou protéiques.
2. Identification des différentes parties du pangénome (partitionnement), soit généralement, le génome cœur et le génome accessoire.
3. Analyses supplémentaires sur la base de ces parties comme la représentation du pangénome sous forme de diagrammes de Venn, le calcul de courbes de raréfaction, le recensement de variants SNP (Single Nucleotide Polymorphism) ou bien encore la construction d'arbres phylogénétiques.

Lorsque le concept de pangénome a émergé, les outils pour construire des familles de gènes homologues existaient déjà mais ceux-ci étaient conçus avec l'objectif de retrouver des gènes orthologues entre espèces. Pour les employer sur des génomes de la même espèce, les seuils et/ou les méthodes utilisées se doivent donc d'être ajustées. Parmi ces outils, on peut penser à OrthoMCL (L. Li et al., 2003), InParanoid (O'Brien et al., 2005 ; Remm et al., 2001), EggNOG (Jensen et al., 2008), RoundUp (Deluca et al., 2006). De plus, parmi les nombreuses ressources visant à répertorier les familles de gènes déjà constituées et les annoter fonctionnellement (sur la base d'algorithmes privés ou publics), on peut citer MBGD (Uchiyama, 2003), OMA (Altenhoff et al., 2018), Homologene (Wheeler et al., 2005), Ensembl compara (Herrero et al., 2016), CMR (Davidsen et al., 2010), FIGfam (Meyer et al., 2009), PIRSF (Wu et al., 2004), COG

(TATUSOV et al., 2003), CATH (KNUDSEN et al., 2010). En fonction des banques de données utilisées, l'ensemble de ces familles est obtenu par le truchement de stratégies très différentes comme l'utilisation des séquences protéiques, des séquences nucléiques, des structures protéiques ou enfin leurs fonctions métaboliques. Plusieurs comparatifs de ces différentes méthodes sont disponibles dont certains ont montré que InParanoid et OrthoMCL possèdent les meilleurs compromis entre la sensibilité et la spécificité (ALTENHOFF et al., 2009 ; CHEN et al., 2007 ; HULSEN et al., 2006). La liste suivante répertorie les principaux outils conçus spécifiquement pour la construction de familles de gènes lors d'analyses pangénomiques.

5.1.1 EDGAR

EDGAR (BLOM et al., 2009) est une application Web semblant être le premier outil dédié à la constitution de familles de gènes d'un pangénome. Elle utilise le célèbre outil BLAST (ALTSCHUL et al., 1990 ; 1997) dans sa version protéique (BLASTP) pour comparer toutes les séquences codantes de tous les génomes entre elles. Parmi ces résultats issus de BLASTP, deux gènes sont considérés comme orthologues entre eux, si dans les 2 sens, l'outil BLASTP rapporte l'autre gène comme la meilleure correspondance possible (BBH : Bidirectional Best Hit). Les auteurs de HULSEN et al., 2006 ont montré que l'existence d'un BBH est un critère suffisant pour établir un lien d'orthologie entre les génomes de 2 espèces.

Néanmoins contrairement à la recherche d'orthologues inter-espèces, entre des génomes d'une même espèce, les gènes sont généralement proches (ANI de 95% en moyenne). Dès lors en cas de duplication d'un gène, le paralogue naissant (c'est-à-dire avant sa néofonctionnalisation) risque d'être très proche du gène père. EDGAR va donc rechercher un seuil cohérent au sein du groupe de génomes pour valider qu'un BBH indique une association qualifiable de lien d'orthologie. Pour cela, EDGAR utilise le ratio des scores BLASTP (BLASTScore Ratio : BSR) introduit par RASKO et al., 2005 pour valider un BBH. Le BSR d'un gène a contre un gène b est calculé par la formule suivante $BSR(a,b) = \frac{S_{a,b}}{S_{a,a}}$ où $S_{a,b}$ est le bit score de l'alignement BLASTP entre le gène a requête et le gène b cible. Sur la base de la distribution de ces valeurs pour un choix de génomes donné, les auteurs obtiennent une distribution bimodale comme l'avait montré LERAT et al., 2003. Le premier pic correspond à de faibles valeurs de similarité, c'est-à-dire probablement des alignements non spécifiques ou entre paralogues, tandis que le second pic est supposé représenter des alignements entre gènes orthologues. À l'aide d'une méthode employant une fenêtre glissante, les auteurs estiment un seuil optimal pour séparer les 2 pics. Les alignements issus du premier pic sont ignorés alors que les BBHs du second sont utilisés pour identifier les orthologues. La base de données conserve tous les scores des résultats BLASTP, de sorte que l'utilisateur est libre de sélectionner le groupe de génomes qu'il souhaite pour obtenir l'ensemble des orthologues optimisés pour ce groupe. L'inconvénient est cependant que, comme tous les résultats sont pré-calculés, l'utilisateur est contraint d'utiliser les génomes déjà intégrés à la base de données. Une fois les orthologues identifiés, l'outil peut aussi identifier le génome cœur (orthologues présents dans tous les génomes) et le génome accessoire (orthologues présents dans quelques génomes ou singletons). Dans la première version, les visualisations se limitaient à la construction d'un arbre sur la base du génome cœur, à la navigation dans les génomes et à la construction d'un phyloprofil. Dans la version 2 de

l'outil (BLOM et al., 2016), de nouvelles figures sont disponibles, il est maintenant possible de produire des courbes de raréfaction selon les méthodes de TETTELIN et al., 2005 ; 2008. De plus des analyses phylogénétiques plus fines sont disponibles comme l'analyse de sous-arbres ou le calcul de l'identité moyenne en acides aminés (Average Amino-acid Identity : AAI).

5.1.2 Panseq

L'outil en ligne de commande Panseq (LAING et al., 2010) n'est pas à proprement parler un outil de construction de familles de gènes homologues mais se charge plutôt d'identifier des régions plus ou moins partagées entre des génomes avec l'objectif d'en découvrir de nouvelles. Ce dernier effectue des alignements nucléotidiques avec l'outil MUMmer (DELCHER et al., 2003) pour les comparaisons de génomes entiers ainsi qu'avec l'algorithme BLASTN (ALTSCHUL et al., 1990 ; 1997) pour les comparaisons locales de séquences. En exploitant ces alignements, l'outil détermine les régions cœurs conservées dans tous les génomes et la distribution des régions accessoires parmi l'ensemble des séquences analysées. L'outil, remarquablement rapide en comparaison de l'état de l'art de l'époque, fournit ainsi une matrice de présence/absence de régions à travers les génomes. Il semble tout de même limité à l'analyse du pangénome d'une petite centaine de génomes au maximum.

5.1.3 The Prokaryotic-genome Analysis Tool (PGAT)

Dans la continuité de EDGAR, PGAT (BRITTNACHER et al., 2011) est également une application Web. Pour construire sa base de données, l'application utilise BLASTP afin d'identifier les familles de gènes. Tout d'abord pour chaque famille de gènes, un gène est choisi comme référence. Les séquences de gènes de référence sont ensuite alignées par rapport à tous les cadres de lecture ouverts (Open Reading Frames : ORFs) de chaque génome. Les gènes ayant un pourcentage d'identité de séquence supérieur à 91 à 92 % sur plus de 80 % de la longueur sont considérés comme des orthologues. Après avoir sélectionné l'une des espèces disponibles sur le site, il est possible d'effectuer certaines analyses comme par exemple obtenir la liste des orthologues de tous les génomes du pangénome. Une option permet d'identifier les gènes présents ou absents dans un sous-ensemble choisi de génomes. Enfin la liste des SNPs obtenus par alignement multiple des orthologues via l'outil MUSCLE (EDGAR, 2004) est disponible. Enfin les voies métaboliques peuvent être comparées avec les voies de référence disponibles dans la base de données KEGG (KANEHISA et al., 2000).

Malheureusement, la base de données de PGAT ne comprend que neuf espèces et ne propose aucun moyen pour que des génomes ou des espèces soient ajoutés. Enfin, le site Web indique que ce projet n'a plus de financement de sorte que l'application peut désormais disparaître à tout moment.

5.1.4 The Pan-genome Analysis Pipeline (PGAP)

PGAP (ZHAO et al., 2012) est la première application permettant la construction de familles de gènes disponibles en ligne de commande. Le grand avantage d'utiliser un programme autonome est que l'utilisateur est libre de choisir les génomes qu'il souhaite intégrer dans ses analyses pangénomiques. Comme les applications précédemment évoquées, PGAP est ainsi capable de trouver les familles de gènes orthologues entre les génomes.

Pour ce faire, l'application utilise au choix 2 méthodes :

— MultiParanoid (MP) : tout d'abord, l'algorithme InParanoid (O'BRIEN et al., 2005 ; REMM et al., 2001) est utilisé pour détecter les orthologues entre chaque paire de génomes. InParanoid détecte les paires de protéines orthologues via une approche basée sur les BBHs. Les paralogues sont eux détectés en utilisant BLASTP pour rechercher des homologues entre les gènes d'un même génome. Ensuite, MultiParanoid (ALEXEYENKO et al., 2006) regroupe ces paires d'orthologues pour constituer les familles de gènes orthologues parmi plusieurs génomes. Toute paire de gènes prise au hasard dans une famille de gènes devra respecter les critères suivants : un alignement local ne doit pas représenter moins de 25 % de la séquence protéique la plus longue et de même pour l'alignement global qui ne doit pas en représenter moins de 50 % ; le score BLASTP doit être au moins de 50 et la valeur *E* de BLASTP inférieure à 10^{-8}.

— Gene Families (GF) : cette méthode est semblable à l'outil OrthoMCL (L. LI et al., 2003). En effet, les séquences protéiques de chaque génome sont alignées les unes contre les autres avec BLASTP. Les résultats de BLASTP sont réutilisés par l'algorithme MCL (Markov Cluster Algorithm : MCL, ENRIGHT et al., 2002) pour constituer les familles. Pour cela, l'algorithme traite un graphe où les nœuds représentent les gènes et les arêtes les similarités d'alignement (pas d'arête en cas d'absence d'alignement après filtrage). La matrice d'adjacence de ce graphe est soumise à des cycles itératifs de multiplication et d'inflation matricielle jusqu'à ce qu'il y ait peu ou pas de changement net dans la matrice. La matrice finale est alors interprétée comme un groupement de familles de protéines. Le paramètre correspondant à la valeur d'inflation de l'algorithme MCL est utilisé pour contrôler la granularité du regroupement.

Comparée à la méthode MP, la méthode GF est décrite comme considérablement plus rapide. Elle demande 20 minutes *versus* 191 minutes pour construire les familles de gènes de 11 génomes de *Streptococcus pyogenes*. Cependant, la méthode GF va davantage détecter les homologues, tandis que la méthode MP est plus spécifique de la détection des orthologues. Au final, la méthode GF semble générer un nombre un peu plus important de familles que la méthode MP [1], ce qui va accélérer la croissance du pangénome d'une part et la décroissance du génome cœur d'autre part dans les courbes de raréfaction.

À la suite de la génération des familles de gènes, l'application propose 4 modules d'analyse fonctionnelle. Ces modules sont très complets et permettent, en autre, de générer le profil du pangénome, les courbes de raréfaction du pangénome et du génome cœur et d'analyser les variations en contenu génique entre les génomes, ainsi que les SNPs des gènes du génome cœur. Sur la base

1. Voir pour illustrer cela les données supplémentaires de l'article ZHAO et al., 2012.

de ces deux types de variations, il est également possible de construire un arbre. Enfin, il propose de détecter les catégories fonctionnelles COG de chaque famille de gènes pour comparer leur différence d'enrichissement fonctionnel entre le pangénome, le génome accessoire et le génome cœur.

PAGP fut un grand succès en termes de diffusion ; il est cité à ce jour environ 215 fois (deuxième outil de pangénomique en termes de citations). Enfin, une interface graphique appelée PGAP-X dédiée à PGAP a par ailleurs été développée (ZHAO et al., 2018).

5.1.5 PanOCT

PanOCT (FOUTS et al., 2012) est un outil de construction de familles de gènes orthologues provenant d'un ensemble de génomes bactériens d'une même espèce (ou relativement proches). Comme de nombreux outils, il est basé sur une comparaison de toutes les CDSs (Coding DNA Sequence) contre toutes les autres. Celui-ci, néanmoins, n'effectue pas de comparaison directement mais réutilise le résultat d'une comparaison au format de BLASTP. PanOCT va ainsi regrouper les gènes en familles de gènes orthologues et exclure les paralogues, en combinant une méthode similaire à EDGAR (c'est-à-dire via l'utilisation du BSR) avec une autre information permettant d'améliorer la prédiction des orthologues. Cette information, c'est la conservation du voisinage des gènes (Conserved Gene Neighborhood : CGN). En effet, il est admis que pour deux gènes orthologues, l'ordre des gènes est conservé (on parle parfois d'orthologues positionnels) dans des espèces proches tandis que ce ne sera pas nécessairement le cas pour des gènes paralogues (HUYNEN et al., 1998). Néanmoins, les positions des gènes ne semblent plus aussi conservées entre des espèces éloignées (ROCHA, 2008 ; TAMAMES, 2001). Sachant cela, la méthode PanOCT utilise des formules *ad hoc* pour agréger les 2 scores de BLASTP avec le score CGN pour chaque paire de gènes comparée [1] afin d'identifier ces orthologues positionnels. Par défaut, le score CGN est calculé en tenant compte des scores d'homologie des 5 gènes en amont et en aval de la cible.

5.1.6 GET_HOMOLOGUES

L'outil GET_HOMOLOGUES (CONTRERAS-MOREIRA et al., 2013) construit les familles de gènes homologues en employant trois méthodes au choix :

— Une approche semblable à l'approche EDGAR, c'est-à-dire basée sur les BBHs comme précédemment décrit à la sous-section 5.1.1 page 81.

— La méthode COGtriangles (KRISTENSEN et al., 2010). Cette méthode se base sur un graphe où les nœuds correspondent à la totalité des gènes de tous les génomes (y compris les gènes en plusieurs copies) et où les arêtes indiquent les BBHs. Dans ce graphe, un triangle correspond à trois gènes reliés par trois arêtes. Les triangles de ce graphe sont ensuite regroupés en COGs s'ils possèdent des arêtes en commun. Pour chaque triangle T_a provenant de la liste T des triangles non regroupés dans un groupe d'orthologue (Cluster of Orthologuous Genes : COG), il faut initialiser un COG graine C en lui ajoutant T_a. Puis pour chaque triangle T_b provenant de T, ajouter T_b à C_1 si T_b a une arête en commun avec les triangles de C_1. Et ainsi de suite jusqu'à vider T et ainsi obtenir un ensemble C de COGs.

1. Pour plus de détails, voir figure 3 de FOUTS et al., 2012

— La méthode OrthoMCL (L. LI et al., 2003) précédemment décrite pour l'outil PGAP à la sous-section 5.1.4 page 83.

En plus d'être utilisées isolément, ces trois méthodes peuvent être employées de concert car GET_HOMOLOGUES est capable de proposer un consensus sur la base de l'ensemble des résultats obtenus.

5.1.7 Integrated Toolkit for Exploration of Microbial Pan-genomes (ITEP)

La méthode ITEP (BENEDICT et al., 2014) correspond à un ensemble de programmes couplé à une base de données ainsi que de nombreuses visualisations. Cette base stocke le résultat d'une comparaison de tous les gènes d'un pangénome entre eux, à l'aide de BLASTP ainsi que BLASTN. Les gènes sont ensuite regroupés via l'algorithme MCL de façon similaire à ce que nous avons vu pour la méthode PGAP à la sous-section 5.1.4 page 83. Les résultats d'autres outils de prédiction de gènes orthologues peuvent également être importés dans ITEP. Une fois la base de données construite, les génomes cœur et accessoire sont extraits et différentes analyses peuvent être effectuées à travers plusieurs visualisations. On trouve notamment des modules pour parcourir les résultats d'alignements multiples, pour construire des arbres phylogénétiques ou même la possibilité de visualiser le voisinage des gènes. L'outil offre enfin la possibilité de rechercher la présence ou l'absence de certains gènes sous l'angle de la phylogénie et de générer des réseaux métaboliques.

5.1.8 Roary

Roary (PAGE et al., 2015) est de loin l'outil de pangénomique le plus cité [1]. La raison de son succès tient à sa simplicité mais surtout à sa rapidité. Roary est en effet capable de traiter des pangénomes basés sur un millier de génomes en 4h30 là où la même analyse avec PanOCT ou PGAP prendrait des semaines. Bien que Roary effectue tous les alignements entre séquences deux à deux, le secret de sa rapidité réside dans l'utilisation de CD-HIT afin d'effectuer la majorité des alignements entre séquences très proches (la méthode est abordée ci-dessous) suivi de BLASTP pour le reste des alignements. Les familles sont ensuite générées par l'utilisation de la méthode MCL (ENRIGHT et al., 2002) décrite à la sous-section 5.1.4 page 83.

Nous reviendrons sur la méthode Roary au chapitre 6 page 99, puisque que celle-ci propose d'effectuer un partitionnement trichotomique du pangénome en utilisant des seuils fixes.

CD-HIT

CD-HIT (FU et al., 2012) est un outil de regroupement de séquences proches (nucléiques ou protéiques) très performant pour réduire la redondance d'un ensemble de séquences. CD-HIT est un algorithme glouton par incrément (« greedy incremental ») qui utilise tout d'abord la séquence la plus longue en tant que premier représentant, puis recherche les séquences similaires à celle-ci. Les séquences similaires sont associées à cette séquence la plus longue et forment une famille. Le même processus est répété jusqu'à épuisement des séquences restantes. Les similitudes entre les

1. Soit 826 citations au 5 novembre 2019 d'après Google Scholar.

séquences sont estimées par un comptage de mots communs. Pour qu'une séquence soit considérée comme similaire à la séquence la plus longue, elle doit avoir un nombre de mots en commun supérieur à un seuil.

5.1.9 Piggy

Le but de Piggy (THORPE et al., 2018) n'est pas de faire le pangénome d'une espèce mais son PanIGROme. Le PanIGROme correspond à l'équivalent du pangénome pour les régions intergéniques (InterGenique Region : IGR). Ces régions sont souvent négligées en pangénomique alors qu'elles contiennent pourtant d'importantes régions comme les promoteurs, les opérateurs, les régions CRISPR et même de possibles gènes en naissance : les protogènes (LEVY, 2019). D'un point de vue technique, Piggy est une façade de Roary effectuant non pas le regroupant des gènes mais de tout ce qui reste dans les génomes après soustraction des gènes.

5.1.10 The Bacterial Pan Genome Analysis (BPGA)

La méthode BPGA (CHAUDHARI et al., 2016) est l'une des plus rapides actuellement pour construire un pangénome. Pour construire les familles de gènes, l'outil propose, au choix, trois outils de regroupement de séquences (USEARCH, CD-HIT et OrthoMCL). Les auteurs montrent que les trois outils génèrent des résultats presque similaires, mais USEARCH (dont la méthode est résumée ci-dessous) est employée par défaut dans l'outil BPGA semble la plus rapide des trois. En se basant sur USEARCH, BPGA se place avec Roary comme l'un des outils les plus rapides pour la génération de familles de gènes à ce jour.

Sur la base des familles de gènes homologues détectées, l'outil génère la matrice de présence/absence, le profil des pangénomes ainsi que la phylogénie des génomes. Les familles de gènes sont par ailleurs annotées fonctionnellement (TATUSOV et al., 2003) par des identifiants COG et KEGG (KANEHISA et al., 2000).

USEARCH

USEARCH est un outil décrit par ses auteurs (EDGAR, 2010) comme 350 fois plus rapide que BLASTP. Pour cela, l'algorithme emploie des heuristiques. Pour une séquence requête donnée, les séquences cibles sont triées par ordre décroissant du nombre de k-mers en commun pour exploiter le fait que des séquences similaires ont tendance à avoir des mots courts en commun. Des alignements classiques sont ensuite effectués dans l'ordre jusqu'à ce que plusieurs alignements de suite échouent à remplir les critères d'identité. À ce moment-ci, le parcours des séquences cibles est avorté. Ainsi, contrairement à des méthodes comme BLAST, de très nombreuses comparaisons sont alors évitées, ce qui explique la célérité de l'outil.

5.1.11 Large Scale Blast Score Ratio (LS-BSR)

La méthode LS-BSR (SAHL et al., 2014) est une méthode de détection de gènes orthologues dont le fonctionnement est basé sur les BLASTScore Ratio de façon similaire à ce qui a été décrit pour la méthode EDGAR (voir sous-section 5.1.1 page 81). Cependant, des approches comme la méthode EDGAR nécessitent des alignements BLASTP entre toutes les séquences ce qui est

limitant pour construire le pangénome de nombreux génomes. Pour pallier à cela, les auteurs de SAHL et al., 2014 utilisent USEARCH ainsi que nous l'avons vu avec la méthode BPGA (à la sous-section 5.1.10 page précédente) afin de pré-filtrer les résultats et accélérer l'algorithme.

5.1.12 micropan

L'outil micropan (SNIPEN et al., 2015) a la particularité d'être un outil très complet en ce sens qu'il se base sur des génomes assemblés, les annote, génère les familles de gènes, puis partitionne le pangénome selon la méthode de SNIPEN et al., 2009. L'algorithme de partitionnement du pangénome utilisé dans la méthode micropan a été décrit à la sous-section 3.3.2 page 54 et nous le ré-aborderons à nouveau à la sous-section 6.1.4 page 105. Nous nous intéresserons donc là, seulement à la méthode utilisée pour constituer les familles de gènes.

Pour identifier les familles de gènes homologues, micropan utilise deux stratégies au choix. D'une part, une approche relativement classique et peu optimisée où toutes les séquences protéiques sont alignées les unes contre les autres avec BLASTP afin de générer une matrice de distance. Les familles de gènes sont ensuite identifiées par un regroupement hiérarchique.

Conscients des limites de leur première approche, les auteurs proposent une approche originale afin d'accélérer le regroupement de séquences en familles en se basant sur l'outil HMMER (MISTRY et al., 2013) et décrite plus en détail dans cette étude de SNIPEN et al., 2012. Pour rappel, l'outil HMMER permet de générer une représentation consensus d'une famille de gènes homologues sous la forme d'un modèle de Markov caché. En alignant des séquences sur une base de données constituée d'un ensemble de modèles HHMER de domaines protéiques, il devient possible de détecter avec une grande sensibilité les domaines présents sur chaque séquence protéique. Ainsi, micropan obtient pour chaque protéine la séquence ordonnée de domaines que la protéine contient. Ils utilisent alors cette information pour regrouper les protéines ayant les mêmes successions de domaines en familles de gènes homologues.

Il est à noter que l'outil de construction de pangénome PanFunPro (LUKJANCENKO et al., 2013) utilise en combinaison avec l'outil CD-HIT, une méthode de génération de familles (FU et al., 2012) s'approchant de celle présentée ici par l'utilisation de l'ordre des domaines comme empreinte protéique.

5.2 Approches à l'échelle de la séquence

Comme nous l'avons vu, un pangénome désigne historiquement un ensemble de familles de gènes homologues. Cependant, le mot pangénome englobe désormais un nouveau sens toujours compatible avec la définition 2.0.1 page 33, mais en pratique relativement différent. Celui-ci peut en effet désigner des méthodes algorithmiques de représentation et d'indexation de l'ensemble de la variabilité nucléotidique de N génomes. Ces variations peuvent être des polymorphismes d'un seul nucléotide (Single Nucleotide Polymorphism : SNP) ou des variations plus conséquentes comme des variants structuraux. Ces différents sens du concept de pangénome sont notamment passés en revue dans l'article de THE COMPUTATIONAL PAN-GENOMICS CONSORTIUM, 2016.

5.3 Les approches d'indexation de génomes basées sur une référence

Les méthodes présentées ci-dessous sont conçues afin d'être en mesure d'effectuer des recherches sur des centaines de génomes, même massifs, avec un simple ordinateur portable. Ce sont donc des méthodes particulièrement optimisées. Un rigoriste pourrait toutefois contester que la logique des approches présentées dans cette section relève de la pangénomique. En effet, celles-ci restent centrées sur une référence. Pourtant, nous évoquons ces méthodes car elles ont vocation à stocker toutes les variations possibles par rapport à une référence de sorte qu'elles peuvent entrer dans le champ de la pangénomique.

5.3.1 Referentially Compressed Search Index (RCSI)

La méthode RCSI (WANDELT et al., 2014) indexe une référence pour stocker un pangénome. Pour cela, la référence est indexée dans un arbre de suffixes compressés. Sur un arbre des suffixes, tous les chemins de la racine aux feuilles de l'arbre listent l'ensemble des suffixes possibles d'un texte. On rappelle qu'un suffixe correspond à tous les caractères situés entre une position donnée et la fin du texte.

Pour ce qui est des autres génomes, ceux-ci sont exprimés sous la forme de listes ordonnées de sous-séquences en correspondance avec la référence. Ces sous-séquences sont appelées les entrées de correspondance référentielle.

L'index est cependant construit d'une manière qui présente des limites aux requêtes susceptibles d'être effectuées sur cette référence. En effet, les séquences requêtes ne peuvent excéder une certaine taille et ne peuvent détecter de correspondances au-delà d'une certaine distance d'édition M.

5.3.2 Multiple Genome Index (MuGI)

L'approche précédente, RSCI, est intéressante pour des génomes très stables mais risque de présenter des limites dans le cas où les parties variables sont largement dominantes par rapport à la référence, de sorte que la méthode semble moins indiquée pour ce type d'emploi (donc notamment à l'indexation de génomes de procaryotes).

Pour répondre à cela, la méthode MuGi (DANEK et al., 2014) utilise une approche très proche de la méthode RSCI à la différence qu'au lieu d'indexer la référence, elle indexe plutôt les variants par rapport à la référence. Même si les auteurs proposent l'algorithme pour le référencement du génome humain, son algorithme semble plus adapté pour le stockage de génomes plus variables.

5.3.3 The Journaled String Tree (JST)

La méthode JST (RAHN et al., 2014) est une autre structure de données dédiée à l'indexation de génomes similaires en se basant sur le stockage des variants du génome par rapport à une référence. La structure est composée d'une référence et d'un tableau des variants de tous les génomes se branchant sur la référence. Chaque génome est stocké sous la forme d'une chaîne journalisée (« journaled string ») qui se compose de deux parties : un tampon d'insertion et un arbre de recherche binaire. Le tampon d'insertion est une chaîne de caractères issue de la concaténation de tous les variants du génome. L'arbre de recherche binaire stocke pour sa part, les positions et longueurs des variants dans le tampon d'insertion par rapport à la référence.

La méthode ne propose pas à proprement parler un algorithme de recherche dédié mais permet à plusieurs algorithmes de venir utiliser cette structure de données. Les auteurs ont notamment utilisé leur structure couplée à ces deux algorithmes (HORSPOOL, 1980 ; MYERS, 1999). Ils montrent que, de cette manière, ils parviennent à stocker et rechercher les variants de 1000 chromosomes I du génome humain.

5.3.4 BWBBLE

La méthode (HUANG et al., 2013) propose elle aussi une structure de données incluant tous les variants détectés entre un ensemble de génomes et une référence. L'outil permet notamment l'alignement de lectures sur cette structure. Pour ce faire, une référence linéaire augmentée est construite, puis compressée à l'aide de l'algorithme de transformation de Burrows-Wheeler Transform (BWT) décrit ci-dessous (BURROWS et al., 1994) et enfin indexée à l'aide d'un index de Ferragina-Manzini décrit ci-dessous également (FM-index, FERRAGINA et al., 2000). La construction de la référence linéaire augmentée consiste à encoder la position des variants en remplaçant des nucléotides de la référence par un encodage IUPAC incluant les variants possibles à une position donnée. Pour rappel, l'encodage IUPAC permet grâce à un alphabet de taille 15 d'encoder des combinaisons de nucléotides alternatifs.

Le gros inconvénient de cette approche est qu'elle permet uniquement de stocker les variants d'un seul nucléotide (SNPs). Dès lors, les autres types de variants comme les variations en nombre de copies (Copy Number Variation : CNV) ou les variants structuraux ne pouvent pas être répertoriés ainsi.

Transformation de Burrows-Wheeler

La transformation de Burrows-Wheeler (BWT) consiste à effectuer des permutations successives des éléments d'une séquence. Ces éléments doivent appartenir à un alphabet admettant une relation d'ordre (exemple $\$ > A > C > G > T$ dans l'ordre lexicographique pour l'alphabet nucléotidique, avec « $ » l'indicateur de la fin d'une séquence). Le but de cette transformation est d'augmenter le nombre de cas où des éléments identiques se retrouveront répétés en tandem sur la séquence ainsi transformée (Burrows-Wheeler Transform : BWT). L'avantage d'augmenter le nombre d'éléments en tandem dans une séquence est qu'il devient ainsi possible de ne plus stocker tous les éléments répétés en tandem mais simplement le nombre de répétitions de l'élément.

Pour effectuer cette transformation, il suffit d'effectuer les l rotations possibles d'une séquence de taille l, puis de trier ces l rotations. La transformée de Burrows-Wheeler correspond ainsi à la dernière colonne de chaque rotation.

Prenons comme exemple la séquence $T = ATGGCTCTAGGACTATAA\$$ où le $ indique la fin de la séquence, celle-ci sera transformée en $L = AATGTT\$TAGGGATACCCA$ (voir figure 5.2 page suivante). Par la suite la séquence $AATGTT\$TAGGGATACCCA$ peut être réduite en $2ATG2T\$TA3GATA3CA$.

Ainsi, comme cette transformation de Burrows-Wheeler est réversible, il devient possible de compresser des séquences via cette méthode.

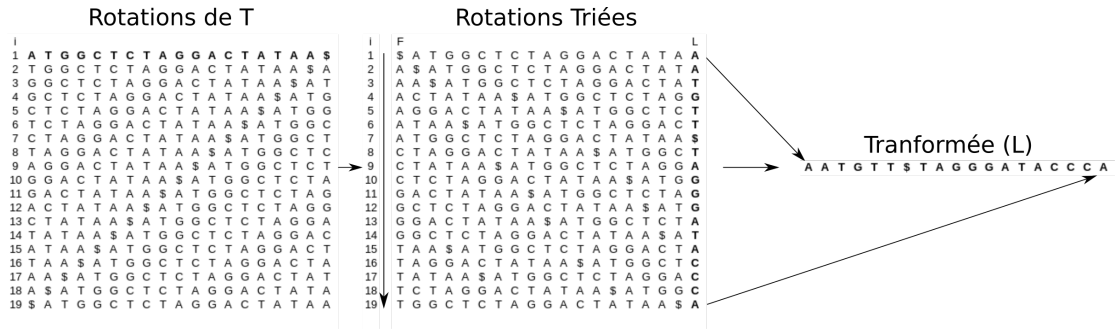


FIGURE 5.2 – Figure illustrant la transformation de Burrows-Wheeler d'une séquence $T = ATGGCTCTAGGACTATAA\$$

Index de Ferragina-Manzini

L'index de Ferragina-Manzini (FM-index, FERRAGINA et al., 2000) est capable de faire très rapidement le lien (en temps constant) entre la séquence originelle T avant transformation de Burrows-Wheeler et sa transformée L.

Pour cela, l'index se base sur un vecteur C et deux fonctions :

— Le vecteur C compte pour chaque élément de l'alphabet, le nombre d'occurrences des éléments inférieurs dans T (ou dans L).

— La fonction $\text{rank}(e,i)$ donne le nombre d'occurrences des éléments e dans les différents préfixes $L[1 \cdot I]$. Les auteurs de FERRAGINA et al., 2000 montrent que cette fonction donne un résultat en temps constant.

— La fonction $LF(i) = C[L[i]] + rank(L[i],i)$.

C

e	$	A	C	G	T
C[e]	0	1	7	10	14

Matrice résultat de la fonction Rank(e,i) de L[1,i]

L	A	A	T	G	T	T	$	T	A	G	G	G	A	T	A	C	C	C	A
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
$	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
A	1	2	2	2	2	2	2	2	3	3	3	3	4	4	5	5	5	5	6
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	3
G	0	0	0	1	1	1	1	1	1	2	3	4	4	4	4	4	4	4	4
T	0	0	1	1	2	3	3	4	4	4	4	4	4	5	5	5	5	5	5

FIGURE 5.3 – Figure illustrant l'index de Ferragina-Manzini de la séquence T utilisée dans la figure 5.2

Pour la cas de la séquence T introduite dans l'exemple précédent, le vecteur C et la matrice résultat de la fonction $\text{rank}(e,i)$ sont indiqués dans la figure 5.3.

Avec ces informations, il est possible de retrouver la séquence T en parcourant la séquence L et en employant la fonction $LF(i)$. Sachant que T est une séquence de longueur $l = 19$, voici les premières étapes.

1. Pour obtenir le dernier élément de T, prenons le premier élément de $L[1] = A$. Cela nous indique que $T[(l-1)\dots l] = A\$$.
2. Via la fonction $LF(1) = C[L[1]] + \text{rank}(L[1],1) = 1+1 = 2$, on sait donc que $T[l-1] = L[2] = A$, donc que $T[(l-2)\dots l] = AA\$$.

3. Via la fonction $LF(2) = C[L[2]] + \text{rank}(L[2],2) = 1+2 = 3$, on sait donc que $T[l-2] = L[3] = T$, donc que $T[(l-3)\dots l] = TAA\$$.
4. Via la fonction $LF(3) = C[L[3]] + \text{rank}(L[3],3) = 14+1 = 15$, on sait donc que $T[l-3] = L[3] = A$, donc que $T[(l-4)\dots l] = ATAA\$$.
5. Via la fonction $LF(15) = C[L[15]] + \text{rank}(L[15],15) = 1+5 = 6$, on sait donc que $T[l-4] = L[6] = T$, donc que $T[(l-5)\dots l] = TATAA\$$.
6. Via la fonction $LF(6) = C[L[6]] + \text{rank}(L[6],6) = 14+3 = 17$, on sait donc que $T[l-6] = L[17] = C$, donc que $T[(l-6)\dots l] = CTATAA\$$.
7. En continuant nous pouvons ainsi retrouver la séquence entière de T grâce au FM-index.

De la même façon qu'il permet de retrouver T, cet index permet surtout la reconnaissance et le comptage de motifs dans L.

5.4 Les graphes de séquence

5.4.1 GenomeMapper

On peut remonter l'origine du concept de graphe de variants (Variation Graph : VG) à l'outil GenomeMapper (SCHNEEBERGER et al., 2009). Cette structure de données stocke l'ensemble des variants de plusieurs génomes sous la forme d'un graphe. Pour cela, la méthode découpe chaque génome en blocs de taille fixe. Les blocs de l'ensemble des génomes correspondent aux nœuds du graphe et ceux-ci sont reliés selon le voisinage des blocs à travers les génomes. Par conséquent les blocs partagés par plusieurs génomes ne sont stockés qu'une seule fois. Comme désiré, on obtient ainsi un graphe.

Enfin, pour accélérer les requêtes sur ce graphe, l'ensemble des k-mers contenus dans le graphe sont indexés dans un dictionnaire attribuant, à chaque k-mer, l'identifiant des blocs qui le contiennent.

5.4.2 Variation Graph (VG)

Quand on parle de graphe de variants, l'outil VG (GARRISON et al., 2018) est la solution logicielle la plus complète à ce jour pour manier cette structure de données. À l'origine développé pour le génome humain, VG comporte de nombreuses commandes permettant de créer, de manipuler et d'utiliser un graphe de variants.

Dans VG, un graphe de variants $G = (N,E,P)$ comprend un ensemble de nœuds $N = (n_1 \dots n_M)$ représentant une séquence de nucléotides n_i, un ensemble d'arcs $E = (e_1 \dots e_L)$ connectant les nœuds n_i correspondant à des séquences connexes dans au moins un génome et un ensemble de chemins $P = (p_1, \dots, p_Q)$ décrivant les chemins propres à chaque génome dans le graphe. Comme l'outil est très complet, on peut, sur la base de ce graphe, ajouter des séquences, en retirer, annoter des portions du graphe et aligner des lectures sur ce graphe.

Pour aligner des séquences sur ce graphe, VG se base sur la bibliothèque « Generalized Compressed Suffix Array 2 » (GCSA2) (SIRÉN, 2017). Cette bibliothèque transforme un alignement multiple de génomes en un automate fini sur lequel il est possible d'aligner des lectures.

En plus de la possibilité d'effectuer des alignements sur ce graphe, il est également possible de visualiser le graphe de variants sous la forme d'une carte de métro grâce à un outil *ad hoc* relié à VG appelé TubeMap (BEYER et al., 2019).

5.4.3 HISAT2

Comme VG, la méthode HISAT2 (KIM et al., 2019) est un outil d'alignement sur un graphe de variants. Cette méthode a été conçue pour étudier les millions de variants du génome humain. L'outil se base sur une BWT adaptée pour un graphe de variants similaire à VG. De plus, la transformée est également indexée par un FM-index lui aussi adapté pour un graphe (Graph FM-index, GFM-index). D'après les auteurs, ce type d'approche basée sur un GFM-index est unique à ce jour. En plus d'utiliser un GFM-index global qui représente tout le graphe, HISAT2 utilise un grand nombre de petits GFM-index qui couvrent collectivement l'ensemble du génome. Pour donner un ordre d'idée, chaque GFM-index représente une région génomique de 56Kb dans l'utilisation présentée par les auteurs sur des génomes humains. Ces petits index (appelés index locaux) combinés à plusieurs stratégies d'alignement permettent un alignement efficace des lectures. Ce nouveau schéma d'indexation est appelé HGFM (Hierarchical Graph FM index). Grâce à cette structure innovante, les auteurs montrent que leur outil est plus rapide que VG pour effectuer un alignement de lectures sur le graphe.

5.4.4 PanCake

L'outil PanCake (ERNST et al., 2013) construit un graphe de séquences à partir d'un ensemble d'alignements de génomes deux à deux. Chaque nœud contient une sous-séquence de référence locale associée à l'ensemble des sous-séquences s'alignant sur cette référence locale. Pour chacune des sous-séquences, une version compressée est stockée de sorte qu'elle puisse être régénérée à partir de la séquence de référence locale. Les occurrences caractéristiques d'un génome sont ordonnées via une liste doublement chaînée de sorte que le génome entier puisse être reconstruit en itérant sur la liste et en concaténant les sous-séquences ainsi reconstruites. Les nouveaux génomes à insérer dans la structure de données sont alignés par paire avec les génomes déjà insérés dans le graphe.

La principale limite de la méthode PanCake est qu'elle nécessite, comme nous l'avons vu, des alignements de tous les génomes deux à deux, ce qui limite son utilisation à quelques dizaines de génomes au maximum.

5.5 Les graphes de De Bruijn

Sachant une liste de tous les mots de taille k (appelés k-mers) présents dans un ensemble de séquences, un graphe de De Bruijn (De Bruijn Graph : DBG, DE BRUIJN, 1946) est un graphe orienté où les nœuds représentent ces mots et où les arcs indiquent la succession chevauchante des mots dans les séquences.

À travers ce type de graphes, très communs en bioinformatique, une immensité d'outils a été développée. La plupart sont dédiés à l'assemblage de lectures afin d'obtenir la ou les séquences contiguës (« contig ») d'un génome. Cependant, il existe des méthodes utilisant cette structure aux fins de représentation et de stockage des pangénomes. Sur ce type de graphe de pangénome, les différents génomes ayant participé à constituer ce graphe sont utilisés afin d'en colorer les nœuds et les arcs.

5.5.1 Cortex

L'une des premières méthodes employant un graphe de De Bruijn afin de représenter les régions variables du génome humain est la méthode Cortex (IQBAL et al., 2012). En 2012, il était difficile de détecter les variants génétiques très différents de la séquence de référence (au-delà des variants SNP). En employant un graphe de De Bruijn coloré, la méthode Cortex est en mesure de détecter et génotyper l'ensemble de ces variants. Pour cela, la méthode va co-assembler plusieurs génomes humains simultanément. Elle permet ensuite d'utiliser ce graphe dans le but de retrouver les chemins communs à différents génomes mais surtout les chemins spécifiques formant des bulles dans le graphe (voir figure 5.4).

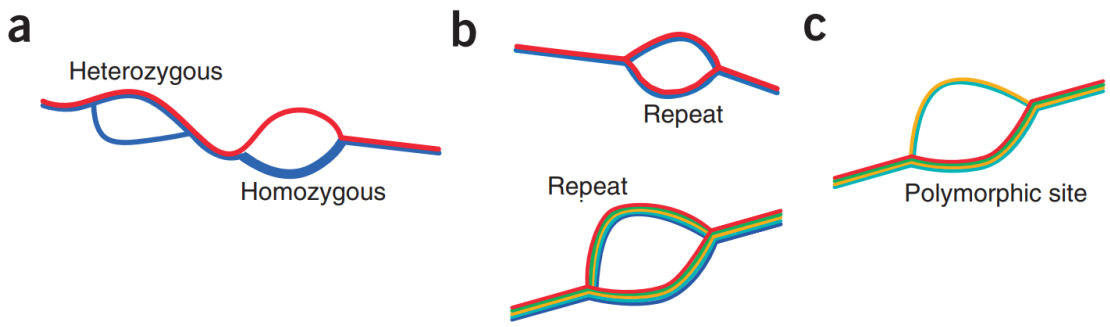


FIGURE 5.4 – Figure reproduite depuis IQBAL et al., 2012 illustrant des chemins alternatifs visualisables avec un pangénome graphe coloré par les génomes. (a) Découverte de variants chez un individu diploïde (en bleu) assemblé avec une séquence de référence (en rouge). (b) Les vrais polymorphismes génèrent des bulles qui divergent de la référence, alors que les structures répétées conduisent à des bulles qui sont également observées dans la référence. (c) Lorsque plusieurs échantillons (de couleur différente) sont combinés, il est possible de distinguer les bulles induites par les répétitions des véritables sites variants.

Grâce à cette approche, les auteurs indiquent avoir notamment découvert, en analysant 1000 génomes, 3Mb de séquences supplémentaires du génome humain qui étaient absentes de la référence.

5.5.2 SplitMEM

La logique des auteurs de SplitMEM (MARCUS et al., 2014) est d'optimiser l'utilisation de ce type de graphe afin d'extraire les régions similaires et spécifiques d'un ensemble de génomes. Notons déjà que les auteurs utilisent une version compressée du graphe de Bruijn.

Pour extraire ces séquences, les auteurs de SplitMEM commencent par mettre en évidence des relations topologiques entre les arbres de suffixe et les graphes de De Bruijn compressés (cDBG), ce qui leur permet d'introduire le concept de saut de suffixes (« suffix skips »). Un cDBG correspond à un DBG dont les successions de nœuds sans ramification sont compressées en un seul nœud (appelé « unitig »). Grâce à la méthode du saut de suffixes, les auteurs parviennent de façon très optimisée à construire le cDBG en se basant sur un arbre de suffixes sans avoir à construire un DBG au préalable. La partie compressée du graphe représente ainsi les correspondances exactes maximales (Maximum Exact Matches : MEMs) entre tous les génomes. Il suffit donc d'extraire ces parties compressées. Néanmoins, la méthode est aussi capable d'extraire des MEMs présents dans seulement un sous-ensemble des génomes. Un MEM peut ainsi correspondre à plusieurs nœuds compressés que l'algorithme est également capable de retrouver.

Les auteurs montrent que leur algorithme est linéaire en temps et en espace en l'employant afin d'extraire le génome cœur et le génome accessoire de 62 génomes d'*E. coli*.

Il faut noter que les MEMs comme leur nom l'indique sont des motifs exacts. Ainsi il y a un risque de fragmentation du nombre de MEMs quand de nombreux génomes sont ajoutés au cDBG.

5.5.3 Bloom Filter Trie (BFT)

Le Bloom Filter Trie (BFT, HOLLEY et al., 2016) est une structure de données compacte, incrémentale, sans référence et dont on peut interroger le contenu sans alignement. Cette structure permet l'indexation et l'interrogation à grande échelle d'un pangénome représenté sous la forme d'un DBG coloré construit à partir de génomes ou de lectures. Le BFT est basé sur un « burst trie » pour indexer les k-mers ainsi que sur un filtre de Bloom. Nous allons donc d'abord décrire ces structures avant de revenir au BFT.

Qu'est-ce qu'un « burst trie » ?

Un « trie » est un arbre stockant un ensemble de séquences et dont les arêtes sont annotées par un caractère de sorte qu'aucune arête commençant au même nœud ne peut avoir le même caractère. Chaque chemin de la racine à la feuille représente la chaîne de caractères obtenue en concaténant tous les caractères sur ce chemin. Le « burst trie » est une optimisation d'un trie qui réduit son nombre d'arêtes par la compression des sous-arbres du trie en feuilles. Ainsi, ses nœuds internes sont étiquetés avec de multiples préfixes de longueur 1, liés aux nœuds enfants. Les feuilles sont étiquetées avec de multiples suffixes de longueur arbitraire. Une feuille a une capacité limitée de suffixes et est éclatée lorsque cette capacité est dépassée. Pour cela le burst trie divise les suffixes d'une feuille en préfixes de longueur 1, liés à de nouvelles feuilles représentant les suffixes restants.

Qu'est-ce qu'un « filtre de Bloom » ?

Un filtre de Bloom (BLOOM, 1970) est une structure de données permettant de stocker un ensemble. Sachant cet ensemble, avec l'aide du filtre du Bloom, il est possible de vérifier si un élément est présent dans cet ensemble avec une complexité constante. Pour cela, les données à insérer dans l'ensemble sont minimisées avec plusieurs fonctions f_i de hachage de sorte que les résultats du hachage servent d'index pour un vecteur binaire. Lors de l'insertion d'un élément dans l'ensemble, les cases ainsi indexées sont positionnées à 1.

Pour savoir si un élément est présent dans le filtre de Bloom, il suffit de le hacher avec les f_i fonctions et d'aller observer si toutes les positions retournées par les différentes fonctions de hachage sont égales à 1 ou non. S'il existe au moins une valeur à 0 à ces positions, l'élément est absent du filtre avec certitude. À l'inverse si toutes les valeurs sont à 1, l'élément est très probablement présent, même si le filtre de Bloom possède un risque de collisions. Le risque dépendra de la taille du vecteur binaire ainsi que du nombre de fonctions de hachage utilisées.

L'avantage du filtre de Bloom est qu'il garantit un usage mémoire stable et un taux d'accès constant. Une table de hachage par exemple aura une taille variable et même si son temps d'insertion est constant en temps normal, en cas de collision celui-ci est supérieur.

Bloom Filter Trie

Le burst trie est construit de telle sorte qu'un chemin de la racine à la feuille représente un k-mer associé à un ensemble de couleurs du DBG. Les nœuds du BFT stockent des listes de conteneurs. Les conteneurs non compressés indexent les suffixes des k-mers (associés à leurs couleurs) et ont une capacité limitée. Aussi, lorsque la capacité en suffixes d'un tel conteneur est dépassée au moment de l'insertion, le conteneur est éclaté (« burst ») ; les préfixes communs aux suffixes des k-mers sont alors indexés dans un nouveau conteneur compressé remplaçant ainsi le conteneur non compressé.

Ici, les filtres Bloom sont employés afin de permettre une navigation efficace entre les conteneurs. De plus, ils optimisent la traversée du DBG coloré en regroupant les k-mers voisins.

Enfin, d'après les auteurs de HOLLEY et al., 2016, le BFT serait la seule structure capable de recevoir de nouveaux génomes sans nécessité de reconstruire l'index.

5.5.4 TwoPaCo

La méthode TwoPaCo (MINKIN et al., 2017) est une méthode massivement parallélisable qui est capable elle aussi de construire directement le cDBG de nombreux génomes. Les auteurs appellent une jonction, un nœud v ayant soit plus d'un arc convergent (arc arrivant au nœud), soit plus d'un arc divergent (arc partant du nœud), soit étant le premier k-mer d'une séquence, soit étant le dernier k-mer d'une séquence. L'outil est basé sur l'observation (dont les auteurs apportent la preuve) qu'il existe une bijection entre les chemins non ramifiés maximaux du graphe de Bruijn et les sous-chaînes dont les jonctions sont exactement les deux k-mers aux extrémités des sous-chaînes. Cette observation réduit le problème de la recherche de chemins maximaux non ramifiés dans le DBG à la recherche, dans les séquences d'entrée, de la position des jonctions, c'est-à-dire des k-mers qui sont ramifiés ou bien des k-mers qui commencent ou finissent une séquence. Pour ce faire, l'algorithme va traiter l'ensemble de positions de jonctions possibles dans les séquences d'entrée. Pour chaque position candidate i, les deux $(k+1)$-mers commençant aux positions i et $i-1$ sont insérés dans un filtre de Bloom D. Ensuite, chacune de ces positions i est à nouveau traitée en interrogeant D pour tous les successeurs et prédécesseurs possibles du k-mer commençant à la position i. Si le k-mer présente à la fois un degré convergent de 1 et un degré divergent de 1, ce n'est pas une jonction. Ces positions sont ensuite utilisées comme entrées pour la deuxième passe

où cette fois D est remplacé par une table de hachage pour supprimer les jonctions fausses positives générées par le filtre de Bloom. Ainsi les auteurs parviennent à identifier les sous-séquences entre les jonctions afin d'obtenir le cDBG.

Sur la base de cette méthode, les auteurs montrent qu'ils sont capables de construire le cDBG de 100 génomes humains en moins d'une journée avec 15 cœurs et seulement 77GB de mémoire vive grâce à l'utilisation des filtres de Bloom.

5.5.5 DBGWAS

La méthode DBGWAS (JAILLARD et al., 2018) utilise un cDBG pour identifier les k-mers correspondant à des variants génétiques susceptibles d'être associées à des phénotypes. Les analyses d'association (Genome Wide Association Study : GWAS) sont très souvent effectuées en alignant des lectures sur une référence afin d'identifier des SNPs. D'autres approches comme Scoary (BRYNILDSRUD et al., 2016) effectuent de la génomique d'association sur la base d'une matrice de présence/absence de gènes d'un pangénome. L'originalité de l'approche présentée ici est de ne pas nécessiter de référence, ni d'alignement. De même l'approche ne recherche pas spécifiquement les variants SNP mais directement les sous-graphes du cDBG associés aux phénotypes. Cela permet donc d'extraire des variants complexes comme ceux résultant de la présence d'un gène entier. De plus en utilisant un cadre graphique, il permet d'explorer les variants significativement associés dans leur contexte génomique (les auteurs ayant d'ailleurs développé un outil de visualisation *ad hoc*). Cela facilite l'identification des îlots génomiques par exemple.

En utilisant leur méthode, les auteurs montrent notamment qu'elle est à même de retrouver des k-mers associés à des résistances à des antibiotiques acquises par transferts horizontaux de gènes dans les espèces *Staphylococcus aureus* et *Pseudomonas aeruginosa*.

Enfin, les auteurs indiquent que leur méthode est plus rapide que les autres approches GWAS basées sur des k-mers comme pyseer (LEES et al., 2018) ou HAWK (RAHMAN et al., 2018).

Partie II

La méthode PPanGGOLiN

6 Motivation de l'approche PPanGGOLiN 99
6.1 Améliorer le partitionnement des pangénomes
6.2 Limites des approches tenant compte des contextes génomiques
6.3 PPanGGOLiN, le chaînon manquant de la pangénomique

7 Construire un graphe de pangénome partitionné 111
7.1 Vue d'ensemble de la méthode PPanGGOLiN
7.2 Les familles de gènes, les points de repère entre les génomes
7.3 La construction d'un graphe de pangénome
7.4 Le partitionnement du graphe de pangénome
7.5 L'emmêlement du graphe de pangénome
7.6 L'écaillage du graphe de PPanGGOLiN
7.7 L'échantillonnage des génomes

8 Description technique de la méthode PPanGGOLiN 139
8.1 Informations sur l'implémentation effectuée
8.2 Fonctionnement de l'application PPanGGOLiN
8.3 Performances de l'application PPanGGOLiN

9 Résultats de la méthode appliquée sur 439 espèces 151
9.1 Analyse des pangénomes de GenBank
9.2 Application sur des génomes issus de métagénomes



Chapitre 6 : Motivation de l'approche PPanGGOLiN

6.1 Améliorer le partitionnement des pangénomes

6.1.1 Les problèmes causés par la dichotomie génome cœur/génome accessoire

Les modélisations statistiques évoquées dans la section 3 page 45, sont très attractives mais à l'exception de celle proposée par Snipen et al., 2009 (implantée dans le paquet R micropan de Snipen et al., 2015), elles restent difficiles à mettre en œuvre. En conséquence, la grande majorité des études pangénomiques, même récentes, semblent très souvent préférer la simple dichotomie entre génome cœur et génome accessoire pour rendre compte de la diversité des espèces qu'ils étudient (comme par exemple Araujo et al., 2019 ; Rasigade et al., 2018). Néanmoins, contrairement à ce que pensaient les auteurs de Tettelin et al., 2005, le concept de génome cœur, entendu en son sens strict, conduit mécaniquement à ce que celui-ci décroît avec le nombre de génomes ajoutés au pangénome. Cela a notamment été montré pour l'espèce *E. coli* par les auteurs de Lukjancenko et al., 2010 en estimant qu'avec 61 génomes, le génome cœur possède 993 familles de gènes, soit bien moins que les 2 344 familles indiquées par les auteurs Rasko et al., 2008 basées sur 17 génomes ou bien que les 1 976 familles retrouvées par les auteurs de Touchon et al., 2009 (basées sur 20 génomes). Ceci est problématique pour plusieurs raisons. D'abord car cela n'a aucun sens biologique car s'il existe un génome cœur de fonctions indispensables à la survie des espèces, il n'a pas de raison de décroître massivement en fonction du nombre de génomes inclus dans le pangénome. De plus, comme le génome cœur est censé correspondre aux éléments stables donc à la signature même des espèces, alors si sa taille est dépendante du nombre de génomes connus, il devient très difficile de comparer des pangénomes entre eux. Enfin, la décroissance du génome cœur est ici très dépendante de la qualité des génomes utilisés. Il suffit d'un seul génome ayant des gènes manquants pour réduire fortement la taille du génome cœur. Une solution simple consisterait dans un meilleur choix des génomes à intégrer dans le pangénome en s'assurant que ceux-ci soient d'une qualité irréprochable. Ceci dit, cela limiterait considérablement la quantité d'information disponible pour estimer la diversité génomique d'une espèce. De plus, cette solution demeurerait un pis-aller car même en ayant un niveau de qualité des génomes irréprochable, le

phénomène n'en serait que déplacé et le génome cœur continuerait de tendre vers sa diminution avec un nombre infini de génomes. En effet, si l'origine de ces pertes peut effectivement être attribuable à des problèmes techniques, il serait faux de penser que cela en est la cause unique.

6.1.2 Causes de la perte des gènes cœurs

Les causes techniques

Pour ce qui est des problèmes méthodologiques menant à une perte de gènes, ceux-ci peuvent advenir à toutes les étapes en amont de la construction du pangénome. En effet, pour obtenir des génomes de qualité, il faut d'abord s'assurer que l'ensemble des molécules d'ADN des génomes soient correctement extraites, en particulier pour les plasmides et les chromides (pour les espèces en possédant). Il faut ensuite s'assurer que les étapes d'amplification ne laissent pas de régions non amplifiées à même de provoquer des ruptures dans les assemblages. Enfin, il faut aussi penser que les gènes se trouvant dans des régions répétées du génome sont susceptibles d'être plus ou moins bien assemblés en fonction de la technologie de séquençage et/ou des méthodes d'assemblage utilisées. Ceci à cause du fait que les logiciels peuvent produire des assemblages fragmentés ou incorrects en cas répétitions de segments excédant la taille des lectures de séquençage.

Pour que les gènes soient bien détectés à l'annotation il est préférable de ne pas avoir d'erreurs de séquence amenant à des insertions et/ou à des délétions susceptibles de provoquer des décalages de phase dans les régions codantes. Ces décalages peuvent en particulier être causés par les technologies de séquençage Roche 454 ou Oxford Nanopore connues pour mal déterminer le nombre de nucléotides successifs dans les régions homopolymériques de l'ADN (Luo et al., 2012 ; Tyler et al., 2018).

De plus, une fois les gènes identifiés, ils se doivent d'être correctement regroupés en familles de gènes homologues, c'est-à-dire qu'ils doivent ne pas avoir trop de mutations et être de taille complète pour que les éventuelles heuristiques de l'algorithme de génération de familles n'oublient aucune séquence. En effet, si des gènes homologues sont de tailles trop divergentes, ils peuvent éventuellement être placés dans des familles différentes.

À travers ces différentes étapes où un gène peut être ignoré, perdu ou mal classifié, on peut ainsi se rendre compte du nombre de paramètres pouvant influer sur la taille du génome cœur. L'ennui étant que plus le nombre de génomes s'accumulant dans le pangénome est grand, plus ces problèmes s'amoncellent causant ainsi la décroissance du génome cœur.

Les causes biologiques : apport de la biologie de synthèse et hypothèse de la reine noire

Au-delà de la technique, la notion de génome cœur est aussi à préciser d'un point de vue biologique. En effet, si certains gènes sont absolument indispensables, et ce quelles que soient les circonstances, à la survie d'une cellule bactérienne comme les gènes impliqués dans la synthèse des protéines ou la plupart de ceux chargés de l'intendance de l'ADN (réplication, réparation), à l'inverse, d'autres fonctions clés n'ont pas toujours besoin d'être présentes dans la cellule. Déterminer la limite entre les gènes indispensables et les gènes quasi-indispensables est difficile. La question se pose notamment dans le domaine de la biologie de synthèse. La raison est que pour concevoir des bactéries d'intérêt industriel, il est nécessaire d'avoir des organismes « châssis » sur

lesquels implanter artificiellement des fonctions. Pour cela, les chercheurs mènent une quête depuis les années 1960 pour le « Saint Graal » [sic] qu'est le génome minimum, c'est-à-dire l'ensemble des gènes strictement essentiels au vivant en général. Néanmoins, cette quête s'avère plus compliquée que prévu car le génome minimal est en pratique intimement lié au contexte de vie de la cellule, de sorte que les auteurs peinent à trouver un consensus sur les gènes constitutifs de ce génome. C'est pourquoi, plutôt que de rechercher les gènes essentiels, les auteurs de ACEVEDO-ROCHA et al., 2013 proposent de considérer le génome minimal comme les gènes généralement essentiels, que les auteurs nomment des gènes persistants [1].

Si des cellules peuvent parfois vivre en l'absence de gènes pourtant persistants, c'est à cause du fait que celles-ci sont placées dans des conditions très spécifiques qui ne sont pas généralisables à l'espèce entière. Localement cela peut même s'avérer sélectivement avantageux si celles-ci perdent un gène dans un contexte où la fonction n'est pas absolument nécessairement. En effet, exprimer un gène inutile ou moins utile que ne l'est son « coût » métabolique peut localement être un désavantage. Ainsi un gène cœur, en temps normal indispensable à l'espèce, peut être perdu. Dès lors, ce qui permettra de dire si le gène est persistant ou non, c'est donc le caractère isolé ou non de cette réduction de génome.

Mais au-delà du gène, il faut bien comprendre que ce qui est nécessaire à la croissance de la cellule, ce n'est pas ontologiquement le gène mais la fonction qu'il assure uniquement. D'après certaines études, il s'avère que la fonction d'un gène pourrait tout à fait être assurée pour une cellule donnée, alors même que le gène codant pour cette fonction en serait absent. Comment cela est-il possible sans verser dans des théories peu convaincantes à base de mémoire de l'eau [2] ? En 2012, les auteurs de MORRIS et al., 2012 étudient deux types de bactéries planctoniques, les cyanobactéries du genre *Prochlorococcus* et la bactérie marine la plus abondante sur Terre *Pelagibacter ubique*. Les auteurs suggèrent qu'une réduction de leur génome (perte de gènes) s'est produite et que ces pertes ne seraient pas dues à la dérive génétique mais bien au fait que cela leur aurait procuré des avantages sélectifs. Pour expliquer ce phénomène, les auteurs postulent qu'il peut être avantageux pour des bactéries vivant en communauté de perdre certains de leurs gènes si les fonctions associées sont disponibles sous la forme de « bien public » [sic] dans l'environnement. On observerait ainsi une microéconomie bactérienne mêlant une division du travail enzymatique avec des échanges de métabolites entre des cellules aidantes et des cellules bénéficiaires. Cette hypothèse nommée l'hypothèse de la reine noire (Black Queen hypothesis : BQH) fait allusion à une hypothèse célèbre en co-évolution, l'hypothèse de la reine rouge [3]. L'hypothèse de la reine noire s'inspire du jeu de carte « la dame de pique » où le but est de gagner le moins de points possible. Lorsque l'on remporte

1. Attention, cela ne doit pas être confondu avec la notion de « persistance » en microbiologie, c'est-à-dire, la capacité d'une sous-population de microbes à survivre à un traitement antibiotique sans pourtant avoir acquis les modifications génétiques de résistance à ce traitement.

2. Dont on peut noter que même un récipiendaire du prix Nobel comme le Pr. Luc Montagnier peut s'en faire le thuriféraire.

3. L'hypothèse de la reine rouge en biologie de l'évolution (VAN VALEN, 1973) fait référence au livre du mathématicien Lewis Caroll « Alice aux pays des merveilles ». Dans ce livre, Alice demande à la reine rouge : « Mais, Reine Rouge, c'est étrange, nous courons vite et le paysage autour de nous ne change pas ? ». Et cette dernière de lui répondre : « Ici il faut courir pour rester à la même place. Pour aller quelque part, il faudrait courir deux fois plus vite ». Métaphoriquement, l'hypothèse postule ainsi la nécessité d'une évolution permanente des espèces pour maintenir leurs adaptations au regard de celles des espèces avec lesquelles elles co-évoluent. Le cas idéal-typique étant le système proie-prédateur.

des plis, les points sont donnés par les cartes cœurs valant 1 point et par la dame de pique valant 14 points (la reine noire). Deux stratégies s'offrent au joueur pour gagner la partie : soit gagner tous les points possibles (26 points) et dans ce cas le joueur ne marque aucun point tandis que tous les autres joueurs gagnent 26 points chacun (avec le risque de manquer le défi d'un seul point et donc de gagner 25 points), soit de faire le moins de plis possibles contenant des cartes cœur ou la dame de pique. C'est la même chose pour les cellules bactériennes, un de leurs comportements peut être d'avoir tous les gènes nécessaires, sans exception, pour survivre dans leur environnement. Ce faisant, elles devront en endurer la totalité du coût métabolique sachant que la moindre absence de gène peut être délétère. C'est le cas par exemple quand une voie métabolique est interrompue à cause de l'absence d'une seule enzyme ; tous les autres gènes de la voie peuvent alors s'en trouver inutiles. Ou bien à l'inverse, une autre approche de survie, comme le postule la BQH, peut consister dans le fait de se répartir les fonctions à assurer entre cellules d'une communauté. C'est au demeurant, la formule qu'adoptent les organismes multicellulaires.

Voilà donc quelques éléments de réponse susceptibles d'expliquer les pertes des gènes chez certains organismes, motivant ainsi l'idée de réfléchir en termes de génome persistant au lieu de génome cœur strict. Ainsi, si nous reprenons à notre compte ce concept pour la pangénomique, on peut mettre à jour la figure 3.10 page 59, en remplaçant le génome cœur par le génome persistant, ce qui nous donne le modèle de partitionnement que nous allons utiliser avec la méthode PPanGGOLiN qui a été développée dans le cadre de cette thèse (figure 6.1).

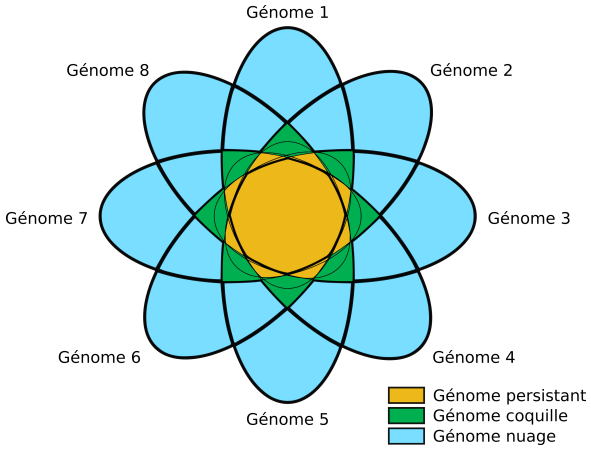


FIGURE 6.1 – Diagramme de Venn illustrant un pangénome ouvert partitionné en trois parties. En comparaison de la figure 3.10 page 59, le génome cœur est ici remplacé par le génome persistant, soit l'ensemble des gènes généralement présents dans l'espèce.

6.1.3 Assouplir la dichotomie entre le génome cœur et le génome accessoire

À la vue de ma critique du génome cœur, on est tenté de me répondre que je procède au sophisme de l'homme de paille. Et c'est vrai que cette remarque se justifie car en pratique de très nombreuses études, conscientes de ce problème n'utilisent pas une conception stricte du génome cœur mais lui préfèrent une version assouplie. Pour cela, les auteurs des publications imposent un seuil de présence minimal pour considérer qu'un gène appartient ou non au génome cœur assoupli.

Ce seuil varie suivant les études et est fixé en général entre 90% (PERIWAL et al., 2015) et 99% de présence (LIVINGSTONE et al., 2018). S'il est vrai que ce seuil règle la majorité des problèmes précédemment indiqués, pour autant il n'est pas exempt de limites comme illustré sur la figure 6.2.

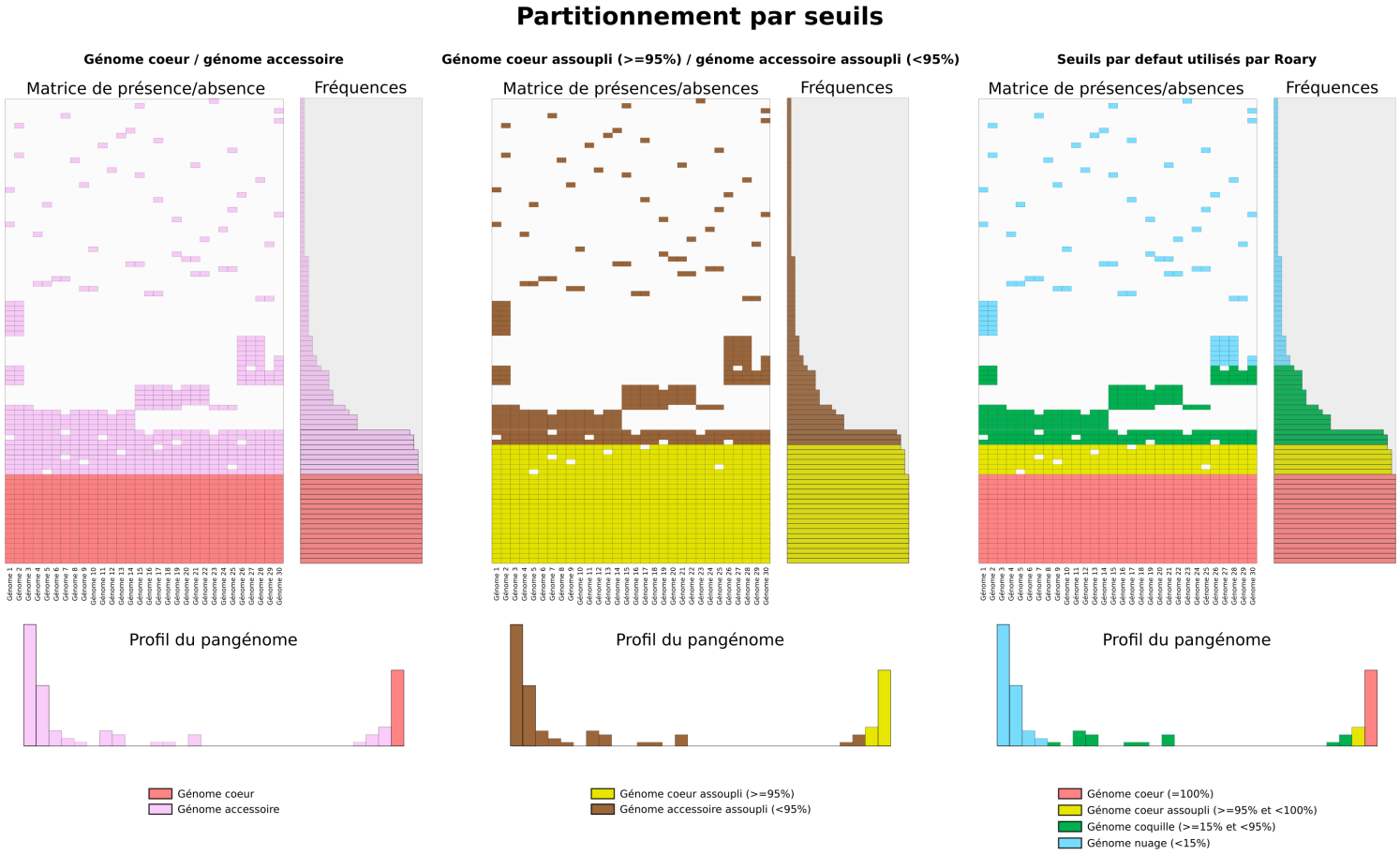


FIGURE 6.2 – Illustration d'un pangénome sous la forme d'une matrice de présence/absence, avec en lignes, les familles de gènes homologues provenant de 30 génomes (correspondant aux colonnes). La fréquence de la présence de chaque famille de gènes parmi les 30 génomes est illustrée sur la droite de la matrice tandis que le profil du pangénome est indiqué sous la matrice. La matrice est partitionnée selon les trois approches de très loin les plus utilisées dans la littérature, c'est-à-dire sur la base de différents seuils de présence.

D'abord ce seuil séparant le génome cœur assoupli du génome accessoire assoupli est difficile à estimer, de sorte qu'en pratique le seuil utilisé est très souvent celui que propose par défaut le logiciel le plus utilisé en pangénomique (Roary, PAGE et al., 2015), c'est-à-dire un seuil de 95% de présence. Il faut ici noter que le terme génome cœur assoupli (« soft core genome ») désigne en fonction des études, l'ensemble des gènes présent à une fréquence supérieur à un seuil, génome cœur compris ou bien la même chose génome cœur exclu (comme dans l'outil Roary par exemple). J'informe le lecteur qu'à partir de maintenant, seul le premier sens sera employé dans ce manuscrit.

Lorsqu'on utilise un seuil de 95%, cela veut donc dire que ce seuil ne devient opérationnel qu'à partir de 20 génomes ($\frac{100}{100-95} = 20$), car en-dessous le génome cœur assoupli est identique au génome cœur. De même, à 20 génomes, une seule absence est autorisée, à 40 génomes, deux absences sont autorisées, etc... Il en résulte ce que j'appelle « l'effet escalier » que l'on peut observer sur la figure 6.3.

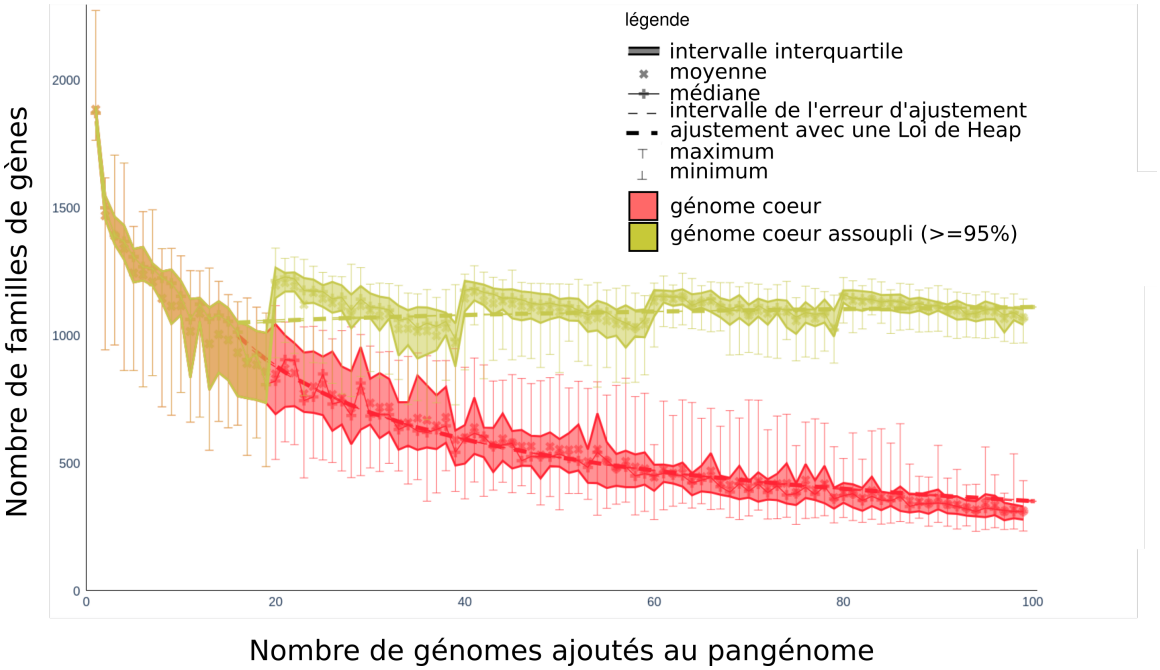


FIGURE 6.3 – Courbe de raréfaction du génome cœur et du génome cœur assoupli de l'espèce *Oenococcus oeni*. Sur la courbe du génome cœur assoupli, on note un « effet escalier » tous les 20 génomes. Même si cet effet semble particulièrement bien marqué sur cette espèce, il est présent sur la majorité des courbes de raréfaction.

On pourrait bien évidemment baisser ce seuil, mais quelle en serait la limite ? La question même de l'universalité de ce seuil se pose. Idéalement, il faudrait le ré-estimer pour chaque pangénome. Certains gènes persistants pouvant être plus facilement perdus que d'autres, ne serait-ce que pour des raisons techniques (présence de segments répétés dans le gène).

Si l'on constate quelques limites à l'approche génome cœur assoupli, cela reste toutefois raisonnablement surmontable dans la majorité des cas, raison pour laquelle ce seuil de présence est toujours utilisé dans un très grand nombre d'études. Cependant, cette remarque est valable uniquement pour des génomes dont les séquences sont obtenues partir d'organismes isolés puis mis en culture, soit dans les meilleures conditions expérimentales pour aboutir à des génomes fiables et complets. Mais quid des génomes provenant de méthodes notoirement connues pour produire des génomes incomplets comme ceux assemblés depuis les amplicons de cellules uniques (Single-cell Amplified Genome : SAG) ou bien ceux assemblés depuis des métagénomes (Metagenome-Assembled Genome : MAG, évoqué plus en détail à la sous-section 9.2.1 page 173). Ici, l'approche

génome cœur assoupli à 95% devient très difficile à soutenir. Là encore, baisser le seuil est toujours possible mais outre les remarques déjà effectuées, la recherche du seuil idéal dans le cas de génomes passablement incomplets devient particulièrement épineuse.

Pour résumer, on constate que bien que l'idée triviale d'un seuil de présence pour assouplir le génome cœur soit attirante, elle ne peut prétendre à devenir la panacée face aux limites évoquées. Pour répondre à ces écueils, il semble préférable d'avoir des méthodes cherchant à optimiser des seuils de présences pour détecter le génome persistant d'un pangénome. C'est ce que font des méthodes comme micropan (SNIPEN et al., 2015), celle des auteurs de TONDER et al., 2014 ou bien encore celle de GUMIERE et al., 2018 en déterminant des seuils de fréquences afin de partitionner le pangénome. Ces deux dernières méthodes restent cependant bien moins polyvalentes que micropan (SNIPEN et al., 2015) en ceci qu'elles sont limitées à la simple dichotomie génome persistant/-génome accessoire, c'est-à-dire à un partitionnement statistique en deux parties uniquement. Or, comme nous allons le voir à la section suivante, le même problème de seuil se pose également pour détecter le génome coquille et le génome nuage ; ce à quoi le paquet R micropan propose une solution contrairement aux méthodes de GUMIERE et al., 2018 ; TONDER et al., 2014.

6.1.4 Détecter le génome coquille et le génome nuage

À ce stade, nous avons vu notamment à la section 3.4 page 57, que la trichotomie génome persistant (au lieu de génome cœur), génome coquille et génome nuage semble mieux rendre compte des diverses gammes d'occurrences des familles de gènes dans un pangénome que les approches dichotomiques. Après avoir constaté les difficultés que posent la séparation du génome cœur assoupli, la question qui se pose dès lors est : comment séparer le génome coquille du génome nuage ? S'il n'est pas si simple de trouver un seuil pour le génome cœur assoupli, le problème devient effectivement prégnant quand il s'agit de trouver le seuil de séparation entre le génome coquille et le génome nuage. En effet, entre le génome persistant et le génome coquille on observe très souvent une chute relativement nette de densité de présence alors que ce n'est pas nécessairement le cas entre le génome coquille et le génome nuage où l'on observe très souvent un gradient. Pour solutionner ce problème, Roary (PAGE et al., 2015) utilise, à nouveau, des seuils fixes ($\geqslant 15\%$ et $< 95\%$ de fréquence de présence dans le pangénome pour le génome coquille) dont les valeurs par défaut ne sont que très rarement ré-adaptées par les auteurs des études pangénomiques l'employant. Pourtant, ce seuil arbitraire est inadapté dans la majorité des cas tant la délinéation entre génome coquille et génome nuage est parfois subtile comme illustré sur la figure 6.2 page 103. En effet, il faut voir que le génome coquille couvre une large gamme de fréquences de présence.

Dans certain cas, il est même préférable que le génome coquille soit divisé en sous parties. En effet, comme l'avait déjà montré les auteurs de SNIPEN et al., 2009 à travers le tableau 3.2 page 56, le nombre de parties dans un pangénome, sans pour autant remettre en cause la trichotomie en général, peut être supérieur à trois. Dès lors, il faut aussi trouver des séparations à l'intérieur du génome coquille. Ainsi, il devient compliqué de se reposer uniquement sur des seuils fixes de présence arbitrés humainement pour partitionner le pangénome, de sorte qu'une méthode statistique *ad hoc* semble indiquée.

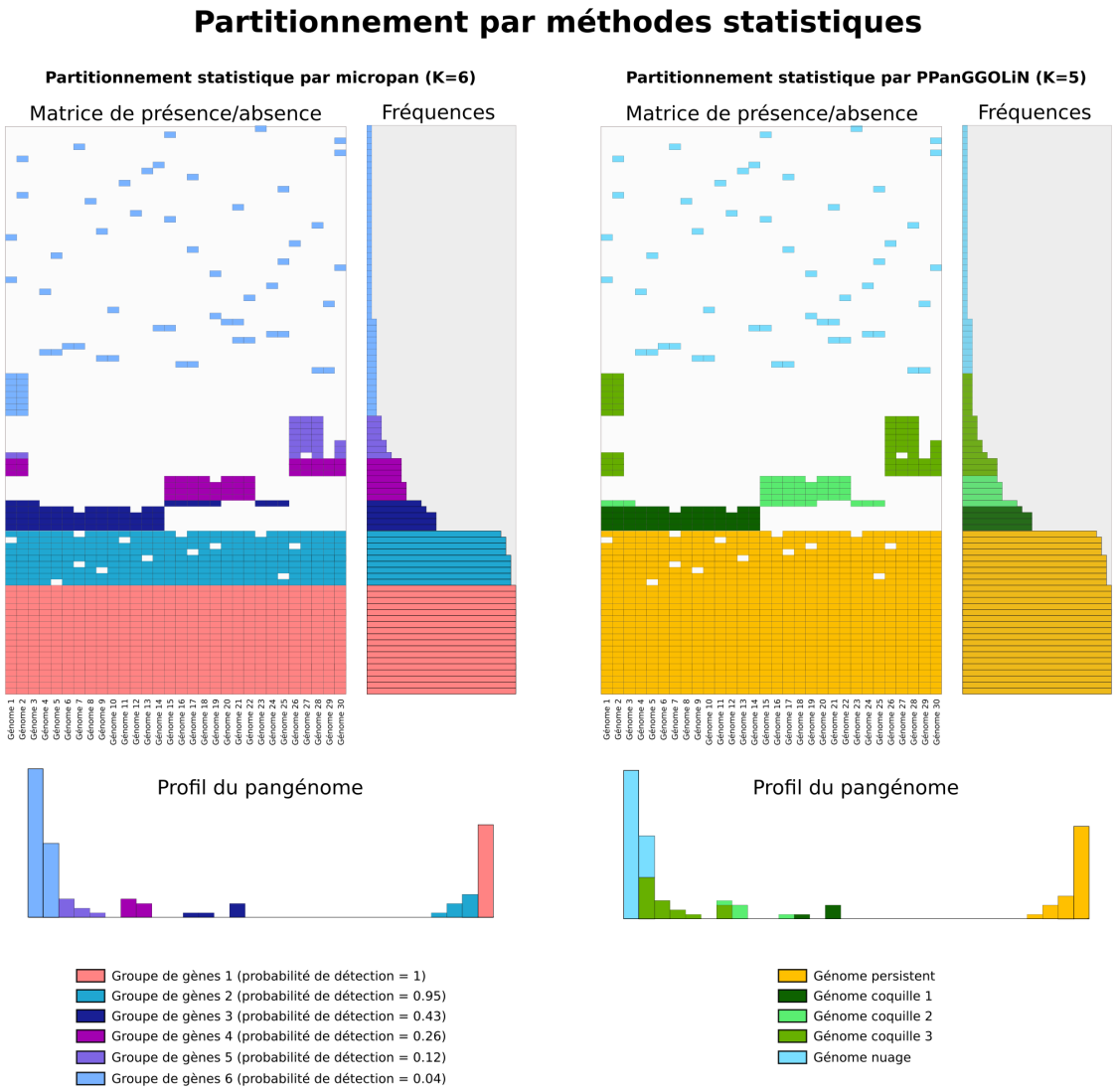


FIGURE 6.4 – Illustration d'un pangénome sous la forme d'une matrice de présence/absence, avec en lignes, les familles des gènes homologues provenant de 30 génomes (correspondant aux colonnes). La fréquence de présence de chaque famille de gènes parmi les 30 génomes est illustrée sur la droite de la matrice tandis que le profil du pangénome est indiqué sous la matrice. La matrice est partitionnée selon deux approches statistiques. La première est la méthode micropan (décrite à la sous-section 3.3.2 page 54) qui partitionne le pangénome sur la base de la fréquence de présence des familles de gènes. La deuxième est la méthode PPanGGOLiN (décrite au chapitre 7 page 111) qui partitionne le pangénome sur la base des schémas (ou vecteur) d'occurrences similaires des familles de gènes. Le nombre K correspond au nombre de parties détectées.

Pour répondre à cela, nous l'avons vu (sous-section 3.3.2 page 54 et 5.1.12 page 87), il existe déjà un outil cherchant à optimiser des seuils de présence entre les différentes parties du pangénome, le paquet R micropan SNIPEN et al., 2015). Cet outil va partitionner le pangénome en constituant des groupes de familles de gènes ayant des fréquences de présence similaires, ce qui semble répondre aux problèmes soulevés jusqu'ici. Toutefois, cette méthode peut manquer une part importante de

l'information en ne se basant que sur la fréquence de présence. En effet, avec des schémas de présence/absence comme illustré dans la figure 6.4 page précédente, il est préférable d'opérer le partitionnement non pas sur la base des fréquences uniquement mais sur la base des schémas de présence/absence dans leur globalité. Dans l'alternative, une part de l'information est manquée et cela mène à un découpage s'ajustant peu à la matrice de présence/absence. Imaginons un exemple où un ensemble de familles de gènes est présent dans la moitié des génomes en étant absent des autres et un autre ensemble de familles de gènes où c'est exactement l'opposé. Dans cet exemple, les deux ensembles ont la même fréquence de présence en ayant pourtant des schémas de présence/absence strictement antagonistes. Cela correspond à ce que les auteurs de MOLDOVAN et al., 2018 nomment des espèces non homogènes, c'est-à-dire dont le profil de pangénome est en forme de **W** comme déjà évoqué à la sous-section 3.5.4 page 65. Comme exemple, on peut notamment citer l'espèce marine photosynthétique *Prochlorococcus marinus* dont le génome coquille contient deux schémas antagonistes. En effet, soit la bactérie possède les gènes pour vivre à la surface de la mer donc avec une haute luminosité, soit à l'inverse elle possède les gènes pour vivre plus en profondeur (DELMONT et al., 2018) notamment jusqu'à accepter moins de 1% de la lumière disponible à la surface. Pourtant avec la méthode micropan, ces deux groupes de familles se retrouveront partitionnés ensemble. À l'inverse, en utilisant une méthode comme PPanGGOLiN se basant sur les schémas de présence/absence, il devient possible de les séparer en deux parties différentes. Enfin, comme nous le verront à la section 6.3 page 110, PPanGGOLiN exploite encore davantage d'information que les schémas de présence/absence des familles de gènes.

6.2 Limites des approches tenant compte des contextes génomiques

6.2.1 Périmètre de l'usage des graphes de pangénome à l'échelle de la séquence

Comme je l'ai déjà indiqué, se développent, en pangénomique depuis quelques années, de nombreuses approches utilisant un graphe à l'échelle de la séquence pour représenter la diversité des variations nucléotidiques dans un pangénome. Ces représentations constituent une rupture avec la pangénomique habituelle basée sur des ensembles de familles de gènes isolées. À l'origine, ces graphes ont principalement été développés pour les génomes eucaryotes où le nombre de variants nucléotidiques mais aussi le nombre de gènes spécifiques par génome (hors duplication) est généralement plus petit que pour les génomes procaryotes (MCINERNEY et al., 2017). Pour rappel, au sein d'une espèce bactérienne, le taux d'identité moyen entre les parties conservées de deux génomes peut descendre jusqu'à 95% alors que chez *Homo sapiens* par exemple, il est de 99.9% entre deux individus (même s'il existe également de nombreux variants structuraux). Certes, ce chiffre de 95% est une moyenne minimale, car pour certaines espèces bactériennes dites clonales, on peut observer des taux d'identité bien supérieurs. Il reste qu'avec un nombre de génomes procaryotes important (des centaines, des milliers), l'exploration d'un graphe de De Bruijn ou d'un graphe de variants reste difficile, sauf à ne se concentrer que sur une région particulière. D'autant que dans le monde microbien, bon nombre de variants nucléotidiques sont a priori sans importance biologique notable. En effet, il faut se rappeler, d'une part que les ADN polymérases bactériennes font $\approx 10^{-8}$ à $\approx 10^{-9}$ erreurs par base (PRAY, 2008) lors de la réplication d'un génome et d'autre

part que certaines bactéries se divisent jusqu'à une fois toutes les $\approx 20-30$ minutes en milieu non limitant, soit $\approx 10^{21}$ cellules générées à partir d'une seule cellule en une seule journée [1]. Ainsi de nombreux variants plus ou moins neutres sont générés, de sorte que la visualisation d'un graphe de séquences pour une espèce bactérienne doit être couplé *a minima* avec un filtrage des variants pour être plus commodément exploitable. On peut penser notamment aux variants synonymes, qui bien que pouvant impacter la vitesse d'expression des gènes (et donc possiblement la structure des protéines RAUSCHER et al., 2018), sont supposés avoir moins d'effet. De plus, dans ces graphes de pangénome, des méthodes statistiques pourraient mettre en évidence les variants abondants, ces derniers étant davantage susceptibles d'être signifiants que les variants rares. Enfin, le mieux semble de pouvoir associer ces variants à des phénotypes par une approche statistique comme le fait une méthode comme DBGWAS (JAILLARD et al., 2018). Cela nécessite cependant d'obtenir des métadonnées sur les organismes comme par exemple : leurs pathogénies, leurs résistances attestées à des antibiotiques ou bien encore leurs environnements d'isolation.

6.2.2 Intérêt de considérer la colocalisation des gènes

Avec une autre perspective, les auteurs de CHAN et al., 2015 proposent l'idée, plutôt que de se baser sur les variants au niveau nucléotidique, de représenter les variations du contenu en gènes dans les génomes via ce qu'ils nomment le « panchromosome » avant d'employer cette approche pour l'étude de l'espèce *Acinetobacter baumannii*. Le panchromosome désigne un chromosome consensus contenant une succession de blocs de gènes cœurs interrompus par des blocs de régions spécifiques à certains génomes, regroupée par tailles, puis disposés de façon concentrique (voir figure 6.5 page ci-contre). Cette analyse basée sur 249 génomes leur a permis de découvrir de nouvelles régions de plasticité génomique, c'est-à-dire, des zones où se produisent d'importantes variations en termes de gains et de pertes de gènes. Pour ce type d'analyse, se placer à l'échelle du gène semble plus naturel pour faire le lien avec les fonctions impliquées dans une région génomique. En l'espèce, les auteurs de CHAN et al., 2015 montraient des régions variables dédiées au métabolisme du carbone, à la production de sidérophores et ou même à la construction de pilus.

Pour un biologiste, le gène a un sens fonctionnel fort : ceux-ci peuvent être mobiles, former des régions, de sorte que la manière dont ils s'organisent, la manière dont ils sont obtenus ou perdus revêt une grande importance. Toutefois, il faut aussi noter qu'avec les graphes de séquences, il est également possible de colorer (annoter) les sous-graphes correspondant aux gènes et ainsi retrouver cette information. Par contre, des régions contenant beaucoup de polymorphismes peuvent générer plusieurs chemins parallèles dans le graphe pour un même enchaînement de gènes et compliquent ainsi leur analyse. De plus, se placer à l'échelle des gènes permet une intégration de l'information, donc diminue la quantité d'objet à traiter que ce soit visuellement ou statistiquement. Enfin, le niveau gène permet d'accéder plus facilement à la variabilité au niveau de régions génomiques comme les îlots ou les opérons. En effet, l'ordre des gènes est en soit une chose très importante pour l'étude de l'évolution des génomes procaryotes car les gènes essentiels sont en général organisés sous la forme de régions ayant un ordonnancement conservé dans les génomes ROCHA et al., 2003

1. À cette vitesse, cela amène notamment à la création de multifourches de réplication d'ADN car l'ADN se réplique plus vite que ne se divisent les cellules.

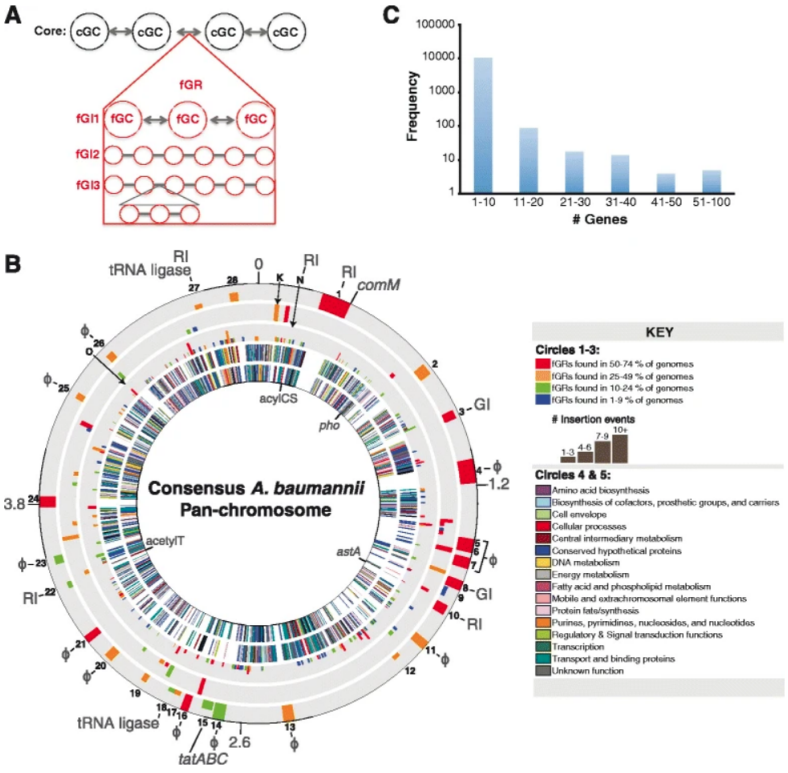


FIGURE 6.5 – Illustration d'un panchromosome reproduit depuis CHAN et al., 2015. Les blocs de gènes cœurs (cGC), les blocs de gènes variables (fGC) et les régions génomiques variables (fGR) du pan-génome (a) d'*A. baumannii* ont été calculés à partir des résultats de PanOCT (abordé à la sous-section 5.1.5 page 84). Ils sont représentés comme des cercles concentriques numérotés de l'extérieur vers l'intérieur (b). Les emplacements des fGR sont représentés dans les cercles 1 (>20 000b), 2 (10 001 à 20 000b) et 3 (1000 à 10 000b) en rapport avec un génome cœur assoupli de gènes dans les cercles 4 (brin positif) et 5 (brin négatif). Enfin, l'histogramme (c) indique la distribution de la taille des îlots génomiques ainsi détectés.

tandis que les parties plus variables, majoritairement issues de gènes transférés horizontalement, se retrouvent souvent dans des points chauds d'intégration OLIVEIRA et al., 2017. Ce constat sera par la suite appelé la « régularité spatiale » des différentes parties du pangénome. La localisation des gènes à l'échelle du chromosome est donc un élément crucial en soi pour nous informer sur leur utilité biologique (génome persistant ou non) et sur leur origine possible (acquis horizontalement ou verticalement). Malheureusement, ce que proposent les auteurs de CHAN et al., 2015 est une analyse ponctuelle qui n'a pas été conçue pour être reproductible sur d'autres espèces en l'état [1]. De manière générale, il existe peu d'approches pangénomiques se plaçant à l'échelle du gène tout en conservant les liens de voisinage chromosomique entre ceux-ci comme le font tacitement les graphes de séquences. On peut toutefois citer l'outil panCEA qui permet de reconstruire ses propres panchromosomes (CLARKE et al., 2018). On peut néanmoins regretter que cette approche, bien que remarquablement synthétique, ne permet pas de représenter précisément les liens entre les différentes régions variables et le génome persistant.

1. Le code source n'a notamment pas été rendu disponible.

6.3 PPanGGOLiN, le chaînon manquant de la pangénomique

Nous l'avons vu, les méthodes actuelles de partitionnement et d'exploration de la topologie génomique des pangénomes présentent des limites. L'outil développé au cours de ce travail de doctorat, PPanGGOLiN (Partitionned PanGenome Graph of Linked Neighbors) a notamment pour objectif d'y répondre par une approche de partitionnement statistique exploitant la totalité de l'information présente dans un pangénome, à commencer par l'intégralité de la matrice de présence/absence et non plus seulement la fréquence de présence des familles de gènes comme c'est généralement le cas. De plus, la méthode réutilise l'information de voisinage des gènes dans les génomes afin de construire ce que nous appelons un graphe de pangénome. Cette solution, tout en centrant son approche sur l'unité biologique fondamentale qu'est la famille de gènes homologues, permet d'explorer bien plus finement l'organisation spatiale des différents génomes que le panchromosome proposé par les auteurs de CHAN et al., 2015 et de CLARKE et al., 2018. À ce jour, je ne connais qu'une approche se rapprochant de ce concept de graphe de pangénome : l'approche MetaPGN publiée durant ma première année de thèse (PENG et al., 2018). Mais cependant, PPanGGOLiN va plus loin car la méthode statistique que nous utilisons est non seulement en mesure d'exploiter les schémas d'occurrence mais également la structuration spatiale des familles de gènes dans le graphe afin d'améliorer le partitionnement du pangénome. En effet, comme nous l'avons déjà évoqué, à la fois les régions persistantes et les régions variables se retrouvent généralement organisées de façon spatialement régulière dans le graphe de pangénome. En quelque sorte, on peut donc dire que cette approche constitue le chaînon manquant entre l'approche historique basée sur des familles de gènes indépendantes et les approches utilisant des graphes de séquences. D'ailleurs, et pour éviter tout malentendu, l'approche que propose PPanGGOLiN est loin de vouloir reléguer les approches au niveau de la séquence, simplement elle considère ces deux niveaux d'information complémentaires. Il pourrait même être pertinent d'emboîter ces deux niveaux pour la compréhension de la dynamique des génomes mais nous y reviendront dans la section 11.6 page 202, dédiée aux perspectives. Enfin, grâce à sa structure partitionnée, nous montrerons au chapitre 9 page 151, que la méthode PPanGGOLiN permet notamment :

— de partitionner le graphe de pangénome en trois parties (ou davantage si nécessaire) en utilisant une méthode statistique prenant en compte les schémas de présence/absence des familles et la topologie du graphe.
— d'explorer la diversité pangénomique d'une façon graphique.
— d'identifier l'ensemble des familles de gènes persistants dans un pangénome.
— de partitionner le pangénome y compris si celui-ci est obtenu en utilisant des génomes reconstruits à partir de métagénome ou de cellules uniques.

Mais avant d'y venir, voyons d'abord comment nous procédons pour obtenir ce graphe de pangénome partitionné.



Chapitre 7 : Construire un graphe de pangénome partitionné

7.1 Vue d'ensemble de la méthode PPanGGOLiN

La méthode PPanGGOLiN (Partitionned PanGenome Graph of Linked Neighbors) est la principale méthode développée et utilisée au cours de cette thèse de doctorat. Cette dernière permet de construire ce que nous appelons un graphe de pangénome partitionné. Un article développant ce concept et l'appliquant sur 439 espèces de procaryotes (comme nous le verrons au chapitre 9 page 151) a notamment été publié dans le journal *PLOS Computional Biology* (GAUTREAU et al., 2020).

Comme cela est illustré sur la figure 7.1 page suivante décrivant la méthode PPanGGOLiN, pour construire un graphe de pangénome partitionné, il faut d'abord obtenir des génomes assemblés et annotés, ayant leurs gènes respectifs regroupés en familles de gènes homologues. Ces génomes doivent être préférentiellement de la même espèce car la conservation de l'ordre des opérons chute rapidement lorsque l'on compare des génomes lointains (ROCHA, 2008). Or cette information est importante pour construire le graphe de pangénome comme nous allons le voir. Néanmoins, comme l'ordre intra-opéronique est tout de même conservé, il reste possible d'utiliser la méthode pour des génomes du même genre ou de la même famille, avec le risque cependant de générer des graphes plus complexes, mais également une plus grande difficulté à constituer des familles de gènes homologues avec des gènes plus distants phylogénétiquement.

L'annotation syntaxique, c'est-à-dire la détection de la position des gènes sur les génomes, peut être effectuée soit par un utilisateur de la méthode, soit par PPanGGOLiN directement à l'aide de l'outil Prodigal (HYATT et al., 2010). En cas de superposition d'un gène codant une protéine, avec un gène d'ARN ribosomique (détecté par PPanGGOLiN avec l'outil Aragorn, LASLETT, 2004) ou bien d'ARN de transfert (détecté par PPanGGOLiN avec l'outil Infernal, E. P. NAWROCKI et al., 2013), le gène prédit par Prodigal est ignoré.

Pour ce qui est des familles de gènes homologues, elles sont construites à l'aide de l'outil MMseqs2 (STEINEGGER et al., 2017) inclus dans PPanGGOLiN. Cependant, là encore les familles de gènes peuvent aussi être fournies par l'utilisateur si celui-ci souhaite utiliser une autre méthode pour regrouper ses séquences.

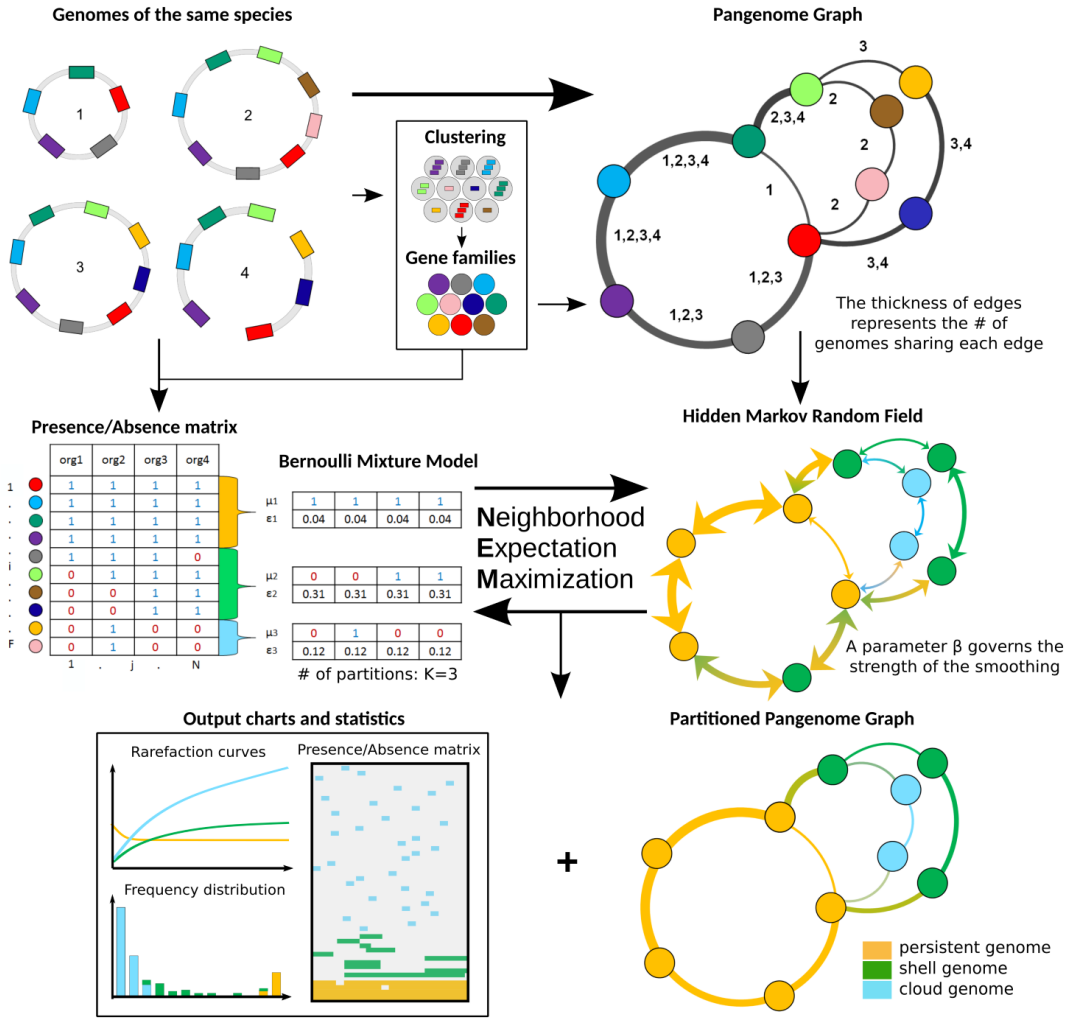


FIGURE 7.1 – Figure reproduite depuis GAUTREAU et al., 2020 illustrant les principales étapes du fonctionnement de l'outil PPanGGOLiN.

Une fois les familles de gènes obtenues, un graphe de pangénome est construit en reliant ces dernières sur la base des voisinages génomiques des gènes qu'elles englobent. Ainsi, deux familles sont reliées par une arête si elles contiennent au moins une paire de gènes adjacents dans un génome. De la sorte, les nœuds représentent les familles de gènes et les arêtes représentent le voisinage génomique. Les arêtes sont étiquetées par les génomes ayant leurs gènes partageant le même voisinage génique. Parallèlement, les familles de gènes sont encodées sous la forme d'une matrice de présence/absence indiquant pour chaque génome si au moins un membre est présent ou non. Afin de rendre compte des différents schémas de présence/absence présents dans cette matrice, nous utilisons un modèle basé sur un mélange de K distributions de Bernoulli multivariées. Le pangénome peut ainsi être partitionné en estimant les meilleurs paramètres de notre modèle via un algorithme de maximisation de vraisemblance. Dès lors, nous obtenons un pangénome divisé en K parties ($K = 3$ dans l'exemple de la figure 7.1), chacune composée de familles de gènes ayant des schémas de présence/absence homogènes. De plus, le partitionnement des familles de gènes est également contraint par un critère de régularité spatiale dans le graphe de pangénome à l'aide

d'un champ aléatoire de Markov caché (Markov Random Field : MRF). Ainsi, un partitionnement basé sur les schémas de présence/absence, mais ne maximisant pas le fait que les familles proches dans le graphe soient assignées dans la même partie, sera pénalisé. Après maximisation de la vraisemblance du modèle de mélange de Bernoulli multivarié, contraint à être spatialement régulier, on obtient alors le graphe de pangénome partitionné. On peut alors observer sur le graphe, les chemins correspondant au génome persistant, au génome coquille et au génome nuage. En outre, le logiciel fournit de nombreux tableaux, graphiques et statistiques. Enfin, le nombre de parties à détecter (K) peut être fourni par l'utilisateur ou déterminé par l'algorithme.

Il est à noter qu'il est conseillé d'utiliser au moins 15 génomes ayant des variations de contenu en gènes pour pouvoir obtenir un partitionnement intéressant. Si des utilisateurs veulent utiliser PPanGGOLiN sur moins de génomes, cela fonctionnera mais la qualité du résultat ne sera pas optimale. Par ailleurs, PPanGGOLiN calcule également le génome cœur non assoupli et assoupli avec un seuil ajustable sachant qu'avec un petit nombre de génomes, le génome cœur peut être considéré comme une métrique suffisante.

Mais tout d'abord, commençons par détailler le fonctionnement de chacune des étapes de la méthode.

7.2 Les familles de gènes, les points de repère entre les génomes

Une fois les gènes détectés, les méthodes de génération des familles de gènes sont cruciales en pangénomique. La plupart des méthodes cherchent à retrouver les orthologues entre les génomes. Pour cela, beaucoup d'auteurs utilisent des méthodes dont le passage à l'échelle sur un grand nombre de génomes (plusieurs milliers) est relativement difficile. On peut penser au fait d'effectuer des recherches de similarité par alignement des séquences de tous les gènes deux à deux pour chaque couple de génomes. C'est la raison pour laquelle, à l'exception de Roary (PAGE et al., 2015) et BPGA (CHAUDHARI et al., 2016), peu de méthodes de construction de pangénome sont à ma connaissance capables de traiter des milliers de génomes procaryotes.

Pour PPanGGOLiN, nous avons choisi la méthode MMseqs2 afin de regrouper les séquences en familles de gènes homologues. Plusieurs raisons ont présidé à ce choix. D'une part, nous souhaitions utiliser la séquence protéique plutôt que la séquence nucléique afin de gagner en sensibilité en étant tolérant aux mutations synonymes dans la construction des familles de gènes homologues. Ainsi, comme MMseqs2 est à la fois un outil spécialisé dans l'alignement de protéines et un outil permettant de les regrouper en familles, il était logique de l'envisager. D'autre part, cette méthode a la particularité d'être remarquablement véloce. Selon les auteurs de MMseqs2 (STEINEGGER et al., 2017), elle est environ 1000 à 100 000 fois plus rapide que BLASTP avec un niveau de sensibilité comparable. Par conséquent, en utilisant MMseqs2 nous disposons ainsi d'une méthode de génération de familles à la fois rapide et capable de supporter des millions de séquences protéiques, ce qui est indispensable pour pouvoir constituer des pangénomes contenant des dizaines de milliers de génomes.

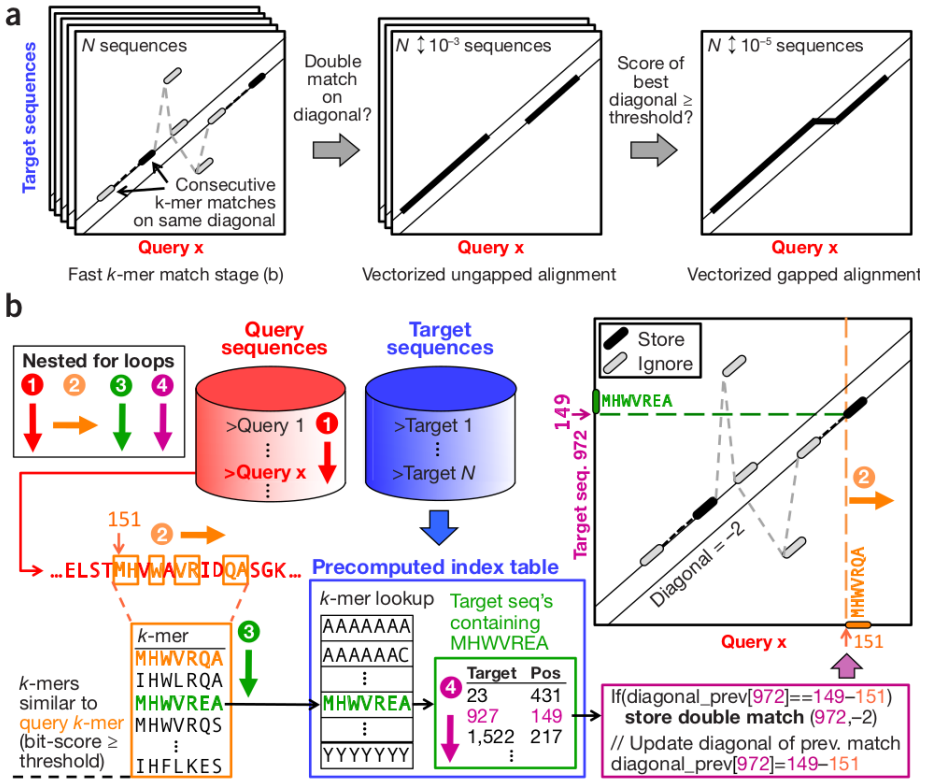


FIGURE 7.2 – Figure reproduite depuis STEINEGGER et al., 2017 illustrant le fonctionnement de l'outil MMseqs2. Une petite erreur est présente dans l'encadré vert, il faut en effet lire « 972 » et non « 927 ».

Des heuristiques d'alignement efficaces

Pour aligner des séquences protéiques contre d'autres, MMseqs2 va rechercher des couples de k-mers d'acides aminés présents sur la même séquence. Si deux couples de k-mers sont présents à la fois dans la séquence requête et dans les séquences cibles en étant séparés par un même nombre de résidus, cela se matérialise par une diagonale lorsqu'on place la séquence requête sur l'axe des abscisses et la séquence cible sur l'axe des ordonnées (**(a)** sur la figure 7.2). MMseqs2 va ainsi chercher l'ensemble des diagonales entre la séquence requête et l'ensemble des séquences cibles présentes dans la base de données. En effet, les séquences cibles sont indexées dans une base de données listant les k-mers présents et indiquant la position où ils se trouvent.

Pour chercher ces couples de k-mers, chaque séquence requête est linéairement parcourue pour produire des k-mers espacés. Sur la base de ces k-mers espacés, une liste de k-mers proches est générée. La proximité des k-mers générés va dépendre du taux de sensibilité choisi par l'utilisateur. Pour chacun de ces k-mers générés, une comparaison est faite avec les k-mers présents dans la base de données. Si une séquence possède une correspondance, alors, pour chacune des occurrences du k-mer dans la séquence, sont calculées puis stockées les différences entre la position du k-mer dans la séquence requête et la position d'une occurrence du k-mer dans la séquence cible.

Après avoir parcouru tous les k-mers approchant dans les séquences cibles, il suffit de rechercher des ensembles de k-mers séparés par un même nombre de résidus, c'est-à-dire étant sur la même diagonale. S'il en existe, alors un alignement sans autoriser d'insertion ou de délétion est effectué sur cette diagonale. Enfin, si entre les deux séquences il y a plusieurs diagonales présentes, celles-ci sont alors regroupées via un second alignement classique de Smith-Waterman (T. F. SMITH et al., 1981).

Ainsi grâce à cette méthode, nous obtenons avec célérité tous les alignements possibles d'une séquence requête sur les séquences cibles, de sorte qu'il devient envisageable d'obtenir l'alignement de milliards de protéines entre elles.

Regroupement de séquences par couverture d'ensemble

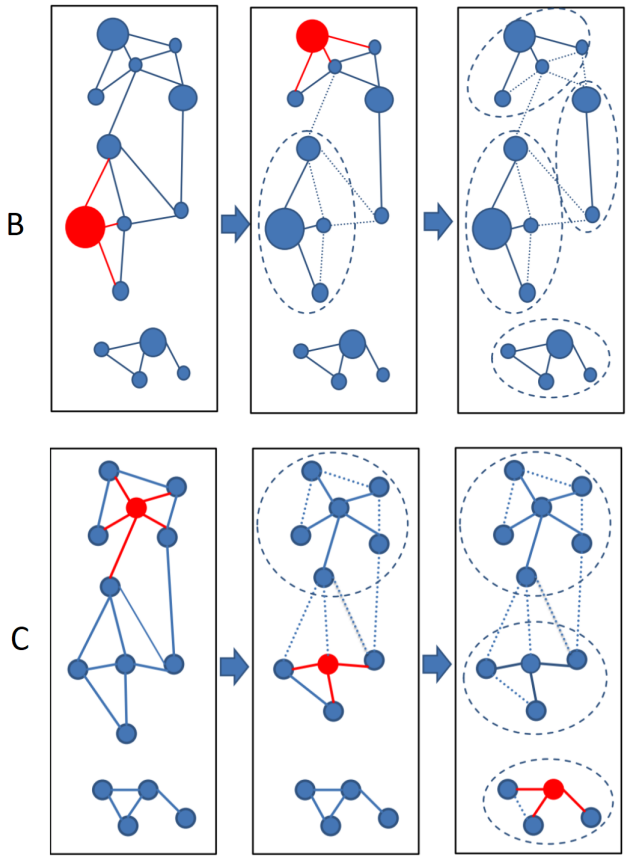


FIGURE 7.3 – Figure reproduite depuis MIRDITA et al., 2019 illustrant deux des trois méthodes de regroupement de séquences utilisées par l'outil. La méthode A : regroupement en simple lien n'est pas illustrée. La méthode B : regroupement glouton par incrément. La méthode C : regroupement glouton par couverture d'ensemble.

Après avoir aligné toutes les séquences les unes contre les autres, MMseqs2 les regroupe ensuite en familles de gènes homologues. Pour cela, nous filtrons les alignements qui sont en-dessous d'un certain seuil de couverture dans les deux sens. Ce seuil est de 80% afin de s'assurer d'avoir des familles de gènes issues de protéines complètes et composées des mêmes domaines. Comme nous l'avons vu, la moyenne d'identité entre deux gènes d'une même espèce est de 95%. Pour prendre

une marge par rapport à cette moyenne afin de prévenir une surestimation du nombre de familles qui séparerait des protéines homologues en de multiples familles, nous utilisons donc un seuil de 80% d'identité en acides aminés (par défaut). De plus, d'autres études ont également choisi ce seuil pour déterminer des gènes homologues dans leurs analyses (BATTY et al., 2018 ; FEDRIZZI et al., 2017 ; IRAOLA et al., 2017).

Après filtrage, les liens de similarité sont ensuite transformés en un graphe de similarité dont sont extraites les familles de gènes selon trois méthodes possibles (dont deux sont illustrées sur la figure 7.3 page précédente) :

A Regroupement en simple lien. Ici, chaque composante connexe formera une famille de gènes. Le gène avec le plus de liens adjacents est le gène central.

B Regroupement glouton par incrément (« greedy incremental »). Cette méthode fonctionne en sélectionnant, pour chaque composante connexe, le nœud correspondant à la séquence la plus longue (indiquée par la taille du nœud sur la figure 7.3 page précédente) ainsi que tous les nœuds connectés à lui pour former une famille. Ce processus est ensuite répété avec les nœuds restants jusqu'à ce que tous les nœuds soient dans une famille. Cette méthode est similaire à celle utilisée dans l'outil CD-HIT (W. LI et al., 2006).

C Regroupement glouton par couverture d'ensemble (« greedy set cover »). Cette méthode fonctionne en sélectionnant, pour chaque composante connexe, le nœud avec le plus de connexions ainsi que tous les nœuds connectés à lui pour former une famille. Ce processus est ensuite répété avec les nœuds restants jusqu'à ce que tous les nœuds soient dans une famille. Une fois les séquences centrales fixées, si nécessaire, les nœuds périphériques peuvent être réassignés à une autre famille si leur score d'alignement est plus élevé avec le centre de cette autre famille.

Pour construire nos graphes de pangénome, nous utilisons des familles obtenues par la méthode C, c'est-à-dire de regroupement glouton par couverture d'ensemble. En effet, nous avons exclu d'utiliser la méthode A car en se basant sur les composantes connexes, nous risquerions de mélanger ensemble des familles par transitivité, c'est-à-dire sur la base par exemple d'un seul gène faisant le pont entre deux familles. Nous avons ensuite préféré la méthode C à la méthode B car aucun élément biologique ne nous amène à penser que la séquence la plus longue dans une famille doit nécessairement être la séquence centrale pour regrouper des gènes en familles.

7.3 La construction d'un graphe de pangénome

Une fois les familles de gènes homologues obtenues, il devient possible de construire le graphe où les nœuds correspondent, comme nous l'avons dit, aux familles de gènes et où les arêtes indiquent le voisinage des gènes composant les familles.

Formellement un graphe de pangénome $G = (V, E)$ est un graphe associant un ensemble de nœuds $V = \{(v_i)_{(1 \leqslant i \leqslant F)}\}$ où F est le nombre de familles de gènes dans le pangénome et un ensemble des arêtes $E = \{e_{i \sim i'}\} = \{(v_i, v_{i'})\}, v_i \in V, v_{i'} \in V$ où les couples de nœuds $(v_i, v_{i'})$ représentent des familles dont certains gènes $(v_{i,j}, v_{i',j})$ sont adjacents [1] sur le génome j. La fonction $a(v_{i,j}, v_{i',j})$

1. La relation de voisinage est indiquée par le symbole $\sim$.

compte les occurrences de ces adjacences dans les N génomes. Ainsi, chaque arête $\{e_{i\sim i'}\}$ a un poids $w_{i\sim i'}$ où $w_{i\sim i'} = \frac{1}{N}\sum_{j=1}^{N} a(v_{i,j}, v_{i',j})$. Les arêtes sont étiquetées par un dictionnaire où les clés correspondent aux génomes et où les valeurs associées indiquent le nombre de fois où le voisinage apparaît pour chaque génome j (les valeurs nulles ne sont pas stockées). Ainsi, le poids total $w_{i\sim i'}$ de chaque arête peut être obtenu en divisant la somme des valeurs stockées dans chaque dictionnaire $\sum_{j=1}^{N} a(v_{i,j}, v_{i',j})$ par le nombre total de génomes utilisés N. Il est à noter que les arêtes réflexives dans le graphe sont considérées de la même façon que les autres arêtes. Ces arêtes réflexives proviendront de gènes appartenant à la même famille répétée en tandem dans les génomes. De plus, si les génomes ajoutés au graphe sont indiqués comme étant circulaires, le graphe peut parfaitement les prendre en compte. Dans ce cas, ceux-ci sont ajoutés dans le graphe de pangénome de façon circulaire également, c'est-à-dire en considérant que le dernier gène de la séquence est voisin du premier, donc en ajoutant une arête entre les nœuds correspondant aux familles auxquelles ils appartiennent.

L'intérêt de ce graphe est qu'il est construit par simple accumulation de génomes sur celui-ci, de sorte que les nouveaux génomes ajoutés n'ont besoin d'aucune information sur les génomes déjà présents. Ainsi la construction d'un graphe de pangénome est d'une complexité linéaire en temps.

Néanmoins, le graphe de pangénome ainsi structuré présente également des inconvénients. En effet, comme les arêtes ne stockent que des voisinages directs, il n'est pas possible d'assurer que les génomes puissent être reconstitués à partir du graphe de pangénome dans tous les cas. Autrement dit, le graphe de pangénome est susceptible de perdre une partie de l'information apportée par les génomes. C'est notamment le cas si un génome possède plusieurs gènes ayant des voisinages différents mais appartenant à une même famille de gènes. Nous y reviendrons notamment dans la section 7.5 page 129 consacrée au démêlement du graphe de pangénome. Enfin, le graphe est non orienté, car à l'échelle pangénomique, comme il n'y a plus de référence, il n'existe pas d'orientation absolue. Ainsi, les gènes, non plus, ne stockent pas l'information de leur orientation. Ceci dit, il serait parfaitement possible de choisir arbitrairement une orientation de référence (par exemple l'orientation majoritaire), puis de stocker l'orientation des gènes, relativement à cette référence arbitraire.

7.4 Le partitionnement du graphe de pangénome

Pour expliquer le fonctionnement de la méthode de partitionnement du pangénome, on abordera d'abord le problème en se basant uniquement sur la matrice de présence/absence puis nous décrirons une évolution de cette approche prenant en compte le graphe de pangénome pour améliorer ce partitionnement.

7.4.1 Modéliser la matrice du pangénome par un modèle de mélange de Bernoulli

Nous l'avons dit, nous souhaitons partitionner en K parties ($K \in \mathbb{N}; K \geqslant 3$) la matrice du pangénome sur la base des schémas de présence/absence de familles de gènes qu'elle présente. Formalisons d'abord cette matrice de présence/absence. Celle-ci est une matrice binaire X ayant F lignes correspondant aux familles de gènes et N colonnes correspondant aux différents génomes.

Pour chaque entrée binaire $x_{i,j}$, $x_{i,j} = 1$ si la famille i ($1 \leqslant i \leqslant F$) possède un ou plusieurs gènes provenant du génome j ($1 \leqslant j \leqslant N$) et dans tout autre cas $X_{i,j} = 0$ (voir illustration sur la figure 7.1 page 112). Ainsi, la partie correspondant au génome persistant sera dense alors en valeurs 1 tandis que le reste du pangénome sera bien plus dense en valeurs 0.

Pour modéliser les présences et les absences dans la matrice nous allons utiliser un modèle de Bernoulli. Un essai de Bernoulli correspond à un évènement binaire, ce qui est le cas de la présence ou de l'absence d'un gène dans un génome j pour une famille i donnée. Des évènements binaires indépendants sont habituellement modélisés par une loi de Bernoulli, soit une distribution discrète paramétrée par la probabilité de succès (généralement appelée p) de chaque essai. Ici chaque famille de gènes peut être considérée comme le résultat de N essais de Bernoulli, soit un par génome, de sorte que cela doit être modélisé par N distributions de Bernoulli ayant des paramètres potentiellement différents. Par conséquent, pour modéliser chaque famille du pangénome nous avons donc besoin d'un modèle multivarié (ou multidimensionnel) de Bernoulli. De plus, nous nous attendons à avoir plusieurs distributions de schémas de présence/absence à l'intérieur de la matrice. En effet, ces distributions correspondent au nombre de parties (K) dans le pangénome, généralement égal à 3 afin de correspondre à la trichotomie du pangénome précédemment discutée, c'est-à-dire : au génome persistant (haute densité en présence), au génome coquille (densité intermédiaire en présence/absence) et au génome nuage (haute densité en absence). Néanmoins, celui-ci peut être supérieur si la structure de la matrice est mieux modélisée par un nombre supérieur de partitions comme nous le verrons à la sous-section 7.4.3 page 123. Ce sera en particulier le cas, s'il existe plusieurs schémas antagonistes de génome coquille dans la matrice. Ces observations faites, nous allons considérer que la matrice est composée de vecteurs de présence/absence $X_i = (x_{i,j})_{1 \leqslant j \leqslant N}$ indépendants et identiquement distribués dont on peut modéliser la distribution par un mélange de distributions multivariées de Bernoulli (Bernoulli Mixture Model : BMM). Cette modélisation peut notamment être exprimée par le formalisme suivant proposé par GOVAERT, 1990 :

$$P(X_i = (x_{i,j})_{1 \leqslant j \leqslant N}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{N} \varepsilon_{k,j}^{|x_{i,j} - \mu_{k,j}|} (1 - \varepsilon_{k,j})^{1 - |x_{i,j} - \mu_{k,j}|}$$

où les paramètres :

- $\pi = (\pi_1, \ldots, \pi_k, \ldots, \pi_K)$ correspondent aux proportions de mélange avec $\pi_k \in [0,1]$ et $(\sum_{k=1}^{K} \pi_k) = 1$. Ainsi π_k nous fournit la proportion de familles de gènes appartenant à chaque partie k.
- $\mu_k = (\mu_{k,j})_{1 \leqslant j \leqslant N} \in \{0;1\}^N$ correspond au vecteur de présence/absence central le plus probable des familles de gènes appartenant à chaque partie k.
- $\varepsilon_k = (\varepsilon_{k,j})_{1 \leqslant j \leqslant N} \in [0, \frac{1}{2}]^N$ correspond au vecteur de dispersion autour du centre μ_k pour les familles de gènes appartenant à chaque partie k. Autrement dit ε_k quantifie la probabilité que la valeur dans une famille i pour un génome j diffère de la valeur centrale $\mu_{k,j}$.

Pour ce qui est des valeurs du vecteur de dispersion ε_k associé à chaque vecteur μ_k, l'algorithme admet deux modes de fonctionnement :

— Un mode où tous les ε_k sont contraints à être identiques pour tous les $\varepsilon_{k,j}$ d'une même partie k afin d'éviter un possible sur-ajustement sur les données. Le vecteur ε_k est donc ainsi réduit à une seule valeur par partie.

— Un mode où tous les ε_k sont libres de varier pour chaque génome j et pour chaque partie k afin de s'ajuster au plus près des données.

Il faut noter que ce paramétrage est atypique, car un modèle multivarié de Bernoulli pourrait être paramétré pour un simple vecteur de probabilité de succès p. Or ici, deux vecteurs sont nécessaires (μ_k et ε_k). La raison pour laquelle ce paramétrage a été proposé par GOVAERT, 1990 est qu'on suppose que, pour chaque composant du mélange, les N variables sont indépendantes et que chacune d'entre elles suit une des deux lois de Bernoulli suivantes :

$$\begin{cases} 1 \text{ avec la probabilité } 1-\varepsilon \text{ et } 0 \text{ avec la probabilité } \varepsilon \\ 1 \text{ avec la probabilité } \varepsilon \text{ et } 0 \text{ avec la probabilité } 1-\varepsilon \end{cases}$$

Ainsi, l'auteur de GOVAERT, 1990 montre que cela permet une simplification facilitant l'estimation des paramètres du modèle par la maximisation d'un critère de vraisemblance classifiante dont il a montré que sous certaines hypothèses celui-ci était convergent. En effet, pour estimer ces paramètres, nous utilisons un algorithme d'Espérance-Maximisation (Expectation-Maximisation : EM, DEMPSTER et al., 1977). Cet algorithme ajuste les paramètres du modèle par rapport aux données de façon à maximiser de façon itérative sa log-vraisemblance (et plus spécifiquement le critère de vraisemblance classifiante) ; ce que nous approfondirons à la sous-section 7.4.3 page 121. D'où l'on comprend donc l'importance d'utiliser ce paramétrage disposant d'un critère convergent pour estimer les meilleurs paramètres du modèle.

Pour débuter l'algorithme, comme nous ne connaissons pas les proportions respectives de chacune des k parties, alors tous les π_k sont initialisés avec des proportions équiprobables égales à $1/K$. Pour ce qui est des vecteurs μ_k et ε_k, pour accélérer l'algorithme EM, nous initialisons le modèle avec des paramètres correspondant à ce que l'on suppose observer systématiquement dans un pangénome. À savoir :

— une partie homogène avec beaucoup de 1, soit le génome persistant.

— une partie homogène avec beaucoup de 0, soit le génome nuage.

— un ou des parties moins homogènes contenant des mélanges de 1 et 0.

Sur la base de ces observations et sachant une variable $s = \frac{1}{\lceil 0.5K \rceil}$, les vecteurs sont donc initialisés ainsi [1] :

$$\begin{aligned} \{\mu_{k,j}\}_{1 \leqslant k \leqslant \frac{K}{2}, 1 \leqslant j \leqslant N} &= 1 \\ \{\mu_{k,j}\}_{\frac{K}{2} < k \leqslant K, 1 \leqslant j \leqslant N} &= 0 \\ \{\varepsilon_{k,j}\}_{1 \leqslant k \leqslant \frac{K}{2}, 1 \leqslant j \leqslant N} &= s \cdot k \\ \{\varepsilon_{k,j}\}_{\frac{K}{2} < k \leqslant K, 1 \leqslant j \leqslant N} &= s \cdot (K - k + 1) \end{aligned}$$

1. Il est à noter que $\lceil x \rceil$ indique l'arrondi au plus petit entier $\geqslant x$, $\lfloor x \rfloor$ indique l'arrondi au plus grand entier $\leqslant x$ et $\lfloor x \rceil$ indique l'arrondi à l'entier le plus proche de x.

Une autre conséquence intéressante de cette initialisation est que le génome persistant correspondra à la partie ayant $k = 1$ tandis que le génome nuage correspondra à la dernière partie $k = K$. La ou les parties du génome coquille correspondront ainsi aux parties ayant $k > 1$ et $k < K$. Ainsi, cette initialisation facilite aussi l'étiquetage des résultats du partitionnement, ce qui n'est pas toujours un problème trivial avec ce type d'algorithme.

7.4.2 Partitionnement spatial au travers du graphe de pangénome

À partir du graphe de pangénome G précédemment décrit à la sous-section 7.3 page 116, nous allons étendre le modèle de partitionnement jusqu'à présent décrit. Pour cela, nous allons nous appuyer sur un champ de Markov caché dont la structure est donnée par le graphe G. Ainsi, nous ajoutons un critère de régularité spatiale dans le partitionnement précédemment décrit. Dans ce modèle actualisé, chaque nœud i du graphe, c'est-à-dire chaque famille de gènes i, appartient à une partie k cachée. Ces parties cachées sont distribuées selon le champ de Markov caché de sorte que la vraisemblance du modèle soit meilleure si deux nœuds voisins dans le graphe appartiennent à la même partie.

À l'origine, l'algorithme NEM (Neighboring Expectation-Maximization) utilisant ce modèle – couplant un modèle de mélange multivarié de Bernoulli avec un champ de Markov caché – fut théorisé par les auteurs de AMBROISE et al., 1997 ; 1998 ; DANG et al., 1998 et implémenté dans le langage C par Mo Dang (comme évoqué dans sa thèse de doctorat DANG, 1998) afin de reconnaître des motifs dans une image. Dans l'approche PPanGGOLiN, nous réemployons cette implémentation en l'adaptant au traitement de données pangénomiques ainsi que cela est détaillé dans le chapitre 8 page 139. Ainsi, NEM va tendre à « lisser », à « étaler » le partitionnement en regroupant préférentiellement ensemble les familles voisines dans le graphe. Il faut noter que l'algorithme NEM prend en compte le poids des arêtes entre les différentes familles de gènes voisines.

Formellement, nous stockons dans une variable latente $\{Z_i\}_{1 \leqslant i \leqslant F}$, pouvant prendre les valeurs d'état $\{1, \ldots, K\}$, la partie du pangénome à laquelle chaque famille i appartient. Cette variable est maintenant distribuée d'après un champ de Markov caché suivant la distribution de Gibbs suivante :

$$\mathbb{P}(\{Z_i\}_{1 \leqslant i \leqslant F}) = W_\beta^{-1} \exp\Big(\underbrace{\sum_{i=1}^{F} \sum_{k=1}^{K} \pi_k 1_{z_i = k}}_{\text{terme d'expansion}} + \underbrace{\beta \frac{F}{\sum_{i \sim i'} w_{i \sim i'}} \sum_{i \sim i'} w_{i \sim i'} 1_{z_i = z_{i'}}}_{\text{terme de régularité spatiale}} \Big)$$

où 1_A est une notation pour indiquer que l'évènement A est réalisé.

Dans cette distribution, il y a deux termes :

— Un terme d'expansion pénalisant un partitionnement équiprobable. Par exemple pour F=100 avec deux parties a et b, si $\pi_a = 0.8$ et $\pi_b = 0.2$ alors au maximum le terme d'expansion peut être de $100 \times 0.8 \times 0.8 + 0.2 \times 0.2 = 68$. Dans une autre situation, si les deux parties ont un π_k égal à 0.5 alors $100 \times 0.5 \times 0.5 + 0.5 \times 0.5 = 50$. On constate bien la pénalisation d'un partitionnement équiprobable.

— Un terme de régularité spatiale dont la valeur est maximisée si des familles appartenant à la même partie sont voisines. Le paramètre β (le coefficient de régularité spatiale) en gouverne la force. Pour s'assurer que β ait la même force relativement au terme d'expansion même quand le nombre de génomes ajoutés au pangénome augmente, un terme de correction $\frac{F}{\sum_{i\sim i'}^{E} w_{i\sim i'}}$ est introduit. En effet, dans un pangénome le nombre de familles de gènes a tendance à croître plus vite que la somme totale des arêtes quand le nombre de génomes ajoutés augmente.

Enfin, tout cela est normalisé par une constante théorique non calculable qui correspond à :

$$W_\beta = \sum_{\{\widetilde{z}_i\}\in\{1\ldots K\}^F} \exp\Big(\sum_{i=1}^{F}\sum_{k=1}^{K} \pi_k 1_{\widetilde{z}_i=k} + \beta \frac{F}{\sum_{i\sim i'} w_{i\sim i'}} \sum_{i\sim i'} w_{i\sim i'} 1_{\widetilde{z}_i=\widetilde{z}_{i'}}\Big)$$

Ici, les vecteurs $(X_i)_{1\leqslant i\leqslant F}$ ne sont plus absolument indépendants comme dans la sous-section 7.4.1 page 117, ils ne le sont que conditionnellement à la variable $(Z_i)_{1\leqslant i\leqslant F}$ distribuée selon le champ de Markov caché. Ainsi la probabilité des vecteurs (X_i) peut désormais être modélisée ainsi :

$$\mathbb{P}(\{X_i\}_{1\leqslant i\leqslant F}|\{Z_i\}_{1\leqslant i\leqslant F}) = \prod_{i=1}^{F}\prod_{j=1}^{N} \varepsilon_{Z_i,j}^{|x_{i,j}-\mu_{Z_i,j}|}(1-\varepsilon_{Z_i,j})^{1-|x_{i,j}-\mu_{Z_i,j}|}.$$

Néanmoins cette formulation est théorique et ne permet pas d'estimer « directement », c'est-à-dire de façon analytique, les paramètres du modèle. Pour estimer ces paramètres, nous allons donc devoir maximiser la vraisemblance classifiante du modèle sachant les données via l'algorithme NEM que nous allons décrire plus en détail à la sous-section suivante.

7.4.3 Estimation des paramètres du modèle et partitionnement par NEM

Pour estimer les paramètres maximisant la vraisemblance classifiante dans un champ de Markov caché, la méthode NEM s'appuie sur une approximation en champ moyen de la distribution des variables latentes $Z_{i1\leqslant i\leqslant F}$ conditionnellement aux observations. En pratique cela prend la forme d'un partitionnement flou où $c_{i,k1\leqslant i\leqslant F,1\leqslant k\leqslant K}$ indiquant les probabilités d'appartenance de chaque famille i à chaque partie k du pangénome.

Tout d'abord, sachant les paramètres d'initialisation fixés à la sous-section 7.4.1 page 117, chaque famille i est associée à sa partie du pangénome la plus probable à l'itération $(q+1)$.

Étape E : affectation d'après des distances de Manhattan

La mise à jour de la classification floue $c_{i,k}^{(q+1)}$ à l'itération $(q+1)$ se fait en appliquant successivement pour toutes les familles $1\ldots i\ldots F$ et pour toutes les parties $1\ldots k\ldots K$:

$$c_{i,k}^{(q+1)} = \frac{exp\big(-d_{\varepsilon_k{}^{(q)}}(x_i,\mu_k^{(q)}) + \sum_{j=1}^{N} log(1-\varepsilon_{k,j}^{(q)}) + \beta\frac{F}{\sum_{i\sim i'} w_{i\sim i'}}\sum_{i\sim i'} w_{i\sim i'}\tilde{c}_{i',k}\big)}{\sum_{h=1}^{K} exp\big(-d_{\varepsilon_k{}^{(q)}}(x_i,\mu_h^{(q)}) + \sum_{j=1}^{N} log(1-\varepsilon_{h,j}^{(q)}) + \beta\frac{F}{\sum_{i\sim i'} w_{i\sim i'}}\sum_{i\sim i'} w_{i\sim i'}\tilde{c}_{i',h}\big)}$$

où $d_{\varepsilon_k{}^{(q)}}$ est une fonction de distance en norme L_1 (distance de Manhattan) pondérée par les poids suivants :

$$log\Big(\frac{1-\varepsilon_{k,1}}{\varepsilon_{k,1}}\Big),\ldots,log\Big(\frac{1-\varepsilon_{k,j}}{\varepsilon_{k,j}}\Big),\ldots,log\Big(\frac{1-\varepsilon_{k,N}}{\varepsilon_{k,N}}\Big)$$

et où $\tilde{c}_{i',k} = c_{i,k}^{(q+1)}$ si $i' < i$ ou bien $\tilde{c}_{i',k} = c_{i,k}^{(q)}$ si $i' > i$.

Cette formule implique ainsi que l'on affecte, d'une part, d'autant plus de familles i à une partie k que l'observation x_i est proche du centre $\mu_k^{(q)}$ de la partie k et d'autre part que la partie k est bien représentée parmi les familles i' voisines de i.

Étape M : centre binaire et dispersions

À l'étape M, on met à jour le paramétrage θ à l'itération $(q+1)$. Pour cela, on cherche les paramètres satisfaisant :

$$(\mu^{(q+1)},\varepsilon^{(q+1)}) = \underset{\mu,\varepsilon}{\arg\max} - \sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i=1}^{F} c_{i,k}^{(q+1)}\left(|(x_{i,j}-\mu_{k,j}|log\Big(\frac{1-\varepsilon_{k,j}}{\varepsilon_{k,j}}\Big) - log(1-\varepsilon_{k,j})\right)$$

Pour ce qui est de l'estimation de $\mu^{(q+1)}$, pour tous les génomes $1\ldots j\ldots N$ et pour toutes les parties $1\ldots k\ldots K$, nous avons :

$$\mu_{k,j}^{(q+1)} = \underset{\mu=\{0,1\}}{\arg\min} \sum_{i=1}^{F} c_{i,k}^{(q+1)}|x_{i,j}-\mu_{k,j}|$$

En posant $n_k^{(q+1)} = \sum_{i=1}^{F} c_{i,k}^{(q+1)}$, si $n_k^{(q+1)} \neq 0$, les auteurs de DANG et al., 1998 montrent que :

$$\begin{aligned}
\mu_{k,j}^{(q+1)} &= \begin{cases} 0 \text{ si } \frac{1}{n_k^{(q+1)}}\sum_{i=1}^{F} c_{i,k}^{(q+1)}x_{i,j} < \frac{1}{2} \\ 1 \text{ si } \frac{1}{n_k^{(q+1)}}\sum_{i=1}^{F} c_{i,k}^{(q+1)}x_{i,j} > \frac{1}{2} \\ 0 \text{ ou } 1 \text{ si } \frac{1}{n_k^{(q+1)}}\sum_{i=1}^{F} c_{i,k}^{(q+1)}x_{i,j} = \frac{1}{2} \end{cases} \\
&= \left\{ \left\lceil \sum_{i=1}^{F} c_{i,k}^{(q+1)}x_{i,j} \right\rfloor \right. \\
&= \left\{ \text{ médiane pondérée (observation, poids) : med}\{(x_{1,j},c_{1,k}^{(q+1)}),\ldots,(x_{i,j},c_{i,k}^{(q+1)}),\ldots,(x_{F,j},c_{F,k}^{(q+1)})\} \right.
\end{aligned}$$

Pour ce qui est de l'estimation de $\varepsilon^{(q+1)}$, pour tous les génomes $1\ldots j\ldots N$ et pour toutes les parties $1\ldots k\ldots K$, nous avons :

$$\varepsilon_{k,j}^{(q+1)} = \underset{\varepsilon\in[0,0.5]}{\arg\min} \underbrace{\sum_{i=1}^{F} c_{i,k}^{(q+1)}|x_{i,j}-\mu_{k,j}|log\Big(\frac{1-\varepsilon}{\varepsilon}\Big) - n_k^{(q+1)}log(1-\varepsilon)}_{g(\varepsilon)}$$

Si l'on annule la dérivée $g'(\varepsilon)$, cela nous donne :

$$\varepsilon_{k,j}^{(q+1)} = \frac{1}{n_k^{(q+1)}} \sum_{i=1}^{F} c_{i,k}^{(q+1)} |x_{i,j} - \mu_{k,j}^{(q+1)}|$$

Dans le cas où nous contraignons tous les ε_k à être égaux alors cela nous donne :

$$\varepsilon_{k}^{(q+1)} = \frac{1}{n_k^{(q+1)}} \sum_{i=1}^{F} c_{i,k}^{(q+1)} |(x_{i,j} - \mu_{k,j}^{(q+1)}|$$

Ainsi, nous obtenons l'ensemble des paramètres du BMM couplés aux résultats de partitionnement « flou », attribuant une probabilité postérieure $c_{i,k}$ à chaque famille de gènes d'être dans une partie k du pangénome. Les itérations E et M sont répétées jusqu'à ce que la classification floue $c_{i,k}$ des familles de gènes soit stabilisée. Néanmoins, pour partitionner effectivement le pangénome, nous allons durcir ce partitionnement flou, c'est-à-dire attribuer chaque famille à une partie k seulement. Nous stockons la partie à laquelle va appartenir chaque famille dans une variable $\{Z_i\}_{1 \leqslant i \leqslant F}$ pouvant prendre les valeurs d'état suivantes $\{1, \ldots, K\}$. Pour cela, une fois les paramètres du modèle optimisés par NEM, la méthode PPanGGOLiN attribue automatiquement les familles de gènes à leur partition la plus probable z_i si sa probabilité postérieure $c_{i,k}$ est supérieure à 0.5 et au génome coquille dans le cas contraire (mais ce dernier cas reste théorique car extrêmement rare en pratique).

Estimation du nombre de parties (K) dans un pangénome

Comme nous l'avons vu, il peut parfois être nécessaire d'admettre plus de parties que trois dans le cas où il existe des schémas antagonistes à l'intérieur du génome coquille.

Pour sélectionner le nombre de parties nécessaires pour partitionner au mieux le pangénome, nous utilisons une métrique presque équivalente au critère d'information bayésien (Bayesian Information Criterion : BIC, SCHWARZ, 1978) appelée le critère de vraisemblance complétée intégrée (Integrated Completed Likelihood : ICL, BIERNACKI et al., 2000).

Comme le BIC, l'ICL se calcule sur la base de la vraisemblance d'un modèle pénalisé par le nombre de paramètres utilisés par celui-ci. En effet, la vraisemblance augmente nécessairement quand on ajoute des parties, c'est-à-dire des paramètres au modèle. Dès lors, le BIC permet de trouver un équilibre parcimonieux entre la maximisation de la vraisemblance et le nombre de parties K. L'ICL est un critère très proche du BIC mais il amène un léger gain dans le cas d'un partitionnement flou en apportant une pénalité sur des résultats qui seraient trop flous. Dans notre cas, l'ICL et le BIC produisent cependant des résultats très similaires.

Pour estimer le nombre de parties dans le pangénome, la vraisemblance du modèle est calculée sans utiliser le champ de Markov caché (c'est-à-dire que $\beta = 0$) car celui-ci ajouterait des paramètres n'ayant pas d'impact sur le nombre de parties. Le critère ICL est calculé ainsi :

$$ICL(K) = BIC(K) - \sum_{k=1}^{K}\sum_{i=1}^{F} p(z_i \mid X, \hat{\theta}, k)\log(p(z_i \mid X, \hat{\theta}, k)) \,; \forall p(z_i \mid X, \hat{\theta}, k) > 0$$

et

$$BIC(K) = \log \mathbb{P}_K(X \mid \hat{\theta}) - \frac{1}{2} dim(K)\log(F)$$

où $\log \mathbb{P}_K(X \mid \hat{\theta})$ correspond à la log-vraisemblance avec $\beta = 0$ du modèle paramétré par $\hat{\theta} = (\{\hat{\pi}_k\}_{1 \leqslant k \leqslant K}, \{\hat{\mu}_{kj}\}_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant N}, \{\hat{\varepsilon}_{kj}\}_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant N})$. Cette log-vraisemblance est calculée ainsi :

$$\log \mathbb{P}_K(X \mid \hat{\theta}) = \sum_{i=1}^{F} \log \left(\sum_{k=1}^{K} \hat{\pi}_q \prod_{j=1}^{N} \hat{\varepsilon}_{kj}^{|x_{ij} - \hat{\mu}_{kj}|} (1 - \hat{\varepsilon}_{kj})^{1 - |x_{ij} - \hat{\mu}_{kj}|} \right)$$

$\hat{\theta}$ correspond ici aux paramètres maximisant la vraisemblance du modèle et $dim(K)$ indique le nombre de paramètres variant dans ce modèle, soit $dim(K) = K(N+2)$, correspondant à : N paramètres $\hat{\mu}_k$, K paramètres $\hat{\pi}_k$ et K paramètres $\hat{\varepsilon}_k$ si le vecteur $\hat{\varepsilon}_k$ est obtenu avec le mode le contraignant à avoir des valeurs identiques. Au contraire, si $\hat{\varepsilon}_k$ est obtenu avec le mode libre, alors $dim(K) = K(2N+1)$, correspondant à : N paramètres $\hat{\mu}_k$, K paramètres $\hat{\pi}_k$ et N paramètres $\hat{\varepsilon}_k$.

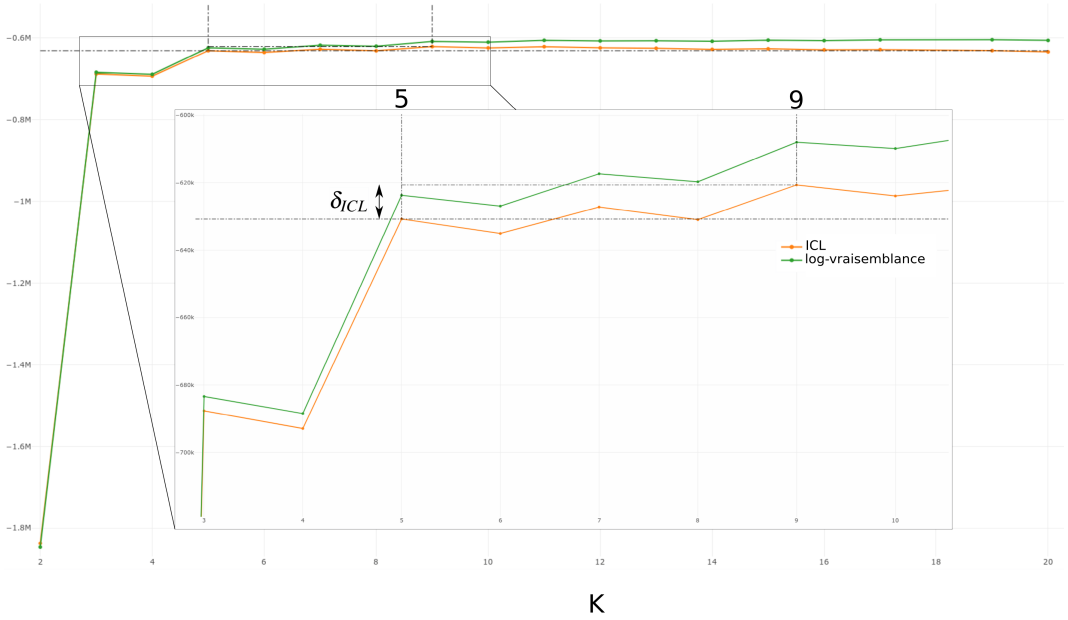


FIGURE 7.4 – Figure illustrant une courbe d'évolution de l'ICL et de la log-vraisemblance en fonction du nombre de parties. Le maximum de l'ICL est à $K = 9$ mais du fait de δ_{ICL}, le nombre de parties choisi est de $K = 5$ car les 4 parties supplémentaires apportent un gain d'ICL minime.

Pour obtenir l'ICL pour toutes les valeurs de K, nous effectuons des proto-partitionnements du pangénome avec un nombre croissant de parties K allant de 2 à 20 (par défaut). Ce que j'appelle ici un proto-partitionnement correspond à un partitionnement volontairement avorté au bout de 10 itérations NEM seulement, de sorte que les paramètres $\hat{\theta}$ ainsi obtenus ne maximisent pas réellement la vraisemblance du modèle. Cependant, à ce niveau, il n'est pas intéressant d'observer dans le détail les familles assignées à chaque partie mais simplement d'observer si, de façon grossière, le partitionnement est plus ou moins vraisemblable pour un nombre donné de parties. Il faut en effet rappeler que toutes les parties k sont initialisées selon la même méthode décrite dans la sous-section 7.4.1 page 117, le seul paramètre variant étant le nombre de parties K. Cette optimisation semble pertinente car on observe systématiquement que les paramètres du modèle varient de façon importante pendant les $\approx$2 premières étapes de l'algorithme EM puis se stabilisent en s'ajustant finement aux données. Néanmoins, cela nous impose de la prudence de sorte qu'une marge est ajoutée, comme nous allons le voir, pour inférer le nombre de parties. Ainsi, le meilleur nombre de parties $\hat{K}$ est donné par le calcul $\hat{K} = \arg\min_{K}((1-\delta_{ICL})ICL(K))$ où δ_{ICL} est une marge évitant d'avoir une valeur de K trop élevée qui serait due à un gain marginal, donc sans apport substantiel par rapport à un K plus petit (comme illustré sur la figure 7.4 page ci-contre). Par défaut, $\delta_{ICL} = 0.05 \times (max(ICL) - min(ICL)))$.

Pour éviter toute confusion, il est à préciser qu'une fois le meilleur nombre de parties K identifié, les paramètres $\hat{\theta}$ sont recalculés avec un partitionnement cette fois-ci normal (et non un proto-partitionnement) et prenant bien entendu en compte le champ de Markov caché.

Dans le cas d'un partitionnement par morceau comme nous allons le voir dans la sous-section suivante, l'estimation du nombre idéal de parties est effectuée sur un morceau pris au hasard. On considère en effet qu'un morceau est censé être représentatif du pangénome global. Mais voyons tout d'abord ce qu'est un partitionnement par morceaux.

7.4.4 Partitionnement par morceaux

Lorsque l'on effectue des tests avec la méthode de partitionnement NEM, on observe que si le nombre de génomes devient trop grand (généralement $N > 800$), l'algorithme se met à attribuer des équiprobabilités d'appartenance à chaque partie. Dès lors, le partitionnement ainsi obtenu ne présente plus d'intérêt.

Après de nombreux tests, analyses, tentatives de débogage et discussions avec les statisticiens Catherine Matias (directrice de recherche CNRS au Laboratoire de Probabilités, Statistique et Modélisation) et Christophe Ambroise (un des concepteurs de la méthode NEM), nous sommes arrivés à la conclusion que ce comportement de l'algorithme NEM serait probablement dû à un problème de saturation de la méthode aux grandes dimensions, c'est-à-dire quand la condition $F >> N$ n'est plus satisfaite.

Dans le cas de la pangénomique, comme nous l'avons vu à travers l'étude des courbes de raréfaction du pangénome, le taux de nouvelles familles découvertes décroît avec l'ajout de nouveaux génomes. Ainsi, F croît moins vite que N ce qui expliquerait notre problème. Des solutions mathématiques semblent exister à ce problème, par exemple, en pondérant les génomes (BOUGUILA,

2010) ou en réduisant la dimensionnalité de la matrice (YAMAMOTO et al., 2015). Néanmoins, ces méthodes contrairement à NEM ne prennent pas en compte le voisinage des familles de gènes dans le pangénome.

Par ailleurs, l'autre inconvénient de la méthode NEM lorsque N et F augmentent est qu'en toute hypothèse, faire croître la matrice infiniment risquerait nécessairement de poser des problèmes de ressources mémoires pour le calcul. Une solution à ce problème de mise à l'échelle pourrait venir de l'utilisation de matrices éparses. Cependant le modèle statistique NEM n'est pas en l'état adapté à ce type de données. Là encore, ce type de méthode existe comme celle proposée par les auteurs de ŚMIEJA et al., 2019. Toutefois cette implantation semble dysfonctionnelle en l'état car d'après mes tests, le partitionnement qu'elle propose est purement aléatoire [1]. Par ailleurs, cette méthode ne prend non plus en compte le voisinage des familles de gènes dans le pangénome.

Comme solution à ce problème, nous proposons dans la méthode PPanGGOLiN d'effectuer un partitionnement par morceau, c'est-à-dire de diviser l'ensemble des génomes en plusieurs sous-ensembles aléatoires de génomes appelés morceaux, puis de les partitionner chacun isolément, avant d'en rassembler les résultats. Ainsi de nombreuses familles de gènes se retrouvent partitionnées plusieurs fois dans des ensembles de génomes différents. Si la famille est toujours attribuée à la même partie (génome persistant, génome coquille ou génome nuage), cela ne pose pas de problème. Néanmoins, pour les familles à la frontière entre deux parties, il est possible qu'elles ballottent d'une partie à l'autre, et notamment du génome nuage au génome coquille. Dès lors, pour déterminer la partie à laquelle appartient une famille, on compte le nombre total d'assignations à chaque partie. Si une partie majoritaire se dégage (>50%), la famille est associée à cette partie. Sinon, pour les quelques cas (très rares) où une majorité n'est pas atteinte, de nouveaux cycles de partitionnement sont effectués jusqu'à ce qu'une majorité se dégage. Le nombre minimum de partitionnements dans laquelle une famille doit être impliquée est au moins égal à $\frac{N}{N_m}$ où N_m correspond à la taille des morceaux. Par défaut, nous utilisons une taille de morceau de 500 pour s'assurer que chaque morceau ait un échantillonnage suffisamment représentatif.

L'autre gros avantage du partitionnement par morceau est de permettre la parallélisation du partitionnement de gros pangénomes. Pour donner un ordre de grandeur, aujourd'hui les analyses pangénomiques dépassent fréquemment le millier de génomes. Par exemple, les auteurs de LAING et al., 2017 utilisent 4893 génomes de *Salmonella enterica* dans une même analyse. Dans les résultats présentés dans le chapitre 9 page 151, on constate que le plus gros pangénome est basé sur 20 656 génomes de *Listeria monocytogenes*. Des pangénomes à plusieurs centaines de milliers de génomes d'une même espèce sont d'ores et déjà envisageables. Dans nos tests de mise à l'échelle de la méthode PPanGGOLiN, nous avons réussi à construire le graphe de pangénome de 50 000 génomes de *Salmonella enterica* et à le partitionner en utilisant des morceaux de 500 génomes (voir la section 8.3 page 148 pour plus de détails sur les performances de l'outil PPanGGOLiN). On imagine alors qu'à un moment donné, même sans le problème de saturation aux grandes dimensions, autoriser la parallélisation des calculs aurait de toute façon été nécessaire. C'est la raison pour laquelle cette solution de partitionnement par morceaux fut retenue et mise en œuvre.

1. Cela a été signalé par mes soins à 2 reprises aux auteurs, sans retour de leur part.

7.4.5 Estimation du coefficient de régularité spatiale (β)

Le coefficient de régularité spatiale (β) pondère la force du lissage du partitionnement à l'aide du graphe de pangénome. Ainsi, dans le cas où $\beta = 0$, le partitionnement ne s'appuie donc que sur la matrice de présence/absence. Pour s'assurer que le lissage ne soit pas trop fort, ni trop faible, il serait souhaitable d'estimer la valeur de β idéale. La question qui se pose alors est comment estimer β ? Répondre à cette question est loin d'être une sinécure et autant le dire tout de suite, la tentative d'y répondre m'a demandé beaucoup de temps, de doutes, de discussions et d'énergie au cours de cette thèse. Je vais donc ici présenter brièvement la démarche suivie et une partie des résultats obtenus mais sans hélas pouvoir y apporter de réponse absolue.

Prédire le coefficient β améliorant le partitionnement des génomes incomplets

L'hypothèse de base que nous avions était que le lissage spatial aurait permis de mieux partitionner le graphe de pangénome dans le cas où celui-ci serait obtenu à partir de génomes de qualité réduite, c'est-à-dire susceptibles d'avoir des gènes manquants dus à des assemblages très fragmentés. Pour vérifier cela nous avons rassemblé un jeu de données constitué de deux ensembles de génomes d'*E. coli* contenant une diversité de souches comparable. Le premier ensemble est constitué de 247 génomes complets (C), c'est-à-dire de génomes parfaitement assemblés et le second de 247 génomes incomplets (I), c'est-à-dire ayant une qualité d'assemblage variable correspondant au cas général des génomes utilisés dans des études pangénomiques. Ce jeu de données a été constitué de sorte qu'il y ait une bijection entre les ensembles C et I. Chaque génome C_j a donc un équivalent dans I_j basé sur une distance génomique MASH entre chaque I_j et C_j inférieure à 0.002 comme illustré sur la figure 7.5.

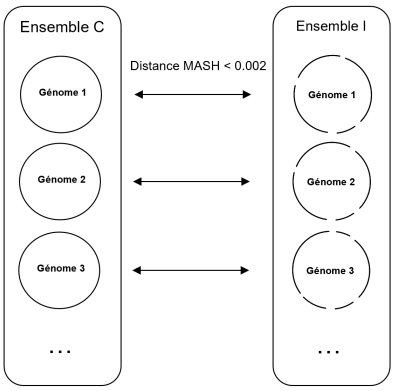


FIGURE 7.5 – Illustration de la bijection entre les ensembles I de génomes incomplets et C de génomes complets. On rappelle que pour considérer que deux génomes soient équivalents entre les deux ensembles donc avec des contenus génomiques extrêmement proches, leur distance MASH doit être inférieure à 0.002. Pour rappel, une distance MASH $\leqslant$0.06 regroupe des génomes de la même espèce.

Les familles de gènes de ces deux ensembles de génomes sont constitués conjointement afin de pouvoir effectuer des comparaisons directes. Les deux ensembles sont ensuite partitionnés séparément ($K = 3$) avec des gradients de β pour l'ensemble I et un $\beta = 0$ pour l'ensemble C. En effet, l'ensemble C étant constitué de génomes complets, nous pouvons considérer que ces génomes

correspondent à des sortes de références ne nécessitant donc pas de recourir au lissage spatial, d'où le choix de $\beta = 0$. À la suite de cela, on observe les différences de partitionnement entre les deux ensembles notamment au niveau du génome persistant. Pour cela, nous allons comparer les familles assignées au génome persistant de l'ensemble *I* avec celles de l'ensemble *C*. Ces résultats sont indiqués dans la table 7.1.

β_I	Nombre de familles communes à l'ensemble *I* et *C*	Nombre de familles spécifiques à l'ensemble *C*	Nombre de familles spécifiques à l'ensemble *I*	Nombre total de familles spécifiques
0	3 576	41	3	44
1	3 574	43	3	46
2	3 573	44	5	49
3	3 579	38	12	50
4	3 580	37	17	54
5	3 580	37	18	55
6	3 576	41	19	60
7	3 583	34	26	60
8	3 588	29	42	71
9	3 596	21	66	87
10	3 600	17	73	90

TABLE 7.1 – Évolution du nombre de familles de gènes persistants en commun et spécifiques entre l'ensemble de génomes *C* partitionné avec $\beta = 0$ et l'ensemble de génomes *I* partitionné avec un gradient de β (appelé β_I).

Sur la base de ce tableau, il est difficile de voir un impact positif du terme de lissage spatial β notamment du fait des faibles effectifs de familles spécifiques par rapport aux familles communes. On note tout de même que si le nombre de familles de gènes persistants en commun entre les deux ensembles est croissant quand β augmente c'est parce que le nombre de familles de gènes persistants spécifiques à l'ensemble complet *C* a tendance à diminuer (de 41 à 17 familles). Cependant, le nombre de familles de gènes spécifiques à l'ensemble *I* augmente lui aussi dans la même cinétique, et ce de façon plus importante encore (de 3 à 73 familles). On en déduit que c'est le génome persistant en général de l'ensemble *I* qui augmente quand on augmente β sans que ce soit prioritairement les familles de gènes persistants spécifiques de l'ensemble *C* qui soient rapatriés dans l'ensemble *I* quand β augmente. Dès lors, le choix d'augmenter β permet certes de repartitionner correctement certaines des familles persistantes présentes dans les références associées mais également et même davantage, d'ajouter des familles de gènes non persistants dans celles-ci, donc du bruit en quelque sorte.

Par souci de concision, je ne détaille pas ici les très nombreux tests du même acabit que j'ai effectués. Par exemple, en effectuant des mélanges de *I* et de *C* avec des proportions variables ou bien en faisant varier le β pour l'ensemble *C* également. Globalement, à la suite de ces expériences, cela penche dans le sens d'un rejet de notre hypothèse. Il n'y aurait pas de lissage du partitionnement susceptible d'améliorer celui-ci pour les génomes incomplets du fait du bruit que génère ce lissage.

Choix du coefficient β stabilisant la prédiction du génome persistant

Pour déterminer la valeur de β à privilégier nous avons donc opté pour une autre stratégie. Nous avons tout d'abord constitué un échantillon contenant 48 espèces ayant plus de 100 génomes et moins de 500 génomes (pour des raisons de temps de calcul) obtenus depuis la banque de données GenBank. Nous avons par la suite effectué un partitionnement de ces 48 espèces avec un gradient de la valeur β allant de 0 à 10.

Comme on peut le voir sur la figure 7.6, le génome persistant varie de façon plus importante avec les petites augmentations de β (au début du gradient) et se stabilise aux environs de $\beta = 2.5$. Ainsi on peut penser que 2.5 est la valeur où la topologie du graphe ne provoque plus de changement majeur dans le partitionnement du génome persistant, c'est donc la valeur par défaut que nous avons considérée dans l'outil PPanGGOLiN. Le lissage topologique doit ainsi être vu non plus comme un moyen d'améliorer le partitionnement des génomes incomplets mais comme la prise en compte de l'empreinte topologique des génomes pour la stabilisation des résultats de partitionnement.

Ceci dit cette topologie peut elle-même souffrir de limites dans sa capacité à représenter le pangénome comme nous allons le voir maintenant.

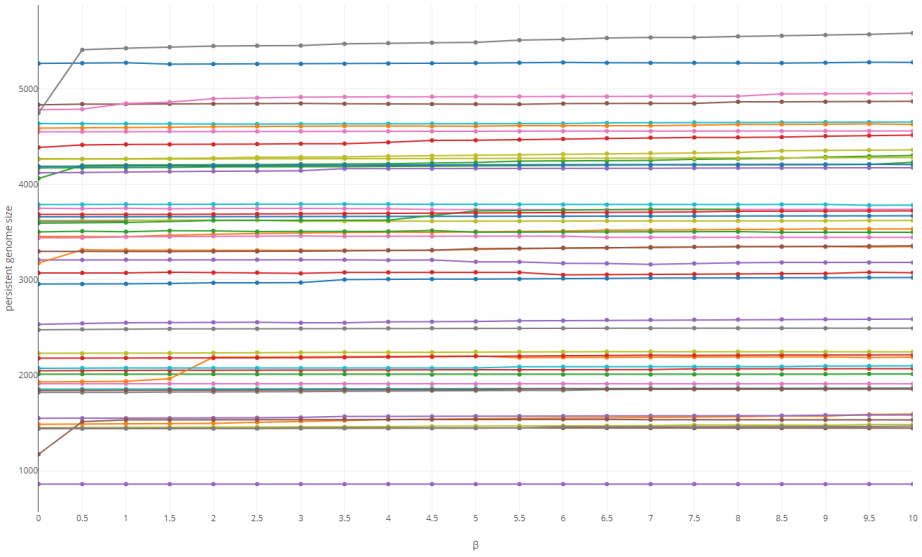


FIGURE 7.6 – Illustration de l'évolution du nombre de familles dans le génome persistant de 48 espèces en fonction d'un gradient du coefficient de régularité spatiale β.

7.5 L'emmêlement du graphe de pangénome

7.5.1 Cas des éléments transposables

Les éléments transposables sont des séquences capables de se déplacer de façon autonome car elles possèdent un gène encodant une enzyme appelée transposase capable de faire se mouvoir la séquence dont elle est issue dans les génomes (les transposons). Il existe de nombreuses transposases et structures de transposons différents. Du fait qu'elles sont mobiles, ces séquences sont susceptibles

de se situer à des positions très variables sur les génomes. De même celles-ci sont généralement en nombreuses copies car elles sont capables, soit de se multiplier elles-mêmes, soit de se sur-répliquer au moment de la réplication du génome (saut d'un élément mobile pendant la réplication).

La conséquence directe de la présence de ces gènes dans les graphes de pangénome est de mener à l'apparition de nœuds hyperconnectés aussi appelés « hubs ». Dans le cas de la méthode PPanGGOLiN, cela peut notamment engendrer les deux types de problèmes que nous avons abordés.

Les problèmes de perte d'information visuelle

La topologie du graphe est perturbée par ce type de nœuds. Par exemple, si une transposase s'insère systématiquement entre deux familles de gènes, nous pouvons perdre le lien visuel entre les deux familles. En effet, même si les deux familles restent bien connectées par le truchement de la transposase dans le graphe, celle-ci est connectée à tellement d'emplacements dans le génome qu'il devient impossible de repérer visuellement que les deux familles sont voisines. Il faut cependant noter que visualiser la connexité d'une transposase dans un graphe de pangénome peut aussi être intéressant en soit. Enfin, les arêtes restent annotées par les génomes qui les supportent ce qui peut permettre de se repérer par la mise en évidence d'un chemin particulier entre les deux familles passant par la transposase. Pour cette raison, par défaut dans la méthode PPanGGOLiN, les familles de gènes codant des transposases sont laissées dans le graphe. Cependant une option est proposée pour les retirer. Mais la question qui se pose d'abord est : comment les détecter ? Trois solutions viennent ici spontanément à l'esprit :

— Construire le graphe puis lister les nœuds (familles de gènes) ayant un degré supérieur à un seuil.
— Compter le nombre de gènes appartenant à une même famille dans les génomes, puis au-delà d'un seuil, les considérer comme transposases. Les transposases ayant généralement un nombre de copies important dans les génomes.
— Aligner les gènes sur des bases de données de séquences de transposases connues comme, par exemple, ISFinder (SIGUIER et al., 2006) ou ITIS (JIANG et al., 2015).

Après les avoir détectées, il est possible de construire (ou de reconstruire pour la première solution) le graphe en sautant les gènes codant des transposases à chaque ajout de nœuds au graphe. Par souci de simplicité, c'est la deuxième solution qui a été choisie. Ceci étant dit il reste loisible à l'utilisateur d'utiliser deux fois PPanGGOLiN dans le cadre de la première option ou bien de pré-filtrer les gènes fournis à PPanGGOLiN pour la troisième option.

Les problèmes de partitionnement provoqués par les nœuds hyperconnectés

Au-delà des difficultés visuelles évoquées, les transposases peuvent également provoquer des problèmes de partitionnement du fait de la méthode statistique employée. En effet, nous l'avons dit, le partitionnement effectué sur la base de la matrice est lissé par la topologie du graphe de pangénome. Or, si dans ce graphe il existe des nœuds hyperconnectés, la partie à laquelle chaque famille de gènes sera assignée aura une grande importance. On imagine, en effet, que du fait de la contrainte de régularité spatiale, un nœud hyperconnecté aura une grande capacité « d'influence » sur les familles voisines. Certes ces liens sont pondérés de sorte que chacun d'eux provoque des

influences subtiles mais en se cumulant, cela peut faire basculer le partitionnement pour des raisons indésirables. Pour contrer cela, les familles de gènes ayant un degré supérieur à un seuil (10 par défaut) sont considérées comme des nœuds isolés pour la méthode statistique. Attention, je parle ici du partitionnement statistique uniquement, cela n'empêche en rien de les visualiser avec leur topologie réelle dans le graphe de pangénome par la suite.

7.5.2 Cas des familles multigéniques

Pour comparer les génomes entre eux, nous l'avons dit, il nous faut des points de repère entre ceux-ci. Ces points de repères sont les gènes regroupés en familles de gènes homologues afin de les associer entre eux sur la base de leur similarité. Pour des raisons d'optimisation, la méthode MMseqs2 que nous utilisons ne cherche pas particulièrement à produire des familles de gènes orthologues, seulement des homologues. Elle peut donc regrouper dans la même famille des gènes issus d'une duplication. Comme nous l'avons vu dans la sous-section 5.1.5 page 84, dédiée à panOCT, une manière de détecter les orthologues est de s'intéresser aux positions des gènes dans les génomes. Les orthologues entre deux génomes étant généralement environnés des mêmes gènes, on parle alors d'orthologues positionnels.

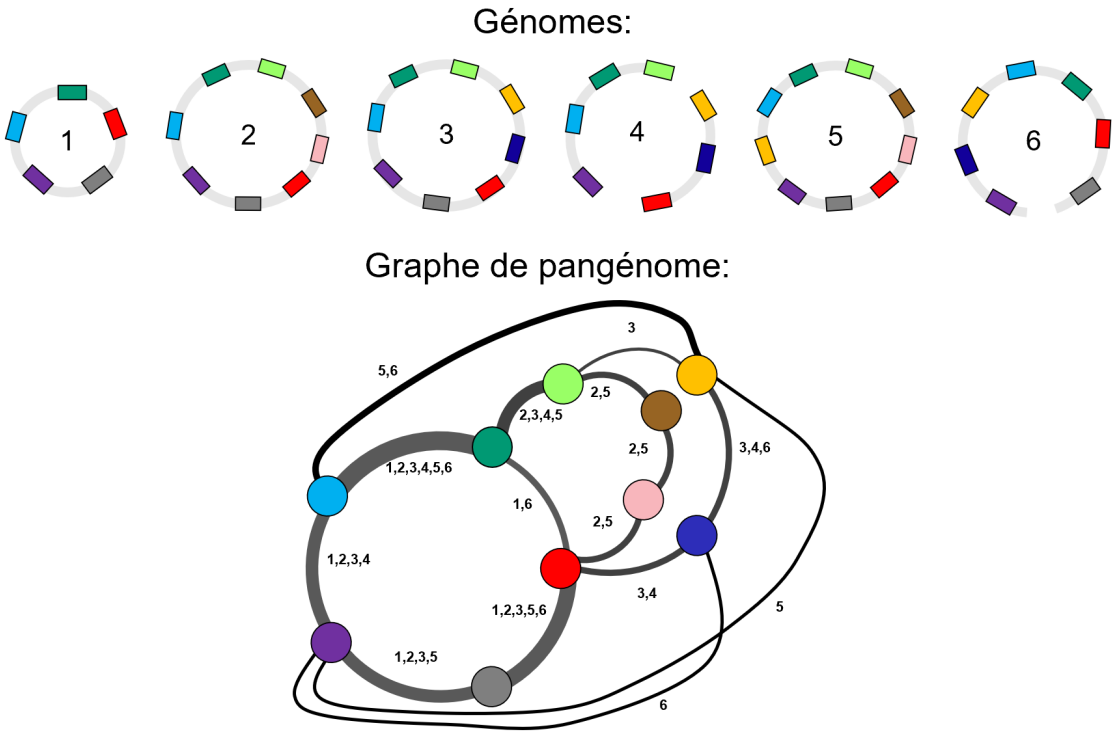


FIGURE 7.7 – Ajout de deux génomes aux génomes de la figure 7.1 page 112 afin de générer des cas d'emmêlements

Une solution pour identifier ces derniers peut alors être de démêler le graphe de pangénome. En effet, en cas de regroupement de paralogues amenant à des familles multigéniques, il y aura des emmêlements dans le graphe comme illustré sur la figure 7.7. Dans cette figure, nous ajoutons deux génomes aux génomes de la figure 7.1 page 112 pour l'exemple. On peut notamment constater que la famille de gènes ● provient de gènes localisés dans différents contextes génomiques. Cela

provoque donc ce que nous appellerons un emmêlement du graphe au niveau de ce nœud. De plus, il est aussi possible d'observer des blocs de gènes localisés dans plusieurs contextes génomiques différents. Dans la figure 7.7 page précédente, c'est le cas du chemin constitué des familles de gènes qui peut être considéré comme présent dans deux contextes génomiques. Comment démêler cette situation ?

Pour cela, après des recherches infructueuses dans la littérature [1], je propose l'ébauche d'un algorithme de démêlement [2]. L'algorithme que je propose est capable de démêler des chemins pangénomiques c, et ce même enchevêtrés avec plusieurs niveaux d'emmêlement, si ceux-ci sont résolubles. Précisons tout d'abord ce qu'est un chemin pangénomique :

Définition 7.5.1 Un chemin pangénomique c est ici une succession de familles de gènes, c'est-à-dire de nœuds dans le graphe de pangénomes G dont la séquence correspond à la manière dont sont ordonnés les gènes appartenant à ces familles dans au moins 1 génome.

Le cœur de l'algorithme est basé sur l'identification d'emmêlements résolubles qui se définissent ainsi :

Définition 7.5.2 Un emmêlement résoluble existe sur un chemin pangénomique c_l de longueur $|c_l| = l$ s'il existe un sous-chemin commun entièrement parcouru par plusieurs chemins pangénomiques de longueur $\geqslant (l+1)$ dont les nœuds aux extrémités de part et d'autre de c_l exclu constituent des sous-ensembles qui après fermeture transitive de ceux-ci admettent plusieurs sous-ensembles de cardinalité $\geqslant 2$. Le nombre de ces sous-ensembles donne le nombre de divisions du chemin pangénomique c à effectuer pour démêler la situation.

Dans l'exemple de la figure 7.7 page précédente, il n'y a pas de chemin pangénomique c_1 de longueur 1 (c'est-à-dire des nœuds) emmêlé et résoluble selon la définition 7.5.2. En effet, si nous listons l'ensemble des extrémités pour tous les chemins pangénomiques de longueur $\geqslant 2$ parcourant chaque nœud, puis que nous effectuons la clôture transitive de ces sous-ensembles, on obtient un seul sous-ensemble de cardinalité $\geqslant 2$. Il faut noter que pour le cas de la famille elle est parcourue par un chemin pangénomique de longueur 2 dont l'unique extrémité est disjointe du reste des couples d'extrémités. Ce cas n'est pas considéré comme résoluble et est ignoré, car si nous acceptions de considérer les sous-ensembles de taille 1 comme des emmêlements résolubles, nous obtiendrions après division des chemins reliés par une seule extrémité, ce que nous ne désirons pas.

Dès lors, si l'on veut démêler le graphe, il faut en venir au démêlement de chemins pangénomiques plus longs, soit des chemins c_2 de longueur 2. Pour cela, il suffit de considérer les chemins préalablement listés comme des graines et de les étendre (si des génomes supportent cette extension). Déroulons l'algorithme comme précédemment en recherchant l'ensemble des sous-ensembles d'extrémités possibles pour tous les chemins pangénomiques de longueur $\geqslant 4$ parcourant chaque chemin pangénomique de longueur 2. Après clôture transitive de ces sous-ensembles, nous observons qu'un chemin pangénomique c_2 ( ) possède deux sous-ensembles de cardinalité $\geqslant 2$

1. Notamment à l'aide des mots-clés « graph untangling », « multigraph untangling ». Cela m'a amené à constater que les moteurs de recherche sont très mal adaptés à la recherche d'algorithmes.
2. Attention, pour qu'il n'y ait pas d'ambiguïté, je précise que je ne propose pas ici un algorithme rigoureusement formalisé et publiable, juste une piste de recherche qui nécessiterait un travail supplémentaire.

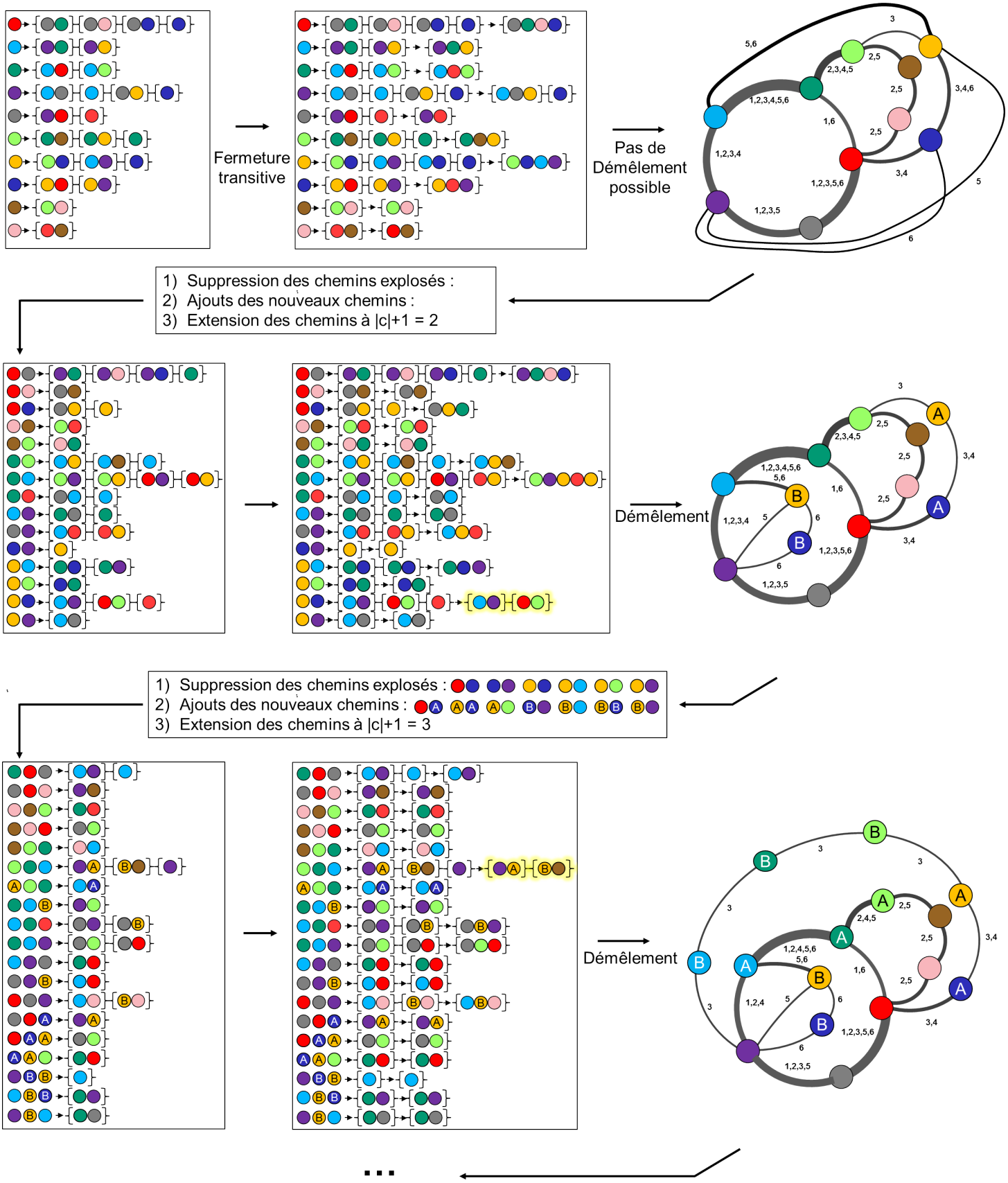


FIGURE 7.8 – Illustration de l'algorithme de démêlement appliqué (jusqu'à $c = 3$) au pangénome graphe emmêlé de la figure 7.7 page 131.

(illuminé en jaune dans la figure 7.8 page précédente). Dans ce cas, ce chemin pangénome va voir les nœuds qui le composent divisés en 2, puis les nouveaux nœuds sont reconnectés à chacun des sous-ensembles d'extrémités.

On peut ainsi recommencer l'algorithme pour les chemins pangénomiques c_3 de longueur 3. On constate alors qu'il est possible de démêler le graphe en séparant le chemin  (illuminé en jaune dans la figure 7.8 page précédente).

Nous constatons qu'en appliquant cet algorithme il est ainsi possible de démêler en particulier les paralogues classifiés dans une même famille ou des segments de plusieurs gènes répétés en tandem dans les génomes. Néanmoins, certains cas restent difficiles à démêler sans risquer d'isoler des familles, de générer des nœuds de degré 1 (donc connectés d'un seul côté seulement) ou de provoquer des démêlements arbitraires. Pour aller plus loin, il pourrait être intéressant de diviser les familles multigéniques en sous-familles suivant les conservations d'identité en séquence afin d'arbitrer les situations non décidables sur la base de la topologie du graphe de pangénome uniquement.

7.6 L'écaillage du graphe de PPanGGOLiN

Un choix a été fait, celui de générer des familles de gènes constituées de gènes très similaires. Ceci, à la fois en termes de séquence (80% d'identité en acides aminés) mais également du fait d'avoir imposé une contrainte d'alignement sur une large portion (80% de couverture) dans les deux sens. Or, il est courant lors de la détection des gènes, d'identifier des pseudogènes correspondant à des séquences protéiques tronquées ou fragmentées. En conséquence, ces gènes présentent vraisemblablement une fonction inactivée, diminuée ou modifiée.

Le type de familles qui en résulte est essentiellement des familles singletons, c'est-à-dire des familles constituées d'un seul gène. Dès lors, en tant que familles rares, elles seront donc assignées au génome coquille. Ainsi, on observe que ces familles se mettent à flotter autour de familles persistantes auxquelles elles auraient pu appartenir dans le graphe de pangénome, ce qui ressemble aux écailles d'un pangolin.

Comme solution, nous aurions pu baisser le taux de couverture afin de pallier à ce problème, néanmoins nous avons préféré garder ce seuil de couverture bidirectionnelle assez haut comme une garantie que les familles de gènes soient bien homologues. Ainsi, l'écaillage du pangénome est facultatif et peut s'effectuer par une méthode *ad hoc*.

Pour cela, nous pouvons réaligner avec MMseqs2 les séquences centrales de chaque famille de gènes sur toutes les autres séquences centrales. Pour filtrer ces alignements, nous pouvons utiliser le même taux d'identité pour la construction des familles (80%) mais cependant sans imposer ici de taux de couverture de la séquence cible mais seulement de la séquence requête (80%) afin de permettre l'alignement des fragments. Il est ainsi possible d'identifier de possibles fragments de gènes et de les réassocier à leur potentielle famille de gènes originelles. Ce travail reste cependant à améliorer, en effet les pseudogènes proviennent très souvent d'une mutation ayant entraîné un

décalage de phase. Dès lors la séquence en acides aminés est très différente et ne permet pas un réalignement sur la séquence centrale. Une solution à cela serait par exemple de réutiliser la séquence nucléotidique des séquences centrales pour effectuer ces alignements.

7.7 L'échantillonnage des génomes

Une grande limite de tout le travail présenté au cours de cette thèse et même potentiellement sa limite principale est que l'on considère uniquement ce qui est présent dans les bases de données. Or, rien ne permet de penser que ce qui est présent dans les banques est représentatif de la « réalité ». On pourrait même penser l'inverse en considérant les biais de focalisation et le mimétisme dans les sciences. Généralement, les chercheurs apprécient de travailler sur quelques organismes modèles bien connus. De même, on pourrait évoquer les difficultés à cultiver certaines espèces ou certaines souches, ce qui limite leur représentation dans les banques. De plus, il faudrait préciser, le sens que l'on donne au mot « réalité ». De quelle réalité parle-t-on ? Peut-on accéder au réel ? La réalité est-elle une construction humaine ? Ces questions sont d'une grande profondeur épistémologique mais je propose au lecteur de les laisser de côté pour la suite de cette thèse.

Pour le bioinformaticien, le problème consiste dans l'échantillonnage des points de mesure de cette réalité. Ce problème pourrait cependant être laissé en suspens en déchargeant la responsabilité sur les utilisateurs de la méthode PPanGGOLiN, c'est-à-dire, en leur laissant le soin de correctement sélectionner les génomes pertinents pour leur analyse pangénomique en fonction de leur problématique. Cet article des auteurs de Brockhurst et al., 2019 donne notamment quelques conseils utiles pour mener à bien cet échantillonnage. Il est toutefois possible de faire certaines opérations, comme le fait de pondérer les génomes dans le partitionnement effectué par PPanGGOLiN en fonction de la diversité qu'ils représentent, afin de faciliter la tâche des biologistes. J'ai donc réfléchi à la question et je vais rendre compte ici de mon raisonnement sur le sujet. Encore une fois, je livre là une piste de recherche que j'ai suivi au tout début de ma thèse et non une méthode aboutie.

Pour pondérer les génomes, il faut d'abord commencer par leur attribuer des poids en fonction de la diversité qu'ils représentent. Pour cela je propose tout d'abord de représenter les génomes sous forme de points disposés sur un espace. Ainsi, j'ai calculé une matrice de distance entre tous les génomes à l'aide de l'outil MASH (Ondov et al., 2016). En me basant sur cette matrice, j'ai effectué une mise à l'échelle multidimensionnelle (MultiDimensional Scaling : MDS) pour obtenir des coordonnées de points respectant cette matrice de distance en dimension 3, donc avec une perte d'information (perte appelée le stress de Kruskal).

Une fois ces points disposés dans l'espace, je souhaitais considérer le volume où il n'y a pas d'autre point plus proche autour d'eux, comme un poids. Pour cela, je vais calculer le volume de ce que l'on appelle des cellules de Voronoï. Pour information, un diagramme de Voronoï est obtenu en pavant un espace avec des cellules (au sens géométrique, c'est-à-dire des polygones) correspondant à la partition de l'espace en zones délimitant le sous-espace le plus proche de chaque point. La principale méthode pour obtenir ces cellules est basée sur des triangulations de Delaunay. Pour cela,

nous traçons des triangles entre tous les points avec pour seule règle que le cercle circonscrit de ces triangles ne doit contenir aucun autre point que les trois points du triangle considéré. Par la suite, les centres des cercles circonscrits sont reliés entre eux afin d'obtenir les cellules.

L'idée est donc de calculer le volume de chaque cellule afin d'en faire un poids que nous pourrions nommer « son volume de représentation ». Ce volume correspond à l'ensemble des génomes absents que le génome associé à cette cellule est susceptible de représenter. Cela suppose donc d'accepter l'hypothèse que l'espace des différences entre les génomes est continu. Ainsi, une cellule avec beaucoup de voisins proches sera donc petite et à l'inverse une cellule avec des voisins lointains sera plus grande. Pour chaque cellule, il est donc nécessaire de calculer le volume de ce polyèdre convexe. Cela ne pose pas de difficulté car il suffit de paver la cellule avec des pyramides et de calculer la somme des volumes de ces pyramides.

Cependant à ce stade du raisonnement un problème advient. En effet, toutes les cellules d'un diagramme de Voronoï ne sont pas finies. Or il est difficile d'accepter qu'un organisme ait un poids infini qui lui soit attribué. Une solution triviale à cet écueil est de borner artificiellement les cellules infinies mais cela est arbitraire et entraîne un traitement différencié entre les cellules. Une autre solution plus élégante est de faire un diagramme de Voronoï sphérique 7.9. J'ai effectué quelques essais en ce sens et cela semble possible avec quelques nouvelles difficultés cependant.

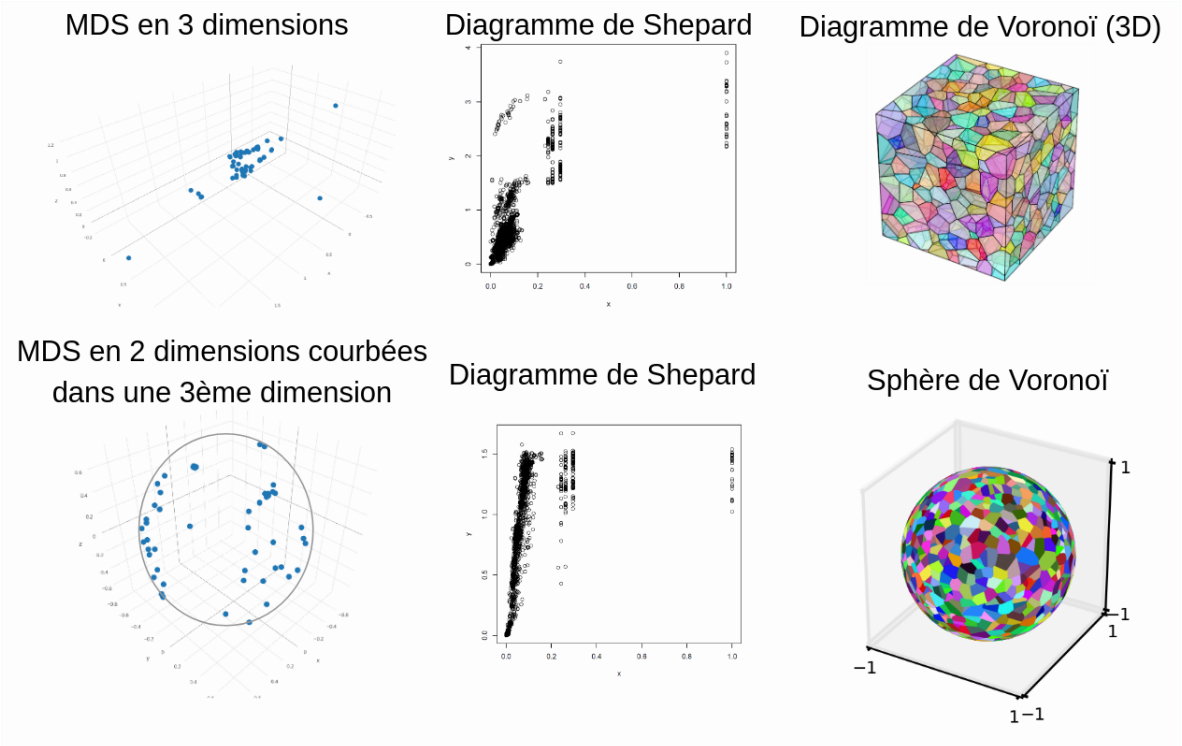


FIGURE 7.9 – Illustration en haut de la réduction d'une matrice en trois dimensions via une MDS (associée à son diagramme de Shepard), suivi du dessin des cellules Voronoï en 3D (les cellules sont seulement illustratives). Sur la partie basse on peut voir la réduction d'une matrice sous la forme d'un plan courbé en sphère dans un espace de dimension 3 suivie d'une illustration de cellules de Voronoï (surfaces quadratiques délimitées par des géodésiques) disposées sur une sphère.

Faire un diagramme de Voronoï sphérique suppose d'avoir préalablement transformé la matrice de distances entre les organismes dans un ensemble de coordonnées disposées sur une sphère et respectant au mieux la matrice de distance. Nous devons donc faire une MDS sphérique. Cela tombe bien car les auteurs de COX et al., 1991 ont proposé une méthode pour faire ce type de MDS et les auteurs de LEEUW et al., 2009 en ont proposé une implémentation au sein de paquet R SMACOF.

Une MDS sphérique est bien plus lente (plusieurs minutes pour une centaine de génomes) mais peut être plus rapide si nous explorons moins de solutions de placement de points. Au niveau de la perte d'information, nous pouvons comparer une MDS conventionnelle réduite à 3 dimensions et une MDS sphérique avec un diagramme de Shepard 7.9 page précédente. Ce diagramme donne la correspondance entre les distances réelles et les distances après réduction. L'idéal étant d'avoir une droite indiquant une parfaite corrélation linéaire entre les distances réelles et les distances obtenues après réduction de la dimensionnalité. De ces figures (basées sur 79 génomes de *Listeria monocytogenes*), nous observons que la MDS sphérique semble bien rendre compte des petites distances mais sature avec les grandes à l'inverse d'une MDS conventionnelle qui semble suréloigner des points parfois proches. À l'aune de ces figures, selon moi la MDS conventionnelle semble mieux rendre compte des distances réelles que la MDS sphérique mais cela est discutable. Quel est le pire ? Éloigner des organismes proches ou rapprocher des organismes lointains ?

Au reste, la dernière version du module Python Scipy (VIRTANEN et al., 2020) est capable de calculer les diagrammes de Voronoï sphériques à partir de coordonnées cartésiennes situées sur une sphère. Et même si le calcul des surfaces de ces cellules de Voironoï quadratiques est plus ardu qu'avec des polygones classiques cela semble réalisable.

Une fois ces poids obtenus, il reste cependant à les transmettre à l'algorithme NEM afin d'améliorer le partitionnement du pangénome. Or l'algorithme n'est pas adapté pour cela en l'état. Une autre solution, peu élégante, est de dupliquer les génomes dans la matrice autant de fois que leur poids le nécessite pour rendre compte de leur représentativité réelle.

Pour limiter la perte de l'information, l'idéal serait, sachant la matrice de présence/absence M de dimension (F,N) de considérer chaque génome comme un point dans un hyperespace de dimension F courbé sous la forme d'une hypersphère d'espace de dimension $F+1$ et de calculer les hypervolumes des cellules de Voronoï associées. Après en avoir discuté avec Catherine Matias et Christophe Ambroise, ceux-ci m'ont conseillé de ne pas m'y aventurer car cela est « horriblement compliqué ». Dont acte, toutefois si quelqu'un lit ce manuscrit de thèse et sait comment faire cela, merci de me contacter. Quelques articles de recherche mathématique semblent d'ailleurs avoir été publiés sur ce sujet (WILSON et al., 2014).

Ainsi, ce travail fut interrompu à ce stade sans produire de résultats plus approfondis. J'ai néanmoins découvert par la suite la méthode panGP proposée par les auteurs de ZHAO et al., 2014 qui semble la seule approche à ma connaissance proposant une stratégie d'échantillonnage de génomes basée sur une matrice de distance aux fins d'analyses pangénomiques.



Chapitre 8 : Description technique de la méthode PPanGGOLiN

8.1 Informations sur l'implémentation effectuée

8.1.1 Informations sur le langage et les dépendances utilisés

D'un point de vue concret, la méthode PPanGGOLiN consiste en une application principalement développée en langage Python et en langage C. Cette application se base de plus sur les quatre dépendances externes suivantes : Prodigal (Hyatt et al., 2010), Aragorn (Laslett, 2004) et Infernal (E. P. Nawrocki et al., 2013) pour ce qui est de l'annotation syntaxique des génomes et MMseqs2 (Steinegger et al., 2017) afin d'obtenir les familles de gènes homologues. Dans PPanGGOLiN, le langage Python est principalement utilisé pour, d'une part, manipuler les fichiers d'entrée et de sortie ainsi que les informations sur les gènes et les familles auxquels ils appartiennent, et d'autre part, pour générer le graphe de pangénome et les diverses statistiques qui y sont associées. Pour cela l'application se base notamment sur plusieurs bibliothèques Python parmi les plus célèbres telles que Pandas (McKinney, 2010), Numpy (Walt et al., 2011) et Scipy (Virtanen et al., 2020).

Pour ce qui est du partitionnement du pangénome PPanGGOLiN utilise la méthode statistique NEM que nous avons exposée en détail à la section 7.4 page 117 du chapitre précédent. Le code source de cette méthode a été développé en C par Mo Dang (Dang, 1998) en 1998 et a été incorporé dans l'application PPanGGOLiN. L'interfaçage entre les deux langages utilisés (Python et C) s'effectue au moyen de la bibliothèque Cython (Behnel et al., 2011) chargée de compiler et de rendre exécutable en Python le code source de la méthode NEM. Ainsi, les informations sur le pangénome sont transmises aux fonctions C de la méthode NEM à l'aide des fonctionnalités de cette bibliothèque mais également via des fichiers temporaires détaillés à la sous-section 8.2.5 page 148, afin de permettre la conservation des étapes intermédiaires des calculs statistiques.

8.1.2 Disponibilité de l'application

Le code source de l'application PPanGGOLiN est disponible sur un dépôt GitHub public (GAUTREAU et al., 2019) sous licence CeCILL 2.1. Ce dépôt GitHub comprend également une documentation détaillant le fonctionnement des sous-commandes décrites à la sous-section précédente.

Enfin, l'outil a également été conçu pour être aisément installable via le gestionnaire de paquets : Conda (CONTINUUM ANALYTICS, 2017). Pour cela Conda se charge de télécharger l'application PPanGGOLiN sur le répertoire Bioconda (GRUNING et al., 2018), de déclencher la compilation et l'installation de l'application PPanGGOLiN, enfin il effectue également l'installation conjointe des quatre dépendances externes de PPanGGOLiN déjà évoquées : Prodigal, Aragorn, Infernal et MMseqs2.

8.1.3 Contributions

Pour ce qui est des contributions au développement de l'application PPanGGOLiN, je suis le concepteur et développeur principal de la méthode dès son origine [1]. De plus, Adelme Bazin a également contribué significativement à son amélioration et à son optimisation notamment par la mise en place d'un stockage des données de PPanGGOLiN au format HDF5 et la séparation de l'application en de multiples sous-commandes au lieu d'une seule. C'est également lui qui s'est chargé d'intégrer les fonctionnalités « panRGP » (dont l'algorithme est décrit au chapitre 11 page 189) à l'intérieur de l'outil PPanGGOLiN.

On notera aussi l'aide spontanée de Sébastien Letort de l'INRIA de Rennes qui a proposé quelques améliorations sur le GitHub de PPanGGOLiN dans le cadre du festival « HacktoberFEST 2019 » [2].

8.2 Fonctionnement de l'application PPanGGOLiN

8.2.1 Interface en ligne de commande de la méthode PPanGGOLiN

Pour ce qui est de l'interaction avec l'utilisateur, l'application s'utilise en ligne de commande à travers diverses sous-commandes offrant chacune ses fonctionnalités propres dans le processus d'obtention du graphe de pangénome partitionné et dans l'exploitation de celui-ci (comme illustré sur la figure 8.1 page suivante).

La sous-commande d'entrée est au choix : `annotate` pour annoter les génomes à intégrer dans le pangénome, `workflow` pour effectuer automatiquement toutes les étapes jusqu'à l'obtention du pangénome, ou enfin `panrgp` pour lancer les sous-commandes `workflow` suivie de la détection des RGPs et des spots (voir chapitre 11 page 189 pour plus d'informations sur ces concepts).

Ces trois sous-commandes génèrent ainsi un fichier HDF5 stockant toutes les informations nécessaires à chaque étape de l'analyse. Ainsi, hormis pour ces trois sous-commandes autonomes, chacune des sous-commandes de l'application PPanGGOLiN nécessite de fournir le fichier au format HDF5 (THE HDF GROUP, 2000-2010) obtenu par une autre sous-commande.

1. Il est également à noter que Jonathan Mercier et Alexandre Renaux avaient mené une réflexion préalablement à cette thèse sur l'utilisation de NEM, en particulier pour le partitionnement de voies métaboliques.
2. Lui permettant ainsi de remporter un t-shirt offert par les organisateurs du festival.

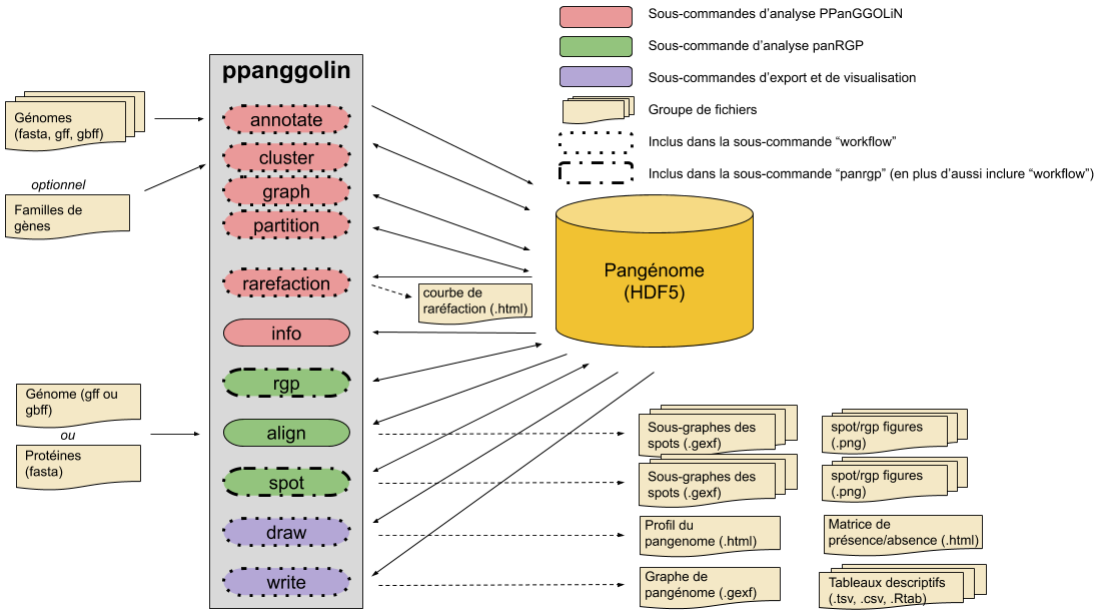


FIGURE 8.1 – Illustration des types de fichier en entrée et en sortie de chaque sous-commande de l'application PPanGGOLiN. La sous-commande `workflow` inclue les sous-commandes représentées en pointillés (en suivant un ordre d'exécution de haut en bas). La sous-commande `panrgp` inclue celles déjà présentes dans `workflow` auxquelles s'ajoutent les sous-commandes représentées en pointillés et tirets (en suivant un ordre d'exécution de haut en bas).

Voici les fonctionnalités résumées de chaque sous-commande :

— `annotate` : permet d'annoter syntaxiquement un ensemble de génome ou d'importer des annotations déjà effectuées.
— `cluster` (requiert le fichier HDF5 obtenu par la commande `annotate`) : permet de construire des familles de gènes homologues depuis un ensemble de génome annotés ou d'importer des familles de gènes déjà constituées.
— `graph` (requiert le fichier HDF5 obtenu par la commande `cluster`) : permet de construire le graphe de pangénome à partir d'un ensemble de génomes annotés et dont les familles de gènes ont été construites.
— `partition` (requiert le fichier HDF5 obtenu par la commande `graph`) : permet de partitionner le graphe de pangénome précédemment obtenu.
— `rarefaction` (requiert le fichier HDF5 obtenu par la commande `graph`) : permet de partitionner des sous-échantillons de taille croissante du graphe de pangénome précédemment obtenus afin de calculer les courbes de raréfaction des différentes parties du pangénome selon une approche basée sur la loi de Heaps (décrite à la section 3.2 page 49).
— `write` : permet de générer des fichiers de résultats comme la matrice de présence/absence au format CSV, le graphe de pangénome au format GEXF, de multiples statistiques concernant les génomes inclus et divers fichiers plats décrivant le pangénome.
— `draw` : permet de générer des figures dynamiques en HTML représentant le pangénome (illustré à la sous-section 8.2.4 page 144).

— `info` : permet d'obtenir des informations sur le pangénome comme les sous-commandes déjà employées sur ce pangénome ou des métriques basiques sur celui-ci (nombre de génomes, nombre de gènes, nombre de familles de gènes...).
— `workflow` : permet d'exécuter en une sous-commande les 6 sous-commandes suivantes successivement : `annotate`, `cluster`, `graph`, `partition`, `write`, `draw` ainsi que la commande `rarefaction` (facultatif).
— `align` (requiert le fichier HDF5 obtenu par la commande `partition`) : permet en fournissant un génome à cette commande d'assigner chaque gène de ce génome à une partie du pangénome même si ce génome n'est pas inclus dans le pangénome.
— `rgp` (requiert le fichier HDF5 obtenu par la commande `partition`) : permet d'exécuter la méthode panRGP décrite au chapitre 10 page 181.
— `spot` (requiert le fichier HDF5 obtenu par la commande `rgp`) : permet de regrouper des RGPs ancrés dans la même zone du pangénome (voir également le chapitre 10 page 181).
— `panrgp` : permet d'exécuter en une sous-commande les 8 sous-commandes suivantes successivement : `annotate`, `cluster`, `graph`, `partition`, `write`, `draw`, `rgp`, `spot` ainsi que la commande `rarefaction` (facultatif).

À partir de ce fichier HFD5 toutes les étapes conduisant au graphe de pangénome partitionné peuvent être donc être relancées isolément en gardant les informations obtenues avec la sous-commande précédente.

8.2.2 Fichiers d'entrée

Pour utiliser la méthode PPanGGOLiN, il suffit de fournir à la sous-commande `annotate` un fichier tabulé listant les génomes à utiliser pour construire le pangénome. Ce fichier comporte pour chaque ligne un identifiant unique pour chaque génome ainsi que le chemin vers le fichier de séquences nucléiques au format FASTA associé à celui-ci. À la suite du chemin chaque ligne peut également lister les identifiants des séquences circulaires afin que PPanGGOLiN prenne en compte l'existence d'un lien de voisinage entre le dernier et le premier gène d'une séquence. Si l'utilisateur veut importer dans PPanGGOLiN des annotations déjà effectuées par d'autres moyens, il peut alors fournir un fichier ayant la même forme, mais où les chemins vers les fichiers, au lieu d'être au format FASTA, sont remplacés par des chemins pointant vers des fichiers au format GBFF. PPanGGOLiN détecte et importe automatiquement ces fichiers FASTA ou GBFF même s'ils sont compressés par l'algorithme GZIP (DEUTSCH, 1996).

Sur la base de ces fichiers seulement, la suite des étapes de l'application peut ainsi se poursuivre. Cependant si l'utilisateur le souhaite il peut également importer un regroupement de séquences sous la forme de familles de gènes obtenues par la méthode de son choix. Pour cela, l'utilisateur devra fournir à la sous-commande `cluster` un fichier tabulé où chaque ligne indique un identifiant de famille de gènes suivi de l'identifiant du gène. Chaque identifiant de gène doit correspondre avec celui présent dans les annotations effectuées par PPanGGOLiN ou importées.

8.2.3 Fichiers de sortie

Fichier HDF5

L'ensemble des informations sur un pangénome à chacune des étapes du processus de la construction du graphe de pangénome partitionné sont stockées sous la forme d'un fichier au format HDF5 (« Hierarchical Data Format 5 », THE HDF GROUP, 2000-2010) manipulé via la bibliothèque Python Pytables (TEAM, 2002). Ce fichier binaire permet notamment de conserver les séquences de chaque génome, les annotations de ceux-ci, les familles de gènes et leur séquence centrale, le graphe de pangénome ainsi que les différentes parties à laquelle chaque famille de gènes appartient. L'avantage principal de cette structure hiérarchique est qu'elle permet de ne charger en mémoire que les informations nécessaires au fonctionnement de chaque sous-commande. Ainsi par exemple, les séquences nucléotides ne sont par exemple pas chargées en mémoire lors de la construction du graphe de pangénome car ces informations sont inutiles pour cette sous-commande.

En plus du fichier HDF5, la méthode est capable de rendre compte des résultats à travers plusieurs fichiers textes dont les principaux sont le graphe d'une part et la matrice de présence/absence d'autre part.

Fichier contenant le graphe de pangénome au format GEXF

Pour sa part, le graphe de pangénome est exportable sous la forme d'un fichier GEXF (« Graph Exchange XML Format »). Ce format de fichier permet de décrire les structures de réseaux complexes couplées aux données et dynamiques associées. Ainsi, en plus de stocker la structure graphique de pangénome, plusieurs informations permettant d'annoter les nœuds et les arêtes du graphe peuvent être stockées dans ce fichier afin d'en faciliter l'exploration. Cela correspond par exemple à la partie à laquelle est assignée chaque famille de gènes, le nombre de gènes dans la famille, la taille médiane des gènes ou bien celle des régions intergéniques. Ce graphe peut être exporté par PPanGGOLiN soit en stockant la totalité des identifiants de chaque gène ou bien sans cette information (version dite « light »). Afin de visualiser le graphe que contient ce fichier je conseille d'importer son contenu dans le logiciel de visualisation de graphe nommé Gephi (BASTIAN et al., 2009) [1].

Fichiers contenant la matrice de présence/absence du pangénome

Pour ce qui est de la matrice de présence/absence, elle peut être exportée au format CSV (« Comma-Separated Values ») où les lignes correspondent aux familles de gènes, les colonnes aux génomes et les intersections des lignes avec les colonnes aux identifiants de gènes (concaténés par un « | » en cas de familles multigéniques). Par ailleurs, les 15 premières colonnes fournissent des informations sur les familles de gènes comme le nombre de présence ou la taille médiane des gènes. Ce format est inspiré du format de fichier généré par l'outil Roary (PAGE et al., 2015) à l'exception de la deuxième colonne. En effet, cette colonne ayant un contenu qui n'est pas contraint dans le format de Roary nommé « non unique gene name », dans PPanGGOLiN cette colonne est réutilisée pour indiquer la partie à laquelle chaque famille de gènes est assignée. En respectant ce format cela facilite l'utilisation de PPanGGOLiN en remplacement de Roary ; cela permettant également

1. On notera que pour les génomes contenant des milliers de génomes, l'outil Gephi aura des difficultés à importer le fichier « light » contenant les identifiants de gènes.

de réutiliser les outils de visualisation et d'analyse déjà développés autour de ce dernier. Enfin, PPanGGOLiN génère également une version réduite de ce fichier CSV au format « Rtab » où les identifiants de gènes sont remplacés par leur nombre de copies dans le génome en cas de présence ou bien par des 0 en cas d'absence.

Autres fichiers de sortie

En plus des trois précédents fichiers, de très nombreux fichiers de statistiques peuvent être générés sur chaque génome afin de mettre en évidence des résultats, le plus souvent déjà présents dans la matrice de présence/absence, sous des formes plus spécifiques et de façon plus condensée. Enfin, l'outil permet également d'exporter les séquences représentatives de chaque famille de gènes (au format nucléique ou protéique).

Il est pour finir à noter que la totalité des fichiers que PPanGGOLiN exporte peuvent également être directement compressés en GZIP (DEUTSCH, 1996) si l'utilisateur le désire (à l'exception du fichier binaire au format HDF5).

8.2.4 Fichiers de visualisation

Au-delà des nombreux fichiers textuels qu'il génère, PPanGGOLiN ouvre également la possibilité de visualiser le profil du pangénome 8.2 page suivante, la matrice de présence/absence d'un pangénome 8.3 page 146 ainsi que les courbes de raréfaction 8.4 page 147. Ces trois types de figures sont générés sous la forme de pages Web de sorte qu'elles soient éditables et que l'utilisateur puisse zoomer sur l'information qu'il souhaite mettre en exergue sans perte de qualité des figures. Ceci est rendu possible grâce à l'utilisation de la bibliothèque graphique Plot.ly (PLOTLY, 2015).

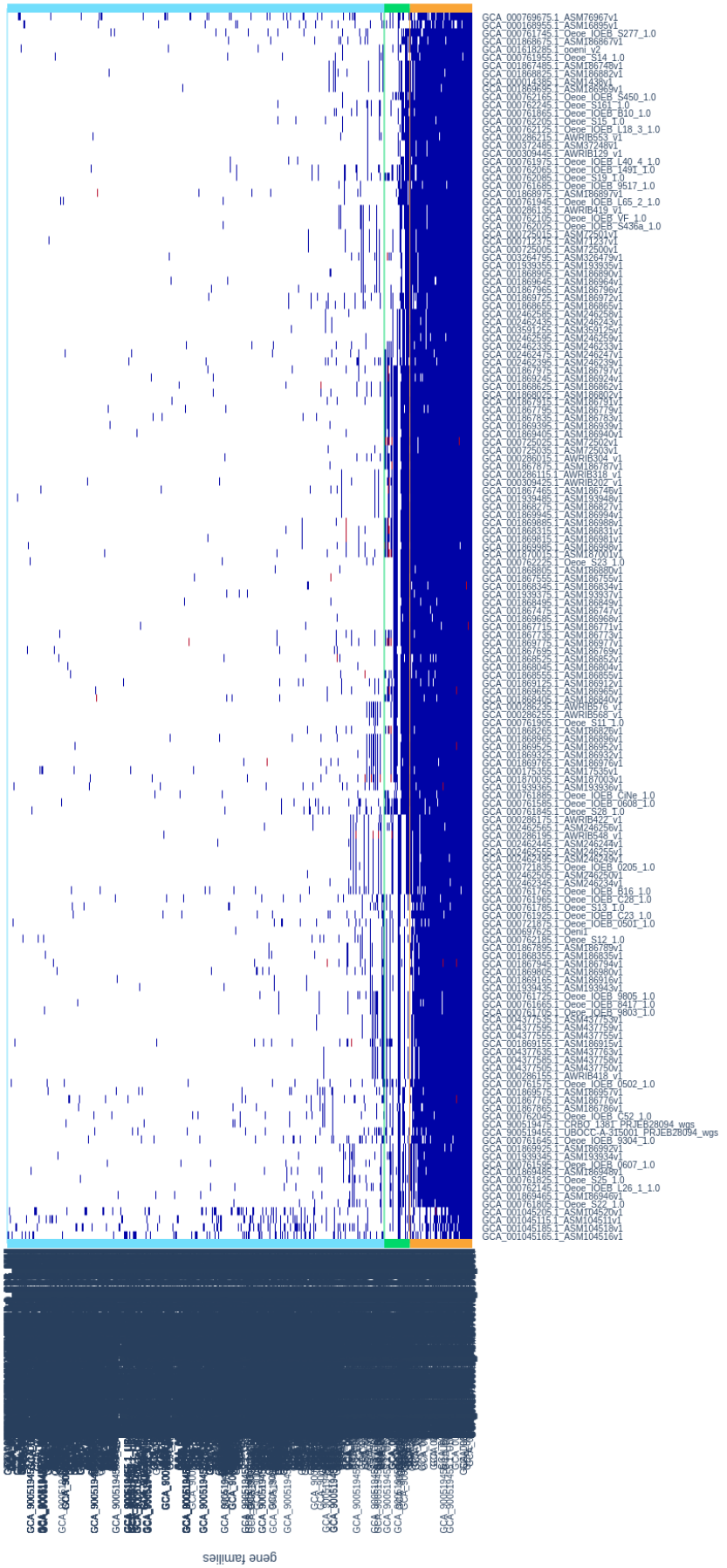


FIGURE 8.2 – Exemple de figure générée par l'outil PPanGGOLiN permettant d'explorer la matrice de présence/absence de l'espèce *Oenococcus oeni*. L'axe des ordonnées indique les noms des familles de gènes tandis que celui des abscisses indique les identifiants des génomes. Les cases bleues indiquent une présence, les cases rouges des présences multigéniques et les cases blanches des absences. Les familles de gènes persistants, coquilles et nuages sont respectivement délimitées par les barres oranges, vertes et cyan. La figure correspond à une page Web dynamique de sorte qu'il est possible de zoomer sur des parties de la matrice notamment afin de pouvoir visualiser le nom de chaque famille de gènes ou leur nombre de copies dans chaque génome.

FIGURE 8.3 – Exemple de figure générée par l'outil PPanGGOLiN permettant de visualiser le profil du pangénome de l'espèce *Oenococcus oeni*. Les barres sont colorées afin d'indiquer le nombre de familles à l'intérieur de chacune d'elles associées à chaque partie du pangénome. Le trait en pointillés indique le seuil de 95% de présence parmi les génomes de l'espèce. La figure correspond également à une page Web dynamique permettant d'obtenir des chiffres plus détaillés en survolant les barres avec la souris.

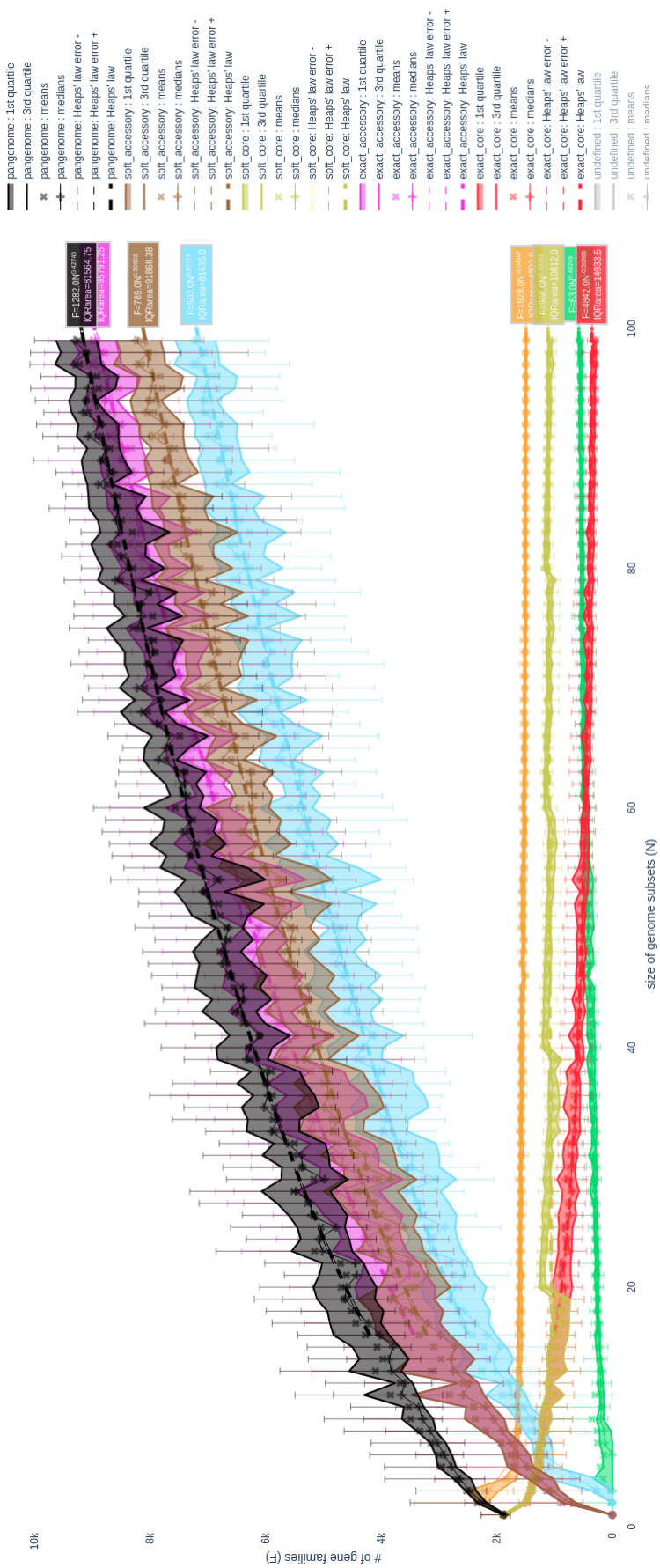


FIGURE 8.4 – Exemple de figure générée par l'outil PPanGGOLiN permettant de visualiser les courbes de raréfaction des différentes parties du pangénome de l'espèce *Oenococcus oeni*. La figure correspond également à une page Web dynamique permettant d'obtenir des chiffres plus détaillés en survolant les points de chaque courbe. Les courbes de raréfaction du pangénome, du génome accessoire, accessoire assoupli, nuage, persistant, cœur assoupli, coquille et enfin du cœur strict sont respectivement représentées en noir, violet, marron, cyan, orange, jaune, vert et rouge.

8.2.5 Fichiers temporaires

Pour l'utilisation de la méthode NEM à travers l'outil PPanGGOLiN plusieurs fichiers temporaires sont employés. Ces fichiers stockent l'ensemble des informations relatives au partitionnement du pangénome. Par défaut, ceux-ci sont créés dans le répertoire temporaire du système d'exploitation avant d'être effacés après utilisation. Le fait de stocker ces informations sous forme de fichiers permet le cas échéant d'analyser plus en détail des résultats de partitionnement.

Ces fichiers sont au nombre de 7 :

— `nem_file.str` : fichier indiquant le nombre de colonnes (nombre de génomes) et de lignes (familles de gènes) dans la matrice de présence/absence.
— `nem_file.nei` : fichier indiquant la structure du voisinage entre les familles de gènes ainsi que le poids de chaque voisinage.
— `nem_file.dat` : fichier indiquant la matrice de présence/absence.
— `nem_file.m` : fichier indiquant les paramètres d'initialisation du modèle.
— `nem_file.mf` : fichier indiquant les paramètres finaux du modèle et notamment sa log-vraisemblance.
— `nem_file.uf` : fichier indiquant les probabilités postérieures finales associant chaque famille de gènes à chaque partie comme indiqué à la sous-section 7.4.3 page 121.
— `nem_file.log` : fichier indiquant l'évolution des différents paramètres du modèle durant les différentes étapes de l'algorithme NEM.

8.3 Performances de l'application PPanGGOLiN

8.3.1 Temps de calcul

Le temps d'exécution d'une méthode est un critère crucial pour privilégier son utilisation par rapport à une autre méthode. Les ressources calculatoires des laboratoires sont en effet limitées et la recherche d'outils optimisés pour le calcul de pangénome est assurément l'un des facteurs qui a présidé au succès de l'outil Roary (PAGE et al., 2015). Via la méthode PPanGGOLiN, nous avons souhaité offrir une méthode à la fois simple d'utilisation car comprenant tous les outils nécessaires à la génération des pangénomes tout en présentant une rapidité similaire à Roary. Pour cela la méthode MMseqs2 a été employée pour construire les familles de gènes comme justifié à la section 7.2 page 113. De plus, dans le cas où la matrice de présence/absence passe un certain seuil (500 génomes par défaut), celle-ci est partitionnée par morceau de façon parallélisée (voir sous-section 7.4.4 page 125). Cela permet ainsi d'éviter que la complexité du partitionnement soit quadratique. Certes comme on peut le constater sur la figure 8.5 page suivante, la méthode possède une certaine variabilité notamment dans le temps de partitionnement du pangénome car le temps nécessaire pour explorer les paramètres du modèle utilisé possède une composante aléatoire. Toutefois, il reste que le temps d'exécution de la méthode reste globalement linéaire même après 10 000 génomes.

8.3.2 Usage mémoire

Pour ce qui est de l'utilisation de la mémoire, la figure 8.6 nous indique une tendance très clairement linéaire permettant d'envisager son utilisation sur des dizaines de milliers de génomes. À titre d'exemple plus pratique, l'usage mémoire de la méthode permet ainsi d'envisager l'utilisation de l'application PPanGGOLiN pour construire le graphe de pangénome d'environ 1000 génomes (13 Go de mémoire utilisé) sur un ordinateur portable conventionnel.

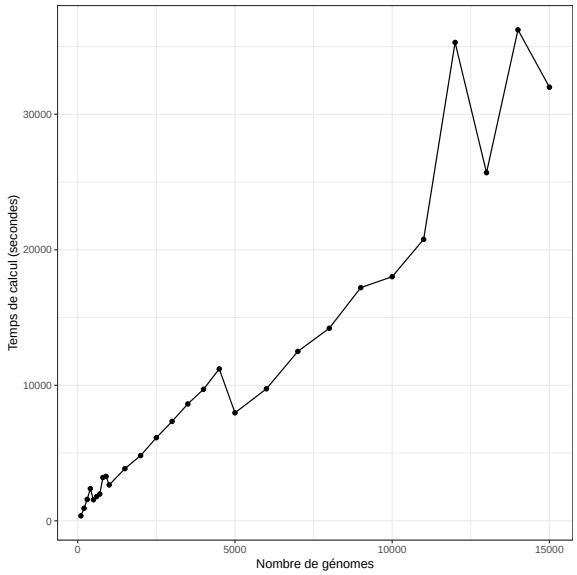


FIGURE 8.5 – Évolution du temps d'exécution de la méthode PPanGGOLiN lors de l'augmentation du nombre de génomes utilisés pour construire le pangénome de l'espèce *Salmonella enterica*. Un processeur comprenant 16 cœurs est ici utilisé quel que soit le nombre de génomes.

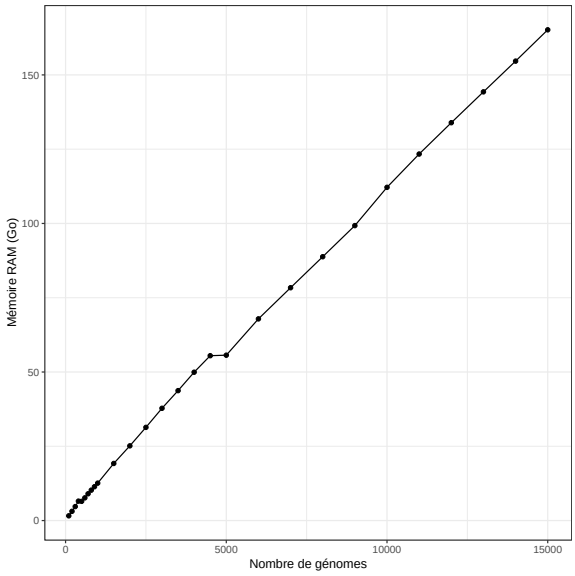


FIGURE 8.6 – Évolution de la quantité de mémoire vive utilisée par la méthode PPanGGOLiN lors de l'augmentation du nombre de génomes utilisés pour construire le pangénome de l'espèce *Salmonella enterica*. Les valeurs d'usage mémoire indiquées correspondent aux pics d'exécution.



Chapitre 9 : Résultats de la méthode appliquée sur 439 espèces

Les résultats présentés dans ce chapitre proviennent pour une large part des résultats de l'article GAUTREAU et al., 2020 [1].

9.1 Analyse des pangénomes de GenBank

Pour illustrer les résultats produits par la méthode PPanGGOLiN, nous allons l'employer sur 439 espèces procaryotes possédant au minimum 15 génomes après une étape de filtrage de la qualité des génomes (décrite au chapitre 10 page 181). Au total, cela correspond à 136 287 génomes assemblés provenant de la base de données GenBank (CLARK et al., 2016). Ces génomes sont majoritairement obtenus à partir de souches isolées, ils ont donc un niveau de complétion a priori élevé avec peu de contamination. Le filtrage qualité permet toutefois de se prémunir contre ces éventuels problèmes mais surtout de s'assurer contre de possibles erreurs d'annotation taxonomique pour certains d'entre eux. De plus, vont être analysées plus en détail au cours de ce chapitre, les espèces représentées par un plus grand nombre de génomes, soit les 88 espèces possédant au minimum 100 génomes. Mais tout d'abord nous allons exposer ce à quoi correspond concrètement un graphe de pangénome partitionné à travers l'exemple de deux espèces parmi les 88 ayant plus de 100 génomes : *Chlamydia trachomatis* et *Acinetobacter baumannii*.

9.1.1 À quoi ressemble un graphe de pangénome partitionné ?

Le pangénome de *Chlamydia trachomatis*

Après avoir énoncé la manière de le construire, voyons maintenant à quoi ressemble un graphe de pangénome partitionné. Pour une première représentation d'un graphe de pangénome, plaçons-nous dans le cas idéal. Nous allons donc commencer par visualiser le graphe de pangénome partitionné d'une espèce relativement simple, pathogène de l'œil et du tractus génital : *Chlamydia trachomatis*. Cette dernière comme la majorité des d'espèces intracellulaires obligatoires présente un petit génome d'environ 1Mb seulement et ne semble posséder aucun gène dupliqué. Avec

1. Les tableaux de données ayant servi à générer les figures présentées dans les sections de ce chapitre sont disponibles dans les informations supplémentaires de cet article.

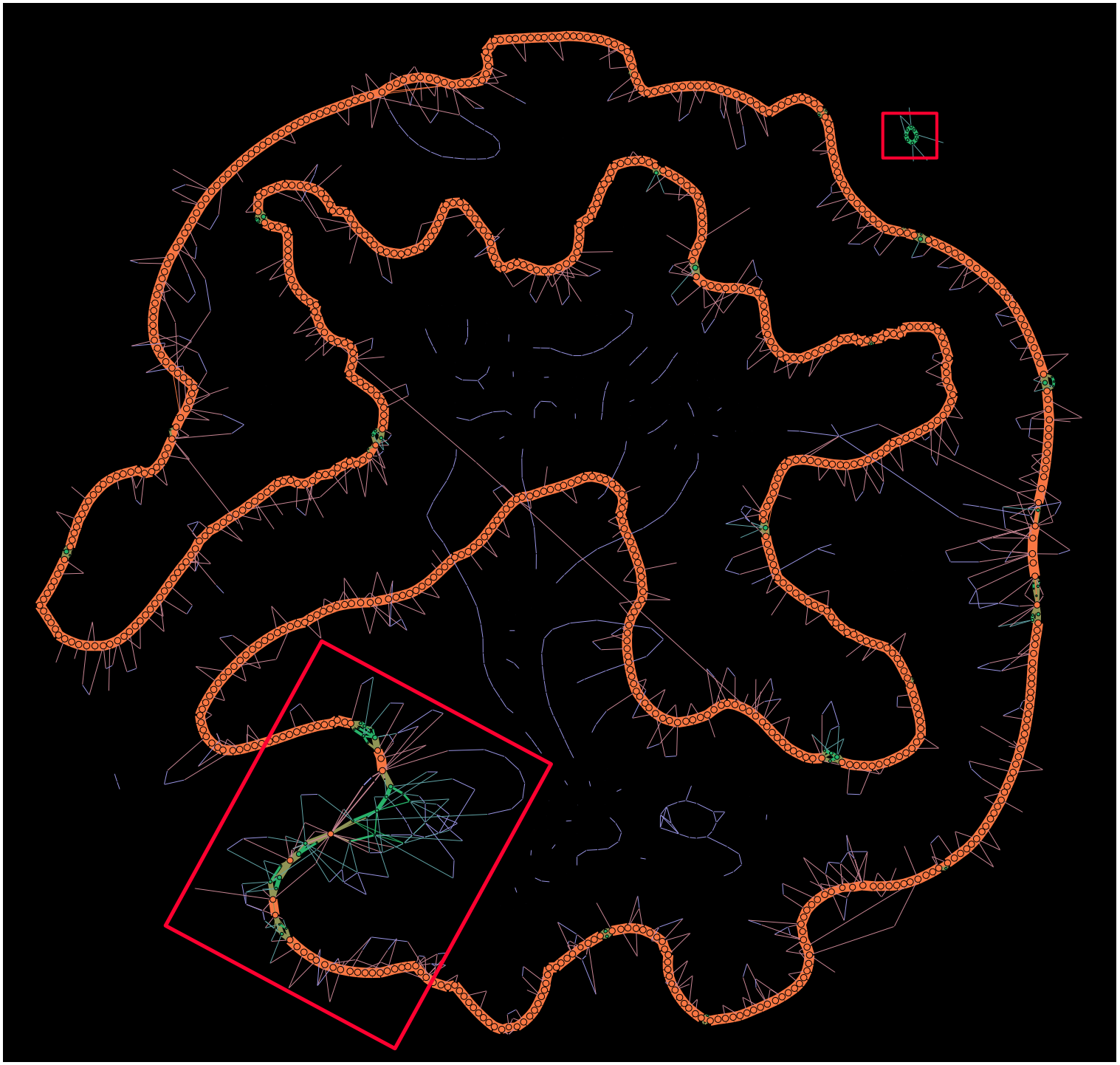


FIGURE 9.1 – Représentation du graphe de pangénome partitionné de 160 génomes de l'espèce *Chlamydia trachomatis*. Les nœuds oranges, verts et bleus correspondent, respectivement, aux familles de gènes du génome persistant, coquille et nuage. La taille des nœuds et l'épaisseur des arêtes correspondent, respectivement, au nombre de gènes présents dans chaque famille et au poids des arêtes. La couleur des arêtes est un mélange de celles des nœuds de départ et d'arrivée.

pourtant 160 génomes utilisés, l'espèce présente un très petit pangénome de seulement 1 833 familles de gènes. Partitionné en trois parties, le pangénome est constitué, respectivement, de 861,

82 et 890 familles de gènes persistants, coquilles et nuages. Par ailleurs, comme nous le reverrons à la sous-section 9.1.2 page 157, chaque génome de cette espèce possède une très grande proportion de gènes persistants (avec une médiane de ≈96% de la totalité des gènes de chaque génome).

Le graphe de pangénome partitionné dépeint par la figure 9.1 page précédente est obtenu en utilisant PPanGGOLiN avant d'être visualisé à l'aide du logiciel Gephi (BASTIAN et al., 2009). Pour dessiner le graphe, nous utilisons l'algorithme ForceAtlas 2 [1] (M. JACOMY et al., 2014). Nous utilisons cet algorithme d'agencement des nœuds car de tous les algorithmes testés, c'est celui qui déploie le mieux les longs chemins linéaires plutôt que de mettre en évidence seulement les regroupements comme la plupart des algorithmes d'agencement de graphes. Les algorithmes en question étant les suivants : ForceAtlas 1 et 2 (M. JACOMY et al., 2014), Fruchterman-Reingold (FRUCHTERMAN et al., 1991), Yifan Hu (Y. F. HU, 2005) et OpenOrd (MARTIN et al., 2011).

Sur l'illustration 9.1 page ci-contre de ce graphe de pangénome partitionné, on ne peut manquer d'apercevoir le chromosome circulaire principalement constitué de familles de gènes oranges appartenant au génome persistant. Celui-ci est à certaines positions interrompu par quelques chemins alternatifs d'une ou de plusieurs familles appartenant au génome coquille (en vert). On note en particulier une région de variabilité génomique plus importante (encadré en rouge au bas de la figure 9.1 page précédente). Cette région contient notamment des familles de gènes codant des lipases dont la fonction est associée à une activité d'hydrolyse de la phosphatidylcholine ainsi que des perforines. Si l'on se rappelle que les membranes plasmiques des cellules sont constituées de phospholipides comme la phosphatidylcholine, on peut trouver un lien avec le fait que cette espèce soit intracellulaire. En effet, la bactérie pathogène de l'humain *Chlamydia trachomatis* envahit les cellules par son intériorisation dans une vacuole la plaçant à l'abri du système immunitaire. La bactérie va ensuite détourner et modifier les phospholipides de l'hôte pour son propre compte (≈40% des lipides de la bactérie semblent provenir d'un détournement d'après les auteurs de WYLIE et al., 1997). On en déduit par conséquent, que les gènes de cette région participent très probablement à l'adaptation de la bactérie afin qu'elle puise au mieux ces ressources dans les différents types de cellules hôtes qu'elle infecte.

Pour ce qui est du génome nuage, on l'observe ici (nœuds bleus) sous plusieurs formes. D'une part, il forme des chemins connectés au chromosome principal ; potentiellement associés à des caractéristiques rares et spécifiques de certaines souches. Mais, on retrouve aussi des familles de gènes du génome nuage parsemées autour des familles du génome persistant, cela correspond très souvent à ce que nous avons précédemment appelé les écailles du PPanGGOLiN, c'est-à-dire, des fragments de gènes persistants. En effet, nous avons ici laissé le graphe brut, sans écaillage ni démêlement. Enfin, on peut également le voir sous une forme satellite autour du chromosome principal contenant des séquences de protéines hypothétiques chez *C. trachomatis* ou des gènes contaminants. On retrouve notamment des fragments provenant des bactéries *Elizabethkingia anophelis*, *Lactobacillus crispatus* et le plasmide « MobC » (souvent présent dans le genre *Staphylococcus*) associé à la fonction de conjugaison, ce dont est de toute évidence incapable l'espèce *C. trachomatis*.

1. Pour cette figure, nous utilisons les paramètres suivants : « Dessuade Hubs : True ; Prevent overlap : True ; Stronger gravity : True ; Scaling : 7000 ; Gravity : 0.15 ; Edge Weight influence : 1.3 ».

Enfin pour terminer on notera quelques plasmides, tous constitués de familles de gènes du génome nuage à l'exception d'un seul constitué de familles de gènes du génome coquille (encadré en rouge en haut de la figure 9.1 page 152). Ce dernier semble notamment connu pour contenir un facteur de virulence correspondant à un régulateur de la transcription du gène nécessaire à la synthèse du glycogène (SONG et al., 2013).

Voyons maintenant le pangénome d'une espèce ayant un pangénome bien plus complexe, celui d'*Acinetobacter baumannii*.

Le pangénome d'*Acinetobacter baumannii*

Nous allons ici illustrer un graphe de pangénome partitionné obtenu par la méthode PPanGGOLiN à partir de 3 117 génomes de l'espèce *Acinetobacter baumannii*. Par rapport au cas précédent, nous avons ici bien davantage de génomes, ceux-ci sont quatre fois plus grands (≈4 Mb) et l'espèce est bien plus diversifiée, de sorte que nous obtenons alors un graphe bien plus complexe. Cette espèce est généralement considérée comme une bactérie d'origine environnementale comme les autres bactéries de son genre, c'est-à-dire présente dans le sol et dans les eaux. Néanmoins, elle est principalement connue comme étant un pathogène opportuniste très souvent impliqué dans des infections nosocomiales. Depuis quelques décennies, on constate une forte augmentation des cas d'infections graves liés à cette bactérie (ASIF et al., 2018). Cela a notamment été associé au retour des soldats ayant eu des plaies infectées lors des guerres du Golfe ainsi qu'en Afghanistan, lui valant parfois le surnom de « Iraqibacter » (HOWARD et al., 2012). Aujourd'hui, cette espèce s'est largement propagée dans les hôpitaux et fait partie des six espèces bactériennes présentant le plus de multirésistances aux antibiotiques.

Selon nos analyses, l'espèce *Acinetobacter baumannii* possède un pangénome constitué de 69 446 familles de gènes obtenues à partir de l'analyse de 3 117 génomes. Le pangénome y est partitionné en trois parties auxquelles appartiennent 3 084, 1 529 et 64 833 familles de gènes, respectivement, pour le génome persistant, le génome coquille et le génome nuage. Chaque génome de cette espèce possède environ ≈82% de génome persistant, ≈13% de génome coquille et ≈5% de génome nuage.

Nous pouvons observer le graphe de pangénome partitionné de cette espèce sur la figure 9.2 page suivante [1]. Contrairement à *Chlamydia trachomatis*, il n'est plus ici possible d'observer la totalité de la circularité du chromosome. En effet, on observe de longs chemins de génomes persistants relativement stables et linéaires qui sont interrompus d'une part par des transposases (nœuds hyperconnectés dont un exemple est entouré en rouge) et d'autre part des chemins plus ou moins ramifiés appartenant au génome coquille. On visualise ainsi, le fait que le génome persistant est plutôt régulier spatialement mais interrompu par des parties variables qui s'insèrent dans des points chauds de variabilité.

Cela peut notamment se mesurer en calculant le degré moyen des nœuds au sein de chaque partie du graphe de pangénome. Si des chemins sont parfaitement linéaires, le degré moyen attendu pour ceux-ci serait très légèrement en dessous de 2 (leurs extrémités ayant un degré de 1). Pourtant,

1. Pour cette figure, nous utilisons les paramètres suivants pour l'algorithme ForceAtlas 2 dans le logiciel Gephi : « Dessuade Hubs : False ; Prevent overlap : False ; Stronger gravity : True ; Scaling : 8000 ; Gravity : 4.0 ; Edge Weight influence : 1.3 ».

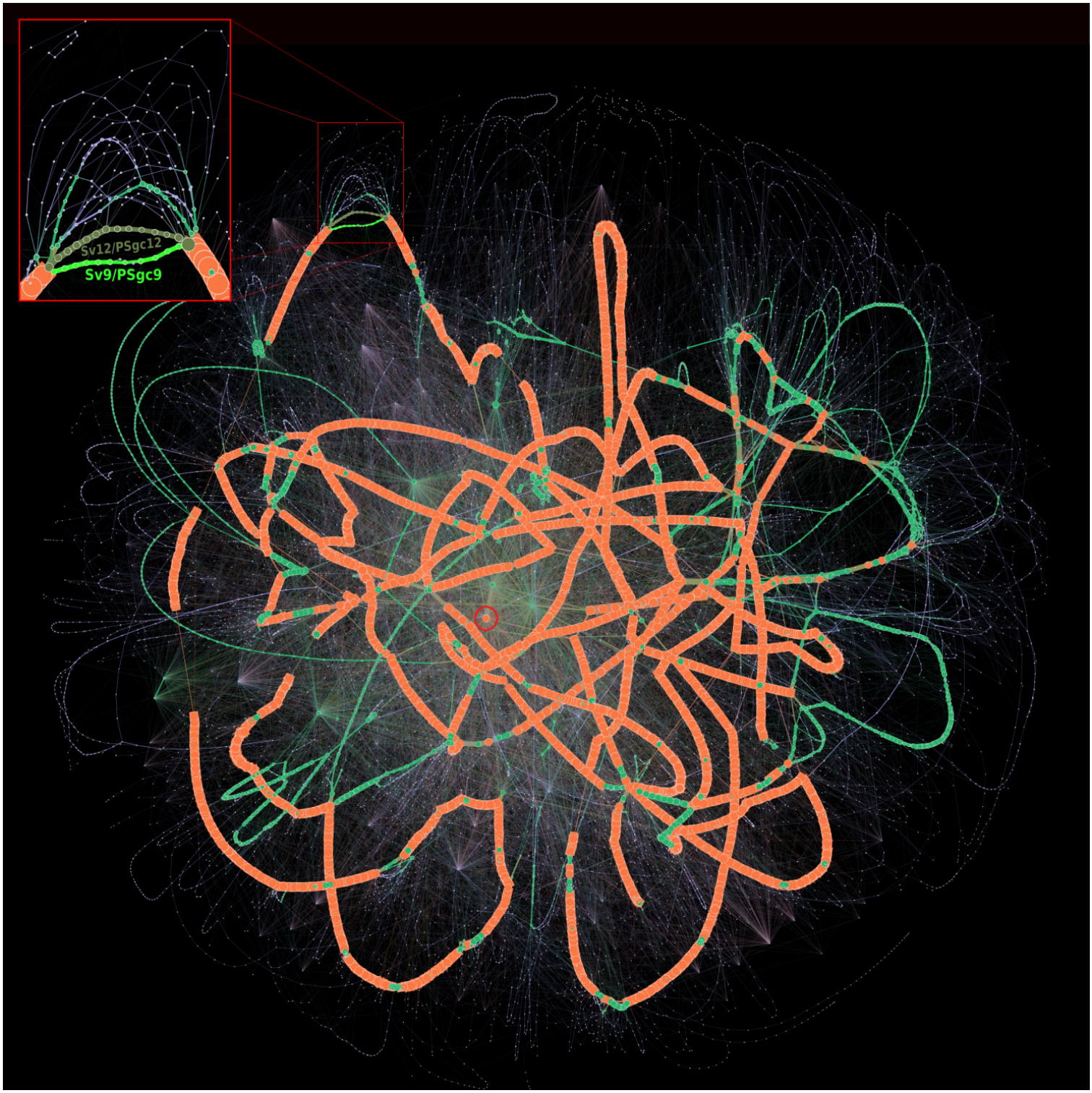


FIGURE 9.2 – Représentation reproduite depuis GAUTREAU et al., 2020 du graphe de pangénome partitionné de 3 117 génomes de l'espèce *Acinetobacter baumannii*. Les nœuds oranges, verts et bleus correspondent, respectivement, aux familles de gènes du génome persistant, coquille et nuage. La taille des nœuds et l'épaisseur des arêtes correspondent, respectivement, au nombre de gènes présents dans chaque famille et au poids des arêtes. La couleur des arêtes est un mélange de celles des nœuds de départ et d'arrivée. Pour que la visualisation soit plus lisible, il a été nécessaire de filtrer les familles de gènes très rares, c'est-à-dire celles ayant moins de 0.006% de présence (<20 gènes) bien qu'elles constituent 84.68% des nœuds (principalement des familles constituées d'un seul gène).

on observe que le degré moyen des nœuds persistant entre eux est de 2.80. En effet, il existe une transposase persistante au centre de la figure (cerclée en rouge) qui si elle est retirée fait passer le degré moyen à 2.5. De même, si l'on ôte les arêtes faiblement couvertes (par moins de 1% des génomes), le degré moyen des nœuds persistants repasse à 2. On peut donc ici confirmer pour cette espèce, l'hypothèse sur laquelle se base notre approche, à savoir la linéarité du génome persistant.

À l'inverse, si l'on s'intéresse au génome coquille, celui-ci est bien plus varié en termes de topologie. Si l'on regarde le degré moyen des nœuds du génome coquille entre eux uniquement, on obtient la valeur de 3.95. Cependant, certaines transposases du génome coquille ont des degrés très importants allant jusqu'à 151. En filtrant du génome coquille, ces nœuds ayant un degré supérieur à 10, on obtient alors un degré moyen de 2.61, ce qui reste plus important que pour le génome persistant. Cela confirme ce que l'on constate visuellement, c'est-à-dire un réseau de gènes plus diversifié en termes de topologie reflétant une composition mosaïque de gènes issus de transferts horizontaux comme le suggèrent les auteurs de OLIVEIRA et al., 2017.

Quant au cas du génome nuage, celui-ci possède un degré moyen de familles de gènes entre elles de de 1.96. Cependant, il serait trompeur d'en rester là. En effet, si l'on s'intéresse aux composantes connexes du génome nuage uniquement, on constate que 60% des nœuds du génome nuage appartiennent à une énorme composante connexe de 39 673 nœuds ayant un degré moyen de 2.65, soit une topologie similaire au génome coquille. Pour le reste, on a 14 623 nœuds appartenant à 3 606 composantes connexes qui sont quasiment linéaires et relativement courtes (de degré moyen 1.45 avec 4.25 nœuds en moyenne. Enfin, on a 10 575 nœuds singletons dans le génome nuage dont le degré est nul car ce sont des gènes isolés sur leur assemblage.

Pour illustrer ce que l'on peut trouver dans un graphe de pangénome partitionné, on peut s'intéresser à une région du graphe en particulier, celle encadrée en rouge dans la figure 9.2 page précédente. Cette région contient des gènes impliqués dans la synthèse des polysaccharides de la capsule. Nous l'avons dit, cette espèce est à l'origine de nombreuses infections nosocomiales. Les infections les plus graves sont dues au fait que ces bactéries sont souvent résistantes au traitement car elles possèdent des îlots génomiques conférant ces résistances. De surcroît, l'espèce *A. baumannii* possède un très grand nombre de sérotypes, résultat d'une capsule faite de polysaccharides possédant une importante variabilité. La capsule joue un rôle clé dans la capacité globale d'adaptation à l'hôte et dans sa pathogénie. En effet la capsule protège la bactérie des stress environnementaux, des réponses immunitaires et peut aussi conférer une résistance aux composés antimicrobiens (SINGH et al., 2018). À ce jour, plus d'une centaine de types de capsule distincts avec chacune leur organisation génomique ont été rapportés pour l'espèce *A. baumannii* (D. HU et al., 2013). En se basant sur l'ensemble des 3 117 génomes de GenBank, nous avons détecté 229 chemins pangénomiques différents partageant chacun de nombreuses sous-portions communes (encadré rouge de la figure 9.2 page précédente). Néanmoins, la plupart de ces chemins restent rares et seulement 24 sont conservés dans plus de 10 génomes. On observe en particulier deux chemins alternatifs très conservés. Si l'on fait le lien avec la nomenclature proposée par les auteurs de D. HU et al., 2013, ces deux chemins alternatifs commencent au niveau du gène « gnaA » et se terminent au gène « weeH » (voir la figure 3 de leur article). Toujours sur la base de cette nomenclature, nous avons un chemin (coloré en vert kaki sur la figure 9.2) correspondant au sérovariant « PSgc12 »

contenant 14 familles de gènes du génome coquille parfaitement conservées dans 581 génomes. Nous avons également un autre chemin (coloré en vert fluo sur la figure 9.2) correspondant au sérotype « PSgc9 » (qui est équivalent au sérovariant « PSgc7 » en termes de pur contenu en gènes) contenant 11 familles de gènes du génome coquille parfaitement conservées dans 408 génomes.

Cet exemple sur l'espèce *Acinetobacter baumannii* illustre comment le graphe de pangénome partitionné peut permettre d'explorer la diversité du contenu génomique de milliers de souches dans un même référentiel. Cependant certaines régions sont très complexes et difficiles à explorer seulement à travers le graphe. Dans ce cas, il peut être nécessaire de filtrer les génomes que l'on souhaite considérer, ce que l'on peut faire avec l'interface graphique de PanGBank que nous verrons au chapitre 10 page 181 mais il peut aussi être nécessaire de reprojeter les résultats issus du partitionnement du graphe de pangénome sur chaque génome ainsi que nous y reviendrons dans le chapitre 11 page 189 consacré à l'outil panRGP.

9.1.2 Résultats du partitionnement de PPanGGOLiN

Comme je l'ai dit précédemment, nous avons appliqué la méthode PPanGGOLiN sur 439 espèces. Cependant, comme il est difficile d'afficher les résultats sur autant d'espèces [1], nous allons ici nous concentrer sur les 88 espèces les plus abondantes dans la banque de données GenBank, c'est-à-dire celles ayant au minimum 100 génomes. Pour commencer cette analyse, nous souhaitons tout d'abord générer une figure permettant de comparer le partitionnement du pangénome de ces espèces. C'est pourquoi, afin obtenir cette figure 9.3 page suivante résumant les proportions des différentes parties du pangénome de ces espèces, nous projetons les résultats du partitionnement issu de PPanGGOLiN sur les génomes. Pour cela, nous associons chaque gène à une partie du pangénome, puis nous comptons le nombre de familles de gènes pour chaque partie du pangénome dans chaque génome des 88 espèces analysées. Nous constatons tout d'abord que malgré que cette figure soit issue de pangénomes générés à partir d'un nombre variable de génomes (de 101 pour *Flavobacterium psychrophilum* à 20 656 génomes pour *Listeria monocytogenes*), les proportions indiquées restent globalement comparables entre les espèces. En effet, il n'y a aucune corrélation notable entre la proportion médiane de génome persistant et le nombre de génomes considérés dans les espèces ($\rho_{\text{Spearman}} = -0.008$ et $R_{\text{Pearson}} = 0.015$) mais nous y reviendrons à la section suivante. Il en est de même concernant le génome coquille ($\rho_{\text{Spearman}} = 0.007$ et $R_{\text{Pearson}} = -0.04$). Cependant pour le génome nuage, il est possible que le nombre de génomes ait une très légère influence ($\rho_{\text{Spearman}} = -0.32$, p-valeur $= 0.002$ et $R_{\text{Pearson}} = -0.086$, p-valeur $= 0.42$).

À travers cette figure, on peut constater que les phyla les plus représentés dans GenBank sont les Proteobacteria, les Firmicutes et les Actinobacteria. À l'inverse deux phyla sont peu représentés, à savoir, les Spirochaetes, les Bacteroidetes et les Chlamydiae et de nombreux autres ne sont pas analysés car aucune espèce de ces phyla n'est représentée par plus de 100 génomes. Nous observons également que la proportion de génome variable (c'est-à-dire l'addition du génome coquille et du génome nuage) dans les génomes varie en général assez peu au sein des espèces ; en revanche elle varie bien davantage entre celles-ci. Cette proportion s'échelonne entre des espèces comme *Brucella*

1. Les tableaux de données sur les 439 espèces sont disponibles dans les informations supplémentaires de GAUTREAU et al., 2020.

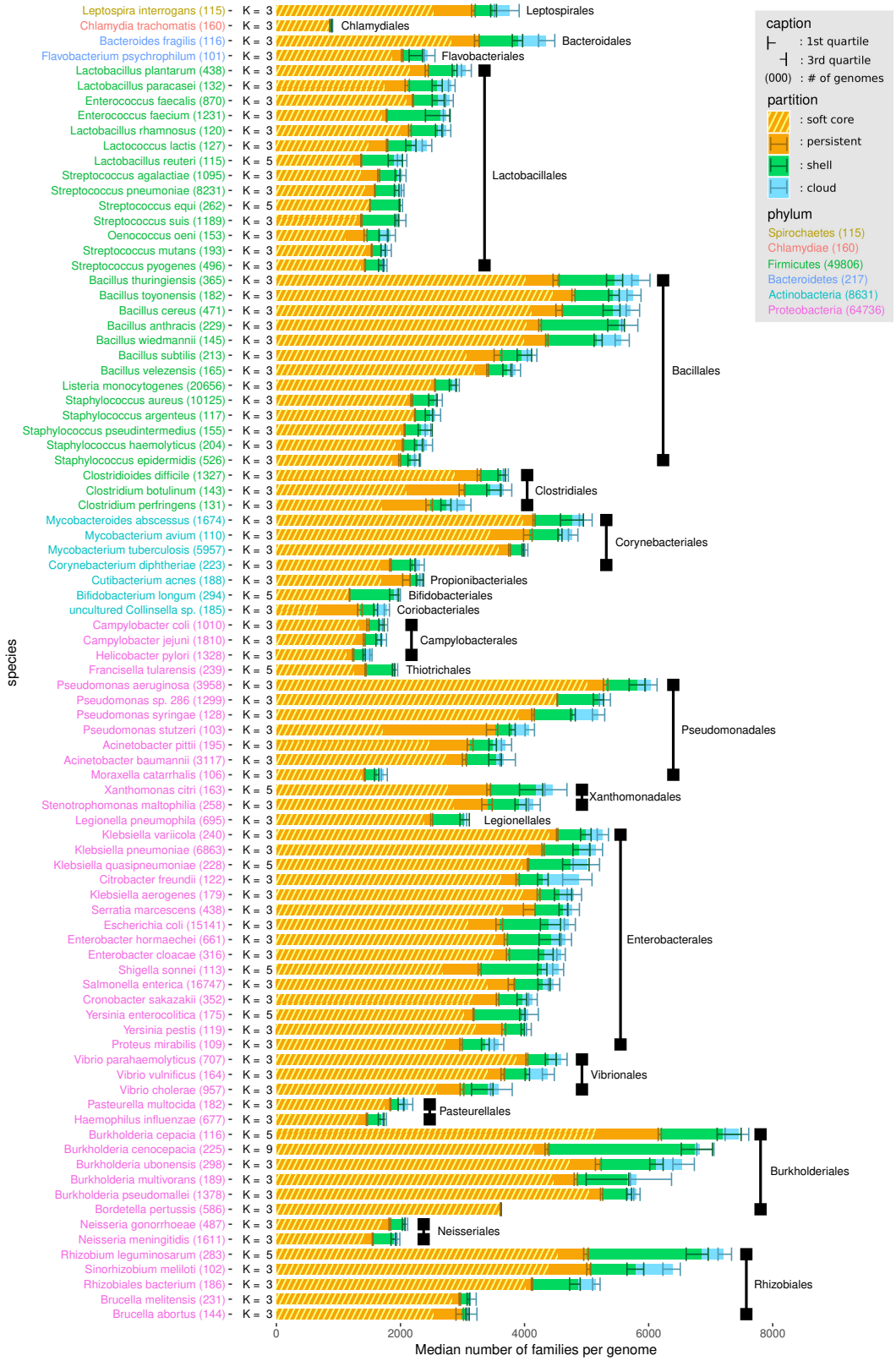


FIGURE 9.3 – Représentation reproduite depuis GAUTREAU et al., 2020 indiquant le nombre médian de familles de gènes par génome, dans 88 espèces, assignées au génome persistant (« persistent »), génome coquille (« shell ») et génome nuage (« cloud »). Le nombre de familles dans le génome cœur assoupli (« soft core », avec un seuil de présence à 95%) est également indiqué et superposé sur le génome persistant.

abortus, *Brucella melitensis*, *Bordetella pertussis* ou *Chlamydia trachomatis* présentant moins de 6% de génome variable et des espèces comme *E. faecium*, *S. suis*, *B. longum* , *R. leguminosarum*, *B. cenocepacia* ou *L. reuteri* ayant plus de 30% de génome variable, avec une moyenne à ≈17.8% en considérant toutes les espèces analysées.

On remarque par ailleurs que dans la totalité des cas, le génome persistant détecté est plus grand que le génome cœur assoupli (en utilisant un seuil de présence de 95%). Avec plusieurs cas notables comme le genre-espèce *Collinsella sp.* [1] et les espèces *Clotridium botulinum*, *Clotridium perfringens*, *Pseudomonas stutzeri* dont le génome cœur assoupli correspond à moins de 70% du génome persistant détecté par PPanGGOLiN.

Enfin pour ce qui est du nombre de parties dans le pangénome, on confirme globalement ce qu'avaient montré les auteurs de KOONIN et al., 2008, à savoir que le pangénome se partitionne généralement en trois parties (c'est le cas pour 385 espèces sur les 439 étudiées et 77 sur les 88 espèces ayant au minimum 100 génomes). Néanmoins, pour certaines espèces, le partitionnement trichotomique peut être remis en cause. Comme on peut le voir sur la figure 9.3 page précédente, c'est ici le cas de 10 espèces pour lesquelles $K = 5$ et pour une autre pour laquelle le partitionnement va jusqu'à $K = 9$ (*Burkholderia cenocepacia*). On constate de plus que la proportion de génome coquille est de ≈23% en moyenne dans les génomes de ces espèces contre ≈13% pour ce qui est de la totalité des 88 espèces. On peut donc dire que ces 11 cas correspondent à des espèces ayant une part de génome coquille plus importante. Enfin, pour ces espèces, on ne peut pas dire que ces parties supplémentaires de génome coquille soit « volées » au génome nuage puisque celui-ci possède une proportion de ≈4.5% pour les 88 espèces alors qu'elle est de ≈3.7% sur ces 11 espèces ; or, cette différence n'est pas significative (p-valeur = 0.2 via un test-t de Wilcoxon). Ces espèces possèdent donc une plus grande variabilité en familles de gènes fréquentes mais non fixées dans l'espèce.

9.1.3 Comparaison du génome persistant avec le génome cœur assoupli

Afin de montrer l'intérêt de la méthode PPanGGOLiN pour la détection de la partie persistante des génomes, nous l'avons comparée à l'approche génome cœur assoupli décrite à la section 6.1.3 page 102. Pour cela, nous avons employé l'approche basée sur la loi de Heaps décrite à la section 3.2 page 49. Ainsi pour chaque espèce, nous avons procédé à de multiples rééchantillonnages des génomes afin de mesurer la variabilité des métriques du pangénome avec un nombre croissant de génomes considérés dans les analyses. Nous obtenons ainsi ce que nous avons déjà appelé des courbes de raréfaction. Ces courbes de raréfaction indiquent si le nombre de familles tend à se stabiliser, à augmenter ou à diminuer lorsque l'on ajoute des génomes. Ainsi pour chacune des 88 espèces ayant au moins 100 génomes, 8 courbes de raréfaction décrivant la dynamique du pangénome d'une part, et des génomes persistant, coquille, nuage, accessoire (assoupli et non assoupli) et cœur (assoupli et non assoupli) d'autre part, ont été calculées pour des sous-ensembles contenant de 1 à 100 génomes avec 30 échantillons pour chaque taille. Dans le but d'optimiser les calculs, le nombre de partitions K est automatiquement déterminé, pour chaque échantillon, entre 3 et la valeur de K obtenue sur tous les génomes connus de l'espèce (et non 20 comme c'est le cas

1. Il est à noter que même si cela n'est pas annoté comme tel dans GenBank, cette espèce *Collinsella sp.* est représentée par des génomes provenant en réalité d'assemblages obtenus à partir de données métagénomiques.

par défaut comme cela est mentionné à la sous-section 7.4.3 page 123). Une régression non linéaire avec la méthode des moindres carrés (algorithme de Levenberg-Marquardt, MARQUARDT, 1963) a été effectuée pour ajuster les paramètres de la loi de Heaps aux courbes de raréfaction. Pour rappel, cette loi exprime l'évolution du nombre de familles de gènes F en fonction de nombre de génomes N à l'aide de deux paramètres : κ et γ, de sorte que $F = \kappa N^{\gamma}$. Le paramètre κ est un paramètre de proportionnalité tandis le second modélise la tendance (ci-après appelée tendance-γ) de la courbe de raréfaction ($\gamma = 0$ implique une tendance stable, $\gamma > 0$ implique une tendance croissante, $\gamma < 0$ implique une tendance décroissante). Les sous-ensembles de génomes de taille $\leqslant 15$ n'ont pas été utilisés pour l'algorithme de Levenberg-Marquardt car ils sont parfois trop variables pour assurer un bon ajustement. Enfin pour chaque taille de sous-ensemble, la médiane, la moyenne, les premier et troisième quartiles, le minimum et le maximum ont été calculés.

Pour comparer le génome cœur assoupli au génome persistant, nous allons évaluer les courbes de raréfaction qui a priori sont supposées avoir des tendances-γ proches de 0, c'est-à-dire correspondant à un ensemble stable de gènes conservés dans les génomes des espèces étudiées. Ces courbes ne peuvent donc que décroître ou au mieux (ce que nous recherchons) rester stable. Mais au-delà de la tendance-γ, il semble aussi important de prendre en compte la variabilité des estimations. En effet même avec une tendance-γ parfaitement stable, on pourrait parfaitement avoir une forte variabilité des prédictions sur de multiples échantillons. Pour cela, nous calculons également la surface du ruban généré en reliant tous les écarts interquartiles (InterQuartile Range area : IQRarea). Ainsi, des rubans fins signifient que les prédictions sont stables et résilientes à l'échantillonnage et inversement pour les rubans larges. Ces rubans sont notamment visibles sur les figures 6.3 page 104 et 8.4 page 147. En l'espèce, comme nous allons effectuer ces comparaisons sur des combinaisons de 100 génomes au maximum pour les 88 espèces évoquées, tout est normalement comparable de sorte qu'il n'y ait pas besoin de normaliser la surface des rubans interquartiles.

Une fois ces calculs effectués sur les 88 espèces, la figure 9.4 page suivante nous montre que la tendance-γ du génome persistant estimé par PPanGGOLiN est plus proche de 0 que celle de l'approche utilisant le génome cœur assoupli. D'abord visuellement, on constate que les points oranges sont plus près de zéro, mais surtout que la moyenne de la valeur absolue de la tendance-γ est de $9.1 \cdot 10^{-3}$ contre $2.5 \cdot 10^{-2}$ (valeur-p = $1.5 \cdot 10^{-9}$ avec un test-t de Student unilatéral à 2 échantillons appariés). De même, les écarts d'ajustement aux données par l'algorithme sont également faibles, ce qui veut dire que les courbes du génome persistant ont une tendance plus lisse (plus facile à modéliser sous la forme d'une loi de Heaps) avec un écart moyen de $5.3 \cdot 10^{-4}$ contre $2.1 \cdot 10^{-3}$ pour le génome cœur assoupli (valeur-p = $9.5 \cdot 10^{-11}$ avec un test-t de Student unilatéral à 2 échantillons appariés). Concrètement cela veut donc dire que le génome cœur assoupli a tendance à croître car sa valeur est réévaluée à la hausse lorsque que le nombre de génomes augmente. Le génome cœur assoupli, pourtant très utilisé dans de très nombreuses études, a ainsi tendance à sous-estimer le « vrai » génome persistant. À l'inverse, le génome persistant de PPanGGOLiN aura plutôt une très légère tendance à la décroissance, cela veut donc dire qu'il réévalue très légèrement à la baisse le génome persistant lorsque le nombre de génomes augmente. Le génome persistant aurait donc une très légère tendance à surestimer le « vrai » génome persistant. Enfin, le ruban de l'écart interquartile de la prédiction avec PPanGGOLiN est également plus fin avec une surface

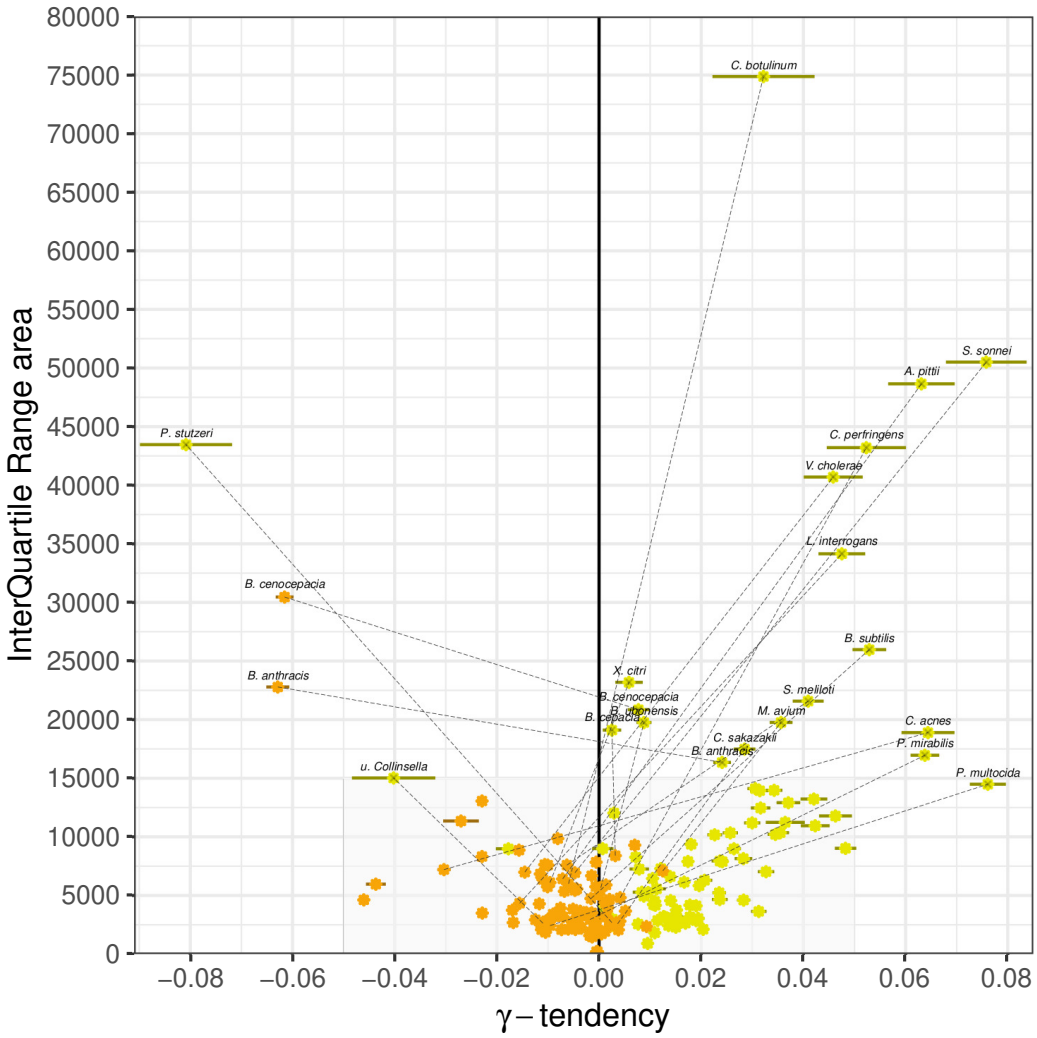


FIGURE 9.4 – Représentation reproduite depuis GAUTREAU et al., 2020 illustrant les valeurs de tendance-γ en fonction de la surface du ruban interquartile (IQRarea) pour chacune des 88 espèces sélectionnées ayant plus de 100 génomes. Les points oranges correspondent à l'estimation pour le génome persistant et les points jaunes à celui du génome cœur assoupli (seuil de présence à 95%) pour chaque espèce. Les barres horizontales sur chaque point indiquent l'erreur d'ajustement de la loi de Heaps aux données. La zone grisée correspond aux courbes de raréfaction relativement stables ($|\gamma| < 0.05$) et peu variables (surface du ruban interquartile < 15 000). Pour chaque espèce, si soit le génome cœur assoupli, le génome persistant ou bien les deux ne sont pas dans la zone grise, alors un trait pointillé est tracé pour aider à la comparaison. Pour des raisons de lisibilité, seuls les noms des espèces en dehors de la zone grise sont indiqués.

moyenne de 4906.6 contre 11645.9 pour le génome cœur assoupli (valeur-p = $8.9 \cdot 10^{-7}$ avec un test-t de Student unilatéral à 2 échantillons appariés). On montre ici que la prédiction du génome persistant est à la fois plus stable en moyenne mais aussi moins variable suivant l'échantillonnage. Enfin la détection du génome persistant n'amène pas à un « effet escalier » comme nous avons pu le voir sur la figure 6.3 page 104.

Toutefois, il faut remarquer que deux espèces ont des caractéristiques peu à l'avantage du génome persistant de PPanGGOLiN : *Bacillus anthracis* et *Burkholderia cenocepacia*.

Pour *Burkholderia cenocepacia*, cela pourrait s'expliquer par la grande part de génome coquille. C'est en effet la seule espèce à avoir une valeur de K aussi élevée ($K = 9$). Ainsi en fonction des échantillonnages, les schémas de présence absence du génome coquille seront plus ou moins bien marqués, de sorte que cela peut faire varier le génome persistant. C'est une limite de la méthode PPanGGOLiN.

Pour *Bacillus anthracis*, la source de variabilité pour définir le génome persistant est le résultat d'une assignation taxonomique discutable dans GenBank. En réalité, nous avons considéré comme l'espèce *B. anthracis*, un groupe d'espèce qui sont, selon la « Genome Taxonomy DataBase » (évoqué à la sous-section 2.2.3 page 38), est en fait *B. anthracis bona fide* accompagné de 17% de génome *Bacillus cereus* et *Bacillus thuringiensis*. Ce problème n'a pas été détecté par notre procédure de contrôle taxonomique parce que ces espèces sont à la limite du seuil de distance génomique conspécifique utilisé (évoqué à la sous-section 10.4 page 183). Certaines familles de gènes persistants de *B. anthracis bona fide* peuvent donc passer d'une partie génome persistant à une partie génome coquille, selon le rééchantillonnage. En excluant ces génomes mal classés, nous avons prédit un génome persistant plus grand que celui de l'ensemble initial de génomes ($\approx$1000 familles de gènes de plus) avec une tendance-γ très proche de 0 (valeur de $1.7 \cdot 10^{-2}$) et une surface de ruban interquartile de 8367, ce qui est meilleur que pour le génome cœur assoupli avec ces génomes ($\gamma = 3.6 \cdot 10^{-2}$ et surface du ruban égale à 32167.1).

Dans l'ensemble, ces résultats suggèrent que notre approche de partitionnement est plus robuste que l'utilisation d'un seuil arbitraire de 95% de présence. Le principal inconvénient de l'utilisation d'une fréquence arbitraire pour partitionner les pangénomes est que même s'il était possible de déterminer le meilleur seuil pour chaque espèce, il ne tiendrait pas compte du fait que certaines familles de gènes persistants peuvent avoir une fréquence atypiquement faible. Cela peut être dû comme nous l'avons vu à des pertes importantes de gènes dans la population ou à des raisons techniques comme l'appartenance à une région génomique difficile à assembler (c'est-à-dire des gènes manquants ou fragmentés dans les projets d'assemblage génomique).

9.1.4 Structure et dynamique du génome coquille et nuage

La tendance-γ du génome coquille nous informe sur la dynamique des pangénomes

Au-delà du génome persistant, nous pouvons aussi évaluer la dynamique d'accumulation des nouvelles familles de gènes notamment par la tendance-γ du pangénome, du génome coquille et du génome nuage. On distingue généralement deux types de dynamique de croissance de la taille des pangénomes : les pangénomes ouverts et les pangénomes fermés. Dès lors, on peut dire qu'une faible valeur de tendance-γ signifie un pangénome plutôt fermé, alors qu'une valeur élevée signifie un pangénome plutôt ouvert. Même si rigoureusement un pangénome fermé devrait signifier un pangénome complètement stabilisé, de nos analyses ne ressort aucune espèce répondant à ce strict critère, c'est-à-dire un pangénome ayant une tendance-γ nulle. Cela suggère qu'au lieu d'utiliser

ce genre de concept binaire, il est plus utile de quantifier le degré d'ouverture des pangénomes étant donné le flux de nouveaux gènes (principalement par transfert horizontal) ainsi que la perte de gènes que connaît chaque espèce.

Comme évoqué à la sous-section 9.1.3 page 159, nous déterminons la tendance-γ pour différentes parties du pangénome sur la base des courbes de raréfaction pour les 88 espèces étudiées. Si l'on s'intéresse à la distribution des tendances-γ des différentes parties prédites par PPanGGOLiN (voir figure 9.5), on remarque que le génome coquille est la partie du pangénome renfermant la plus grande amplitude de dynamique de croissance (moyenne de 0.41 avec un écart-type de 0.20) en comparaison des autres parties du pangénome telles que le génome nuage (moyenne de 0.55 avec un écart-type de 0.08). Ceci indique que les principales différences en matière de dynamique du pangénome entre les espèces semblent résider dans le génome coquille plutôt que dans le génome nuage.

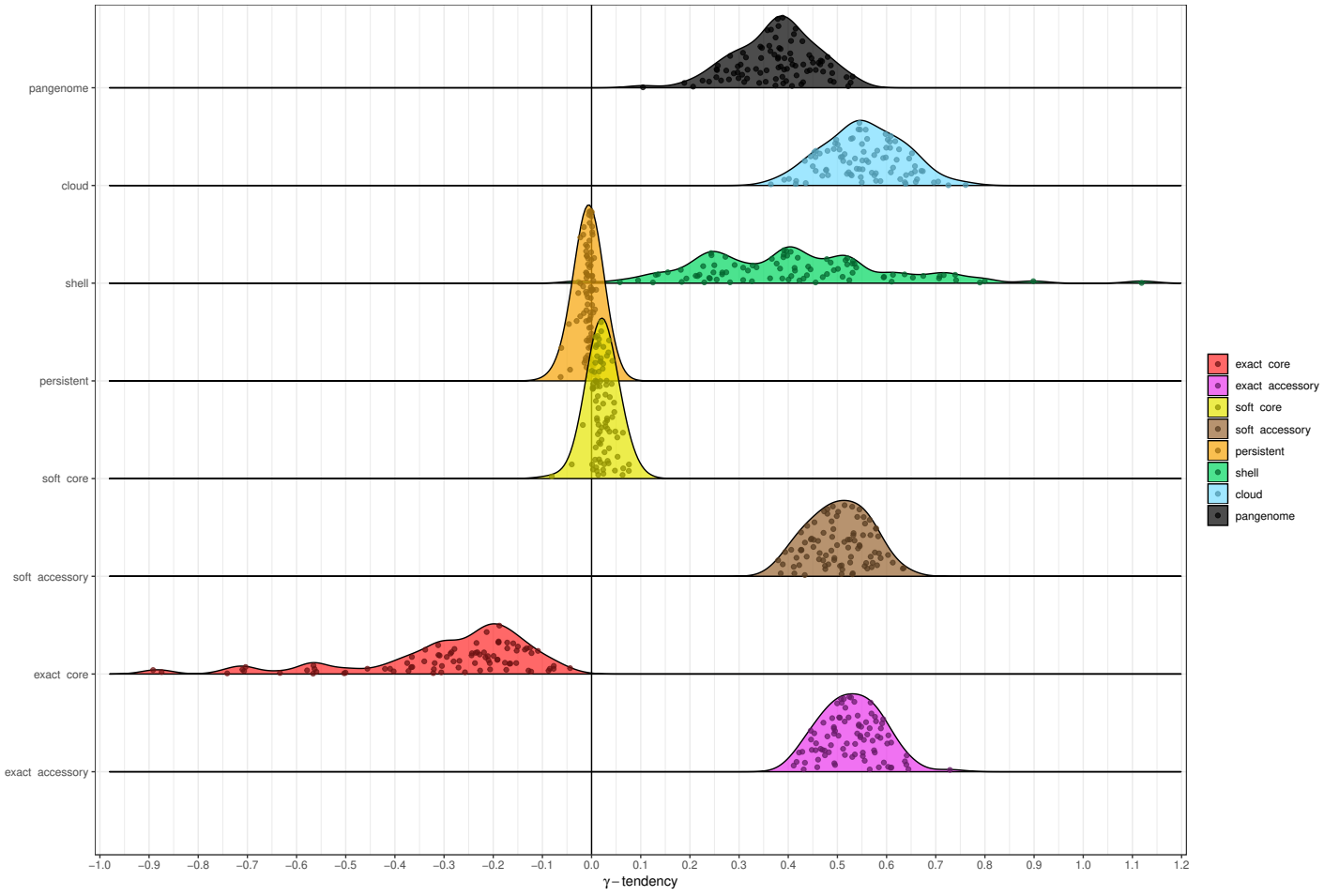


FIGURE 9.5 – Représentation reproduite depuis GAUTREAU et al., 2020 illustrant la distribution des tendances-γ pour les 88 espèces ayant plus de 100 génomes. Le partitionnement est effectué à la fois selon la dichotomie génome cœur/génome accessoire, la dichotomie génome cœur assoupli/génome accessoire assoupli et la trichotomie de PPanGGOLiN (génome persistant, génome coquille, génome nuage).

Le rapport entre le génome coquille total et le génome coquille dans les génomes semble conservé

Comme on peut le prévoir, la taille totale du génome coquille (en nombre de familles de gènes) est corrélée avec le nombre médian de gènes du génome coquille par génome ($\rho_{\text{Spearman}} = 0.81$, p-valeur $= 2.2 \cdot 10^{-16}$ et $R_{\text{Pearson}} = 0.86$, p-valeur $= 2.2 \cdot 10^{-16}$). Cependant, ce qui étonne c'est davantage la force de ce lien. En effet, cette corrélation linéaire de Pearson est solide ($R^2 = 0.73$, soit 73% de la variance expliquée), de sorte qu'on peut estimer les paramètres de la fonction affine reliant la taille totale du génome coquille C_t avec sa taille dans les génomes C_m. On obtient ainsi l'équation : $C_t = 3.324 \cdot C_m + 168.4$. On en déduit donc qu'on retrouve généralement entre $\approx \frac{1}{3}$ et $\approx \frac{1}{4}$ du génome coquille total de la population représenté dans chaque génome pour les 88 espèces étudiées ayant au moins 100 génomes. Un résultat similaire est de plus observé sur la totalité des 439 espèces avec un $R_{\text{Pearson}} = 0.86$ et $C_t = 3.011 \cdot C_m - 62.637$.

La proportion des différentes parties du pangénome dans les génomes est indépendante de la taille des génomes

Variables corrélées avec la taille médiane des génomes ?	ρ_{Spearman}	p-valeur$_{\text{Spearman}}$	R_{Pearson}	p-valeur$_{\text{Pearson}}$
Nombre médian de gènes du génome persistant dans les génomes	0.98	$2.2 \cdot 10^{-16}$	0.98	$2.2 \cdot 10^{-16}$
Nombre médian de gènes du génome coquille dans les génomes	0.7	$5.2 \cdot 10^{-14}$	0.68	$5.2 \cdot 10^{-13}$
Nombre médian de gènes du génome nuage dans les génomes	0.63	$4.3 \cdot 10^{-11}$	0.62	$9.8 \cdot 10^{-11}$
Nombre médian de gènes du génome variable dans les génomes	0.77	$2.2 \cdot 10^{-16}$	0.77	$2.2 \cdot 10^{-16}$
Fraction médiane du génome persistant dans les génomes	0.01	0.92	−0.04	0.71
Fraction médiane du génome coquille dans les génomes	−0.05	0.66	0.01	0.95
Fraction médiane du génome nuage dans les génomes	0.11	0.29	0.1	0.30
Fraction médiane du génome variable dans les génomes	0.01	0.95	0.04	0.70

TABLE 9.1 – Corrélation des différentes parties du pangénome après projection sur les génomes des 88 espèces ayant plus de 100 génomes en fonction de la taille médiane des génomes (en nombre de gènes).

Comme attendu, nous constatons une importante corrélation entre la taille des génomes et le nombre de gènes médians de l'ensemble des différentes parties du pangénome (voir tableau 9.1) : génome persistant, génome coquille, génome nuage et génome variable (génome coquille et génome nuage additionnés). Sachant cela, se pose la question de savoir si ces corrélations se conservent si l'on étudie les fractions (les proportions relatives) de ces parties en fonction de la taille des génomes. Comme on le voit sur le tableau 9.1, on n'observe aucune corrélation notable quelles que soient les parties étudiées (un exemple de l'absence de relation est illustré sur la figure 9.6 page suivante). Le même type de résultats est par ailleurs observé sur les 439 espèces. Ainsi, le fait que les génomes soient plus grands ne semble pas influer sur les proportions propres à chaque espèce. Un résultat négatif similaire avait déjà été reporté par les auteurs de ANDREANI et al., 2017 en comparant la fluidité des génomes (assimilable à la proportion de génome variable dans les génomes) avec leur taille (sur 90 espèces en l'occurrence). Ainsi, non seulement la proportion médiane des parties, comme nous l'avons déjà indiqué à la sous-section 9.1.2 page 157, n'a pas de lien avec le nombre de génomes considérés mais il n'y a pas non plus de lien avec la taille des génomes. Sachant l'influence de la taille des génomes et du nombre de génomes écartée, il est dès lors approprié de dessiner la figure 9.7 page 166 afin de comparer les proportions des différentes parties du pangénome entre les espèces.

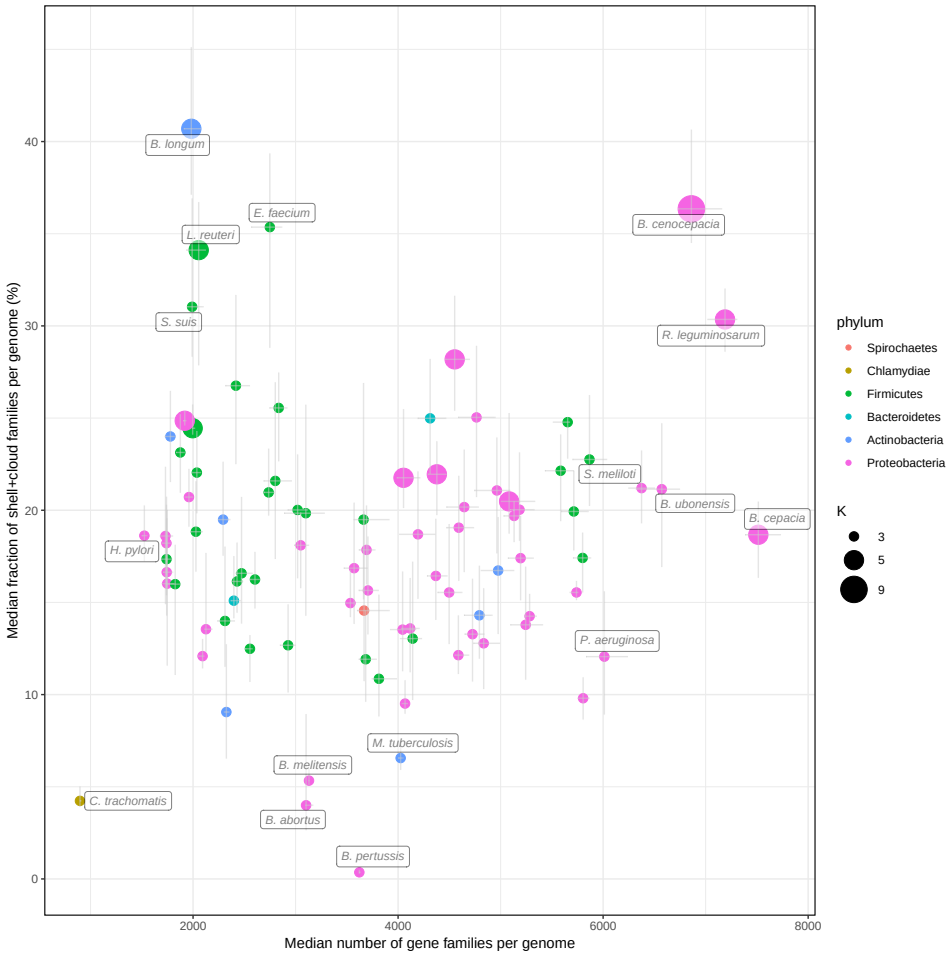


FIGURE 9.6 – Représentation reproduite depuis GAUTREAU et al., 2020 illustrant la fraction médiane de génome variable dans les génomes en fonction du nombre de gènes dans les génomes. Les résultats pour les 88 espèces les plus abondantes dans GenBank sont représentés. Les barres d'erreur montrent les intervalles interquartiles de la fraction de génome variable. La taille du point correspond au nombre de partitions (K) utilisées et les points sont colorés par phylum.

Comment dès lors expliquer ces équilibres propres à chaque espèce ? Ces équilibres dépendent de la capacité des espèces à acquérir de nouveaux gènes depuis leur environnement et à les fixer dans la population. Ce que nous pouvons dire, comme nous l'avons vu, c'est que la taille totale du génome coquille semble généralement correspondre à environ 3 à 4 fois celle retrouvée dans les génomes individuellement. Pour étudier plus en détail ces équilibres, il serait intéressant de les coupler à des informations environnementales. Néanmoins, les données environnementales manquent pour proposer une approche systématique des liens entre les environnements et la variabilité dans les pangénomes. On peut toutefois proposer quelques commentaires au cas par cas.

Si l'on s'intéresse par exemple aux bactéries associées à des hôtes comme les trois espèces : *Bifidobacterium longum*, *Enterococcus faecium* et *Streptococcus suis*, on constate qu'elles possèdent une fraction élevée des gènes du génome coquille ($> 28\%$) mais une faible tendance-γ. Deux d'entre elles, (*B. longum* et *E. faecium*), se trouvent dans l'intestin des mammifères et la troisième (*S. suis*) dans les voies respiratoires supérieures du porc.

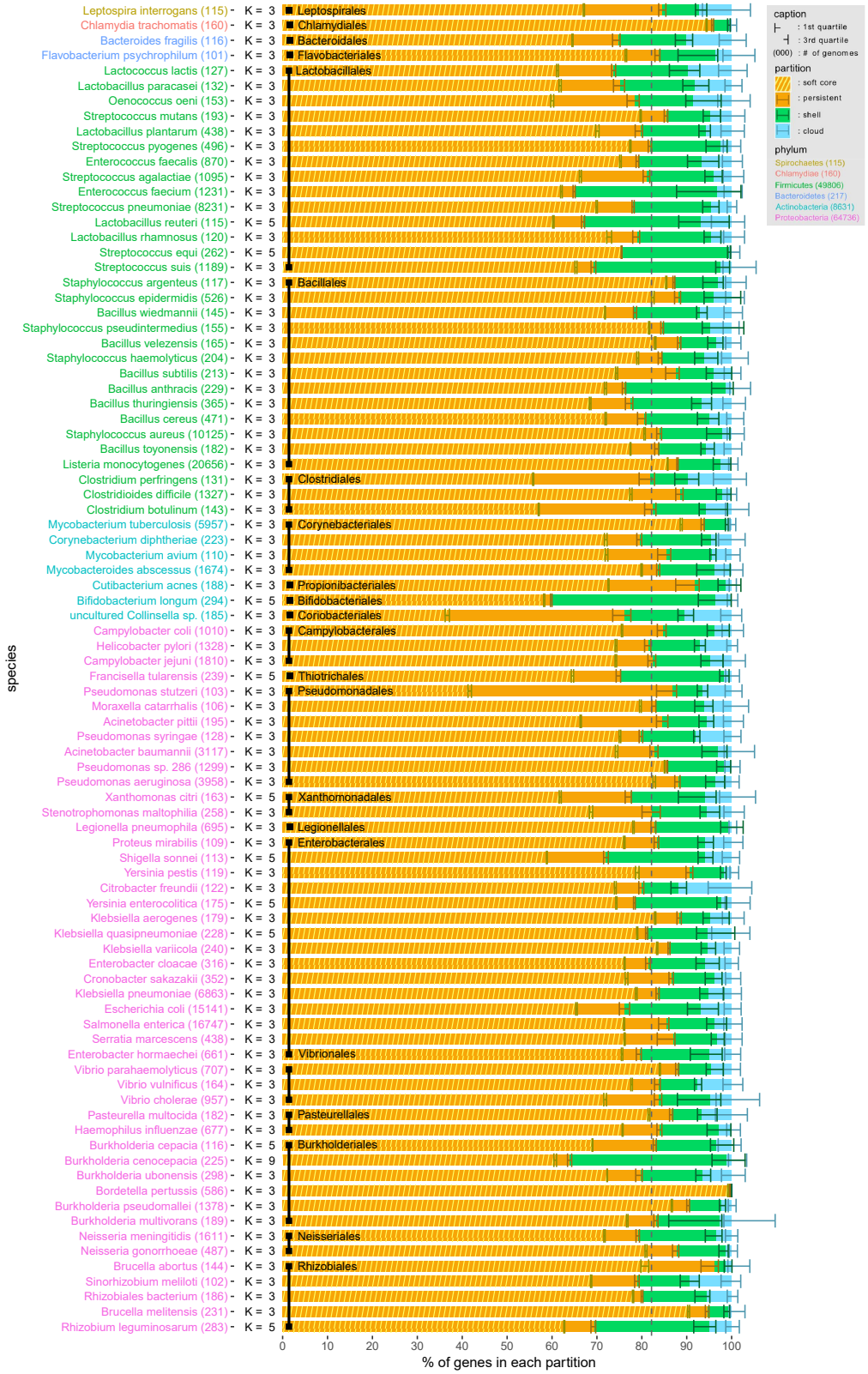


FIGURE 9.7 – Représentation indiquant la part relative des gènes, dans les génomes de 88 espèces, appartenant à des familles assignées à chaque partie du pangénome : le génome persistant (« persistent »), le génome coquille (« shell »), le génome nuage (« cloud »). Le génome cœur assoupli (avec seuil de présence à 95%) est également superposé sur le génome persistant. Un trait vertical pointillé indique la proportion médiane de gènes persistants dans les génomes (82.3%).

Elles diffèrent des autres espèces associées à un hôte (dans notre ensemble de données) qui correspondent principalement à des pathogènes humains (bactéries des genres *Corynebacterium*, *Neisseria*, *Streptococcus*, *Staphylococcus* et *Chlamydia*) et qui ont une faible fraction de génome coquille ($< 20\%$ des génomes). Il est possible que ces trois espèces (*B. longum*, *E. faecium* et *S. suis*) se soient spécialisées dans leur niche écologique tout en conservant un bassin important et stable de gènes coquilles pour leur adaptation aux stress environnementaux. Une analyse plus poussée serait nécessaire pour confirmer cette hypothèse. Comme évoqué, ce type d'information manque à cette échelle. On peut toutefois citer l'article en pré-publication des auteurs de MAISTRENKO et al., 2019 qui propose une analyse plus globale. Cette étude montre notamment que des environnements comme le sol favorisent l'expansion du pangénome, tandis que les habitats associés à l'hôte entraînent sa réduction. Comme on le voit cela ne semble pas être le cas pour les trois espèces précitées. Enfin les auteurs montrent que l'environnement semble avoir un impact plus important sur les caractéristiques du pangénome que la phylogénie. Intéressons-nous donc à ce lien entre la variabilité du pangénome et le signal phylogénétique.

9.1.5 Liens entre la phylogénie et les schémas de présence du génome coquille

Nous souhaitons estimer l'importance du lien entre la phylogénie de l'espèce et les schémas de présence/absence des familles du génome coquille. Le but étant ainsi de savoir quelle part de la variabilité pangénomique des espèces est liée à la variabilité héritée verticalement et quelle part ne semble pas y être reliée et est donc potentiellement explicable par des échanges récents par transfert horizontal. À cette fin, nous avons effectué pour chacune des 88 espèces des corrélations de Spearman « case à case » entre deux matrices de dissimilarité qui correspondent à :

— Une matrice de dissimilarité générée sur la base des distances de Jaccard-Needham entre les schémas de présence/absence des familles de gènes coquilles. Pour rappel, la distance de Jaccard-Needham divise le nombre de valeurs 1 aux mêmes positions entre deux vecteurs binaires (de même longueur) par le nombre de positions ayant pour valeur 1 dans au moins un des deux vecteurs.

— Une matrice de distance génomique obtenue par des comparaisons MASH entre paires de génomes (ONDOV et al., 2016). Cette métrique est une approximation de la distance phylogénétique. En effet pour des génomes dont la distance se situe à proximité du rang d'espèce, les distances MASH s'avèrent être une bonne estimation des distances évolutives (CRISCUOLO, 2019). Dans notre cas, nous utilisons 5 000 « sketchs » MASH pour calculer cette distance. De plus, des tests ont été effectués nous permettant de montrer que les distances MASH calculées sur des génomes entiers produisent les mêmes résultats que des distances MASH calculées sur le génome persistant, confirmant ainsi que la distance MASH ne semble pas être affectée par la variabilité pangénomique.

Cette corrélation a notamment été examinée par rapport à la fraction des familles de gènes qui font partie du génome coquille pour chaque espèce (voir la figure 9.8 page suivante). Nous observons que les espèces ayant une fraction élevée de génome coquille ($> 20\%$ de leur génome) ont également une structure du génome coquille qui s'explique principalement par la phylogénie. En ce qui concerne ces cas, les schémas de présence/absence du génome coquille sont fortement

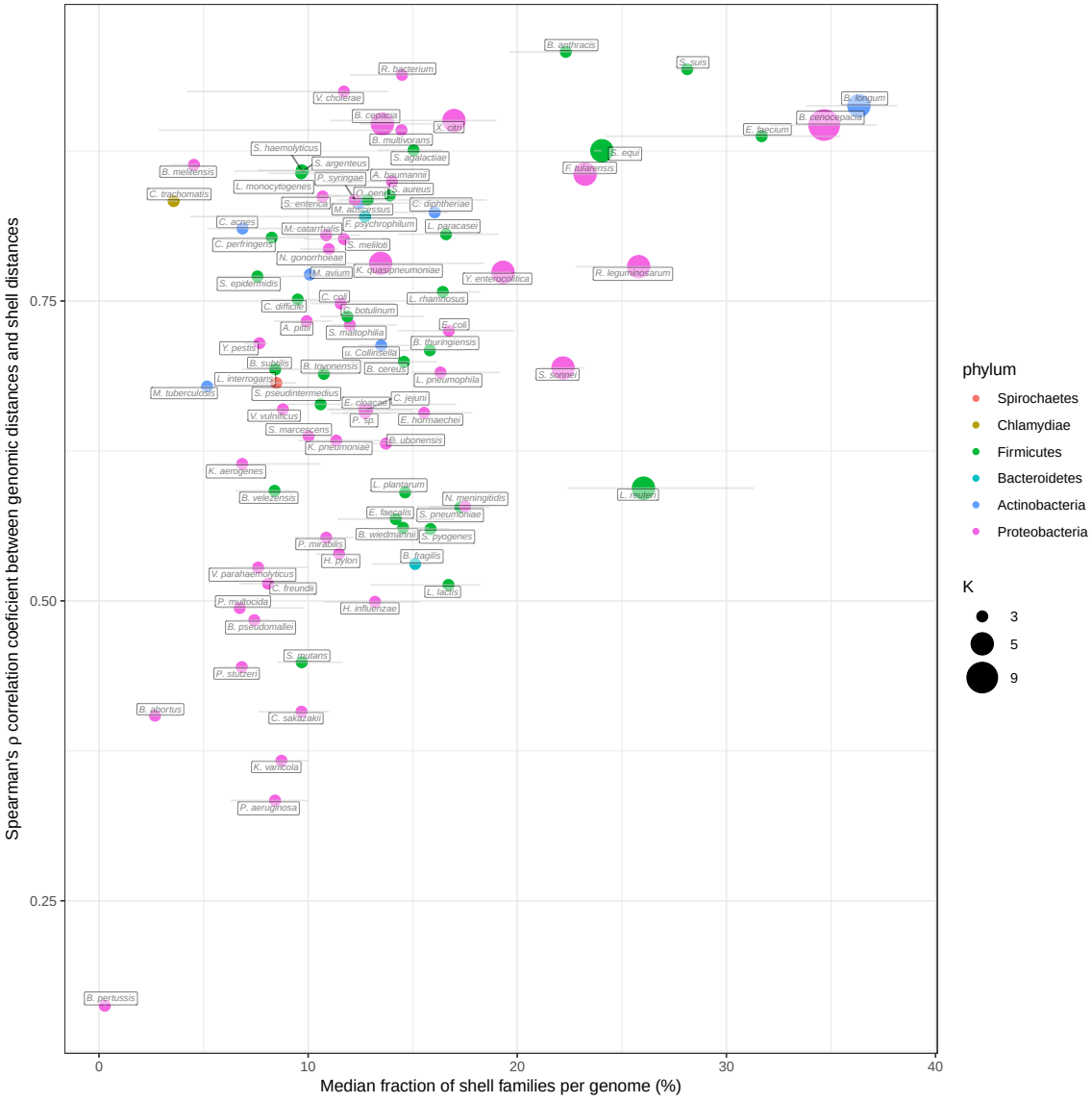


FIGURE 9.8 – Représentation reproduite depuis GAUTREAU et al., 2020 illustrant les coefficients de corrélation ρ de Spearman (entre les schémas de présence/absence du génome coquille et des distances génomiques MASH) par rapport à la fraction de génome coquille par génome. Les résultats pour les 88 espèces les plus abondantes dans GenBank sont représentés. Les barres d'erreur montrent les intervalles interquartiles de la fraction de génome coquille. La taille du point correspond au nombre de partitions (K) utilisées et les points sont colorés par phylum.

corrélés avec les distances MASH, avec un coefficient ρ de Spearman > 0.75. De plus, ces espèces correspondent généralement à celles pour lesquelles PPanGGOLiN prédit un nombre de partitions $K > 3$. Ainsi, leur génome coquille est plus hétérogène entre les sous-clades qui se structurent en plusieurs sous-parties propres. Cela va à l'encontre du cas général des espèces où une seule partie de génome coquille est détectée ($K = 3$) et qui présentent donc un génome coquille moins structuré, soit de nombreux échanges de gènes entre les souches de différentes lignées dans l'espèce. Dès lors, cette structuration en plusieurs parties pourrait ainsi être le révélateur d'un possible phénomène de spéciation allopatrique en cours.

Parmi les neuf espèces ayant une grande fraction de génome coquille (en excluant *B. anthracis* en raison des erreurs d'assignation taxonomique mentionnées précédemment), seules deux d'entre elles (*Shigella sonnei* et *Lactobacillus reuteri*) présentent une corrélation relativement faible de la structure de leur génome coquille avec la phylogénie (voir figure 9.8 page précédente). Pour *S. sonnei*, ceci pourrait s'expliquer par un nombre élevé de pertes de gènes dans le génome coquille de cette espèce. Ces pertes de gènes pourraient résulter d'une convergence évolutive provoquée par des transposons ainsi que le suggèrent les auteurs de cet article en pré-publication (HAWKEY et al., 2019). Pour le cas de *L. reuteri*, ces bactéries colonisent le tractus gastro-intestinal d'une grande variété d'espèces de vertébrés et se sont diversifiées en clades phylogénétiques distincts reflétant davantage l'hôte où les souches ont été isolées que leur provenance géographique (OH et al., 2010). Comme illustré sur la figure 9.9 page suivante, le génome coquille de *L. reuteri* contient plusieurs schémas de présence/absence qui ne sont que partiellement expliqués par la phylogénie de l'espèce ($\rho_{Spearman} = 0.59$, $R_{Pearson} = 0.80$). En effet, même si la phylogénie semble le principal facteur structurant le génome coquille, nous observons des groupes de familles de gènes présents dans des souches de lignées distinctes qui pourraient contenir des facteurs d'adaptation à un même hôte. En revanche si l'on s'intéresse à une autre espèce du tractus gastro-intestinal comme *B. longum*, on constate que la structure de son génome coquille dépend fortement des distances phylogénétiques ($\rho_{Spearman} = 0.91$, $R_{Pearson} = 0.92$). Ainsi, on observe une nette délimitation des souches évoluant chez un hôte adulte et chez le porc (*B. longum subsp. longum* et *B. longum subsp. suis*) d'une part et chez l'enfant d'autre part (*B. longum subsp. infantis*) comme on peut le voir sur la figure 9.10 page 171. Dès lors, il y aurait une séparation à la fois phylogénétique et pangénomique en 3 sous-espèces spécialisées dans 3 environnements. Cela pourrait s'expliquer par le fait que les différentes sous-espèces posséderaient des voies métaboliques dédiées au catabolisme des différents polysaccharides liés au régime alimentaire de leur hôte respectif (Homme adulte et enfant, porc). Cela a notamment été établi pour la sous-espèce *B. longum subsp. infantis* qui se serait spécialisée dans le catabolisme des sucres provenant du lait consommé abondamment par l'enfant (UNDERWOOD et al., 2015). À l'inverse pour des espèces comme *Klebsiella variicola*, la corrélation avec la phylogénie est bien plus faible ($\rho_{Spearman} = 0.36$, $R_{Pearson} = 0.46$), ce qui s'illustre sur la figure 9.11 page 172 par une matrice de présence/absence du génome coquille bien plus désordonnée par rapport à la phylogénie. Enfin pour terminer, il faut négliger le résultat indiquant une absence de corrélation pour une espèce comme *B. pertussis* car cette dernière possède un génome coquille composé de 24 familles de gènes seulement.

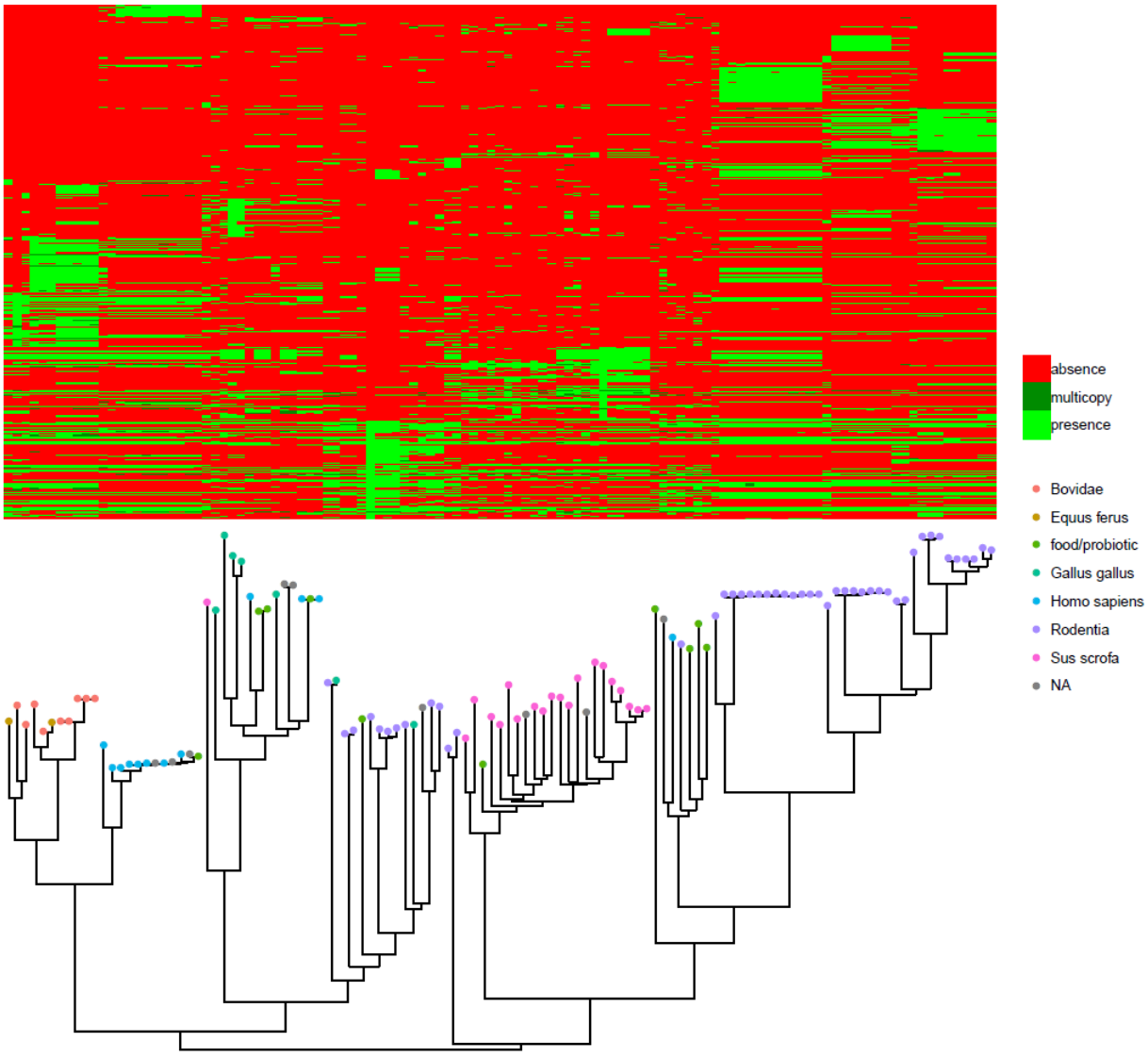


FIGURE 9.9 – Représentation reproduite depuis GAUTREAU et al., 2020. Matrice de présence/absence du génome coquille de *B. longum* ordonné par un arbre obtenu par l'algorithme « Neighbor Joining » (SAITOU et al., 1987) se basant sur la matrice de distance MASH. Les feuilles de l'arbre sont colorées par l'environnement (hôte ou origine) d'isolation. Ces informations sont extraites depuis les métadonnées de GenBank (« host » et « isolation source »).

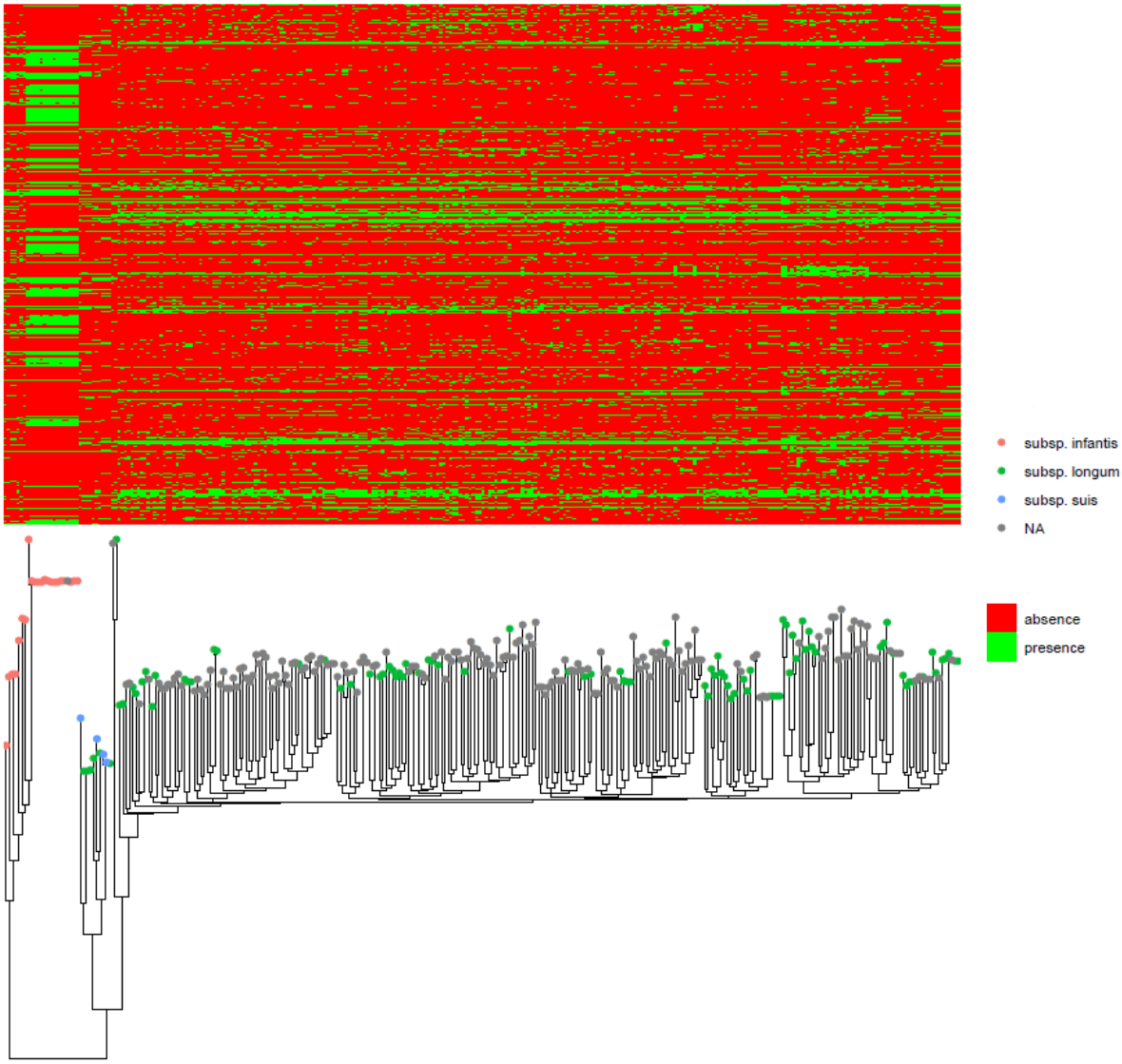


FIGURE 9.10 – Représentation de la matrice de présence/absence du génome coquille de *B. longum* ordonnée par un arbre obtenu via l'algorithme « Neighbor Joining » (SAITOU et al., 1987) se basant sur la matrice de distance MASH. Les feuilles de l'arbre sont colorées par sous-espèce l'appartenance à la sous-espèce *B. longum subsp. infantis*, *B. longum subsp. suis* ou *B. longum subsp. longum*. Cette information a été obtenue à l'aide de la base de données GTDB (version R04-RS89, PARKS et al., 2018). Les points rouges indiquant les génomes n'appartenant pas à *B. longum subsp. infantis* peuvent inclure des génomes absents de GTDB.

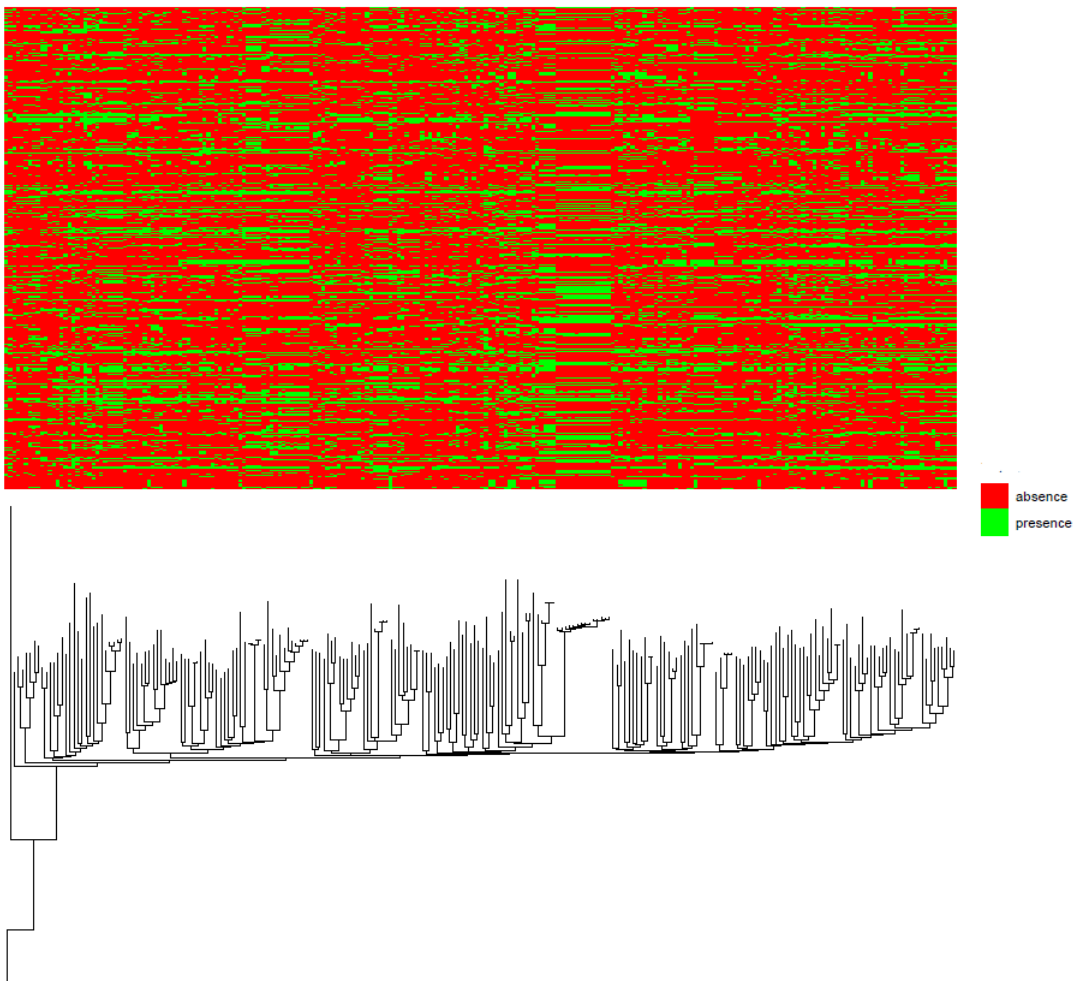


FIGURE 9.11 – Représentation d'une matrice de présence/absence du génome coquille de *K. variicola* ordonnée par un arbre obtenu via l'algorithme « Neighbor Joining » (SAITOU et al., 1987) se basant sur la matrice de distance MASH.

9.2 Application sur des génomes issus de métagénomes

9.2.1 Rappels sur la notion de métagénomique

Pour étudier les génomes d'espèces difficiles à isoler et à cultiver, il existe un ensemble de méthodes, dites « métagénomiques ». Ces dernières permettent de séquencer le contenu génomique de tout un échantillon sans nécessité d'isoler les organismes y étant présents. L'inconvénient de ces méthodes est que tout le contenu génomique présent dans l'échantillon est alors mélangé lors du séquençage. Il faut donc reconstituer les génomes. Une technique très employée pour y parvenir consiste tout d'abord dans l'assemblage de toutes les lectures d'ADN obtenues avec des assembleurs adaptés (comme par exemple metaSPAdes, NURK et al., 2017). Cependant, les contigs issus de ces assemblages restent très fragmentés et mélangés, de sorte que pour les regrouper en des ensembles issus des mêmes génomes, nous allons réaligner les lectures d'ADN sur ceux-ci. Grâce à cela, on peut par la suite classifier les contigs selon deux critères, d'une part, leur signature fréquentielle en k-mers (généralement des tétranucléotides) qui est spécifique de chaque espèce, et d'autre part, la profondeur de couverture des contigs qui est reliée à l'abondance du génome dans l'échantillon. Nous obtenons ainsi des groupes de contigs homogènes en composition nucléotidique et en couverture, dès lors susceptibles d'être considérés comme provenant du même génome à l'origine. Les outils les plus connus pour effectuer ce regroupement sont CONCOCT (ALNEBERG et al., 2014) et MetaBAT (KANG et al., 2015). Cependant, comme ces algorithmes automatiques peuvent parfois faire des erreurs, il existe aussi des outils comme Anvi'o (EREN et al., 2015) pour effectuer des regroupements manuels assistés par une interface graphique très ergonomique. Au final, l'inconvénient de cette approche est qu'elle fournit très souvent des génomes incomplets et particulièrement fragmentés en comparaison des génomes obtenus à partir d'isolats. On appelle généralement ces génomes assemblés depuis des métagénomes des MAGs (Metagenome-Assembled Genomes).

9.2.2 Protocole de la comparaison entre les MAGs et les isolats de GenBank

Nous pensons que l'approche PPanGGOLiN grâce à sa structure graphique peut être résiliente aux incomplétudes des MAGs. En effet, on peut s'attendre à ce que le partitionnement du pangénome sur la base de schémas de présence/absence rende notre outil bien plus résilient aux absences de gènes dans les données génomiques que les méthodes usuelles basées sur des seuils de présence comme le génome cœur assoupli.

Pour tester cette hypothèse, nous avons construit les pangénomes sur la base de MAGs regroupés au niveau taxonomique de l'espèce et nous souhaitons les comparer aux pangénomes que nous avons construit à partir des isolats de GenBank. Ces MAGs proviennent du récent article écrit par les auteurs de PASOLLI et al., 2019. Cette étude a aggloméré 154 723 MAGs en 4 930 groupes de génomes de la même espèce (Species Genome Bins : SGB) sur la base d'un total de 9 428 échantillons (issus de 13 études) provenant de microbiomes humains, principalement de la matière fécale, mais aussi de prélèvements sur la bouche, la peau et le vagin. Ces groupes de génomes rassemblent des MAGs ayant une diversité génétique de 5%, c'est-à-dire supposés appartenir à la même espèce. Pour identifier le nom des SGBs, les génomes représentatifs des espèces connues

dans RefSeq (sous-ensemble de la base de données GenBank) ont été ajoutés aux MAGs avant le regroupement en SGB. L'espèce à laquelle appartient chaque SGB est ensuite déterminée par rapport au génome représentatif associé à celui-ci. Si aucun génome représentatif n'y est présent, le SGB est considéré comme une nouvelle espèce potentielle. Cependant, avec leur méthode de regroupement hiérarchique des SGBs, les auteurs ont parfois plusieurs SGBs nommés avec le même nom d'espèce. Pour pouvoir faire la comparaison avec les 439 espèces obtenues depuis GenBank, nous avons regroupé ensemble les SGBs ayant le même nom d'espèce. Nous avons cependant constaté des erreurs d'assignation taxonomique pour certains SGB, en particulier pour deux espèces *Blautia obeum* et *Chlamydia trachomatis*. En conséquence, nous les avons exclus de nos analyses. Nous n'avons pas non plus effectué l'étape de contrôle de la qualité des génomes puisque celle-ci a été effectuée par les auteurs. Nous avons ensuite généré les pangénomes avec PPanGGOLiN en suivant rigoureusement la même procédure que celle utilisée précédemment pour les génomes provenant de GenBank. Le seul paramètre qui diffère est le nombre de parties K. En effet, il est plus difficile de détecter le nombre de parties sur des génomes incomplets. Le génome persistant peut en effet, présenter beaucoup d'absences dans les MAGs, de sorte, qu'une fraction de ce dernier risque d'induire une partie additionnelle au moment de l'estimation du nombre de parties et donc d'être considéré comme du génome coquille. C'est pourquoi nous avons choisi de fixer le nombre de parties à $K = 3$.

Pour effectuer notre comparaison, nous allons imposer la même contrainte que pour GenBank en filtrant les SGBs ayant moins de 15 génomes. En regardant les espèces communes aux deux ensembles de données, nous obtenons un total de 78 espèces (correspondant à 151 SGBs) pouvant être comparées. Si le nombre est si faible en comparaison des 4 930 SGBs générés par les auteurs de PASOLLI et al., 2019, c'est qu'une très faible proportion des organismes présents dans le microbiome est en mesure d'être isolée en utilisant les protocoles microbiologiques actuels ; dès lors, ils sont faiblement représentés dans une banque de données comme GenBank.

Pour effectuer notre comparaison, nous mesurons la quantité de familles de gènes assignées au génome persistant pour chacune des 78 espèces en fonction du mode d'obtention des données (MAGs contre GenBank). Ainsi, pour chaque espèce, les séquences centrales des familles de gènes issus des MAGs sont alignées sur celles des familles de GenBank (voir la sous-section 7.2 page 115, pour plus de détails concernant la manière de déterminer la séquence protéique centrale au sein des familles de gènes). En alignant avec MMseqs2 (STEINEGGER et al., 2017) ces séquences centrales issues de chaque espèce provenant de GenBank sur leur équivalent dans les MAGs, nous pouvons ainsi calculer le nombre de familles communes entre les différentes parties du pangénome. Nous ne retenons ici que l'alignement ayant le meilleur score de similarité entre les MAGs et les isolats de GenkBank. De plus, un alignement est validé uniquement si celui-ci a un taux d'identité de séquence d'au minimum de 80% ainsi que s'il couvre au minimum 80% de la séquence cible. Pour ce qui est du génome persistant, nous nous attendons à une forte bijection entre les familles de gènes provenant des MAGs et celles provenant de GenBank. À l'inverse, il n'y a aucune raison pour que les génomes coquilles soient identiques entre les données provenant de GenBank et les MAG, sachant les différences entre les environnements considérés. Néanmoins, il semble intéressant de quantifier ces différences afin de mesurer la quantité de diversité apportée par les MAGs.

9.2.3 Résultat de la comparaison des génomes persistants

D'après nos résultats illustrés sur la figure 9.12, nous observons que la taille du génome persistant estimée des MAGs est similaire à celle des génomes de GenBank pour la plupart des espèces. Le terme « similaire » indique ici que pour chaque espèce la différence entre le plus grand et le plus petit génome persistant (en nombre de familles de gènes) est de moins de 20%, ce qui correspond à une valeur de « fold-change » <1.2.

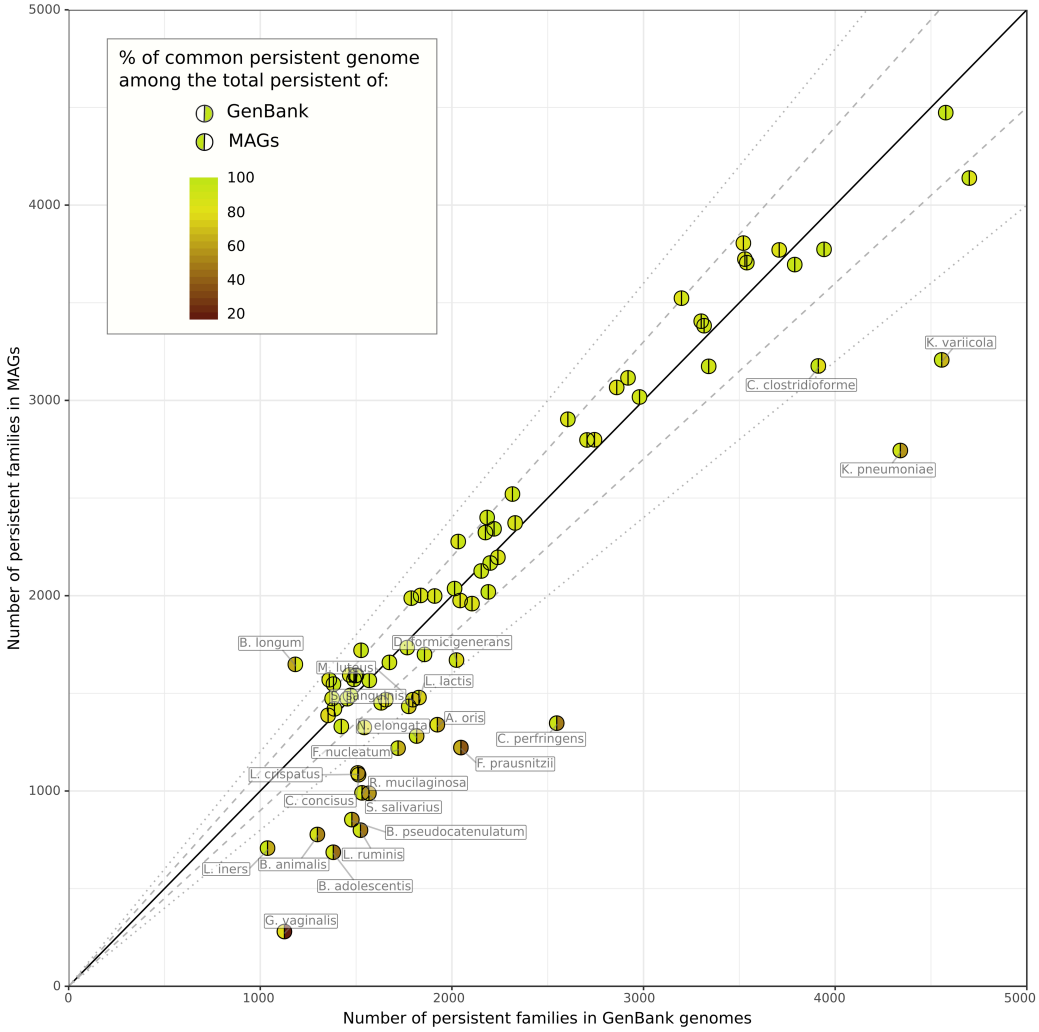


FIGURE 9.12 – Représentation reproduite depuis GAUTREAU et al., 2020 illustrant la comparaison du génome persistant de 78 espèces entre les génomes MAGs et les génomes de GenBank. Les couleurs des hémisphères donnent le pourcentage de familles de gènes persistants communes parmi l'ensemble des familles persistantes des MAGs (hémisphère gauche) ou GenBank (hémisphère droit). Les lignes pleines, les pointillés longs et les pointillés fins indiquent, respectivement, l'identité en nombre de familles de gènes, une gamme de 10% de différence et une gamme de 20% de différence.

Pour 55 des 78 espèces, le fold-change de la taille du génome persistant est inférieur à 1.2 et ≈90% (écart-type de ≈5.21%) de son contenu est commun entre les génomes MAGs et de GenBank. Pour être très précis, cela correspond à 90.63% (écart-type de 5.21%) de familles communes par rapport à GenBank et 89.59% (écart-type de 5.11%) de familles communes par rapport aux MAGs.

Les 23 autres espèces présentent cependant des différences plus importantes. En ce qui concerne ces cas, le génome persistant est fortement sous-estimé car seulement 60.12% (écart-type de 15.10%) du génome persistant de GenBank est présent dans les MAGs. Pour ces 23 espèces, la méthode PPanGGOLiN a manqué une fraction du génome persistant en raison notamment du caractère incomplet des MAGs. En effet, les familles de gènes manquantes sont principalement classées dans le génome coquille des MAGs ; ce dernier contenant 31.36% (écart-type de 11.65%) de familles du génome persistant de GenBank. Néanmoins, 88.8% (écart-type de 9.42%) des familles assignées au génome persistant des MAGs correspondent effectivement à des familles présentes dans GenBank, ce qui signifie que PPanGGOLiN a correctement attribué ces familles, même si celui-ci reste incomplet pour ces 23 espèces. Cependant, deux espèces (*Bifidobacterium longum* et *Faecalibacterium prausnitzii*) ont moins de 75% de leurs familles MAGs persistantes en commun avec celles de GenBank. Pour *B. longum*, cela pourrait s'expliquer par le fait que les MAGs ont principalement été obtenus à partir d'échantillons humains adultes alors que les souches de cette espèce dans les banques de données proviennent d'une gamme d'hôtes bien plus large (notamment depuis des nourrissons et des porcs). Cela signifie donc que le génome persistant des MAGs peut contenir des gènes supplémentaires liés à l'hôte. D'ailleurs, 412 familles de gènes du génome persistant des MAGs de l'espèce *B. longum* (25.11% du génome persistant total des MAGs) se trouvent dans le génome coquille de GenBank, confirmant ainsi notre hypothèse. Enfin, pour *F. prausnitzii*, les différences s'expliquent plutôt par une mauvaise estimation du génome persistant en raison du faible nombre de génomes considérés dans GenBank (17 génomes contre 4232 MAGs).

9.2.4 Quels résultats aurait-on obtenu avec le génome cœur assoupli ?

Certes les résultats produits par PPanGGOLiN sont imparfaits mais pour remettre cela en perspective, nous avons calculé les résultats que nous aurions obtenus en utilisant le génome cœur assoupli à la place du génome persistant détecté par PPanGGOLiN. En effet, avec un seuil de présence classique à 95%, on obtient un génome cœur assoupli de ≈98 familles de gènes en moyenne sur les MAGs des 78 espèces étudiées. De plus, seulement 4 espèces possèdent un génome cœur assoupli ayant plus de 500 familles de gènes. En descendant à 90% de présence, on a en moyenne ≈410 familles de gènes (écart-type de ≈415) dont seulement 7 espèces ont plus de 1000 familles de gènes. Ces chiffres semblent manifestement trop faibles, ce qui disqualifie à l'évidence l'approche génome cœur assoupli sur ce type de donnée.

On pourrait objecter qu'il est encore possible d'abaisser davantage ce seuil, mais pour information, sur le génome persistant (qui lui-même rappelons-le, semble déjà sous-estimé pour 23 espèces), on constate en moyenne une fréquence de présence de 78.8% avec un écart-type de 17.5%. De telles valeurs compliquent la distinction entre le génome coquille et le persistant et montrent les limites des méthodes se basant sur la définition d'un seuil basé sur la fréquence des familles de gènes.

9.2.5 Quelles différences du point de vue du génome coquille ?

Pour ce qui est du génome coquille et a fortiori du génome nuage des MAGs, il faut être prudent. En effet, nous l'avons dit, les assemblages sont regroupés du point de vue de leur composition et de leur abondance. Or les gènes récemment acquis possèdent des taux de bases GC souvent très différents du reste du génome. De plus, si un échantillon contient un mélange de plusieurs souches de la même espèce, certes le génome persistant aura une abondance stable mais le génome coquille risque d'avoir une abondance bien plus variable. Dès lors, les algorithmes de regroupement, évoqués à la sous-section 9.2.1 page 173, peinent généralement à regrouper le génome coquille des différentes souches et surtout le génome nuage avec le reste du génome des MAGs.

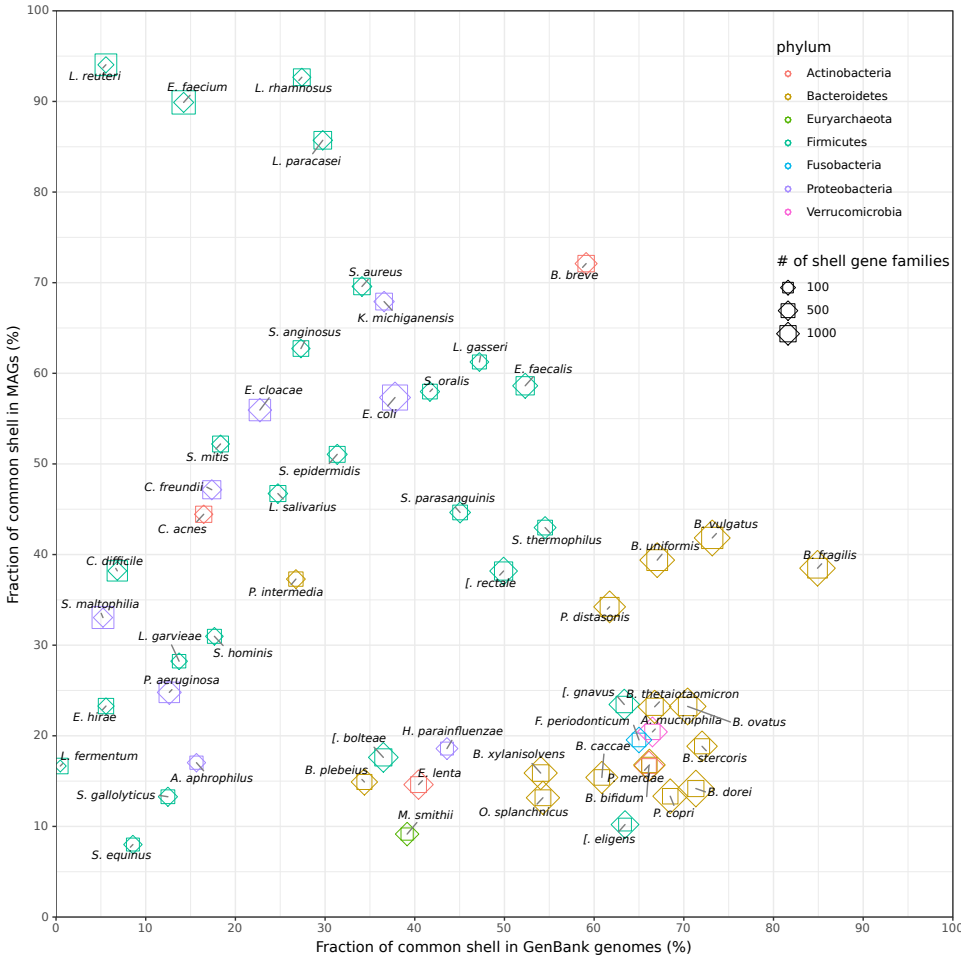


FIGURE 9.13 – Représentation reproduite depuis GAUTREAU et al., 2020 illustrant la comparaison du génome coquille de 55 espèces entre les génomes MAGs et les génomes de GenBank. L'axe des abscisses représente le pourcentage de génome coquille en commun par rapport aux génomes de GenBank tandis que l'axe des ordonnées représente ce pourcentage par rapport aux MAGs. Les diamants représentent les MAGs alors que les carrés représentent les génomes de GenBank.

Ces réserves faites, voyons les résultats que nous obtenons. Pour explorer cette diversité au sein des pangénomes, nous avons comparé le génome coquille des génomes de GenBank et des MAGs pour les 55 espèces ayant des génomes persistants similaires. Il est intéressant de noter que pour l'ensemble des 55 espèces, nous n'avons pu observer qu'un chevauchement partiel du contenu

du génome coquille des MAGs avec celui de GenBank (figure 9.13 page précédente). En effet, étant donné que les MAGs ont été obtenus à partir d'un environnement spécifique (c'est-à-dire le microbiome humain), la diversité de GenBank n'est pas entièrement capturée par les MAGs. C'est particulièrement le cas pour la plupart des Firmicutes et des Protéobactéries dont les espèces se retrouvent dans de multiples écosystèmes. Inversement, la plupart des MAGs du phylum des Bacteroidetes couvrent plus de la moitié de la diversité de GenBank tout en contenant une grande fraction de gènes coquilles spécifiques (c'est-à-dire que moins de 45 à 25% des familles sont représentées dans le génome coquille de GenBank). Comme l'avaient déjà rapporté les auteurs PASOLLI et al., 2019, cela confirme que les MAGs améliorent considérablement l'estimation de la diversité génétique des Bacteroidetes, ces acteurs-clés du microbiome intestinal.

9.2.6 Conclusion sur le partitionnement du pangénome des MAGs

En résumé, nous avons montré que PPanGGOLiN est capable de fournir une estimation du génome persistant et du génome coquille en utilisant uniquement des MAGs et ce malgré que ces derniers manquent un nombre significatif de gènes. Cela permet donc d'envisager de nombreuses analyses supplémentaires sur des espèces encore non cultivées et donc dépourvues de génomes de référence.

Sur la base du génome persistant, on peut notamment imaginer identifier les principales voies métaboliques utilisées par l'espèce. Sachant ce métabolisme de base, il devient ainsi possible de pourvoir aux besoins nutritionnels des souches afin de pouvoir les isoler en laboratoire. Cela autorise des études plus approfondies sur sa biologie mais aussi la possibilité de générer des mutants pour des analyses fonctionnelles ou d'intérêt industriel.

Enfin, l'identification des régions génomiques issues de transferts entre des organismes par l'étude directe de métagénomes reste un problème difficile (DOUGLAS et al., 2019). Grâce à l'utilisation du graphe, couplé au partitionnement des familles de gènes, il devient possible en se basant sur les résultats de PPanGGOLiN d'explorer les MAGs à la recherche des régions d'intérêt qu'ils possèdent. De méthodes supplémentaires restent toutefois nécessaires afin d'extraire ces régions. Pour ce faire, une méthode se basant sur les résultats générés par PPanGGOLiN est présentée au chapitre 11 page 189.

Partie

III

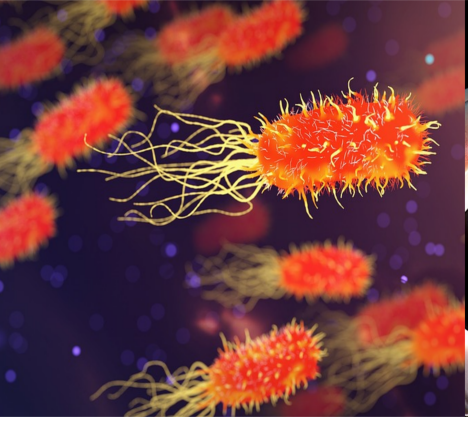
Applications basées sur la méthode PPanGGOLiN

10 Vers la constitution d'une banque de pangénome 181

10.1 L'idée d'une banque de graphes de pangénomes partitionnés
10.2 Limites des ressources en ligne existantes
10.3 Obtenir les génomes de GenBank
10.4 Filtrer les erreurs dans la banque de données GenBank
10.5 Une interface pour explorer les graphes de pangénomes partitionnés

11 Détecter les régions de plasticité génomique 189

11.1 Limites des approches actuelles de détection des RGPs
11.2 panRGP : utiliser un pangénome partitionné pour détecter les GIs
11.3 Résultats de la méthode panRGP
11.4 Conclusion sur la méthode panRGP





Chapitre 10 : Vers la constitution d'une banque de pangénome

10.1 L'idée d'une banque de graphes de pangénomes partitionnés

Pour générer les pangénomes des 439 espèces que nous avons étudiées précédemment, nous avons conçu un ensemble de méthodes automatisées. À l'origine, ces programmes étaient conçus pour être utilisés une unique fois afin d'obtenir les résultats présentés au chapitre 9 page 151. Pourtant, nous nous sommes demandé s'il était bien raisonnable, sachant ces méthodes en place, de ne les utiliser qu'une seule fois. La réponse fut évidemment non. Il serait préférable de compléter les pangénomes au fur et à mesure que les génomes s'accumulent dans les banques de données. L'objectif est ainsi de proposer une ressource automatiquement mise à jour et rendue disponible à la communauté scientifique. Pour ne pas laisser penser que j'ai mené ce projet tout seul, je précise que le travail présenté dans ce chapitre est basé sur des programmes que j'avais certes conçus moi-même pour la plupart mais qui ont été très profondément modifiés, améliorés et réécrits par quatre ingénieurs (Mathieu Gachet, Laura Burlot, Mathieu Dubois et Rémi Planel) et un doctorant (Adelme Bazin). Merci à eux. Les développements réalisés utilisent le cadre informatique de Snakemake pour la gestion des flux d'exécution (KOSTER et al., 2012) et le logiciel BioMAJ (FILANGI et al., 2008) pour la synchronisation des données de la banque GenBank. Actuellement, un étudiant en alternance (Paul Amours) a repris le flambeau de ce projet et je lui adresse mes encouragements.

10.2 Limites des ressources en ligne existantes

Outils	Avantages	Limites
La plateforme MicroScope VALLENET et al., 2019	— Libre choix des génomes — Deux taux d'identité (50% et 80%) pour générer les familles de gènes — Lien avec le reste de la plateforme — Possibilité de générer les courbes de raréfaction	— Limité à 300 génomes par pangénome — Se base uniquement sur la dichotomie (génome cœur/génome accessoire) — Paramètre de couverture fixé à 80% pour générer les familles de gènes
panWeb PANTOJA et al., 2017	— Libre choix des génomes — Libre choix du taux d'identité/de couverture pour générer les familles de gènes — Possibilité d'ajouter ses propres génomes — Possibilité de générer les courbes de raréfaction	— Pas de temps réel donc présence d'une liste d'attente (assez long ≈1 heure)
panX DING et al., 2018	— Énormément d'informations disponibles dans une interface riche et ergonomique — Seuil de présence ajustable pour la dichotomie génome cœur assoupli/génome accessoire assoupli	— Ne fonctionne que sur des pangénomes précalculés
PanGFR-HM CHAUDHARI et al., 2018	— Énormément d'informations disponibles — Beaucoup de génomes issus du microbiome avec des métadonnées sur les sites d'isolation — Libre choix du taux d'identité pour générer les familles de gènes — Possibilité de comparer jusqu'à quatre pangénomes entre eux — Possibilité de générer les courbes de raréfaction	— Limité à 200 génomes par pangénome — Peu de génomes disponibles pour certaines espèces (seulement 32 génomes pour *E. coli* par exemple) — Pas de choix du paramètre de couverture pour générer les familles de gènes

TABLE 10.1 – Comparaison des quatre principaux outils permettant de facilement explorer des pangénomes en ligne.

À ce jour, il existe plusieurs sites Web permettant d'explorer des pangénomes. Même si chaque ressource a ses spécificités et ses limites ainsi que détaillé dans la table 10.1, on constate globalement que :

— Aucune ressource n'utilise systématiquement toutes les données présentes dans GenBank.
— Aucune ressource ne propose d'interface pour la programmation d'application (Application Programming Interface : API).
— Aucune ressource ne va au-delà de la dichotomie génome cœur assoupli/génome accessoire assoupli.
— Et bien évidemment, aucune ne propose d'explorer en ligne le graphe de pangénome des espèces proposées.

L'objectif est donc de concevoir une ressource, nommée PanGBank, utilisant tous les génomes disponibles dans les banques de données, en filtrer les erreurs, puis de construire le pangénome, les métriques associées, le graphe de pangénome partitionné et enfin de permettre sa visualisation.

10.3 Obtenir les génomes de GenBank

Télécharger des génomes depuis une banque de données comme GenBank, cela peut paraître simple. C'était mon a priori au début en tout cas. J'utilisais notamment un outil très pratique appelé ncbi-genome-download (BLIN, 2019).

Mais pour gérer une banque de données dont il faut assurer la mise à jour régulière dans la durée, il nous fallait des outils plus robustes. Pour cela, deux ingénieurs, Laura Burlot et Mathieu Dubois, ont implémenté le téléchargement de génomes depuis les serveurs du NCBI grâce à l'outil BioMàJ (FILANGI et al., 2008) qui permet de réaliser des mises à jour incrémentales et autorise des pré et post-traitements.

10.4 Filtrer les erreurs dans la banque de données GenBank

Comme cela a déjà été dit, nous téléchargeons nos données depuis la banque de données de génomes assemblés la plus vaste existante à ce jour, j'ai nommé GenBank. Néanmoins, comme les soumissionnaires de génomes assemblés peuvent déposer leurs génomes dans cette base sans un contrôle approfondi, alors celle-ci peut comporter des erreurs.

10.4.1 Filtrer les erreurs d'assignation taxonomique

La principale des erreurs pouvant perturber la constitution d'un pangénome est l'erreur d'assignation taxonomique. Comme la banque de données GenBank est conçue et gérée par le NCBI, cette ressource unifiée réutilise la classification taxonomique du NCBI (FEDERHEN, 2012). Ainsi, chaque génome est assigné à un identifiant taxonomique généralement au niveau de l'espèce pour les souches isolées mais parfois à un niveau plus précis (sous-espèce). Néanmoins, comme les données ne sont pas vérifiées, il est fréquent que des erreurs d'assignation taxonomique soient faites. Nous l'avons notamment constaté à la sous-section 9.1.3 page 159, pour l'espèce *B. anthracis*. Mais il y a pire, on retrouve par exemple dans la base de données Genbank, l'espèce *Clostridium botulinum* composée de groupes de génomes très différents entre eux. On peut constater sur la figure 10.1 page suivante que deux groupes (correspondant à presque 40% des génomes) ont une valeur d'ANI (définie à la sous-section 2.2.2 page 37) inférieure à 80% par rapport au groupe principal. On rappelle que le seuil habituel pour considérer que deux génomes proviennent de la même espèce est de 94-95% d'ANI. Cela est probablement dû au fait que les biologistes se basent sur la présence du gène de la toxine botulique pour effectuer l'assignation et non sur la totalité du génome. Comme le gène de la toxine est sur un plasmide, d'autres espèces comme *Clostridium sporogenes* et *Clostridium butyricum* sont susceptibles de l'acquérir (E. M. NAWROCKI et al., 2018). Il y a là de très nombreuses erreurs d'assignation dans GenBank, même si *C. botulinum* est a priori l'espèce ayant le plus d'erreurs de ce type à jour. Il est à noter que le NCBI semble avoir pris conscience des difficultés causées par ces erreurs mais bien qu'intéressantes, les solutions qu'il propose reste encore peu matures ou en développement (CIUFO et al., 2018) de sorte que nous devions envisager notre propre solution.

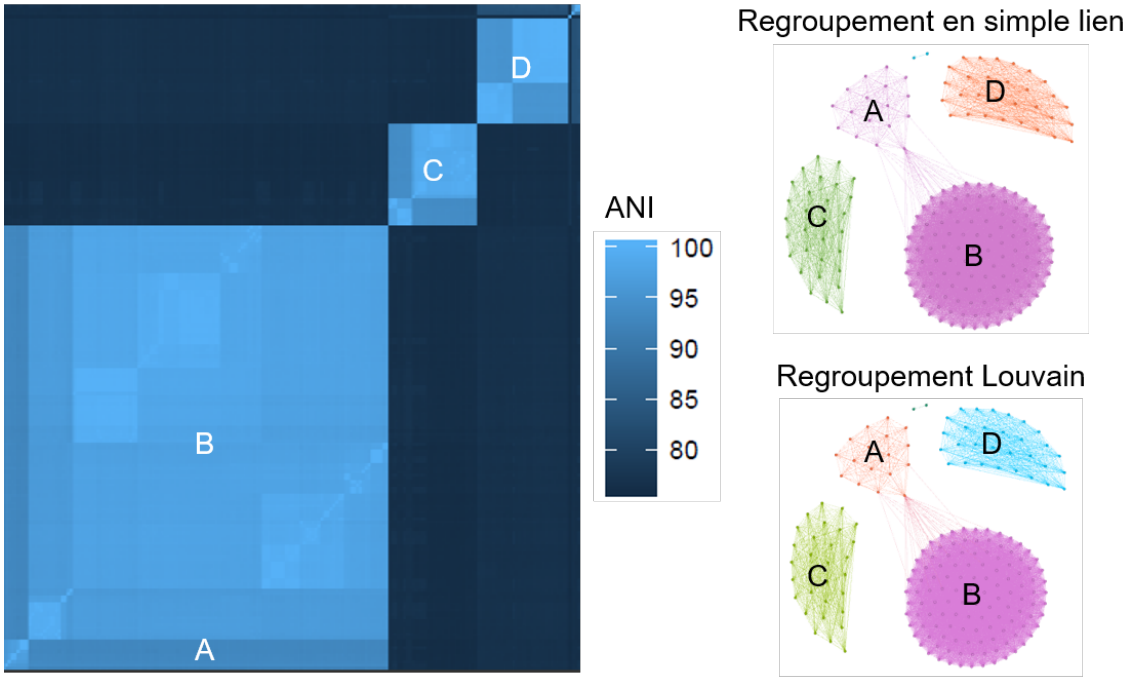


FIGURE 10.1 – Sur la gauche : illustration d'une matrice des calculs de taux d'identité moyens (ANI) entre 200 génomes de *C. botulinium*. Sur la droite : graphe connectant les génomes si leur distance est < 0.06 et les regroupant en 4 composantes connexes ou bien en 5 groupes avec l'algorithme de recherche de communautés Louvain (BLONDEL et al., 2008).

Nous avons donc besoin d'écarter les génomes ne correspondant pas à l'espèce tel qu'indiqué par le NCBI. Au-delà de cela nous avons aussi besoin d'éviter d'intégrer des génomes incomplets (ceux-ci ne posent pas trop de problèmes à la méthode PPanGGOLiN tant qu'ils restent minoritaires) ou des génomes contaminés par d'autres génomes (génomes chimériques correspondant au mélange de deux souches ou deux génomes d'espèces différentes présents dans une même entrée GenBank). L'objectif est également que ce contrôle qualité soit rapide car il doit pouvoir être effectué en routine. Après des essais avec l'outil CheckM (PARKS et al., 2015), nous nous sommes rendus compte que bien que produisant des résultat intéressants (en détectant les génomes contaminés), celui-ci est bien trop lent pour être utilisable en routine sur des centaines de milliers de génomes. En effet, l'étape de détection des gènes marqueurs à utiliser pour le contrôle qualité dans CheckM est basé sur l'outil PPlacer (MATSEN et al., 2010) qui n'est pas parallélisé. Au vu de la croissance du nombre de génomes dans GenBank, le million de génomes procaryotes sera très probablement atteint en 2020. Notre objectif étant de concevoir une méthode pérenne, l'outil CheckM a dès lors été écarté. Nous avons à l'inverse préféré une méthode simple et rapide, à savoir l'utilisation de l'outil de calcul de distance entre génomes MASH (ONDOV et al., 2016) que nous avons évoqué à la sous-section 2.2.2 page 37.

Les résultats des calculs de distance de MASH sont rendus sous forme d'une matrice qui est ensuite transformée en graphes (exemple sur la partie droite de la figure 10.1) où les nœuds sont les génomes et les arêtes les distances MASH $\leqslant 0.06$ (soit $\approx$94% d'ANI). Je précise que nous utilisons 5000 « sketchs » pour effectuer ce calcul de distance. Par la suite, on extrait la composante

connexe (regroupement en simple lien) contenant le plus de génomes avec l'aide de l'outil Silix (MIELE et al., 2011). On considère que celle-ci correspond aux génomes correctement assignés et les autres composantes connexes correspondent à des erreurs d'assignation, de sorte qu'elles doivent être ignorées. L'inconvénient de ne se baser que sur les composantes connexes est qu'il suffit d'un seul lien pour que par transitivité de nombreux génomes soient reliés à la composante connexe principale (exemple du groupe A avec le groupe B sur la figure 10.1 page précédente). Pour tenter d'y remédier, nous avons également tenté d'effectuer un regroupement avec l'algorithme de détection de communautés Louvain (BLONDEL et al., 2008). Néanmoins celui-ci était un peu trop lent pour une méthode à utiliser en routine car cela aurait impliqué d'effectuer la détection de communautés à chaque ajout de génome. Or nous nous attendons à devoir être en mesure de gérer des graphes de ce type ayant des centaines de milliers de génomes. L'autre inconvénient de l'approche en simple lien est que l'on ignore les génomes absents de la composante connexe principale. Ceux-ci sont pourtant des génomes potentiellement intéressants qui sont supposés appartenir à d'autres espèces, il faudrait donc pouvoir les réassigner.

Ainsi, pour la suite du projet, nous envisageons de cesser d'utiliser la hiérarchie du NCBI et nous lui préférons celle proposée par GTDB (PARKS et al., 2018). Ainsi tous les génomes seraient comparés à l'aide d'une méthode de type MASH ou fastANI (JAIN et al., 2018) contre les génomes de référence pour chaque espèce de GTDB. Ce travail est actuellement mené par Paul Amours. Enfin, l'utilisation de l'algorithme Louvain, bien qu'écartée pour le moment pour des raisons de performance, reste envisageable si les problèmes provoqués par les regroupements en simple lien se généralisent.

10.4.2 Filtrer les redondances

Il est intéressant de noter que dans les bases de données il existe de très nombreux génomes presque identiques, c'est-à-dire des clones avec parfois quelques variants SNPs mais n'ayant aucune variation en termes de contenu en gènes. Dès lors, ces génomes présentent des schémas de présence/absence très similaires. Plutôt que de considérer plusieurs fois ces génomes ayant un contenu en gènes quasi-identique, nous filtrons les génomes de ce type en se basant là encore sur la distance MASH. En effet, dans le cas où nous trouvons des génomes ayant une distance MASH entre eux < 0.0001, nous choisissons un seul des deux génomes comme représentant. Ce génome choisi correspond à celui ayant les meilleures métriques d'assemblage, c'est-à-dire le plus petit nombre de contigs.

10.4.3 Filtrer les assemblages fragmentés

Enfin, pour éviter d'ajouter des génomes constitués d'un trop grand nombre de fragments, nous filtrons ces cas ayant un L90 > 100 et un nombre total de contigs > 1000. Pour information, le L90 est une métrique indiquant le nombre de contigs nécessaires pour stocker 90% de la taille totale de l'assemblage.

Certes la méthode fonctionne correctement en utilisant des génomes fragmentés comme nous l'avons constaté avec les MAGs. Néanmoins, par souci d'hygiène, nous avons tout de même préféré filtrer les génomes trop fragmentés, notamment afin de ne pas ajouter trop de génome nuage correspondant à des fragments de gènes.

10.5 Une interface pour explorer les graphes de pangénomes partitionnés

Une fois les filtres qualité appliqués, les génomes restants sont traités avec PPanGGOLiN afin d'obtenir un graphe de pangénome partitionné pour chaque identifiant d'espèce de GenBank. Entre-temps, pour ce qui est de l'annotation des génomes et la construction des familles de gènes, il est à rappeler que ces étapes sont également effectuées par PPanGGOLiN ainsi que décrit au chapitre 7 page 111.

Après avoir constitué notre ressource de graphes de pangénomes partitionnés, notre objectif est bien d'en faire une ressource utilisable par la communauté des bioinformaticiens. Pour cela, nous avons opté pour une modélisation sous la forme d'une base de données orientée graphe et choisi la technologie orientDB (ORIENTDB LTD, 2019) pour la rendre interrogeable à distance.

Mais au-delà de l'accessibilité technique, pour que ces graphes de pangénome soient faciles d'accès pour des microbiologistes, j'ai également conçu un prototype d'interface utilisateur d'exploration de graphes de pangénome. Ce prototype est une application Web basée sur la technologie de visualisation de graphes Sigma.js (A. JACOMY et al., 2019). Ce prototype a ensuite été grandement amélioré par Rémi Planel en retravaillant l'interface avec l'emploi de la technologie de conception d'interfaces React.js (WALKE, 2019). Le résultat est consultable sur la figure 10.2 page 188. Sur la première capture, on peut contempler l'interface de parcours du pangénome. Grâce à cet écran, on peut observer le graphe dont les nœuds sont colorés en fonction des différentes parties du pangénome, mais l'utilisateur peut aussi recalculer la disposition des nœuds et empêcher que les nœuds se superposent. De plus, un panneau de configuration permet de personnaliser les paramètres de l'algorithme ForceAtlas2 (M. JACOMY et al., 2014) utilisé pour disposer les nœuds. Sur la deuxième capture, on constate qu'il est possible de zoomer avec la molette de la souris sur une région d'intérêt puis sélectionner un nœud pour obtenir des méta-informations sur cette famille de gènes comme par exemple le nom du gène représentatif de la famille. Un clic sur la famille pourrait ouvrir une fenêtre contextuelle avec des informations supplémentaires comme la fonction du gène, enfin l'idéal serait de combiner cette interface avec la plateforme MicroScope afin de bénéficier d'annotations additionnelles comme, par exemple, le contexte métabolique du gène. Cela n'est pas montré sur la figure mais une barre de recherche permet également de rechercher une famille par un nom ou un identifiant de gène et de sélectionner le nœud correspondant dans le graphe. De plus, les nœuds sont amovibles afin de laisser la possibilité à l'utilisateur d'affiner lui-même le visuel du graphe. Enfin, la troisième capture montre qu'il est possible de filtrer les nœuds que l'on souhaite visualiser en fonction d'une ou plusieurs souches d'intérêt, ou bien par exemple de filtrer les nœuds du génome nuage afin de faciliter la visualisation.

Bien que cette interface soit un prototype et qu'elle mérite d'être enrichie on mesure le potentiel qu'elle pourrait avoir si elle était rendue disponible à la communauté scientifique. Avec la prétention ambitieuse qu'à terme PanGBank devienne à la pangénomique microbienne un peu ce qu'est le visualisateur du génome humain de l'université de Californie (KENT et al., 2002).

Néanmoins, pour y arriver il faudra dépasser plusieurs contraintes et ajouter de nombreuses fonctionnalités. Tout d'abord, il est à ce jour très difficile d'afficher plus de 20 000 à 30 000 nœuds avec la technologie Sigma.js. Il est donc souhaitable de filtrer les familles de gènes très rares (moins de 1% de présence) lorsque qu'on parcourt la vue globale et de ne les faire apparaître qu'au moment où l'on zoomerait sur des régions plus précises du graphe. Deuxièmement, si le calcul de la disposition des nœuds de petits graphes (1000 à 3000 nœuds) ne semble pas poser trop de problèmes au travers du navigateur de l'utilisateur, cela n'est plus le cas avec des graphes de 20 000 à 30 000 nœuds. La disposition des nœuds via l'algorithme ForceAtlas2 doit donc être précalculée (ce que nous effectuons déjà) mais ne pourrait plus, dès lors, être ajustée par l'utilisateur. Des solutions seront peut-être à trouver dans l'emploi de technologies plus récentes comme le WebGL [1] (PARISI, 2012) ou le WebAssembly (HAAS et al., 2017). Ces technologies visent à faire du navigateur Web un véritable système d'exploitation où il devient possible d'utiliser le processeur graphique et d'exécuter des programmes bas niveau, autorisant ainsi potentiellement des calculs plus lourds.

1. Il est à noter qu'une partie de Sigma.js est déjà compatible avec la technologie WebGL.

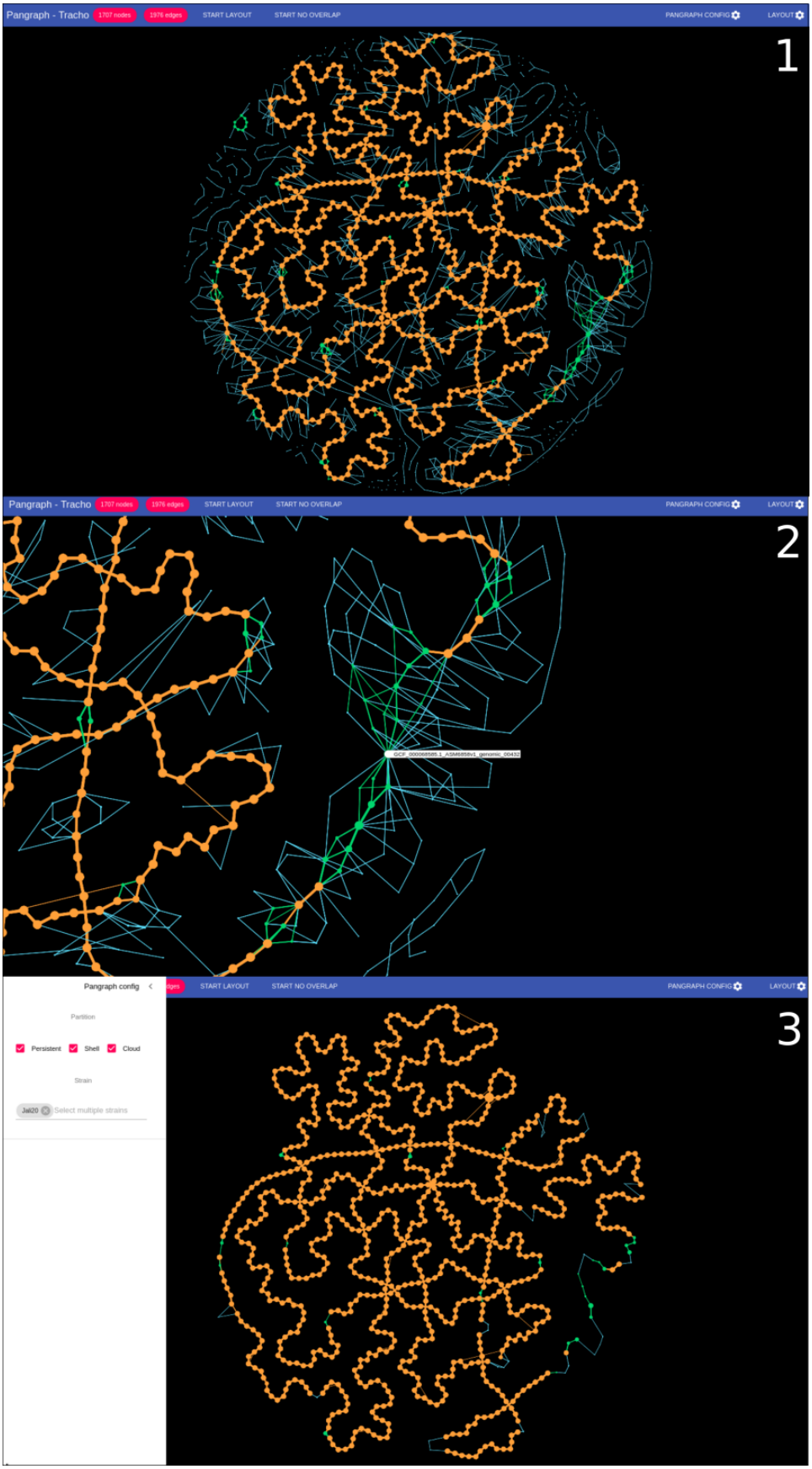


FIGURE 10.2 – Captures d'écran de l'interface Web PanGBank (basée sur le même pangénome que la figure 9.1 page 152).



Chapitre 11 : Détecter les régions de plasticité génomique

11.1 Limites des approches actuelles de détection des RGPs

À eux seuls les gènes obtenus par transferts horizontaux semblent constituer la principale source de diversité pangénomique (Treangen et al., 2011). Ainsi, les génomes procaryotes possèdent des régions variables souvent appelées îlots génomiques (Genomic Island : GI). En effet, les gènes d'origine externe semblent s'insérer préférentiellement dans ces points chauds de variabilité au sein des génomes (Oliveira et al., 2017).

La nature adaptative de ces gènes est encore débattue (cela a notamment été abordé au chapitre 4 page 71), toutefois, on constate que ces gènes obtenus par transfert vont majoritairement rester à de faibles fréquences, constituant ainsi l'ensemble des gènes rares retrouvés dans le pangénome ; ce que nous avons appelé le génome nuage. Néanmoins, ces gènes apportent parfois des fonctions avantageuses aux organismes leur permettant de s'adapter à leur environnement. Ainsi, les gènes se fixent peu à peu dans la population, de sorte qu'ils peuvent potentiellement se retrouver dans le génome coquille.

L'identification de ces régions de plasticité génomique (Regions of Genome Plasticity : RGP) revêt une grande importance. En effet, elles correspondent notamment à des îlots impliqués dans la pathogénicité, la résistance aux antibiotiques, ou encore, à des îlots d'adaptation au mode de vie de l'organisme comme le saprophytisme ou la symbiose. Pour identifier ces GIs dans les génomes procaryotes, il existe alors deux catégories d'approches : les méthodes détectant les différences en termes de composition de séquence et les méthodes fondées sur la génomique comparée. Par ailleurs, ces méthodes affinent parfois leurs résultats en se basant aussi sur le fait que les îlots génomiques sont très souvent bordés par des séquences répétées comme des séquences d'insertion avec leur transposase (Insertion Sequence : IS) ou des séquences d'ARNt.

11.1.1 Limites des approches de détection de GIs par approche compositionnelle

Les méthodes reposant sur le différentiel de composition nucléotidique sont efficaces pour détecter les GIs issus de transferts récents depuis des espèces éloignées dans l'arbre du vivant. En effet, ces gènes provenant de transferts horizontaux ont tendance à présenter une composi-

tion nucléotidique différente des portions d'ADN héritées verticalement dans les génomes. La composition en nucléotides d'un fragment constitue ainsi une signature de son espèce d'origine. C'est notamment sur cette propriété que l'on peut se baser pour regrouper des contigs issus d'un métagénome ainsi que nous l'avons évoqué à la sous-section 9.2.1 page 173.

Une méthode très commune par exemple pour détecter ces différences de composition est la méthode SIGI-HMM (WAACK et al., 2006). Sachant un ensemble de gènes détectés dans des génomes, cette méthode compte les occurrences de tous les codons utilisés afin de calculer la fréquence de l'usage des codons pour chaque gène. Sur la base de cette information, l'outil peut ainsi effectuer des comparaisons entre les gènes et utiliser des chaînes de Markov cachées afin de retrouver les régions ayant un usage atypique des codons. Comme l'avait déjà montré les auteurs de MÉDIGUE et al., 1991, il existe pour *E. coli* par exemple, trois types d'usages des codons permettant de distinguer : les gènes fortement exprimés, les gènes dits natifs et les gènes majoritairement obtenus par transfert. Sur la base de ce constat, SIGI-HMM est ainsi capable de distinguer les gènes potentiellement aliens [sic], c'est-à-dire issus de transferts horizontaux, du reste des gènes.

Cependant ces méthodes présentent des limites ainsi que l'ont montré les auteurs de KOSKI et al., 2001. Tout d'abord, les gènes ayant pour origine un transfert ancien seront de plus en plus difficiles à détecter à mesure qu'ils se stabiliseront dans la population. En effet, ces gènes présenteront une composition nucléotidique qui tendra à s'homogénéiser avec la composition basale de l'espèce au cours du temps (LAWRENCE et al., 1997). De même, si des gènes sont acquis horizontalement à partir d'un donneur ayant un usage des codons proche, la différence de composition nucléotidique sera plus difficilement détectable.

11.1.2 Limites des approches de détection de GIs par la génomique comparée

Les méthodes de génomique comparée cherchent à détecter des gènes ayant une distribution phylogénétique différente de celle à laquelle on pourrait s'attendre dans l'hypothèse où les gènes seraient hérités verticalement. Cela correspond le plus souvent à des gènes présents dans une souche alors que ceux-ci sont complètement absents des autres souches proches.

L'une des méthodes de référence de ce type d'approche est IslandPick (intégré dans IslandViewer4, BERTELLI et al., 2017). Cette méthode détecte les îlots génomiques dans un génome Q en effectuant la sélection d'un ensemble de génomes de référence R_n avec lesquels comparer les régions en commun. Ces génomes doivent être du même groupe taxonomique, idéalement de la même espèce, mais pas non plus trop proches pour éviter d'analyser des génomes issus de souches quasi-clonales et donc ayant un contenu en gènes quasi-identique. Enfin, en comparaison de ces R_n références, la méthode détermine les régions de Q susceptibles de correspondre à des îlots génomiques.

Le principal avantage de ces méthodes est qu'elles sont plus adaptées que les méthodes basées sur la composition pour détecter les transferts de gènes horizontaux intra-espèces ou intra-genres et de même pour les transferts anciens dont la composition en séquence des gènes s'est homogénéisée. Pour autant, ces méthodes utilisent souvent des comparaisons de génomes deux à deux (en les alignant entièrement ou en comparant leur contenu en gènes), ce qui a pour conséquence qu'elles sont difficilement extensibles à plus de quelques dizaines de souches.

11.2 panRGP : utiliser un pangénome partitionné pour détecter les GIs

11.2.1 Motivation de l'approche panRGP

Conscients des limites des approches précédemment présentées, nous proposons une troisième catégorie de méthodes utilisant une approche pangénomique. On s'attend, en effet, à ce que la plupart des gènes provenant de transferts horizontaux soient dans le génome nuage (pour les transferts récents) ou le génome coquille (pour les transferts maintenus dans la population et donc plus anciens) du graphe de pangénome partitionné par PPanGGOLiN (voir la partie II page 99 pour plus d'information). Ainsi, comme nous l'avons vu pour les gènes de synthèse de la capsule de *Acinetobacter baumannii* à la sous-section 9.1.1 page 154, des chemins dans le graphe contenant des successions de familles de gènes nuages ou coquilles constituent potentiellement des îlots génomiques. Nous appelons ces chemins des régions de plasticité génomique (RGPs) car elles contiennent les GIs mais également des plasmides et potentiellement des régions perdues dans une grande majorité de souches suite à des évènements de réduction de génomes qui sont anciens dans l'évolution de l'espèce.

Néanmoins, l'extraction de tous les chemins constituant les RGPs dans le graphe de pangénome est complexe, surtout pour des régions trop enchevêtrées, et peut générer des chemins qui n'existent pas dans les génomes initiaux. Une solution serait alors de démêler le graphe comme nous l'avons vu à la section 7.5 page 129.

Cependant, nous avons ici suivi une autre option en développant une méthode, appelée panRGP (BAZIN et al., 2020), capable de détecter les RGPs directement dans les génomes à partir du partitionnement effectué par PPanGGOLiN. En effet pour détecter les RGPs, nous projetons les résultats du partitionnement sur les génomes afin d'associer chaque gène à une partie du pangénome (génome persistant, génome coquille, génome nuage). La méthode va ensuite utiliser à la fois l'ensemble des parties associées aux gènes mais aussi leur ordre le long des génomes. Pour cela, un score va être attribué à chaque gène lors du parcours des contigs de chaque génome. Ce score dépendra à la fois de la partie à laquelle le gène est associé ainsi que du score du gène précédent. Une fois les scores des gènes calculés, la méthode peut ensuite extraire les RGPs respectant des critères de scores minimaux pour les gènes qu'elles contiennent. Entrons maintenant plus en détail dans le fonctionnement de la méthode.

11.2.2 Étape d'attribution des scores

Cette étape va consister à initialiser les scores attribués à chaque gène. Pour cela, nous allons parcourir les contigs en calculant pour chaque gène, le score suivant [1] :

$$s_g = (s_{g-1} + f(g))^+$$

où s_g est le score du gène g, s_{g-1} est le score du gène précédent de g et $f(g)$ est la fonction attribuant la part spécifique du gène g au score s_g. Si un gène n'a pas de voisin précédent, s_{g-1} est égal à 0 donc $s_g = f(g)^+$.

1. La notation x^+ utilisée dans la fonction symbolise une fonction renvoyant le maximum entre 0 et x.

Introduisons la fonction $p(g)$, où $p(g) = 1$ si le gène est à la fois persistant et qu'il n'appartient pas à une famille multigénique et $p(g) = 0$ s'il est autre chose (génome nuage, génome coquille, génome persistant multigénique).

Enfin, introduisons n qui est un compteur dont la valeur est :

$$n = \begin{cases} n+1 \text{ si } p(g) > 0 \\ 0 \text{ sinon.} \end{cases}$$

Une fois ces éléments introduits, nous avons alors une fonction f :

$$f(g) = \begin{cases} -(3^n) \text{ si } p(g) > 0 \\ 1+\varepsilon \text{ sinon.} \end{cases}$$

où $\varepsilon = \frac{1}{\infty}$ de sorte que l'on considère que $\varepsilon \approx (\varepsilon + \varepsilon)$ (bien que $\varepsilon > 0$). Cela permet ainsi d'éviter qu'il puisse exister des situations non décidables car le résultat du parcours d'un contig dans un sens serait différent du parcours dans un autre sens. Grâce à l'ajout de ε cette ambiguïté est levée.

Pour identifier les familles de gènes multigéniques, nous utilisons une méthode qui est certes peccable mais qui semble suffire pour ce type d'analyse. En effet, nous recherchons tout simplement dans si celles-ci sont multigéniques dans plus de 5% de tous les génomes où elles sont présentes.

En cas d'assemblage circulaire, le calcul du score commencera par le premier gène g_1 puis il parcourra tous les gènes. Si au moins un gène de score nul a été rencontré alors un deuxième passage sera réalisé jusqu'au dernier gène ayant eu un score nul lors du premier passage. Ce cas peut se rencontrer dans des plasmides circulaires composés principalement de gènes du génome coquille ou nuage et de quelques gènes du génome persistant seulement.

Enfin, pour améliorer la compréhension, un exemple d'attribution de scores à des gènes présents sur un contig, suivi de l'identification des RGPs est proposé sur la figure 11.1 page suivante. Voyons donc maintenant comment ces scores sont utilisés pour l'identification des RGPs.

11.2.3 Étape d'identification des RGPs

Pour identifier les RGPs, deux paramètres sont fournis par l'utilisateur : un critère de score minimal pour considérer une RGP comme valide (4 par défaut et au minimum 1) et un critère de taille minimale en nucléotides (3Kb par défaut). Le score d'une RGP correspond à l'arrondi (c'est-à-dire la suppression du terme ε) du score du gène le plus à l'extrémité droite de la RGP.

Pour détecter les RGPs, l'algorithme boucle sur trois étapes appliquées sur chaque contig :

1. Sélection du gène ayant le plus haut score. En cas d'égalité, le gène le plus proche de l'extrémité 3' est sélectionné.
   - Si son score s_g est supérieur ou égal au score minimal pour une RGP défini par l'utilisateur alors une nouvelle RGP est initiée.
   - Sinon, il y arrêt de la boucle.

2. Les gènes précédant le gène sélectionné sont successivement ajoutés dans le sens 3'→5' dans la RGP en mettant à 0 leur score s_g, ceci jusqu'à trouver un gène ayant déjà un score nul (on rappelle que $\varepsilon > 0$).
3. Recalculer les scores s_g à partir de l'extrémité 3' de la RGP extraite jusqu'à trouver un gène ayant déjà un score nul, puis revenir à l'étape 1.

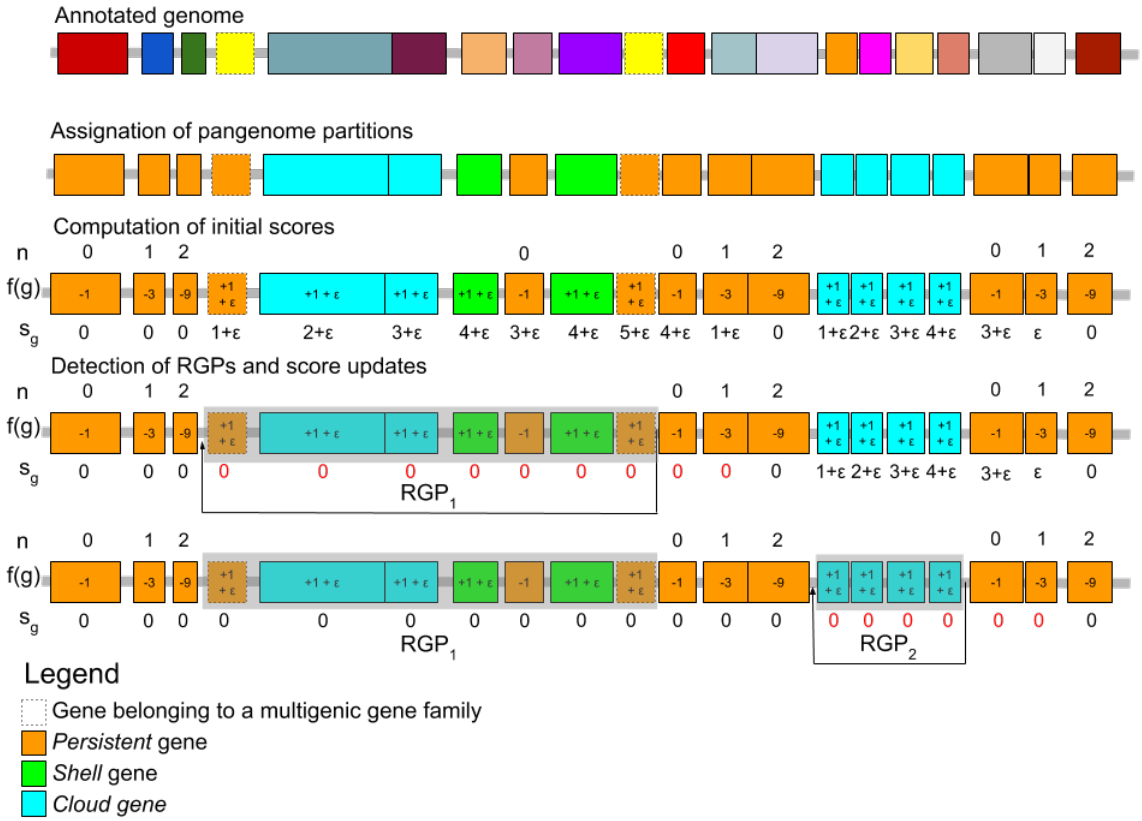


FIGURE 11.1 – Figure reproduite depuis BAZIN et al., 2020 illustrant la méthode d'identification des RGPs sur un contig. Le partitionnement PPanGGOLiN est tout d'abord projeté sur les gènes du contig et les gènes multigéniques sont identifiés. Des scores sont ensuite attribués aux gènes en commençant par l'extrémité gauche du contig. Une fois tous les scores attribués, le gène ayant le score maximum ($5+\varepsilon$ dans l'exemple) sert d'origine à l'extraction de la RGP_1 dont le score est ainsi de 5. Les scores des gènes à l'extrémité 3' de la RGP_1 sont ensuite recalculés jusqu'à atteindre un gène possédant déjà un score $s_g = 0$. Le gène ayant le nouveau score maximum ($4+\varepsilon$ dans l'exemple) sert d'origine à l'extraction de la RGP_2 dont le score est ainsi de 4. Les scores des gènes suivants la RGP_2 sont recalculés jusqu'à atteindre un gène possédant déjà un score $s_g = 0$. À ce moment, tous les scores sont à égalité ($s_g = 0$), l'algorithme s'arrête.

Après la fin de l'algorithme, les RGPs ayant une taille inférieure à celle souhaitée par l'utilisateur sont filtrées. Il est à noter que l'algorithme présenté ici est symétrique, c'est-à-dire qu'il donnera les mêmes RGP si nous inversons la droite et la gauche pour le parcours des gènes et d'identification des RGPs. Du fait des scores utilisées, il est attendu que les RGPs identifiées puissent contenir

quelques gènes persistants afin d'assurer la contiguïté des RGPs. Néanmoins ceux-ci resteront dans tous les cas marginaux car les fonctions de score utilisées imposent un nombre important de gènes coquilles ou nuages pour accepter l'ajout d'un gène persistant dans une RGP.

11.2.4 Regroupement des RGPs appartenant au même spot

Pour étudier l'évolution de la diversité des îlots génomiques présents dans une espèce ou une population, il peut être intéressant d'étudier les « points chauds d'insertion » où de nombreuses RGPs alternatives sont présentes. Pour cela, il est nécessaire de regrouper les RGPs détectées avec l'approche précédemment décrite, sur la base de leurs extrémités communes. Dans l'article décrivant la méthode panRGP (BAZIN et al., 2020), nous avons décidé d'utiliser le terme « spot » pour désigner ces ensembles de RGPs présentant des extrémités similaires de sorte qu'on puisse considérer qu'elles proviennent très probablement d'un même point chaud d'insertion dans les génomes.

Pour la détection de ces spots nous procédons à partir d'un ensemble de génomes dont les RGPs ont été prédites ainsi que cela est détaillé à la sous-section précédente et illustré sur la figure 11.2 page ci-contre. Les spots d'appartenance des RGPs sont ensuite déterminés en comparant la similarité des gènes flanquants non inclus à l'intérieur de chaque RGP. Plus précisément, nous sélectionnons, aux deux extrémités de chaque RGP, les 3 gènes consécutifs qui sont à la fois des gènes persistants et non multigéniques afin de les ordonner en fonction de leur distance par rapport à la RGP puis de les convertir en leur famille de gènes correspondante.

Les extrémités d'une RGP sont ainsi définies comme des paires d'ensembles ordonnés de familles de gènes. Pour décrire les correspondances entre les RGPs, nous construisons un graphe d'extrémité $G(V,E)$ où chaque nœud v représente le couple d'extrémités de chaque RGP et où chaque arête indique la similarité entre ces couples (voir figure 11.2 page suivante).

Une extrémité du couple v_i et une autre extrémité du couple v_j sont similaires si l'une de ces 2 conditions est respectée :

— Si les premières familles de gènes des 2 ensembles ordonnés sont identiques.
— Si au moins 2 familles de gènes sur 3 sont présentes dans les 2 ensembles ordonnés correspondant à chaque extrémité.

Les critères de similarité ici indiqués ont été déterminés empiriquement mais peuvent être adaptés si nécessaire.

En cas de similarité entre les 2 couples d'extrémités v_i et v_j de deux RGP comparés i et j, nous ajoutons alors une arête entre les nœuds correspondants dans le graphe d'extrémités. Une fois le graphe entièrement construit, toutes les composantes connexes du graphe sont extraites, chacune d'elles correspond alors à un spot. Par la suite, plusieurs métriques sont calculées sur ces spots comme par exemple le nombre de RGPs qu'ils possèdent ou bien le nombre de chemins différents présents entre les deux extrémités.

Les RGPs qui n'ont pas suffisamment de gènes aux deux extrémités pour effectuer la comparaison ne sont pas pris en compte pour la prédiction des spots. Ce type de RGP correspond soit à des RGPs incomplètes car elles se terminent aux frontières d'un contig, soit à des plasmides n'ayant donc pas de contexte génomique dans le génome.

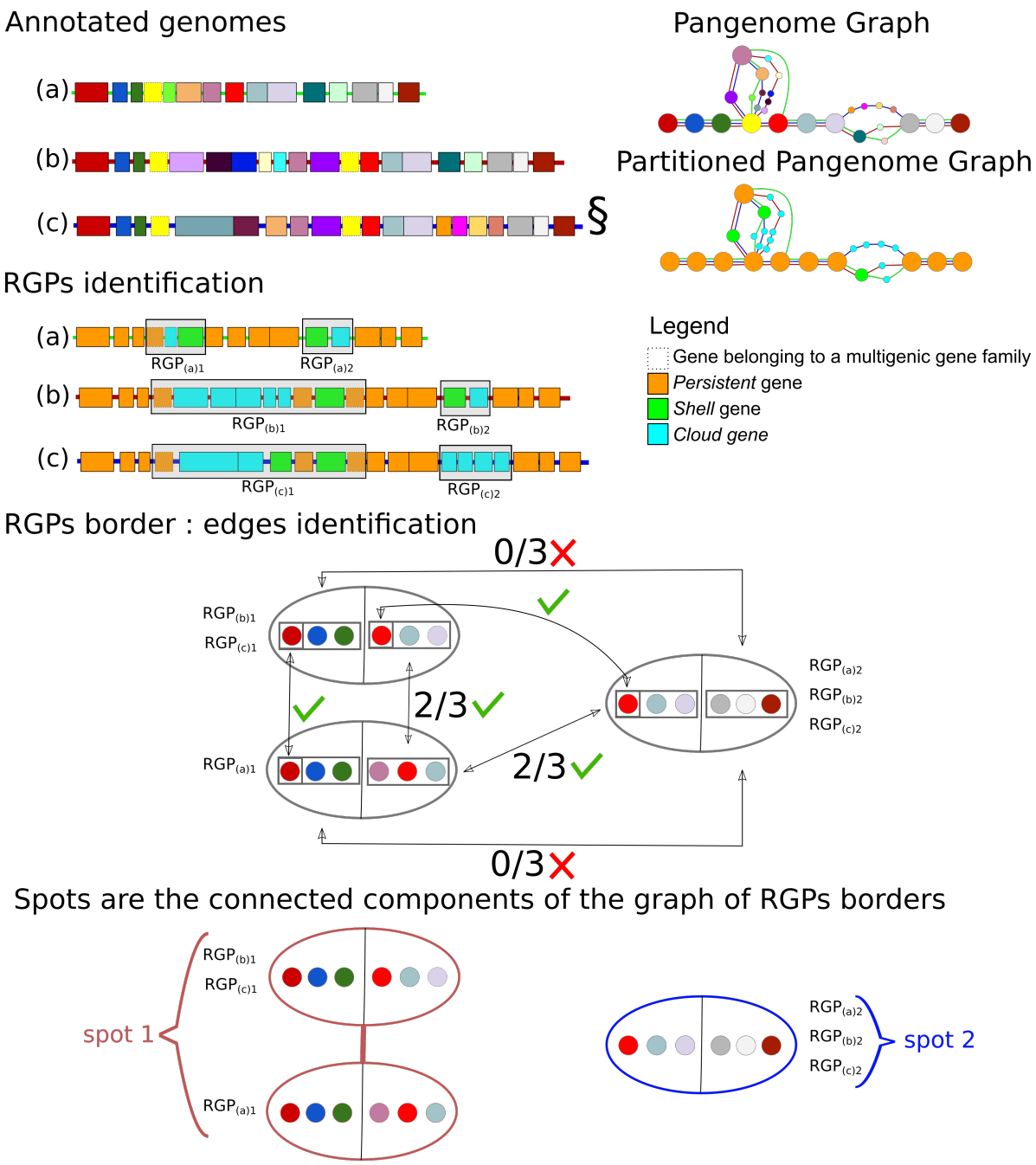


FIGURE 11.2 – Figure supplémentaire non publiée de l'article BAZIN et al., 2020 illustrant le regroupement des RGPs appartenant à un même spot. Pour cela chaque extrémité est comparée à toutes les autres et une arête est produite entre 2 RGPs si les 2 couples d'extrémité respectent au moins une des conditions indiquées pour être considéré comme similaire. Nous obtenons alors le graphe d'extrémité. Pour alléger la figure, les RGPs $RGP_{(b)1}$ et $RGP_{(c)1}$ sont fusionnés car elles possèdent des extrémités identiques et donc appartiennent trivialement au même spot. Les composantes connexes du graphe d'extrémité nous fournissent alors les spots de RGPs (ici 2 spots sont identifiés). Il est à noter que la RGP affublée d'un symbole § correspond à la RGP dont l'identification est détaillée sur la figure 11.1 page 193.

11.3 Résultats de la méthode panRGP

11.3.1 Exemple d'une RGP extraite de la souche d'*A. baumannii* AYE

À travers la figure 9.2 page 155, nous avons vu qu'il était possible de parcourir le graphe de pangénome à la recherche de chemins pangénomiques. Pourtant, certaines régions détectées par panRGP (à l'image de celle présentée sur la figure 11.3) auraient très difficilement pu être identifiées dans le graphe brut car elles comprennent de très nombreux gènes dont les familles sont enchevêtrées dans le graphe. Plus particulièrement, cette RGP décrite plus en détail dans l'article de FOURNIER et al., 2006 correspond à un îlot de résistance aux antibiotiques comprenant 83 gènes (dont 45 sont connus pour conférer des résistances) pour une taille estimée de 86Kb d'après la méthode panRGP, ce qui est conforme à la taille de la région qu'avait identifiée les auteurs en 2006. Avec ses 45 gènes de résistance aux antibiotiques, cet îlot de résistance semble être un des plus importants à avoir été identifié dans la littérature. La souche *A. baumannii* AYE dont est extraite cette RGP a été identifiée comme multirésistante aux traitements lors d'infections nosocomiales notamment dans des hôpitaux français (POIREL et al., 2003).

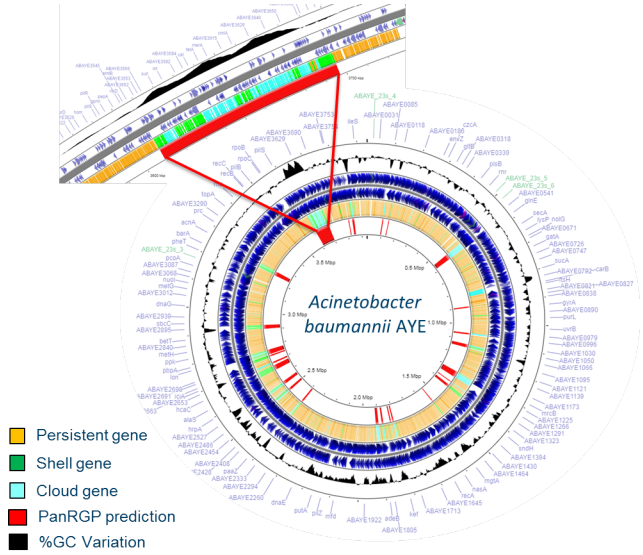


FIGURE 11.3 – Figure illustrant une RGP de 86kb comportant 83 gènes extraits de la souche *A. baumannii* AYE grâce à la méthode panRGP. La RGP peut être consultée plus en détail sur le site Web de la plateforme MicroScope (VALLENET et al., 2019). On constate par ailleurs de multiples RGPs identifiées sur ce même génome.

Cet exemple permet d'illustrer les résultats que peut produire la méthode panRGP sur la base du partitionnement effectué par PPanGGOLiN mais des analyses plus globales sont nécessaires pour évaluer la méthode panRGP.

11.3.2 Comparaison avec d'autres outils

Afin de s'assurer de la fiabilité des résultats de la méthode panRGP un seul exemple ne suffit pas. Nous avons donc comparé les résultats issus de l'application de la méthode sur les génomes de la base de données NCBI RefSeq (sous-partie réannotée de GenBank) à un jeu de données dont les

îlots génomiques ont été vérifiés par des recherches bibliographiques par les auteurs de (LANGILLE et al., 2008) avant d'être mis à jour par les auteurs de (BERTELLI et al., 2018). Nous avons également utilisé d'autres méthodes couramment utilisées pour détecter les îlots génomiques afin de mesurer les résultats que ces méthodes produisent sur ce jeu de données. Celui-ci est constitué de six génomes complets [1] dont les auteurs ont inspecté les régions correspondant effectivement à des îlots génomiques (régions positives) et ainsi que celles qui ne peuvent absolument pas être des îlots génomiques (régions négatives). Entre ces deux types de régions, il existe des régions dont le statut est incertain et qui ne sont pas prises en compte pour l'évaluation des outils. Ainsi, une bonne méthode de détection doit à la fois détecter les régions positives sans détecter les régions négatives.

Nous mesurons la fiabilité de ces différentes méthodes (résultats présentés dans le tableau 11.1) en calculant le taux de précision, le taux de sensibilité et le score F1. Ce dernier correspondant à la moyenne harmonique entre précision et sensibilité.

Outil	Type d'approche	Précision	Sensibilité	score-F1
panRGP (BAZIN et al., 2020)	pangénomique	**1.0**	0.884	**0.932**
xenoGI (BUSH et al., 2018)	génomique comparée	0.935	**0.924**	0.917
Islandviewer4 (BERTELLI et al., 2017)	hybride	0.998	0.669	0.791
IslandCafe (JANI et al., 2019)	compositionnelle	1.0	0.574	0.715
GI-Cluster (LU et al., 2018)	compositionnelle	0.87	0.714	0.743
PredictBias (PUNDHIR et al., 2008)	compositionnelle	0.856	0.771	0.805
IslandPath-DIMOB (BERTELLI et al., 2018)	compositionnelle	0.998	0.479	0.636
SIGI-CRF (WAACK et al., 2006)	compositionnelle	0.993	0.434	0.52
AlienHunter (VERNIKOS et al., 2006)	compositionnelle	0.753	0.57	0.642
SIGI-HMM (WAACK et al., 2006)	compositionnelle	0.817	0.325	0.444
ZislandExplorer (WEI et al., 2017)	compositionnelle	0.833	0.18	0.278

TABLE 11.1 – Comparaison de la fiabilité de la détection des îlots génomiques de onze outils sur un jeu de données vérifié par les auteurs de LANGILLE et al., 2008 (constitué de six génomes).

Les résultats indiquent que panRGP se place avec xenoGI parmi les méthodes les plus fiables de l'état de l'art pour la détection des îlots génomiques. Nous pensions que comme panRGP identifie indistinctement les régions en réduction et les régions acquises par transfert horizontaux, nous aurions pu constater une baisse de la précision de PPanGGOLiN due à une surprédiction de RGPs. Or il n'en est rien, car seules les réductions massives de blocs entiers de gènes apparaissent dans le génome coquille. En effet, PPanGGOLiN est capable de détecter un génome persistant stable même en cas d'absence de gènes due à des réductions isolées.

Bien qu'ayant des performances relativement similaires à panRGP sur le tableau 11.1, la méthode xenoGI (BUSH et al., 2018) est très gourmande en ressources. Par exemple, les auteurs indiquent dans leur article que l'analyse de 40 souches demande 20 heures de calcul sur 50 cœurs (en utilisant 500 Go de mémoire vive). À l'inverse, la méthode panRGP est pour sa part quasiment instantanée sur un seul cœur une fois le pangénome partitionné. Le partitionnement du pangénome est en effet à effectuer au préalable mais seulement une seule fois pour tous les génomes d'une espèce. Celui-ci peut potentiellement se baser pour sa part sur des dizaines de milliers de génomes. À titre d'exemple, la méthode PPanGGOLiN utilisée sur 40 génomes de l'espèce *Shigella dysenteriae* requiert un temps de calcul d'environ 10 minutes sur 8 cœurs.

1. *Escherichia coli* O157 : H7 str. Sakai, *Escherichia coli* CFT073, *Salmonella enterica subsp. enterica* serovar Typhi str. CT18, *Staphylococcus aureus str.* MW2, *Streptococcus pyogenes* str. MGAS315 et *Vibrio parahaemolyticus* RIMD 2210633.

11.4 Conclusion sur la méthode panRGP

À l'origine, j'ai conçu la méthode panRGP avec un étudiant en stage de Master 2 se prénommant Valentin Sabatet. Ce travail de stage avait abouti à une première version de l'outil [1] détectant correctement les RGPs mais qui avait tendance à les interrompre plus facilement lors de la rencontre de gène persistants. Sur la base de ce travail, une nouvelle version a été conçue et développée par Adelme Bazin afin de considérablement améliorer la contiguïté des RGP prédites au travers de la maximisation d'un score global de la RGP et non local comme dans la version précédente. C'est cette méthode qui est présentée dans la section 11.2 page 191. De plus, Adelme Bazin a également été à l'origine de la comparaison de panRGP aux autres outils présentés à la sous-section 11.3.2 page 196. En sus du tableau 11.1 page précédente, Adelme a aussi effectué d'autres tests supplémentaires sur davantage de génomes confirmant que panRGP semble au niveau des meilleures méthodes de l'état de l'art. Cela est en particulier détaillé dans l'article BAZIN et al., 2020 dont je suis co-auteur. Il faut avoir à l'esprit qu'une fois que les pangénomes de toutes les espèces sont constitués et rendus disponibles (ce que nous avons abordé aux chapitres 10 page 181), panRGP devient une méthode d'une rapidité qui sera très difficilement atteignable par des méthodes concurrentes. Cela autorise donc l'utilisation de panRGP sur des grandes quantités de données et même d'ajouter les RGPs détectés comme métadonnées à la ressource PanGBank. Il faut d'ailleurs dire que la méthode panRGP est d'ores et déjà intégrée à la plateforme MicroScope avec une interface graphique pour explorer les résultats de panRGP ainsi que cela est décrit dans l'article (VALLENET et al., 2019) dont je suis également co-auteur.

Enfin, il n'existe à ce jour aucune méthode d'identification des îlots génomiques à même d'utiliser des MAGs ou des SAGs. Or nous l'avons vu, la méthode PPanGGOLiN est en mesure de partitionner le pangénome provenant de ces données. Une prochaine étape de ce travail sera donc de détecter les îlots génomiques dans les MAGs directement, au besoin d'ailleurs en utilisant le graphe de pangénome afin de positionner et orienter les gènes appartenant à plusieurs contigs, c'est-à-dire, de procéder au « scaffolding » des MAGs.

1. Cette version est décrite dans le rapport de stage de Valentin Sabatet (disponible sur demande).



Conclusion, bilan et perspectives

11.5 Bilan des contributions scientifiques

L'étude de la biologie des procaryotes révèle que leurs génomes sont soumis à des flux incessants de gains notamment par transferts horizontaux et de pertes de gènes pouvant conduire à la réduction des génomes. De ce fait, on constate généralement une bien plus importante différence en termes de contenu génique, même entre génomes proches, que l'on ne retrouve de diversité de mutations dans les gènes du génome cœur, pourtant traditionnellement utilisés pour construire la phylogénie des espèces. L'exemple le plus frappant, même s'il reste anecdotique, étant probablement celui rapporté par les auteurs de MARTINEZ-MURCIA et al., 1992 sur l'espèce *Aeromonas trota* où le taux d'identité pour l'ARNr 16S est de 99.9% malgré des différences de contenu en gènes de 70% entre certaines souches du fait de pertes massives de gènes. Dans le cas général, cette variation peut atteindre des valeurs jusqu'à ≈30-35% du contenu en gènes des génomes pour certaines espèces. Face à ce constat, la pangénomique des procaryotes propose alors de répertorier, de modéliser et de visualiser l'étendue de cette diversité en gènes notamment entre génomes proches sous un paradigme aréférentiel afin de comprendre le rôle évolutif que peuvent avoir ces gènes, notamment sur la manière dont les organismes s'adaptent à leur environnement.

11.5.1 PPanGGOLiN

La pangénomique des procaryotes palpite au cœur du travail de doctorat dont ce manuscrit parachève l'exercice (la soutenance est pour sa part disponible en annexe A.1 page 207). À travers les analyses effectuées sur la base de 136 287 génomes provenant majoritairement d'isolats (439 espèces) et 34 210 génomes reconstruits à partir de métagénomes (78 espèces), nous avons pu illustrer la manière dont le graphe de pangénome partitionné peut être utile pour étudier la variabilité du pangénome des espèces mais également de leurs régions de plasticité génomique. À ma connaissance, ces résultats constituent l'étude pangénomique la plus massive menée jusqu'à ce jour. Pour pouvoir mener cette étude, nous devions développer la méthode PPanGGOLiN

afin de répondre à certaines des limites soulevées par l'utilisation des approches historiques en pangénomique. Nous devions en effet stabiliser le génome persistant et séparer les familles de gènes à croissance lente et rapide.

Un génome est en quelque sorte une photographie d'un individu dans une population qui varie dans le temps et dans l'espace. Ces multiples images pour chaque espèce permettent de rendre compte des diverses caractéristiques qu'un génome peut accepter sachant ses limites idiosyncratiques. Notamment, certains des gènes partageant une longue histoire sont destinés à rester associés car ils participent à une fonction commune tandis que d'autres possèdent une existence transitoire ou limitée à des environnements spécifiques. Comment distinguer ces catégories ? La caractéristique principale de la méthode PPanGGOLiN est l'utilisation d'un modèle statistique multivarié de Bernoulli pour effectuer le partitionnement du pangénome en trois parties (génome persistant, génome coquille, génome nuage) ou plus, en se basant sur les schémas de présence/absence des familles de gènes et sur l'organisation des gènes dans les génomes.

Sur la base de la méthode PPanGGOLiN, nous avons montré que notre modélisation statistique permet ainsi une estimation d'un génome persistant stable contrairement aux approches classiques fondées sur la fréquence des familles de gènes. L'utilisation du génome coquille a été un élément clé de nos analyses afin de comprendre la dynamique du pangénome en ce qu'elle reflète les capacités d'adaptation des espèces par l'acquisition de nouveaux gènes qui sont maintenus dans la population. Nous avons observé notamment que la proportion de génome coquille n'augmente pas avec la taille du génome mais représente une grande fraction des génomes des espèces lorsque ce génome est structuré en plusieurs parties. Le fait que, pour autant, les génomes ayant plus de gènes n'aient pas une proportion supérieure de génome variable semble indiquer qu'il y aurait bien une fixation de gènes s'opérant à long terme dans le génome persistant. Cet équilibre est propre à chaque espèce et notamment à sa capacité à acquérir de nouveaux gènes depuis son environnement et à les fixer dans la population.

La méthode PPanGGOLiN est la seule méthode à ce jour à être en mesure de détecter plusieurs parties dans le génome coquille sur la base de leur schéma de présence/absence. Cela permet ainsi de mettre en évidence de potentielles structures dans la population où le génome coquille tendrait à devenir persistant, indiquant ainsi un probable phénomène de spéciation allopatrique en cours.

À l'exception des travaux de DELMONT et al., 2018 ; PLAZA ONATE et al., 2019, très peu d'études avaient jusqu'ici étudié le pangénome à partir de données métagénomiques. La méthode PPanGGOLiN est ainsi à notre connaissance la première méthode capable de partitionner le contenu pangénomique des espèces issues de génomes assemblés depuis des métagénomes malgré des taux de présences – pour les gènes appartenant normalement au génome persistant – parfois très bas. Cela permet donc d'envisager de nombreuses analyses supplémentaires sur des espèces encore non cultivées et donc dépourvues de génomes de référence. On peut penser à la reconstruction du métabolisme central à partir du génome persistant afin de prédire les milieux de culture et ainsi pouvoir les étudier en laboratoire ou bien à l'étude du paysage des gènes transférés horizontalement au sein des espèces présentes dans un écosystème.

C'est fort de ces conclusions que nous avons décidé de publier un article décrivant la méthode ainsi que les résultats obtenus dans le journal *PLOS Computational Biology* (GAUTREAU et al., 2020). Enfin, l'approche a également été présentée sous la forme d'un poster (annexe A.3 page 209) ainsi qu'à travers six présentations orales dans des conférences et séminaires dont les JOBIM 2017 et 2018 (Journées Ouvertes en Biologie Informatique et Mathématique) ainsi que « Genome Informatics » édition 2018 à Cambridge (présentation disponible à l'annexe A.2 page 208).

11.5.2 panRGP

Grâce aux résultats de partitionnement de la méthode PPanGGOLiN, nous avons également développé la méthode panRGP permettant de prédire des régions de plasticité génomique comme les îlots génomiques notamment. Cette méthode se base sur la succession des différentes parties (issues du partitionnement du pangénome) dans les génomes, afin d'identifier ces régions de plasticité génomique, c'est-à-dire enrichies en génome nuage et en génome coquille. En comparant les résultats de la méthode panRGP aux méthodes les plus utilisées sur un jeu de données indépendant, nous montrons que panRGP présente un taux de détection des îlots génomiques meilleur ou équivalent aux autres outils tout en étant considérablement plus rapide et surtout en mesure d'utiliser davantage de génomes (jusqu'à des centaines de milliers). Les résultats de cette méthode sont présentés dans l'article de BAZIN et al., 2020 dont je suis co-auteur. Cet article a été accepté en tant que « proceeding » pour l'« European Conference on Computational Biology » (ECCB2020) qui sera publié dans la revue *Oxford Bioinformatics*. Par ailleurs, en plus de cette conférence ECCB2020, nous avons eu l'opportunité de présenter la méthode aux conférences JOBIM 2019 et FEMS 2019 sous la forme d'un poster présenté en annexe A.5 page 211.

11.5.3 MicroScope

La méthode PPanGGOLiN et la méthode panRGP sont d'ores et déjà intégrées à la plateforme MicroScope comme évoqué dans les articles VALLENET et al., 2019 et MÉDIGUE et al., 2017 dont je suis co-auteur. Ces articles décrivent les dernières fonctionnalités développées au sein du laboratoire afin que celles-ci puissent bénéficier aux 4 500 utilisateurs répartis à travers le monde. Enfin, j'ai eu la chance de présenter les fonctionnalités pangénomiques de la plateforme sous la forme d'un poster (annexe A.6 page 212) lors des éditions 2017, 2018 et 2019 des JOBIM.

11.5.4 PanGBank

PanGBank est une banque de données couplée à une interface de visualisation qui est en cours de développement. Nous avons fait le constat que cette ressource ne correspond à aucune banque déjà existante et qu'elle aurait un fort potentiel pour la communauté scientifique. En effet, le partitionnement de PPanGGOLiN prend toute sa puissance couplé aux graphes de pangénome. Singulièrement, le parcours des génomes de milliers de souches en une seule interface ressemblant en quelque sorte à une carte du métropolitain semble susciter un certain intérêt quand l'idée est évoquée aux biologistes. Et on le comprend, grâce à sa structure compacte dans laquelle les gènes sont regroupés en familles tout en préservant l'information sur leur voisinage génomique, le graphe résume la diversité de milliers de génomes en une seule image et permet d'analyser efficacement

les différents chemins que peuvent parcourir les génomes, les régions ou les gènes d'intérêt. Cette approche permet alors de fournir un cadre d'analyse nouveau pour la génomique comparée à très grande échelle.

À ce jour, la ressource PanGBank a été présentée sous la forme d'un poster A.4 page 210 aux JOBIM 2019 et reste toujours en cours de développement avec l'objectif de la rendre publique d'ici 2020-2021. L'objectif est également d'écrire un article décrivant ses fonctionnalités dans la parution dédiée aux bases de données du journal *Nucleid Acid Research*.

11.5.5 Réserves sur l'approche pangénomique proposée

Toutefois ce travail pourrait être amélioré sur bien des points comme je l'ai déjà discuté à plusieurs occasions dans le manuscrit. Je souhaiterais insister sur un point crucial en pangénomique afin qu'il soit bien à l'esprit des lecteurs de ce document. En utilisant la méthode PPanGGOLiN, comme avec beaucoup de méthodes en pangénomique ou même en phylogénie, on est aisément tenté de supposer que la diversité d'un ensemble de génomes à analyser conjointement est représentative de la « réalité ». Or, les bases de données présentent de nombreux biais de représentation. Des solutions de normalisation de la diversité génomique susceptibles d'aider à résoudre ce type d'écueils ont été discutées dans la sous-section 7.7 page 135 mais les intégrer à la méthode PPanGGOLiN aurait nécessité un travail plus approfondi. Il reste que même avec une normalisation, la représentativité d'un échantillon est relative, de sorte qu'on peut dire que cela constitue la principale objection pouvant être faite à l'approche pangénomique si elle est naïvement comprise. Ainsi, avant de tirer des conclusions hâtives en pangénomique, il est souhaitable de se demander si la variabilité des génomes inclus dans l'analyse est réellement susceptible de répondre à la question scientifique posée. Et plus généralement on peut aussi regretter qu'avec l'essor de « buzz words » [sic] comme « Big Data » ou « Machine Learning », il y ait parfois l'illusion qu'accumuler des données (comme le fait PPanGGOLiN) soit ontologiquement suffisant pour être en mesure de répondre à tous les problèmes hors de toute démarche hypothético-déductive.

11.6 Perspectives à approfondir ou à envisager

En perspective, cela va sans dire, j'aimerais tout d'abord pouvoir suivre les pistes que j'ai évoquées dans le manuscrit comme le démêlement du graphe, ou bien la normalisation de l'échantillonnage de la diversité des génomes. Je souhaiterais également continuer le développement des outils basés sur PPanGGOLiN comme panRGP ou PanGBank. J'aimerais aussi collaborer plus étroitement avec des biologistes afin d'imaginer avec eux comment améliorer les différentes approches en fonction de leurs besoins. En effet, une des limites de cette thèse est potentiellement d'avoir eu une perspective trop large sur la pangénomique des procaryotes. Si ce point de vue élargi semble nécessaire pour développer une méthode générique, il reste qu'avoir une compréhension détaillée de la biologie d'une espèce me semble indispensable si l'on veut tirer le meilleur parti du graphe de pangénome partitionné.

Mais avant de terminer ce manuscrit de thèse, il me faut suggérer des développements nouveaux, des fonctionnalités susceptibles d'étendre l'écosystème de PPanGGOLiN. Développements que bien sûr, j'aurais à cœur de conduire si d'aventure les conditions de travail du métier de chercheur me permettaient d'y pourvoir librement.

11.6.1 Association des chemins à des données externes

Une première chose qui me semble importante pour tirer parti de tout le potentiel du graphe de pangénome partitionné est de pouvoir associer des chemins pangénomiques à des métadonnées externes sur les organismes. En effet, si l'on suppose que des gènes confèrent des adaptations à des environnements ou qu'ils procurent des phénotypes particuliers, il est possible de relier leur présence dans les génomes à ces traits caractéristiques.

J'ai mené des tentatives en ce sens, elles consistaient à diviser le génome coquille en autant de groupes que de familles de gènes ayant des vecteurs de présence/absence différents. L'objectif est ainsi d'effectuer un sous-partitionnement plus fin que celui de NEM. Pour cela, une matrice de dissimilarité de Jaccard-Needham est calculée entre les familles de gènes du génome coquille sur la base de la matrice de présence/absence du pangénome. Cette matrice est filtrée pour les distances supérieures à un seuil et sert de matrice d'adjacence à un graphe de distance. Sur ce graphe, les composantes connexes permettent d'identifier des groupes de familles du génome coquille ayant des schémas de présence/absence similaires. À la place des composantes connexes, nous avons aussi testé les algorithmes Louvain (Blondel et al., 2008) et MCL (Enright et al., 2002), ce dernier semblant donner de bons résultats. Le génome coquille a été utilisé car nous pensions que le génome nuage ainsi que le génome persistant avaient trop peu de variabilité (trop de valeurs 0 pour le premier et trop de valeurs 1 un pour le second) pour permettre des conclusions statistiques. Par la suite, l'association de ces groupes de familles de gènes ayant des vecteurs de présence/absence similaires (un vecteur médian est calculé pour chaque groupe) avec des variables externes a été testée en effectuant des corrélations (de Pearson ou de Spearman) entre les vecteurs binaires et les valeurs numériques des métadonnées. Ainsi, on peut lister des gènes dont la présence/absence covarie avec un phénotype et trier les résultats sur la base de valeurs R ou ρ.

Des associations ont également été testées cette fois avec des variables catégorielles en utilisant un test V de Cramér (Cramér, 1946). Ce test correspond à un test d'indépendance du χ^2 normalisé pour quantifier l'intensité de la relation entre deux variables catégorielles (présence/absence et un phénotype). En effet, le problème d'un test d'indépendance du χ^2 classique est qu'il indique seulement la probabilité de se tromper en rejetant l'hypothèse d'indépendance entre deux variables alors qu'elle est vraie, mais il ne permet pas, en soit, de quantifier ce lien de dépendance en cas de rejet de H_0.

Bien que cette tentative soit naïve par rapport à des approches statistiques plus sophistiquées comme des modèles linéaires généralisés avec pénalisation de Lasso souvent utilisés pour les études d'association (Genome Wide Association Studies : GWAS), j'ai tout de même effectué des tests notamment avec le jeu de données des auteurs de Scoary (Brynildsrud et al., 2016). Ce jeu de données comprend 3085 isolats de *Streptococcus pneumoniae* séquencés lors d'une épidémie en Thaïlande couplés aux résultats binaires de susceptibilité à l'érythromycine. Les auteurs ont

notamment mis en évidence grâce à leur méthode Scoary, un gène de résistance à l'érythromycine : le gène *erm*. La méthode que j'ai décrite a été employée pour les gènes du génome coquille associés aux résultats binaires de susceptibilité via le test V de Cramér. Les résultats ont été décevants car les gènes associés significativement (p-valeur < 0.05) par un test du χ^2 (même après correction des tests multiples avec la méthode de Bonferroni) l'étaient avec une valeur V très faible et ne correspondaient pas au gène *erm*. Après une analyse plus poussée, il se trouve que le gène *erm* était dans le génome nuage et non dans le génome coquille. Il est donc rassurant que le test V de Cramér n'ait pas donné un score V élevé à d'autres gènes. À l'inverse, en ajoutant le génome nuage dans l'analyse, le gène *erm* est effectivement associé avec une valeur V très importante.

Là où on aurait pu penser que le génome nuage était composé de trop de valeurs nulles pour permettre une association, on constate ici que cela, semble-t-il, n'a pas causé de problème pour l'association. On remarquera cependant que pour ce jeu de données nous avons tout de même pu obtenir 3085 génomes ayant un statut phénotypique, ce qui est un cas idéal mais potentiellement difficile à généraliser sur de plus petits jeux de données. Cela montre tout de même qu'il ne faut pas faire l'erreur de sous-estimer l'importance du génome nuage dans les analyses pangénomiques.

Ces tests informels dessinent des perspectives à portée de main. Il serait intéressant de reproduire ces résultats et mûrir l'ébauche de méthode statistique ici discutée afin de l'intégrer à l'écosystème PPanGGOLiN, notamment pour pouvoir explorer ces résultats d'association via le graphe de pangénome partitionné.

11.6.2 Faire le lien entre PPanGGOLiN et les approches à l'échelle de la séquence

Si d'aventure certains avaient pu penser que je critiquais trop sévèrement les graphes de pangénome à l'échelle de la séquence à la sous-section 6.2.1 page 107, je tiens ici à les rassurer. Pour éviter tout malentendu, je me dois donc en effet, de préciser que l'approche proposée par PPanGGOLiN est loin de vouloir écarter les approches au niveau de la séquence, simplement elle considère ces deux niveaux d'information comme complémentaires. En outre, au lieu de faire coexister indépendamment ces deux approches, il semble au contraire très pertinent de les emboîter pour étendre les possibilités de compréhension de la dynamique des génomes.

On peut ici faire le lien avec ce que développe actuellement pour la microbiologie l'équipe de Zamin Iqbal via l'outil Pandora (Colquhoun *et al.*, en préparation). Pionnier dans l'usage de graphe de De Bruijn pour la pangénomique (IQBAL et al., 2012), celui-ci développe désormais des outils pour la microbiologie et semble lui aussi s'orienter vers une approche emboîtant ces deux niveaux d'information pour la compréhension de la dynamique des génomes. En effet, pour l'outil Pandora, le pangénome est bien représenté à l'échelle de la séquence mais, au lieu de construire un énorme graphe contenant toutes les séquences, l'outil se base sur des familles de gènes indépendantes en construisant puis utilisant le graphe de séquences de chacune de celles-ci. L'outil Pandora construit également les graphes des régions intergéniques obtenues avec l'outil Piggy (THORPE et al., 2018) que nous avons évoqué à la sous-section 5.1.9 page 86. Cependant, l'intégration des deux niveaux d'information n'est pas complète dans Pandora. En effet, les graphes que l'outil manipule pour chaque famille (et région intergénique) sont ici « floating » [sic] (indépendants). Mais il serait tout à fait possible de les relier à un niveau d'abstraction supérieur comme dans le graphe

de PPanGGOLiN, de sorte qu'on imagine ici la forte complémentarité possible entre les deux méthodes [1]. Par exemple, en alignant les lectures issues du séquençage d'un génome inconnu sur ces graphes locaux, il est possible d'avoir un niveau de compréhension en termes de présence/absence de gènes par rapport aux génomes connus d'une part, et un niveau de compréhension en termes de génotypage et de polymorphisme nucléotidique d'autre part. Grâce au niveau gène, il est possible d'accéder au niveau de compréhension supérieur, celui de la détection des régions de plasticité génomique ainsi que nous l'avons vu dans le chapitre 11 page 189.

11.6.3 Autres perspectives

Après un alignement sur le pangénome, en comptant le nombre de lectures associées à chaque famille de gènes, il devient aussi possible de faire de la métagénomique quantitative en utilisant le génome persistant pour déterminer l'abondance de l'espèce et le génome coquille pour identifier et quantifier la présence éventuelle de différentes souches d'une même espèce dans un échantillon. De plus avec des longues lectures, il semble imaginable d'ajouter directement des chemins au graphe de pangénome. J'ai eu l'occasion de participer à deux réunions de recherche avec un étudiant en thèse, Kévin Da Silva, sous la direction de Pierre Peterlongo et Nicolas Pons qui mènent actuellement des recherches dans ce sens. Je lui adresse mes encouragements.

De même avec la complémentarité des approches à l'échelle de la famille de gènes et de la séquences, il devient loisible d'imaginer que nous puissions améliorer l'interface de visualisation de PanGBank pour explorer avec différents niveaux de zoom ces différents niveaux. En effet, il serait possible d'aller de l'échelle de l'espèce jusqu'au niveau variant nucléotidique en passant par le niveau région génomique et gène. Si l'on ajoute à cela des métainformations fonctionnelles, sur les populations, sur la génomique d'association, on conçoit là un outil susceptible de présenter un fort intérêt pour la communauté scientifique.

11.7 Conclusion générale

À la lecture de ces perspectives on peut avoir l'impression d'une liste au père Noël et je me dois d'avouer qu'il y a de ça en effet. Pourtant cette liste demeure crédible car, en définitive, la pangénomique est un paradigme de mise en relation de toutes les composantes de la génomique. Selon moi, l'intégration des différents niveaux de compréhension sera le point capital à traiter dans la pangénomique à venir.

Au-delà des procaryotes, l'approche PPanGGOLiN serait probablement bénéfique à d'autres domaines du vivant. En témoigne le fait d'avoir été invités avec mon directeur de thèse David Vallenet à participer à un séminaire de pangénomique végétale, à l'invitation du « Groupement d'Intérêt Scientifique Biotechnologies Vertes » (GIS BV) afin de présenter l'approche PPanGGOLiN. Ainsi, les progrès en pangénomique microbienne semblent également irriguer l'imaginaire des biologistes du végétal.

1. J'ai notamment été à l'initiative d'une réunion avec Zamin Iqbal qui s'est soldée par la promesse de rester en contact pour imaginer cette complémentarité.

On peut aussi constater ce phénomène dans le contexte de la génomique humaine où l'on découvre maintenant que l'Homme possède finalement un génome accessoire. Avec une taille de 29.5Mb (DUAN et al., 2019), il reste certes proportionnellement modeste en comparaison des procaryotes, de même son rôle fonctionnel reste à élucider. Mais par-delà le génome accessoire, les approches basées sur des graphes permettent également de mieux rendre compte des différentes variations structurales contenues dans les génomes. Cela semble essentiel pour appréhender plus finement les différences de génotype entre les individus.

Pour épilogue, j'exprime la conviction qu'au LABGeM, PPanGGOLiN est entre de bonnes mains afin de libérer le maximum de son potentiel bioinformatique et biologique. Avec panRGP, PanGBank, et l'intégration dans MicroScope, je laisse un travail dont je sais l'héritage assuré. Je pars avec l'espoir que ces méthodes infuseront dans la communauté. Je dis à Claudine, David, Adelme, Paul, Alexandra, prenez soin de PPanGGOLiN, davantage que l'Homme n'a su prendre soin de l'animal éponyme !



Chapitre A : Annexes des contributions scientifiques

A.1 Soutenance de thèse de doctorat

La soutenance de thèse de doctorat dont le lien est disponible ci-dessous a été enregistrée le 27/02/2020. Malgré mes quintes de toux répétées, la soutenance permet de résumer les principales notions parcourues au cours du manuscrit. N'hésitez pas à ajouter un commentaire, j'y répondrai dans la mesure du possible.

FIGURE A.1 – https://youtu.be/DwRtP7QA6rY

A.2 Présentation orale de PPanGGOLiN

J'ai présenté la conférence dont le lien est disponible ci-dessous à six reprises :

— Journées Ouvertes Biologie, Informatique et Mathématiques 2017, Lille
— Mini-congrès de l'Unité Mixte de Recherche 8030, 2018, Evry
— Séminaire sur invitation à l'INRA de Jouy-en-Josas (Equipe MetaGenopolis), 2017
— Journées Ouvertes Biologie, Informatique et Mathématiques 2018, Marseille
— Genome Informatics 2018, Cambridge (l'enregistrement a été effectué à cette occasion)
— Séminaire Symbiose sur invitation à INRIA/IRISA de Rennes (Equipe GenScale), 2019

La conférence a également été donnée par David Vallenet sur invitation à une conférence thématique dédiée à la pangénomique végétale par le GIS Biotechnologies Vertes (2019).

FIGURE A.2 – https://youtu.be/sc4-MROoLJI

A.3 Poster PPanGGOLiN

Le poster ci-dessous a été présenté lors des Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM) 2017 de Lille et lors du mini-congrès 2017 de l'Unité Mixte de Recherche 8030 à Evry.

GenomeOnRails: depicting microbial species diversity via a pangenome graph

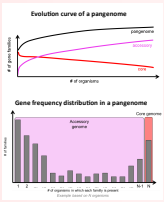
Guillaume GAUTREAU[1], Rémi PLANEL[1], Amandine PERRIN[2], Marie TOUCHON[2], Eduardo ROCHA[2], Christophe AMBROISE[3], Catherine MATIAS[4], Stéphane CRUVEILLER[1], Claudine MEDIGUE[1] and David VALLENET[1]

1 CEA, Genoscope, 91000 Evry, France — 2 Microbial Evolutionary Genomics Unit, Institut Pasteur, 75724 Paris, France — 3 Laboratoire de Mathématiques et Modélisation d'Evry (LaMME), 91000 Evry, France — 4 Laboratoire de Probabilités et Modèles Aléatoires (LPMA), 75252 Paris, France

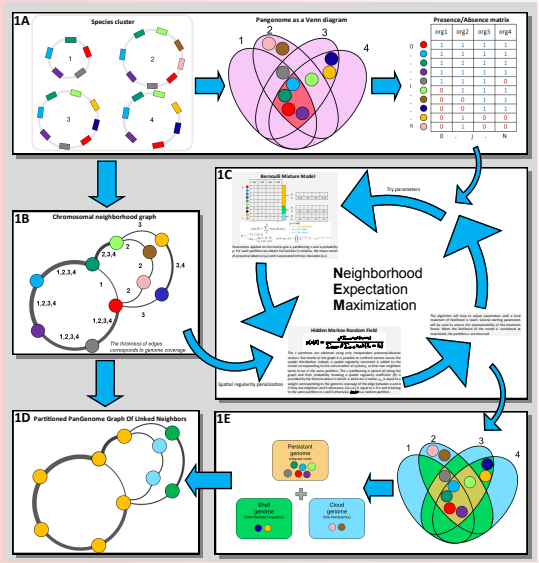
Introduction to pangenomics

Comparative genomics approaches in microbiology now rely on thousands of genomes to analyze a given species in different environmental or medical contexts. By collecting and comparing these genomic sequences, many studies focus on the overall gene content of a species (i.e. the pangenome) to understand its evolution in terms of core and accessory parts [1]. The core genome is defined as the set of genes shared by all the organisms of a taxonomic unit (generally a species). Accessory part (variable regions) is crucial to understand the adaptive potential of bacteria and contains genomic regions that are exchanged between strains by horizontal transfers (i.e. the mobilome). As recently suggested [3,4], a consensus representation of multiple genomes would provide a better analysis framework than using individual reference genomes. Here, we introduce an extension of this concept, giving it a formal mathematical representation using a graph model built up from genes clustered into families.

Some tools have been developed to characterize a pangenome like in [4,5], however none performs statistical computation to estimate the size of core genome (hereafter called persistent genome). Indeed, when new draft genomes are added to a pangenome, some genes can be missed due to potential assembly gaps or annotation artefacts. In consequence, the core genome could drop to unreasonably low number. Our approach is designed to scale to thousands of genomes without lowering the size of the persistent genome. Moreover as shown by [6], the gene frequency distribution in pangenomes bears an asymmetric U-shape regardless of the phylogenetic level. This distribution justifies a smarter partitioning approach than the traditional dichotomy between core and accessory genome. As proposed by [6], we therefore decided to split the pangenome into persistent, shell and cloud genomes.



The method generating the PPanGGOLiN



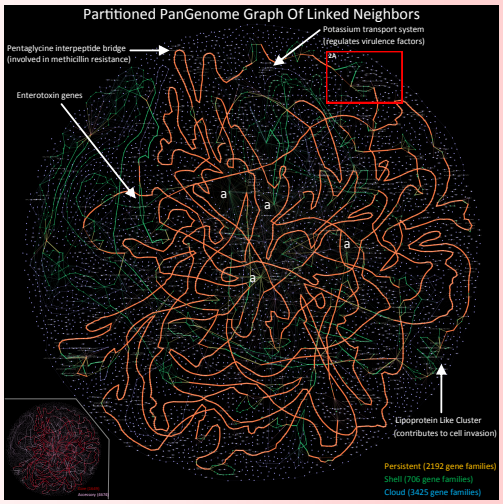
(1A) Genomes of the same species (or species clusters) are annotated and homologous genes are grouped together into gene families (colored balls) corresponding to the pangenome. Families can be organized using a Venn diagram or a presence/absence matrix. **(1B)** The physical links between genes are used to build a graph of chromosomal neighborhood (the organisms covering each edge are stored) **(1C)** The pangenome is divided by estimating the best partitioning parameters through an Expectation-Maximisation algorithm. The method involves the maximization of the likelihood of a Bernoulli Mixture Model smoothed by a Hidden Markov Random Field. This field spreads the partitions all along the graph of gene neighbors and can penalize badly classified families according to the graph. **(1E)** This approach returns a pangenome split in three partitions:

- *persistent genome:* equivalent to a relaxed core genome (genes conserved in all but a few genomes)
- *shell genome:* genes having intermediate frequencies corresponding to moderately conserved genes potentially associated to environmental adaptation capabilities
- *cloud genome:* genes found at very low frequency

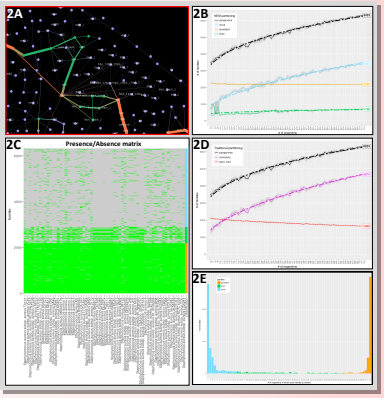
(1D) These partitions are then projected on the neighborhood graph in order to obtain what we called the **P**artitioned **PanG**enome **G**raph **O**f **LI**nked **N**eighbors.

Contact: ggautrea@genoscope.cns.fr

Results



The figure above built up from 72 *S. aureus* strains is called a PPanGGOLiN. Families have been generated using a threshold of 85% of aminoacid identity over at least 80% of the protein length. Highly connected nodes (**a**) correspond to transposable elements. Orange paths draw a kind of backbone while green and blue show alternative ones. These alternative paths can harbor genomic or pathogenic islands. The frame (**2A**) shows a zoom into a region having multiple alternative paths involved in the synthesis of fibronectin-binding protein (implicated in various infections).



The evolution curves (**2B** and **2D**) show the evolution of the partition sizes when new organisms are added to the *S. aureus* pangenome. In this representation, persistent genome remains stable whatever the number of organisms while strict core slightly decreases. This stability allows the comparison of pangenomes.
The gene frequency distribution (**2E**) and the presence/absence matrix (**2C**) confirm the relevance of the used partitioning approach.



Conclusion

Pangenomics is a relevant strategy for very large scale comparative genomics. Graph-based pangenomics tools, as described here, lead to a paradigm shift from reference centric approach to pangenome centric approach. Further developments should provide solid bases to compare pangenomes and to study their dynamics. This approach could also be useful to build almost exhaustive references for metagenomics studies.

References

1- Tettelin, H., Riley, D., Cattuto, C., & Medini, D. (2008). Comparative genomics: the bacterial pan-genome. Current Opinion in Microbiology.
2- The computation Pan-Genomics Consortium(2016). Computational pan-genomics: status, promises and challenges. Briefing in Bioinformatics.
3- Chan, A. P., Sutton, G., Depew, J., Krishnakumar, R., Choi, Y., Huang, X.-Z., ... Fouts, D. E. (2011). A novel method of consensus pan- chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii.
4- Fouts D. E., Brinkac L., Beck E., Inman J. and Sutton G (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res.
5- Zhao Y., Wu J., Yang J., Sun S., Xiao J. and Yu J (2012). PGAP: pan-genomes analysis pipeline. Bioinformatics.
6- Lobkovsky, A. E., Wolf, Y. I., & Koonin, E. V. (2013). Gene frequency distributions reject a neutral model of genome evolution. Genome Biology and Evolution.
7- Ambroise, C., Dang, M., & Govaert, G. (1997). Clustering of spatial data by the EM Algorithm. geoENV I —Geostatistics for Environmental Applications, 493–504.

FIGURE A.3 – Poster PPanGGOLiN

A.4 Poster PanGBank

Le poster ci-dessous a été présenté lors des JOBIM 2019 de Nantes.

PanGBank: depicting microbial species diversity via PPanGGOLiN

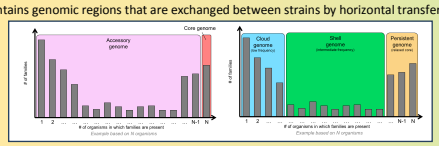
Guillaume GAUTREAU[1], Adelme BAZIN[1], Rémi PLANEL[1], Laura BURLOT[1], Mathieu DUBOIS[1], Mathieu GACHET[1], Amandine PERRIN[2], Marie TOUCHON[2], Eduardo ROCHA[2], Christophe AMBROISE[3], Catherine MATIAS[4], Claudine MEDIGUE[1] and David VALLENET[1]

1 LABGeM, Génomique Métabolique, CEA, Genoscope, CNRS, Evry University, 91000 Evry, France — 2 MEG, Institut Pasteur, CNRS, UMR3525, 75015 Paris, France — 3 LaMME, Evry University, 91000 Evry, France — 4 LPMA, Sorbonne University, 75005 Paris, France

Introduction to pangenomics

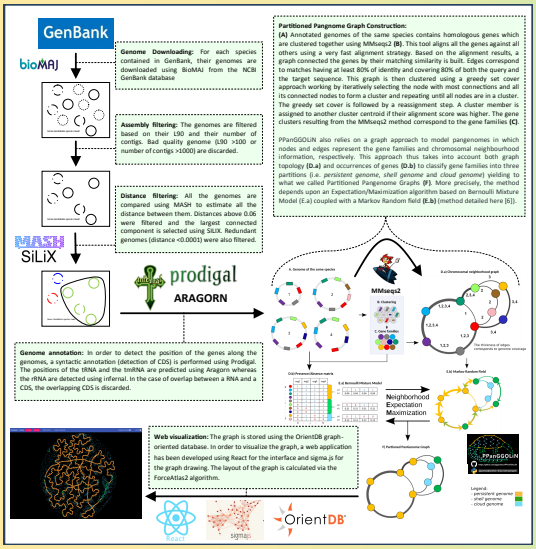
Comparative genomics approaches in microbiology now rely on thousands of genomes to analyze a given species in different environmental or medical contexts. By collecting and comparing these genomic sequences, many studies focus on the overall gene content of a species (i.e. the **pangenome**) to understand its evolution in terms of core and accessory parts [1]. The **core genome** is defined as the set of genes shared by all the organisms of a taxonomic unit (generally a species). **Accessory genome** (variable regions) is crucial to understand the adaptive potential of bacteria and contains genomic regions that are exchanged between strains by horizontal transfers.



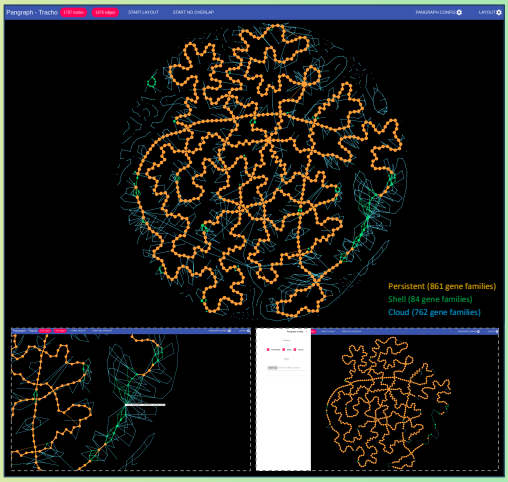


However, the concept of accessory genome lacks subtlety as it gathers genes having variable ranges of frequency. Indeed, the gene frequency distribution in pangenomes bears an asymmetric U-shape regardless of the phylogenetic level justifying a smarter partitioning approach than the traditional dichotomy between core and accessory genome. Some tools have been developed to characterize a pangenome [2,3], including several web platforms [4,5]. None performs statistical computations to partition the content of pangenome. Moreover, in these approaches, genes are considered only as a set of topologically independent objects neglecting the physical relationship linking them. To overcome these issues the method PPanGGOLiN partitions the pangenome into in persistent (equivalent to a relaxed core), shell (genes having intermediate frequencies corresponding to moderately conserved genes potentially associated to environmental adaptation capabilities) and cloud genome (genes found at a low frequency) and introduce a compact modelization (i.e. **Partitioned Pangenome Graph**) of multiple genomes using a neighbourhood graph coupled with a statistical method. PanGBank applies this method on all the species available in the GenBank database (encompassing ≈600 species containing ≈200 000 genomes) to obtain a database of PPGs. This in-development resource provides a wide view of the different range of gene frequencies and chromosomal topologies among the microbial world thanks to a web visualization tool.

The PanGBank Workfow



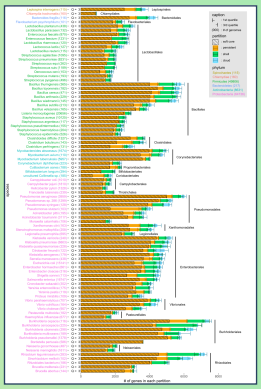
Partitioned Pangenome Graph Viewer



The figure above shows a web application to browse a Partitioned Pangenome Graph built up from 160 genomes of *Chlamydia trachomatis*. Orange paths (persistent) draw a kind of backbone of the species while green (shell) and blue (cloud) show alternative ones corresponding to regions of genomic plasticity like genomic islands or regions in genomic reduction. The left frame shows a zoom into a region having multiple alternative paths. Hovering a node allows to show the name of the gene families and several metadata like its biological function. The frame on the left illustrates that is it possible to selected only a particular genome or a bunch of genomes in order to compare their genomic topology.

PanGBank global results

The figure on the right shows the number of genes (median and interquartile range) belonging to the different partitions (persistent, shell, cloud) in the 88 most populated species of GenBank (having at least 100 genomes). The proportion of the soft core genome (hatched area) corresponds to the gene families present in 95% of the genomes of the species.



The figure shows a large range of variation in terms of proportions of the different partitions in the genomes of each species (even in the same taxonomy order). For example, in the Burkholderiales, one species (*Bordetella Pertussis*) is only composed of persistent genome whereas others contain up to 35% of shell and cloud genome.

Conclusion

Pangenomics is a relevant strategy for very large scale comparative genomics. As described here, graph-based pangenomics tools lead to a paradigm shift from a reference-genome centric approach to a pangenome centric approach. In this context, drawing genomes on rails like a subway map may help biologists to compare their genomes of interest to the overall species diversity. Further developments should provide solid bases to compare pangenomes and to study their dynamics. This approach could also be useful to build comprehensive references for metagenomics studies.

Contact

ggautrea@genoscope.cns.fr
@GUILLAUMEGAUTRE
abazin@genoscope.cns.fr
@axbazin
vallenet@genoscope.cns.fr
@vallenet

References

1- Tettelin H., Riley D., Cattuto C., and Medini D. (2008). Comparative genomics: the bacterial pan-genome. Current Opinion in Microbiology.
2- Fouts D. E., Brinkac L., Beck E., Inman J. and Sutton G (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res.
3- Zhao Y., Wu J., Yang J., Sun S., Xiao J. and Yu J. (2012). PGAP: pan-genomes analysis pipeline. Bioinformatics.
4- Médigue C., Calteau A., Cruveiller S., Gachet M., Gautreau G., Josso A., Lajus A., Langlois J., Pereira H., Planel R., Roche D., Rollin J., Rouy Z., Vallenet D. (2017). MicroScope-an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data.
5- Chaudhari N.M., Gautam A., Gupta V.K., Kaur G., Dutta C., Paul S. (2018). PanGFR-HM: A Dynamic Web Resource for Pan-Genomic and Functional Profiling of Human Microbiome With Comparative Features.
6- Ambroise C., Dang M., & Govaert G. (1997). Clustering of spatial data by the EM Algorithm. geoENV I—Geostatistics for Environmental Applications, 493–504.

FIGURE A.4 – Poster PanGBank

A.5 Poster panRGP

Le poster ci-dessous a été présenté lors des JOBIM 2019 de Nantes et lors de la conférence de la « Federation of European Microbiology Societies « (FEMS) 2019 à Glasgow.

Comparative microbial pangenomics to explore mobilome dynamics

Adelme Bazin[1], Guillaume Gautreau[1] , Claudine Médigue[1], Alexandra Calteau[1] and David Vallenet[1]

[1]LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, CNRS, Université d'Évry, Université Paris-Saclay, Evry, France.

Regions of genomic plasticity

Regions of genomic plasticity (RGP) are areas of the genome that encompass most of the strain variabilities within a genome including Genomic Islands (GI) such as pathogenicity islands, antibiotic resistance genes, environmental adaptations and secondary metabolites. For this reason they have raised a lot of attention from the scientific community, and numerous tools have been created to find them [1]. Most of them use sequence composition only and expect to differentiate them from the 'core' genome as they tend to not be fully adapted to the genome's nucleotide composition. With the advent of Next Generation Sequencing and the increase of available genomes, methods based on comparative genomics were developed to further improve RGP detection as those approaches can detect actual variations between genomes. However, as of today, no method is truly scalable to cope with the massive increase of available genomic data.

Pangenomics

For the last decade, pangenomics has provided new tools for researchers to estimate genomic diversity by partitioning gene families in terms of *core* and *accessory* genome [2]. The *core* genome consists in ubiquitous genes within the taxonomic group being studied and the *accessory* genome are the genes present in one or some individuals but not all. However, the content of the *core* genome is highly dependent on the number of genomes included in a study limiting the relevance of comparisons between studies. Moreover, the concept of *accessory* genome lacks subtlety as it gathers genes with a large range of frequencies.

Recently, a new tool named **PPanGGOLiN** (Gautreau *et al.*, in preparation) was developed to exploit gene neighborhood, gene frequency and population structure to classify gene families using a graph-based approach. 3 classes are obtained : p*ersistent*, which is a relaxed definition of the *core* genome, *shell* which corresponds to genes belonging to some individuals of the population and associated to environmental adaptations, and *cloud* which are genes present at very low frequencies.

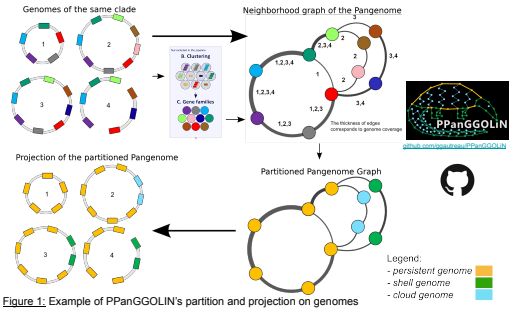


Figure 1: Example of PPanGGOLIN's partition and projection on genomes

Detecting RGP through pangenomics

RGP have a high turnover in terms of gene content. At the scale of the pangenome, we expect them to be made in majority of *shell* and *cloud* gene families. Thus, we can easily extract RGP by searching through the genomes for regions mostly *variable*. We detect them by projecting the pangenome partition on a genome and annotating genes with the partition of their corresponding gene families. Then we use a global scoring approach named **panRGP** where we try to maximize a score penalized by *persistent* genes and improved by *variable* genes to extract potential RGP in each genomes. Its simplicity makes it easily scalable to thousands of genomes as long as the pangenomic analysis scales as well.

To evaluate this approach, we tested it on a curated literature-based dataset from [1]. We built the pangenome of each of the involved species using all the genomes from genbank that have less than 1000 contigs and L90 < 100. We used the GI predictions available from [1] for the other tools.

Tool	MCC	F1-score	Accuracy	Precision	Recall
panRGP	**0.833**	**0.912**	**0.909**	**1.0**	**0.839**
IslandViewer 4	0.71	0.826	0.832	0.996	0.705
IslandPath-DIMOB	0.573	0.696	0.736	0.995	0.535
Islander	0.383	0.442	0.595	**1.0**	0.284
SIGI-HMM	0.332	0.372	0.55	**1.0**	0.228

Table 1: Quality metrics for each tool on a literature dataset

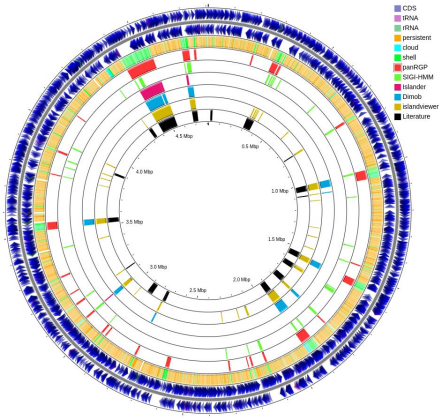


Figure 2: RGP predictions in *Salmonella enterica subsp. enterica serovar Typhi* str. CT18 (NC_003198.1) From inside out the different circles are : Literature dataset, IslandViewer 4, IslandPath-Dimob, Islander, SIGI-HMM, panRGP, PPanGGOLiN partitions, forward strand CDS, reverse strand CDS. Made with [5]

Modularizing regions and detecting co-localized and co-occurring gene sets

Genes involved in the same function tend to be co-localized in genomes [3]. Functions encoded by colocalized genes can be transferred in a single horizontal gene transfer event. If the function provides a fitness advantage it will remain active in the genomes and the genes will be kept by the organisms. Therefore we expect gene families that are in co-localized RGP and co-occurrent in the pangenome to be functionally linked. We use a cover Itemset Mining algorithm [4] using a Jaccard distance to extract gene sets that match defined criteria.

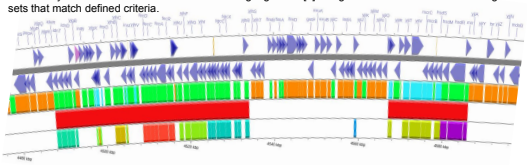
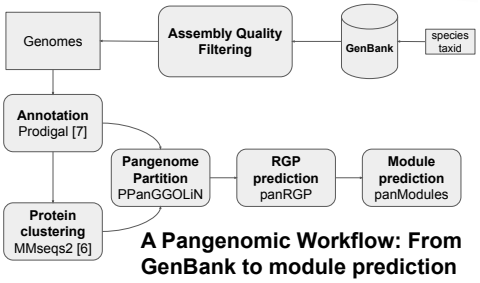


Figure 3: Example of Module predictions in RGP of *Escherichia coli* K-12. From up to down : gene annotation, forward strand CDS, reverse strand CDS, PPanGGOLiN partitions, panRGP predictions, module predictions. Made using [5]



References

[1] Bertelli, C., Tilley, K. E., & Brinkman, F. S. (2018). Microbial genomic island discovery, visualization and analysis. *Briefings in bioinformatics*.
[2] Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., ... & DeBoy, R. T. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39), 13950-13955.
[3] Overbeek, R., Fonstein, M., D'souza, M., Pusch, G. D., & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*, *96*(6), 2896-2901.
[4] Segond, M., & Borgelt, C. (2011). Item set mining based on cover similarity. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 493-505). Springer, Berlin, Heidelberg.
[5] Grant, J. R., & Stothard, P. (2008). The CGView Server: a comparative genomics tool for circular genomes. *Nucleic acids research*, *36*, W181-W184.
[6] Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, *35*(11), 1026.
[7] Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, *11*(1), 119.

FIGURE A.5 – Poster panRGP

A.6 Poster MicroScope

Le poster ci-dessous a été présenté (en tant que co-auteur) lors des JOBIM 2017 de Lille, 2018 de Marseille et 2019 de Nantes.

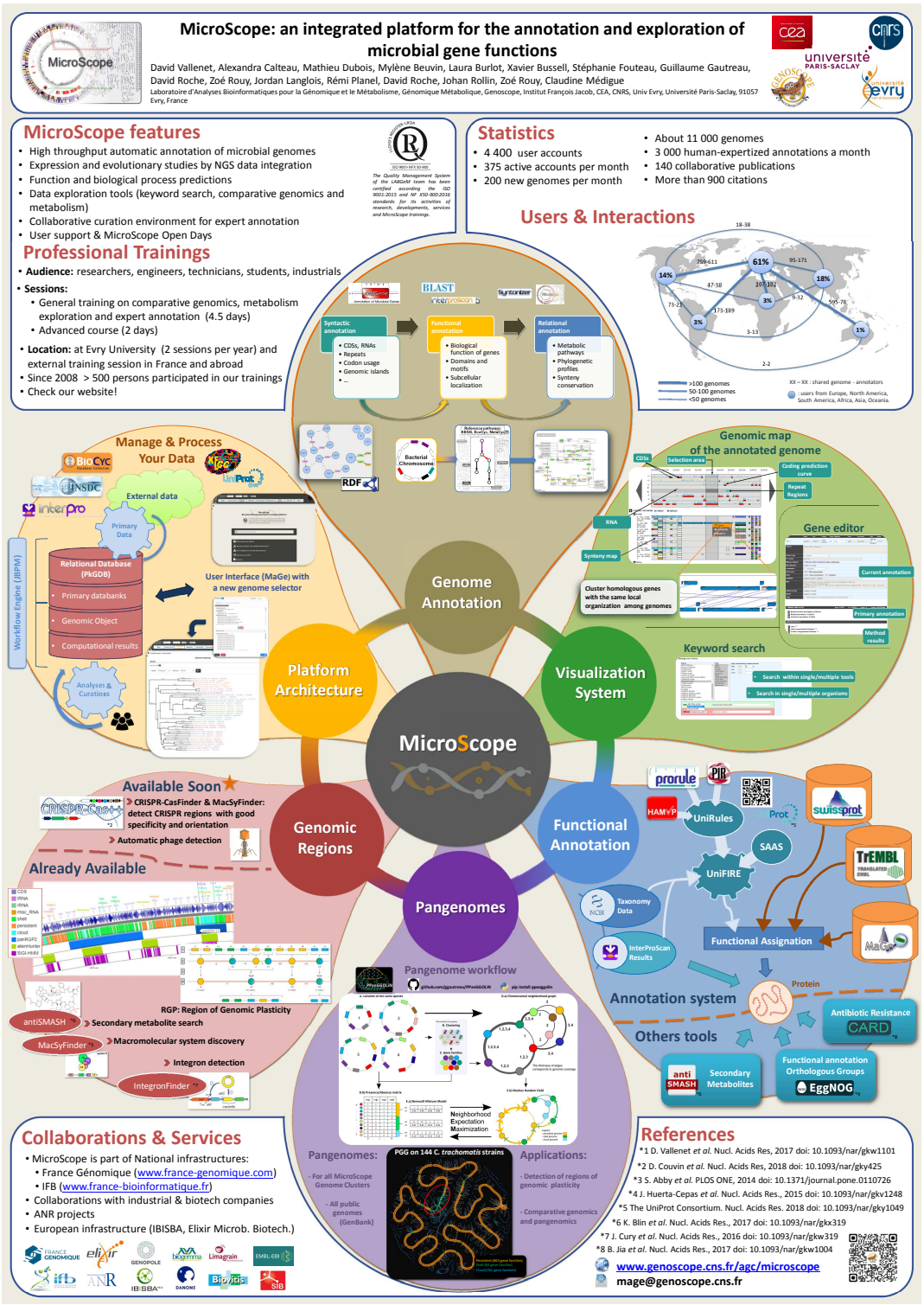


FIGURE A.6 – Poster MicroScope



Bibliographie

ACEVEDO-ROCHA, C. G. et al. (mai 2013). “From essential to persistent genes : a functional approach to constructing synthetic life”. In : *Trends Genet.* 29.5, pages 273-279 (cf. page 101).

ADDO, M. M. et al. (fév. 2003). “Comprehensive epitope analysis of human immunodeficiency virus type 1 (HIV-1)-specific T-cell responses directed against the entire expressed HIV-1 genome demonstrate broadly directed responses, but no correlation to viral load”. In : *J. Virol.* 77.3, pages 2081-2092 (cf. page 34).

AKOPYANTS, N. S. et al. (oct. 1998). “PCR-based subtractive hybridization and differences in gene content among strains of Helicobacter pylori”. In : *Proc. Natl. Acad. Sci. U.S.A.* 95.22, pages 13108-13113 (cf. page 39).

ALEXEYENKO, A. et al. (juil. 2006). “Automatic clustering of orthologs and inparalogs shared by multiple proteomes”. In : *Bioinformatics* 22.14, pages 9-15 (cf. page 83).

ALNEBERG, J. et al. (nov. 2014). “Binning metagenomic contigs by coverage and composition”. In : *Nat. Methods* 11.11, pages 1144-1146 (cf. page 173).

ALTENHOFF, A. M. et C. DESSIMOZ (jan. 2009). “Phylogenetic and functional assessment of orthologs inference projects and methods”. In : *PLoS Comput. Biol.* 5.1, e1000262 (cf. page 81).

ALTENHOFF, A. M. et al. (jan. 2018). “The OMA orthology database in 2018 : retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces”. In : *Nucleic Acids Res.* 46.D1, pages D477-D485 (cf. page 80).

ALTSCHUL, S. F. et al. (oct. 1990). “Basic local alignment search tool”. In : *J. Mol. Biol.* 215.3, pages 403-410 (cf. pages 81, 82).

ALTSCHUL, S. F. et al. (sept. 1997). “Gapped BLAST and PSI-BLAST : a new generation of protein database search programs”. In : *Nucleic Acids Res.* 25.17, pages 3389-3402 (cf. pages 81, 82).

AMBROISE, C., M. DANG et G. GOVAERT (1997). “Clustering of Spatial Data by the EM Algorithm”. In : *geoENV I — Geostatistics for Environmental Applications*. Sous la direction d’Amílcar SOARES, Jaime GÓMEZ-HERNANDEZ et Roland FROIDEVAUX. Dordrecht : Springer Netherlands, pages 493-504. ISBN : 978-94-017-1675-8 (cf. page 120).

AMBROISE, C. et G. GOVAERT (1998). “Convergence of an EM-type algorithm for spatial clustering”. In : *Pattern Recognition Letters* 19.10, pages 919-927. ISSN : 0167-8655. DOI : https://doi.org/10.1016/S0167-8655(98)00076-2. URL : http://www.sciencedirect.com/science/article/pii/S0167865598000762 (cf. page 120).

ANDREANI, N. A., E. HESSE et M. VOS (juil. 2017). “Prokaryote genome fluidity is dependent on effective population size”. In : *ISME J* 11.7, pages 1719-1721 (cf. pages 72, 73, 164).

ARAUJO, C. L. et al. (juin 2019). “Prediction of new vaccine targets in the core genome of Corynebacterium pseudotuberculosis through omics approaches and reverse vaccinology”. In : *Gene* 702, pages 36-45 (cf. page 99).

ASIF, M., I. A. ALVI et S. U. REHMAN (2018). “Insight into Acinetobacter baumannii : pathogenesis, global resistance, mechanisms of resistance, treatment options, and alternative modalities”. In : *Infect Drug Resist* 11, pages 1249-1260 (cf. page 154).

AUTON, A. et al. (oct. 2015). “A global reference for human genetic variation”. In : *Nature* 526.7571, pages 68-74 (cf. page 38).

AVERY, O. T., C. M. MACLEOD et M. MCCARTY (fév. 1944). “Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III”. In : *J. Exp. Med.* 79.2, pages 137-158 (cf. page 23).

BALTRUS, D. A. (août 2013). “Exploring the costs of horizontal gene transfer”. In : *Trends Ecol. Evol. (Amst.)* 28.8, pages 489-495 (cf. page 74).

BAPTESTE, E. et al. (sept. 2009). “Prokaryotic evolution and the tree of life are two different things”. In : *Biol. Direct* 4, page 34 (cf. page 36).

BASTIAN, Mathieu, Sebastien HEYMANN et Mathieu JACOMY (2009). “Gephi : An Open Source Software for Exploring and Manipulating Networks”. In : URL : http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154 (cf. pages 143, 153).

BATTY, E. M. et al. (juin 2018). “Long-read whole genome sequencing and comparative analysis of six strains of the human pathogen Orientia tsutsugamushi”. In : *PLoS Negl Trop Dis* 12.6, e0006566 (cf. page 116).

BAUMDICKER, F., W. R. HESS et P. PFAFFELHUBER (2012). “The infinitely many genes model for the distributed genome of bacteria”. In : *Genome Biol Evol* 4.4, pages 443-456 (cf. pages 59, 61, 63, 65-67).

BAZIN, Adelme et al. (2020). “panRGP : a pangenome-based method to predict genomic islands and explore their diversity”. In : *bioRxiv*. DOI : 10.1101/2020.03.26.007484. eprint : https://www.biorxiv.org/content/early/2020/03/26/2020.03.26.007484.full.pdf. URL : https://www.biorxiv.org/content/early/2020/03/26/2020.03.26.007484 (cf. pages 191, 193-195, 197, 198, 201).

BEHNEL, S. et al. (2011). “Cython : The Best of Both Worlds”. In : *Computing in Science Engineering* 13.2, pages 31-39. ISSN : 1521-9615. DOI : 10.1109/MCSE.2010.118 (cf. page 139).

BENDALL, M. L. et al. (juil. 2016). “Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations”. In : *ISME J* 10.7, pages 1589-1601 (cf. page 76).

BENEDICT, M. N. et al. (jan. 2014). “ITEP : an integrated toolkit for exploration of microbial pan-genomes”. In : *BMC Genomics* 15, page 8 (cf. page 85).

BERGEY, D.H. et al. (1923). *Bergey’s Manual of Determinative Bacteriology, 1st ed.* The Williams Wilkins Co, Baltimore (cf. page 37).

BERTELLI, C. et F. S. L. BRINKMAN (juil. 2018). “Improved genomic island predictions with IslandPath-DIMOB”. In : *Bioinformatics* 34.13, pages 2161-2167 (cf. page 197).

BERTELLI, C. et al. (juil. 2017). “IslandViewer 4 : expanded prediction of genomic islands for larger-scale datasets”. In : *Nucleic Acids Res.* 45.W1, W30-W35 (cf. pages 190, 197).

BEYER, Wolfgang et al. (août 2019). “Sequence tube maps : making graph genomes intuitive to commuters”. In : *Bioinformatics*. btz597. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btz597. eprint : http://oup.prod.sis.lan/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btz597/29161637/btz597.pdf. URL : https://doi.org/10.1093/bioinformatics/btz597 (cf. page 92).

BIERNACKI, C., G. CELEUX et G. GOVAERT (juil. 2000). “Assessing a mixture model for clustering with the integrated completed likelihood”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.7, pages 719-725. ISSN : 0162-8828. DOI : 10.1109/34.865189 (cf. page 123).

BLIN, Kai. (2019). *ncbi-genome-download*. [Online ; consulté le 30-Novembre-2019 à 18 :18 GMT]. URL : https://github.com/kblin/ncbi-genome-download (cf. page 183).

BLOM, J. et al. (mai 2009). “EDGAR : a software framework for the comparative analysis of prokaryotic genomes”. In : *BMC Bioinformatics* 10, page 154 (cf. page 81).

BLOM, J. et al. (juil. 2016). “EDGAR 2.0 : an enhanced software platform for comparative gene content analyses”. In : *Nucleic Acids Res.* 44.W1, W22-28 (cf. page 82).

BLONDEL, Vincent D et al. (oct. 2008). “Fast unfolding of communities in large networks”. In : *Journal of Statistical Mechanics : Theory and Experiment* 2008.10, P10008. ISSN : 1742-5468. DOI : 10.1088/1742-5468/2008/10/p10008. URL : http://dx.doi.org/10.1088/1742-5468/2008/10/P10008 (cf. pages 184, 185, 203).

BLOOM, Burton H. (juil. 1970). “Space/Time Trade-offs in Hash Coding with Allowable Errors”. In : *Commun. ACM* 13.7, pages 422-426. ISSN : 0001-0782. DOI : 10.1145/362686.362692. URL : http://doi.acm.org/10.1145/362686.362692 (cf. page 94).

BOUGUILA, Nizar (2010). “On multivariate binary data clustering and feature weighting”. In : *Computational Statistics and Data Analysis* 54.1, pages 120-134. ISSN : 0167-9473. DOI : https://doi.org/10.1016/j.csda.2009.07.013. URL : http://www.sciencedirect.com/science/article/pii/S0167947309002618 (cf. page 125).

BRITTNACHER, M. J. et al. (sept. 2011). “PGAT : a multistrain analysis resource for microbial genomes”. In : *Bioinformatics* 27.17, pages 2429-2430 (cf. page 82).

BROCKHURST, Michael A. et al. (2019). “The Ecology and Evolution of Pangenomes”. In : *Current Biology* 29.20, R1094-R1103. ISSN : 0960-9822. DOI : https://doi.org/10.1016/j.cub.2019.08.012. URL : http://www.sciencedirect.com/science/article/pii/S0960982219310280 (cf. page 135).

BRYNILDSRUD, O. et al. (nov. 2016). “Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary”. In : *Genome Biol.* 17.1, page 238 (cf. pages 96, 203).

BURROWS, M. et D. J. WHEELER (1994). *A block-sorting lossless data compression algorithm*. Rapport technique (cf. page 89).

BUSH, E. C. et al. (fév. 2018). “xenoGI : reconstructing the history of genomic island insertions in clades of closely related bacteria”. In : *BMC Bioinformatics* 19.1, page 32 (cf. page 197).

CAMPBELL, A. (1981). “Evolutionary significance of accessory DNA elements in bacteria”. In : *Annu. Rev. Microbiol.* 35, pages 55-83 (cf. pages 34, 40, 41).

CENTERS FOR DISEASE CONTROL AND PREVENTION (2005). *Active Bacterial Core Surveillance (ABCs) Report, Emerging Infections Program Network, group B streptococcus*. [Online ; consulté le 22-Octobre-2019 à 22 :07 GMT]. URL : https://www.cdc.gov/abcs/reports-findings/survreports/gbs05.pdf (cf. page 42).

CHAN, A. P. et al. (juil. 2015). “A novel method of consensus pan- chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii”. In : *Genome Biol.* 16, page 143 (cf. pages 108-110).

CHAO, A. (déc. 1987). “Estimating the population size for capture-recapture data with unequal catchability”. In : *Biometrics* 43.4, pages 783-791 (cf. page 56).

CHARLESWORTH, B. (mar. 2009). “Fundamental concepts in genetics : effective population size and patterns of molecular evolution and variation”. In : *Nat. Rev. Genet.* 10.3, pages 195-205 (cf. page 76).

CHAUDHARI, N. M., V. K. GUPTA et C. DUTTA (avr. 2016). “BPGA- an ultra-fast pan-genome analysis pipeline”. In : *Sci Rep* 6, page 24373 (cf. pages 86, 113).

CHAUDHARI, N. M. et al. (2018). “PanGFR-HM : A Dynamic Web Resource for Pan-Genomic and Functional Profiling of Human Microbiome With Comparative Features”. In : *Front Microbiol* 9, page 2322 (cf. page 182).

CHAUVIN, A. et al. (nov. 2019). “Ocular injuries caused by less-lethal weapons in France”. In : *Lancet* 394.10209, pages 1616-1617 (cf. page 3).

CHEN, F. et al. (avr. 2007). “Assessing performance of orthology detection strategies applied to eukaryotic genomes”. In : *PLoS ONE* 2.4, e383 (cf. page 81).

CIUFO, Stacy et al. (2018). “Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI”. In : *International Journal of Systematic and Evolutionary Microbiology* 68.7, pages 2386-2392. ISSN : 1466-5026. DOI : https://doi.org/10.1099/ijsem.0.002809. URL : https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.002809 (cf. page 183).

CLARK, K. et al. (jan. 2016). “GenBank”. In : *Nucleic Acids Res.* 44.D1, pages 67-72 (cf. page 151).

CLARKE, T. H. et al. (juin 2018). “PanACEA : a bioinformatics tool for the exploration and visualization of bacterial pan-chromosomes”. In : *BMC Bioinformatics* 19.1, page 246 (cf. pages 109, 110).

CLAVERYS, J. P., M. PRUDHOMME et B. MARTIN (2006). “Induction of competence regulons as a general response to stress in gram-positive bacteria”. In : *Annu. Rev. Microbiol.* 60, pages 451-475 (cf. page 31).

COHAN, F. M. (mai 1994a). “Genetic exchange and evolutionary divergence in prokaryotes”. In : *Trends Ecol. Evol. (Amst.)* 9.5, pages 175-180 (cf. page 34).

— (1994b). “The Effects of Rare but Promiscuous Genetic Exchange on Evolutionary Divergence in Prokaryotes”. In : *The American Naturalist* 143.6, pages 965-986. ISSN : 00030147, 15375323. URL : http://www.jstor.org/stable/2462891 (cf. page 34).

COLLINS, R. E. et P. G. HIGGS (nov. 2012). “Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome”. In : *Mol. Biol. Evol.* 29.11, pages 3413-3425 (cf. pages 63, 65-67).

CONTINUUM ANALYTICS Inc. (dba Anaconda, Inc.) (2017). URL : https://%20conda.io (cf. page 140).

CONTRERAS-MOREIRA, B. et P. VINUESA (déc. 2013). “GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis”. In : *Appl. Environ. Microbiol.* 79.24, pages 7696-7701 (cf. page 84).

COX, Trevor F. et Michael A.A. COX (1991). “Multidimensional scaling on a sphere”. In : *Communications in Statistics - Theory and Methods* 20.9, pages 2943-2953. DOI : 10.1080/03610929108830679. eprint : https://doi.org/10.1080/03610929108830679. URL : https://doi.org/10.1080/03610929108830679 (cf. page 137).

CRAMÉR, Harald (sept. 1946). *Mathematical Methods of Statistics. (PMS-9)*. Princeton University Press. ISBN : 0691080046. URL : https://www.xarg.org/ref/a/0691080046/ (cf. page 203).

CRICK, F. (août 1970). “Central dogma of molecular biology”. In : *Nature* 227.5258, pages 561-563 (cf. page 22).

CRISCUOLO, Alexis (2019). “A fast alignment-free bioinformatics procedure to infer accurate distance-based phylogenetic trees from genome assemblies”. In : *Research Ideas and Outcomes* 5, e36178. DOI : 10.3897/rio.5.e36178. eprint : https://doi.org/10.3897/rio.5.e36178. URL : https://doi.org/10.3897/rio.5.e36178 (cf. page 167).

DANEK, A., S. DEOROWICZ et S. GRABOWSKI (2014). “Indexes of large genome collections on a PC”. In : *PLoS ONE* 9.10, e109384 (cf. page 88).

DANG, M. (1998). “Classification de donnees spatiales : modeles probabilistes et criteres de partitionnement”. 1998COMP1173. Thèse de doctorat, 250 p. URL : http://www.theses.fr/1998COMP1173 (cf. pages 120, 139).

DANG, M. et G. GOVAERT (1998). “Spatial Fuzzy Clustering using EM and Markov Random Fields”. In : *InternationalJournal of System Research and Information Science*, pages 183-202 (cf. pages 120, 122).

DAVIDSEN, T. et al. (jan. 2010). “The comprehensive microbial resource”. In : *Nucleic Acids Res.* 38.Database issue, pages D340-345 (cf. page 80).

DAWKINS, R (1976). *The Selfish Gene*. Oxford University Press, Oxford, UK (cf. page 75).

DE BRUIJN, N. G. (1946). “A combinatorial problem”. In : *Proc. Koninklijke Nederlandse Academie van Wetenschappen* 49, pages 758-764. URL : https://ci.nii.ac.jp/naid/10019660672/en/ (cf. page 92).

DELCHER, A. L., S. L. SALZBERG et A. M. PHILLIPPY (fév. 2003). “Using MUMmer to identify similar regions in large sequence sets”. In : *Curr Protoc Bioinformatics* Chapter 10, Unit 10.3 (cf. page 82).

DELMONT, T. O. et A. M. EREN (2018). “Linking pangenomes and metagenomes : the Prochlorococcus metapangenome”. In : *PeerJ* 6, e4320 (cf. pages 107, 200).

DELUCA, T. F. et al. (août 2006). “Roundup : a multi-genome repository of orthologs and evolutionary distances”. In : *Bioinformatics* 22.16, pages 2044-2046 (cf. page 80).

DEMPSTER, A. P., N. M. LAIRD et D. B. RUBIN (1977). “Maximum likelihood from incomplete data via the EM algorithm”. In : *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 39.1, pages 1-38 (cf. page 119).

DEUTSCH, P. (1996). *RFC1952 : GZIP File Format Specification Version 4.3*. USA (cf. pages 142, 144).

DIDELOT, X. et D. FALUSH (mar. 2007). “Inference of bacterial microevolution using multilocus sequence data”. In : *Genetics* 175.3, pages 1251-1266 (cf. page 61).

DING, W., F. BAUMDICKER et R. A. NEHER (jan. 2018). “panX : pan-genome analysis and exploration”. In : *Nucleic Acids Res.* 46.1, e5 (cf. page 182).

DOBRINDT, U. et al. (mai 2004). “Genomic islands in pathogenic and environmental microorganisms”. In : *Nat. Rev. Microbiol.* 2.5, pages 414-424 (cf. page 29).

DOOLITTLE, W. F. (déc. 1999). “Lateral genomics”. In : *Trends Cell Biol.* 9.12, pages 5-8 (cf. page 36).

DOUGLAS, G. M. et M. G. I. LANGILLE (oct. 2019). “Current and Promising Approaches to Identify Horizontal Gene Transfer Events in Metagenomes”. In : *Genome Biol Evol* 11.10, pages 2750-2766 (cf. page 178).

DUAN, Z. et al. (juil. 2019). “HUPAN : a pan-genome analysis pipeline for human genomes”. In : *Genome Biol.* 20.1, page 149 (cf. page 206).

EDGAR, R. C. (2004). “MUSCLE : multiple sequence alignment with high accuracy and high throughput”. In : *Nucleic Acids Res.* 32.5, pages 1792-1797 (cf. pages 65, 82).

— (oct. 2010). “Search and clustering orders of magnitude faster than BLAST”. In : *Bioinformatics* 26.19, pages 2460-2461 (cf. page 86).

ENRIGHT, A. J., S. VAN DONGEN et C. A. OUZOUNIS (avr. 2002). “An efficient algorithm for large-scale detection of protein families”. In : *Nucleic Acids Res.* 30.7, pages 1575-1584 (cf. pages 83, 85, 203).

ERDOS, G. et al. (juil. 2003). “Development and characterization of a pooled Haemophilus influenzae genomic library for the evaluation of gene expression changes associated with mucosal biofilm formation in otitis media”. In : *Int. J. Pediatr. Otorhinolaryngol.* 67.7, pages 749-755 (cf. page 41).

EREN, A. M. et al. (2015). “Anvi’o : an advanced analysis and visualization platform for ’omics data”. In : *PeerJ* 3, e1319 (cf. page 173).

ERNST, Corinna et Sven RAHMANN (2013). “PanCake : A Data Structure for Pangenomes”. In : *German Conference on Bioinformatics 2013*. Sous la direction de Tim BEISSBARTH et al. Tome 34. OpenAccess Series in Informatics (OASIcs). Dagstuhl, Germany : Schloss

Dagstuhl–Leibniz-Zentrum fuer Informatik, pages 35-45. ISBN : 978-3-939897-59-0. DOI : 10.4230/OASIcs.GCB.2013.35. URL : http://drops.dagstuhl.de/opus/volltexte/2013/4231 (cf. page 92).

EVANS, Marc A., Douglas G. BONETT et Lyman L. MCDONALD (1994). "A General Theory for Modeling Capture-Recapture Data from a Closed Population". In : *Biometrics* 50.2, pages 396-405. ISSN : 0006341X, 15410420. URL : http://www.jstor.org/stable/2533383 (cf. page 54).

FALKOW, S. et al. (juin 1971). "The problems of drug-resistant pathogenic bacteria. The replication of R-factor DNA in Escherichia coli K-12 following conjugation". In : *Ann. N. Y. Acad. Sci.* 182, pages 153-171 (cf. page 36).

FEDERHEN, S. (jan. 2012). "The NCBI Taxonomy database". In : *Nucleic Acids Res.* 40.Database issue, pages D136-143 (cf. page 183).

FEDRIZZI, T. et al. (mar. 2017). "Genomic characterization of Nontuberculous Mycobacteria". In : *Sci Rep* 7, page 45258 (cf. page 116).

FEIL, E. J. et al. (jan. 2001). "Recombination within natural populations of pathogenic bacteria : short-term empirical estimates and long-term phylogenetic consequences". In : *Proc. Natl. Acad. Sci. U.S.A.* 98.1, pages 182-187 (cf. page 35).

FERRAGINA, P. et G. MANZINI (2000). "Opportunistic Data Structures with Applications". In : *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*. FOCS '00. Washington, DC, USA : IEEE Computer Society, pages 390-. ISBN : 0-7695-0850-2. URL : http://dl.acm.org/citation.cfm?id=795666.796543 (cf. pages 89, 90).

FILANGI, O. et al. (août 2008). "BioMAJ : a flexible framework for databanks synchronization and processing". In : *Bioinformatics* 24.16, pages 1823-1825 (cf. pages 181, 183).

FLEISCHMANN, R. D. et al. (juil. 1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". In : *Science* 269.5223, pages 496-512 (cf. page 43).

FOURNIER, P. E. et al. (jan. 2006). "Comparative genomics of multidrug resistance in Acinetobacter baumannii". In : *PLoS Genet.* 2.1, e7 (cf. page 196).

FOUTS, Derrick E. et al. (2012). "PanOCT : automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species". In : *Nucleic Acids Research* 40.22, e172 (cf. page 84).

FRUCHTERMAN, Thomas M. J. et Edward M. REINGOLD (1991). "Graph drawing by force-directed placement". In : *Software : Practice and Experience* 21.11, pages 1129-1164. DOI : 10.1002/spe.4380211102. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/spe.4380211102. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.4380211102 (cf. page 153).

FU, L. et al. (déc. 2012). "CD-HIT : accelerated for clustering the next-generation sequencing data". In : *Bioinformatics* 28.23, pages 3150-3152 (cf. pages 85, 87).

GALTON, D. (mar. 2009). "Did Darwin read Mendel ?" In : *QJM : An International Journal of Medicine* 102.8, pages 587-589. ISSN : 1460-2725. DOI : 10.1093/qjmed/hcp024. eprint : http://oup.prod.sis.lan/qjmed/article-pdf/102/8/587/4531009/hcp024.pdf. URL : https://doi.org/10.1093/qjmed/hcp024 (cf. page 22).

GARRISON, E. et al. (oct. 2018). “Variation graph toolkit improves read mapping by representing genetic variation in the reference”. In : *Nat. Biotechnol.* 36.9, pages 875-879 (cf. page 91).

GAUTREAU, G. et A. BAZIN (2019). *PPanGGOLiN*. [Online ; consulté le 07-Novembre-2019 à 20 :18 GMT]. URL : https://github.com/labgem/PPanGGOLiN (cf. page 140).

GAUTREAU, G. et al. (mar. 2020). “PPanGGOLiN : Depicting microbial diversity via a partitioned pangenome graph”. In : *PLoS Comput. Biol.* 16.3, e1007732 (cf. pages 111, 112, 151, 155, 157, 158, 161, 163, 165, 168, 170, 175, 177, 201).

GIBBS, R. S., S. SCHRAG et A. SCHUCHAT (nov. 2004). “Perinatal infections due to group B streptococci”. In : *Obstet Gynecol* 104.5 Pt 1, pages 1062-1076 (cf. page 42).

GILLESPIE, J. H. (sept. 1984). “Molecular evolution over the mutational landscape”. In : *Evolution* 38.5, pages 1116-1129 (cf. page 67).

GOGARTEN, J. P. et J. P. TOWNSEND (sept. 2005). “Horizontal gene transfer, genome innovation and evolution”. In : *Nat. Rev. Microbiol.* 3.9, pages 679-687 (cf. page 71).

GOUY, M. et C. GAUTIER (nov. 1982). “Codon usage in bacteria : correlation with gene expressivity”. In : *Nucleic Acids Res.* 10.22, pages 7055-7074 (cf. page 35).

GOVAERT, G. (1990). “Classification binaire et modèles”. fr. In : *Revue de Statistique Appliquée* 38.1, pages 67-81. URL : http://www.numdam.org/item/RSA_1990__38_1_67_0 (cf. pages 118, 119).

GRUNING, B. et al. (juil. 2018). “Bioconda : sustainable and comprehensive software distribution for the life sciences”. In : *Nat. Methods* 15.7, pages 475-476 (cf. page 140).

GUINDON, S. et al. (mai 2010). “New algorithms and methods to estimate maximum-likelihood phylogenies : assessing the performance of PhyML 3.0”. In : *Syst. Biol.* 59.3, pages 307-321 (cf. page 65).

GUMIERE, Thiago et al. (2018). “A probabilistic model to identify the core microbial community”. In : *bioRxiv*. DOI : 10.1101/491183. eprint : https://www.biorxiv.org/content/early/2018/12/10/491183.full.pdf. URL : https://www.biorxiv.org/content/early/2018/12/10/491183 (cf. page 105).

HAAS, Andreas et al. (juin 2017). “Bringing the Web up to Speed with WebAssembly”. In : *SIGPLAN Not.* 52.6, pages 185-200. ISSN : 0362-1340. DOI : 10.1145/3140587.3062363. URL : https://doi.org/10.1145/3140587.3062363 (cf. page 187).

HACKER, J. et al. (mar. 1997). “Pathogenicity islands of virulent bacteria : structure, function and impact on microbial evolution”. In : *Mol. Microbiol.* 23.6, pages 1089-1097 (cf. page 36).

HAEGEMAN, B. et J. S. WEITZ (mai 2012). “A neutral theory of genome evolution and the frequency distribution of genes”. In : *BMC Genomics* 13, page 196 (cf. pages 63, 66).

HAWKEY, Jane et al. (2019). “Impact of insertion sequences on convergent evolution of Shigella species”. In : *bioRxiv*. DOI : 10.1101/680777. eprint : https://www.biorxiv.org/content/early/2019/06/24/680777.full.pdf. URL : https://www.biorxiv.org/content/early/2019/06/24/680777 (cf. page 169).

HEAPS, H. S. (1978). “Information Retrieval : Computational and Theoretical Aspects”. In : *Academic Press.* (cf. page 49).

HEIN, Jotun, Mikkel H. SCHIERUP et Carsten WIUF (2005). *Gene Genealogies, Variation and Evolution : A Primer in Coalescent Theory*. Oxford University Press. ISBN : 0198529961 (cf. page 59).

HELGASON, E. et al. (juin 2000). "Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis–one species on the basis of genetic evidence". In : *Appl. Environ. Microbiol.* 66.6, pages 2627-2630 (cf. page 27).

HERDAN, G. (1960). "Type-Token Mathematics. A Textbook of Mathematical Linguistics." In : *Mouton and Cie, The Hague* (cf. page 49).

HERRERO, J. et al. (2016). "Ensembl comparative genomics resources". In : *Database (Oxford)* 2016 (cf. page 80).

HERSHEY, A. D. et M. CHASE (mai 1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage". In : *J. Gen. Physiol.* 36.1, pages 39-56 (cf. page 23).

HILLER, N. L. et al. (nov. 2007). "Comparative genomic analyses of seventeen Streptococcus pneumoniae strains : insights into the pneumococcal supragenome". In : *J. Bacteriol.* 189.22, pages 8186-8195 (cf. page 54).

HOGG, J. S. et al. (2007). "Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains". In : *Genome Biol.* 8.6, R103 (cf. pages 51, 53, 54, 56, 57, 59, 62, 63).

HOLLEY, G., R. WITTLER et J. STOYE (2016). "Bloom Filter Trie : an alignment-free and reference-free data structure for pan-genome storage". In : *Algorithms Mol Biol* 11, page 3 (cf. pages 94, 95).

HOLLRICHER, K (2007). "Microbial systematics : Species Don't Really Mean Anything in the Bacterial World". In : *Lab Times* 5, pages 22-25 (cf. page 37).

HORSPOOL, R. Nigel (1980). "Practical fast searching in strings". In : *Software : Practice and Experience* 10.6, pages 501-506. DOI : 10.1002/spe.4380100608. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/spe.4380100608. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.4380100608 (cf. page 89).

HOWARD, A. et al. (mai 2012). "Acinetobacter baumannii : an emerging opportunistic pathogen". In : *Virulence* 3.3, pages 243-250 (cf. page 154).

HU, D. et al. (2013). "Diversity in the major polysaccharide antigen of Acinetobacter baumannii assessed by DNA sequencing, and development of a molecular serotyping scheme". In : *PLoS ONE* 8.7, e70329 (cf. page 156).

HU, Y. F. (2005). "Efficient and high quality force-directed graph drawing". In : *The Mathematica Journal* 10, pages 37-71. URL : http://www.mathematica-journal.com/issue/v10i1/contents/graph_draw/graph_draw.pdf (cf. page 153).

HUANG, L., V. POPIC et S. BATZOGLOU (juil. 2013). "Short read alignment with populations of genomes". In : *Bioinformatics* 29.13, pages i361-370 (cf. page 89).

HULSEN, T. et al. (2006). "Benchmarking ortholog identification methods using functional genomics data". In : *Genome Biol.* 7.4, R31 (cf. page 81).

HUSON, D. H. et M. STEEL (sept. 2004). "Phylogenetic trees based on gene content". In : *Bioinformatics* 20.13, pages 2044-2049 (cf. page 60).

HUYNEN, M. A. et P. BORK (mai 1998). “Measuring genome evolution”. In : *Proc. Natl. Acad. Sci. U.S.A.* 95.11, pages 5849-5856 (cf. page 84).

HYATT, D. et al. (mar. 2010). “Prodigal : prokaryotic gene recognition and translation initiation site identification”. In : *BMC Bioinformatics* 11, page 119 (cf. pages 111, 139).

ICNP (jan. 2019). “International Code of Nomenclature of Prokaryotes”. In : *Int. J. Syst. Evol. Microbiol.* 69.1A, S1-S111 (cf. page 37).

IQBAL, Z. et al. (jan. 2012). “De novo assembly and genotyping of variants using colored de Bruijn graphs”. In : *Nat. Genet.* 44.2, pages 226-232 (cf. pages 93, 204).

IRAOLA, G. et al. (nov. 2017). “Distinct Campylobacter fetus lineages adapted as livestock pathogens and human pathobionts in the intestinal microbiota”. In : *Nat Commun* 8.1, page 1367 (cf. page 116).

JACOMY, A. et G. PLIQUE (2019). *Sigma.js*. [Online ; consulté le 01-Décembre-2019 à 14 :57 GMT]. URL : https://github.com/jacomyal/sigma.js (cf. page 186).

JACOMY, Mathieu et al. (juin 2014). “ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software”. In : *PLOS ONE* 9.6, pages 1-12. DOI : 10.1371/journal.pone.0098679. URL : https://doi.org/10.1371/journal.pone.0098679 (cf. pages 153, 186).

JAILLARD, M. et al. (nov. 2018). “A fast and agnostic method for bacterial genome-wide association studies : Bridging the gap between k-mers and genetic events”. In : *PLoS Genet.* 14.11, e1007758 (cf. pages 96, 108).

JAIN, C. et al. (nov. 2018). “High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries”. In : *Nat Commun* 9.1, page 5114 (cf. pages 38, 185).

JANI, M. et R. K. AZAD (oct. 2019). “IslandCafe : Compositional Anomaly and Feature Enrichment Assessment for Delineation of Genomic Islands”. In : *G3 (Bethesda)* 9.10, pages 3273-3285 (cf. page 197).

JENSEN, L. J. et al. (jan. 2008). “eggNOG : automated construction and annotation of orthologous groups of genes”. In : *Nucleic Acids Res.* 36.Database issue, pages D250-254 (cf. pages 58, 80).

JIANG, C. et al. (mar. 2015). “ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data”. In : *BMC Bioinformatics* 16, page 72 (cf. page 130).

JORDAN, I. K. et al. (avr. 2001). “Lineage-specific gene expansions in bacterial and archaeal genomes”. In : *Genome Res.* 11.4, pages 555-565 (cf. page 39).

KANEHISA, M. et S. GOTO (jan. 2000). “KEGG : kyoto encyclopedia of genes and genomes”. In : *Nucleic Acids Res.* 28.1, pages 27-30 (cf. pages 82, 86).

KANG, D. D. et al. (2015). “MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities”. In : *PeerJ* 3, e1165 (cf. page 173).

KARCAGI, I. et al. (mai 2016). “Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining”. In : *Mol. Biol. Evol.* 33.5, pages 1257-1269 (cf. page 75).

KATSIR, L. et al. (jan. 2018). “Genome Sequence of "Candidatus Carsonella ruddii" Strain BT from the Psyllid Bactericera trigonica”. In : *Genome Announc* 6.4 (cf. page 24).

KENT, W. J. et al. (mai 2002). “The Human Genome Browser at UCSC”. In : *Genome Research* 12.6, pages 996-1006. DOI : 10.1101/gr.229102. URL : https://doi.org/10.1101/gr.229102 (cf. page 187).

KETTLER, G. C. et al. (déc. 2007). “Patterns and implications of gene gain and loss in the evolution of Prochlorococcus”. In : *PLoS Genet.* 3.12, e231 (cf. page 47).

KIM, D. et al. (août 2019). “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. In : *Nat. Biotechnol.* 37.8, pages 907-915 (cf. page 92).

KIMURA, Motoo (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. DOI : 10.1017/CBO9780511623486 (cf. pages 72, 73).

KINGMAN, J.F.C. (1982). “The coalescent”. In : *Stochastic Processes and their Applications* 13.3, pages 235-248. ISSN : 0304-4149. DOI : https://doi.org/10.1016/0304-4149(82)90011-4. URL : http://www.sciencedirect.com/science/article/pii/0304414982900114 (cf. page 60).

KISLYUK, A. O. et al. (jan. 2011). “Genomic fluidity : an integrative view of gene diversity within microbial populations”. In : *BMC Genomics* 12, page 32 (cf. page 72).

KNUDSEN, M. et C. WIUF (fév. 2010). “The CATH database”. In : *Hum. Genomics* 4.3, pages 207-212 (cf. page 81).

KONSTANTINIDIS, K. T. et J. M. TIEDJE (fév. 2005). “Genomic insights that advance the species definition for prokaryotes”. In : *Proc. Natl. Acad. Sci. U.S.A.* 102.7, pages 2567-2572 (cf. page 38).

KOONIN, E. V. (2011). *The Logic of Chance : The Nature and Origin of Biological Evolution (FT Press Science)*. FT Press, pages 72-73 (cf. page 58).

KOONIN, E. V., K. S. MAKAROVA et L. ARAVIND (2001). “Horizontal gene transfer in prokaryotes : quantification and classification”. In : *Annu. Rev. Microbiol.* 55, pages 709-742 (cf. page 35).

KOONIN, E. V. et Y. I. WOLF (déc. 2008). “Genomics of bacteria and archaea : the emerging dynamic view of the prokaryotic world”. In : *Nucleic Acids Res.* 36.21, pages 6688-6719 (cf. pages 24, 25, 57, 63, 159).

KOSKI, L. B., R. A. MORTON et G. B. GOLDING (mar. 2001). “Codon bias and base composition are poor indicators of horizontally transferred genes”. In : *Mol. Biol. Evol.* 18.3, pages 404-412 (cf. page 190).

KOSTER, J. et S. RAHMANN (oct. 2012). “Snakemake–a scalable bioinformatics workflow engine”. In : *Bioinformatics* 28.19, pages 2520-2522 (cf. page 181).

KRISTENSEN, D. M. et al. (juin 2010). “A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches”. In : *Bioinformatics* 26.12, pages 1481-1487 (cf. page 84).

KRYAZHIMSKIY, S. et J. B. PLOTKIN (déc. 2008). “The population genetics of dN/dS”. In : *PLoS Genet.* 4.12, e1000304 (cf. page 74).

KUO, C. H. et H. OCHMAN (juin 2009). “Deletional bias across the three domains of life”. In : *Genome Biol Evol* 1, pages 145-152 (cf. page 74).

LACHENAUER, C. S. et al. (avr. 1999). “Serotypes VI and VIII predominate among group B streptococci isolated from pregnant Japanese women”. In : *J. Infect. Dis.* 179.4, pages 1030-1033 (cf. page 42).

LAING, C., M. D. WHITESIDE et V. P. J. GANNON (2017). “Pan-genome Analyses of the Species Salmonella enterica, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar”. In : *Front Microbiol* 8, page 1345 (cf. page 126).

LAING, C. et al. (sept. 2010). “Pan-genome sequence analysis using Panseq : an online tool for the rapid analysis of core and accessory genomic regions”. In : *BMC Bioinformatics* 11, page 461 (cf. page 82).

LAN, R. et P. R. REEVES (jan. 1996). “Gene transfer is a major factor in bacterial evolution”. In : *Mol. Biol. Evol.* 13.1, pages 47-55 (cf. page 35).

LANGILLE, M. G., W. W. HSIAO et F. S. BRINKMAN (août 2008). “Evaluation of genomic island predictors using a comparative genomics approach”. In : *BMC Bioinformatics* 9, page 329 (cf. page 197).

— (mai 2010). “Detecting genomic islands using bioinformatics approaches”. In : *Nat. Rev. Microbiol.* 8.5, pages 373-382 (cf. page 52).

LAPIERRE, P. et J. P. GOGARTEN (mar. 2009). “Estimating the size of the bacterial pan-genome”. In : *Trends Genet.* 25.3, pages 107-110 (cf. pages 58, 59, 63).

LARTIZIEN, R. et al. (août 2019). “Yellow vests protests : facial injuries from rubber bullets”. In : *Lancet* 394.10197, pages 469-470 (cf. page 3).

LASLETT, D. (jan. 2004). “ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences”. In : *Nucleic Acids Research* 32.1, pages 11-16. DOI : 10.1093/nar/gkh152. URL : https://doi.org/10.1093/nar/gkh152 (cf. pages 111, 139).

LAWRENCE, J. G. et H. HENDRICKSON (2005). “Genome evolution in bacteria : order beneath chaos”. In : *Current Opinion in Microbiology* 8.5. Antimicrobials / Edited by Malcolm Page and Christopher T Walsh · Genomics / Edited by Stephan C Schuster and Gerhard Gottschalk, pages 572-578. ISSN : 1369-5274. DOI : https://doi.org/10.1016/j.mib.2005.08.005. URL : http://www.sciencedirect.com/science/article/pii/S1369527405001219 (cf. pages 40, 42).

LAWRENCE, J. G. et H. OCHMAN (avr. 1997). “Amelioration of bacterial genomes : rates of change and exchange”. In : *J. Mol. Evol.* 44.4, pages 383-397 (cf. page 190).

— (août 1998). “Molecular archaeology of the Escherichia coli genome”. In : *Proc. Natl. Acad. Sci. U.S.A.* 95.16, pages 9413-9417 (cf. pages 35, 40).

LEE, M. C. et C. J. MARX (2012). “Repeated, selection-driven genome reduction of accessory genes in experimental populations”. In : *PLoS Genet.* 8.5, e1002651 (cf. page 76).

LEES, J. A. et al. (déc. 2018). “pyseer : a comprehensive tool for microbial pangenome-wide association studies”. In : *Bioinformatics* 34.24, pages 4310-4312 (cf. page 96).

LEEUW, Jan de et Patrick MAIR (2009). “Multidimensional Scaling Using Majorization : SMACOF in R”. In : *Journal of Statistical Software, Articles* 31.3, pages 1-30. ISSN : 1548-7660. DOI : 10.18637/jss.v031.i03. URL : https://www.jstatsoft.org/v031/i03 (cf. page 137).

LEFÉBURE, Tristan et Michael J. STANHOPE (mai 2007). “Evolution of the core and pan-genome of Streptococcus : positive selection, recombination, and genome composition”. In : *Genome Biology* 8.5, R71. ISSN : 1474-760X. DOI : 10.1186/gb-2007-8-5-r71. URL : https://doi.org/10.1186/gb-2007-8-5-r71 (cf. page 47).

LERAT, E., V. DAUBIN et N. A. MORAN (oct. 2003). “From gene trees to organismal phylogeny in prokaryotes : the case of the gamma-Proteobacteria”. In : *PLoS Biol.* 1.1, E19 (cf. page 81).

LEVIN, B. R. (sept. 1981). “Periodic selection, infectious gene exchange and the genetic structure of E. coli populations”. In : *Genetics* 99.1, pages 1-23 (cf. page 34).

LEVY, A. (oct. 2019). “How evolution builds genes from scratch”. In : *Nature* 574.7778, pages 314-316 (cf. page 86).

LI, L., C. J. STOECKERT et D. S. ROOS (sept. 2003). “OrthoMCL : identification of ortholog groups for eukaryotic genomes”. In : *Genome Res.* 13.9, pages 2178-2189 (cf. pages 80, 83, 85).

LI, W. et A. GODZIK (juil. 2006). “Cd-hit : a fast program for clustering and comparing large sets of protein or nucleotide sequences”. In : *Bioinformatics* 22.13, pages 1658-1659 (cf. page 116).

LIGUE CONTRE LE CANCER (2019). *Cartes d’Identité des Tumeurs*. [Online ; consulté le 18-October-2019 à 15 :01 GMT]. URL : https://www.ligue-cancer.net/article/la-recherche/carte-d-identite-des-tumeurs (cf. page 33).

LIVINGSTONE, P. G., R. M. MORPHEW et D. E. WHITWORTH (2018). “Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets”. In : *Front Microbiol* 9, page 3187 (cf. page 103).

LOBKOVSKY, Alexander E., Yuri I. WOLF et E. V. KOONIN (2013). “Gene frequency distributions reject a neutral model of genome evolution”. In : *Genome Biology and Evolution* (cf. pages 66-69).

LOCEY, K. J. et J. T. LENNON (2016). “Scaling laws predict global microbial diversity”. In : *Proceedings of the National Academy of Sciences* 113.21, pages 5970-5975. ISSN : 0027-8424. DOI : 10.1073/pnas.1521291113. eprint : https://www.pnas.org/content/113/21/5970.full.pdf. URL : https://www.pnas.org/content/113/21/5970 (cf. page 38).

LU, B. et H. W. LEONG (juin 2018). “GI-Cluster : Detecting genomic islands via consensus clustering on multiple features”. In : *J Bioinform Comput Biol* 16.3, page 1840010 (cf. page 197).

LUKJANCENKO, O., T. M. WASSENAAR et D. W. USSERY (nov. 2010). “Comparison of 61 sequenced Escherichia coli genomes”. In : *Microb. Ecol.* 60.4, pages 708-720 (cf. pages 78, 99).

LUKJANCENKO, O. et al. (2013). “PanFunPro : PAN-genome analysis based on FUNctional PROfiles [version 1 ; peer review : 3 approved with reservations]”. In : *F1000Research* 2.265. DOI : 10.12688/f1000research.2-265.v1 (cf. page 87).

LUO, C. et al. (2012). “Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample”. In : *PLoS ONE* 7.2, e30087 (cf. page 100).

MAIDEN, M. C. et al. (mar. 1998). “Multilocus sequence typing : a portable approach to the identification of clones within populations of pathogenic microorganisms”. In : *Proc. Natl. Acad. Sci. U.S.A.* 95.6, pages 3140-3145 (cf. page 35).

MAIONE, D. et al. (juil. 2005). “Identification of a universal Group B streptococcus vaccine by multiple genome screen”. In : *Science* 309.5731, pages 148-150 (cf. pages 42, 43, 45).

MAISTRENKO, Oleksandr M. et al. (2019). “Disentangling the impact of environmental and phylogenetic constraints on prokaryotic strain diversity”. In : *bioRxiv*. DOI : 10.1101/735696. eprint : https://www.biorxiv.org/content/early/2019/08/15/735696.full.pdf. URL : https://www.biorxiv.org/content/early/2019/08/15/735696 (cf. page 167).

MAKAROVA, K. S. et al. (juil. 1999). “Comparative genomics of the Archaea (Euryarchaeota) : evolution of conserved protein families, the stable core, and the variable shell”. In : *Genome Res.* 9.7, pages 608-628 (cf. page 58).

MAKAROVA, K. S. et al. (nov. 2007). “Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea”. In : *Biol. Direct* 2, page 33 (cf. pages 57, 58, 63).

MALHOTRA, S., S. C. VEDITHI et T. L. BLUNDELL (août 2017). “Decoding the similarities and differences among mycobacterial species”. In : *PLoS Negl Trop Dis* 11.8, e0005883 (cf. page 30).

MARCUS, S., H. LEE et M. C. SCHATZ (déc. 2014). “SplitMEM : a graphical algorithm for pan-genome analysis with suffix skips”. In : *Bioinformatics* 30.24, pages 3476-3483 (cf. page 93).

MARQUARDT, Donald W. (juin 1963). “An Algorithm for Least-Squares Estimation of Nonlinear Parameters”. In : *Journal of the Society for Industrial and Applied Mathematics* 11.2, pages 431-441. DOI : 10.1137/0111030. URL : https://doi.org/10.1137/0111030 (cf. page 160).

MARTIN, Shawn et al. (2011). “OpenOrd : an open-source toolbox for large graph layout”. In : *Visualization and Data Analysis 2011*. Sous la direction de Pak Chung WONG et al. Tome 7868. International Society for Optics et Photonics. SPIE, pages 45-55. DOI : 10.1117/12.871402. URL : https://doi.org/10.1117/12.871402 (cf. page 153).

MARTINEZ-MURCIA, A. J., S. BENLLOCH et M. D. COLLINS (juil. 1992). “Phylogenetic interrelationships of members of the genera Aeromonas and Plesiomonas as determined by 16S ribosomal DNA sequencing : lack of congruence with results of DNA-DNA hybridizations”. In : *Int. J. Syst. Bacteriol.* 42.3, pages 412-421 (cf. pages 37, 199).

MASSON-DELMOTTE, V. et al. (2018). “Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty”. In : *Intergovernmental Panel on Climate Change* (cf. page 3).

MATSEN, F. A., R. B. KODNER et E. V. ARMBRUST (oct. 2010). “pplacer : linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree”. In : *BMC Bioinformatics* 11, page 538 (cf. page 184).

MAYR, E (1942). *Systematics and the Origin of Species from the Viewpoint of a Zoologist*. Harvard University Press. ISBN : 0674862503 (cf. page 36).

MCDANIEL, T. K. et J. B. KAPER (jan. 1997). “A cloned pathogenicity island from enteropathogenic Escherichia coli confers the attaching and effacing phenotype on E. coli K-12”. In : *Mol. Microbiol.* 23.2, pages 399-407 (cf. page 36).

MCINERNEY, James O, Alan MCNALLY et Mary J O CONNELL (2017). "Why prokaryotes have pangenomes". In : *Nature Publishing Group* 2.March, pages 1-5. DOI : 10.1038/nmicrobiol.2017.40. URL : http://dx.doi.org/10.1038/nmicrobiol.2017.40 (cf. pages 71, 74-78, 107).

MCKINNEY, Wes (2010). "Data Structures for Statistical Computing in Python". In : *Proceedings of the 9th Python in Science Conference*. Sous la direction de Stéfan van der WALT et Jarrod MILLMAN, pages 51-56 (cf. page 139).

MÉDIGUE, C. et al. (déc. 1991). "Evidence for horizontal gene transfer in Escherichia coli speciation". In : *J. Mol. Biol.* 222.4, pages 851-856 (cf. pages 35, 190).

MÉDIGUE, C. et al. (sept. 2017). "MicroScope—an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data". In : *Briefings in Bioinformatics* 20.4, pages 1071-1084. ISSN : 1477-4054. DOI : 10.1093/bib/bbx113. eprint : https://academic.oup.com/bib/article-pdf/20/4/1071/31614296/bbx113.pdf. URL : https://doi.org/10.1093/bib/bbx113 (cf. page 201).

MEDINI, D. et al. (déc. 2005). "The microbial pan-genome". In : *Curr. Opin. Genet. Dev.* 15.6, pages 589-594 (cf. pages 33, 41-43).

MEYER, F., R. OVERBEEK et A. RODRIGUEZ (nov. 2009). "FIGfams : yet another set of protein families". In : *Nucleic Acids Res.* 37.20, pages 6643-6654 (cf. page 80).

MIELE, V., S. PENEL et L. DURET (avr. 2011). "Ultra-fast sequence clustering from similarity networks with SiLiX". In : *BMC Bioinformatics* 12, page 116 (cf. page 185).

MINKIN, I., S. PHAM et P. MEDVEDEV (déc. 2017). "TwoPaCo : an efficient algorithm to build the compacted de Bruijn graph from many complete genomes". In : *Bioinformatics* 33.24, pages 4024-4032 (cf. page 95).

MIRDITA, M., M. STEINEGGER et J. SOEDING (2019). *MMseqs2*. [Online ; consulté le 30-Novembre-2019 à 16 :18 GMT]. URL : https://github.com/soedinglab/MMseqs2 (cf. page 115).

MISTRY, J. et al. (juil. 2013). "Challenges in homology search : HMMER3 and convergent evolution of coiled-coil regions". In : *Nucleic Acids Res.* 41.12, e121 (cf. page 87).

MOLDOVAN, Mikhail A. et Mikhail S. GELFAND (2018). "Pangenomic Definition of Prokaryotic Species and the Phylogenetic Structure of Prochlorococcus spp." In : *Frontiers in Microbiology* 9, page 428. ISSN : 1664-302X. DOI : 10.3389/fmicb.2018.00428. URL : https://www.frontiersin.org/article/10.3389/fmicb.2018.00428 (cf. pages 66, 107).

MONOD, J. et F. JACOB (1961). "Teleonomic mechanisms in cellular metabolism, growth, and differentiation". In : *Cold Spring Harb. Symp. Quant. Biol.* 26, pages 389-401 (cf. page 22).

MORAN, PAP. (1962). "The statistical processes of evolutionary theory". In : *Clarendon Press* (cf. page 67).

MORRIS, J. J., R. E. LENSKI et E. R. ZINSER (2012). "The Black Queen Hypothesis : evolution of dependencies through adaptive gene loss". In : *MBio* 3.2 (cf. page 101).

MULLIS, K. et al. (1986). "Specific enzymatic amplification of DNA in vitro : the polymerase chain reaction". In : *Cold Spring Harb. Symp. Quant. Biol.* 51 Pt 1, pages 263-273 (cf. pages 27, 39).

MUNTZING, A. (1945). “Cytological studies of extra fragment chromosomes in rye ; transmission and multiplication of standard fragments and iso-fragments”. In : *Hereditas* 31.3-4, pages 457-477 (cf. page 34).

MYERS, Gene (mai 1999). “A Fast Bit-vector Algorithm for Approximate String Matching Based on Dynamic Programming”. In : *J. ACM* 46.3, pages 395-415. ISSN : 0004-5411. DOI : 10.1145/316542.316550. URL : http://doi.acm.org/10.1145/316542.316550 (cf. page 89).

NAKAMURA, Y. et al. (juil. 2004). “Biased biological functions of horizontally transferred genes in prokaryotic genomes”. In : *Nat. Genet.* 36.7, pages 760-766 (cf. page 75).

NAWROCKI, E. M., M. BRADSHAW et E. A. JOHNSON (fév. 2018). “Botulinum neurotoxin-encoding plasmids can be conjugatively transferred to diverse clostridial strains”. In : *Sci Rep* 8.1, page 3100 (cf. page 183).

NAWROCKI, E. P. et S. R. EDDY (sept. 2013). “Infernal 1.1 : 100-fold faster RNA homology searches”. In : *Bioinformatics* 29.22, pages 2933-2935. DOI : 10.1093/bioinformatics/btt509. URL : https://doi.org/10.1093/bioinformatics/btt509 (cf. pages 111, 139).

NEEDLEMAN, S. B. et C. D. WUNSCH (mar. 1970). “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In : *J. Mol. Biol.* 48.3, pages 443-453 (cf. page 28).

NIEHUS, R. et al. (nov. 2015). “Migration and horizontal gene transfer divide microbial genomes into multiple niches”. In : *Nat Commun* 6, page 8924 (cf. pages 76, 77).

NIELSEN, K. M., A. M. BONES et J. D. VAN ELSAS (oct. 1997). “Induced Natural Transformation of Acinetobacter calcoaceticus in Soil Microcosms”. In : *Appl. Environ. Microbiol.* 63.10, pages 3972-3977 (cf. page 31).

NURK, S. et al. (mai 2017). “metaSPAdes : a new versatile metagenomic assembler”. In : *Genome Res.* 27.5, pages 824-834 (cf. page 173).

O’BRIEN, K. P., M. REMM et E. L. SONNHAMMER (jan. 2005). “Inparanoid : a comprehensive database of eukaryotic orthologs”. In : *Nucleic Acids Res.* 33.Database issue, pages D476-480 (cf. pages 80, 83).

OCHMAN, H., J. G. LAWRENCE et E. A. GROISMAN (mai 2000). “Lateral gene transfer and the nature of bacterial innovation”. In : *Nature* 405.6784, pages 299-304 (cf. page 36).

OH, P. L. et al. (mar. 2010). “Diversification of the gut symbiont Lactobacillus reuteri as a result of host-driven evolution”. In : *ISME J* 4.3, pages 377-387 (cf. page 169).

OLIVEIRA, P. H. et al. (oct. 2017). “The chromosomal organization of horizontal gene transfer in bacteria”. In : *Nat Commun* 8.1, page 841 (cf. pages 109, 156, 189).

ONDOV, Brian D. et al. (juin 2016). “Mash : fast genome and metagenome distance estimation using MinHash”. In : *Genome Biology* 17.1, page 132. ISSN : 1474-760X. DOI : 10.1186/s13059-016-0997-x. URL : https://doi.org/10.1186/s13059-016-0997-x (cf. pages 38, 135, 167, 184).

ORIENTDB LTD (2019). *orientDB*. [Online ; consulté le 01-Décembre-2019 à 14 :22 GMT]. URL : https://orientdb.com/ (cf. page 186).

PAGE, A. J. et al. (nov. 2015). “Roary : Rapid large-scale prokaryote pan genome analysis”. In : *Bioinformatics* 31.22, pages 3691-3693 (cf. pages 85, 103, 105, 113, 143, 148).

PANTOJA, Y. et al. (2017). “PanWeb : A web interface for pan-genomic analysis”. In : *PLoS ONE* 12.5, e0178154 (cf. page 182).

PARISI, Tony (2012). *WebGL : Up and Running*. 1st. O’Reilly Media, Inc. ISBN : 144932357X (cf. page 187).

PARKS, D. H. et al. (juil. 2015). “CheckM : assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes”. In : *Genome Res.* 25.7, pages 1043-1055 (cf. page 184).

PARKS, D. H. et al. (nov. 2018). “A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life”. In : *Nat. Biotechnol.* 36.10, pages 996-1004 (cf. pages 38, 39, 171, 185).

PASOLLI, E. et al. (jan. 2019). “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle”. In : *Cell* 176.3, pages 649-662 (cf. pages 173, 174, 178).

PENG, Y. et al. (nov. 2018). “MetaPGN : a pipeline for construction and graphical visualization of annotated pangenome networks”. In : *Gigascience* 7.11 (cf. page 110).

PERIWAL, V. et al. (2015). “Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of Mycobacterium tuberculosis pangenome”. In : *PLoS ONE* 10.4, e0122979 (cf. page 103).

PERNA, N. T. et al. (jan. 2001). “Genome sequence of enterohaemorrhagic Escherichia coli O157 :H7”. In : *Nature* 409.6819, pages 529-533 (cf. page 40).

PLAZA ONATE, F. et al. (mai 2019). “MSPminer : abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data”. In : *Bioinformatics* 35.9, pages 1544-1552 (cf. page 200).

PLOTLY (2015). *Collaborative data science*. URL : https://plot.ly (cf. page 144).

POIREL, L. et al. (août 2003). “Outbreak of extended-spectrum beta-lactamase VEB-1-producing isolates of Acinetobacter baumannii in a French hospital”. In : *J. Clin. Microbiol.* 41.8, pages 3542-3547 (cf. page 196).

PRAY, L. A. (2008). *DNA Replication and Causes of Mutation*. [Online ; consulté le 22-Novembre-2019 à 14 :01 GMT]. URL : https://www.nature.com/scitable/topicpage/dna-replication-and-causes-of-mutation-409/# (cf. page 107).

PUNDHIR, S., H. VIJAYVARGIYA et A. KUMAR (2008). “PredictBias : a server for the identification of genomic and pathogenicity islands in prokaryotes”. In : *In Silico Biol. (Gedrukt)* 8.3-4, pages 223-234 (cf. page 197).

RAHMAN, A. et al. (juin 2018). “Association mapping from sequencing reads using k-mers”. In : *Elife* 7 (cf. page 96).

RAHN, R., D. WEESE et K. REINERT (déc. 2014). “Journaled string tree-a scalable data structure for analyzing thousands of similar genomes on your laptop”. In : *Bioinformatics* 30.24, pages 3499-3505 (cf. page 88).

RASIGADE, J. P., F. HOLLANDT et T. WIRTH (nov. 2018). “Genes under positive selection in the core genome of pathogenic Bacillus cereus group members”. In : *Infect. Genet. Evol.* 65, pages 55-64 (cf. page 99).

RASKO, D. A., G. S. MYERS et J. RAVEL (jan. 2005). “Visualization of comparative genomic analyses by BLAST score ratio”. In : *BMC Bioinformatics* 6, page 2 (cf. page 81).

RASKO, D. A. et al. (oct. 2008). “The pangenome structure of Escherichia coli : comparative genomic analysis of E. coli commensal and pathogenic isolates”. In : *J. Bacteriol.* 190.20, pages 6881-6893 (cf. pages 47, 99).

RAUSCHER, R. et Z. IGNATOVA (août 2018). “Timing during translation matters : synonymous mutations in human pathologies influence protein folding and function”. In : *Biochem. Soc. Trans.* 46.4, pages 937-944 (cf. page 108).

REMM, M., C. E. STORM et E. L. SONNHAMMER (déc. 2001). “Automatic clustering of orthologs and in-paralogs from pairwise species comparisons”. In : *J. Mol. Biol.* 314.5, pages 1041-1052 (cf. pages 80, 83).

ROCHA, E. P. C. (2008). “The Organization of the Bacterial Genome”. In : *Annual Review of Genetics* 42.1. PMID : 18605898, pages 211-233. DOI : 10.1146/annurev.genet.42.110807.091653. eprint : https://doi.org/10.1146/annurev.genet.42.110807.091653. URL : https://doi.org/10.1146/annurev.genet.42.110807.091653 (cf. pages 75, 84, 111).

ROCHA, E. P. C. et A. DANCHIN (nov. 2003). “Gene essentiality determines chromosome organisation in bacteria”. In : *Nucleic Acids Res.* 31.22, pages 6570-6577 (cf. page 108).

SAHL, J. W. et al. (2014). “The large-scale blast score ratio (LS-BSR) pipeline : a method to rapidly compare genetic content between bacterial genomes”. In : *PeerJ* 2, e332 (cf. pages 86, 87).

SAITOU, N. et M. NEI (juil. 1987). “The neighbor-joining method : a new method for reconstructing phylogenetic trees”. In : *Mol. Biol. Evol.* 4.4, pages 406-425 (cf. pages 170-172).

SANGER, F., S. NICKLEN et A. R. COULSON (déc. 1977). “DNA sequencing with chain-terminating inhibitors”. In : *Proc. Natl. Acad. Sci. U.S.A.* 74.12, pages 5463-5467 (cf. pages 28, 47).

SCHNEEBERGER, K. et al. (2009). “Simultaneous alignment of short reads against multiple genomes”. In : *Genome Biol.* 10.9, R98 (cf. page 91).

SCHNEIKER, S. et al. (nov. 2007). “Complete genome sequence of the myxobacterium Sorangium cellulosum”. In : *Nat. Biotechnol.* 25.11, pages 1281-1289 (cf. page 24).

SCHOEN, C. et al. (mar. 2008). “Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis”. In : *Proc. Natl. Acad. Sci. U.S.A.* 105.9, pages 3473-3478 (cf. page 47).

SCHWARZ, Gideon (mar. 1978). “Estimating the Dimension of a Model”. In : *Ann. Statist.* 6.2, pages 461-464. DOI : 10.1214/aos/1176344136. URL : https://doi.org/10.1214/aos/1176344136 (cf. pages 56, 123).

SELA, I., Y. I. WOLF et E. V. KOONIN (oct. 2016). “Theory of prokaryotic genome evolution”. In : *Proc. Natl. Acad. Sci. U.S.A.* 113.41, pages 11399-11407 (cf. page 74).

SELANDER, R. K. et al. (mai 1986). “Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics”. In : *Appl. Environ. Microbiol.* 51.5, pages 873-884 (cf. page 35).

SIGAUX, F. (2000). “Cancer genome or the development of molecular portraits of tumors”. In : *Bull. Acad. Natl. Med.* 184.7, pages 1441-1447 (cf. pages 33, 34).

SIGUIER, P. et al. (jan. 2006). “ISfinder : the reference centre for bacterial insertion sequences”. In : *Nucleic Acids Research* 34.suppl$_1$, pages D32-D36. ISSN : 0305-1048. DOI : 10.1093/nar/gkj014. eprint : http://oup.prod.sis.lan/nar/article-pdf/34/suppl_1/D32/3923468/gkj014.pdf. URL : https://doi.org/10.1093/nar/gkj014 (cf. page 130).

SINGH, J. K., F. G. ADAMS et M. H. BROWN (2018). “Diversity and Function of Capsular Polysaccharide in Acinetobacter baumannii”. In : *Front Microbiol* 9, page 3301 (cf. page 156).

SIRÉN, Jouni (2017). “Indexing Variation Graphs”. In : *2017 Proceedings of the Ninteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 13-27. DOI : 10.1137/1.9781611974768.2. eprint : https://epubs.siam.org/doi/pdf/10.1137/1.9781611974768.2. URL : https://epubs.siam.org/doi/abs/10.1137/1.9781611974768.2 (cf. page 91).

ŚMIEJA, Marek, Krzysztof HAJTO et Jacek TABOR (juin 2019). “Efficient mixture model for clustering of sparse high dimensional binary data”. In : *Data Mining and Knowledge Discovery*. ISSN : 1573-756X. DOI : 10.1007/s10618-019-00635-1. URL : https://doi.org/10.1007/s10618-019-00635-1 (cf. page 126).

SMITH, J. M., E. J. FEIL et N. H. SMITH (déc. 2000). “Population structure and evolutionary dynamics of pathogenic bacteria”. In : *Bioessays* 22.12, pages 1115-1122 (cf. page 35).

SMITH, T. F. et M. S. WATERMAN (mar. 1981). “Identification of common molecular subsequences”. In : *J. Mol. Biol.* 147.1, pages 195-197 (cf. page 115).

SNEL, B., P. BORK et M. A. HUYNEN (jan. 2002). “Genomes in flux : the evolution of archaeal and proteobacterial gene content”. In : *Genome Res.* 12.1, pages 17-25 (cf. page 39).

SNIPEN, L., T. ALMOY et D. W. USSERY (août 2009). “Microbial comparative pan-genomics using binomial mixture models”. In : *BMC Genomics* 10, page 385 (cf. pages 54, 56, 57, 59, 63, 87, 99, 105).

SNIPEN, L. et K. H. LILAND (mar. 2015). “micropan : an R-package for microbial pan-genomics”. In : *BMC Bioinformatics* 16, page 79 (cf. pages 57, 87, 99, 105, 106).

SNIPEN, L. et D. W. USSERY (2012). “A domain sequence approach to pangenomics : applications to Escherichia coli”. In : *F1000Res* 1, page 19 (cf. page 87).

SONG, L. et al. (mar. 2013). “Chlamydia trachomatis plasmid-encoded Pgp4 is a transcriptional regulator of virulence-associated genes”. In : *Infect. Immun.* 81.3, pages 636-644 (cf. page 154).

STACKBRANDT, E. et B. M. GOEBEL (1994). “Taxonomic Note : A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology”. In : *International Journal of Systematic and Evolutionary Microbiology* 44.4, pages 846-849. ISSN : 1466-5026. DOI : https://doi.org/10.1099/00207713-44-4-846. URL : https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-44-4-846 (cf. page 37).

STACKEBRANDT, E. et J. EBERS (nov. 2006). “Taxonomic parameters revisited : tarnished gold standards”. In : *Microbiol Today* 33, pages 152-155 (cf. page 37).

STEINEGGER, Martin et Johannes SOEDING (2017). “Sensitive protein sequence searching for the analysis of massive data sets”. In : *Nat Biotech*. DOI : doi:10.1038/nbt.3988 (cf. pages 111, 113, 114, 139, 174).

TAMAMES, J. (2001). “Evolution of gene order conservation in prokaryotes”. In : *Genome Biol.* 2.6, RESEARCH0020 (cf. page 84).

TATUSOV, R. L. et al. (sept. 2003). “The COG database : an updated version includes eukaryotes”. In : *BMC Bioinformatics* 4, page 41 (cf. pages 81, 86).

TEAM, PyTables Developers (2002). *PyTables : Hierarchical Datasets in Python*. URL : https://www.pytables.org/ (cf. page 143).

TETTELIN, H. et al. (sept. 2005). “Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae : implications for the microbial "pan-genome"”. In : *Proc. Natl. Acad. Sci. U.S.A.* 102.39, pages 13950-13955 (cf. pages 33, 41-43, 45-49, 51, 65, 66, 82, 99).

TETTELIN, H. et al. (oct. 2008). “Comparative genomics : the bacterial pan-genome”. In : *Curr. Opin. Microbiol.* 11.5, pages 472-477 (cf. pages 47, 49, 51, 62, 63, 82).

TETZ, V. V. (juil. 2005). “The pangenome concept : a unifying view of genetic information”. In : *Med. Sci. Monit.* 11.7, Y24-29 (cf. page 42).

THE COMPUTATIONAL PAN-GENOMICS CONSORTIUM (oct. 2016). “Computational pan-genomics : status, promises and challenges”. In : *Brief. Bioinformatics* (cf. page 87).

THE HDF GROUP (2000-2010). *Hierarchical data format version 5*. URL : http://www.hdfgroup.org/HDF5 (cf. pages 140, 143).

THORPE, H. A. et al. (avr. 2018). “Piggy : a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria”. In : *Gigascience* 7.4, pages 1-11 (cf. pages 86, 204).

TONDER, A. J. van et al. (août 2014). “Defining the estimated core genome of bacterial populations using a Bayesian decision model”. In : *PLoS Comput. Biol.* 10.8, e1003788 (cf. page 105).

TONDER, A. J. van et al. (2019). “Genomic Analyses of >3,100 Nasopharyngeal Pneumococci Revealed Significant Differences Between Pneumococci Recovered in Four Different Geographical Regions”. In : *Frontiers in Microbiology* 10, page 317. ISSN : 1664-302X. DOI : 10.3389/fmicb.2019.00317. URL : https://www.frontiersin.org/article/10.3389/fmicb.2019.00317 (cf. page 57).

TOUCHON, M. et al. (jan. 2009). “Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths”. In : *PLoS Genet.* 5.1, e1000344 (cf. pages 75, 99).

TREANGEN, Todd J. et Eduardo E. P. C. ROCHA (jan. 2011). “Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes”. In : *PLOS Genetics* 7.1, pages 1-12. DOI : 10.1371/journal.pgen.1001284. URL : https://doi.org/10.1371/journal.pgen.1001284 (cf. pages 35, 189).

TYLER, A. D. et al. (juil. 2018). “Evaluation of Oxford Nanopore’s MinION Sequencing Device for Microbial Whole Genome Sequencing Applications”. In : *Sci Rep* 8.1, page 10931 (cf. page 100).

UCHIYAMA, I. (jan. 2003). “MBGD : microbial genome database for comparative analysis”. In : *Nucleic Acids Res.* 31.1, pages 58-62 (cf. page 80).

UNDERWOOD, M. A. et al. (jan. 2015). “Bifidobacterium longum subspecies infantis : champion colonizer of the infant gut”. In : *Pediatr. Res.* 77.1-2, pages 229-235 (cf. page 169).

VALLENET, David et al. (oct. 2019). “MicroScope : an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis”. In : *Nucleic Acids Research*. gkz926. ISSN : 0305-1048. DOI : 10.1093/nar/gkz926. eprint : http://oup.prod.sis.lan/nar/advance-article-pdf/doi/10.1093/nar/gkz926/30294889/gkz926.pdf. URL : https://doi.org/10.1093/nar/gkz926 (cf. pages 16, 182, 196, 198, 201).

VAN HAM, Roeland C. H. J. et al. (2003). “Reductive genome evolution in Buchnera aphidicola”. In : *Proceedings of the National Academy of Sciences* 100.2, pages 581-586. ISSN : 0027-8424. DOI : 10.1073/pnas.0235981100. eprint : https://www.pnas.org/content/100/2/581.full.pdf. URL : https://www.pnas.org/content/100/2/581 (cf. pages 41, 43).

VAN VALEN, L (1973). “A new evolutionary law”. In : *Evol Theory* 1, pages 1-30. URL : https://ci.nii.ac.jp/naid/10019629455/en/ (cf. page 101).

VERNIKOS, G. S. et J. PARKHILL (sept. 2006). “Interpolated variable order motifs for identification of horizontally acquired DNA : revisiting the Salmonella pathogenicity islands”. In : *Bioinformatics* 22.18, pages 2196-2203 (cf. page 197).

VILLANEA, F. (2019). *The n-coalescent*. [Online ; consulté le 28-October-2019 à 06 :43 GMT]. URL : https://hub.wsu.edu/fernandovillanea/the-n-coalescent/ (cf. page 60).

VIRTANEN, Pauli et al. (2020). “SciPy 1.0 : Fundamental Algorithms for Scientific Computing in Python”. In : *Nature Methods*. DOI : https://doi.org/10.1038/s41592-019-0686-2 (cf. pages 137, 139).

VOS, M. et A. EYRE-WALKER (déc. 2017). “Are pangenomes adaptive or not ?” In : *Nat Microbiol* 2.12, page 1576 (cf. pages 76, 78).

WAACK, S. et al. (mar. 2006). “Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models”. In : *BMC Bioinformatics* 7, page 142 (cf. pages 190, 197).

WALKE, Jordan (2019). *React.js*. [Online ; consulté le 01-Décembre-2019 à 14 :26 GMT]. URL : https://reactjs.org/ (cf. page 186).

WALT, Stéfan van der, S. Chris COLBERT et Gaël VAROQUAUX (2011). “The NumPy Array : A Structure for Efficient Numerical Computation”. In : *Computing in Science & Engineering* 13.2, pages 22-30. DOI : 10.1109/MCSE.2011.37. eprint : https://aip.scitation.org/doi/pdf/10.1109/MCSE.2011.37. URL : https://aip.scitation.org/doi/abs/10.1109/MCSE.2011.37 (cf. page 139).

WANDELT, S. et al. (2014). “RCSI : Scalable similarity search in thousand(s) of genomes”. In : tome 3, pages 1534-1545 (cf. page 88).

WARD, D. M. et D. M. WARD (juin 1998). “A natural species concept for prokaryotes”. In : *Curr. Opin. Microbiol.* 1.3, pages 271-277 (cf. page 37).

WATSON, J. D. et F. H. CRICK (avr. 1953). “Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid”. In : *Nature* 171.4356, pages 737-738 (cf. page 23).

WAYNE, L. G. et al. (1987). “Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics”. In : *International Journal of Systematic and Evolutionary Microbiology* 37.4, pages 463-464. ISSN : 1466-5026. DOI : https://doi.org/10.1099/00207713-37-4-463. URL : https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-37-4-463 (cf. page 37).

WEI, W. et al. (mai 2017). “Zisland Explorer : detect genomic islands by combining homogeneity and heterogeneity properties”. In : *Brief. Bioinformatics* 18.3, pages 357-366 (cf. page 197).

WELCH, R. A. et al. (déc. 2002). “Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli”. In : *Proc. Natl. Acad. Sci. U.S.A.* 99.26, pages 17020-17024 (cf. pages 40, 43, 80).

WHEELER, D. L. et al. (jan. 2005). “Database resources of the National Center for Biotechnology Information”. In : *Nucleic Acids Res.* 33.Database issue, pages 39-45 (cf. page 80).

WIKIPEDIA (2019a). *Arrangement*. [Online ; consulté le 9-Juillet-2019 à 16 :49 GMT]. URL : https://fr.wikipedia.org/w/index.php?title=Arrangement&oldid=160769713 (cf. page 45).

— (2019b). *Bacteria display a large diversity of cell morphologies and arrangements*. [Online ; consulté le 1-Juillet-2019 à 10 :19 GMT]. URL : https://en.wikipedia.org/w/index.php?title=Bacterial_cellular_morphologies&oldid=918139503 (cf. page 26).

— (2019c). *Emploi du trait d’union pour les préfixes en français*. [Online ; consulté le 15-mai-2019 à 11 :58 GMT]. URL : https://fr.wikipedia.org/w/index.php?title=Emploi_du_trait_d%5C%27union_pour_les_pr%5C%C3%5C%A9fixes_en_fran%5C%C3%5C%A7ais&oldid=159285221 (cf. page 33).

WILLENBROCK, H. et al. (2007). “Characterization of probiotic Escherichia coli isolates with a novel pan-genome microarray”. In : *Genome Biol.* 8.12, R267 (cf. pages 47, 48).

WILSON, R. C. et al. (2014). “Spherical and Hyperbolic Embeddings of Data”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.11, pages 2255-2269 (cf. page 137).

WOESE, C. R. et G. E. FOX (nov. 1977). “Phylogenetic structure of the prokaryotic domain : the primary kingdoms”. In : *Proc. Natl. Acad. Sci. U.S.A.* 74.11, pages 5088-5090 (cf. page 37).

WOLF, Y. I. et E. V. KOONIN (sept. 2013). “Genome reduction as the dominant mode of evolution”. In : *Bioessays* 35.9, pages 829-837 (cf. page 76).

WU, C. H. et al. (jan. 2004). “PIRSF : family classification system at the Protein Information Resource”. In : *Nucleic Acids Res.* 32.Database issue, pages D112-114 (cf. page 80).

WYLIE, J. L., G. M. HATCH et G. MCCLARTY (déc. 1997). “Host cell phospholipids are trafficked to and then modified by Chlamydia trachomatis”. In : *J. Bacteriol.* 179.23, pages 7233-7242 (cf. page 153).

YAMAMOTO, M. et K. HAYASHI (2015). “Clustering of multivariate binary data with dimension reduction via L1-regularized likelihood maximization”. In : *Pattern Recognition* 48.12. cited By 6, pages 3959-3968. DOI : 10.1016/j.patcog.2015.05.026. URL : https://www.scopus.com/inward/record.uri?eid=2-s2.0-84941415103&doi=10.1016%2fj.patcog.2015.05.026&partnerID=40&md5=8b66575b80daaa30d61dad33792c44c8 (cf. page 126).

ZHAO, Y. et al. (2012). “PGAP : pan-genomes analysis pipeline”. In : *Bioinformatics* 28.3, pages 416-418 (cf. page 83).

ZHAO, Y. et al. (mai 2014). “PanGP : a tool for quickly analyzing bacterial pan-genome profile”. In : *Bioinformatics* 30.9, pages 1297-1299 (cf. page 137).

ZHAO, Y. et al. (jan. 2018). “PGAP-X : extension on pan-genome analysis pipeline”. In : *BMC Genomics* 19.Suppl 1, page 36 (cf. page 84).

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Conceptualisation et exploitation d'un graphe de pangénome partitionné comme représentation compacte de la diversité du répertoire génique des espèces procaryotes

Mots clés : pangénome, graphe de pangénome, microbiologie, métagénome, génomique

Résumé : Introduites en microbiologie en 2005, les approches pangénomiques visent à compiler l'ensemble de la diversité génomique d'une espèce. Dans ces études, on distingue généralement à l'intérieur du pangénome, le génome cœur, c'est-à-dire l'ensemble des familles de gènes où les représentants géniques sont présents dans tous les organismes, et d'autre part, le génome accessoire qui correspond aux gènes spécifiques à certains organismes seulement. Cependant, on constate que le concept de génome cœur est limitant avec un grand nombre d'organismes car des gènes bien que fonctionnellement indispensables peuvent être absents de certains génomes. Pour limiter ce phénomène la quasi-totalité des études utilisent un seuil arbitraire de présence pour définir un génome cœur assoupli. De plus, cette dichotomie entre le génome cœur et accessoire ne rend pas compte des nombreuses gammes de fréquence d'apparition des gènes dans un pangénome. Ce travail de thèse a pour objectif de proposer une approche statistique basée sur un modèle mixé multivarié de Bernoulli couplé à un champ de Markov caché partitionnant le pangénome afin d'être résilient aux absences de gènes et de mieux distinguer leurs différents schémas de présence/absence. En parallèle, plusieurs structures de données basées sur des graphes de pangénomes ont été développées ces dernières années. En effet, exploiter la totalité des informations disponibles dans un génome et non plus seulement la présence de gènes isolés est désormais crucial pour correctement rendre compte de l'organisation des génomes et notamment des régions de plasticité génomique dans les espèces. Cette approche se veut le chaînon manquant entre ces nouvelles approches graphiques à l'échelle de la séquence et les approches originelles en familles de gènes isolés. Pour y parvenir, ce travail de thèse s'intéresse donc à la définition, au partitionnement statistique et à l'exploitation d'un graphe d'un pangénome comme représentation compacte du répertoire génomique des espèces procaryotes. À travers une analyse effectuée sur la base de 136 287 génomes provenant d'isolats (439 espèces) et 34 210 génomes reconstruits à partir de métagénomes (78 espèces), ce travail illustre la manière dont le graphe de pangénome partitionné peut être utile pour estimer la variabilité pangénomique des espèces procaryotes ainsi que de leurs régions de plasticité génomique.

Title: Conceptualization and exploitation of a partitioned pangenome graph as a compact representation of the diversity of the genic repertoire of prokaryote species

Keywords: pan-genome, pangenome, pan-genome graph, microbiology, metagenome, genomic

Abstract: Introduced in microbiology in 2005, pangenome approaches aim to compile the entire genomic diversity of a species. In these studies, we generally distinguish within the pangenome, the core genome, i.e. the set of gene families where gene representatives are present in all organisms; and on the other hand, the accessory genome which corresponds to genes specific to certain organisms only. However, we noticed that the concept of core genome is limiting with many organisms because genes, although functionally essentials, may be absent from some genomes. To deal with this issue, almost all studies use an arbitrary threshold of presence to define a soft core genome. Moreover, this dichotomy between the core and accessory genome does not account for the many ranges of frequencies at which genes appear in a pangenome. The main goal of this thesis work is to introduce a statistical approach based on a multivariate Bernoulli mixture model coupled with a hidden Markov field to partition the pangenome in order to be resilient to gene absences and to better distinguish the gene presence/absence patterns. In parallel, several data structures based on pangenome graphs have been developed in recent years. Indeed, exploiting all the information available in genomes and not just the presence of isolated genes is crucial to highlight genomic organization and particularly the regions of genomic plasticity in species. This approach is intended to be the missing link between these new graphic approaches at the sequence scale and the original approaches in isolated gene families. To achieve this, this thesis work therefore focuses on the definition, statistical partitioning and exploitation of a graph of a pangenome as a compact representation of the genomic repertoire of prokaryotic species. Through an analysis based on 136 287 genomes from isolates (439 species) and 34 210 genomes assembled from metagenomes (78 species), this work illustrates how the partitioned pangenome graph can be useful for estimating the pangenomic variability of prokaryotic species and their regions of genomic plasticity.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France