université PARIS-SACLAY

Méthodes d'analyse comparée des pangénomes procaryotes : explorer la diversité génomique inter-espèces pour une meilleure compréhension du métabolisme

Methods for comparative analysis of prokaryotic pangenomes : exploring interspecies genomic diversity for a better understanding of metabolism

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
Graduate School : Life Sciences and Health. Référent : Université d'Évry Val d'Essonne

Thèse préparée dans l'unité de recherche **Génomique métabolique** (Université Paris-Saclay, Univ Evry, CNRS, CEA) sous la direction de **David VALLENET**, Directeur de recherche, la co-direction d'**Alexandra CALTEAU**, Chercheuse

Thèse soutenue à Paris-Saclay, le 19 mai 2025, par

Jérôme ARNOUX

Composition du jury

Membres du jury avec voix délibérative

Stéphanie BURY-MONÉ Professeure, I2BC, Université Paris Saclay	Présidente
Lucie BITTNER Maîtresse de conférence, HDR, ISYEB, Sorbonne Université	Rapportrice
François SABOT Directeur de recherche, IRD, Université de Montpellier	Rapporteur
Jean CURY Chargé de recherche, CNRS, Institut Pasteur, Sorbonne université, PSL	Examinateur

THESE DE DOCTORAT

NNT : 2025UPASL020

À ma femme Helen et notre petite fille à naitre.

À ma grand-mère Bernadette, qui a toujours été un modèle et une femme que j'admire.

À l'espoir que les hommes comprennent que détruire la nature, c'est se détruire nous-même : "Actuellement, l'homme mène une guerre contre la nature. S'il gagne, il est perdu." - Hubert Reeves

Aux générations futures : "Nous n'héritons pas de la terre de nos ancêtres, nous l'empruntons à nos enfants." - Antoine de Saint-Exupéry

Table des matières

Remerciements **XIII**

Introduction **1**

I Procaryotes : de la biologie cellulaire à la génomique moderne **5**

1 Caractérisation et classification des procaryotes : de la cellule au génome **9**

1.1 La classification des microorganismes : des critères phénotypiques à la biologie moléculaire 9

1.2 Taxonomie des procaryotes : un problème non résolu ? 10

1.3 Espèce procaryote : génomique et phylogénie peuvent-elles trancher ? 12

1.4 Systématique : l'homologie et ses déclinaisons 15

2 Génomique des procaryotes : organisation, évolution et fonctions **17**

2.1 Structure et organisation des génomes procaryotes 17

2.1.1 Composants du génome : séquences codantes et non codantes 17

2.1.2 Réplicons et mécanismes de réplication dans les génomes procaryotes 18

2.2 Dynamique évolutive des génomes 21

2.2.1 Mécanismes d'évolution par transfert vertical 21

2.2.2 Mécanismes d'évolution par transfert horizontal 24

2.3 Du génome aux processus cellulaires 29

2.3.1 Gènes : Régulations et fonctions 29

2.3.2 Îlots génomiques et points chauds d'insertion 31

3 Génomique comparée des procaryotes **33**
3.1 Analyse comparative des génomes : méthodes et applications . . . 33
3.1.1 Alignement des séquences 34
3.1.2 Utilisation des graphes en génomique comparée 35
3.1.3 Modèle statistique pour l'alignement des séquences 38
3.2 Analyse des Systèmes biologiques 43
3.2.1 Définition et intérêt 43
3.2.2 Méthodes de détection 44
3.2.3 Les systèmes de défense aux phages 46
3.3 Génomique à l'ère du Big Data 52
3.3.1 Base de données génomiques 52
3.3.2 Bases de données orientées graphe et données biologiques 55
3.3.3 L'intelligence artificielle au service de la génomique comparée 57

4 Pangénomique : état des lieux, enjeux et ambitions **65**
4.1 Origine et concept 65
4.1.1 Modélisation de la croissance des pangénomes 67
4.1.2 Les différents types de pangénomes 69
4.2 Pangénome de séquences 70
4.2.1 Pangénome basé sur une séquence représentative 70
4.2.2 Pangénome graphe 71
4.3 Pangenome de gènes 76
4.3.1 Généralités et concepts 76
4.3.2 Méthodes et outils de pangénome de gènes 79
4.3.3 Analyses à partir de pangénomes de gènes 80
4.4 Conclusion sur les pangénomes et éléments de réflexions 82

II Du génome au pangénome **85**

1 La suite logicielle PPanGGOLiN : construction et analyse d'un pangénome **89**

1.1 La méthode PPanGGOLiN 90

1.1.1 Construction du graphe de pangénome 92

1.1.2 Partitionnement du graphe 92

1.2 La méthode PanRGP 94

1.2.1 Identification des régions de plasticité génomique 94

1.2.2 Prédiction des spots d'insertion 94

1.3 La méthode PanModule 95

2 Évolution de PPanGGOLiN : présentation de la version 2 **97**

2.1 Nouvelles fonctionnalités et amélioration méthodologique 97

2.1.1 Developpement de nouvelles méthodes d'analyse 97

2.1.2 Amélioration des procédures d'analyses 101

2.2 Optimisation technique 104

2.2.1 Amélioration de l'efficacité de PPanGGOLiN 104

2.2.2 Optimisation du code : lisibilité, maintenance, tests et processus de mise à jour 106

3 Application à l'étude de la dégradation du D-Apiose **111**

3.1 Recherche du contexte génomique chez les procaryotes 111

3.2 Analyse du pangénome de *Escherichia coli* 116

3.3 Identification et annotation de la voie de dégradation dans une nouvelle souche : BVN-ST131 117

4 Conclusion et perspectives **121**

4.1 PPanGGOLiN : bilan de la version 2.0 121

4.2 Évolution de PPanGGOLiN : vers une version 3.0 ? 122

III De la génomique comparée à la pangénomique comparée 125

1 PANORAMA : un nouvel outil pour la pangénomique comparée 129

1.1 Prédiction des systèmes biologiques dans les pangénomes 129

1.2 Comparaison des pangénomes 129

2 Article : PANORAMA 131

3 Conclusion et perspectives 155

3.1 Prédiction de systèmes biologiques 155

3.2 Comparaison de pangénomes 156

IV Base de données de graphe de pangénomes 157

1 Intégration de pangénomes dans une base de données orientée graphe 159

2 Article : Integrating Complex Pangenome Graphs 163

3 Conclusion et perspectives 169

V Conclusion et perspectives 171

1 Conclusions sur le travail de thèse 173

2 Perspectives sur les méthodes développées 177

2.1 Critique et amélioration possible des méthodes 177

2.1.1 PPanGGOLiN et recherche de contexte génomique 177

2.1.2 PANORAMA et prédiction des systèmes 177

2.2 Perspectives et projets autour de la pangénomique 178

3 Perspectives sur la pangénomique et la génomique comparée 181

Bibliographie **182**

Annexes **207**

Table des figures

I.1.1 Schéma cellules eucaryotes et procaryotes 10
I.1.2 Morphologie et arrangement cellulaire procaryote . . . 11
I.1.3 Variation du score d'ANI au niveau de l'espèce 13
I.1.4 Représentation schématique des différentes approches utilisées en métagénomique 15
I.1.5 Schéma représentatif des différents types d'homologie . 16

I.2.1 Représentation des gènes et de leurs produits 18
I.2.2 Évolution d'un plasmide en chromid 19
I.2.3 Distribution de la taille des génomes chez les procaryotes 20
I.2.4 Schéma de la dynamique évolutive des génomes procaryotes 21
I.2.5 Identification des SNP et indels entre 2 génomes 22
I.2.6 Réarrangement et implication 24
I.2.7 Schéma du fonctionnement de la conjugaison 26
I.2.8 Schéma du mécanisme de transformation 27
I.2.9 Schéma synthétique de la transduction 28
I.2.10 Exemple de l'opéron lactose 30
I.2.11 Îlots génomiques et leur caractéristique 31
I.2.12 Cycle de vie d'un îlot génomique 32

I.3.1 Exemple de graphe 36
I.3.2 Principe de l'alignement de MMSeqs2. 40
I.3.3 Algorithmes de clustering de MMSeqs2 41
I.3.4 Exemple de modélisation HMM d'une séquence 42
I.3.5 Exemple de modélisation de systèmes dans MacSyFinder 45
I.3.6 Diversité morphologique parmi les phages 47
I.3.7 Cycle de vie des phages 47

I.3.8 Diversité des systèmes de défenses aux phages 50
I.3.9 Nombre de génomes cumulés par an dans GenBank . . 53
I.3.10 Comparaison de l'homogénéité des rangs taxonomiques entre le NCBI et GTDB 54
I.3.11 Comparaison des modèles entre une base de données relationnelle et une base de données orientées graphe . 56
I.3.12 Schéma général d'une application de machine learning . 58
I.3.13 Exemple d'application de méthode non supervisé 59
I.3.14 Représentation d'un réseau de neurone à 3 couches . . 60

I.4.1 Bibliométrie pangénome 66
I.4.2 Schéma de croissance du pangénome 67
I.4.3 Évolution du pangénome : visualisation de la croissance et de la raréfaction du contenu génique selon la loi de Heap 69
I.4.4 Différents types de pangénomes 70
I.4.5 Exemple d'un graphe de De Bruijn 73
I.4.6 Partitionnement des pangénomes. 77
I.4.7 Représentation d'un pangénome de gènes sous forme de graphe 78

II.1.1 Aperçu général de la méthode PPanGGOLiN 91
II.1.2 Graphe de pangénome de *Acinetobacter baumannii* . . . 93
II.1.3 PanRGP : vue d'ensemble de la méthode de détection des RGP 95

II.2.1 Méthode de recherche du contexte génomique dans un graphe de pangénome 98
II.2.2 Principe de fonctionnement de la méthode de projection 100
II.2.3 Clustering des RGPs 102
II.2.4 Évaluation des performances de calcul des scores de connexions 104
II.2.5 Taille des fichiers de pangénome entre la version 1 et 2 de PPanGGOLiN 106

II.3.1 Voie de dégradation du D-Apiose par la transcétolase non oxydante 112
II.3.2 Arbre taxonomique des procaryotes étiqueté par la présence de la voie de dégradation du D-Apiose 114
II.3.3 Prédiction du contexte de dégradation du D-apiose dans les pangénomes des Enterobacteriaceae 115
II.3.4 Graphe de pangénome de *E. coli* 116
II.3.5 Visualisation du spot 181 dans les génomes de *E. coli* . . 117
II.3.6 Projection du pangénome de *E. coli* sur le génome de la souche BVN-ST131 119

Liste des tableaux

I.3.1 Outils de clustering des séquences 39
I.3.2 Méthodes de Machine learning 61
I.3.3 Méthode de deep learning 62

I.4.1 Outils de pangénomique basés sur les séquences 75
I.4.2 Outils de pangénomique basés sur les familles de gènes 82

II.3.1 Composition du pangénome de *E. coli* 116

IV.1.1 Description des données pangénomiques intégrées dans la base de données graphe. 160

Remerciements

Je commencerai par remercier toutes les personnes qui ont contribué de près ou de loin à la conclusion de ce chapitre de ma vie. La thèse de doctorat est une expérience unique, que j'ai pu vivre en y prenant beaucoup de plaisir. Je retiendrai de ces années que le monde de la recherche est un monde vivant et dynamique, où chaque personne contribue à la réussite et au progrès collectif. À toutes les personnes qui ne verraient pas leurs noms cités, veuillez m'excuser par avance et sachez que je garde en mémoire toutes les personnes qui ont pu me soutenir ces dernières années.

Je me dois d'abord (mais je le fais bien volontiers) de remercier David Vallenet, directeur du LABGeM et directeur de ma thèse. Merci de m'avoir fait confiance pour mener à bien ces projets de recherches. J'ai beaucoup appris à tes côtés, sur le plan technique, mais aussi sur les rouages du monde académique. Je me souviendrai particulièrement de ma première conférence internationale à FEMS où tu m'as accompagné.

Je remercie aussi Alexandra Calteau, chercheuse au LABGeM et co-directrice de ma thèse. Au-delà de toutes les connaissances biologiques et bioinformatiques que tu m'as partagées, je garderai l'image d'une personne extrêmement humaine qui a su me présenter des opportunités pour me former à devenir un chercheur dans tous ces aspects. Tu as su être plus que compréhensive sur des événements personnels qui ont certainement impacté ma thèse et pour ça je te remercie énormément.

Merci à Jean Mainguy, tu as été d'une aide plus que bienvenue pendant tous mes travaux de thèse. Je pense sincèrement que tu as abattu le travail de plusieurs personnes à toi tout seul et sans ça je ne sais pas si les outils de pangénomique du LABGeM seraient dans leur état actuel.

À Adelme Bazin et Guillaume Gautreau, je vous remercie de ne pas avoir quitté le bateau après votre thèse et d'être restés disponibles pour mes questions et pour tout le travail que nous avons pu accomplir sur PPanGGOLiN. Je vous tiens en haute estime, vous m'impressionnez toujours par vos connaissances, vos compétences et aussi par votre sympathie.

Pour Laura Bry et Quentin Fernandez De Grado, j'espère avoir su être un bon encadrant et que vous gardiez un bon souvenir de votre passage au LABGeM. Sachez que j'ai aussi beaucoup appris à vos côtés. Je vous souhaite toute la réussite possible pour votre avenir et j'espère vous croiser à l'occasion.

À Eddy Élisée, merci pour ta joie de vivre et pour avoir été toujours été le premier arrivé au laboratoire pour prendre le premier café. Tu es une personne rayonnante avec qui on aimerait travailler plus souvent ou même juste faire tourner quelques molécules.

Alexandre Protat, tu restes un membre honoraire du LABGeM. Merci d'avoir organisé tous ces GenoPub, et merci pour toutes ces discussions politiques où même si nos opinions étaient parfois opposées, nous avons pu parler sans amertume.

Merci à tous les autres membres du LABGeM. David Roche pour ta sympathie, ton calme et ta bienveillance. Stéphanie Fouteau, pour ton sourire et ta gentillesse. Zoé Rouy, pour tes conseils et ton savoir. Aurélie Génin-Lajus, pour ton énergie à revendre.

Marc Stam, pour nos discussions autour du café. À tous les membres passés et présents, encore merci.

Je remercie également les membres de mon comité de suivi : Hélène Chiapello, Sophie Abby et Vincent Lacroix. Vous avez été de précieux conseils et vous avez toujours jugé mon travail avec honnêteté et bienveillance. J'espère que le résultat final sera à la hauteur des promesses et que nous pourrons nous revoir à l'occasion.

Je remercie tous mes amis qui m'ont supporté pendant ces 3 dernières années. Merci à Maud Repellin pour avoir été présente lorsque j'en avais besoin. Merci à Chloé Beaumont, Florian Jeanneret et Alba Caparros-Roissard pour nos moments de partage d'expérience de doctorant.

Merci à tous les membres de ma famille qui ont su me soutenir, chacun à leur manière. Merci à mon frère et ma sœur qui, même si mes travaux de recherches ne les passionnaient pas, ont bien essayé de me supporter. Merci à mes parents, d'avoir cru en moi et de m'avoir permis de saisir les opportunités qui se présentaient à moi. Merci à mes grands-parents pour leur soutien moral dans toutes les difficultés.

Pour terminer, je ne peux que remercier ma femme Helen qui a toujours été présente pour me soutenir dans toutes les difficultés que j'ai pu rencontrer pendant ma thèse. Même si ce n'était pas toujours évident de comprendre pourquoi j'ai choisi ce métier et cette voie, tu as toujours pensé d'abord à moi et à mon bonheur. Au moment où j'écris ces lignes, tu me prépares le plus beau des cadeaux et je n'ai pas les mots pour t'exprimer ma reconnaissance.

Merci bien sûr à vous, lecteur de ce manuscrit de thèse, j'espère que celui-ci sera à la hauteur de vos attentes.

Introduction

Cette introduction a pour objectif de faire le panorama scientifique et historique des différents sujets qui seront abordés dans ce manuscrit de thèse. Elle fait aussi office d'entrée en matière pour les personnes non expertes qui liront ce manuscrit. Pour ces quelques lignes, je me permettrai donc quelques facilités et imprécisions scientifiques.

Les conditions qui ont permis à la vie de naître sur Terre suscitent encore de nombreuses questions et sont à l'origine de débat scientifique passionnant. Néanmoins, les premières traces de vie retrouvées remontent à 4 milliards d'années et correspondent à des microorganismes, des êtres invisibles à l'œil nu. Ces microorganismes colonisent la Terre depuis des milliards d'années et représentent aujourd'hui la proportion d'êtres vivants la plus importante en termes de nombre et de diversité. Ils jouent un rôle crucial dans les écosystèmes, les cycles biogéochimiques et la santé de la planète comme de la nôtre. En effet, ces microbes sont connus pour poser des problèmes de santé publique (épidémie, hygiène…), de contamination des plantes et des sols, ou encore de dégradation des matériaux. En contrepartie, ils peuvent aussi améliorer notre santé (les probiotiques par exemple), fertiliser les sols et épurer les eaux et être utile dans l'industrie et les biotechnologies (fermentation des fromages et des bières). Pourtant, la microbiologie, l'étude des microorganismes, reste une science relativement récente. Même s'il existe bien, dans l'Antiquité, certains savants et philosophes qui avaient déjà imaginé ces "animaux invisibles", marquant une compréhension primitive de la transmission des maladies infectieuses, il faudra attendre l'invention du microscope par Leeuwenhoek, au XVII^{e} siècle, pour qu'il fasse les premières observations d'*animaculum*, marquant la naissance de la microbiologie. La microbiologie du XVII^{e} au XX^{e} siècle a amené de grandes découvertes et révolution scientifique, notamment en médecine. Nous pouvons citer les travaux de Louis Pasteur qui a prouvé, en 1877, que les maladies infectieuses étaient causées par des microorganismes (staphylocoque, pneumocoque et streptocoque), ou encore d'Alexander Fleming qui découvrit, en 1928, la pénicilline, le premier agent antibiotique.

En s'éloignant quelque peu de la microbiologie, toujours entre le XVII^{e} et XX^{e} siècle, les chimistes s'intéressent aux molécules du vivant. Autour des années 1800, Le Français Antoine Fourcroy va faire la première description de substances azotées dans les organismes vivants, qu'il appelait "substances animales". C'est ensuite, en 1835, que le chimiste néerlandais Gérardus Johannes Mulder découvre des chaînes de substance azotées, qui seront introduites sous le terme de protéine par le chimiste suédois Jöns Jacob Berzelius en 1838. Le mot vient du grec *proteios*, qui signifie "de première importance", soulignant l'intérêt fondamentale de ces molécules composées de carbone, hydrogène, azote et oxygène, avec des proportions spécifiques, dans les organismes vivants. Enfin, en 1894, le chimiste allemand, Emil Fischer, démontra que les protéines sont composées d'acides aminés, unité de base des protéines, liés par des liaisons peptidiques. Il déterminera la composition et la structure de plusieurs d'entre eux. La fin du XIX^{e} siècle voit aussi la découverte d'une autre molécule du vivant, l'acide désoxyribonucléique, mieux connue sous l'acronyme ADN. C'est le biologiste suisse Friedrich Miescher qui découvre, en 1869, une substance riche en phosphore dans les cellules du pus, qu'il appelle "nucléine". Il faudra attendre près d'un demi-siècle (1929) pour que Phoebus Levene, biochimiste russe-américain, identifie les composants de base de l'ADN : les nucléotides. Plus tard, en pleine Seconde Guerre mondiale, les chercheurs

Oswald Avery, Colin MacLeod et Maclyn McCarty, confirment l'hypothèse de Miescher, en montrent que l'ADN est la substance qui transfère les caractères héréditaires et que l'ADN est le support de l'information génétique. Pour terminer, les travaux de James Watson, Francis Crick et Rosalind Franklin, ont permis de décrire la structure de la molécule d'ADN. Toutes ces découvertes ont ouvert la voie à de nombreuses autres dans tous les domaines : médecine, agroalimentaire, biotechnologie, et sont le socle de la génétique moderne.

Les développements technologiques de la seconde moitié du XX^{e} siècle, et notamment l'apparition du séquençage et de l'informatique, amènent les chercheurs à créer une nouvelle discipline pour l'étude de la structure et de la composition des molécules du vivant : la bioinformatique. En 1955, Frederick Sanger séquencera la première protéine, l'insuline. Cette découverte, récompensée par un prix Nobel, a établi la base du séquençage. Peu de temps après, Margaret Dayhoff, une pionnière de la bioinformatique, développe l'un des premiers programmes informatiques pour analyser les séquences de protéines. Elle publiera d'ailleurs, en 1969, le premier atlas de séquences protéiques, jetant les bases de l'analyse des séquences biologiques. À partir de là, la bioinformatique ne cessera d'évoluer avec les techniques de séquençage. En 1970, Saul Needleman et Christian Wunsch introduisent un algorithme pour l'alignement global des séquences, qui est toujours utilisé aujourd'hui. En 1977, Sanger va à nouveau révolutionner le domaine de la biochimie en proposant une méthode de séquençage de l'ADN qui portera son nom. Elle devient rapidement la méthode de référence en raison de sa précision. Dans les années 80, la méthode s'automatise, devient plus rapide et précise. On voit alors se développer les premières bases de données accessibles au public pour stocker des séquences génétiques et protéiques. En 1990, est lancé le Projet du Génome Humain (HGP), un effort international visant à séquencer l'intégralité du génome humain. Ce projet catalyse de nombreux développements en bioinformatique, notamment dans la gestion et l'analyse des grandes quantités de données générées. Enfin, au début des années 2000, les technologies de séquençage sont de plus en plus performantes et abordables, faisant entrer la bioinformatique dans l'âge du *Big Data*, la rendant essentielle dans de nombreux domaines d'étude en biologie.

La microbiologie et, pour ce qui va nous intéresser ici, l'étude de la génétique des microorganismes profitent de toutes ces nouvelles technologies pour développer ces connaissances. Elle va aussi subir cette explosion de la quantité d'informations disponibles dans les bases de données. C'est pourquoi, microbiologistes et bioinformaticiens sont toujours à la recherche de nouvelles méthodes pour l'analyse de ces données. Alors que les programmes bioinformatiques s'attachaient à représenter et à étudier un génome en tant qu'une séquence indépendante des autres, un nouveau concept de représentation des génomes est apparu : le pangénome. Il permet de regrouper l'ensemble des génomes en une seule entité et donc de rendre une représentation globale de l'ensemble de l'information contenue dans les génomes. Le pangénome garantit une meilleure représentation de la diversité des génomes, tout en étant plus adapté à l'analyse de grandes quantités de données.

C'est dans ce cadre que s'inscrit mon travail de thèse, articulé autour de trois objectifs principaux. Le premier visait à développer de nouvelles approches pour l'analyse des pangénomes, en particulier pour la recherche de contexte génomique et la prédiction de systèmes biologiques. Le deuxième objectif consistait à renforcer les outils de pangénomique sur le plan technique, afin de faire face à l'augmentation exponentielle du nombre de génomes disponibles. Cela impliquait l'amélioration de PPanGGOLiN, un outil dédié à la construction et à l'analyse de graphes de pangénomes procaryotes. Enfin, le troisième objectif visait à ouvrir la voie à la comparaison directe entre pangénomes, un domaine encore peu exploré, mais essentiel pour mieux comprendre la diversité inter-espèces. Pour cela, j'ai mis au point de nouvelles méthodes, que j'ai intégrées dans un outil : PANORAMA, conçu pour analyser et comparer plusieurs pangénomes. L'ensemble de ces travaux s'inscrit dans une démarche conforme aux principes FAIR et à l'open science, avec pour ambition de rendre ces outils accessibles à l'ensemble de la communauté des microbiologistes et bioinformaticiens.

Ce manuscrit sera divisé en plusieurs parties comme suit. Une première partie sera consacrée à contextualiser et à rendre compte des problématiques auxquelles répond ce travail de thèse. Dans cette partie, je donnerai la définition précise des termes que j'utiliserai et je reviendrai sur l'état de l'art en génomique comparée des procaryotes. Je poursuivrai par une seconde partie sur les développements méthodologiques que j'ai pu réaliser en pangénomique, notamment dans la suite logicielle PPanGGOLiN. La troisième partie sera consacrée au cœur de mon sujet de thèse, *i.e.*, aux développements de méthodes pour la comparaison de pangénomes. Enfin, je présenterai une nouvelle approche utilisant les bases de données orientées graphe comme solution pour le stockage et l'étude des pangénomes. Pour terminer, je présenterai une discussion critique sur le travail réalisé pendant ces trois ans et demi. Je me dois également de rappeler aux lecteurs que la nature, *mirabile dictu*, se distingue par une diversité extraordinaire, et que certaines règles ou affirmations généralement vérifiées peuvent souffrir d'exceptions.

Je vous souhaite bonne lecture de ce manuscrit, qui, je l'espère, fait preuve de toute la rigueur scientifique attendue et rend compte du travail réalisé pendant ces trois ans et demi de manière authentique.

CHAPITRE I PROCARYOTES : DE LA BIOLOGIE CELLULAIRE À LA GÉNOMIQUE MODERNE

Ce chapitre marque le début du manuscrit et posera les bases conceptuelles et méthodologiques, biologiques et bioinformatiques essentielles à la compréhension des travaux menés dans le cadre de cette thèse. Il s'agit ici de contextualiser les enjeux de la génomique des procaryotes et de la pangénomique, tout en abordant les principaux concepts et méthodes utilisés dans ce domaine.

Nous commencerons par un rapide retour sur ce qu'est un procaryote, un élément clé pour définir les bornes et le contexte d'application de nos recherches. Cette partie est essentielle pour comprendre les spécificités de ces organismes et leurs impacts sur les approches méthodologiques adoptées. Cette introduction permettra également de les situer dans la classification du vivant, notamment en revenant sur la structure cellulaire et l'organisation du génome, tout en apportant les éléments pour discuter de la notion d'espèce procaryote.

Une fois ce cadre biologique posé, nous aborderons les bases de la génomique comparée, en se focalisant sur l'application aux procaryotes. Ce moment sera l'occasion de clarifier l'utilisation de simplifications ou de choix algorithmiques, souvent nécessaires en raison des caractéristiques propres à ces génomes. Ces éléments permettront de mieux comprendre l'approche bioinformatique qui sous-tend la comparaison des génomes, et ce, de la comparaison de séquences en allant jusqu'à l'approche par graphe, en passant par les modèles statistiques et les méthodes d'intelligence artificielle.

Le chapitre poursuivra en contextualisant la pangénomique, un domaine en pleine expansion qui permet de saisir la diversité génétique des populations microbiennes et qui est le centre des travaux de recherche ici réalisés. Nous mettrons en lumière l'évolution des données biologiques, tant sur le plan quantitatif que qualitatif, et soulignerons les défis posés par la gestion et l'analyse de ces données, en particulier dans le cadre de leur représentation, pour conclure par la manière dont la pangénomique a pu répondre à ces difficultés.

À la fin de ce chapitre, le lecteur aura tous les éléments théoriques et méthodologiques pour aborder les travaux de recherche développés, tout en disposant du cadre dans lequel s'inscrit la thèse et des enjeux actuels de la génomique des procaryotes et de la pangénomique.

1 - Caractérisation et classification des procaryotes : de la cellule au génome

Caractériser et classer un organisme comprend deux approches complémentaires : la taxonomie, qui consiste à regrouper les individus en catégories appelées taxons, et la systématique, qui vise à reconstruire les relations évolutives entre ces individus. Ces deux approches ont conduit à l'élaboration de classifications du vivant, dont l'une des plus largement adoptées est celle fondée sur trois grands domaines : Bactérie, Archée et Eucaryote [1]. Dans la suite, nous nous appuierons sur cette classification.

1.1 . La classification des microorganismes : des critères phénotypiques à la biologie moléculaire

On regroupe dans le terme **microorganisme** tous les êtres vivants (organismes) de taille microscopique. Cette définition inclut des organismes très diversifiés appartenant aux trois domaines du vivant : les bactéries, les archées et des eucaryotes (comme les protozoaires, certaines algues et champignons microscopiques). Elle englobe également les virus, bien qu'ils ne soient pas considérés comme des organismes vivants à proprement parler. Les premières classifications des microorganismes se sont appuyées sur des critères phénotypiques, *i.e.*, des caractéristiques observables. Bien que ces premières tentatives aient été limitées par la petite taille des organismes et les technologies disponibles pour les observer et les analyser, elles ont permis de distinguer plusieurs grands groupes.

Pour commencer, certains microorganismes sont pluricellulaires, comme les champignons du genre *Penicillium*, qui sont des eucaryotes, tandis que d'autres, tels que la bactérie *Escherichia coli* [2], ne sont constitués que d'une seule cellule et sont qualifiés d'unicellulaires. Dans la suite, nous nous concentrerons exclusivement sur les organismes unicellulaires [3]. La première distinction majeure qui a été établie pour diviser le vivant en deux grands domaines repose sur la présence ou l'absence de noyau (figure I.1.1). Le noyau est une structure interne de la cellule qui va contenir l'ensemble du matériel génétique. Les organismes (unicellulaires ou non) qui ont un noyau sont qualifiés d'eucaryotes. Pour ceux dont le matériel génétique est librement dispersé dans le cytoplasme, ils sont catégorisés comme appartenant aux procaryotes, qui regroupent les domaines Bactérie et Archée [4]. Ce sont ces derniers qui vont nous intéresser.

1. Les virus ne sont pas inclus dans cette classification, leur statut d'êtres vivants restant controversé.
2. Bactérie modèle présente dans l'intestin humain, où elle peut être inoffensive, bénéfique ou pathogène.
3. Certains procaryotes montrent des formes primitives de coopération et de différenciation cellulaire, évoquant une multicellularité, mais leurs cellules restent fonctionnellement indépendantes (Wilpiszeski *et al.*, 2019).
4. Une ancienne classification, proposé par le biologiste français Édouard Chatton, divisait les organismes en 2 empires Eucaryote et Procaryote. L'empire Procaryote est aujourd'hui divisé en 2 domaines Bactérie et Archée.

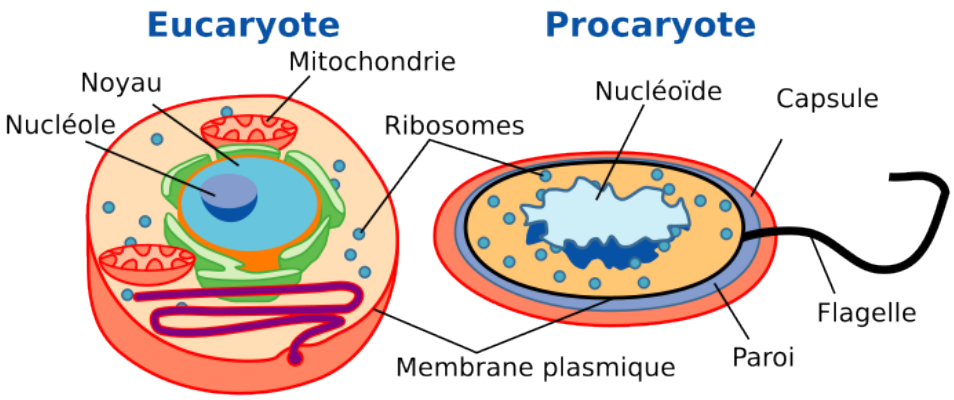


Figure I.1.1 – **Schéma représentant une cellule eucaryote (à gauche) et une cellule procaryote (à droite).** La cellule eucaryote est identifiable par un noyau entouré d'une membrane, ainsi que par la présence de mitochondries, de petits organites responsables de fournir de l'énergie chimique à la cellule. En revanche, dans la cellule procaryote, le matériel génétique (nucléoïde) est librement dispersé dans le cytoplasme, sans être isolé par une membrane. Extrait et adapté de Wikipédia `https://commons.wikimedia.org/wiki/File:Celltypes.svg` Crédit image : Creative Commons.

Le développement de la biologie moléculaire a permis d'affiner et de corriger les classifications précédentes en analysant la physiologie et la biochimie des cellules procaryotes, ainsi que les séquences d'ADN des génomes. C'est notamment en étudiant les gènes codant l'ARN 16S que Carl Woese mit en évidence en 1977 que l'ensemble des procaryotes ne formait pas un groupe monophylétique, mais qu'ils étaient séparés en deux domaines, Bactérie et Archée (Woese et Fox, 1977). Longtemps considéré comme des bactéries extrémophiles, il est aujourd'hui clair que les archées représentent un domaine à part entière avec toute sa singularité, comme la composition de leur membrane par exemple (Albers et Meyer, 2011). Malgré toute la fascination que nous pouvons avoir pour les archées, et que toutes les méthodes qui seront présentées peuvent s'appliquer aux espèces Archée, nous ne présenterons que très peu de résultats les concernant. C'est pourquoi dans la suite, même si nous parlerons de procaryote, nous considérerons plutôt les bactéries avec un prolongement possible aux archées.

1.2 . Taxonomie des procaryotes : un problème non résolu ?

La classification et la définition d'espèce procaryote ne fait pas consensus dans la communauté des microbiologistes (Chun et Rainey, 2014; Adl *et al.*, 2019). Toutefois, les méthodes de classification se basent sur le principe de caractères partagés entre les individus (Aldhebiani, 2018). Ces caractères peuvent être soit phénétiques, *i.e.*, reposant sur la similarité d'un trait, sans s'intéresser au lien évolutif qui pourrait les relier, soit phylogénétiques, *i.e.*, reposant sur l'hérédité du caractère indépendamment de son état actuel.

Les premières tentatives de classification des bactéries reposaient sur des approches phénétiques, utilisant des critères basés sur les caractéristiques observables de ces organismes : morphologie, physiologie et biochimie. D'un point de vue morphologique, les microbiologistes examinaient des paramètres tels que la taille des cellules, leur mode de croissance et leur capacité à former des agrégats spécifiques (figure I.1.2). La présence ou l'absence de structures spécialisées, telles que les flagelles, était également un critère de différenciation. Les caractéristiques physiologiques permettaient, quant à elles, de classer les bactéries selon leur mode de vie, leur métabolisme (ana-

bolisme et catabolisme) et leurs réponses aux conditions environnementales. L'étude de la composition cellulaire offrait par ailleurs de nouveaux outils pour affiner ces classifications sur le plan biochimique. Par exemple, la coloration de Gram, méthode emblématique, permet de différencier les bactéries en deux grands groupes : les Gram-positives, caractérisées par une paroi épaisse de peptidoglycane, et les Gram-négatives, qui présentent une paroi plus fine associée à une membrane externe lipidique. Selon le contexte d'étude, d'autres critères peuvent être intégrés. Dans le domaine médical, la pathogénicité (capacité à induire une maladie) et le sérogroupage (basé sur la composition antigénique de la capsule bactérienne) sont particulièrement utilisés pour identifier et classifier les bactéries d'intérêt clinique.

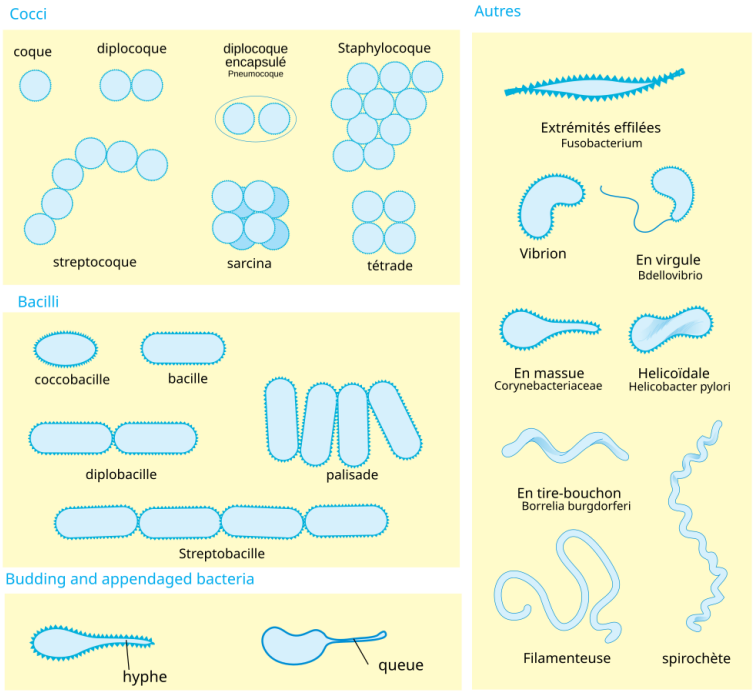


Figure I.1.2 – **Diversité morphologique des cellules procaryotes et leur arrangement.** Par Mariana Ruiz Villarreal `https://commons.wikimedia.org/w/index.php?curid=9908634`

L'ADN et ses propriétés biochimiques ont également été utilisés en tant que critère phénétique. Dans ces propriétés, il y a d'abord le pourcentage de guanine-cytosine (GC) qui permet de différencier 2 souches appartenant à 2 genres différents si leur taux de GC respectif diffère de plus de 10 %, mais il faut noter qu'une composition en GC proche n'implique pas forcément que les souches soient proches (Schleifer, 2009). Une approche visant à définir formellement une espèce procaryote a été adoptée en 1987 par un comité d'experts (Moore *et al.*, 1987). Il propose que des souches appartiennent à une même espèce si leur ADN s'hybride [5] à plus de 70 % et que le ΔT_m [6] diffère de 5 degrés ou moins.

Toutes ces approches permettent de classer les procaryotes en taxon et sont toujours d'actualité [7]. Avec l'arrivée de la génomique, du séquençage de génomes et de la bioinformatique, ces classifications intègrent des critères génomiques.

5. Appariement de 2 brins d'ADN par complémentarité des bases
6. température à laquelle la moitié de l'ADN est dénaturés
7. On les retrouve notamment dans des ouvrages comme le *Bergey's Manual of Systematic Bacteriology* qui référence toutes les espèces connues et leur caractéristique.

1.3 . Espèce procaryote : génomique et phylogénie peuvent-elles trancher ?

Les approches phénétiques présentées précédemment ont l'intérêt de s'appliquer au laboratoire et donc de regrouper et d'identifier les souches directement. Néanmoins, elles restent relativement approximatives et sont parfois coûteuses (en temps et en moyens). De plus, elles ne peuvent s'appliquer qu'aux souches que l'on peut cultiver et analyser en laboratoire, ce qui empêche leur utilisation sur les espèces qui vivent dans des conditions non reproductibles, *i.e.*, que les souches ne peuvent pas être cultivées ou isolées, en laboratoire. Enfin, même si elles répondent aux problèmes de la taxonomie, et donc de ranger les bactéries dans des taxons, elles ne répondent pas à la question du lien entre les différents taxons et comment représenter ce lien, *i.e.*, à la question de la systématique.

Le développement des techniques de séquençage d'ADN, initié par F. Sanger en 1977 et sa méthode éponyme (Sanger *et al.*, 1977), a permis de séquencer les premiers génomes [8]. Ainsi, le premier génome complet procaryote (aussi le premier génome complet d'un organisme cellulaire), celui de la bactérie *Haemophilus influenzae* [9], est séquencé en 1995 (Fleischmann *et al.*, 1995). Dans les années qui suivent, de plus en plus de génomes sont séquencés (*cf.* section 3.3). Cette nouvelle source d'information, amène le développement de méthodes basées sur les gènes et les génomes pour classer les organismes. Une de ces méthodes, qui reste encore largement utilisée en routine aujourd'hui, compare les souches à partir d'un gène marqueur qui reflète à la fois la similarité entre les souches, permettant leur regroupement, et les événements dits de spéciation ayant conduit à leur séparation en espèces distinctes. On va privilégier l'utilisation de gènes hautement exprimés qui assurent une fonction essentielle à la vie de l'organisme : les gènes de ménage (*house-keeping genes*). Le gène codant l'ARNr 16S a été largement utilisé en tant que marqueur depuis les années 1970 pour des analyses phylogénétiques (Balch *et al.*, 1977; Stackebrandt *et al.*, 1983), il a la particularité d'être présent chez tous les procaryotes. Avec l'amélioration des méthodes et des technologies, ainsi que l'augmentation du nombre de génomes disponibles, de plus en plus d'analyses utilisent des arbres à différents niveaux taxonomiques (Ludwig *et al.*, 1990; Chun et Goodfellow, 1995; Lee *et al.*, 2000; Imhoff, 2003). En 2007, un arbre du vivant de toutes les espèces a été reconstruit à partir d'un arbre d'ADNr 16S comprenant toutes les souches types séquencées d'espèces de bactéries et d'archées publiées jusqu'à la fin de l'année 2007 (Yarza *et al.*, 2008). En allant encore plus loin, des analyses *multilocus sequence analysis* MLSA ont été proposées (Stinear *et al.*, 2000; Kuhnert et Korczak, 2006; Pascual *et al.*, 2010; Glaeser et Kämpfer, 2015). Ces analyses prennent en compte plusieurs gènes marqueurs pour réaliser la taxonomie. L'utilisation de plusieurs gènes augmente le niveau d'information et réduit les biais. Toutefois, il n'y a pas de recommandation universelle pour réaliser l'analyse et chaque MLSA est réalisé en fonction des souches de départ. La sélection des gènes et leur nombre sont des paramètres qui ont un impact encore peu évalué sur la taxonomie. Il en va de même pour la taille des fragments considérés pour chaque gène, qui ne représente qu'une partie de la séquence du gène. Enfin, expérimentalement, il est souvent difficile, voire impossible, de concevoir des amorces facilitant l'amplification des gènes dans toutes les souches considérées. Malgré ces critiques, l'utilisation de gènes marqueurs est encore aujourd'hui utilisée, mais est peu à peu remplacée par des méthodes considérant l'ensemble du génome.

8. le premier génome séquencé est celui du virus de bactérie MS2 (Fiers *et al.*, 1976)
9. Bactérie pathogène, responsable de maladie respiratoire ou de méningites et bactériémie.

Au début des années 2000, avec les nombreux projets autour du séquençage et de l'analyse des génomes, comme le projet génome humain (Lander *et al.*, 2001), les technologies de séquençage sont de plus en plus précises et de moins en moins coûteuses, amenant la génomique dans l'ère "moderne" : une augmentation exponentielle du nombre de séquences et des séquences plus longues et de meilleure qualité (Hugenholtz *et al.*, 2021; Hu *et al.*, 2021). Cette avancée est notamment due à l'émergence des technologies de séquençage de nouvelle génération (*Next-Generation Sequencing* en anglais, NGS), qui permettent le séquençage massif en parallèle d'ADN à grande échelle, mais aussi d'obtenir des séquences plus longues. Avec ces nouvelles séquences, il est possible d'obtenir des génomes complets (*Whole Genome Sequencing* -WGS) et de les analyser. Il est alors possible de considérer le génome complet dans les approches d'assignation taxonomique d'organismes. Une de ces approches est l'*Average Nucleotide Identity* (ANI), qui rend compte de la similarité (cf section 3.1) entre 2 séquences nucléotidiques. Le score d'ANI va d'ailleurs remplacer celui de l'hybridation, où un ANI inférieur à 95 % permet de différencier les espèces à la place d'une hybridation à 70 % (Goris *et al.*, 2007). Plus récemment, le seuil de 95 % a été confirmé par les auteurs de FastANI (Jain *et al.*, 2018), utilisant plus de 90 000 génomes. Ils ont montré l'existence d'un *gap*, espace où l'ANI diminue fortement avant 95 % (figure I.1.3).

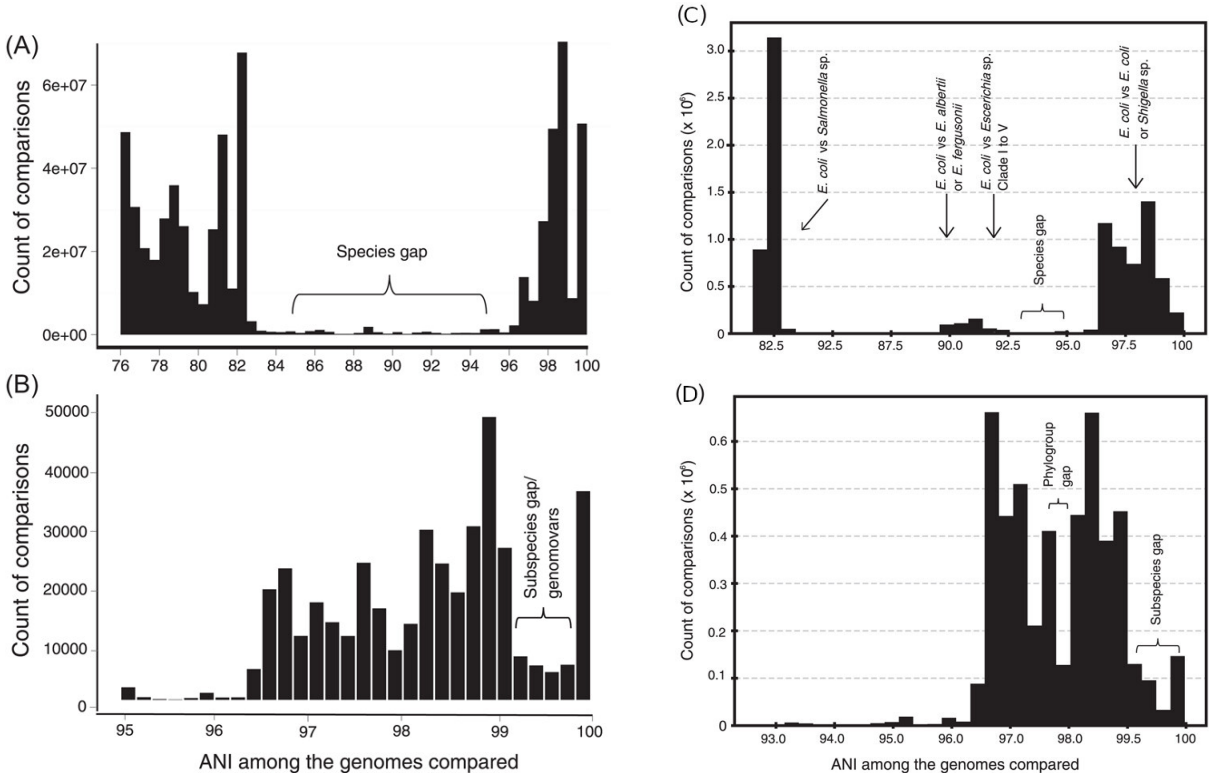


Figure I.1.3 – **Variation du score d'ANI au niveau de l'espèce.** (A-B) Les histogrammes sont basés sur des comparaisons par paire effectuées avec FastANI. (A) Le Score d'ANI représenté au niveau de l'espèce se base sur les données de Jain *et al*. On y retrouve un *gap* entre 84 et 95 % d'ANI. (B) Score d'ANI représenté au niveau intra-espèce sur les données de Rodrigues-R *et al*. On retrouve un *gap* entre 99,2 et 99,8 % d'ANI. (C-D) Score d'ANI au niveau du groupe *Escherichia coli*. Le nombre de génomes utilisés est le suivant : *E. coli* : 2815 ; *Salmonella enterica* : 1351 ; *Escherichia fergusonii* : 57 ; *Escherichia albertii* : 70 ; et *Shigella flexneri* : 93 (tous les génomes complets disponibles au NCBI en juillet 2023). (C) Comparaison de l'ANI entre *E.Coli* et d'autres espèces. Le seuil de 95 % délimitant l'espèce est retrouvé. Un *gap* à 97 % existe entre *E.coli* et *Shigella flexneri* (une espèce d'*E.Coli* particulière pour ces propriétés infectieuse). (D) Analyse de l'ANI au sein des génomes de *E.Coli*. L'écart d'ANI de 99,5 % est aussi prononcé, par rapport aux barres adjacentes, que l'écart d'ANI de 98 %-97 % qui correspond à l'écart entre les phylogroupes d'*E. coli*, un groupe distinct et bien reconnu au sein d'*E. coli*. Figures et légende adaptées de (Konstantinidis, 2023)

En parallèle, de nouvelles techniques de séquençage émergent, pour permettre de séquencer l'ADN d'échantillons environnementaux (figure I.1.4.A). En 1998, Handelsman *et al.* (Handelsman *et al.*, 1998) proposent une méthode pour séquencer l'ADN des bactéries du sol, sans avoir à cultiver les cellules, comme c'est le cas des méthodes précédentes. Pour cela, les bactéries sont isolées et leur ADN extrait, puis des enzymes de restriction permettent de récupérer des régions génomiques d'intérêt, qui seront amplifiées par PCR. Ces régions vont contenir des gènes marqueurs pour l'identification des organismes. Dans leur article, ils désignent ce séquençage direct d'ADN d'un milieu par le terme **métagénomique**.

Avec le développement des NGS, des projets d'étude métagénomique à grande échelle voient le jour, comme le projet Tara Oceans par exemple (Karsenti *et al.*, 2011). Les nouvelles techniques de séquençage permettent d'obtenir l'intégralité de l'ADN de l'environnement. Dans ce cadre, deux approches complémentaires ont émergé pour reconstituer des génomes complets à partir de données métagénomiques (Chang *et al.*, 2024) : les *Metagenome-Assembled Genomes* (MAGs) et les *Single-Amplified Genomes* (SAGs). Les MAGs (figure I.1.4.B) sont reconstruits en assemblant des fragments d'ADN issus de séquençage métagénomique en génomes cohérents, sans nécessiter l'isolement des cellules. Les SAGs, quant à eux, sont obtenus en isolant et en amplifiant le génome d'une seule cellule avant séquençage, permettant une meilleure résolution pour les organismes rares ou difficiles à assembler par approche métagénomique classique.

La métagénomique permet donc d'obtenir une part importante de gènes jusqu'alors inconnus (Bickhart *et al.*, 2022; Rinke *et al.*, 2013). Elle présente un intérêt en médecine, en agriculture, en écologie, mais aussi en phylogénie. Ces méthodes ont considérablement enrichi notre compréhension de la diversité microbienne, en révélant de nombreuses lignées jusqu'alors inconnues et en affinant les classifications existantes (Hug *et al.*, 2016).

Pourtant, la communauté n'est toujours pas arrivée à un consensus sur la classification des procaryotes en espèces et même sur l'existence d'espèces procaryotes. On peut d'abord critiquer l'approche et les résultats des études utilisant l'ANI, qui se limitent aux génomes de bonne qualité et complets, ce qui *de facto* limite le nombre de génomes et d'espèces potentielles pris en compte, tout en augmentant la redondance et limitant la diversité et la variabilité. De plus, la démarche apporte le biais d'utiliser une taxonomie déjà existante. Il faut aussi considérer que la dynamique évolutive des procaryotes, que nous détaillerons dans le chapitre suivant (section 2.2), n'est pas linéaire et héréditaire, mais que les procaryotes sont capables de recevoir et d'échanger de l'ADN. C'est pourquoi des auteurs soutiennent une définition plus écologique de l'espèce procaryote (Luo *et al.*, 2011), intégrant ces échanges qui agissent sur la valeur sélective [10] (*fitness* en anglais), des organismes dans leur environnement. Récemment, l'équipe Phil Hugenholtz a proposé une classification normalisée des espèces procaryotes, basée uniquement sur des critères génomiques (*cf.* section 3.3), disponible dans la base de données GTDB (Parks *et al.*, 2018).

On peut donc convenir qu'il n'est pas encore communément admis de parler d'espèce procaryote. Il existe toutefois des caractéristiques communes et spécifiques aux procaryotes ainsi que des traits propres à chaque taxon. De nombreuses méthodes et démarches scientifiques parviennent à construire une phylogénie des procaryotes, mais celle-ci doit être replacée dans son contexte d'étude pour prendre sens. Notamment dans les travaux de recherche que j'ai réalisés, où il était nécessaire de se baser

10. la capacité d'un individu à survivre et à transmettre son patrimoine génétique

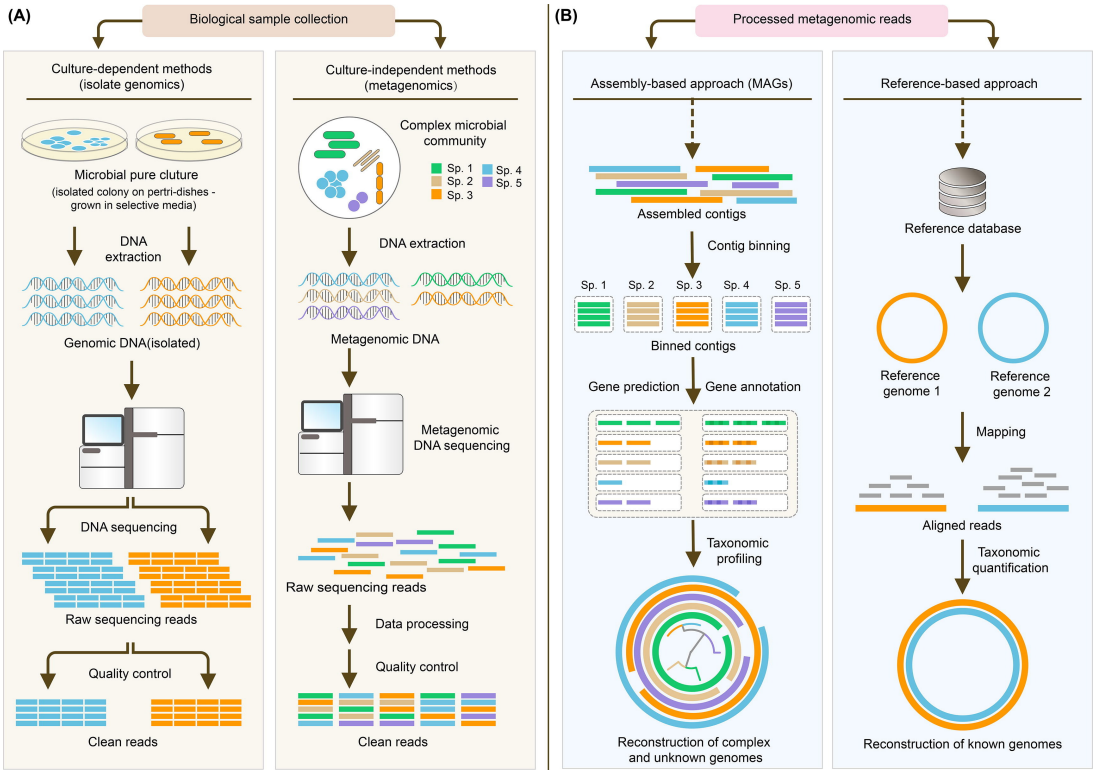


Figure I.1.4 – **Représentation schématique des différentes approches utilisées en métagénomique.** (A) Comparaison des étapes de génération des données de séquençage pour les méthodes culture-dépendante et culture-indépendante (métagénomique). (B) Comparaison entre les approches basées sur l'assemblage et celles basées sur une référence pour l'analyse des données de séquençage métagénomique. Extrait de (Yang *et al.*, 2021)

sur une classification des génomes en espèces. Dans le contexte de nos travaux, la similarité des séquences l'emporte comme critère de classification, nous utiliserons donc des génomes provenant de bases de données utilisant des critères comme l'ANI ou des gènes marqueurs pour construire des pangénomes.

1.4 . Systématique : l'homologie et ses déclinaisons

Pour retracer l'évolution des organismes et représenter leur lien, une première approche intuitive est de rechercher une origine commune entre les gènes, appelée gène ancestral. Si un tel gène existe, on dit que les gènes issus de ce gène ancestral sont homologues. Dans un second temps, il convient d'étudier les évènements évolutifs qui ont conduit à la séparation des gènes pour préciser le type d'homologie (figure I.1.5). Le premier évènement est un évènement dit de spéciation et conduit à l'émergence d'une espèce. Dans ce cas, si les gènes sont uniquement séparés par des spéciations, on dit qu'ils sont orthologues. Le second évènement est une duplication des gènes sur le même génome (cf. sous-sous-section 2.2.1.2). Les gènes sont alors dits paralogues et vont évoluer de façon indépendante dans le génome. Un autre évènement évolutif, fréquent chez les procaryotes, est celui du transfert horizontal, *i.e.*, l'échange de matériel génétique entre organismes (cf. sous-section 2.2.2). Lorsque des gènes sont transférés horizontalement, ils sont dits xenologues.

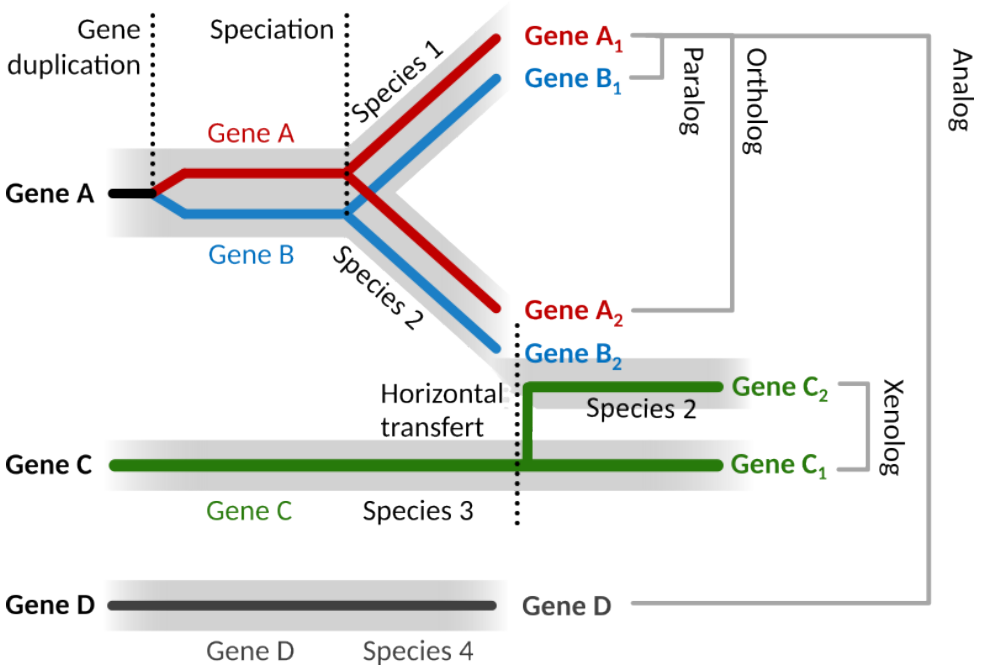


Figure I.1.5 – **Schéma représentatif des différents types d'homologie.** Figure extraite et adaptée de https://en.m.wikipedia.org/wiki/Sequence_homology sous licence Creative Common

Toutes ces notions sont essentielles pour poser des hypothèses de travail qui seront utilisées dans les analyses de génomique comparée (Koonin, 2005; Stamboulian *et al.*, 2020). Une de ces hypothèses que nous utiliserons dans nos analyses est celle que des gènes homologues vont coder pour des fonctions similaires[11]. Plus précisément, on suppose que des gènes orthologues vont coder pour une même fonction et vont donc faire également partie du même processus. Les gènes paralogues peuvent avoir des fonctions différentes, mais qui restent proches, par exemple pour des enzymes, le substrat va changer, mais la réaction rendra un produit chimiquement proche (Mirny et Gelfand, 2002).

11. *N.B* : 2 gènes codant pour la même fonction ne sont pas nécessairement homologues. Il existe des cas de convergence évolutive, *i.e.*, une fonction similaire sans origine commune. Il est aussi possible que des séquences courtes ou peu complexe semble homologue, mais cette apparente homologie serait liée au hasard.

2 - Génomique des procaryotes : organisation, évolution et fonctions

Les génomes procaryotes sont souvent décrits comme plus simples et plus faciles à étudier que les génomes eucaryotes. La simplicité apparente de ces génomes cache en réalité des mécanismes complexes. Dans cette partie, je décrirai les mécanismes les plus connus et les plus répandus.

2.1 . Structure et organisation des génomes procaryotes

2.1.1 . Composants du génome : séquences codantes et non codantes

Le génome se divise en 2 catégories : les séquences codantes, représentant la majorité du génome (entre 85 et 90 %), et les séquences non codantes. Les séquences codantes sont divisées en unités appelées gènes. Les gènes jouent un rôle essentiel puisqu'ils contiennent l'information nécessaire à la production des protéines, impliquées dans toutes les réactions cellulaires. Ils renferment également les séquences pour produire des ARN ribosomiques (ARNr) et des ARN de transfert (ARNt), indispensables à la production des protéines (cf. sous-section 2.3.1).

Dans le génome, les gènes ne sont pas répartis aléatoirement. Ceux qui sont impliqués dans un même processus biologique sont souvent regroupés dans un contexte génomique. La conservation de l'ordre des gènes, appelée aussi synténie, peut varier entre les génomes, mais les gènes restent dans le même contexte (Lathe *et al.*, 2000), on parle alors de contexte conservé ou de synténie conservée. De plus, la position des gènes par rapport à l'origine de réplication (Ori : région où commence la réplication de l'ADN) a aussi son importance. Il a été montré que chez les bactéries avec un fort taux de division, les gènes ayant un rôle essentiel sont plus proches de l'Ori afin d'être plus fortement exprimés (Sharp *et al.*, 1989; Vieira-Silva et Rocha, 2010).

Pour finir, les gènes peuvent être classés selon l'importance de leur fonction pour la survie de la cellule. Les gènes indispensables au cycle de vie d'une cellule, par exemple la réplication de l'ADN, la transcription, ou la traduction, sont dits "essentiels" et se distinguent des gènes "accessoires", qui codent pour des fonctions d'adaptation à des conditions particulières, comme la résistance aux antibiotiques, la défense contre les virus ou des réactions métaboliques spécifiques à une condition environnementale.

L'ADN non codant constitue une part tout de même non négligeable du génome et selon l'adage "la nature a horreur du vide"[1]. Il n'est donc pas inutile et renferme des fonctions essentielles à la vie de la cellule, comme les microARN et les ARN interférents (miARN et siARN). Ces ARN sont aujourd'hui considérés comme des acteurs clés dans la régulation des fonctions biologiques (Backofen *et al.*, 2014; Watkins et Arya, 2019), mais aussi dans d'autres processus comme le système immunitaire (Bobadilla Ugarte *et al.*, 2023).

1. Citation d'Aristote qui, répondant à Démocrite, rejetait l'idée du vide dans l'univers. On sait aujourd'hui que Démocrite avait raison, mais dans le cas des génomes procaryote, l'idée fonctionne.

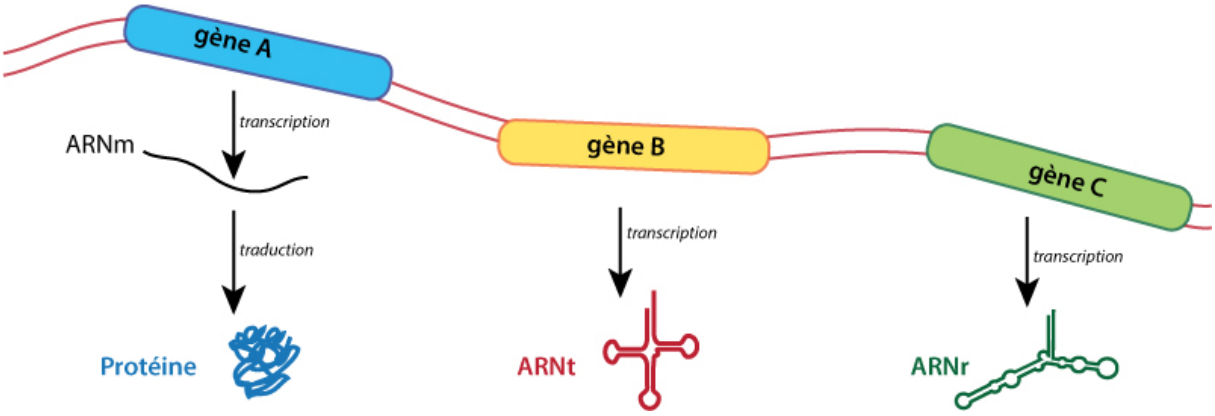


Figure I.2.1 – **Représentation des gènes et de leurs produits : protéines et ARN.** Un gène est d'abord transcrit en ARN. Si l'ARN transcrit est dit messager (ARNm), il sera ensuite traduit en protéine, sinon l'ARN produit (ARNt ou ARNr) aura un rôle spécifique dans des processus cellulaires. Copié de RNBio, Sorbonne université. https://rnbio.sorbonne-universite.fr/genetique_genotype1

L'ADN non codant n'a pas uniquement le rôle de contenir les séquences transcrites en ARN, il contient aussi d'autres éléments régulateurs de l'expression des gènes contenus dans l'espace intergénique (cf. sous-section 2.3.1). On retrouve aussi dans l'ADN non codant des séquences répétées, en bordure des séquences d'insertion (IS) (qui se déplacent dans le génome), ou dans les séquences CRISPR (Régions composées de répétitions palindromiques, régulièrement séparées par des séquences appelées *spacers*, impliquées dans le système immunitaire adaptatif des bactéries) (Jansen *et al.*, 2002; Bolotin *et al.*, 2005). Il existe tout de même une partie d'ADN non codant qui n'a aucun rôle, ces séquences peuvent faire partie de l'espace intergénique, ou être des vestiges d'anciens gènes qui, au cours de l'évolution, ont perdu leur fonction (cf. section 2.2). Pour terminer, c'est aussi dans le non codant que l'on va retrouver des éléments essentiels dans la réplication et l'évolution des génomes procaryotes, comme l'origine de réplication (Ori).

2.1.2 . Réplicons et mécanismes de réplication dans les génomes procaryotes

Le terme réplicon désigne l'ensemble des molécules d'ADN capables de se répliquer de façon autonome. Un réplicon contient ainsi tous les éléments nécessaires à l'exécution et à la régulation de la réplication. Une cellule va contenir au moins un réplicon, mais elle en contient souvent plus. Les réplicons sont souvent circulaires, mais ils peuvent aussi être linéaires.

La forme de réplicon qui est toujours présente dans la cellule est le chromosome. Le chromosome, souvent circulaire et replié, constitue le plus grand réplicon en termes de paires de bases [2]. Une cellule peut contenir plusieurs chromosomes, dans ce cas, le plus grand sera considéré comme le chromosome principal et les autres comme secondaires. Par exemple, chez *Rhodobacter sphaeroides* [3] et *Vibrio cholerae* [4], un second chromosome a été identifié (Suwanto et Kaplan, 1989; Trucksis *et al.*, 1998).

2. La taille d'une séquence ou d'un génome se mesure en bases (b) ou paires de bases (pb).
3. Bactérie présente dans les lacs profonds et les eaux stagnants. Capable de réaliser la photosynthèse et avec un métabolisme versatile, elle est largement utilisée en biotechnologie
4. Bactérie à l'origine du choléra, présente dans l'eau et transmissible entre humains, notamment via la transpiration.

Une seconde forme de réplicon, connue pour son rôle dans l'évolution (voir sous-section 2.2.2), est le plasmide (Lederberg et Tatum, 1946; Lederberg et Tatum, 1953). Les plasmides sont souvent circulaires et de petite taille (par rapport au chromosome). Bien que certains plasmides puissent coder leurs propres protéines de réplication, et donc répondent à la définition de réplicon, ils peuvent dépendre de la machinerie de la cellule pour se répliquer. Quoi qu'il en soit, la réplication est indépendante du chromosome, ce qui leur permet d'être présents sous un grand nombre de copies. L'origine de réplication des plasmides diffère de celle des chromosomes. Elles sont généralement plus courtes et spécifiques, tandis que celle du chromosome est plus complexe et conservée. Par ailleurs, les plasmides peuvent accumuler de nouvelles séquences et augmenter en taille, prenant alors la forme de mégaplasmides (figure I.2.2).

Chez la majorité des procaryotes, le chromosome contient les gènes essentiels, tandis que les plasmides portent des gènes accessoires. Cependant, certaines formes de réplicons oscillent entre chromosome et plasmide, que ce soit en termes de taille ou de contenu en gène. Le chromide (Harrison *et al.*, 2010) est une de ces formes. Sa taille est intermédiaire entre un plasmide et un chromosome principal, et il peut contenir des gènes essentiels à la cellule. Ces gènes présentent une proximité phylogénétique avec les espèces du même genre, contrairement à ceux du chromosome principal, qui sont conservés au-delà du genre. En revanche, en termes de mécanismes de réplication et de séquences Ori, les chromides utilisent des systèmes de type plasmidique.

L'usage des termes chromosome secondaire, chromide et mégaplasmide demeure actuellement peu standardisé dans la littérature (Hall *et al.*, 2021). Plusieurs critères permettent néanmoins de les distinguer. Le premier repose sur le contenu génétique : les mégaplasmides n'abritent pas de gènes essentiels, contrairement aux chromosomes secondaires et aux chromides. Le second critère est la composition en nucléotides, qui est plus proche de celle du chromosome principal pour les chromides et les chromosomes secondaires. Enfin, leur origine évolutive les différencie : le chromosome secondaire résulte de la scission d'un chromosome ancestral en un chromosome principal et un secondaire, tandis que le chromide dérive d'un ancien mégaplasmide ayant perdu sa capacité de mobilité (voir sous-section 2.2.2) et qui a intégré des gènes essentiels (figure I.2.2). Les chromides auraient donc plutôt un rôle de réservoir de gènes d'intérêt et d'adaptation améliorant la *fitness* des organismes. Cette vision vertueuse de l'accumulation de gènes s'oppose directement à la vision plus ancienne des plasmides non mobilisables décrits comme parasitant la cellule (Levin, 1993; Lili *et al.*, 2007).

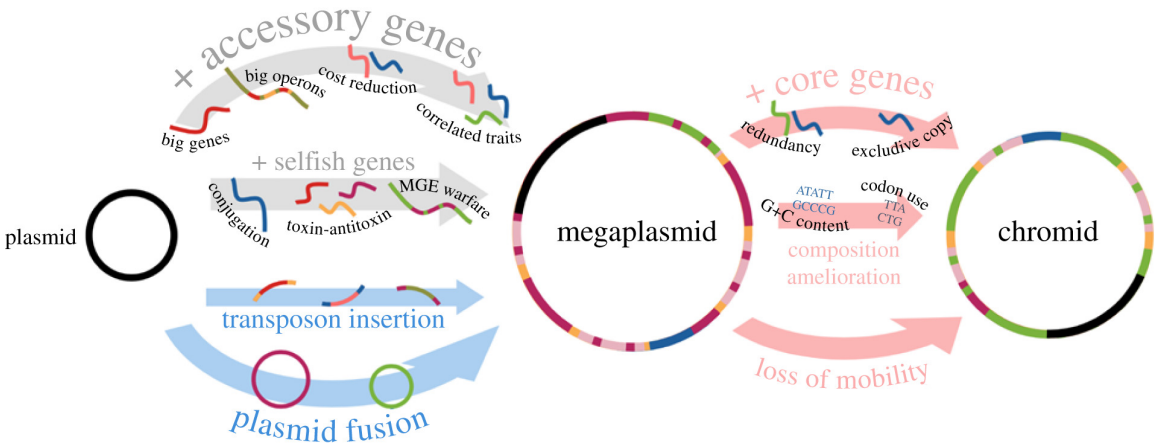


Figure I.2.2 – **Schéma simplifié de l'évolution d'un plasmide en megaplasmide et de megaplasmide à chromide.** Figure extraite de (Hall *et al.*, 2021)

Le génome procaryote correspond à l'ensemble des réplicons présents dans la cellule. La taille des génomes est comprise entre quelques centaines de milliers de bases à plusieurs millions de bases pour certains génomes [5] (figure I.2.3). Cette relative petite taille est optimisée par la structure des génomes et la proportion de séquence codante. Elle est aussi liée au mode de vie : les organismes endosymbiotiques ou pathogènes obligatoires, fortement dépendants de leur hôte, ont souvent un génome réduit en raison de la perte de gènes non essentiels. À l'inverse, les bactéries à vie libre possèdent généralement un génome plus large, leur permettant une plus grande autonomie métabolique et une meilleure adaptation aux variations environnementales.

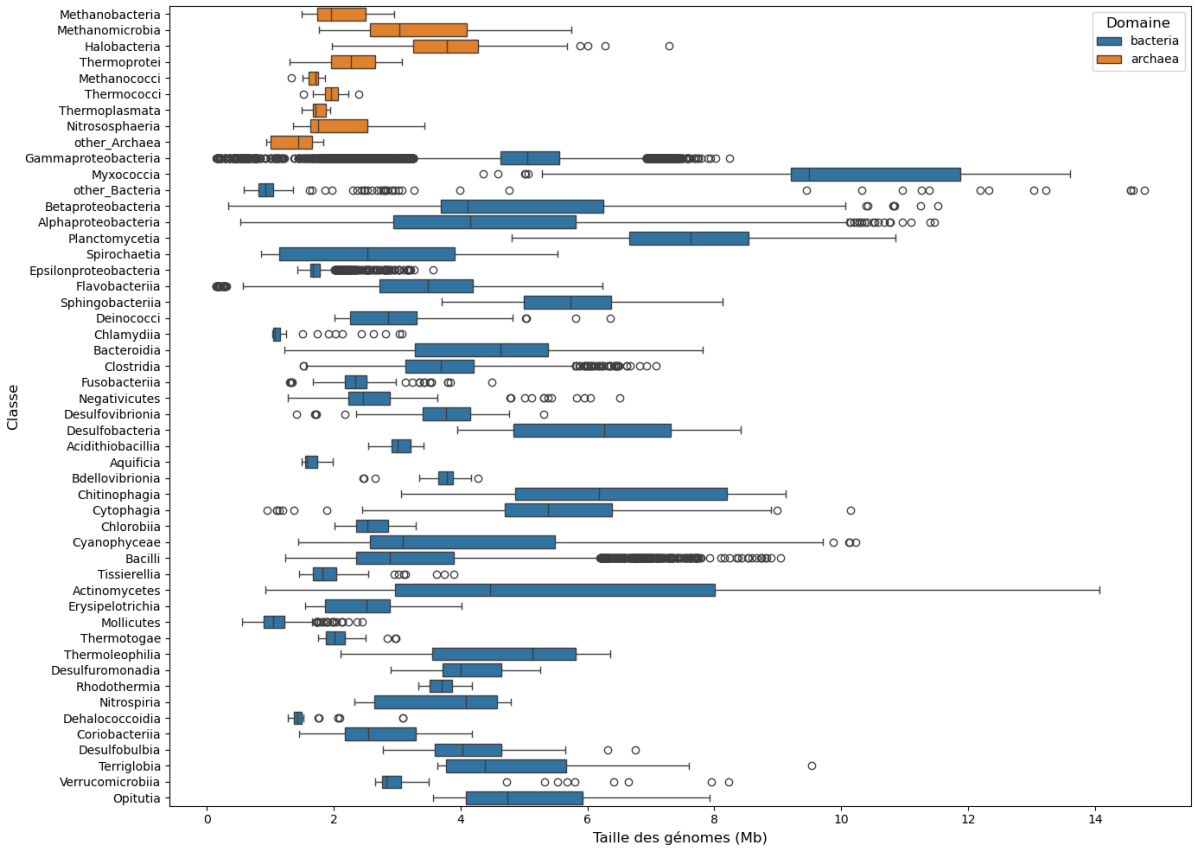


Figure I.2.3 – **Distribution de la taille des génomes (en base) par classe chez les procaryotes.** Les données utilisées proviennent de RefSeq version 28 janvier 2025.

5. Un génome procaryote est compris entre 100 kb et 15 Mb. Pour comparaison, le génome humain mesure environs 3 Gb.

2.2 . Dynamique évolutive des génomes

La dynamique évolutive des procaryotes est caractérisée par des processus continus de gain, perte et modification de gènes (figure I.2.4). La taille des génomes étant restreinte, la perte de gènes peut optimiser le génome en éliminant les séquences redondantes ou non essentielles, favorisant ainsi une efficacité accrue dans des environnements spécifiques. Les modifications génétiques, quant à elles, jouent un rôle crucial dans l'adaptation fine des procaryotes face aux pressions sélectives variées. L'acquisition de nouveaux gènes introduit une diversité génétique, pouvant conférer des traits avantageux, tels que la résistance aux antibiotiques ou la capacité à métaboliser de nouvelles sources de nutriments.

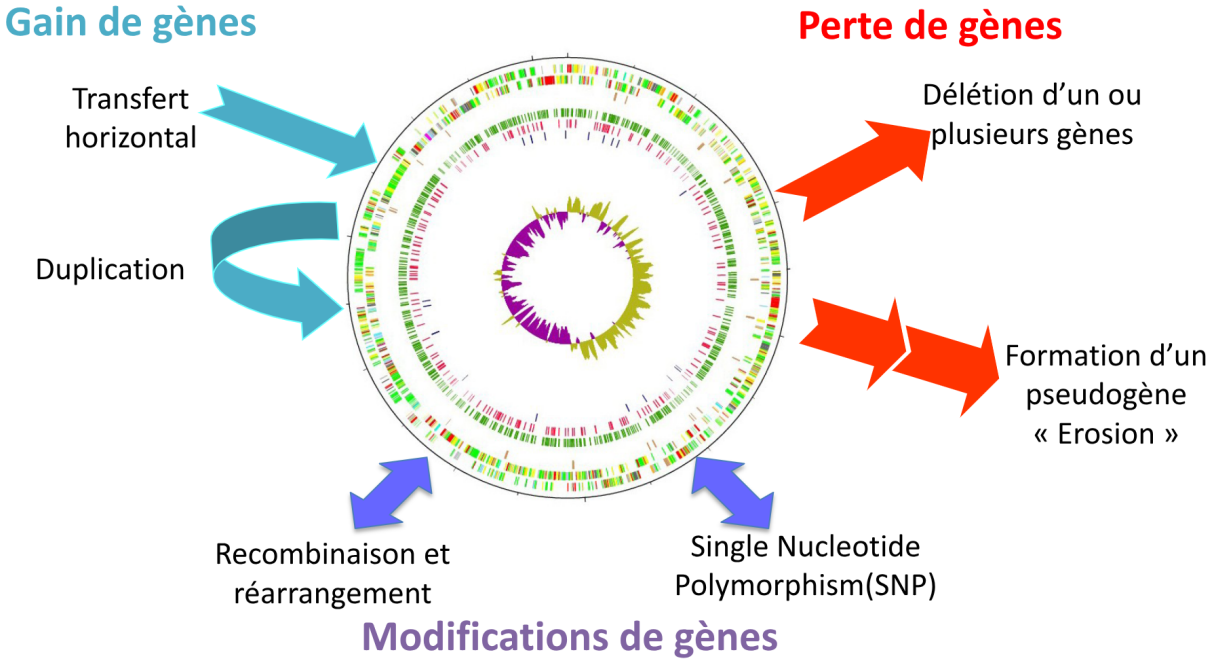


Figure I.2.4 – **Schéma résumant la dynamique évolutive des génomes procaryotes.** Source LABGeM

Les gènes doivent ensuite être transmis dans la population. Les **transferts verticaux** permettent de transférer les gènes de génération en génération, assurent la continuité et la stabilité des traits essentiels. Les **transferts horizontaux** permettent l'échange de gènes entre les organismes, favorisant une diversification rapide des génomes, qui peut radicalement transformer les capacités adaptatives des lignées procaryotes. Cette dynamique complexe façonne la biodiversité procaryote et témoigne de la capacité évolutive exceptionnelle de ces organismes à coloniser une multitude d'écosystèmes.

2.2.1 . Mécanismes d'évolution par transfert vertical

Les mécanismes d'évolution par héritage regroupent les processus menant à une modification du génome entre la cellule mère et la cellule fille. Théoriquement, lors de la division cellulaire, la cellule mère se divise en 2 cellules filles possédant exactement la même information génétique qu'elle. Pourtant, malgré un ensemble de mécanismes de protection et de correction de l'ADN, le génome peut différer entre les cellules mère et fille. Ce sont ces "erreurs" qui vont nous intéresser, car ce sont elles qui sont à l'origine de l'innovation et de la diversité génétique.

2.2.1.1 . Impact des mutations génétiques : SNPs, Indels et pseudogènes

a. *Single Nucleotid Polymorphism*

Un *Single Nucleotide Polymorphism* (SNP) correspond à une modification de la séquence induite par la mutation d'un nucléotide en un autre. Étant donné que le code génétique est dégénéré [6], la mutation peut ne pas avoir d'impact sur la séquence de la protéine, on dit alors que la mutation est silencieuse ou même sens. Si la modification entraîne un changement d'acide aminé dans la séquence protéique, on parle de mutation faux-sens. Enfin, une mutation est qualifiée de non-sens lorsqu'elle introduit prématurément un codon STOP, interrompant ainsi la traduction et conduisant à une perte de fonction de la protéine. Une telle mutation peut également affecter un site fonctionnel clé (comme un site actif), compromettant l'activité de la protéine. Lorsque l'introduction d'un codon STOP précoce rend un gène non fonctionnel, ce dernier devient un **pseudogène**, un vestige génomique dépourvu de rôle biologique actif, un phénomène appelé pseudogénisation.

Sur la figure I.2.5, la première mutation implique un changement de glutamine en histidine, des acides aminés aux propriétés de polarité et de charge différentes. Il s'agit donc d'une mutation faux-sens, qui aura probablement un impact significatif sur la structure de la protéine. En revanche, les deux autres SNPs ne modifient pas l'acide aminé codé, ils sont donc silencieux.

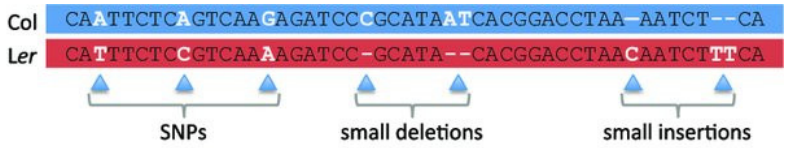


Figure I.2.5 – **SNP et InDels entre deux génomes.** On suppose que le premier codon commence par le premier nucléotide. Figure extraite et adaptée de (Qi *et al.*, 2014)

b. Indels : insertion, délétion et pseudogènes

Un indel correspond à l'insertion (In) ou la délétion (del) [7] d'un ou plusieurs nucléotides dans la séquence d'un gène. Lorsque la taille de l'indel est un multiple de 3 (insertion ou délétion d'un codon), la séquence protéique peut soit être allongée, soit raccourcie d'un acide aminé, soit coupée de façon précoce si le codon est un codon STOP.

Si la taille de l'indel n'est pas un multiple de 3, il y aura un décalage du cadre de lecture ou *frameshift*. Ce décalage va induire un changement de tous les acides aminés de l'indel à la fin du gène, provoquant avec lui un changement dans la fonction de la protéine ou une inactivation de la fonction. La partie du gène qui n'est pas décalée est alors considérée comme un fragment du gène initial, il est alors qualifié de pseudogène. À nouveau, cette mutation peut être délétère pour la cellule. Sur la figure I.2.5, les indels sont de taille 1 et 2, elles ne provoquent pas l'apparition d'un codon STOP précoce, mais l'ensemble des acides aminés est modifié.

Les indels vont donc transformer la séquence protéique traduite, pouvant nuire à la fonction de cette dernière et être délétère pour l'organisme. Pour éviter les problèmes liés aux *frameshifts*, il a été montré qu'il existe un fort taux de codon STOP hors du cadre de lecture (Tse *et al.*, 2010). Cette adaptation permettrait de limiter la traduction

6. Un acide aminé peut être codé par plusieurs codons différents.
7. On regroupe l'insertion et la délétion, car sans une analyse phylogénétique, il est impossible de les différencier par comparaison de séquence.

des protéines mutantes et d'ainsi limiter le coût énergétique pour la cellule. Il a aussi été montré que les *frameshifts* pourraient être à l'origine d'un réservoir d'adaptation à l'environnement (Koch, 2004). Lors d'un changement dans l'environnement créant une nouvelle pression de sélection, un *frameshift* pourrait produire une protéine qui permet à l'organisme de s'adapter à son environnement et donc d'améliorer sa *fitness* [8]. Une fois que l'élément perturbateur de l'environnement disparaît, un nouveau *frameshift* pourrait ramener le cadre de lecture à sa place d'origine. Ce mécanisme, en accord avec la petite taille des génomes, aurait l'intérêt de ne pas perdre des gènes d'adaptation à l'environnement, même s'ils ne sont nécessaires que ponctuellement.

2.2.1.2 . Réarrangement génomique : un moteur de l'évolution

Les génomes évoluent également suite à des événements de réarrangement. Ils impliquent des segments d'ADN plus importants. La forme du génome obtenue, appelée variant structural (SV pour *Structural variant* en anglais), est plus difficile à détecter que les SNP et les indels (Periwal et Scaria, 2015).

Le mécanisme de recombinaison est à l'origine des réarrangements. Une recombinaison implique l'échange de 2 portions d'ADN entre 2 molécules ou 2 régions d'ADN. La recombinaison peut être homologue, se produisant entre des séquences similaires, ou non-homologue, impliquant des séquences différentes. Elle est souvent médiée par des enzymes spécialisées comme RecA ou des intégrases, qui permettent l'intégration, la réparation ou le réarrangement précis des séquences. La recombinaison homologue est cruciale pour la réparation des cassures de l'ADN, les réarrangements et également dans l'acquisition de nouveaux gènes par transfert horizontal (cf. sous-section 2.2.2) (Eisenstark, 1977).

Les réarrangements de l'ADN correspondent donc à un échange entre 2 segments du génome, induisant une insertion, une délétion ou une modification de l'ordre des nucléotides (figure I.2.6). Les réarrangements sont fréquents dans les génomes procaryotes (Sun *et al.*, 2012) et peuvent être spontanés ou facilités par la présence d'éléments mobiles, tels que les transposons, qui sont des séquences d'ADN capables de se déplacer au sein du génome. Ils sont composés de gènes codant pour une transposase, l'enzyme responsable de son déplacement, ainsi que de séquences répétées aux extrémités, nécessaires à la reconnaissance et à l'excision du transposon.

L'ordre des gènes étant important dans l'expression des gènes et la fonction des protéines, le SV résultant peut conduire à une modification de l'expression génique ou à un changement dans la fonction de la protéine. Il existe 3 formes de réarrangement : symétrique, asymétrique et au sein d'un réplicon. Ces formes ne sont pas toutes équiprobables, car elles affectent plus ou moins la structure du génome. Aussi, les réarrangements proches de l'Ori sont plus fréquents que ceux proches du site de terminaison (Darling *et al.*, 2008).

Les recombinaisons peuvent également conduire à la duplication de gènes ou de régions génomiques, un mécanisme clé dans l'évolution des procaryotes en générant une redondance génétique. Cette redondance offre une opportunité évolutive : tandis qu'une copie du gène conserve sa fonction initiale, l'autre peut accumuler des mutations, potentiellement aboutissant à une nouvelle fonction, sans compromettre la survie de l'organisme. En outre, la duplication peut jouer un rôle dans la régulation de l'expression génique. Par exemple, les gènes codant pour les pompes à efflux, impliquées

8. Le *fitness* correspond à la capacité d'un individu de survivre dans son environnement et à se reproduire

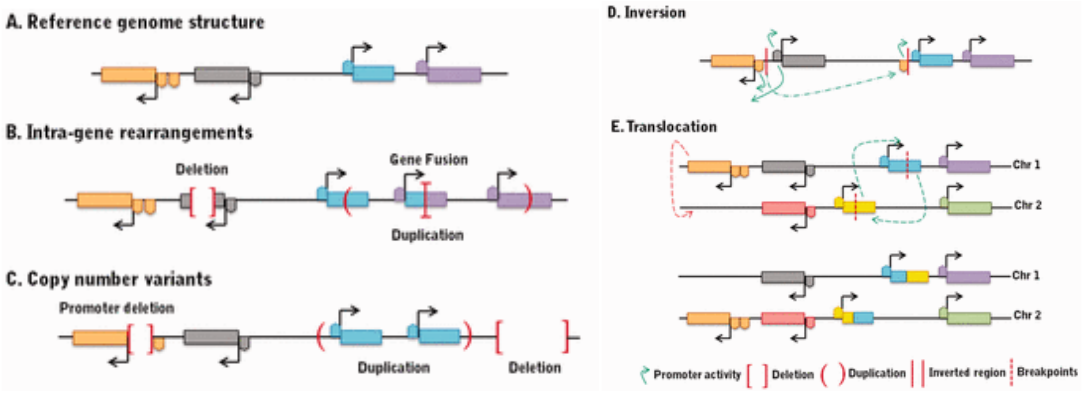


Figure I.2.6 – **Réarrangement et conséquences des variants structuraux.** (A) Région génomique sans SV. Les rectangles représentent les gènes et les petits connecteurs à côté représentent le promoteur du gène concerné. (B) Réarrangement intragénique illustrant la délétion et la fusion de gènes à la suite d'une duplication partielle du gène. Les régions codantes modifiées produisent des transcrits aberrants. La délétion ou la duplication peut entraîner une modification du nombre des gènes dans des régions par ailleurs fonctionnellement intactes. (C) Délétion du promoteur, la régulation est modifiée et une duplication/délétion qui modifie le nombre de copies des gènes. (D) Inversions affectant la structure du gène, le gène est inversé, retourné et réarrangé, ce qui éloigne l'un des promoteurs du premier gène (orange). (E) Translocations affectant le contexte génique. Figure extraite et adaptée de (Periwal et Scaria, 2015)

dans l'évacuation des antibiotiques hors de la cellule, sont fréquemment dupliqués, favorisant ainsi une meilleure résistance aux traitements (Maddamsetti *et al.*, 2024). Toutefois, les événements de duplication restent moins fréquents que les transferts horizontaux de gènes dans les génomes procaryotes (Tria et Martin, 2021). Cette rareté s'explique en partie par les mécanismes d'élimination de la redondance, qui optimisent la compacité et l'efficacité des génomes bactériens.

Les mécanismes qui viennent d'être décrits apportent de l'innovation dans les génomes procaryotes, qui doit ensuite être transmise dans la population. Avec le transfert vertical, cette transmission se fait uniquement d'une génération à l'autre, un processus limité par le temps de génération, qui varie selon l'espèce (*E. coli* : 20 min, *Lactobacillus acidophilus* : 80 min, *Mycobacterium tuberculosis* : 800 min). Un temps de génération plus long semble aussi réduire le taux de mutation spontanée de l'ADN (Weller et Wu, 2015). Pour contourner ces contraintes, les procaryotes échangent de l'ADN avec leur environnement (autres bactéries, virus, eucaryotes, ADN libre…), par un ensemble de processus regroupé sous le terme de **transfert horizontal**, qui leur permet d'acquérir de nouvelles fonctions génétiques.

2.2.2 . Mécanismes d'évolution par transfert horizontal

Les transferts horizontaux de gènes (*Horizontal Gene Transfert* en anglais, HGT) constituent un phénomène central dans l'évolution des procaryotes, permettant l'échange de matériel génétique entre organismes sans nécessiter une relation de lignage directe. La proportion de gènes acquis par transfert horizontal varie considérablement selon les espèces et les environnements, mais elle peut représenter une part significative du génome procaryote. On estime que 20 % des gènes en moyenne ont été acquis par HGT, certaines études montent même jusqu'à 25 % pour certaines bactéries (Ochman *et al.*, 2000; Popa *et al.*, 2011). Cette proportion élevée témoigne de l'importance des HGT dans l'évolution et l'adaptation des procaryotes.

Les gènes sont transférés via des éléments génétiques mobiles (MGE), incluant les plasmides, les transposons et les phages (virus de bactérie, cf. sous-sous-section 3.2.3.1), chacun possédant des capacités uniques pour mobiliser les gènes. Ces vecteurs facilitent le transfert et l'intégration de l'ADN étranger dans le génome hôte. Les séquences répétées, telles que les insertions et les répétitions en tandem, jouent également un rôle, en servant de sites d'intégration pour les MGE.

Il existe 3 grands mécanismes de HGT : la **transformation**, la **conjugaison** et la **transduction**, chacun facilitant le mouvement de gènes entre cellules de manière distincte.

2.2.2.1 . Conjugaison : la sexualité des procaryotes

La conjugaison a été découverte en 1946 par Joshua Lederberg et Edward L. Tatum (Lederberg et Tatum, 1953), qui décrivent ce mécanisme comme la manière sexuée des bactéries d'échanger de l'ADN. En effet, par analogie, la conjugaison demande un contact direct entre une cellule donneuse et une cellule receveuse pour l'échange de matériel génétique [9]. Il existe 2 catégories d'éléments génétiques mobiles conjugatifs : les plasmides et les éléments intégratifs et conjugatifs (ICEs, *Intergrative and Conjugative elements* en anglais). Sur la figure I.2.7 est représenté l'échange d'un plasmide par conjugaison. Les ICEs (Johnson et Grossman, 2015), contrairement aux plasmides, sont directement intégrés au chromosome, ce qui rend leur réplication dépendante de celui-ci. Toutefois, cette intégration favorise un transfert vertical plus stable au cours des générations. Les ICEs pour être échangés doivent suivre un schéma circulaire : excision du chromosome, circularisation, réplication, transfert et réintégration dans le chromosome. Lors de l'étape d'excision, il peut arriver que des gènes flanquant l'ICEs soient excisés aussi, apportant une nouvelle forme à l'ICE (Gibbons *et al.*, 2011).

Plasmides et ICEs sont généralement de petite taille, mais ils contiennent des gènes clés d'adaptation à l'environnement. La présence de ces gènes dans les éléments mobiles permet à des colonies de répondre efficacement et rapidement aux nouvelles conditions environnementales, comme la présence de métaux lourds ou d'antibiotiques (Botelho et Schulenburg, 2021). Toutefois, tous les MGEs ne sont pas forcément conjugatifs (Valentine *et al.*, 1988), ils vont profiter de la conjugaison codée par un autre élément pour se transférer. Dans ces conditions, la bactérie receveuse ne devient pas conjugative à son tour, même si elle reçoit l'élément mobile. Ces éléments mobilisables sont appelés des IMEs (élément intégratif mobilisable). Il est d'ailleurs à noter que tous les plasmides ne sont pas mobilisables, il y aurait d'ailleurs autant de plasmides conjugatifs que de plasmides non mobilisables (Smillie *et al.*, 2010).

La conjugaison est un mécanisme majeur de transfert horizontal de matériel génétique, qui a la caractéristique de rapidement répandre les éléments mobiles. Il a toutefois le défaut de limiter le transfert de gènes entre cellules procaryotes et donc de limiter le transfert aux innovations génétiques déjà intégrées par un autre organisme procaryote. De plus, tous les organismes ne sont pas capables de réaliser la conjugaison, ce qui réduit d'autant plus la capacité de transfert au niveau des communautés.

9. N.B : Le transfert est unidirectionnel, la cellule donneuse ne peut recevoir de l'ADN et la receveuse ne peut en donner.

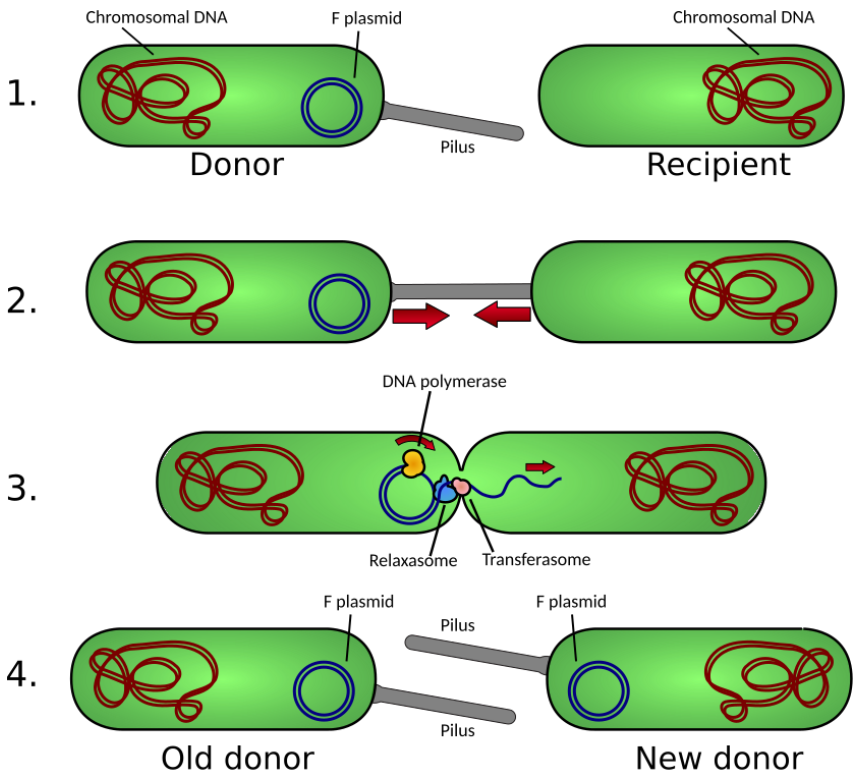


Figure I.2.7 – **Schéma du fonctionnement de la conjugaison, dans le cas d'un plasmide conjugatif.** (1) Formation d'un pili sexuel par la bactérie donneuse. (2) Contact direct entre les 2 bactéries via le pili. (3) Réplication de l'ADN plasmidique et transfert à la bactérie donneuse. (4) Terminaison de la conjugaison et nouvelle formation d'un pili pour la receveuse devenue donneuse. Image sous licence Creative Commons 3.0 https://commons.wikimedia.org/wiki/File:Conjugation.svg

2.2.2.2 . Transformation : recycler l'ADN environnant

La transformation correspond à l'intégration d'un fragment d'ADN étranger dans le génome de l'organisme. Les bactéries pouvant réaliser la transformation sont dites compétentes. Ce qui différencie la transformation de la conjugaison, c'est que l'ADN intégré est libre dans l'environnement [10]. De plus, la transformation est la seule forme de HGT, totalement contrôlée par la cellule receveuse (Huang *et al.*, 2021). Assez peu d'espèces sont connues pour être capables de réaliser la transformation de manière naturelle, toutefois un nombre plus important contient la machinerie nécessaire à sa réalisation (Johnston *et al.*, 2014). De plus, au sein d'une espèce, le taux d'individu compétent peut varier, par exemple chez *S. pneumoniae*, 66 % des individus sont capables de la réaliser (Evans et Rozen, 2013). Pour terminer, les mécanismes de la transformation, notamment l'incorporation de l'ADN dans la cellule (figure I.2.8), sont bien décrits dans la littérature (Johnston *et al.*, 2014; Dubnau et Blokesch, 2019). Toutefois, ils varient d'une espèce procaryote à l'autre, tout comme la proportion d'individus capables de réaliser cette transformation (Stewart et Carlson, 1986). Nous ne reviendrons donc pas sur les mécanismes, mais seulement sur des exemples d'application.

10. La découverte de la transformation en 1928 par Fred Griffith (Griffith, 1928), précède de nombreuses années celle qui a mis en évidence que l'ADN est le porteur de l'information génétique (Avery *et al.*, 1944). La transformation est donc une preuve anticipée et un socle pour démontrer le rôle de l'ADN.

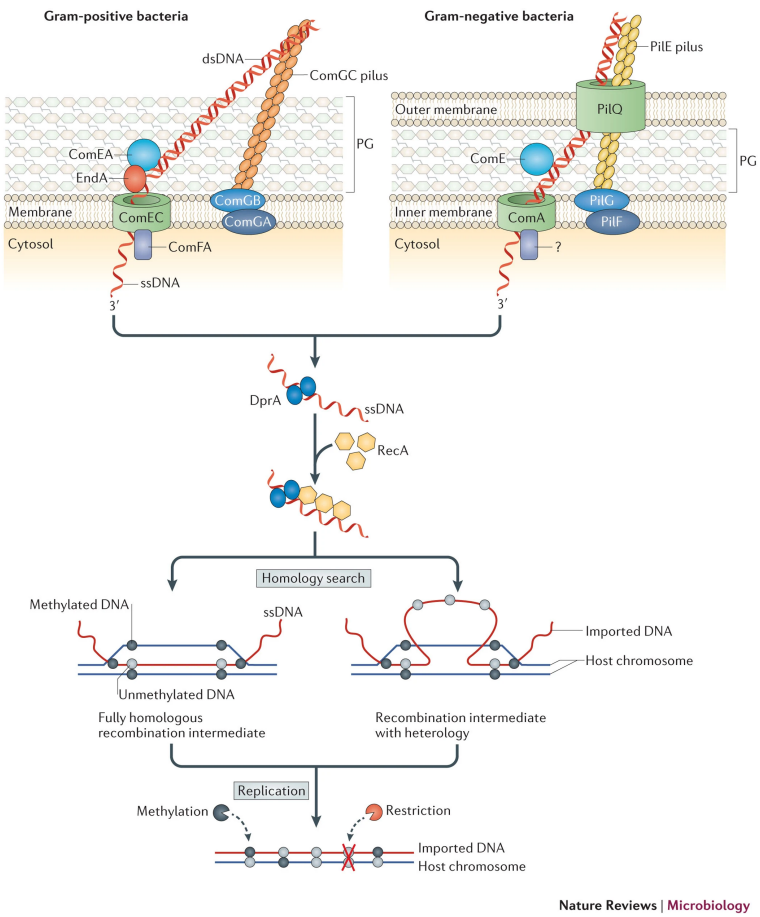


Figure I.2.8 – **Schéma du mécanisme de transformation.** Extrait de (Johnston *et al.*, 2014)

Les bactéries du genre *Nesseria* et particulièrement *N. gonorrhoeae* [11] reconnaissent préférentiellement une séquence d'ADN non palindromique de leur propre ADN (Goodman et Scocca, 1988; Duffin et Seifert, 2010). Ce système permet d'intégrer uniquement l'ADN de souches proches, ainsi que des gènes d'adaptation, comme des gènes de résistance aux antibiotiques (Centers for Disease Control and Prevention (CDC), 2007). Ainsi, les gènes d'adaptation d'intérêt sont préférentiellement distribués dans l'espèce.

Streptococcus pneumoniae [12] utilise la transformation comme mécanisme de réparation de l'ADN, car cette espèce ne possède pas de système de réparation SOS (Gasc *et al.*, 1980). Les souches de *S. pneumoniae* s'engagent alors dans une "guerre fratricide" pour récupérer l'ADN des autres souches de leur espèce (Claverys et Håvarstein, 2007).

Pour terminer, chez *Bacillus subtilis* [13], la transformation entre individus de la même espèce, mais de souche éloignée, est privilégiée (Lyons *et al.*, 2016). Les bactéries vont sécréter dans l'environnement des antibiotiques, auxquels elles sont résistantes, pour tuer les autres individus de l'espèce. L'ADN récupéré est donc différent de celui de la bactérie et donc potentiellement source de nouvelles fonctions.

11. Ce genre bactérien, vivant dans les muqueuses des mammifères, est non pathogène à l'exception de *N. meningitidis*, impliqué dans la méningite et *N. gonorrhoeae*, responsable de la gonorrhée, une infection sexuellement transmissible.

12. Bactérie connue pour son rôle d'agent pathogène dans les pneumonies et responsable de co-infection pendant la grippe espagnole

13. Bactérie du sol, mais qu'on retrouve dans de nombreux habitat dû à ses capacités d'adaptation. Elle est utilisée comme modèle d'étude des bactéries Gram+.

Ces exemples montrent aussi une opposition dans la philosophie des mécanismes de conjugaison et de transformation. La transformation demande que l'ADN soit libre dans l'environnement et donc que les bactéries environnantes soient détruites, alors que la conjugaison laisse les 2 cellules en vie.

2.2.2.3 . Transduction : les virus mis à profit

La transduction est un mécanisme reposant sur l'intervention d'un virus pour transporter et transférer le matériel génétique d'une cellule procaryote à l'autre (figure I.2.9). Les virus de bactéries, surnommés (bacterio)phages, vont infecter la cellule donneuse pour répliquer leur ADN. Lors de la réplication, de l'ADN de la cellule donneuse peut se trouver intégré à celui du phage. Lorsqu'il infectera une cellule receveuse, la portion d'ADN de la donneuse pourra reprendre une forme plasmidique (si c'est un plasmide qui a été transféré) ou être intégrée au génome de la cellule par recombinaison homologue. La transduction est aujourd'hui largement utilisée en génétique et microbiologie pour transférer de l'ADN et modifier les génomes (Wang *et al.*, 2024a).

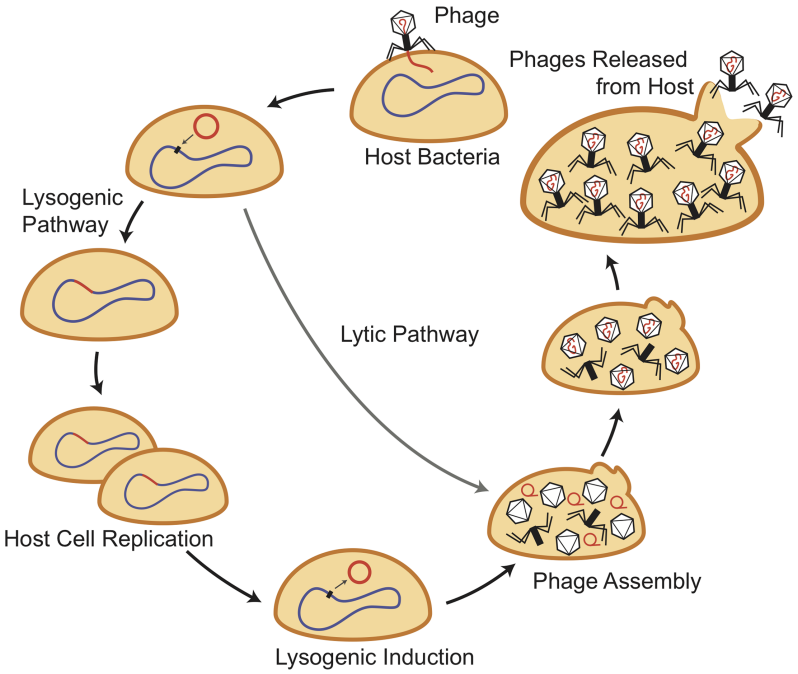


Figure I.2.9 – **Schéma représentant les étapes de transduction.** Extrait de (Chiang *et al.*, 2019)

La première forme de transduction identifiée décrivait le transfert de n'importe quel gène de la donneuse à la receveuse par le phage. Cette forme a donc été nommée transduction généralisée (Zinder et Lederberg, 1952). Une seconde forme dite spécifique a été découverte en étudiant le phage λ infectant les *E. coli* (Morse *et al.*, 1956). Le transfert se limite à un ensemble de gènes définis. Enfin, une dernière forme, la transduction latérale, a récemment été découverte (Chen *et al.*, 2018). Là où les formes générale et spécifique peuvent être vues comme une erreur et un événement lié au hasard, la transduction latérale fait partie du cycle de vie du phage, menant à un taux de transfert beaucoup plus important.

2.3 . Du génome aux processus cellulaires

2.3.1 . Gènes : Régulations et fonctions

Les réactions qui se produisent dans les cellules procaryotes sont souvent complexes et impliquent une multitude de réactifs et de produits. Toutes ces réactions nécessitent la présence de protéines spécifiques pour être réalisées. Ces protéines sont produites et dégradées par la cellule en fonction des conditions rencontrées. C'est pourquoi l'information est stockée dans une structure durable et transmissible, le gène. Chaque gène sera transcrit en une molécule d'ARN messager (ARNm) par l'ARN polymérase, qui sera traduite en protéine par le ribosome (impliquant l'ARNr et l'ARNt).

Dans une cellule, les protéines ont un temps de "vie" allant de quelques minutes à quelques heures. Il est donc nécessaire de produire les protéines régulièrement, toutefois cette production a un coût pour la cellule. C'est pourquoi il existe des mécanismes de régulation de l'expression des gènes et donc de la production des protéines. Dans la sous-section 2.1.1, nous avons vu qu'il existait notamment des petits ARN régulateurs de l'expression. Ils agissent en modulant la stabilité ou la traduction des ARN messagers cibles, jouant ainsi un rôle clé dans l'adaptation aux stress environnementaux, la régulation du métabolisme ou encore la virulence par exemple. Dans l'ADN non codant, on retrouve également une séquence promotrice (ou promoteur) près d'un gène qui permet la fixation de l'ARN polymérase. La fixation et l'activation de l'ARN polymérase au niveau du promoteur sont régulées par des facteurs de transcription qui se lient spécifiquement à des séquences régulatrices en amont du promoteur, les *enhancer* et *silencer*.

Les protéines peuvent agir en collaboration, soit dans des réactions successives (cas des systèmes biologiques, cf. section 3.2), soit en formant des complexes protéiques interagissant pour métaboliser un produit. Les gènes codant pour des protéines impliquées dans les mêmes processus cellulaires sont situés dans le même contexte génomique (cf. sous-section 2.1.1). Ils vont alors être régulés par les mêmes éléments de régulation. L'opéron, une structure spécifique des procaryotes découverte par François Jacob et Jacques Monod en 1960 [14](Jacob et Monod, 1961), permet de produire un seul ARNm pour un ensemble de gènes codant pour des protéines impliquées dans le même processus cellulaire. Dans l'opéron, se trouve une nouvelle séquence de régulation, l'opérateur, où va se lier une molécule régulatrice qui va activer ou inhiber la transcription (figure I.2.10). L'ensemble de l'opéron permet de synchroniser la régulation et l'expression de gènes qui collaborent dans le même processus cellulaire.

Pour terminer, l'expression des gènes peut aussi être régulée par le niveau de repliement et de condensation de l'ADN. L'ADN est condensé notamment grâce à des protéines spécialisées et à la méthylation de l'ADN. L'ADN replié ne pourra pas être accessible pour la transcription des gènes et donc ils seront inactifs. Les mécanismes liés à la méthylation de l'ADN sont l'affaire de l'épigénétique. Des études récentes ont mis en lumière le rôle de la méthylation dans la régulation de la virulence bactérienne et dans la capacité des procaryotes à coloniser leurs hôtes (Oliveira, 2021), soulignant ainsi l'importance de ces mécanismes dans la survie et l'adaptation des bactéries.

14. Découverte qui leur a valu le prix Nobel de médecine en 1965

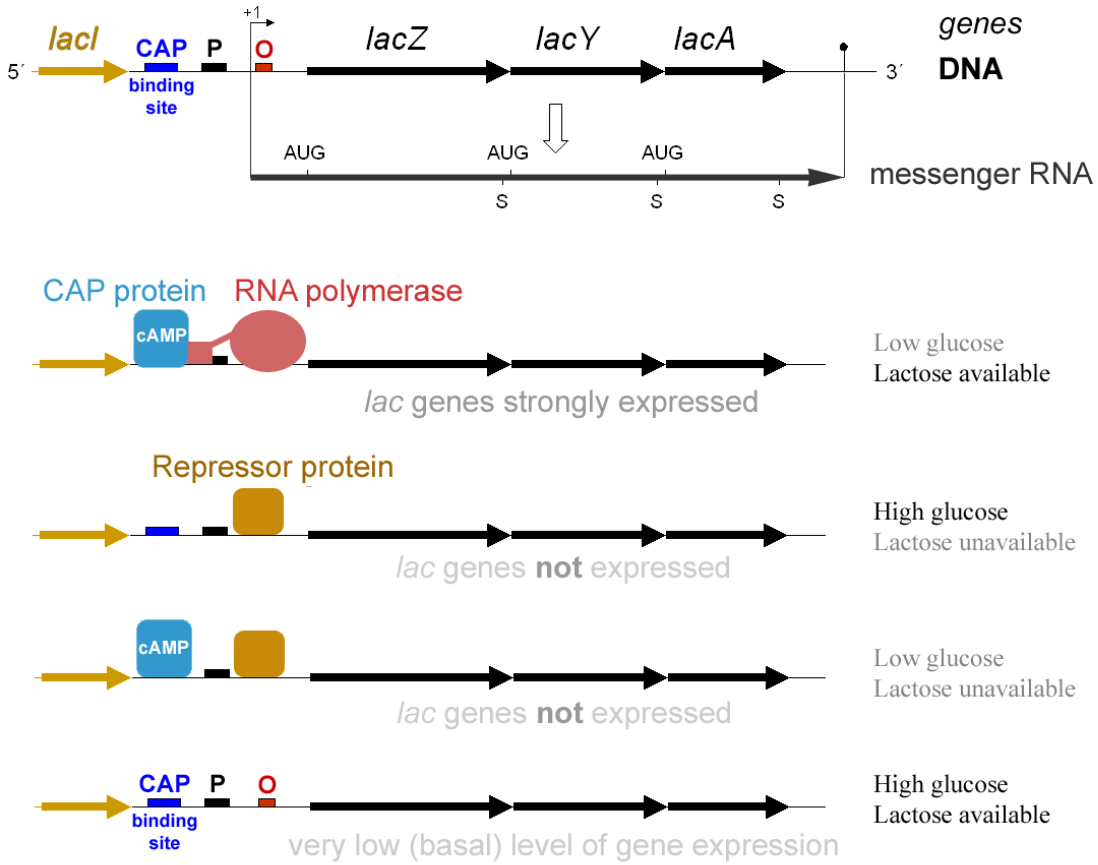


Figure I.2.10 – **Schéma du fonctionnement de l'opéron lactose.** Sur la partie haute est représenté la structure génétique de l'opéron. Les 4 lignes suivantes représentent chacune une configuration de réponse à des conditions de présence, absence de glucose et de lactose. si le taux de glucose est faible, une protéine activatrice (CAP) va se fixer en amont du promoteur pour aider à la fixation de l'ARN polymérase, et si du lactose est disponible, les gènes seront alors fortement exprimés. Si le lactose n'est pas disponible, une protéine de répression va se fixer à l'opérateur et elle empêchera l'ARN polymérase de se fixer même si le taux de glucose est faible. Auteur : G3pro. Sous licence Creative Commons 2.0. Disponible à l'adresse : `https://commons.wikimedia.org/wiki/File:Lac_operon-2010-21-01.png`.

2.3.2 . Îlots génomiques et points chauds d'insertion

Les îlots génomiques (GI, pour *Genomic Island en anglais*) sont des régions spécifiques du génome qui jouent un rôle clé dans l'évolution, l'adaptation et l'acquisition de fonctions spécifiques. Les GIs sont retrouvés chez quasiment tous les organismes procaryotes. Ils sont généralement acquis par transfert horizontal (cf. sous-section 2.2.2) et transportent des gènes accessoires. Ils vont conférer à l'organisme de nouvelles fonctions qui impacteront de façon positive sa *fitness*. Le premier îlot génomique décrit était lié à la capacité de la bactérie *E. coli* de provoquer des maladies et a donc été nommé îlot de pathogénicité (Hacker *et al.*, 1990). Depuis, d'autres classes d'îlots ont été découvertes : métabolique, résistance, symbiotique...(figure I.2.11).

Les îlots génomiques sont des régions assez larges, entre 5 et 200 kb (mais certaines sont beaucoup plus grandes) et présentent des caractéristiques spécifiques. (*i*) Les GIs ont un taux de GC qui diffère par rapport au reste du génome, résultant en un biais d'usage des codons [15] (figure I.2.11). (*ii*) dans les régions flanquantes des GIs, on retrouve des gènes de mobilité : transposases et intégrases, mais aussi d'IS qui peuvent se dégrader rapidement après l'intégration de l'îlot. (*iii*) Dans les gènes flanquants, on retrouve des gènes codant l'ARNt dont l'origine serait à relier à la prévalence des gènes de phages et des ICEs qui utilisent les ARNt comme site d'intégration dans les génomes (Dobrindt *et al.*, 2004). (*iv*) Les protéines contenues dans les GIs ont souvent des fonctions inconnues. (*v*) Dans la partie flanquante, on trouve des séquences répétées directes [16].

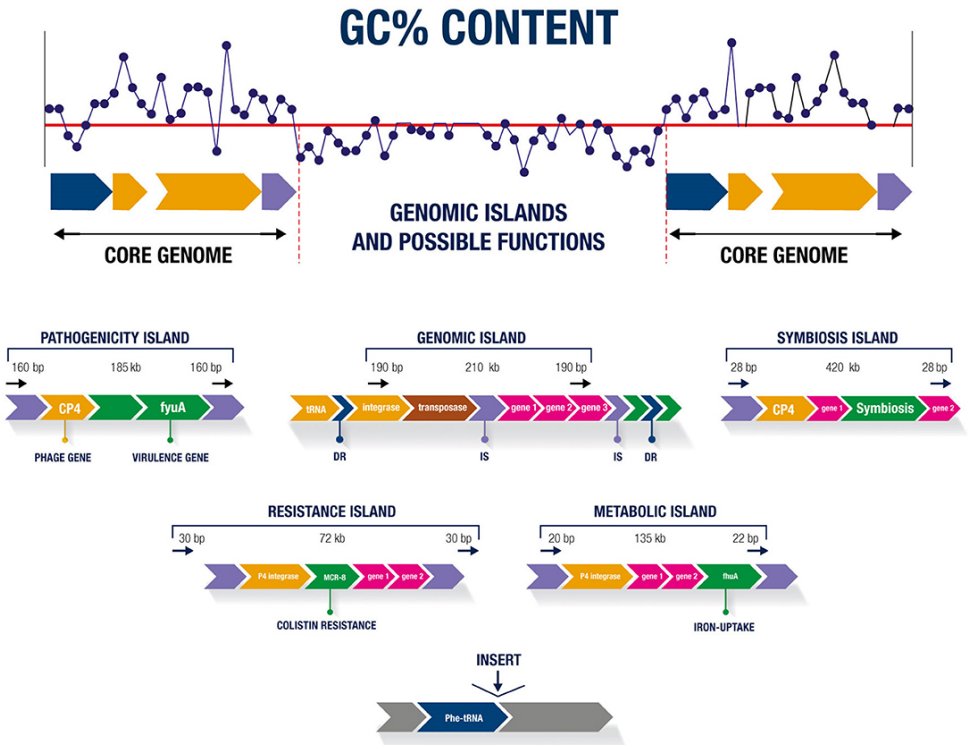


Figure I.2.11 – **Îlots génomiques et leur caractéristique.** Extrait de (da Silva Filho *et al.*, 2018)

15. Un biais d'usage des codons, désigne la fréquence d'utilisation préférentielle de certains codons parmi les codons synonymes pour coder un même acide aminé.

16. Séquences identiques présentes en plusieurs copies dans la même molécule d'ADN et ayant la même orientation.

Ces GIs sont complexes à étudier, car ils concentrent les variations, même entre génomes proches. L'histoire évolutive est souvent difficile à reconstituer, tant des éléments ont été intégrés et éliminés au cours du temps (figure I.2.12). En plus de s'échanger avec d'autres organismes (Buchrieser *et al.*, 1998), les GIs peuvent se déplacer au sein du génome (Karaolis *et al.*, 1999).

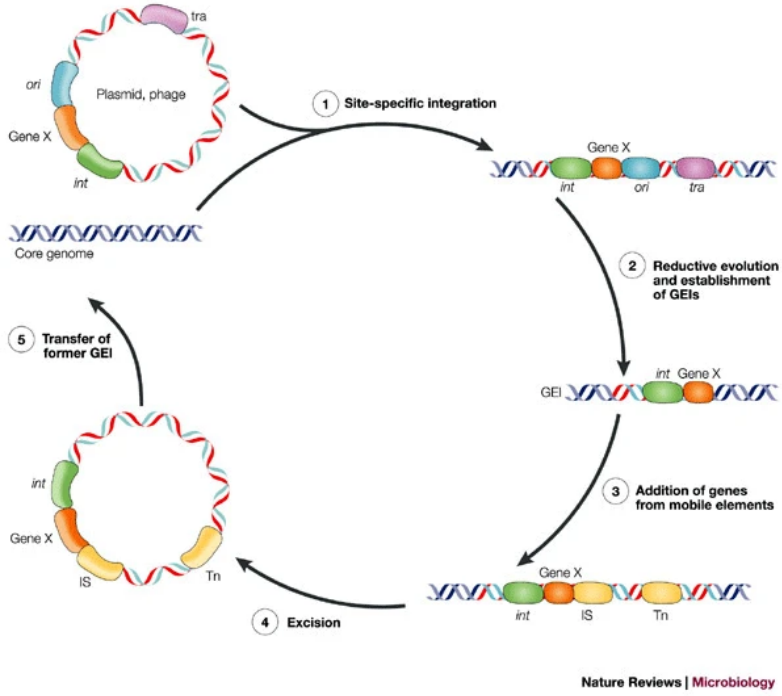


Figure I.2.12 – Cycle de vie d'un îlot génomique. Extrait de (Dobrindt *et al.*, 2004)

Les GIs ne s'insèrent pas n'importe où dans les génomes. On les retrouve fréquemment dans des zones où de nombreux éléments se sont insérés au cours de l'évolution d'un taxon. Ces régions sont appelées : point chaud d'insertion (*hotspot* en anglais). À l'intérieur des *hotspots*, on retrouve une grande variabilité du contenu génique entre les génomes. Les *hotspots* sont également caractérisés par des bordures composées de gènes communs à l'ensemble des génomes.

Ils présentent également une recombinaison homologue accrue dans les gènes flanquant les *hotspots*, avec 50 % d'événements de recombinaison et 30 % d'incongruence phylogénétique [17] par rapport à l'arbre des espèces. Ces *hotspots* contiennent 50 % des gènes acquis par HGT (Oliveira *et al.*, 2017). Ils sont enrichis en gènes liés à la motilité, à la défense, à la transcription, à la réplication et à la réparation de l'ADN (Flores Ramos *et al.*, 2021).

Le contenu génique du *hotspot* provient d'une accumulation progressive de gènes, comme le suggère le faible pourcentage (8 %) de *hotspots* composés uniquement de gènes spécifiques à une souche (Oliveira *et al.*, 2017). Cette accumulation peut se faire par bloc de gènes. Ces blocs, conservés dans le *hotspot*, sont appelés modules (Lescat *et al.*, 2009). Cette modularité pourrait expliquer l'organisation complexe des îlots génomiques (Touchon *et al.*, 2009). Les *hotspots* sont donc communs à un groupe d'organismes, et définis au niveau d'un taxon. Il est donc nécessaire de mener des études de comparaison des génomes pour les identifier.

17. L'incongruence phylogénétique désigne une discordance entre l'arbre phylogénétique d'un gène spécifique et l'arbre phylogénétique global construit à partir d'un grand nombre de gènes conservés.

3 - Génomique comparée des procaryotes

L'analyse comparée des génomes regroupe une grande diversité d'analyses et de thématiques. On peut diviser ces analyses en 3 grands domaines : l'analyse des séquences, l'analyse des structures et l'analyse fonctionnelle. Avec la croissance exponentielle du nombre de génomes disponibles dans les banques (cf. section 3.3), et des informations de structures et d'annotations liées, il est essentiel de comparer les nouvelles séquences à celles déjà connues. Cette comparaison permettra de déduire, entre autres, les fonctions qu'elles contiennent et leur lien évolutif.

Dans cette partie, j'aborderai les concepts informatiques et les algorithmes utilisés dans les méthodes de génomique comparée. Je reviendrai aussi sur les notions présentées précédemment et comment elles sont appliquées dans les outils. Pour terminer, je présenterai un type d'analyse de génomique comparée qui est largement présent dans mon travail de thèse, l'analyse de systèmes biologiques.

3.1 . Analyse comparative des génomes : méthodes et applications

Pour comparer les séquences, on mesure leur similarité. Si la similarité des séquences est significativement élevée, alors on peut faire l'hypothèse que les séquences sont homologues [1]. Ce principe de base simple se révèle être un problème non trivial étant donné l'ensemble des mécanismes gouvernant l'évolution des génomes procaryotes. Pour comparer les génomes sur la similarité des séquences, on peut utiliser les séquences nucléotidiques, mais aussi en fonction du contexte, les séquences d'ARN ou de protéines. Pour l'ADN et l'ARN, la similarité va se confondre avec la notion d'identité ; par contre, pour les séquences d'acides aminés, ces termes n'ont pas le même sens. Lorsqu'on mesure l'identité de séquences entre 2 protéines, on mesure le pourcentage de résidus identiques entre 2 séquences alignées. Pour la similarité, si les 2 acides aminés ont les mêmes propriétés physico-chimiques, ils seront considérés comme similaires. La mesure d'identité entre 2 séquences d'acides aminés sera toujours inférieure ou égale à celle de sa similarité.

Dans la suite, je ferai une revue (non exhaustive) des méthodes et des outils de génomique comparée. L'évolution de ces outils est intimement liée à l'évolution des techniques de séquençage, augmentant le volume de génomes disponibles, et à l'amélioration des technologies informatiques, augmentant les ressources disponibles et leur utilisation.

1. *N.B* : L'homologie est une conclusion qualitative de l'observation quantitative de la similarité. On considère qu'une similarité de 30 % permet de dire que 2 séquence sont homologues.

3.1.1 . Alignement des séquences

L'alignement des séquences peut être : pair ou multiple. Dans les deux cas, l'objectif est de trouver l'alignement qui maximise la correspondance entre les résidus. Pour cela, dans la majorité des cas, les séquences ne seront pas alignées entre leur début et leur fin, elles seront décalées. Il existe alors 2 stratégies pour l'alignement : global et local. Dans un alignement global, on fait l'hypothèse que les séquences sont relativement similaires et donc on peut les aligner sur toute leur longueur. Pour un alignement local, on ne fait pas cette hypothèse et on cherche les régions dans la séquence qui ont le plus de similarité sans considérer la séquence dans sa globalité. Les algorithmes et outils que je vais présenter ensuite peuvent généralement s'appliquer aux 2 types de stratégies, qu'on choisit en fonction du contexte.

a. Alignement par paire

Dès les années 1960, on commence à voir des développements autour de l'idée de comparer 2 séquences (protéiques), mais c'est en 1970 que Needleman et Wunch présentent leur algorithme fondateur des approches de génomique comparée (Needleman et Wunsch, 1970). Leur algorithme d'alignement global repose sur la construction d'une matrice de similarité, représentant en ligne une séquence et en colonne la seconde, et inclut une pénalité de trou (*gap* en anglais). Ainsi, il est possible de déterminer l'alignement optimal en considérant tous les *gap* sans énumérer toutes les possibilités. Cet algorithme sera revu par Smith et Waterman qui, en 1981, proposent un nouvel algorithme, cette fois pour l'alignement local (Smith et Waterman, 1981). Ces 2 algorithmes ont l'intérêt de donner un résultat de comparaison exact et sont d'ailleurs encore utilisés aujourd'hui. Toutefois, avec l'augmentation du volume de séquences, la comparaison de paires de séquences utilisant des algorithmes exhaustifs [2] pose un problème de complexité quadratique [3].

En 1985 et 1988, les programmes FASTP et FASTA [4] (Lipman et Pearson, 1985; Pearson et Lipman, 1988) sont publiés et marqueront un tournant en utilisant une approche heuristique [5]. Le principe est de chercher quelles séquences peuvent être similaires en comparant des mots de taille k (*k-mer*), pour ensuite ne faire l'alignement exact que sur ce sous-ensemble de séquences. Dans la suite, en 1990, le programme BLAST (Altschul *et al.*, 1990) paraît et suit aussi cette approche heuristique. Il sera intégré comme outil dans les bases de données du NCBI, faisant sa renommée.

Toujours lié à l'augmentation du volume de données, les outils utilisant ces approches heuristiques vont se perfectionner pour permettre l'alignement de paires de séquences de manière rapide et efficace, comme LAST (Kiełbasa *et al.*, 2011) ou DIAMOND (Buchfink *et al.*, 2015).

2. Un algorithme exhaustif recherche toutes les solutions possibles pour trouver celle qui exacte ou optimale

3. La complexité d'un algorithme mesure la consommation de ressources (temps ou espace) nécessaire pour son exécution.

4. Le format de données de l'outil FASTA est aujourd'hui utilisé comme format standard pour écrire les séquences. Les fichiers ont donc pris l'extension ".fasta".

5. Un algorithme heuristique fournit un résultat rapidement, mais qui n'est pas nécessairement optimal ou exact.

b. Alignement multiple

L'alignement multiple des séquences (MSA pour *Multiple Sequence Alignment* en anglais) vise à aligner plusieurs séquences simultanément. C'est une extension de l'alignement en paire. Ces alignements ont l'intérêt de révéler des régions conservées et ainsi d'identifier des relations évolutives ; par contre, la complexité est accrue et il est donc nécessaire d'introduire des algorithmes plus puissants.

Les premiers algorithmes étaient des algorithmes exhaustifs (Stoye, 1998), et tout comme pour l'alignement de paires de séquences, rapidement des algorithmes heuristiques ont été publiés. En 1988, Higgins et Sharp publient CLUSTAL (Higgins et Sharp, 1988), une méthode d'alignement progressive pour obtenir un alignement multiple. Elle construit l'alignement en assemblant progressivement les séquences selon une hiérarchie basée sur une matrice de distance ou un arbre guide. D'autres méthodes adopteront cette approche, comme MUSCLE (Edgar, 2004) mais ce dernier apporte un côté itératif. Ces méthodes sont rapides, mais peuvent être sensibles aux erreurs accumulées dans les étapes initiales.

D'autres méthodes, que je décrirai ensuite, s'appuient sur des éléments de statistique ou sur des algorithmes de graphes pour être plus efficaces. Il est à noter que toutes ces méthodes ont leur avantage et leurs inconvénients, qui doivent être évalués en fonction du contexte.

3.1.2 . Utilisation des graphes en génomique comparée

Les graphes sont largement utilisés en bioinformatique (Pavlopoulos *et al.*, 2011) et ce dans des domaines très divers : interactions protéine-protéine, expression des gènes, modélisation du métabolisme…

Dans mes travaux de thèse, je me suis largement appuyé sur les méthodes de graphes, il est donc essentiel de revenir sur la terminologie et les concepts liés à la théorie des graphes. Nous utiliserons le graphe de la figure I.3.1 pour illustrer les principes suivants.

3.1.2.1 . Définitions et concepts

Un graphe est constitué d'un ensemble de **nœuds** (cercles) reliés par un ensemble d'**arêtes** (segments gris). Mathématiquement, tous les graphes ne possèdent pas les mêmes propriétés et donc les théorèmes associés changent. Dans la suite, nous utiliserons les symboles mathématiques suivants :

- V : ensemble de nœuds
- E : ensemble d'arêtes
- $G(V, E)$: un graphe composé d'un ensemble de nœuds V et d'arête E
- u et v : 2 nœuds distincts dans le graphe
- $e_{(u,v)}$: une arête reliant u et v.

a. Orientation du graphe

Un graphe peut être **orienté**, *i.e.*, que les arêtes ont une direction. Dans ce cas, il peut exister une arête de u vers v ($e_{(u,v)}$) sans qu'il n'y ait nécessairement une arête $e_{(v,u)}$. Si le graphe est non orienté, si $e_{(u,v)}$ existe, $e_{(v,u)}$ également. Dans notre exemple, le graphe est non orienté.

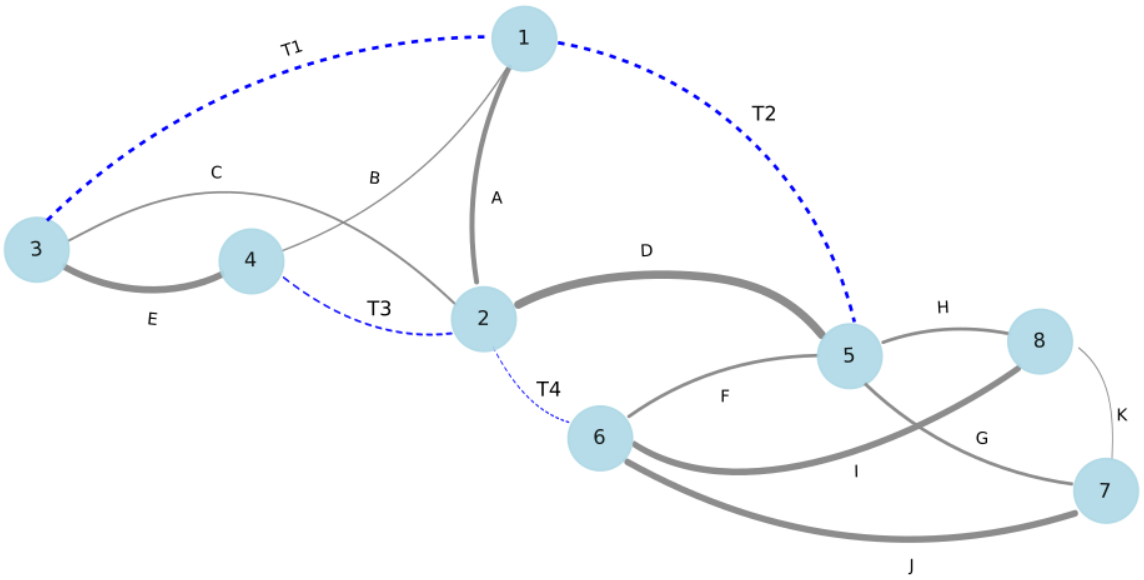


Figure I.3.1 – **Exemple de graphe.** Les nœuds sont représentés par des cercles et étiquetés par des numéros. Les arêtes, illustrées par des lignes pleines grises et étiquetées par une lettre, définissent les connexions entre les nœuds. L'épaisseur des arêtes est proportionnelle à leur poids, indiquant ainsi la valeur associée à chaque connexion. Les arêtes en pointillés bleus représentent les fermetures transitives du graphe, elles sont également étiquetées et pondérées pour expliciter les relations indirectes créées par la transitivité.

b. graphe pondéré et étiqueté

En bioinformatique, il est courant d'ajouter de l'information sur le graphe. Ces informations peuvent servir à modifier le graphe, le filtrer ou l'analyser, par exemple.

On peut ajouter un **poids** aux nœuds (w_u) et aux arêtes ($w_{(u,v)}$), le graphe est alors dit **pondéré**. Le poids est quantifiable et correspond généralement à un nombre. Dans notre exemple, chaque arête a une épaisseur correspondant à son poids. On peut alors filtrer le graphe pour ne conserver que les arêtes les plus épaisses.

D'autres informations peuvent être ajoutées aux nœuds et aux arêtes sous forme d'annotation. Dans ce cas, l'annotation peut être qualitative et on dira que le graphe est **étiqueté** [6]. Dans le graphe exemple, les arêtes sont étiquetées par une lettre et les nœuds par un chiffre. Cet étiquetage peut notamment correspondre à un identifiant.

c. Voisinage et chemin dans le graphe

Dans cette thèse, nous parlerons de nœuds **voisins**, *i.e.*, des nœuds qui sont reliés par un ensemble d'arêtes ($E_{(u,v)}$), cet ensemble d'arêtes est appelé **chemin**. Lorsque u et v sont reliés par une seule arête (chemin de taille 1), on dit qu'ils sont dans un **voisinage direct**. Dans notre exemple, le nœud 1 est un voisin direct des nœuds 2 et 4 et est voisin des nœuds 3 et 5 par un chemin de taille 2. Lorsque tous les nœuds sont voisins les uns des autres, on dit que le graphe est **connexe**.

6. Dans la littérature bioinformatique, on retrouve aussi le terme "coloré", mais qui est utilisé à tort si on se réfère à la théorie des graphes.

d. Transitivité

La **transitivité** dans les graphes est une propriété qui s'applique aux relations entre les nœuds. Un graphe est dit **transitif** si, pour tous nœuds u, v et w, l'existence des arêtes $e_{u,v}$ et $e_{v,w}$, implique qu'il existe $e_{u,w}$. En d'autres termes, si u et v sont reliés et que v et w aussi, alors u et w sont reliés. Dans un graphe orienté transitif, s'il existe $e_{u,v}$ et $e_{v,w}$, alors il existe $e_{u,w}$, mais pas obligatoirement $e_{w,u}$. Cette propriété est particulièrement importante dans les graphes orientés, où elle peut être utilisée pour modéliser des relations hiérarchiques ou des dépendances.

Dans notre exemple, le graphe n'est pas transitif. Pour le rendre transitif, on ajoute des **fermetures transitives** entre les nœuds. Ces arêtes de transitivité (en pointillés bleus) permettent de compléter le graphe pour le rendre transitif, facilitant ainsi l'analyse des relations et des dépendances implicites entre les nœuds.

e. Sous-ensemble du graphe

Lorsqu'on va analyser un graphe, on peut chercher à retrouver des structures d'intérêt. Pour commencer, on peut chercher à identifier un **sous-graphe**. Le sous-graphe est une fraction du graphe qui contient un sous-ensemble de nœuds de $G(V, E)$ et les arêtes reliant ces nœuds.

Une autre structure est la **clique**, qui correspond à un sous-ensemble de nœuds tous connectés entre eux. La détection et l'analyse de cliques a de nombreuses applications en bioinformatique, notamment l'identification de groupes de gènes coexprimés. Dans notre exemple, les nœuds 5,6,7 et 8 représentent une clique.

Pour terminer, une forme de sous-ensemble que j'ai largement utilisée, est la **composante connexe** qui correspond à un ensemble de nœuds tel que, quel que soit u, v, il existe un chemin qui les relie [7].

f. Partitionnement du graphe

Partitionner un graphe consiste à diviser les nœuds du graphe en groupes. Chacun de ces groupes est appelé une **partie** et l'ensemble des parties est appelé **partition**. En fonction de l'algorithme utilisé, la partition sera alors différente. Dans ce manuscrit, nous utiliserons cette notion de partition à de nombreuses reprises.

3.1.2.2 . Application dans la comparaison des génomes

L'utilisation des graphes pour la comparaison de génomes est de plus en plus courante.

Une première application possible est d'améliorer les méthodes de MSA. Des outils comme MUSCLE ou MAFFT (Katoh et Standley, 2013) utilisent des arbres guides pour améliorer les performances de l'alignement. Ces arbres sont des graphes particuliers, où les séquences sont des nœuds et les relations de similarité sont des arêtes.

7. La clique est une composante connexe spéciale où tous les nœuds sont reliés par un chemin de taille 1.

Une seconde utilisation des graphes concerne l'étude des SNPs, indels et SVs. Ces graphes, appelés graphes de variants, représentent d'une manière flexible les différences entre les génomes. Chaque nœud représente une séquence ou un k-mer, les arêtes vont représenter la colocalisation dans le génome. Ainsi, chaque chemin permet de reconstruire un génome, tout en ayant toutes les variations génétiques. Des outils comme VG toolkit (Garrison *et al.*, 2018) et Minigraph (Li *et al.*, 2020), permettent notamment d'améliorer l'alignement des lectures en sortie de séquençage, mais aussi d'enrichir la représentation des génomes procaryotes présentant une forte diversité.

Une autre application proche des graphes de variants est celle des graphes de réarrangements. Dans ces graphes, les nœuds représentent des synténies conservées, les arêtes vont relier ces synténies en fonction de l'ordre et de l'orientation dans les génomes. L'outil Sibelia (Minkin *et al.*, 2013) est un outil d'alignement et d'analyse des réarrangements de génomes procaryotes. Il permet d'étudier les différences évolutives et de reconstruire l'histoire des réplicons.

3.1.3 . Modèle statistique pour l'alignement des séquences

Pour comparer de plus en plus de génomes, de nouvelles méthodes vont s'appuyer sur la modélisation statistique des séquences pour améliorer les performances de l'alignement et aligner de grands jeux de données. Chaque modèle représentera un ensemble de séquences et établit une fréquence ou probabilité d'un résidu pour chaque position. Ce modèle peut être assimilé à une séquence "consensus" du groupe.

3.1.3.1 . Partitionnement des séquences par similarité.

Les méthodes de partitionnement, ou *clustering* en anglais, reposent sur les méthodes d'alignement pour déterminer la similarité des séquences, et les graphes pour représenter les liens de similarité entre chaque séquence.

De manière générale, on va regrouper les séquences en groupes d'homologues en utilisant un seuil de similarité plus ou moins élevé. Les outils sont régulièrement présentés en utilisant la séquence protéique plutôt que nucléique pour calculer la similarité. Ce choix permet de réduire la complexité tout en étant plus précis sur l'évaluation de la similarité fonctionnelle et structurelle, mais aussi d'identifier des homologies plus lointaines. Dans ce cas, il faudra faire attention à la nuance entre similarité et identité. Ce qui va varier entre les méthodes, c'est l'algorithme de partitionnement utilisé. Le tableau I.3.1 présente un aperçu des méthodes et des outils existants.

Outil	Description	Avantages	Inconvénients
COGs (Tatusov *et al.*, 1997)	Classification basée sur l'évolution. Les clusters obtenus sont des groupes de protéines orthologues	Base de données bien documentées et largement utilisée	Méthode statique, mise à jour peu fréquente.
CD-HIT (Li *et al.*, 2001)	Clustering rapide basé sur la longueur des séquences, en ordonnant les protéines de la plus longue à la plus courte	Très rapide et efficace pour réduire la redondance	Sensibilité limitée pour les faibles identités de séquence
InParanoid (Remm *et al.*, 2001)	Détection des orthologues et paralogues en comparant deux génomes	Fiable pour détecter des orthologues proches, distingue bien orthologues et paralogues	Moins adapté aux comparaisons multi-génomes
OrthoMCL (Li *et al.*, 2003)	Identification des orthologues et paralogues récents via une approche basée sur les graphes	Bonne précision, adaptable à divers organismes	Consommation élevée en ressources pour les grands ensembles de données
UCLUST (Edgar, 2010)	Alignement et clustering rapide des séquences protéiques	Très rapide, faible consommation mémoire	Moins précis que BLAST pour certaines comparaisons
Proteinortho (Lechner *et al.*, 2011)	Détection rapide d'orthologues à grande échelle	Évolutif et performant pour l'analyse de nombreux génomes	Moins détaillé sur les relations fonctionnelles des protéines
BUSCO (Simão *et al.*, 2015)	Évaluation de la complétude des génomes en utilisant des ensembles de gènes conservés	Référence fiable pour les génomes récemment séquencés	Ne permet pas une recherche extensive d'orthologues
OMA (Altenhoff *et al.*, 2019)	Méthode évolutive d'identification des orthologues	Haute précision sur les génomes bien annotés	Temps de calcul élevé pour les grands jeux de données
SwiftOrtho (Hu et Friedberg, 2019)	Similaire à OrthoMCL, utilisant des k-mers longs pour améliorer la rapidité	Très rapide et nécessite peu de ressources	Peut être moins précis pour des génomes très divergents
SonicParanoid (Cosentino et Iwasaki, 2019)	Extension de InParanoid pour la détection rapide d'orthologues	Améliore la vitesse et réduit les besoins en ressources	Moins précis pour des génomes très distants évolutivement
OrthoFinder (Emms et Kelly, 2019)	Détection des orthologues basée sur une approche évolutive	Haute précision grâce à l'utilisation de scores de similarité normalisés	Temps de calcul élevé pour les grands ensembles de données
OrthoPhy (Watanabe *et al.*, 2023)	Intègre les informations taxonomiques dans l'identification des orthologues	Minimise les erreurs de prédiction et améliore la fiabilité	Exige davantage de ressources et est plus lent sur de grandes bases de données

Table I.3.1 – **Présentation des principaux outils de clustering de séquences avec leurs descriptions, avantages et inconvénients.**

3.1.3.2 . MMSeqs2

Un outil largement utilisé pour l'alignement et le *clustering* de grands jeux de données est l'outil MMSeqs2 (Steinegger et Söding, 2017). L'objectif de MMSeqs2 est de partitionner les séquences en groupes d'homologues, de manière rapide et efficace. MMSeqs2 s'appuie sur les technologies informatiques, tant matérielles que logicielles, pour optimiser les ressources utilisées, et sur un nouvel algorithme de recherche de k-mer similaire.

MMSeqs2 ne va pas faire des comparaisons exactes de k-mers, mais il va chercher des k-mers similaires. Cette différence permet de comparer les k-mers plus rapidement tout en utilisant des k-mers de plus grandes tailles, améliorant sa sensibilité [8]. Comme présenté sur la figure I.3.2, les k-mers utilisés sont "espacés", ce qui permet un recouvrement plus important de la séquence et donc de réduire les alignements liés au hasard de k-mers successifs entre 2 séquences non homologues. S'appuyant sur cette caractéristique, les auteurs de MMSeqs2 vont supposer que si les séquences ont des k-mers similaires, séparés par le même nombre de résidus, alors la zone entre les k-mers a des chances de s'aligner, ce qui permet d'étendre les zones alignables (diagonale). Enfin, un score associé aux diagonales va être utilisé pour filtrer les séquences qui ont le plus de probabilités de s'aligner.

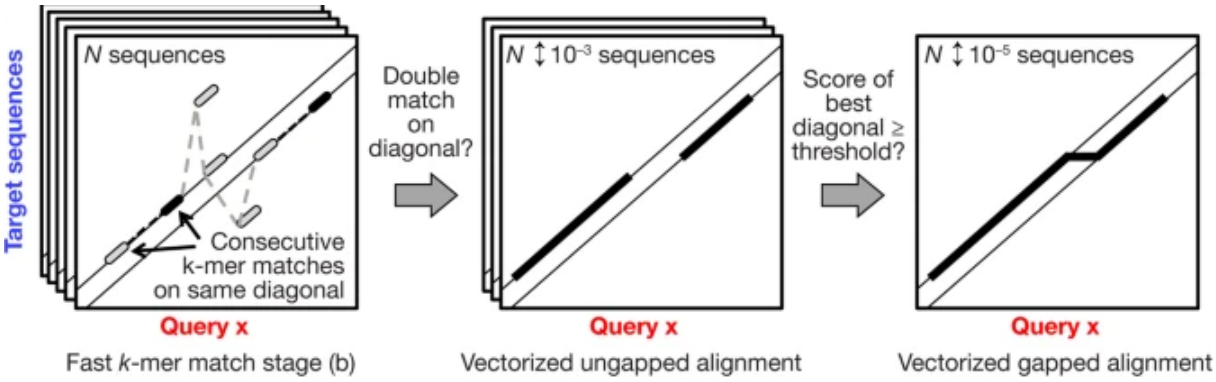
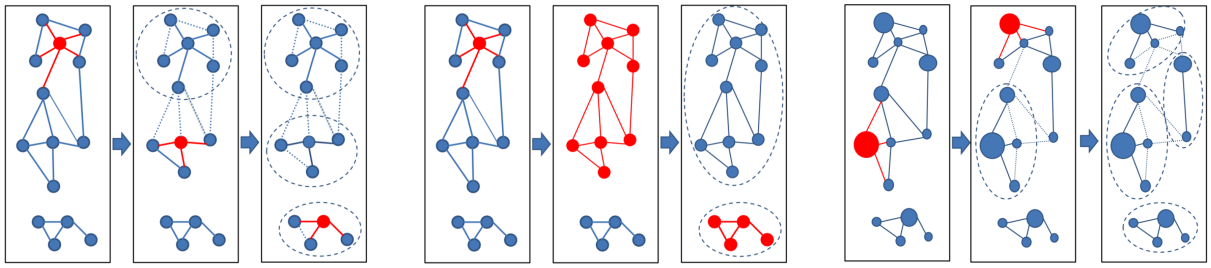


Figure I.3.2 – **Principe de l'alignement de MMSeqs2.** Extrait de (Steinegger et Söding, 2017)

Une fois l'étape d'alignement terminée, MMSeqs2 intègre plusieurs algorithmes pour partitionner les séquences. Dans chacun de ces algorithmes, chaque groupe de séquences similaires (partie) verra une des séquences (nœud) utilisée comme référente. (*i*) L'algorithme Set-cover (figure I.3.3a) sélectionne le nœud avec le plus d'arêtes comme référent et forme une partie avec tous les nœuds dans un voisinage direct, puis de manière itérative reproduit le schéma jusqu'à ce que tous les nœuds soient dans une partie. (*ii*) L'algorithme *Connected Component* (figure I.3.3b) fonctionne comme Set-cover, mais partitionne tous les nœuds pour lesquels il existe un chemin avec le nœud référent. (*iii*) l'algorithme *CD-hit like* (figure I.3.3c) prend pour référence le nœud dont le poids (taille de la séquence) est le plus élevé, puis forme une partie avec tous les voisins directs. Ces algorithmes répondent chacun à des problématiques différentes que nous pourrons illustrer dans la suite.

8. Dans MMSeqs2 la sensibilité correspond au nombre de k-mers similaires qui sont alignées par position dans la séquence. Plus la sensibilité est élevée, plus le k-mer associé à une position sera précis.



(a) Set-cover (b) Connected component (c) CD-hit like

Figure I.3.3 – **Les algorithmes de clustering de MMSeqs2.**

Depuis 2018, MMSeqs2 intègre une nouvelle méthode appelée Linclust (Steinegger et Söding, 2018). Celle-ci vise à proposer un clustering dont la durée évolue linéairement avec le nombre de séquences. Pour ce faire, les séquences ne sont pas alignées entre elles. Dans un graphe, chaque séquence forme un nœud et est représentée par des k-mers, eux-mêmes répartis en groupes. La séquence la plus longue de chaque groupe est comparée à toutes les autres. Si l'alignement dépasse un seuil prédéfini, une arête de similarité est établie entre les séquences. Ensuite, un algorithme de partitionnement est appliqué au graphe résultant pour obtenir le partitionnement final. Cette optimisation permet de partitionner rapidement de grands jeux de données.

3.1.3.3 . Modélisation des séquences similaires : matrices de position, profils et chaînes de Markov

Une fois les séquences regroupées par similarité, il est possible de créer un modèle statistique représentant les séquences, sous forme de "séquence" consensus. L'idée générale de ces modèles va être, pour chaque position de la séquence consensus, d'associer pour chaque type de résidus une fréquence ou probabilité d'apparition, basée sur un alignement multiple des séquences.

Les premiers modèles statistiques correspondaient à des matrices de score à position spécifique (PSSM, position-specific scoring matrices), représentant la probabilité du résidu à une position donnée. Ainsi, la matrice obtenue reflète pour un score positif une correspondance de résidus similaires parmi les séquences, ou pour un score négatif un résidu non conservé. Ces matrices ont été utilisées dans des outils comme CLUSTAL (Higgins et Sharp, 1988) ou MATCH^{TM} (Kel *et al.*, 2003) pour la recherche de facteurs de transcription dans les séquences d'ADN, ou encore dans l'algorithme ESAsearch (Beckstette *et al.*, 2006) pour rechercher des séquences dans les PSSMs. Ces outils vont également amener une variante aux PSSMs qui comble un défaut de ces dernières. En effet, le score dépend du nombre et de la divergence des séquences utilisées dans le MSA. Si la matrice est constituée de peu de séquences ou si des séquences proches sont surreprésentées, alors le score sera biaisé. C'est pourquoi un poids est appliqué pour réduire l'impact des séquences proches et augmenter celui des séquences divergentes.

Pour construire une PSSM, les MSA doivent être continus (sans *gap*), ce qui est rarement le cas. Une nouvelle forme de PSSM, appelée profil, va prendre en compte les *gap* en appliquant des pénalités. Un profil est donc une PSSM intégrant les possibles indels sous forme de pénalités [9]. Les profils sont utilisés, notamment dans le contexte des bases de données, pour rechercher des séquences homologues à un

9. Dans la littérature, les PSSM sont souvent également appelées profils.

groupe de séquences sans aligner chacune des séquences du groupe. PSI-BLAST (Altschul *et al.*, 1997), développé par les auteurs de BLAST, permet de construire des profils et de rechercher des séquences contre un profil. Bien que PSI-BLAST soit reconnu pour sa haute précision, il demeure sensible aux erreurs d'assignation initiales, lesquelles peuvent introduire des biais dans les profils générés et impacter la fiabilité des cycles et des itérations suivants.

Une dernière forme de modèle s'appuie sur les chaînes de Markov cachées (*Hidden Markov Model* en anglais, HMMs). Une chaîne de Markov décrit la probabilité de transition vers un état en fonction des états précédents. Dans nos modèles, cela correspondrait à calculer la probabilité d'un résidu (état) à une position donnée en fonction des résidus des positions précédentes. Une chaîne de Markov cachée inclut, en plus, l'existence de facteurs non observables sur la probabilité de transition. Dans nos modèles, ces facteurs cachés peuvent être les *gaps* qui ne correspondent à aucun résidu (état) mais influencent la probabilité de transition (figure I.3.4). On peut alors obtenir une probabilité pour chaque résidu à chaque position.

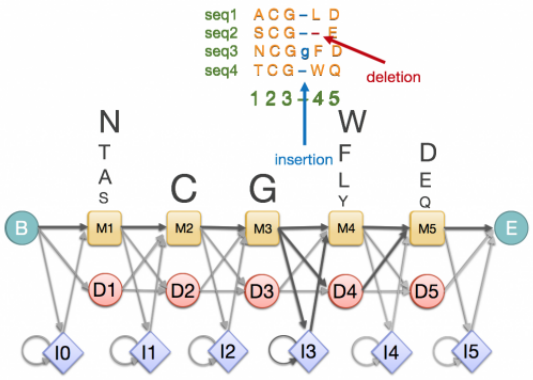


Figure I.3.4 – **Exemple de modélisation HMM d'une séquence.** Les cases jaunes représentent les états de correspondance (M), où la distribution de probabilité est déterminée par la fréquence des acides aminés à cette position. La rangée d'états en forme de losange correspond aux états d'insertion (I) et les états circulaires désignent les états de suppression (D). Ce modèle probabiliste permet d'estimer les fréquences observées des acides aminés à chaque position et de représenter les transitions entre eux, sur la base de l'occupation observée des positions dans un alignement de séquences multiples. Extrait de `https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/what-are-profile-hidden-markov-models-hmms/`

Les modèles HMMs semblent donc tout indiqués pour représenter l'alignement des séquences similaires. Les HMMs, ont l'intérêt de pouvoir différencier les événements d'insertion des événements de délétion par rapport aux profils. Cet avantage les rend plus robustes que les profils.

Un outil largement utilisé pour construire des HMMs et rechercher des séquences homologues contre une base de données HMMs est HMMER (`http://hmmer.org/`). Un autre outil, HH-suite (Steinegger *et al.*, 2019) intègre la possibilité de faire des comparaisons HMM/HMM. Ces outils, dans leurs versions récentes, ont été optimisés pour combler la complexité sous-jacente de l'utilisation de tels modèles. D'autres outils récents, comme ApHMM (Firtina *et al.*, 2024) ou le package pyhmmer (Larralde et Zeller, 2023), proposent des améliorations techniques pour augmenter l'efficacité et la sensibilité des comparaisons aux HMMs, notamment pour ApHMM en s'appuyant sur les nouvelles technologies matérielles et logicielles, et en optimisant les calculs opérés par les algorithmes.

Les modèles HMMs sont couramment utilisés pour la recherche de séquences homologues dans les bases de données. Ils trouvent également des applications dans d'autres domaines, tels que la classification et l'annotation des protéines, ainsi que la prédiction de gènes et de promoteurs (Dimri *et al.*, 2024).

3.2 . Analyse des Systèmes biologiques

La notion de système biologique est vaste et dépend du domaine et du contexte scientifique. Dans cette section, je vais définir les systèmes dans le cadre de la génomique comparée des procaryotes, en lien avec les processus métaboliques et cellulaires. J'aborderai l'état de l'art des méthodes bioinformatiques utilisées pour identifier les systèmes et je terminerai en décrivant un type particulier de système biologique : les systèmes de défense contre les phages, qui ont été au cœur de mes développements méthodologiques.

3.2.1 . Définition et intérêt

Un système biologique est constitué d'un ensemble de protéines interagissant pour réaliser un processus spécifique. Ces processus sont souvent régulés au sein d'opérons ou de groupes de gènes colocalisés. Les systèmes sont classés et nommés en fonction de leur rôle, comme ceux impliqués dans la conjugaison, regroupés sous l'appellation de **système conjugatif**.

La description et l'étude de ces systèmes est essentielle, car une fois caractérisés, ils permettent de comprendre les capacités métaboliques et les capacités d'adaptation des organismes (Alberts, 1998). De plus, certains systèmes sont associés à des îlots génomiques, comme les systèmes de sécrétion de type III et VI associés aux îlots de pathogénicités (Pallen et Wren, 2007). Leur identification dans les GIs est essentielle à la compréhension de l'adaptation et de la diversité des écosystèmes procaryotes.

Les systèmes biologiques présentent une grande diversité de composition et d'organisation. Premièrement, certains gènes peuvent être facultatifs ou spécifiques à certaines niches écologiques. Par exemple, la réparation de l'ADN repose sur RecA, une protéine clé de la recombinaison homologue, mais peut aussi emprunter des voies alternatives, comme les systèmes RecBCD chez *Escherichia coli* ou AddAB chez *Helicobacter pylori*[10] (Dillingham et Kowalczykowski, 2008). Un autre exemple concerne le système de sécrétion T2SS (Korotkov *et al.*, 2012), présent dans un grand nombre de bactéries Gram-négatives pathogènes et non pathogènes. Il est composé de 4 protéines essentielles à son fonctionnement : gspD, gspE, gspF et gspG, mais peut aussi être trouvé dans les organismes avec des protéines supplémentaires facultatives : gspC, gspH, gspI, gspJ, gspK, gspL, gspM et gspN. Ces protéines facultatives peuvent être absentes dans certains taxons, comme chez les Chlamydiae pour T2SS (Abby *et al.*, 2016). Ensuite, des gènes peuvent avoir des homologues avec d'autres systèmes, parfois même très différents, rendant leur classification complexe. C'est notamment le cas du système de sécrétion de type VI, qui présente des similitudes structurelles avec les phages à queue contractile, suggérant une origine évolutive commune (Coulthurst, 2013). La dynamique évolutive des systèmes est également hétérogène. Certains composants sont fortement conservés, tandis que d'autres évoluent rapidement sous l'effet de pression de

10. Bactérie pathogène connue pour son rôle dans les infections gastriques et notamment dans les ulcères de l'estomac.

sélection. C'est le cas des systèmes de défense (cf. sous-section 3.2.3), tels que les systèmes CRISPR-Cas, dont la diversité des protéines Cas (permettent de découper l'ADN viral) reflète une adaptation continue contre les virus (Makarova *et al.*, 2013). Cette variabilité complique alors l'identification des homologues par la seule comparaison de séquences.

3.2.2 . Méthodes de détection

l'identification de systèmes repose sur la combinaison de la recherche des gènes et sur leur organisation en contexte. En s'appuyant sur ces propriétés, on peut alors identifier des systèmes connus ou proches chez les organismes.

Avant les années 2000, la recherche de systèmes biologiques était basée sur des approches phylogénétiques, en recherchant des homologues, ou par de l'annotation manuelle de régions d'intérêt comme les GIs (Buchrieser *et al.*, 1998). Des outils ont ensuite été développés pour détecter différents systèmes : les systèmes conjugatifs, les systèmes de sécrétion, les systèmes de défenses contre les phages (sous-section 3.2.3) et les systèmes métaboliques. Leur évolution a suivi une trajectoire marquée par des avancées méthodologiques, passant de simples bases de données statiques à des modèles probabilistes et des approches d'intelligence artificielle.

a. Systèmes de sécrétion et de conjugaison

Les premiers outils pour la recherche de systèmes spécifiques, tels que les systèmes de sécrétion, annotaient fonctionnellement les génomes pour identifier les gènes codant des fonctions connues des systèmes. En 2008, l'outil ICEberg (Bi *et al.*, 2012) a été développé pour identifier les ICEs (cf. sous-section 2.2.2) à partir d'une base de données de protéines annotées manuellement et d'un alignement avec BLASTp sur les génomes. Bien que régulièrement mise à jour (Wang *et al.*, 2024b), cette base de données n'est pas adaptée à la détection des ICEs divergents, limitant ainsi son utilisation à des systèmes bien caractérisés. Pour pallier ces limites, l'outil ICEscreen (Lao *et al.*, 2022) a été développé afin de détecter les ICEs et IMEs (*Integrative and Mobilizable Elements*) des Firmicutes [11], qu'ils soient isolés ou intégrés dans des structures composites. Contrairement aux approches basées uniquement sur des bases de données de séquences connues, ICEscreen utilise des stratégies de détection basées sur des signatures génétiques et des relations de synténie, permettant ainsi d'identifier des éléments plus divergents ou moins bien caractérisés.

SecReT4 (Bi *et al.*, 2013) a proposé une base de données spécifique pour la détection des systèmes de type IV (T4SS). Bien que cette approche soit utile, elle est limitée par la nécessité de mises à jour régulières et une couverture incomplète des systèmes de sécrétion existants.

En 2014, MacSyFinder (Abby *et al.*, 2014) a introduit une nouvelle approche en utilisant une base de données HMM pour annoter les gènes dans les génomes. Il a également introduit la notion de modèle de système pour décrire les composants du système et son organisation génomique (figure I.3.5). Par exemple, les modèles TXSScan (Abby *et al.*, 2016) permettent de détecter les systèmes de sécrétion, tandis que TFFScan (Denise *et al.*, 2019) et CONJScan (Cury *et al.*, 2017) sont utilisés pour les filaments de type IV et les systèmes de conjugaison, respectivement.

11. Firmicutes, renommé en 2021 Bacillota est un phylum

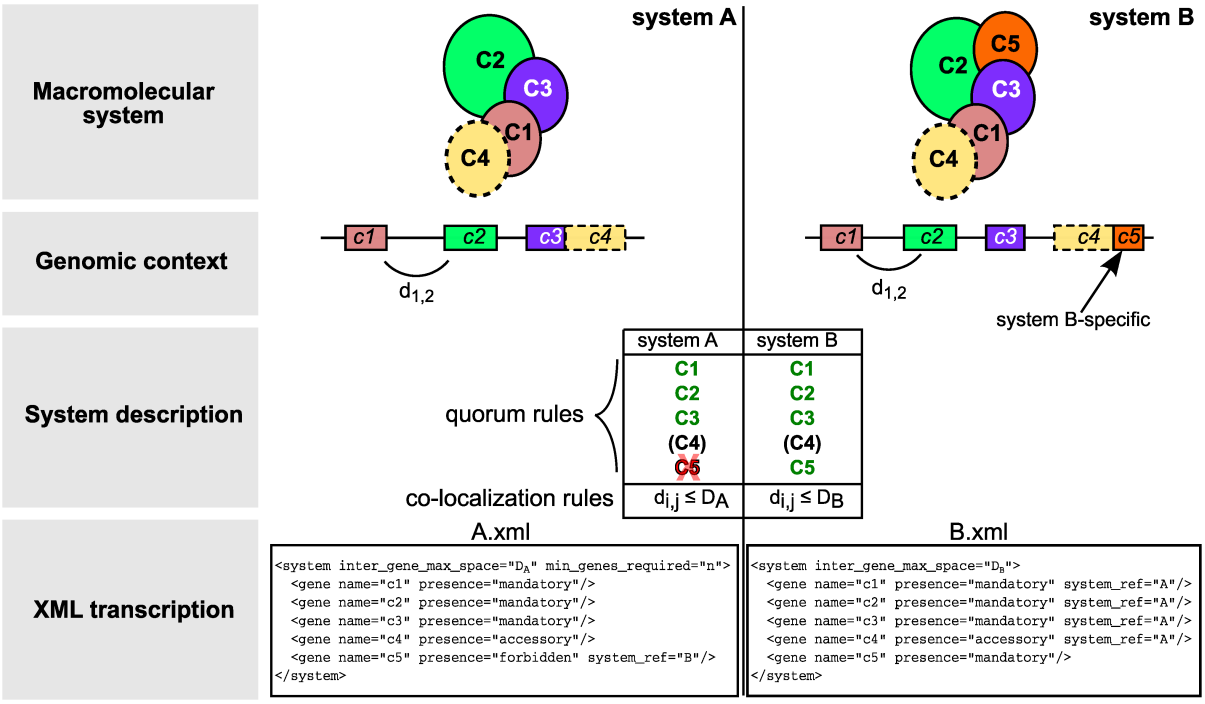


Figure I.3.5 – **Exemple de modélisation de systèmes dans MacSyFinder.** Extrait de (Abby *et al.*, 2014)

Plus récemment, T4SEpre (Wang *et al.*, 2014) et T4SEpp (Hu *et al.*, 2024) ont utilisé des modèles et des méthodes d'apprentissage automatique pour améliorer la sensibilité et la spécificité de la prédiction des systèmes T4SS.

b. Systèmes impliqués dans le métabolisme secondaire

En 2011, antiSMASH (Medema *et al.*, 2011) a marqué une avancée majeure en automatisant l'identification des groupes de gènes colocalisés impliqués dans une voie de biosynthèse, appelés *Biosynthetic Gene Clusters* (BGCs). Cette identification repose sur une vaste base de données de profils HMM, permettant la détection d'un large éventail de BGCs, notamment les NRPS (peptides non ribosomiques), les PKS (polykétides), les RIPP (*Ribosomally Synthesized and Post-translationally Modified Peptides*) ainsi que d'autres métabolites secondaires.

Des méthodes récentes, telles que DeepBGC (Hannigan *et al.*, 2019) et GECCO (Carroll *et al.*, 2021), utilisent des modèles de *deep learning* et des approches de traitement du langage naturel pour prédire les BGCs. Ces méthodes permettent une classification plus précise, bien qu'elles nécessitent une puissance de calcul importante et soient moins accessibles aux microbiologistes que les outils classiques.

3.2.3 . Les systèmes de défense aux phages

Dans leur environnement naturel, les procaryotes sont fréquemment exposés à l'infection par des **phages**. Au fil de l'évolution, ces virus ont développé une remarquable diversité de mécanismes leur permettant d'infecter un éventail plus ou moins large d'espèces. Ces dernières années, l'intérêt pour les phages s'est considérablement accru, passant de 452 publications mentionnant le terme phage dans les MeSH Terms de PubMed en 2000 à 1250 en 2024. Cet engouement est notamment porté par la reconsidération de la **phagothérapie** [12], une approche permettant de combattre les infections bactériennes (Boniver *et al.*, 2022). Bien que cette stratégie thérapeutique ait été utilisée dès l'après Première Guerre mondiale, elle a progressivement été délaissée au profit des antibiotiques. Toutefois, la montée alarmante des souches bactériennes multirésistantes conduit aujourd'hui à réexaminer les phages comme une alternative thérapeutique.

Face à ces infections, les procaryotes ont, eux aussi, développé une vaste panoplie de mécanismes regroupés sous le terme **systèmes de défense contre les phages**. Cette compétition entraîne une course coévolutive, menant à une diversification continue des stratégies d'infection des phages et des systèmes de défense. Les microbiologistes s'intéressent de plus en plus à ces interactions complexes, non seulement pour leurs applications pratiques en génétique moléculaire, comme l'exploitation des enzymes de restriction, mais aussi pour leur potentiel dans le développement de la phagothérapie. La connaissance des mécanismes permettant à une souche de résister à une gamme spécifique de phages étant essentielle pour concevoir des traitements efficaces et adaptés.

3.2.3.1 . Phages : retour sur les virus de bactéries

Les virus infectant les bactéries, connus sous le nom de bactériophages ou phages, ont été observés pour la première fois en 1915 et décrits officiellement par Félix d'Hérelle. Ces phages, dont la taille varie entre 20 et 200 nanomètres, présentent une grande diversité de formes. Leur matériel génétique peut être constitué d'ADN ou d'ARN, en simple ou double brin (figure I.3.6). Chaque phage possède un spectre d'hôtes spécifique, *i.e*, qu'il ne peut infecter qu'un nombre restreint et défini d'espèces procaryotes.

Les phages ne sont pas capables de répliquer leur propre matériel génétique, c'est pourquoi ils infectent les cellules procaryotes, afin d'utiliser les systèmes de réplication de l'hôte. Une fois que le matériel a été répliqué (des milliers de fois), les nouveaux phages seront libérés dans l'environnement en lysant la cellule (ouverture de la paroi). Le cycle d'infection, réplication, libération existe sous 2 formes définissant 2 catégories de phages (figure I.3.7). Le cycle lytique, réalisé par les phages virulents, correspond à un cycle court où le phage détruit l'hôte à la fin de sa réplication. Le cycle lysogénique, opéré par les phages tempérés, réfère à un phage qui va rester dans la cellule pendant plusieurs cycles de réplication de l'hôte. Dans ce cas, le matériel génétique peut s'intégrer au chromosome de l'hôte et se répliquer avec lui, on parle de région prophagique, ou rester dans le cytoplasme sous forme d'épisome et se répliquer indépendamment comme un plasmide.

12. Utilisation des phages pour traiter certaines maladies en ciblant sélectivement les cellules.

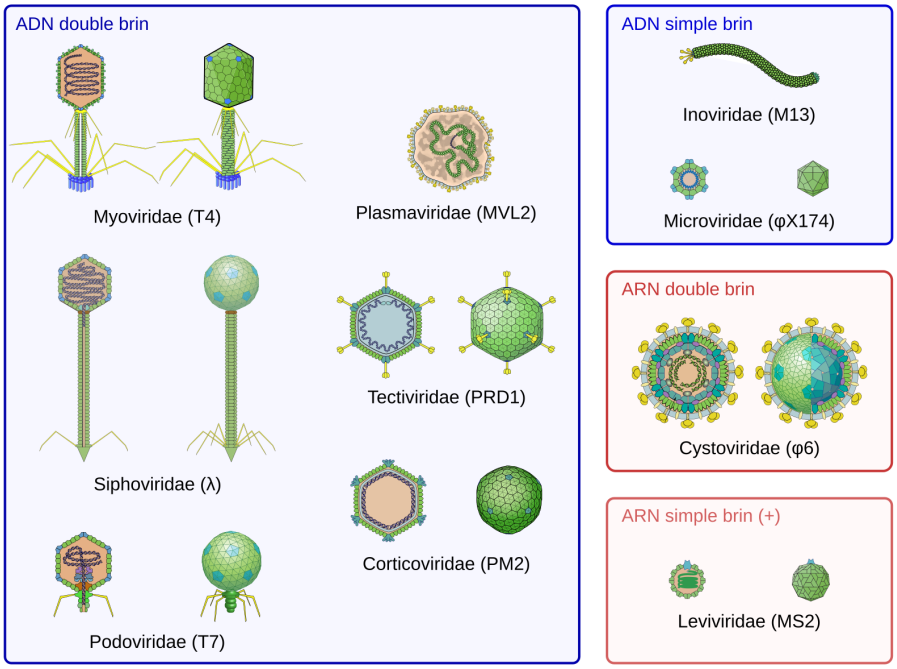


Figure I.3.6 – Diversité morphologique des phages. Auteur : Philippe Le Mercier - ViralZone SIB Swiss Institute of Bioinformatics

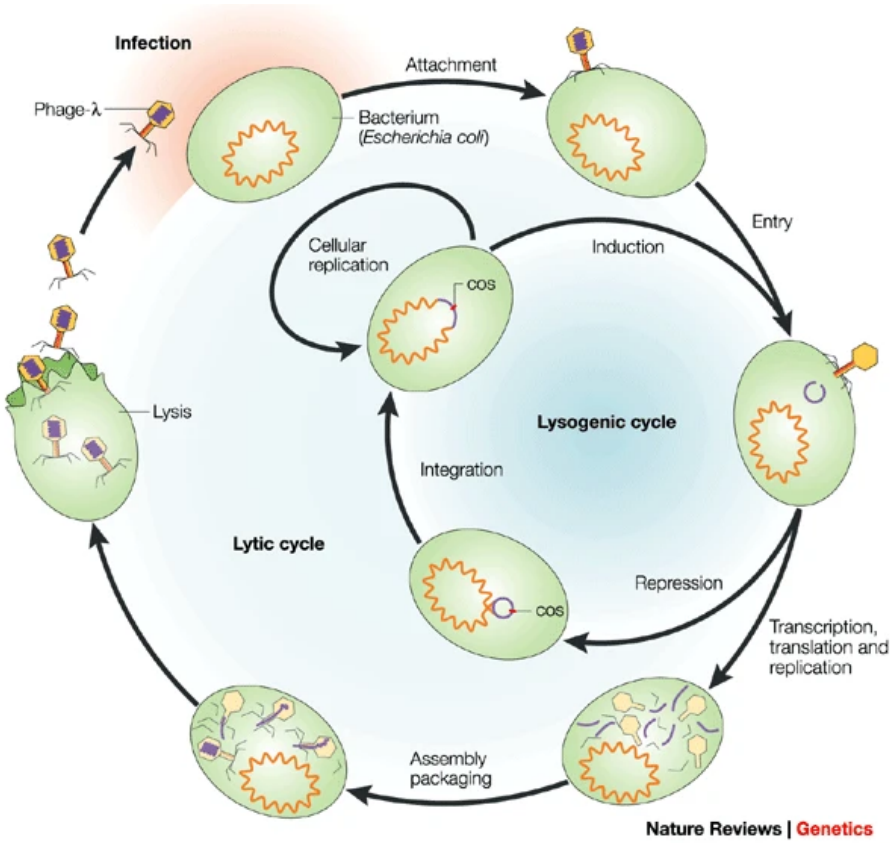


Figure I.3.7 – Cycle de vie des phages. Extrait de (Campbell, 2003)

3.2.3.2 . Mécanismes de défense contre les phages

Pour se défendre contre les phages, les procaryotes ont développé un arsenal pour se protéger : les systèmes de défense contre les phages [13] (Makarova *et al.*, 2013). Un système de défense correspond à un ensemble de protéines qui vont empêcher l'infection du phage et donc empêcher la destruction de la cellule. Ils peuvent agir de manière très diverse et à différents moments du cycle de vie du phage.

Les premiers systèmes de défense ont été identifiés dans les années 50, il s'agit des systèmes de restriction-modification (RM) (Bertani et Weigle, 1953). Ces systèmes sont composés de deux fonctions principales, généralement assurées par deux protéines distinctes : la reconnaissance et la coupure de l'ADN étranger (REase), et la modification par méthylation (MTase) pour protéger l'ADN de la coupure. La REase n'étant pas spécifique, l'action de la MTase permet de prévenir et de protéger les réplicons de l'hôte contre les coupures.

C'est à partir des années 2000 que de nouveaux systèmes de défense ont été identifiés. Les systèmes CRISPR-Cas, connus notamment aujourd'hui pour leur application en médecine et en génétique en tant que ciseaux moléculaires(Haft *et al.*, 2005; Barrangou *et al.*, 2007) [14] pour découper des séquences d'ADN cible. Les CRISPRs correspondent à des clusters de séquences palindromiques répétées et régulièrement espacées par des régions appelées *spacer*. Les séquences CRISPR sont associées à des protéines Cas dont la première fonction est de se lier à des transcrits de *spacer* pour identifier spécifiquement l'ADN étranger dans la cellule et de le découper. La seconde fonction va être de récupérer cet ADN pour l'intégrer dans le chromosome entre des séquences CRISPR et en faire un nouveau *spacer*. Certains de ces *spacers* correspondent à des séquences d'ADN phagique et seront utilisés par des protéines Cas pour combattre l'infection virale. Les systèmes CRISPR-Cas permettent donc à la cellule de répondre efficacement aux infections par des phages connus, mais aussi de construire une mémoire des infections phagiques.

Il existe également des systèmes d'infection abortive (Abi, pour *Abortive infection* en anglais) qui entraînent la mort de l'hôte avant la réplication du phage (Molineux, 1991). Contrairement aux mécanismes précédents qui protègent l'hôte de l'infection, ces mécanismes permettent de protéger les bactéries environnantes en empêchant le phage de se multiplier. Récemment, la découverte récente de nouveaux systèmes Abi a mené à revoir leur définition et leur classification en tant que mécanisme de défense est discutée. Dans leur article, Aframian et Eldar soutiennent que Abi ne doit pas être considéré comme un système de défense, mais comme une issue possible pour l'organisme, qu'il peut emprunter dans certaines conditions (Aframian et Eldar, 2023).

Aujourd'hui, plus de 150 systèmes sont référencés et pour la majorité, ils ont été découverts dans les 10 dernières années, suite à l'intérêt croissant pour les phages et leur application, mais aussi au développement de méthodes pour les détecter. En 2018, Doron, Melamed *et al.* (Doron *et al.*, 2018) ont étudié les gènes localisés à proximité de systèmes de défense. Les systèmes de défense étant concentrés dans les îlots génomiques (îlots de défense)(Makarova *et al.*, 2011), ils ont spécifiquement étudié ces régions. Ils ont ainsi pu identifier 26 nouveaux systèmes de défense, dont 9 qui ont pu être confirmés expérimentalement. Les études suivantes, qui ont permis d'identifier de nouveaux systèmes, se basent sur la même stratégie.

13. que nous raccourcirons en systèmes de défense dans cette partie
14. Emmanuelle Charpentier et Jennifer A. Doudna ont reçu le prix Nobel de chimie en 2020 pour avoir découvert les ciseaux génétiques CRISPR/Cas9

Les systèmes de défense peuvent être classés en 3 grandes catégories (figure I.3.8) :

(i) Les systèmes qui reconnaissent l'ADN des phages, utilisent des séquences d'ADN pour identifier et dégrader l'ADN viral, offrant ainsi une immunité adaptative ;

(ii) Les systèmes qui reconnaissent les protéines de phages, tels que les systèmes AVAST (Gao *et al.*, 2020), ciblent et inactivent les protéines essentielles des phages, empêchant ainsi leur réplication ;

(iii) les systèmes surveillant l'intégrité de la cellule, comme le système toxine-antitoxine : ToxIN (Guegler et Laub, 2021), déclenchent des réponses suicidaires ou de dormance cellulaire en réponse à des dommages ou stress induits par les phages, limitant ainsi la propagation de l'infection.

D'autres systèmes, bien que moins caractérisés, jouent un rôle tout aussi important, incluant des mécanismes diversifiés qui interfèrent avec différentes étapes du cycle de vie des phages. Par exemple, le système CBASS (Cyclic oligonucleotide-Based Antiphage Signaling System), déclenche une réponse suicidaire contrôlée en cas d'infection virale, empêchant ainsi la propagation du phage. Un autre exemple est celui des systèmes viperin [15], inhibent la réplication virale en produisant des analogues de nucléotides modifiés qui bloquent la transcription de l'ADN phagique en agissant comme des chaînes de terminaison précoce. Ces stratégies contribuent à la résistance bactérienne globale contre les infections virales.

15. Système homologue à celui des eucaryotes

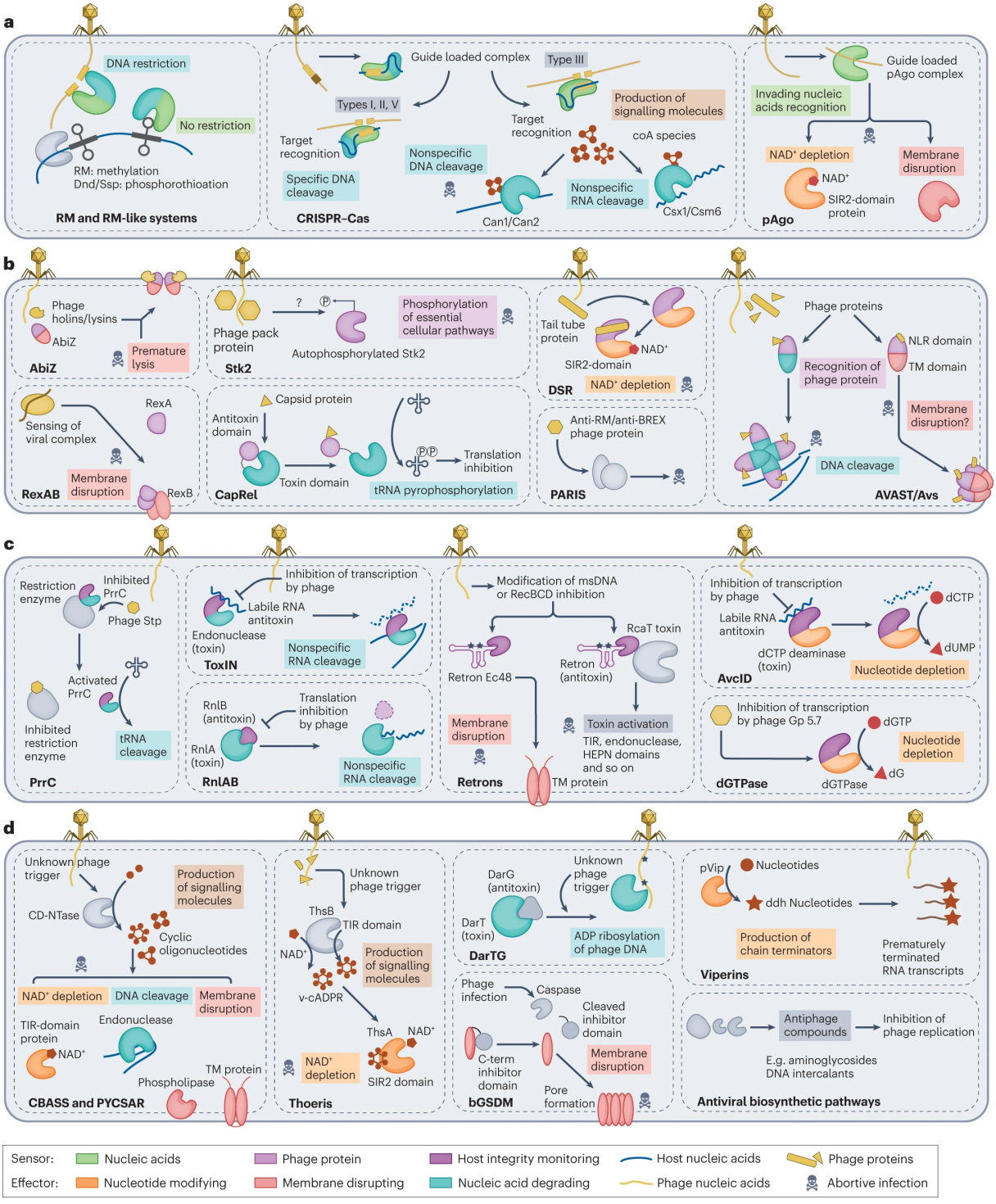


Figure I.3.8 – **Diversité des systèmes de défenses aux phages. a**) Systèmes de détection d'ADN étranger. **b**) Systèmes sensibles aux protéines phagiques. **c**) Systèmes de surveillance de l'intégrité de l'hôte. **d**) Autres mécanismes. Extrait de (Georjon et Bernheim, 2023)

Dans le même temps, avec l'émergence d'outils de détection (cf. sous-sous-section 3.2.3.3), on s'intéresse à la distribution de ces systèmes dans les espèces procaryotes. Bernheim et Sorek (Bernheim et Sorek, 2020) ont montré qu'au sein d'une espèce toutes les souches ne présentent pas les mêmes systèmes et que les organismes s'échangent des systèmes par transfert horizontal. Cette propriété permet aux organismes de rapidement s'adapter aux phages présents dans l'environnement. Le système immunitaire doit donc être considéré comme l'ensemble des systèmes présents dans les organismes de l'environnement. En 2022, Tesson *et al.* a montré que la composition en systèmes de défense varie entre les espèces, mais aussi selon la taille du génome, le risque d'infection et le mode de vie (Tesson *et al.*, 2022). La composition en systèmes de défense est aussi étroitement liée aux phages qui peuvent infecter la bactérie, et réciproquement (Srikant *et al.*, 2022). Pour terminer, Beavogui *et al.* se sont intéressés au système immunitaire dans les données de génomique environnementale et ont montré une distribution différente des systèmes de défense en fonction de l'habitat et de la géographie (Beavogui *et al.*, 2024).

Toutes ces études ont été permises par l'arrivée de méthodes et d'outils de détection automatique des systèmes de défense dans les génomes.

3.2.3.3 . Méthodes et outils de détection

Les premiers outils de détection dans les génomes, étaient spécialisés dans l'identification des systèmes CRISPR. Leur approche reposait sur la recherche de séquences répétées intercalées de séquences uniques, grâce à des méthodes d'alignement. PILER-CR (Edgar, 2007) identifie d'abord toutes les séquences répétées palindromiques, sélectionne celles correspondant aux CRISPR (24 à 48 pb, séparées par des séquences uniques), puis affine leur détection grâce à une approche basée sur l'analyse de graphes et le partitionnement. L'outil CRT (Bland *et al.*, 2007), utilise des k-mers pour rechercher des séquences répétées d'une taille donnée, éloignées d'une distance définie et dont la séquence est unique. Ces 2 méthodes sont rapides et ont l'intérêt de détecter toutes les séquences répétées candidates pour être des CRISPR. L'outil CRISPRFinder (Grissa *et al.*, 2007), va suivre un schéma similaire aux outils précédents, mais va introduire une notion de score, qui prend en compte le nombre de répétitions, leur taille, la régularité et la taille des espacements. De plus, une fois les séquences candidates filtrées, pour améliorer sa précision, CRISPRFinder peut comparer les candidats à sa base de données de CRISPRs validées.

Avec l'accumulation des connaissances autour des CRISPRs et des séquences environnantes qui les composent, les outils vont intégrer de nouveaux critères de détection. Des outils comme CRISPRstrand(Alkhnbashi *et al.*, 2014), CRISPRDirection(Biswas *et al.*, 2014) utilisent les séquences d'ARNcr [16], d'autres utilisent les séquences leader [17] comme CRISPRleader (Alkhnbashi *et al.*, 2016). La première version de MacSyFinder (Abby *et al.*, 2014) intégrait une base de données de profil HMM et de modèles CasFinder, pour identifier les protéines Cas et autres séquences connues proches pour identifier les systèmes CRISPR-Cas. En 2018, une version hybride entre CRISPRFinder et CasFinder est proposée : CRISPRCasFinder (Couvin *et al.*, 2018). Cet outil permet de prendre en compte la structure des CRISPR et des *spacers*, ainsi que les gènes environnants, pour détecter finement les systèmes CRISPR-Cas.

16. Les ARNcr, sont un type d'ARN contenant le transcrit d'une partie du CRISPR et le spacer. Ils sont utilisés dans la reconnaissance spécifique de l'ADN étranger.
17. Une séquence séparant les CRISPR des gènes codant pour les Cas.

En 2021, la découverte de nombreux nouveaux systèmes de défense a conduit au développement d'outils, comme PADLOC (Payne *et al.*, 2021), pour leur identification dans les génomes. PADLOC s'appuie sur une base de données HMM et de modèles décrivant les systèmes inspirés de la grammaire des modèles de MacSyFinder (Abby *et al.*, 2014). Peu après, DefenseFinder (Tesson *et al.*, 2022) a été publié, adoptant une approche méthodologique similaire reposant sur MacSyFinder pour la détection. Bien que ces outils partagent un même principe de fonctionnement, ils diffèrent principalement dans la construction des profils HMM et dans les règles de détection des systèmes. PADLOC génère des HMMs entièrement *de novo*, *ie* qu'il construit sa propre base de données de profils, tandis que DefenseFinder s'appuie en partie sur des bases de données de protéines existantes, comme Pfam (Mistry *et al.*, 2021). Par ailleurs, PADLOC privilégie une approche fondée sur des modèles plus généralistes, intégrant des règles plus flexibles afin de faciliter l'identification de systèmes proches de ceux connus. À l'inverse, DefenseFinder adopte des modèles plus stricts, intégrant un plus grand nombre de paramètres pour affiner la classification des systèmes identifiés. Ainsi, le choix entre ces deux outils doit être guidé par les objectifs spécifiques de l'étude. PADLOC constitue une solution privilégiée pour les analyses exploratoires visant à détecter de nouveaux systèmes proches, tandis que DefenseFinder se révèle plus adapté aux études nécessitant une identification précise et rigoureuse des systèmes déjà caractérisés.

3.3 . Génomique à l'ère du Big Data

Avec l'évolution des méthodes de séquençage et la diminution des coûts, de plus en plus de projets de recherche s'appuient sur le séquençage des génomes pour faire des analyses de génomique comparée. Les chercheurs peuvent ensuite rendre ces séquences publiques et les déposer dans des banques de données (BD) de génomes, comme GenBank (Burks *et al.*, 1985). GenBank est la BD de séquence du *National Center for Biotechnology Information* (NCBI), toutes les séquences qui y sont soumises passent un contrôle d'intégrité et de qualité, avant d'être annotées automatiquement. Entre 2010 et 2025, le nombre de génomes stockés dans GenBank a connu une croissance exponentielle, passant de quelques milliers à plus de 2,5 millions de génomes. (figure I.3.9).

À partir des génomes annotés, il est possible d'obtenir la traduction des gènes en séquences protéiques, prédire leurs fonctions, leurs structures…Toutes ces informations vont également être contenues dans des BD spécifiques pour aider aux analyses. Les BD peuvent aussi être thématiques, en contenant uniquement les génomes d'une espèce ou d'une région géographique, par exemple.

Avec l'essor du Big Data, les méthodes de génomique comparée vont être améliorées et adaptées afin d'analyser efficacement ce nouvel immense volume de données.

3.3.1 . Base de données génomiques

L'existence et l'accessibilité des bases de données biologiques sont fondamentales pour l'annotation, la classification et l'analyse des séquences génomiques et protéiques. Parmi ces ressources, plusieurs se distinguent par leur spécialisation.

Parmi les bases de données de référence de génomes, RefSeq (Pruitt *et al.*, 2007), développée et maintenue par le NCBI, se distingue par la qualité et la fiabilité de ses annotations. Contrairement à GenBank, alimentée par des soumissions d'utilisateurs, RefSeq compile des séquences rigoureusement validées, incluant des ARN et des protéines

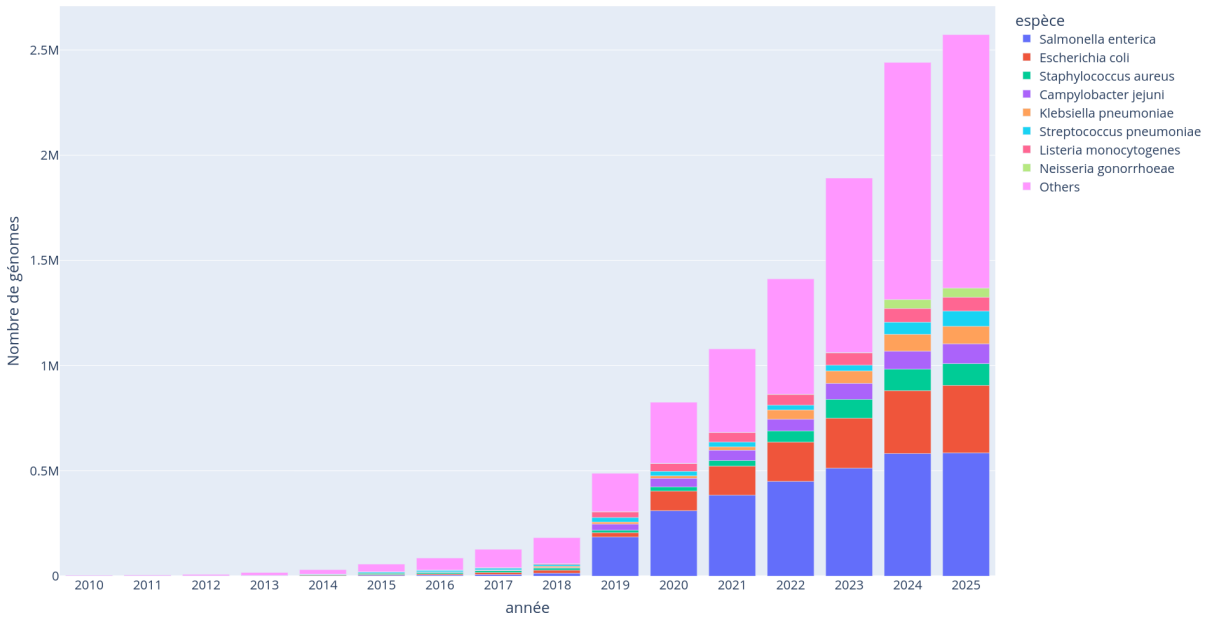


Figure I.3.9 – **Nombre de génomes cumulés par an dans GenBank.** Les génomes pris en compte sont uniquement ceux des procaryotes. Les espèces indiquées en légende sont les plus représentées dans GenBank. Construit avec l'outil drawbank `https://github.com/axbazin/drawbank`

issus d'une grande diversité d'organismes. Son processus de standardisation strict en fait une ressource incontournable pour la génomique comparative et une base essentielle pour des outils phares du NCBI, comme BLAST. En complément des séquences individuelles, RefSeq propose des ensembles complets de génomes annotés, notamment pour des organismes modèles et des pathogènes.

Dans le domaine des protéines, **UniProt** (The UniProt Consortium, 2025) est une ressource incontournable pour l'annotation et l'analyse des séquences protéiques. Elle est développée et maintenue par un consortium composé de l'*European Bioinformatics Institute* (EMBL-EBI), le *Swiss Institute of Bioinformatics* (SIB) et la *Protein Information Resource* (PIR). Elle se divise en 3sections : (*i*) **UniProtKB** (*KnowledgeBase*), qui comprend des annotations détaillées sur les protéines et est divisé en **Swiss-Prot** (annotations manuelles et validées) et **TrEMBL** (annotations automatiques) (Apweiler *et al.*, 2004; Bairoch *et al.*, 2004) ; (*ii*) **UniRef** (*UniProt Reference Clusters*), qui regroupe des séquences non redondantes à différents seuils de similarité (Suzek *et al.*, 2007) ; (*iii*) **UniParc** (*UniProt Archive*), une archive complète de toutes les séquences protéiques connues, indépendamment de leur source. Cette base de données est essentielle pour la modélisation structurale des protéines, la découverte de cibles thérapeutiques et la génomique fonctionnelle.

GTDB (Genome Taxonomy Database) (Parks *et al.*, 2018) propose une taxonomie révisée des génomes procaryotes, fondée sur des analyses phylogénomiques. Contrairement aux approches classiques reposant sur des critères phénotypiques, GTDB utilise une approche purement génomique basée sur l'analyse comparée de 120 gènes marqueurs afin d'harmoniser la classification. Comme l'illustre la figure I.3.10, cette approche permet d'obtenir une taxonomie plus homogène que celle du NCBI. De plus, GTDB intègre les génomes environnementaux et non cultivables issus de métagénomes, facilitant ainsi l'identification de nouvelles espèces et l'exploration de la diversité microbienne selon des critères plus objectifs.

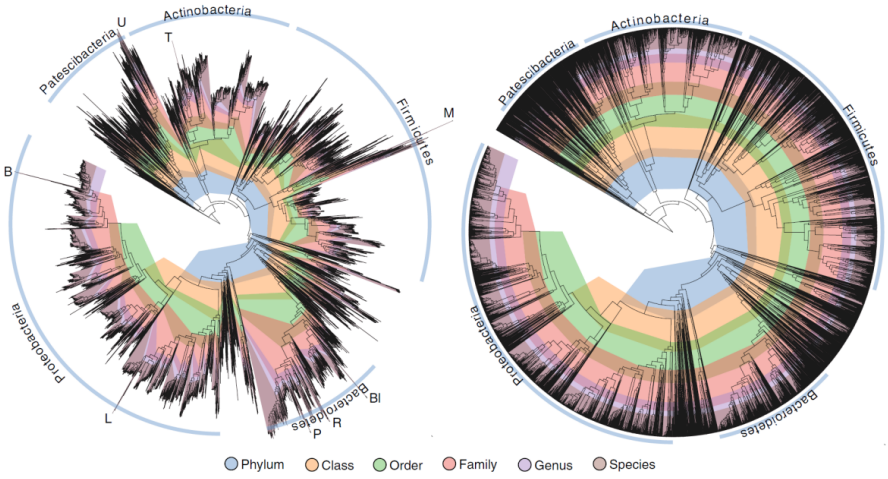


Figure I.3.10 – **Comparaison de l'homogénéité des rangs taxonomiques entre le NCBI et GTDB.** Les 2 figures représentent le même arbre avec à gauche la taxonomie proposée par le NCBI et à droite celle proposée par GTDB. Copié de (Gautreau, 2020) et adapté de (Parks *et al.*, 2018)

L'analyse des communautés microbiennes complexes repose sur des bases de données spécialisées comme **MGnify**[18] (Richardson *et al.*, 2023), une plateforme développée par l'EMBL-EBI dédiée à l'étude des données métagénomiques. Contrairement aux bases centrées sur des séquences génétiques isolées, MGnify s'intéresse aux microbiomes présents dans divers environnements tels que le sol, les océans, les eaux douces, le microbiote humain ou encore les milieux extrêmes. En plus de stocker des données, elle propose des outils bioinformatiques avancés pour l'assemblage, l'annotation et l'analyse fonctionnelle des communautés microbiennes. MGnify permet ainsi d'identifier les espèces présentes dans un échantillon via des approches de taxonomie basées sur l'ARNr 16S/18S, de prédire les fonctions métaboliques des microbiomes et d'étudier leurs interactions avec l'environnement. Cette ressource est devenue incontournable en écologie microbienne et en biotechnologie.

Dans le domaine de la résistance aux antibiotiques, CARD (*Comprehensive Antibiotic Resistance Database*) (McArthur *et al.*, 2013) est une base de données spécialisée qui regroupe des informations sur les gènes de résistance, les mutations associées et les mécanismes moléculaires impliqués. Elle est notamment utilisée pour identifier les gènes de résistance dans des échantillons génomiques en utilisant l'outil RGI (Alcock *et al.*, 2023), et métagénomiques grâce à l'outil CARPDM (Hackenberger *et al.*, 2024). CARD adopte une nomenclature standardisée basée sur l'ontologie ARO (*Antibiotic Resistance Ontology*) pour classifier les résistances selon leur mode d'action et leur mode de transmission, facilitant ainsi l'étude et la surveillance des résistances aux antibiotiques.

Enfin, pour les chercheurs spécialisés dans l'étude des bactéries du genre Pseudomonas, la **Pseudomonas Genome Database** (PGD) constitue une ressource précieuse (Winsor *et al.*, 2016). Cette base de données centralise des informations sur le séquençage et l'annotation des génomes des différentes espèces de Pseudomonas, un groupe bactérien d'importance en médecine, en agriculture et en biotechnologie. PGD fournit des annotations génomiques détaillées, des informations fonctionnelles et des descriptions précises des voies métaboliques. De plus, les données y sont validées par des experts du domaine. La plateforme permet également des analyses comparatives entre souches, offrant ainsi un outil puissant pour l'étude de la diversité génétique et fonctionnelle de ces bactéries.

18. anciennement nommé EBI Metagenomics (Hunter *et al.*, 2014)

Chacune de ces bases de données joue un rôle clé dans l'analyse des génomes et des protéines, en apportant des solutions adaptées aux besoins des microbiologistes. Leur complémentarité permet une exploration approfondie des données biologiques et contribue aux avancées scientifiques dans de nombreux domaines.

3.3.2 . Bases de données orientées graphe et données biologiques

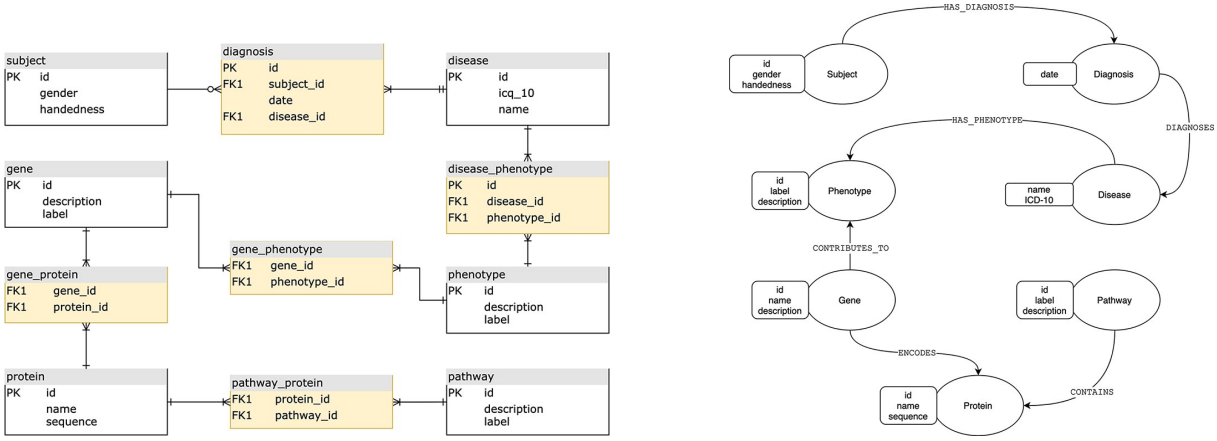
Les premières bases de données biologiques ont été construites sur un modèle relationnel, fondé sur l'organisation des données en tables où chaque ligne représente un élément et chaque colonne un attribut de cet élément. Pour établir des relations entre ces entités, chaque élément se voit attribuer un identifiant unique, qui est ensuite utilisé dans des tables de correspondance reliant différents éléments entre eux. Ce modèle s'avère particulièrement pertinent lorsque les données sont relativement stables et que les relations entre éléments sont peu nombreuses et secondaires par rapport aux attributs eux-mêmes. Il permet ainsi de stocker efficacement un génome et ses métadonnées associées, telles que l'organisme d'origine, le laboratoire de séquençage, la date d'obtention ou encore les publications qui y sont liées.

Toutefois, avec l'accroissement massif des données biologiques, ces BD relationnelles, appelées aussi SQL [19], atteignent leurs limites. Lorsque les données deviennent extrêmement connectées, interroger ces BD nécessite des requêtes complexes et des ressources computationnelles importantes, ce qui peut ralentir l'accès et l'analyse des informations (Hsu *et al.*, 2014). Pour pallier ces contraintes, de nouveaux modèles émergents, privilégiant une approche non seulement relationnelle (NoSQL). Il existe plusieurs types de BD NoSQL, ici, nous nous focaliserons sur un type particulier : les bases de données orientées graphe (BDG).

Contrairement aux bases relationnelles, les BDG modélisent les données sous forme de graphes, où les nœuds représentent les éléments et les arêtes définissent leurs relations. Ce modèle offre une représentation plus intuitive des connexions complexes, ce qui est particulièrement utile pour l'analyse des réseaux biologiques, comme les interactions entre protéines, les relations entre gènes ou encore les mécanismes de résistance aux antibiotiques. De plus, bien que ces bases reposent sur une structure différente, il reste possible de traduire un graphe en une base relationnelle, comme illustré sur la figure I.3.11.

Parmi les bases de données et outils qui exploitent ces nouveaux modèles, GDM (Genomic Data Model) et GMQL (Genomic Metadata Query Language) illustrent bien l'évolution vers des architectures plus flexibles adaptées aux vastes ensembles de données génomiques (Masseroli *et al.*, 2015; Masseroli *et al.*, 2016). GDM est un modèle conçu pour gérer efficacement des données hétérogènes issues de la génomique, en intégrant à la fois des informations sur les séquences et leurs annotations. Dans cette perspective, GMQL se présente comme un langage de requête avancé permettant d'interroger ces bases de manière optimisée, facilitant l'exploration et l'intégration de données. Grâce à son interaction avec un modèle orienté graphe, GMQL améliore la capacité à interroger et à analyser de grands ensembles de données interconnectées, rendant possible l'identification rapide de corrélations complexes entre différentes sources d'information génomique.

19. SQL réfère au langage de requête de ces bases, mais par abus de langage, on parle de base SQL



(a) Base de données relationnelle

(b) Base de données graphe

Figure I.3.11 – **Comparaison des modèles entre une base de données relationnelle et une base de données orientées graphe.** (a) Les tables contenant les informations sont représentées en blanc et les tables de jointure en jaune. (b) Les nœuds représentent les informations et les arêtes les relations (jointure). Extrait de (Timón-Reina *et al.*, 2021)

Dans le domaine du Web sémantique, BioRDF (Cheung *et al.*, 2009) permet de modéliser et d'interconnecter des ressources issues des DB, biomédicales en particulier. Reposant sur le *Ressource Description Framework* (RDF), cette approche permet de relier des concepts biologiques de manière standardisée, permettant ainsi une interopérabilité accrue entre les systèmes et l'extraction de nouvelles connaissances à partir de réseaux complexes. RDF est un modèle de données qui représente l'information sous forme de graphes. Chaque déclaration RDF est un triplet composé d'un sujet, d'un prédicat et d'un objet, ce qui forme naturellement une structure de graphe. BioRDF est utilisé dans diverses applications biologiques, y compris la création de *workflows* bioinformatiques, l'annotation de séquences protéiques, et la modélisation de voies biologiques. Des outils comme Taverna (Oinn *et al.*, 2004) sont utilisés pour composer et exécuter ces workflows.

Un autre exemple de base de données orientée graphe est KEGG (*Kyoto Encyclopedia of Genes and Genomes*), une ressource essentielle pour l'analyse des voies métaboliques, des interactions moléculaires et des relations entre gènes et maladies (Kanehisa *et al.*, 2025). Contrairement aux bases relationnelles classiques, KEGG structure ses données sous forme de graphes de réseaux métaboliques, où les nœuds représentent des gènes, des protéines ou des petites molécules, interconnectés à travers des réactions biochimiques représentées par les arêtes. Cette organisation permet d'étudier le fonctionnement des systèmes biologiques à grande échelle et d'identifier des cibles potentielles pour le développement de nouvelles thérapies.

Ces bases de données sont également utilisées par des outils pour réaliser des analyses et des prédictions. C'est le cas de Spfy (Le *et al.*, 2018), qui permet de prédire des phénotypes bactériens sur de nouveaux échantillons à partir d'une base de données MongoDB [20] (Guo, 2017). Spfy est capable de prédire des caractéristiques phénotypiques importantes telles que le sérotype, le sous-type de la toxine Shiga (toxine sécrétée par les bactéries *E. coli*), ainsi que la présence de facteurs de virulence et de déterminants de résistance aux antimicrobiens. Actuellement, la base de données contient plus de 10 000 génomes de *Escherichia coli*.

20. MongoDB, une base orientée documents, adapté aux ensembles de données évolutifs et hétérogènes.

Enfin, l'utilisation des bases de données orientées graphe s'est intensifiée avec des systèmes de gestion de base de données orientée graphe open-source comme Neo4J, qui a été largement exploité dans des projets tels que CovidGraph (Gütebier *et al.*, 2022). Ce dernier est un réseau de connaissances conçu pour analyser la littérature scientifique, les bases de données biomédicales et les publications liées au SARS-CoV-2. En structurant ces informations sous forme de nœuds et d'arêtes, Neo4J permet d'explorer efficacement les relations complexes entre les études, les auteurs, les interactions moléculaires et les traitements potentiels contre la maladie.

Ces nouvelles approches et ces outils illustrent la transition vers des modèles de bases de données plus dynamiques et adaptés aux défis de la biologie moderne. Que ce soit par l'utilisation de langages de requête spécialisés, de technologies sémantiques, de modèles en graphe ou encore de bases NoSQL, ces solutions permettent une gestion plus efficace des données biomédicales et ouvrent la voie à des analyses plus approfondies dans le domaine des sciences de la vie.

3.3.3 . L'intelligence artificielle au service de la génomique comparée

3.3.3.1 . Définitions et concepts

Avec l'accumulation massive de données, la génomique est désormais confrontée aux défis du *Big Data*, tant en termes de volume que de complexité. Les bioinformaticiens se tournent donc vers des méthodes développées dans la science des données (*data science*) et particulièrement vers l'intelligence artificielle (IA). L'intelligence artificielle regroupe un ensemble de techniques permettant aux machines de reproduire certaines facultés cognitives humaines, telles que l'apprentissage, le raisonnement et la résolution de problèmes. Elle s'appuie sur divers domaines de recherche définissant le traitement des données, la prise de décision et l'adaptation des algorithmes aux informations reçues. Ici, nous nous intéresserons à un champ de recherche particulièrement utilisé en bioinformatique, celui de l'apprentissage automatique (*machine learning en anglais*, ML).

Les méthodes de ML correspondent à des méthodes qui s'améliorent par l'apprentissage ou l'entraînement. Pour fonctionner, elles ont besoin de grands jeux de données, bien définis. Elles sont donc bien adaptées à la génomique du Big Data. Dans une application de ML (figure I.3.12), après avoir bien défini les données d'entrée et la prédiction attendue, une phase d'apprentissage permet de construire le modèle le plus adapté. La première étape de l'apprentissage consiste à entraîner le modèle sur un jeu de données (d'entraînement) pour sélectionner les caractéristiques les plus pertinentes et estimer les meilleurs paramètres du modèle. Une deuxième étape d'évaluation (ou de test) mesure les performances du modèle, *i.e.*, la divergence avec le résultat attendu [21]. En fonction de l'évaluation, un nouvel apprentissage peut être relancé pour améliorer le modèle. Même si le modèle ne sera jamais idéal, le choix des jeux d'entraînement et de test permet de s'approcher au mieux de la prédiction souhaitée.

21. Cette divergence est mesurée par la fonction de perte ou de coût

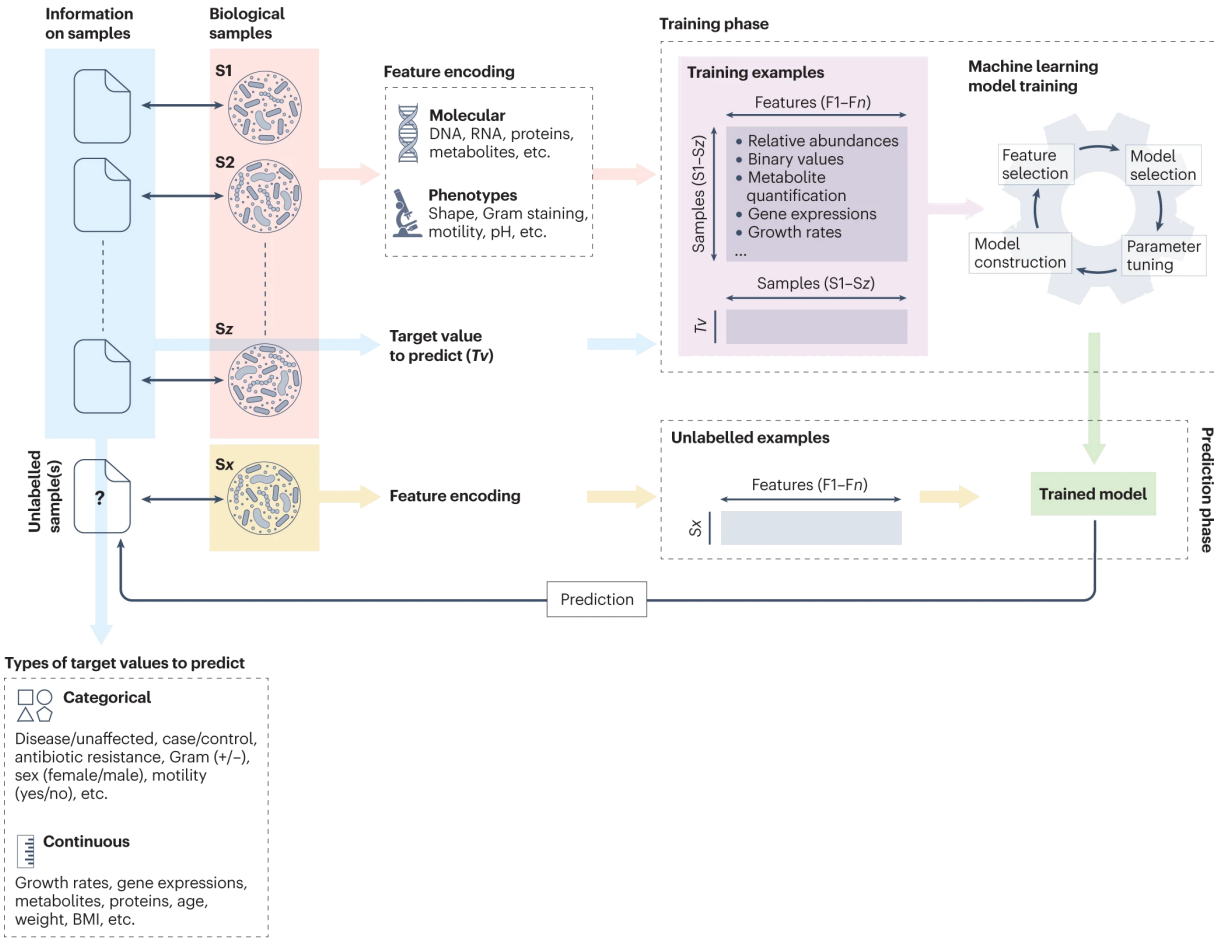


Figure I.3.12 – **Schéma général d'une application de machine learning pour l'assignation de caractéristiques moléculaires et phénotypiques.** Le schéma correspond à l'application d'un modèle supervisé où les données d'entrée sont étiquetées et dont la prédiction attendue est, elle aussi, une étiquette. Extrait de (Asnicar *et al.*, 2024)

Les méthodes de machine learning peuvent être divisées en 2 grandes catégories : (*i*) l'apprentissage **supervisé** ; (*ii*) l'apprentissage **non supervisée**.

Les modèles supervisés permettent d'assigner des **étiquettes** aux données. Dans ce cas, les données du jeu d'entraînement et de test sont étiquetées par le résultat attendu (à priori). La figure I.3.12, présente un ensemble d'échantillons dont on possède les informations d'intérêt. Le modèle entraîné est appliqué sur un nouvel échantillon et permet donc de prédire les informations jusqu'alors inconnues. Par exemple, en entraînant un classificateur supervisé (tel qu'un modèle de régression) sur des assignations génome-espèce connues, et en utilisant la présence ou l'absence de gènes marqueurs comme caractéristiques, il devient possible d'attribuer une taxonomie à un nouveau génome.

Les modèles non supervisés permettent de rechercher des structures dans des données sans nécessité d'étiquettes. Dans ce cas, après la phase d'entraînement, il y a une autoévaluation du modèle basée sur un renforcement positif. Par exemple, si l'on s'intéresse aux gènes impliqués dans la croissance bactérienne, les modèles peuvent partitionner les groupes de gènes présentant des profils d'expression génique similaires reflétant la croissance cellulaire. Les modèles non supervisés ont l'intérêt de pouvoir résoudre des problèmes plus complexes, sans avoir besoin de données annotées ; par contre, ils sont beaucoup plus imprévisibles que les modèles supervisés.

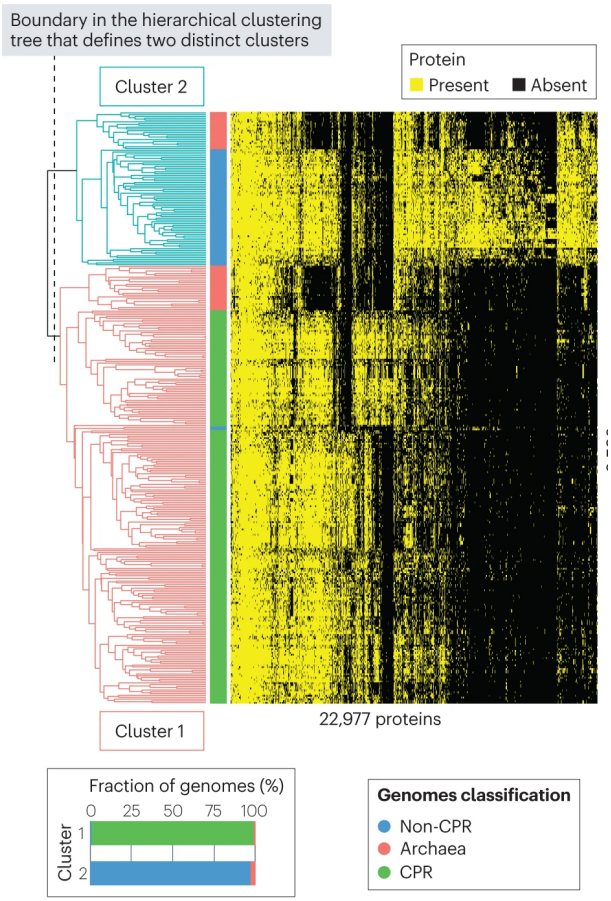


Figure I.3.13 – **Exemple d'application de méthode non supervisée.** Le *HeatMap* représente pour chaque souche la présence/absence d'un ensemble de protéines. La méthode de clustering (ici une Analyse en composantes principales) permet d'identifier 2 groupes dans les souches, correspondant à 2 groupes taxonomiques. Extrait et adapté de (Asnicar *et al.*, 2024)

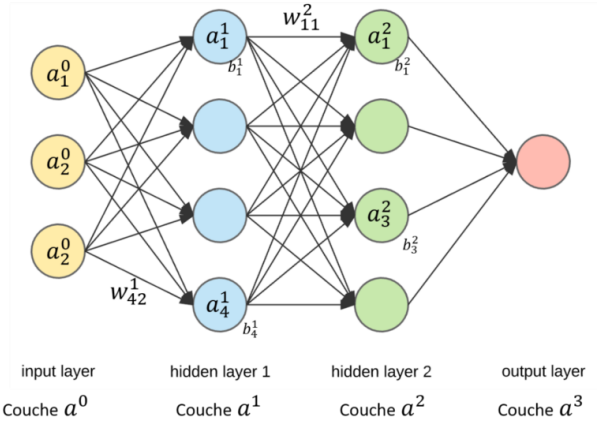


Figure I.3.14 – **Représentation d'un réseau de neurones à 3 couches.** Extrait de https://www.aspexit.com/reseau-de-neurones-on-va-essayer-de-demystifier-un-peu-tout-ca-1/

Dans les années 2010, un sous-domaine de l'apprentissage automatique gagne en popularité, l'apprentissage profond (*deep learning* en anglais DL)[22]. Les algorithmes de DL sont basés sur le concept de **réseaux neuronaux** (artificiels). Ces réseaux peuvent être représentés sous forme d'un graphe orienté, pondéré et étiqueté, comme sur la figure I.3.14. Ce réseau est organisé en couches successives, avec une couche d'entrée (à gauche) correspondant aux données brutes, des couches cachées (au centre) qui effectuent des transformations des données et extraient des caractéristiques, et enfin une couche de sortie (à droite) qui contient la prédiction. L'intérêt du DL réside dans sa capacité à extraire automatiquement les caractéristiques pertinentes lorsque les données passent dans les couches cachées. Plus le nombre de couches est élevé, plus le modèle peut apprendre, plus les questions que l'on pose peuvent être complexes et les réponses précises. Cependant, en augmentant la complexité, on augmente aussi la difficulté d'interprétation[23].

Réaliser une étude de ML/DL peut être un problème épineux. Il faut d'abord avoir une bonne compréhension des données pour s'orienter vers des méthodes supervisées ou non, et choisir entre l'apprentissage traditionnel et profond (cf. tableau I.3.2 & tableau I.3.3). Il convient également de s'assurer de la qualité des données ; si le modèle apprend sur des données corrompues, les prédictions seront fausses. Ceci fait, on choisit un modèle et des paramètres, dont on évalue les performances. L'indicateur doit être choisi judicieusement pour refléter correctement l'efficacité du modèle. L'évaluation du modèle doit aussi être reproductible, *i.e.* testée sur de nouveaux jeux de données via des approches comme la validation croisée entre ensembles distincts. Un autre point d'attention est celui du sur-apprentissage (*overfitting* en anglais). Un modèle trop complexe risque de capturer non seulement les relations sous-jacentes, mais aussi le bruit des données, menant ainsi à des résultats biaisés. Tous ces points sont suffisamment importants pour que finalement "Le choix d'un algorithme d'apprentissage automatique particulier [est] moins important que son application et son utilisation correctes et différents algorithmes d'apprentissage automatique appliqués de la bonne manière devraient fournir des résultats cohérents." (Asnicar *et al.*, 2024).

22. Les premières approches de DL datent des années 1940. Le terme regagne en popularité en 2006 suite à l'article de Geoffrey Hinton and Ruslan Salakhutdinov (Hinton *et al.*, 2006)

23. On parle parfois de boite noire dans les modèles les plus complexes dans le sens où, dans les couches cachées, il est souvent difficile de savoir comment les données ont été transformées et classées.

Table I.3.2 – **Méthodes de Machine learning.** Extrait et adapté de (Greener *et al.*, 2022)

Méthode	**Type de données**	**Exemples d'applications**	**Avantages**	**Inconvénients**
Régression Ridge (et LASSO / élastique)	Étiquetées Nombre de caractéristiques fixe	Prédiction de l'expression des gènes en réponse à des antibiotiques	Facile à interpréter Facile à entraîner Bon benchmark	Ne peut pas apprendre des relations complexes entre caractéristiques Sur-apprend avec un grand nombre de caractéristiques
Machine à vecteurs de support		Classification des gènes en fonction de leur fonction	Peut effectuer à la fois la classification et la régression linéaire et non linéaire	Difficile à adapter à de grands ensembles de données
Forêt aléatoire		Identification des mutations génomiques associées à un phénotype	Apprend l'importance de chaque caractéristique pour la prédiction Les arbres de décision individuels sont lisibles par l'humain Moins sensible à l'échelle et à la normalisation des caractéristiques, donc plus facile à entraîner et à ajuster	Moins approprié pour la régression De nombreux arbres de décision sont difficiles à interpréter
Boosting de gradient (ex. XGBoost)		Profilage de l'expression des gènes	Apprend l'importance de chaque caractéristique Arbres de décision lisibles par l'humain Moins sensible à l'échelle et à la normalisation	Peut avoir du mal à apprendre le signal sous-jacent en présence de bruit Moins adapté à la régression
Clustering	Non étiquetées Nombre de caractéristiques fixe	Groupement des gènes bactériens en fonction de leurs profils d'expression dans différentes conditions environnementales	Bon clustering facilement identifiable pour les données de faible dimension Métriques de validation de clustering disponibles	Difficile à appliquer à de grands ensembles de données Les ensembles de données bruités peuvent produire des résultats contradictoires
Réduction de dimensionnalité	Non étiquetée Grand nombre de caractéristiques fixe	Visualisation des relations entre différentes souches bactériennes basées sur leurs génomes	Fournit une représentation visuelle des données Évaluations de l'ajustement souvent disponibles	Difficile de préserver à la fois les différences locales et globales Difficile à appliquer à un grand nombre d'échantillons

Table I.3.3 – **Méthodes de Deep Learning.** Extrait et adapté de (Greener *et al.*, 2022)

Méthode	type de données	**Exemples d'applications**	**Avantages**	**Inconvénients**
Réseau de neurones convolutionnel (CNN)	Données spatiales disposées dans une grille Permet une taille d'entrée variable	Identification des motifs régulateurs dans les séquences d'ADN bactérien	Taille d'entrée variable Apprend des motifs indépendamment de leur localisation	Champ réceptif limité Difficile à entraîner pour des architectures profondes
Perceptron multicouche	Étiqueté Nombre fixe de caractéris-tiques	Prédiction des interactions protéine-protéine	Moins de couches nécessaires que les CNN, donc plus rapide et plus facile à entraîner	Facile à sur-apprendre Grand nombre de paramètres Difficile à interpréter
Réseau de neurones récurrent (RNN)	Données séquentielles Permets une taille d'entrée variable	Prédiction des séquences d'ARN non codant fonctionnel chez les procaryotes	Taille d'entrée variable Les séquences sont fréquentes en biologie	Long temps d'entraînement Exige beaucoup de mémoire
Réseau de neurones convolutionnel sur graphe (GCN)	Données caractérisées par des connexions entre entités Permets une taille d'entrée variable	Modélisation des interactions entre protéines dans les complexes multiprotéiques bactériens	Modélise les interactions complexes Flexible pour différents types de relations	Difficile à interpréter Peut être exigeant en termes de calcul
Autoencodeurs	Données étiquetées ou non Taille d'entrée fixe ou variable	Ingénierie des protéines et des gènes Prédiction de la méthylation de l'ADN	L'espace latent fournit une représentation à faible dimension qui peut être utilisée pour visualiser les données d'entrée Peut générer de nouveaux échantillons, ce qui est utile dans des domaines tels que la conception de protéines	Espace latent spécifique aux données de l'ensemble d'entraînement et peut ne pas être approprié à d'autres ensembles de données Le test des échantillons nouvellement générés nécessite souvent des expériences en laboratoire humide

3.3.3.2 . Application de la génomique comparée pour l'étude des procaryotes

L'application des méthodes de ML à l'étude des génomes procaryotes a permis d'améliorer l'identification et l'annotation des séquences génétiques, notamment grâce à la capacité des algorithmes ML à reconnaître des motifs complexes et à traiter de grandes quantités de données. Plusieurs outils exploitant ces techniques ont émergé, chacun se concentrant sur des aspects spécifiques de l'analyse génomique.

Nucleic Transformer (He *et al.*, 2023) est un outil basé sur l'apprentissage profond conçu pour l'analyse et la classification des séquences d'acides nucléiques. Il utilise une combinaison de mécanismes d'**auto-attention** et de **convolutions** pour identifier des motifs complexes dans l'ADN et l'ARN. L'auto-attention permet au modèle de capturer des relations à longue distance entre les bases nucléiques, tandis que les convolutions sont efficaces pour détecter des motifs locaux récurrents. Son architecture permet d'analyser de grands ensembles de données génomiques tout en maintenant une précision élevée. L'une des analyses réalisables avec Nucleic Transformer est l'identification des promoteurs bactériens. Dans l'article de He *et al.*, Nucleic Transformer est entraîné et testé sur un jeu de données de 5 720 séquences, dont la moitié sont des promotrices. Les prédictions du modèle surpassent celles des outils classiques d'environ 2 %. D'autres applications sont possibles, comme classifier des génomes viraux ou identifier des éléments régulateurs dans les génomes, facilitant ainsi l'étude des réseaux de régulation et des adaptations microbiennes.

ResFinder(Zankari *et al.*, 2012) est un outil initialement développé sans modèle de ML. ResFinder fournissait une base de données de gènes de résistance aux antibiotiques et les séquences étaient alignées (avec BLAST) sur cette base de données. Dans sa version 4.0 (Bortolaia *et al.*, 2020), il intègre la notion de combinaisons entre des résistances et des espèces bactériennes. Cette méthode permet d'obtenir des prédictions fiables pour les organismes et les gènes de résistance bien connus. Cependant, pour ceux qui sont moins étudiés, les prédictions sont moins précises. En 2022, ResFinder intègre des méthodes de ML pour combler ce manque (Florensa *et al.*, 2022). Grâce à des algorithmes d'apprentissage supervisé, il peut identifier des signatures génétiques associées à des résistances spécifiques. L'avantage principal de ResFinder réside dans sa DB de haute qualité, garantissant un apprentissage fiable des modèles. D'autres méthodes utilisaient déjà des modèles de ML pour identifier les gènes de résistance, par exemple le modèle de classification *Random-forest* (Aytan-Aktug *et al.*, 2021), ou des réseaux de neurones (Aytan-Aktug *et al.*, 2020).

Kaiju (Menzel *et al.*, 2016) est un classificateur taxonomique qui utilise des approches de machine learning pour identifier rapidement des microorganismes à partir de données métagénomiques. Contrairement aux méthodes d'alignement classiques, il repose sur une approche fondée sur les k-mers et des techniques de classification, ce qui lui permet d'annoter efficacement des séquences, y compris lorsqu'elles sont courtes et fragmentées. Kaiju exploite des structures algorithmiques avancées, notamment la Burrows-Wheeler Transform (BWT), qui réorganise les séquences pour optimiser la recherche rapide de motifs, et les Maximums Exact Matches (MEMs), qui détectent les plus longues sous-séquences exactes partagées entre un fragment et une base de référence. Ces méthodes permettent d'accélérer l'identification des séquences en comparant directement les fragments d'ADN à une base de données de génomes de référence. Cette stratégie réduit le besoin d'alignement global, rendant l'analyse plus rapide et plus adaptée pour de grands volumes de données. Grâce à cette architecture

hybride combinant algorithmes efficaces et modèles d'apprentissage, Kaiju est particulièrement adapté à l'analyse d'échantillons environnementaux et cliniques.

LookingGlass (Hoarfrost *et al.*, 2022) est un outil utilisant des modèles de DL pour capturer la complexité des relations fonctionnelles et phylogénétiques entre les séquences grâce à une architecture de type réseau de neurones récurrents à mémoire longue et courte (LSTM). Ce type de réseau neuronal est particulièrement adapté aux données séquentielles comme l'ADN, car il permet de modéliser des dépendances à long terme en conservant en mémoire des informations clés sur de longues distances dans la séquence. En exploitant l'apprentissage non supervisé, le modèle est capable d'identifier des relations évolutives au-delà des similarités de séquence directes, permettant ainsi la reconnaissance de fonctions moléculaires dans des séquences non annotées. LookingGlass intègre aussi de l'apprentissage par transfert, *i.e.* qu'il peut transférer l'apprentissage acquis sur une prédiction à d'autres prédictions. LookingGlass a été éprouvé sur l'identification de nouvelles oxydoréductases, la prédiction de températures optimales d'enzymes, ou encore la détection des cadres de lecture dans des fragments d'ADN courts. Cette approche permet également d'étudier une partie inexplorée de la diversité microbienne, *i.e.* les séquences non caractérisées qui constituent la majeure partie du monde microbien (*microbial dark matter*). LookingGlass ouvre ainsi la voie à une annotation plus rapide et plus exhaustive des métagénomes, facilitant la compréhension des réseaux fonctionnels microbiens et leur impact sur les écosystèmes et la santé humaine.

Les outils de prédiction des systèmes de défense contre les phages bénéficient aussi des développements des modèles d'apprentissage. L'outil CRISPRidentify(Mitrofanov *et al.*, 2021) commence par identifier les séquences CRISPR candidates. La seconde phase, extrait différentes caractéristiques parmi les candidates (stabilité des ARNcr, similarité avec les CRISPRs connus, tailles des *spacers*. . ., contenu en nucléotide AT). Enfin, un algorithme de classification, basé sur le modèle ExtraTrees, permet de valider et d'assigner un score de confiance aux candidates. Ce modèle d'apprentissage automatique repose sur un ensemble d'arbres de décision construits de manière aléatoire et indépendante, optimisant ainsi la robustesse et la précision des prédictions. Comparé aux autres outils de détection CRISPR, CRISPRidentify est plus sensible et retourne moins de faux positifs. D'autres outils, comme DeepDefense (Hauns *et al.*, 2024) et DeepPredictor(Hauns *et al.*, 2024) utilisent des méthodes de *deeplearning*, pour la prédiction de systèmes de défense. L'intérêt de ces méthodes basé sur l'apprentissage machine est qu'elles détectent plus de systèmes et même des systèmes inconnus. Toutefois, ces méthodes sont très sensibles aux données d'apprentissage. De plus, elles demandent beaucoup de ressources de calcul, et notamment, dans les exemples présentés, le nombre de génomes utilisés reste plutôt limité par rapport au nombre de génomes disponibles dans les banques.

4 - Pangénomique : état des lieux, enjeux et ambitions

La pangénomique est un domaine d'étude en plein essor, qui a permis d'explorer et d'analyser les génomes procaryotes sous un nouveau point de vue. Mon travail de thèse s'est concentré sur l'analyse et la comparaison de pangénomes. Dans cette partie, je reviendrai d'abord sur l'origine, les concepts et les défis que pose la pangénomique. Je présenterai ensuite les différentes modélisations permettant de représenter les génomes en pangénomique, pour poursuivre sur les méthodes de construction de pangénome. Pour terminer, je développerai les méthodes d'analyse existantes en pangénomique. Cette partie sera aussi l'occasion de faire l'état de l'art des outils en pangénomique et de présenter l'outil PPanGGOLiN sur lequel j'ai pu travailler et que j'ai utilisé dans mes développements de thèse.

4.1 . Origine et concept

Bien que le terme "pangénome" soit utilisé dans des articles avant 2005, en microbiologie, on s'accorde sur une origine du concept de pangénome proposé dans 2 articles fondateurs (Medini *et al.*, 2005; Tettelin *et al.*, 2005). L'idée est de ne pas représenter chaque génome individuellement, mais d'utiliser une structure mathématique permettant de les représenter tous simultanément. Le pangénome représente l'union de toutes les séquences présentes dans un ensemble de génomes. En bioinformatique, la structure, les algorithmes, les méthodes d'analyses des pangénomes, ont constitué un nouveau champ de recherche, la pangénomique.

À partir du pangénome, Tettelin *et al.* proposent de séparer les gènes en 2 catégories, les gènes "*core*" communs à tous les génomes, des gènes "*dispensable*" (ou *accessory*) trouvés dans un sous-ensemble de génomes. En généralisant, le pangénome permet de distinguer l'ensemble des séquences communes à tous les organismes des variations présentes chez certains groupes d'individus, voire spécifiques à un organisme. De ce postulat a émergé l'idée de remplacer les génomes de référence dans les bases de données par des pangénomes de référence (The Computational Pan-Genomics Consortium, 2018). Toutefois, ce changement de paradigme n'a pas encore été opéré, car aucune méthode n'a encore réussi à s'imposer comme solution optimale. Trouver une méthode globale est un défi, car la pangénomique est appliquée dans de nombreux domaines de recherche, pour répondre à une grande diversité de questions.

En 2018, le "Computational Pan-Genomics Consortium" met en avant le rôle de la pangénomique dans le développement de solutions applicatives répondant à des problématiques communes à plusieurs disciplines (The Computational Pan-Genomics Consortium, 2018). En retour, la pangénomique bénéficie des avancées en phylogénie, métagénomique et intelligence artificielle. En phylogénie, les méthodes de comparaison génomique à grande échelle et les techniques de construction d'arbres phylogénétiques ont été intégrées aux approches pangénomiques. Réciproquement, la pangénomique permet une meilleure prise en compte des variations génétiques à l'échelle de l'ensemble des génomes, plutôt que de se limiter à un génome de référence, offrant ainsi une vision plus fine de la dynamique évolutive (Bazinet, 2017). Les données métagénomiques représentent un challenge pour la pangénomique. À partir des métagénomes, le pangénome doit être construit en étudiant les relations de co-occurrence des gènes,

et non les relations évolutives. Ce changement représente un défi, notamment lorsque les lectures sont courtes. Toutefois, la pangénomique permet d'approfondir l'analyse de la diversité génétique des communautés microbiennes, et de mettre en évidence des adaptations communes à l'environnement ou des co-évolutions et des interactions entre les organismes (The Computational Pan-Genomics Consortium, 2018). L'intelligence artificielle joue également un rôle clé en améliorant l'annotation et la prédiction fonctionnelle des gènes. L'apprentissage automatique est appliqué à la pangénomique pour détecter des motifs génétiques pertinents, prédire des phénotypes et identifier des associations entre mutations et traits phénotypiques (Her et Wu, 2018). Ces méthodes, souvent développées pour d'autres disciplines, ont donc favorisé l'essor de la pangénomique en optimisant l'analyse des données, la reconstruction des génomes et l'interprétation des résultats.

La pangénomique représente une solution à l'analyse de grands volumes de données, à l'heure où le nombre de génomes disponibles dans les banques augmente de façon exponentielle. Entre 2006 et 2024, ce ne sont pas moins de 3 500 articles qui référencent le terme [1], dont près de 800 concernant les procaryotes (figure I.4.1).

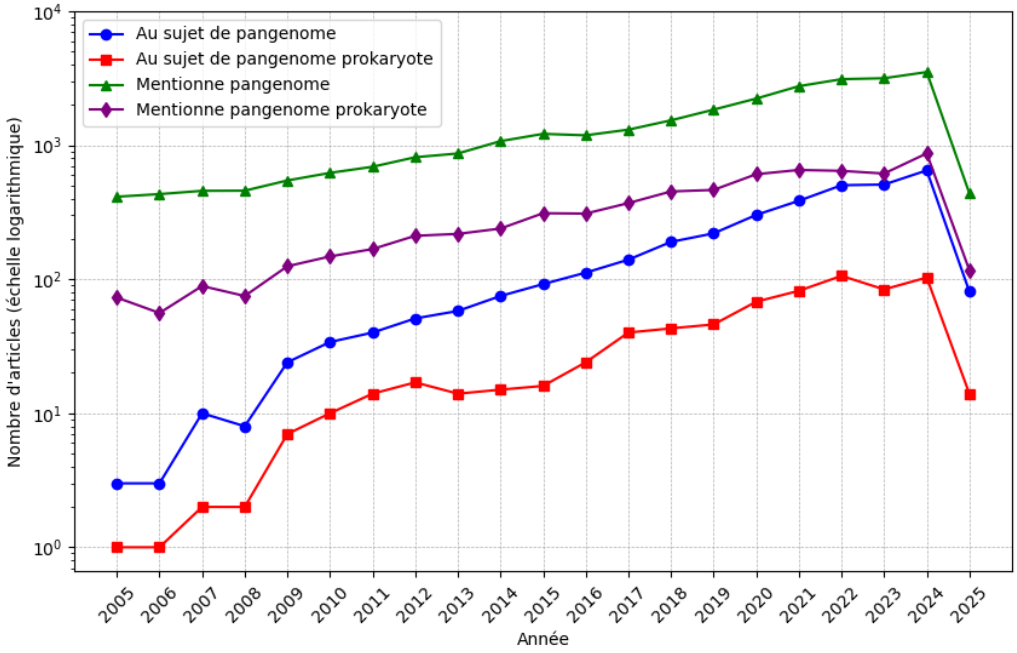


Figure I.4.1 – **Nombre d'articles, référencés dans PubMed, par année, à propos de pangénome du 1er janvier 2004 au 10 février 2025**. La courbe bleue représente le nombre d'articles contenant le terme pangénome dans le titre ou l'abstract : Query=("pan-genome"[Title/Abstract] OR "pangenome"[Title/Abstract] OR "pan-genome"[Title/Abstract]) AND (2004 :2025[pdat]). La courbe rouge limite aux publications concernant les procaryotes : Query=("procaryote"[Title/Abstract] OR "bacteria"[Title/Abstract] OR "archeae"[Title/Abstract]) AND ("pan-genome"[Title/Abstract] OR "pangenome"[Title/Abstract] OR "pan-genome"[Title/Abstract]) AND (2004 :2025[pdat]). La courbe verte représente tous les articles ou le terme pangénome est trouvé : Query=((pangenome) OR (pan genome)) OR (pan-genome) AND (2004 :2025[pdat]). La courbe violette filtre les publications concernant les procaryotes : Query=(((procaryote) OR (bacteria)) OR (archeae)) AND (((pan-genome) OR (pangenome)) OR (pan genome)) AND (2004 :2025[pdat]).

1. Ce chiffre doit être revu à la baisse dû à l'utilisation erronée du terme dans certaines études et une utilisation parfois abusive pour profiter de l'intérêt croissant pour ces analyses

4.1.1 . Modélisation de la croissance des pangénomes

Dans l'article original de Tettelin *et al.* (Tettelin *et al.*, 2005), les auteurs se sont intéressés à la distribution *core/dispensable* en fonction du nombre de génomes de *Streptococcus agalactiae* [2] que contient le pangénome. Ils observent que lorsque le nombre de génomes augmente, la part de *core genome* décroît de façon exponentielle. Ce résultat les amène à modéliser la croissance du *core genomes* selon une équation exponentielle décroissante. Le modèle permet alors d'estimer la taille du *core genome* pour un nombre de génomes en théorie infinie. Il est alors possible d'estimer la taille du *core genome* d'une espèce à partir d'un échantillon de génome.

À partir de ce modèle, il est également possible d'estimer la taille du pangénome, *i.e.*, le nombre de gènes uniques que contient le pangénome. Ils définissent alors 2 types de pangénomes en fonction de l'estimation de la taille : les **pangénomes ouverts** et les **pangénomes fermés**. Les pangénomes sont considérés comme ouverts lorsque le nombre de gènes ajoutés au pangénome augmente pour chaque nouveau génome. Le nombre de gènes est donc théoriquement infini pour un pangénome ouvert avec une infinité de génomes. Les pangénomes fermés, quant à eux, voient le nombre de nouveaux gènes progressivement diminuer lors de l'ajout de nouveaux génomes. La courbe de prédiction permet d'identifier un plateau théorique du nombre maximal de familles que contiendra le pangénome avec un nombre de génomes infinis. Biologiquement, le pangénome ouvert est attendu pour les espèces sympatriques [3] et qui présentent un fort taux de transferts horizontaux, tandis que les espèces vivant dans des niches écologiques ou qui ont une faible capacité d'acquisition de gènes extérieurs vont avoir un pangénome fermé, comme *Staphylococcus lugdunensis* [4] (Argemi *et al.*, 2018).

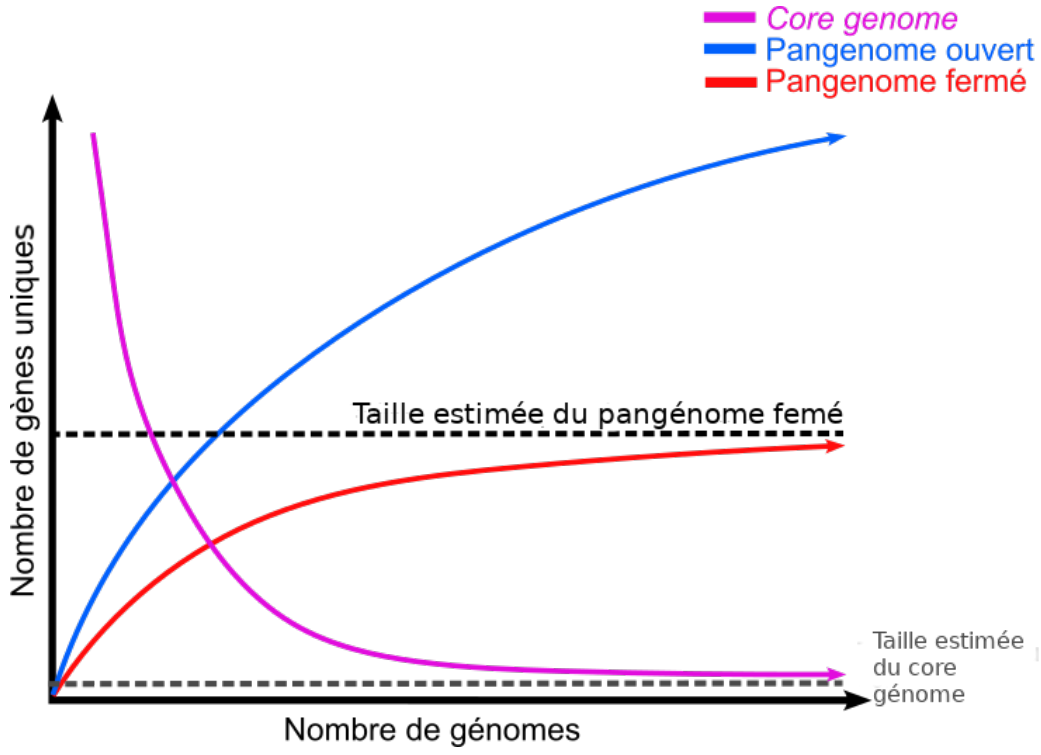


Figure I.4.2 – **Schéma de croissance du pangénome.**

2. Bactérie du microbiote intestinale humain et animal, qui est également associé à des infections graves.
3. Espèces vivant dans le même environnement que d'autres espèces.
4. Bactérie de la flore commensale cutanée et muqueuse, c'est aussi un pathogène opportuniste

Le modèle proposé par Tettelin *et al.* repose sur l'hypothèse que pour un nombre suffisant de génomes, le nombre de nouveaux gènes apportés par un génome devient constant à partir d'un certain nombre de génomes (Tettelin *et al.*, 2005). Cette hypothèse implique alors que la taille du pangénome est infinie. Cette hypothèse sera questionnée par Hogg *et al.* dans leur étude du pangénome de *Haemophilus influenzae* (Hogg *et al.*, 2007). Ils vont alors proposer une modélisation basée sur l'hypothèse que le pangénome est fini. Dans leur modèle, chaque gène est associé à une variable aléatoire de Bernoulli, dont la probabilité correspond à la fréquence du gène dans les génomes. Un génome est ainsi généré en observant ces variables : un gène est présent si l'essai est un succès, sinon il est absent. Bien que certains gènes ne soient pas indépendants en raison de structures comme les îlots génomiques, cette hypothèse est conservée pour simplifier le modèle. Les fréquences réelles des gènes étant inconnues, elles sont modélisées de manière probabiliste en répartissant les gènes en K classes distinctes, chacune ayant une fréquence de présence spécifique. À partir de ce modèle, sur le pangénome de *H. influenzae* avec $K = 7$, la taille du pangénome est estimée à 5 000 gènes (contre 2 800 gènes dans les 13 génomes de base). Ce modèle sera ensuite amélioré par Snipen *et al.* (Snipen *et al.*, 2009), qui proposeront une détermination automatique du nombre de classes K et de la fréquence théorique des gènes pour chaque classe. Les modèles binomiaux proposent une perspective dans laquelle la diversité en gènes est finie et qu'il existe un nombre de génomes suffisamment grand pour que tout le répertoire génique soit connu. Cette vision semble de prime abord logique, car le nombre de combinaisons possibles de nucléotides ou d'acides aminés est fini. Pourtant, on peut y opposer que ce nombre, sans le calculer, semble démesuré et qu'il peut être considéré comme infini. De plus, les génomes évoluent continuellement et de nouveaux gènes apparaissent sans cesse. L'utilisation des modèles binomiaux semble alors plus appropriée à des espèces de niche, isolées et présentant un faible taux de transferts horizontaux.

En 2008, Tettelin *et al.* vont proposer une nouvelle modélisation basée sur la loi de Heaps [5] (Tettelin *et al.*, 2008). On estime le nombre n de gènes distincts, en fonction du nombre N de génomes étudiés, selon la relation :

$$n = kN^{\gamma}, 0 < \gamma < 1, k \geq 1 \tag{4.1}$$

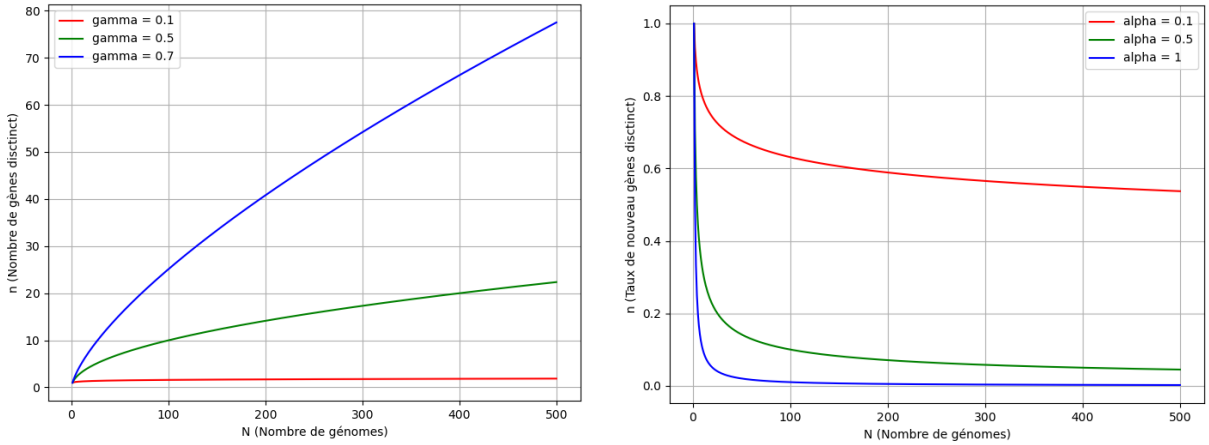
Le paramètre k est une constante de proportionnalité tandis que γ reflète la tendance de la fonction. Ainsi, plus γ est proche de 0 plus la croissance en gènes distincts est lente, et plus γ est proche de 1 plus la croissance est rapide (figure I.4.3a).

Selon la loi de Heap, le nombre de nouveaux gènes découverts diminue à mesure que l'on ajoute des génomes. On peut formuler ceci selon l'équation :

$$p(n) = kN^{(\gamma-1)} = kN^{-\alpha}, \alpha = 1 - \gamma \tag{4.2}$$

Ainsi, sur la figure I.4.3b, lorsque $0 < \alpha < 1$, le taux de nouveaux gènes décroît en ajoutant des génomes, sans jamais être nul. Dans ce cas, le nombre de gènes distincts est croissant. Ce qui implique que si $0 < \alpha < 1$, le pangénome est ouvert. À partir d'un ensemble de génomes, il est possible d'estimer k et α (ou γ) et donc de caractériser si le pangénome est ouvert. Si $\alpha \geq 1$, alors le taux de nouveaux gènes atteint 0, ce qui correspond à un pangénome fermé.

5. Définit de manière empirique en linguistique, cette loi permet de décrire le nombre de mots d'une langue à partir d'un ensemble de documents.



(a) Courbe de croissance selon la loi de Heap

(b) Courbe de raréfaction selon la loi de Heap

Figure I.4.3 – **Évolution du pangénome : visualisation de la croissance et de la raréfaction du contenu génique selon la loi de Heap.**

4.1.2 . Les différents types de pangénomes

Les pangénomes peuvent être divisés en 2 catégories en fonction de l'unité choisie pour les construire. Le premier type, celui présenté par Tettelin *et al.* (Tettelin *et al.*, 2005), utilise les gènes comme unité de base du pangénome (figure I.4.4.B). En regroupant les gènes par homologie (appelé famille de gènes, cf. sous-sous-section 3.1.3.1), il est possible d'obtenir la présence/absence de gènes similaires dans les génomes. Ces pangénomes ont l'avantage d'être moins coûteux en ressources de calcul pour être construits. De plus, ils sont faciles à interpréter, car les gènes sont des unités déjà bien définies et parfois, ils sont même annotés fonctionnellement. Néanmoins, en utilisant les gènes, la méthode d'annotation a un impact important sur le pangénome et il est sensible aux erreurs d'annotation. De plus, les régions non codantes ne sont pas prises en compte dans cette approche. Enfin, les SNPs peuvent passer inaperçus après le regroupement, ainsi que les variants structuraux (SV).

L'autre type de pangénome est basé sur les séquences brutes des génomes. Bien que le terme pangénome n'ait pas encore été employé à l'époque, Chiapello *et al.* (Chiapello *et al.*, 2005) ont proposé une méthode de segmentation des génomes en deux composantes : la "colonne vertébrale", représentant les régions conservées, et les "boucles", qui correspondent aux parties variables. Plus tard, l'outil GenomeMapper (Schneeberger *et al.*, 2009), a explicitement introduit la notion de pangénome de séquence. Son approche repose sur un alignement global des séquences, analysées à travers des k-mers pour différencier les segments conservés des segments variables (figure I.4.4.C,D). Cette approche a l'intérêt de prendre en compte toute la diversité des génomes (codant, non codant, SNPs et SV). Toutefois, la construction de ces pangénomes est plus coûteuse en ressources. De plus, l'interprétation est plus complexe, car le pangénome n'est pas annoté au départ. Pour terminer, certaines méthodes de construction utilisent un génome de référence comme séquence de base (figure I.4.4.C), ce qui peut aussi introduire un biais.

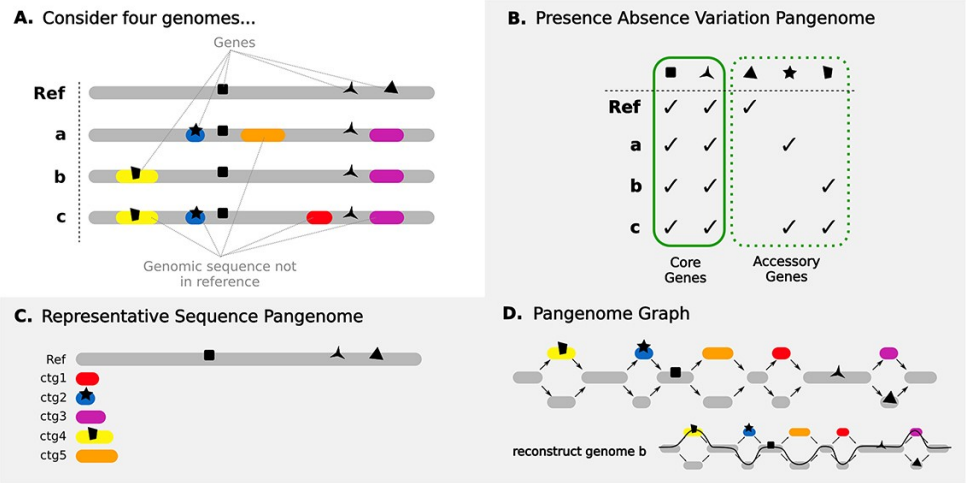


Figure I.4.4 – **Différents types de pangénomes.** Extrait de (Matthews *et al.*, 2024)

4.2 . Pangénome de séquences

4.2.1 . Pangénome basé sur une séquence représentative

Un pangénome basé sur les séquences correspond à un ensemble de génomes dont l'alignement minimise le nombre de régions homologues tout en rendant compte de toute la diversité. L'objectif derrière ces pangénomes est d'obtenir une séquence pangénomique de référence. De façon contre-intuitive (par rapport à la définition "sans-référence" des pangénomes), pour construire ces pangénomes, on utilise une séquence représentative comme base. Toutes les séquences seront alignées à partir de cette base, et les segments non redondants détectés dans au moins un génome seront intégrés à la référence non redondante (NRR, Non-Redundant Reference en anglais). L'ensemble, séquence représentante et NRRs, forme alors la séquence pangénomique de référence.

4.2.1.1 . Méthode de construction

Pour construire ces pangénomes, il faut d'abord identifier une séquence représentative. Les autres séquences, en général des séquences non assemblées (lectures ou *reads* en anglais), sont alignées contre la représentante et les séquences non alignables sont considérées comme des NRRs potentiels. Les NRRs de taille inférieure à 500 pb sont exclues, ainsi que celles dont la similarité avec la représentante est supérieure à un seuil (90 % en général). Les NRRs restantes sont comparées à des bases de données pour retirer tous les contaminants potentiels. De ce schéma général, on peut identifier 4 méthodes différentes pour l'identification des NRRs potentiels :

- **Assemblage de type métagénomique** : les lectures non alignées sur la référence sont regroupées et assemblées *de novo*. Les contigs obtenus sont ajoutés à la séquence représentante. Cette méthode est efficace même avec une faible couverture des lectures.
- **Assemblage itératif** : Dans un premier temps, les lectures non alignées du premier échantillon sont assemblées et ajoutées au génome de référence. Ce génome mis à jour sert ensuite de base pour l'assemblage des échantillons suivants. Ce processus est répété pour tous les échantillons.

- **Assemblage indépendant des *reads* non alignés** : Toutes les lectures non alignées sont séparées par échantillon [6] et assemblées *de novo* indépendamment. Les contigs obtenus sont regroupés selon leur similarité. Dans chaque groupe, un contig référent est identifié et est intégré à la séquence référente. Cette méthode nécessite une couverture d'au moins 10×, pour obtenir des contigs de taille suffisante.
- **Assemblage génomique indépendant** : chaque échantillon est assemblé indépendamment, et les contigs obtenus sont alignés à la référence. Les contigs non alignés sont regroupés par similarité et un contig référent est ajouté à la séquence référente.

Le choix de la méthode dépend du type et de la quantité des données disponibles. Avec une faible couverture (<10×) et un grand nombre d'échantillons, l'approche métagénomique est recommandée, bien qu'elle puisse produire des contigs chimériques. Avec une couverture plus élevée (>10×), l'assemblage indépendant ou l'approche itérative sont préférables. Cette dernière est plus lente, mais facilite l'ajout de nouveaux échantillons. Enfin, si plusieurs assemblages de haute qualité existent déjà, l'assemblage génomique indépendant est la meilleure option. Ces méthodes peuvent être combinées pour optimiser l'utilisation des données disponibles.

4.2.1.2 . Domaines d'application des pangénomes basés sur une séquence représentative

Ces pangénomes sont particulièrement utiles lorsque les données de départ sont des *lectures*. En utilisant ces modèles, il est possible de revenir à une séquence linéaire qui peut être utilisée dans les outils classiques de génomique. De plus, il peut également être utilisé comme étape préliminaire à la construction d'autres types de pangénomes, en réduisant rapidement la redondance dans un sous-ensemble proche de génomes.

L'outil NGSPanPipe (Kulsum *et al.*, 2018) est un pipeline intégré conçu pour l'identification du pangénome à partir de lectures courtes (short reads) issues du séquençage de nouvelle génération (NGS). Contrairement à d'autres méthodes nécessitant des prétraitements des lectures, NGSPanPipe permet une analyse directe des reads bruts pour identifier le pangénome. Il ne génère pas de séquence pangénomique linéaire, mais il permet de reconstruire des contigs à partir des lectures en utilisant un génome de référence. Les contigs obtenus à partir des lectures alignées, permettent de calculer la couverture du génome par rapport au pangénome. Les lectures non alignées sont comparées à des bases de données de *reads* pour identifier de nouveaux *reads*, puis ils sont assemblés en contigs. L'ensemble des contigs (de lectures alignées et non alignées) sont annotés et utilisés pour construire une matrice binaire représentant la présence ou l'absence des gènes dans la séquence de référence.

4.2.2 . Pangénome graphe

Les graphes de séquences sont un modèle de pangénome permettant de visualiser la diversité génomique, qu'elle soit basée sur une séquence de référence ou non. Dans tous les cas, des segments de séquences vont constituer les nœuds du pangénome et les arêtes seront étiquetées par des informations permettant de retrouver le lien entre les segments (comme l'organisation dans les génomes). Ce modèle pangénomique a l'intérêt de représenter toute la diversité, codant et non codant.

6. Ensemble de lectures obtenues simultanément

4.2.2.1 . Méthodes et outils de construction

a. Graphe de variant prédéterminé

La première méthode de construction des graphes de pangénome se base sur l'utilisation d'une séquence référente et d'un fichier contenant les variations connues dans les autres séquences par rapport à cette référence. Cette méthode a l'intérêt de demander peu de ressources, car les variations sont prédéterminées et données en entrée. Toutefois, pour obtenir un graphe fiable et précis, un génome complet de bonne qualité est requis.

L'outil VG (*Variation Graph toolkit*) (Garrison *et al.*, 2018), contient un ensemble d'outils permettant de générer un graphe de variants. À partir de ce graphe, qui peut être assimilé à un graphe de pangénome, il est possible de détecter les variants structuraux (SVs) et les SNPs rapidement. Le graphe est indexé, rendant les recherches et l'alignement plus efficaces, notamment dans l'alignement de lectures ou dans la recherche de variants génétiques (*variant calling*). L'outil a d'abord été développé pour la génomique humaine, mais il est tout à fait possible de l'utiliser avec des génomes procaryotes.

L'outil Minigraph (Li *et al.*, 2020), lui aussi développé pour le variant calling sur le génome humain, propose une méthode demandant moins de ressources que VG. Le graphe est plus léger, sans annotation, permettant de construire des graphes de pangénome de grande taille, en utilisant peu de mémoire de calcul et de stockage. Minigraph permet de capturer les grandes variations génomiques, mais est moins performant sur la détection des SNPs par rapport à VG.

b. Graphe d'alignement multiple

Une méthode, proche de la précédente, est celle basée sur l'alignement multiple des séquences (MSA [7]) entre elles. Cette méthode n'est pas dépendante d'une séquence référente. Le MSA permet de déterminer les variations entre les séquences, ce qui augmente le coût en ressources par rapport au graphe de variants prédéterminé. Toutefois, cette méthode est plus adaptée dans le cas où plusieurs séquences de bonne qualité sont disponibles pour construire le pangénome. En effet, le MSA permet de se passer du biais de la séquence référente dans la construction du graphe et d'ainsi mieux représenter la diversité génomique.

L'outil Harvest (Treangen *et al.*, 2014), permet de comparer des génomes étroitement apparentés. Pour optimiser l'étape d'alignement, il utilise l'outil progressiveMauve (Darling *et al.*, 2010), qui fait un alignement progressif des séquences. Après l'alignement, il identifie le *core genome* dans le pangénome et génère une phylogénie basée sur une matrice des SNPs. Bien qu'étant rapide et efficace, il n'est pas adapté aux génomes très divergents et il ne permet pas d'analyser les éléments mobiles (MGE).

PGGB (Garrison *et al.*, 2024), utilise des algorithmes de graphes de préfixes minimaux (MPHF [8]), pour compresser le graphe et optimiser l'alignement. Il est capable d'identifier et de représenter les SNPs, SV, et les MGEs de manière efficace. PGGB est conçu pour mener des études pangénomiques à grande échelle, prenant en compte de grandes quantités de séquences, ce qui demande des ressources disponibles importantes. De plus, c'est un outil assez complet pour les analyses, ce qui peut le rendre difficile d'accès.

7. cf. paragraphe b.
8. Minimal Perfect Hash Function (MPHF) est une fonction qui associe de manière unique chaque élément d'un ensemble sans collisions et avec un espace mémoire minimal

c. Graphe de De Bruijn

Les graphes de De Bruijn (De Bruijn Graph : DBG) sont des graphes orientés dont les nœuds représentent des k-mers et les arêtes le chevauchement entre le suffixe et le préfixe (de taille k-1) des k-mers (figure I.4.5). Ainsi, en suivant un chemin, il est possible de reconstituer une séquence. C'est pourquoi les DBG sont utilisés dans de nombreuses applications en bioinformatique (assemblage, correction des erreurs de séquençage, métagénomique…) et notamment en pangénomique.

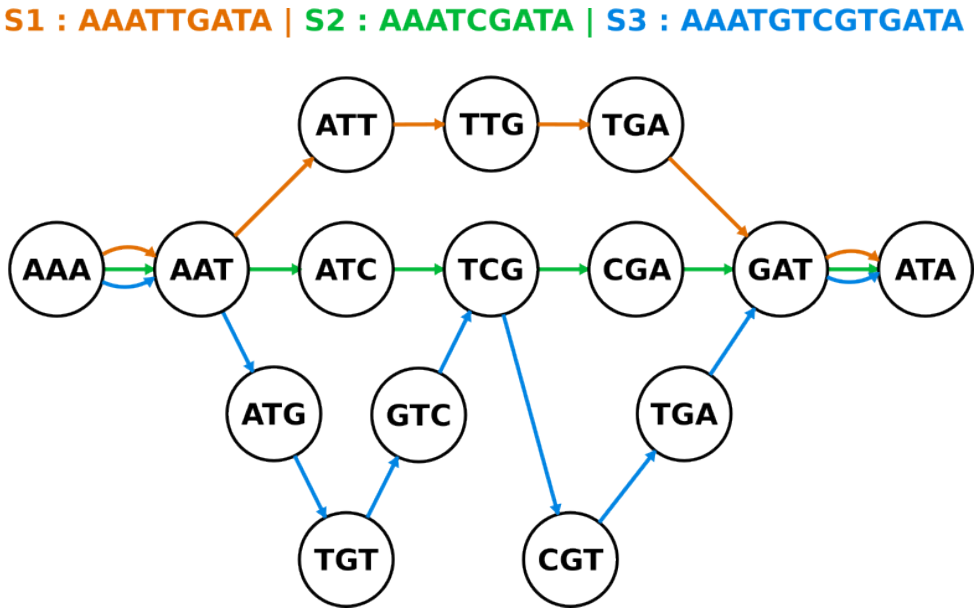


Figure I.4.5 – **Exemple d'un graphe de De Bruijn.** Ici $k = 3$, ce graphe permet de représenter et de reconstruire 3 séquences.

Les DBG, permettent d'avoir une structure compacte des séquences du pangénome. Les nœuds et les arêtes sont colorées en fonction des génomes dans lesquels ils sont retrouvés. Les DBG peuvent être compactés en cDBG, en fusionnant chaque région *core*, *i.e.* chaque suite de nœuds avec une seule arête entre chaque nœud. Ces nouveaux nœuds fusionnés sont appelés "*unitig*" et seront étiquetés avec la séquence combinée des k-mers.

L'une des premières méthodes développées utilisant des DBG est la méthode Cortex (Iqbal *et al.*, 2012), qui construit un DBG "coloré" (les arêtes et les nœuds sont étiquetés par les échantillons dans lesquels ils sont trouvés). À partir de ce DBG coloré, il est possible d'identifier les variants et de les associer à un génotype. Des outils plus récents, comme Bifrost (Holley et Melsted, 2020), améliorent les méthodes de coloration de DBG, permettant d'augmenter le volume de données pris en compte et supportant la mise à jour du graphe. Les auteurs de Bifrost ont notamment appliqué leur méthode sur une collection de plus de 100 000 génomes de *Salmonella* (Luhmann *et al.*, 2021), leur permettant d'identifier des gènes reliés à des îlots de pathogénicité et à une résistance aux fluoroquinolones [9].

9. Classe d'antibiotique utilisée pour traiter les infections bactériennes graves.

SplitMEM (Marcus *et al.*, 2014), permet de construire rapidement et efficacement des cDBG en intégrant une méthode appelée "saut de suffixe" [10], qui permet de construire le cDBG sans passer par un DBG. L'outil permet ensuite de détecter dans l'ensemble des génomes ou dans un sous-ensemble de génomes les régions compressées (appelées *Maximum Exact Matches* : MEMs), correspondant au *core genome*. Cet outil est linéaire en temps et en espace pour identifier le *core genome*, mais ne permet pas de mener d'autres analyses. De plus, la méthode a été testée sur un jeu de 62 génomes de *E. coli*, le caractère linéaire est donc à vérifier sur de plus grands jeux de données.

PanTools (Sheikhizadeh *et al.*, 2016), est un outil complet qui a largement évolué depuis sa publication. Il permet la construction de pangenomes basés sur des cDBG généralisés. PanTools est robuste à l'utilisation de grands volumes de données, que ce soit en temps, en mémoire ou en stockage. Il intègre également des méthodes d'annotation structurale et fonctionnelle, de partitionnement, d'alignement, de phylogénie, d'identification du *core genome* et de visualisation.

DBGWAS (Jaillard *et al.*, 2018), construit également les pangénomes avec des cDBG. Son originalité réside dans l'association de phénotypes (*Genome Wild Association Study* : GWAS). L'intérêt d'utiliser le graphe de pangénome est qu'il n'est pas nécessaire d'utiliser une séquence de référence, contrairement aux approches classiques de GWAS. De plus, les phénotypes ne sont pas associés à des SNPs mais à des sous-graphes, permettant d'extraire des variants plus longs ou plus complexes. DBGWAS intègre une partie visualisation, permettant d'explorer les variants associés au phénotype dans leur contexte génomique pour identifier des régions variables plus larges comme les îlots génomiques.

Le DBG (et ses dérivés) est une structure de données puissante, permettant de calculer, analyser et stocker, rapidement et efficacement, de grandes quantités de données. Néanmoins, ce qui fait la force de cette approche (l'utilisation de k-mer) est aussi sa faiblesse. Le choix de la taille du k-mer va influencer le graphe et donc la découverte des variations. De plus, cette structure est limitée dans l'identification et l'étude des régions répétées. C'est pourquoi des auteurs proposent des méthodes pour lier des informations au DBG (Turner *et al.*, 2018).

4.2.2.2 . Application des graphes de pangénome.

L'utilisation de pangénomes de séquence est très utile à partir de lectures courtes pour améliorer le génotypage. En utilisant le pangénome, contenant des variants connus, on améliore la couverture des lectures et donc on améliore le génotypage de ces lectures. Par rapport aux méthodes utilisant un génome de référence, le pangénome réduit le biais en faveur de la séquence de référence, particulièrement pour les grandes insertions/délétions et les SV. Le pangénome améliore aussi le *variant calling* (VC) en augmentant sa précision, et à partir des DBG de faire du VC sans référence.

Les graphes de séquences sont également utilisés en métagénomique. L'outil MetaKallisto (Schaeffer *et al.*, 2017) utilise notamment une base de données de séquences représentantes qu'il représente sous forme de DBG coloré afin de faire de l'assignation taxonomique et de la quantification de séquences métagénomiques.

10. Le cDBG est relié à des arbres de suffixes, un saut de suffixe permet depuis un suffixe à l'extrémité d'une branche de l'arbre de sauté vers un même suffixe plus proche de la racine. Les sauts se poursuivent jusqu'à atteindre le suffixe le plus proche de la racine. Le chemin restant correspond au chemin le plus court sans ramification, entre la racine et le suffixe.

L'utilisation des graphes de séquences pour les GWAS permet de détecter finement des variations dans les populations associées à un phénotype. Chaguza *et al.* (Chaguza *et al.*, 2020) ont mené une étude sur 909 échantillons de souche hyper virulente de *Streptococcus pneumoniae* (serotype 1). Ils ont pu identifier, grâce à l'outil DBGWAS, des mutations de certaines protéines associées à des phénotypes spécifiques (âge de l'hôte, géographie, structure des populations...). L'utilisation de graphes de pangénome a permis de mener une étude à large échelle, tout en prenant en compte toute la diversité sans nécessiter de référence.

Nom	Méthode	Référence
NGSPanPipe	Séquence représentative	(Kulsum *et al.*, 2018)
Spine	Séquence représentative	(Ozer *et al.*, 2014)
VG toolkit	Variant prédéterminé	(Garrison *et al.*, 2018)
Minigraph	Alignement sur graphe	(Li *et al.*, 2020)
PanVC	Variant prédéterminé	(Norri *et al.*, 2021)
Minigraph-Cactus	MSA	(Hickey *et al.*, 2024)
Harvest	MSA	(Treangen *et al.*, 2014)
PGGB	MSA	(Garrison *et al.*, 2024)
Cortex	graphe de De Bruijn	(Iqbal *et al.*, 2012)
Bifrost	graphe de De Bruijn	(Holley et Melsted, 2020)
SplitMEM	graphe de De Bruijn	(Marcus *et al.*, 2014)
PanTools	graphe de De Bruijn	(Sheikhizadeh *et al.*, 2016)
twoPaCo	graphe de De Bruijn	(Minkin *et al.*, 2017)
DBGWas	graphe de De Bruijn	(Jaillard *et al.*, 2018)
PanVA	Visualisation	(van den Brandt *et al.*, 2024)

Table I.4.1 – **Liste non exhaustive d'outils de pangénomique basés sur les séquences.**

4.3 . Pangenome de gènes

4.3.1 . Généralités et concepts

Les outils basés sur des pangénomes de gènes, aussi appelés *Presence-absence variation pangenome*s (en anglais, PAV), représentent une grande part des outils de pangénomique procaryote. Le génome des procaryotes étant majoritairement codant, et ce type de pangénome étant plus facile à manipuler et à interpréter, de nombreuses études utilisent les gènes comme unité pour construire les pangénomes. Pour construire ces pangénomes, on commence par regrouper les gènes en familles de gènes, puis on partitionne [11] les familles en fonction de leur présence dans les génomes (*core*, *accessory*...).

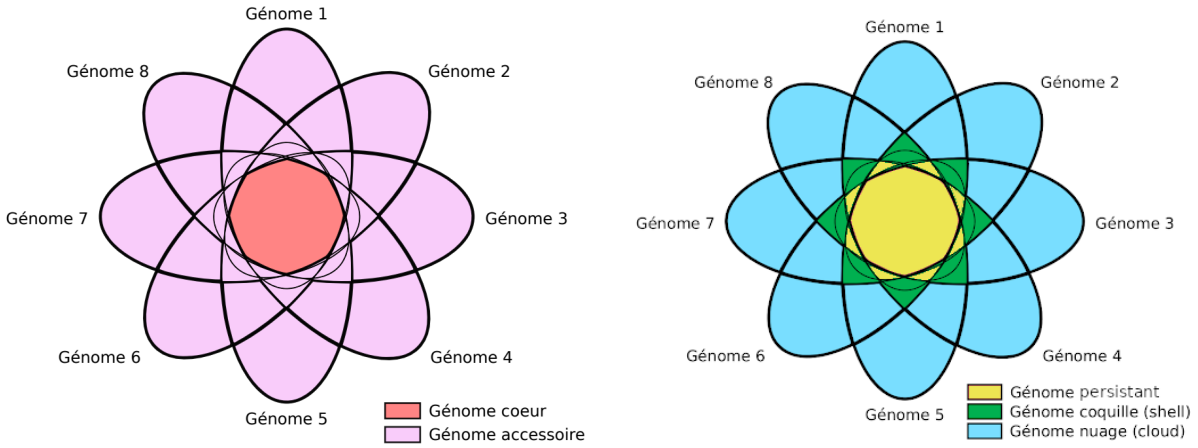
a. Construction des familles de gènes

La construction des familles de gènes consiste à appliquer une méthode de clustering des gènes par similarité, que nous avons vue en sous-sous-section 3.1.3.1. Le choix de la méthode et des seuils appliqués dans le clustering auront un impact important sur le pangénome. L'outil de clustering influencera aussi l'interprétation. Tout d'abord, les outils d'analyse peuvent s'appuyer sur différents niveaux d'information, tels que la séquence nucléotidique ou protéique, la structure tridimensionnelle des protéines ou encore la fonction biologique associée. Le choix de cette base de comparaison influence de manière déterminante le calcul de la similarité, et par extension, l'inférence du caractère homologue. De plus, en utilisant un outil qui construit des clusters (et donc des familles) d'orthologues, comme orthoMCL (Li *et al.*, 2003) ou la base de données COG (cluster of orthologous genes), on retrouvera dans la même famille les gènes ayant suivi les mêmes événements de spéciation. Si l'outil permet de différencier les paralogues des orthologues comme InParanoïd (Remm *et al.*, 2001), il y aura plus de familles que si on prenait en compte uniquement l'homologie. Le choix de la méthode de clustering est donc essentiel.

b. Partitionnement du pangénome

Une fois que les génomes ont été annotés et les familles de gènes construites, les familles de gènes sont partitionnées en fonction de leur présence/absence dans les génomes. Dans les premières analyses, les familles étaient séparées en 2 parties (figure I.4.6a), les familles qui sont présentes dans tous les génomes sont dites "cœur" (*core* en anglais) et les autres sont dites "accessoire" (*accessory* ou *dispensable* en anglais). Cette dichotomie en *core genome* et *accessory genome* est liée au caractère essentiel ou non des fonctions codées par les gènes. Les familles "*core*" sont impliquées dans les processus cellulaires vitaux, ce qui crée une forte pression de sélection de leurs gènes et une forte conservation dans l'ensemble des génomes. À l'inverse, les familles accessoires sont plutôt liées à des adaptations à l'environnement, à un mode de vie.... Leurs gènes sont donc moins soumis à la pression de sélection et donc moins conservés dans les génomes.

11. N.B : Dans la suite, pour ne pas faire de confusion entre le partitionnement des familles et le partitionnement des gènes en famille de gènes, nous utiliserons le terme clustering pour parler de la construction des familles de gènes.



(a) Pangénome bipartie

(b) Pangénome tripartie

Figure I.4.6 – **Partitionnement des pangénomes.** Extrait et adapté de (Gautreau, 2020)

Ce partitionnement du pangénome en 2 parties, bien que largement utilisé, est une vision très limitée de la distribution des gènes dans les génomes, qui peut amener à des erreurs d'interprétation. Il faut d'abord prendre en compte que même si le nombre de génomes disponibles est de plus en plus conséquent, il n'est toutefois pas possible d'avoir l'ensemble des génomes d'une espèce (cf. sous-section 4.1.1), ce qui implique qu'il est plus que probable que des gènes soient identifiés comme accessoires alors qu'ils sont *core* et inversement. De plus, les techniques de séquençage et les outils bio-informatiques ne sont pas infaillibles, et donc une erreur d'assemblage, d'annotation, de regroupement en familles, ou encore l'utilisation de génomes partiels, peut entraîner le mauvais classement d'une famille. Pour répondre à ce problème, Lapierre et Gogarten (Lapierre et Gogarten, 2009) suggèrent de définir un cœur relâché (*soft-core* en anglais), qui contient les familles présentes dans 95 % des génomes [12]. Une autre proposition, de Snippen *et al.* (Snipen *et al.*, 2009) raffinant un modèle proposé par Hogg *et al.* (Hogg *et al.*, 2007), rendrait le nombre de partitions variable en fonction du contenu du pangénome. Cette dernière proposition permet de ne pas utiliser de seuil fixe pour partitionner les familles. En parallèle, Koonin *et al.*, dans une analyse de l'ensemble des génomes procaryotes disponibles en 2008 (Koonin et Wolf, 2008), et Makarova *et al.*, en étudiant l'ensemble des génomes d'archées disponibles en 2007 (Makarova *et al.*, 2007), proposent une vision tripartie du pangénome (figure I.4.6b). Les 2 articles suivent une méthodologie similaire : après une annotation fonctionnelle des génomes, ils comptabilisent le nombre de génomes associés à chaque fonction (COGs pour Makarova et EggNOGs (Jensen *et al.*, 2008) pour Koonin). Les résultats obtenus révèlent une distribution en forme de courbe en U, où chaque extrémité correspond à une catégorie spécifique de fonctions, tandis que la base regroupe une autre catégorie distincte. Ils redéfinissent alors le core genome comme l'ensemble des gènes présents dans la quasi-totalité des génomes. Ce *core genome* relâché et flexible (sans seuil) est aussi appelé *soft-core genome*, ou encore *persistent genome*, dans certains articles pour le différencier du *core genome* strict défini en premier. L'*accessory genome* sera lui divisé en 2, le *cloud genome* correspondant aux gènes partagés par un faible nombre de génomes, et le *shell genome* correspondant aux gènes ayant une fréquence intermédiaire dans les génomes. Ces différents partitionnements, qui ne sont pas incompatibles, sont de plus en plus utilisés, dû au nombre croissant de génomes disponibles.

12. ce pourcentage peut varier en fonction des études.

c. Modélisation et représentation des pangénomes de gènes

Pour représenter les pangénomes de gènes, il est possible d'utiliser une matrice de présence/absence des gènes (figure I.4.4B). Cette représentation permet de rapidement identifier le *core genome* ou de trouver les gènes spécifiques à un génome d'intérêt par exemple. Une seconde représentation est celle du diagramme de Venn (figure I.4.6). À partir du diagramme, on peut rapidement avoir une idée de la proportion de chaque partie, et aussi de la "croissance" du pangénome. Ces 2 représentations ont l'intérêt d'être simples à calculer et à interpréter, néanmoins, lorsque le nombre de génomes devient trop important, il n'est plus possible de les visualiser correctement. De plus, elles se focalisent exclusivement sur le contenu en gènes des génomes, sans fournir d'informations sur leur arrangement ou leur structure organisationnelle.

Afin d'intégrer l'organisation des gènes en plus de leur simple présence, une approche alternative repose sur une représentation où les gènes et leurs relations sont modélisés sous forme de graphe. Dans ce graphe, les familles de gènes constituent les nœuds et les relations de voisinage entre les gènes correspondent aux arêtes. Sur la figure I.4.7, on peut voir que dans cette représentation, plus les familles ont des gènes voisins, plus le poids de l'arête (épaisseur) augmente. Le graphe de pangénome permet alors d'identifier des structures ou des chemins de familles conservées, ou à l'inverse des régions fortement variables.

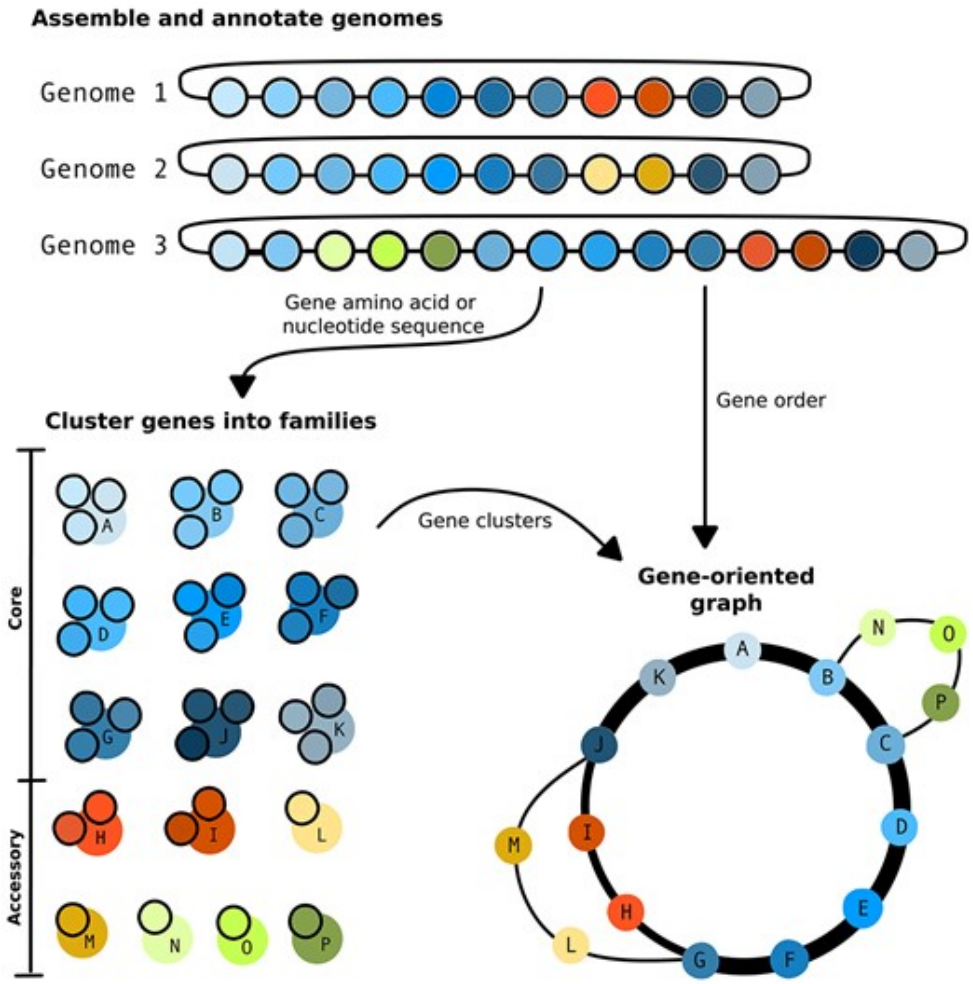


Figure I.4.7 – **Représentation d'un pangénome de gènes sous forme de graphe.** Extrait de (Matthews *et al.*, 2024)

4.3.2 . Méthodes et outils de pangénome de gènes

Le premier outil dédié à la construction de pangénomes de famille de gènes est Edgar (Blom *et al.*, 2009). Disponible en ligne, ce n'est pas un outil à proprement parler, mais plutôt une ressource de résultats d'analyse de pangénome. Dans sa méthode, Edgar clusterise les gènes en familles d'orthologues, en utilisant le BBH [13]. Les génomes étant relativement proches dans les analyses, les paralogues récents pourraient être regroupés avec des orthologues. Les familles sont donc raffinées en utilisant un système de score pour valider les BBH. Dans sa version actuelle, l'outil permet d'identifier le *core genome* et l'*accessory genome*, rechercher des synténies conservées, de construire des arbres phylogénétiques, ou encore d'annoter fonctionnellement les gènes à partir de bases de données de référence.

Le premier outil permettant la construction de pangénome en ligne de commande est PGAP (Zhao *et al.*, 2012). Le premier intérêt de PGAP, est qu'il permet à l'utilisateur de construire un pangénome avec ses propres génomes. PGAP construit lui aussi des familles d'orthologues. À partir de ces familles, il propose différentes analyses, comme la courbe de raréfaction, le profil du pangénome [14], l'identification du *core genome*, l'analyse de variants. Une interface PGAP-X (Zhao *et al.*, 2018) rend l'outil plus accessible et améliore la visualisation des résultats.

PanOCT (Fouts *et al.*, 2012) est un autre outil disponible en ligne de commande. Il propose aussi un clustering en famille orthologues, similaire à EDGAR, mais améliore l'algorithme en ajoutant l'information de contexte conservé (*Conserved Gene Neighborhood*, CGN). Les gènes orthologues ont tendance à conserver leur organisation génomique dans les espèces proches contrairement aux espèces éloignées et aux gènes paralogues (Huynen et Bork, 1998; Rocha, 2008). En combinant le BBH et le CGN, les familles de gènes orthologues sont de meilleure qualité. L'outil PanACEA (Clarke *et al.*, 2018) récupère les familles de PanOCT et permet de visualiser le pangénome, mais aussi de l'annoter, notamment avec des informations liées à l'antibiorésistance.

PanFunPro (Lukjancenko *et al.*, 2013), propose une méthode originale pour construire des familles de gènes homologues, en se basant sur l'annotation fonctionnelle des protéines. Pour annoter les protéines, il utilise des profils HMM de différentes bases de données de domaines protéiques. En définissant l'homologie selon la fonction plutôt que la séquence, cette approche propose un cadre d'analyse alternatif, privilégiant les similarités fonctionnelles des protéines et offrant ainsi une vision complémentaire aux méthodes traditionnelles.

Roary (Page *et al.*, 2015) est l'outil de pangénomique certainement le plus utilisé et le plus célèbre. Sa popularité vient de sa capacité à générer des pangénomes de façon rapide en demandant peu de ressources par rapport à ses concurrents de l'époque. Pour ça, il utilise un algorithme CD-Hit pour grouper les séquences proches avant d'aligner les séquences représentantes entre elles avec BLASTP. Les familles sont construites avec l'algorithme MCL et raffinées en utilisant les informations de colocalisation. Roary est aussi un des premiers programmes à représenter les pangénomes de gènes sous forme d'un graphe dans lequel les nœuds sont les gènes et les arêtes représentent une relation de voisinage dans les génomes.

13. *Bidirectional Best Hit*
14. Le nombre de gènes (y) présents dans x génomes.

D'autres outils vont reprendre ce modèle de graphe de gènes, comme Panaroo (Tonkin-Hill *et al.*, 2020). À partir du graphe, il corrige les erreurs d'annotation et d'assemblage. Il dispose également de plusieurs modules d'analyse : GWAS, SV, phylogénie et visualisation du pangénome. Panaroo est un outil performant, mais demande des ressources de calcul relativement importantes et une expertise bioinformatique plus importante que d'autres outils. De plus, en nettoyant le graphe, il est possible qu'il élimine des évènements évolutifs récents, et donc ne pas être adapté à des taxons dans lesquels les taux de HGT sont élevés par exemple. Panakeia (Beier et Thomson, 2022) est un outil reposant aussi sur un modèle de graphe de pangénome, mais propose une analyse plus "universelle". L'outil propose entre autres l'identification de chemins particuliers dans le graphe, correspondant à des structures biologiques, comme les îlots génomiques.

Les outils présentés jusqu'ici, bien que non limités théoriquement, sont généralement appliqués à la construction de pangénomes au niveau de l'espèce. Des méthodes, comme RIBAP (Lamkiewicz *et al.*, 2024), proposent de construire des pangénomes au niveau du genre. RIBAP construit un pangénome en utilisant ROARY à 95 % d'identité. En parallèle, il utilise MMSeqs2 pour aligner l'ensemble des gènes. Le résultat de l'alignement est ensuite utilisé pour raffiner les familles pour construire des familles homologues au niveau du genre. De cette manière, la partie de *core genome* est plus importante. Avec ce nouveau partitionnement, RIBAP propose de construire un arbre phylogénétique des souches présentes dans le pangénome.

L'utilisation de pangénomes de gènes est bien adaptée à la génomique comparée des procaryotes. Leur simplicité de calcul et d'interprétation les rend accessibles pour tous les utilisateurs. Néanmoins, cette simplicité est liée à une vision gène centrée, qui ne prend pas en compte les régions non codantes. Des outils, comme Piggy (Thorpe *et al.*, 2018), permettent de pallier ce problème en proposant une analyse complémentaire à un pangénome de gène généré par Roary. Ainsi, la complémentarité des outils permet d'avoir une étude plus complète.

En 2020, PPanGGOLiN (Gautreau *et al.*, 2020) introduit une stratégie de partitionnement, s'appuyant sur un algorithme de *machine learning* pour classifier les gènes du pangenome. Cette approche repose sur une analyse des relations de voisinage et du clustering, éliminant ainsi la nécessité de fixer des seuils stricts. Par conséquent, PPanGGOLiN permet d'identifier de manière dynamique les trois composantes du pangenome — *persistent genome*, *shell genome* et *cloud genome* — sans présupposer de leur répartition.

4.3.3 . Analyses à partir de pangénomes de gènes

Les pangénomes de gènes sont utilisés pour mener des études de phylogénomique. En 2020, Gaba *et al.* (Gaba *et al.*, 2020) s'intéressent aux archées de la classe des Halobacteria, des archées extrêmement halophiles [15]. Ils vont construire un pangénome à l'aide de l'outil GET_HOMOLOGUE (Contreras-Moreira et Vinuesa, 2013), un outil similaire à EDGAR. À partir de ce pangénome, ils ont pu identifier les gènes *core* des gènes *accessory*. Sur la base de ce partitionnement, ils ont pu mettre en évidence un fort taux de transferts horizontaux au sein de la classe des Halobacteria. Ils ont alors construit une phylogénie basée sur le pangénome et sa partition, mettant en évidence une évolution étroite entre l'ordre des Natrialbales et celui des Halobacteriales, suggérant alors l'existence d'un superordre les regroupant. Plusieurs méthodes permettent d'ailleurs

15. Capable de vivre dans des milieux à haute concentration en sel

d'identifier les régions où les transferts horizontaux de gènes se produisent préférentiellement. Parmi elles, la méthode panRGP(Bazin *et al.*, 2020) permet de détecter les régions de plasticité génomique (RGP), *i.e.* des segments du génome échangés entre différentes souches par transfert horizontal ou perdus de manière différentielle selon les lignées. Cette approche repose sur l'analyse du graphe de pangénome et de sa partition pour identifier les régions variables. Elle permet également d'identifier des spots, en repérant les familles de gènes persistantes flanquant les RGP. Une autre méthode, panModule(Bazin *et al.*, 2021), vise quant à elle à détecter des groupes de familles de gènes variables dans le pangénome, qui sont organisés en blocs de synténie au sein des génomes. Ces deux méthodes sont intégrées à la suite logicielle PPanGGOLiN et exploitent le graphe de pangénome généré par cette dernière.

Dans les exemples cités, les études étaient centrées sur des groupes taxonomiques et donc les génomes appartenaient au même taxon. Pourtant, comme nous l'avons déjà vu, l'étude d'environnements et les données métagénomiques sont étroitement liées aux études pangénomiques. Vera-Ponce de León *et al.* (Vera-Ponce de León *et al.*, 2024), produisent un atlas génomique du microbiote des saumons. Pour caractériser les espèces présentes et construire le pangénome, ils utilisent l'outil mOTUpan (Buck *et al.*, 2022). mOTUpan permet de clusteriser les métagénomes à un fort niveau d'identité (95 %), définissant ainsi des unités taxonomiques opérationnelles métagénomiques (mOTUs, *metagenomic Operational Taxonomic Unit* en anglais). Chaque mOTU peut être associé à une espèce (ou un taxon) et donc permettre de séparer les gènes appartenant à la même espèce dans une même mOTU. L'intérêt de mOTUpan est qu'il utilise un modèle bayésien pour estimer le génome *core* et le génome accessoire. Ce modèle prend en compte la complétion (souvent faible) des métagénomes et améliore la qualité du partitionnement sur ces données. À partir du pangénome, les auteurs ont pu identifier 14 ordres bactériens différents, représentant 35 genres distincts, ils ont également identifié 29 nouvelles espèces encore non répertoriées.

Les méthodes de machine learning peuvent aussi être utilisées pour analyser les pangénomes. Dans leur article, Kavvas *et al.* (Kavvas *et al.*, 2018) proposent d'utiliser des méthodes de machine learning pour identifier des gènes de résistance aux antibiotiques dans 1 595 souches de *Mycobacterium tuberculosis*. Le pangénome est construit à partir de familles homologues, et les gènes connus pour être des gènes de résistance aux antibiotiques sont annotés. En utilisant des méthodes de machine learning (SVM, MI, ANOVA…), les auteurs ont pu identifier 24 nouveaux gènes de résistance, de plus le signal de certains gènes déjà connus est plus fort que mesuré précédemment. Les auteurs ont poursuivi en analysant la structure des nouveaux gènes et de leurs protéines, ainsi qu'en menant une étude géographique de la répartition des gènes pour établir des liens entre population hôte et résistance. Il faut néanmoins rappeler une nouvelle fois que la découverte de ces nouveaux gènes doit être vérifiée et que leur identification n'est possible que sur la base d'une base de données de gènes connus.

Nom	Méthode	Référence
EDGAR	matrice présence/abscence	(Blom *et al.*, 2009)
PGAP	matrice présence/abscence	(Zhao *et al.*, 2012)
PanOCT	Clustering d'orthologue	(Fouts *et al.*, 2012)
GET_HOMOLOGUES	Clustering d'orthologue	(Contreras-Moreira et Vinuesa, 2013)
PanACEA	Visualisation	(Clarke *et al.*, 2018)
PanFunPro	Familles de profil fonctionnelle	(Lukjancenko *et al.*, 2013)
Roary	Graphe de famille de gène	(Page *et al.*, 2015)
Piggy	Analyse région intergénique	(Thorpe *et al.*, 2018)
Ptolemy	Graphe de famille de gène	(Thorpe *et al.*, 2018)
PIRATE	Graphe de famille de gène	(Bayliss *et al.*, 2019)
Panaroo	Graphe de famille de gène	(Tonkin-Hill *et al.*, 2020)
PPanGGOLiN	Graphe de famille de gène	(Gautreau *et al.*, 2020)
PanACoTA	Graphe & Phylogénie	(Perrin et Rocha, 2021)
Panakeia	Graphe de famille de gène	(Beier et Thomson, 2022)
PanPhlAn	Graphe de données métagénomique	(Scholz *et al.*, 2016)
MSPminer	Graphe de données métagénomique	(Plaza Oñate *et al.*, 2019)
mOTUpan	Graphe de données métagénomique	(Buck *et al.*, 2022)
Pan-Tetris	Visualisation	(Hennig *et al.*, 2015)
PanViz	Visualisation	(Pedersen *et al.*, 2017)
Panache	Visualisation	(Durant *et al.*, 2021)

Table I.4.2 – **Liste non exhaustive d'outils de pangénomique basés sur les familles de gènes**

4.4 . Conclusion sur les pangénomes et éléments de réflexions

Nous avons présenté une grande variété de méthodes et d'outils de pangénomique (*cf.* tableaux I.4.1 et I.4.2) et exploré plusieurs champs d'application des pangénomes. Cependant, plusieurs défis majeurs restent à relever en pangénomique.

Un premier défi concerne la représentation et la visualisation des pangénomes. En effet, ceux-ci doivent être visualisables par l'œil humain tout en intégrant l'ensemble de l'information génomique. Différentes approches ont été développées pour répondre à cette problématique, notamment des outils de visualisation interactifs. À titre d'exemple, Pan-Tetris(Hennig *et al.*, 2015) permet de visualiser et de modifier la composition des groupes de gènes grâce à une technique inspirée du jeu Tetris. PanViz(Pedersen *et al.*, 2017) facilite la comparaison des génomes individuels aux pangénomes avec une navigation basée sur l'ontologie des gènes. PanVa(van den Brandt *et al.*, 2024), utilisant les pangénomes de PanTools(Sheikhizadeh *et al.*, 2016), propose une approche centrée sur la variabilité structurale, permettant une analyse flexible des pangénomes en tenant compte des variations génomiques complexes. Enfin, PANACHE(Durant *et al.*, 2021) propose une visualisation linéarisée des pangénomes, affichant la présence ou l'absence des blocs de séquences sous forme de navigateur génomique.

Un second défi crucial est celui du stockage et de la gestion des données pangénomiques. Contrairement aux génomes individuels, généralement stockés sous forme de texte et reliés à des bases de données, les pangénomes sont des structures plus complexes qui ne peuvent être stockées sous un format linéaire. Les BD doivent donc permettre un accès rapide et efficace aux informations, tout en étant mises à jour régulièrement pour suivre l'augmentation du volume des données génomiques. Des BD comme panKB (Sun *et al.*, 2025), permettent d'avoir accès à des informations sur des pangénomes précalculés, reliés à des métadonnées comme : les publications de référence, les informations sur l'origine des génomes…Cependant, le pangénome en lui-même n'est pas disponible et seuls les résultats d'analyse sont disponibles.

Enfin, un défi essentiel concerne la construction du pangénome. Le choix des méthodes influence fortement le résultat final, mais les données d'entrée jouent également un rôle fondamental. L'objectif étant de représenter au mieux la diversité génomique, il est important de maximiser cette diversité tout en évitant les biais, tels que la surreprésentation de génomes pathogènes dans les bases de données. Un équilibre est nécessaire : trop de variations dans les données d'entrée peuvent complexifier l'analyse, en rendant par exemple les graphes de pangénome difficilement exploitables. La qualité des génomes, des annotations et des bases de données, sont également des facteurs déterminants pour garantir la robustesse des analyses pangénomiques.

Pour répondre à ces défis, les méthodes et les outils de pangénomique sont en constante évolution. Avec ces progrès, divers domaines de la microbiologie voient progressivement s'intégrer des analyses pangénomiques en routine. Cette intégration repose en partie sur des plateformes qui rendent ces outils accessibles et exploitables par la communauté scientifique. Par exemple, MicroScope(Vallenet *et al.*, 2020) permet de construire des pangénomes procaryotes à l'aide de PPanGGOLiN (Gautreau *et al.*, 2020) et d'identifier les régions de plasticité génomique (RGP) grâce à la méthode panRGP(Bazin *et al.*, 2020). MicroScope constitue alors un point d'intersection entre les données génomiques, la production de pangénomes et leur utilisation effective, contribuant ainsi à relever progressivement les défis liés à la visualisation, au stockage et à la construction des pangénomes.

CHAPITRE II DU GÉNOME AU PANGÉNOME

Avec l'essor de la pangénomique, de plus en plus d'outils ont été développés. En 2020, le LABGeM propose son outil de construction et de partitionnement de graphe de pangénome : PPanGGOLiN (Gautreau *et al.*, 2020), développé dans le cadre de 2 thèses. La première, menée par Guillaume Gautreau, a conduit à la création du graphe de pangénome et à son partitionnement. La seconde, réalisée par Adelme Bazin, a contribué au développement des méthodes précédentes, avec également l'analyse des parties variables du graphe. Ces travaux ont conduit au développement de la méthode panRGP (Bazin *et al.*, 2020) pour l'identification des régions de plasticité génomique (RPG) et des spots d'insertion ; suivie de la méthode panModule (Bazin *et al.*, 2021) pour la prédiction de modules conservés.

Au cours de ma thèse, j'ai intégré de nouvelles méthodes dans PPanGGOLiN, notamment pour la recherche de contextes génomiques conservés. Puis j'ai initié et participé à l'amélioration du logiciel et de son environnement : (réusinage de code, maintenance, documentation…) aboutissant à la distribution d'une version 2.0 de PPanGGOLiN. Ces développements seront présentés dans ce chapitre.

1 - La suite logicielle PPanGGOLiN : construction et analyse d'un pangénome

Les premières approches pangénomiques se fondaient sur une dichotomie entre le génome *core*, regroupant les gènes conservés dans toutes les souches d'une espèce, et le génome *accessoire*, constitué des gènes variables (Tettelin *et al.*, 2005). Toutefois, avec l'augmentation du nombre de génomes disponibles, les pangénomes intègrent une quantité croissante de séquences. Cependant, ces séquences peuvent être incomplètes ou contenir des erreurs de séquençage, entraînant une absence apparente de certains gènes dans des génomes où ils devraient pourtant être présents. Par conséquent, un gène qui devrait faire partie du génome *core* peut être mal classé comme appartenant au génome *accessory*. Ce phénomène contribue à une réduction apparente de la taille du génome *core*. Pour compenser cet effet, une version "relâchée" du génome *core* a été proposée. Ce *core* relâché, aussi appelé **génome *persistent***, correspond aux gènes les plus fréquents dans les génomes, *i.e.* dont la présence est supérieure à un seuil minimal défini pour être considéré comme conservé. Par exemple, Roary (Page *et al.*, 2015) applique un seuil de 95 %, bien que celui-ci doive être ajusté en fonction du contexte biologique spécifique (Tonder *et al.*, 2014).

Cependant, cette bipartition du pangénome devient de plus en plus limitante à mesure que le nombre de génomes séquencés augmente. Les travaux de Makarova *et al.* (Makarova *et al.*, 2007), suivis par ceux de Koonin et Wolf (Koonin et Wolf, 2008), montrent que la répartition des clusters de familles homologues, COGs (Tatusov *et al.*, 2000) et EggNogs (Jensen *et al.*, 2008) (respectivement), en fonction du nombre d'organismes, suit une structure en trois catégories. C'est pourquoi, pour mieux refléter la diversité génétique, certains outils, comme micropan (Snipen et Liland, 2015), proposent une classification en trois catégories : (*i*) le génome *persistent*, (*ii*) le génome *shell*, correspondant aux gènes présents dans un sous-ensemble de génomes, et (*iii*) le génome *cloud*, qui regroupe les gènes rares ou spécifiques à certaines souches. Cette trichotomie offre une vision plus nuancée de la distribution des gènes, mais elle reste limitée lorsqu'il s'agit d'espèces à forte hétérogénéité génomique (Moldovan et Gelfand, 2018).

Pour surmonter ces limitations, PPanGGOLiN (Gautreau *et al.*, 2020) propose une approche alternative combinant trois éléments clés : (*i*) l'utilisation des matrices de présence/absence des familles de gènes, (*ii*) l'identification automatique du nombre optimal de partitions et (*iii*) la prise en compte du voisinage génomique des gènes. Ainsi, PPanGGOLiN construit un graphe de pangénome, où les nœuds représentent les familles de gènes et les arêtes reflètent leur relation de voisinage dans les génomes. Ce graphe est ensuite partitionné, grâce à un algorithme statistique et de machine learning, sans seuil et sans limites du nombre de parties. Ces parties sont reclassées en 3 catégories de génome : le *persistent* (analogue au *core* relâché), le *shell* et le *cloud*. Cette approche permet de regrouper les familles en fonction de leur distribution dans les génomes et de leur contexte génomique.

À partir de ce graphe de pangénome partitionné, PPanGGOLiN intègre 2 méthodes pour l'analyse du pangénome. La première méthode panRGP (Bazin *et al.*, 2020) permet d'identifier efficacement les portions variables du pangénome, correspondant aux îlots génomiques, aux plasmides et aux gènes différentiellement perdus dans l'évolution. De plus, en identifiant ces **régions de plasticités génomiques** (RGPs), panRGP peut identifier des bordures de familles persistantes, partagées par plusieurs RGPs, correspondant à des **spots d'insertion**. PanRGP offre ainsi une solution prenant en compte toute la diversité génomique dépassant les limites des méthodes traditionnelles de génomique comparative.

La seconde méthode, panModule (Bazin *et al.*, 2021), exploite le graphe de pangénome pour identifier des **modules** génomiques. Ces modules correspondent à un ensemble non chevauchant de familles de gènes variables cooccurrentes et colocalisées.

Ces méthodes ont été intégrées dans la suite logicielle PPanGGOLiN, un outil open source écrit en Python et C, compatible avec les systèmes Linux et MacOS. PPanGGOLiN fonctionne en ligne de commande et met à disposition une série d'analyses permettant d'exploiter le graphe de pangénome pour une meilleure compréhension de la diversité et de la dynamique des génomes. Grâce à sa modularité, chaque commande permet d'exécuter une analyse spécifique et de croiser les résultats obtenus afin d'affiner l'interprétation des données génomiques. En tant que logiciel open source, il est librement accessible sur GitHub (**?**) et peut être facilement installé via Conda (**?**), facilitant ainsi son intégration dans divers environnements bioinformatiques.

1.1 . La méthode PPanGGOLiN

Pour construire un graphe de pangénome partitionné, PPanGGOLiN suit une liste d'étapes présentées sur la figure II.1.1. En données d'entrée, PPanGGOLiN prend un ensemble de génomes annotés [1]. Les gènes des génomes seront clusterisés [2] en familles de gènes homologues. Ces familles seront utilisées comme nœuds pour construire le graphe de pangénome et la relation de voisinage entre les gènes des familles dans les génomes comme arêtes. À partir du graphe de pangénome et de la matrice de présence/absence des familles dans les génomes, PPanGGOLiN applique des algorithmes statistiques et de machine learning pour déterminer le nombre de partitions et assigner les familles à une partie (*persistent*, *shell* ou *cloud*). Le partitionnement est reporté sur le graphe de pangénome pour obtenir le graphe de pangénome partitionné.

Le logiciel PPanGGOLiN intègre une ligne de commande correspondant à chacune de ces étapes, et un workflow pour construire et partitionner automatiquement le pangénome à partir des fichiers de génome. Le pangénome est enregistré dans un fichier au format HDF5 assurant une compréhension et une structuration des données efficaces. Une fois le graphe de pangénome obtenu, plusieurs commandes permettent de réaliser des analyses automatiques, avec des rapports sous forme de tableaux ou de figures.

1. Il est conseillé d'utiliser au moins 15 génomes ayant des variations de contenue en gènes pour partitionner correctement le graphe de pangénome.

2. Rappel : nous utilisons le mot cluster pour parler du regroupement des gènes en familles par similarité et partitionnement pour l'assignation des familles à une partie dans le pangénome (*core*, *shell*, *cloud*...)

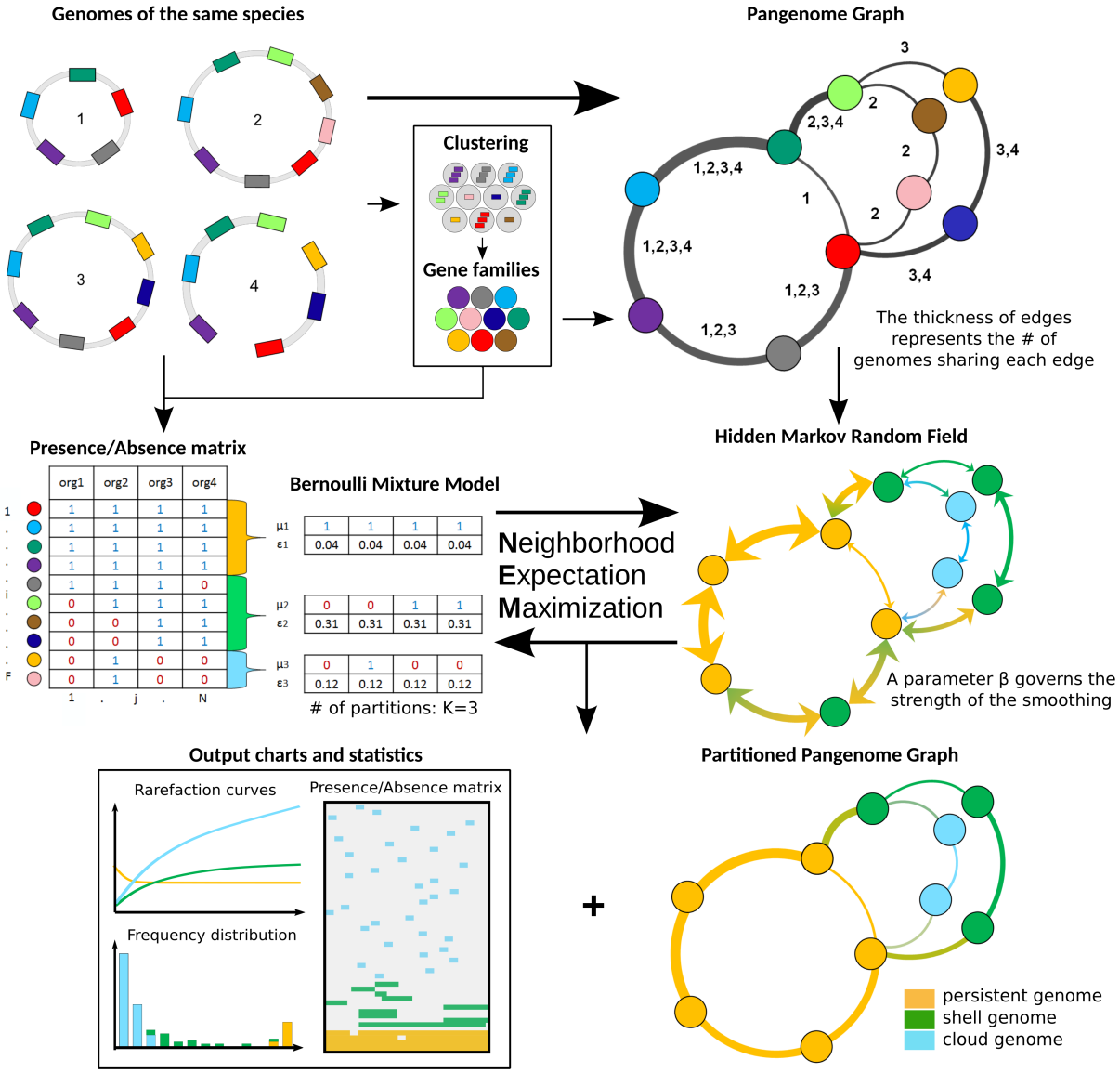


Figure II.1.1 – **Aperçu général de la méthode PPanGGOLiN.** L'exemple comprend 4 génomes annotés, dont les gènes sont représentés par des rectangles de couleurs. Une même couleur indique que les gènes sont homologues. Les gènes sont clusterisés en familles représentées par des cercles de couleur correspondant aux gènes qu'elles contiennent. Dans le graphe, elles constituent les nœuds et sont reliées par des arêtes représentant leur relation de voisinage dans les génomes. Le poids des arêtes représente le nombre de génomes où le voisinage existe. Parallèlement, les familles de gènes sont codées sous la forme d'une matrice de présence/absence qui indique pour chaque famille si elle est présente ou non dans les génomes. Le pangénome est ensuite divisé en K partitions (K = 3 dans cet exemple) en estimant les meilleurs paramètres de partitionnement par un algorithme statistique. PPanGGOLiN renvoie un graphe du pangénome partitionné où les partitions sont superposées au graphe de voisinage. En outre, de nombreux tableaux, graphiques et statistiques sont fournis par le logiciel. Extrait de (Gautreau, 2020)

1.1.1 . Construction du graphe de pangénome

PPanGGOLiN utilise comme données d'entrée un jeu de génomes procaryotes de la même espèce [3]. Ces génomes peuvent déjà être annotés et donc contenir les gènes (format GFF ou GBFF) ou alors contenir uniquement la séquence nucléique. Dans le second cas, PPanGGOLiN prédira les gènes en utilisant la méthode Prodigal (Hyatt *et al.*, 2010). Il détecte également les ARNs à l'aide des logiciels ARAGORN et Infernal (Laslett et Canback, 2004; Nawrocki et Eddy, 2013).

PPanGGOLiN est un outil reposant sur les familles de gènes pour la construction du pangénome. Il regroupe les gènes en familles en utilisant l'outil MMSeqs2 (Steinegger et Söding, 2017), qui applique un workflow automatique de clustering. Ce processus permet de regrouper les gènes homologues en familles avec des seuils (par défaut) de 80 % d'identité (Fedrizzi *et al.*, 2017; Iraola *et al.*, 2017; Batty *et al.*, 2018) et 80 % de couverture (Sjölander *et al.*, 2011) sur les deux séquences, en s'appuyant sur l'algorithme de clustering *Connected Component*. Bien que cette méthode soit efficace, elle présente une limite dans la gestion des fragments de gènes. Pour y remédier, une étape de défragmentation est intégrée afin de réassigner ces fragments à leurs familles d'origine. Cette correction repose sur un réalignement des séquences représentantes de chaque famille avec MMSeqs2, en conservant les mêmes paramètres d'identité et de couverture, mais en adaptant la couverture à la plus petite des deux séquences comparées. Enfin, un algorithme de clustering, prenant en compte la taille des familles et des séquences, est appliqué pour regrouper les fragments avec leur séquence complète correspondante.

Le graphe de pangénome est ensuite construit à partir des familles (nœuds du graphe) et de leur relation de voisinage dans les génomes (arêtes). Deux nœuds sont connectés si les familles de gènes correspondantes contiennent au moins une paire de gènes adjacents dans un génome. Les arêtes sont étiquetées avec les identifiants des génomes correspondants et pondérées par la proportion de génomes partageant ce lien.

1.1.2 . Partitionnement du graphe

Pour partitionner le graphe, PPanGGOLiN commence par classer les familles de gènes en K partitions ($K \geq 2$) à partir d'une matrice binaire indiquant la présence (1) ou l'absence (0) d'un gène dans un génome donné. Le partitionnement s'appuie sur un modèle de mélange de Bernoulli (BMM) estimé via l'algorithme d'Expectation-Maximization (EM), appelé BinEM. En plus des catégories *persistent* et *cloud*, $K-2$ partitions définissent le génome *shell* [4].

Dans ce modèle, chaque famille de gènes suit une distribution de mélange où les proportions d'appartenance aux partitions sont estimées. Pour éviter le sur-ajustement, une contrainte initiale impose une dispersion homogène au sein d'une même partition. Chaque famille de gènes est affectée à une partition unique en fonction de sa probabilité d'appartenance, assignée automatiquement si cette probabilité dépasse 0,5 ; sinon, elle est placée dans la partition de fréquence intermédiaire (si K=3, cette partition correspond au *shell*). Pour choisir K, le modèle exécute plusieurs partitionnements et cal-

3. Cette description correspond à l'utilisation prévue par PPanGGOLiN. Néanmoins, PPanGGOLiN a déjà été utilisé avec des génomes appartenant à une même Famille ou encore sur des familles de phages (Pfeifer *et al.*, 2021)

4. Le nombre de partitions K peut être fixé par l'utilisateur, sinon il est déterminé automatiquement

cule le critère ICL (*Integrated Completed Likelihood*) [5], qui équilibre la qualité du modèle et la complexité, permettant ainsi de sélectionner la valeur optimale de K.

A partir du graphe de pangénome, l'algorithme de *machine learning* **NEM** (Neighboring Expectation - Maximization) (Ambroise *et al.*, 1997) améliore le partitionnement en intégrant les informations de contiguïté dans les génomes via un champ de Markov caché, favorisant le regroupement de gènes voisins dans la même partition. Ce modèle lisse la classification en prenant en compte la structure du graphe, bien que la détermination du nombre optimal de partitions K repose d'abord sur BinEM.

Enfin, le partitionnement obtenu est reporté sur le graphe de pangénome. Ce graphe partitionné peut être visualisé et exploré via un fichier de sortie, généré par PPanGGOLiN et visualisable dans l'outil Gephi (Bastian *et al.*, 2009). Un exemple de graphe partitionné, correspondant au pangénome de *Acinetobacter Baumannii*, est présenté sur la figure II.1.2. Sur le graphe, on peut voir de longs chemins de familles persistantes (orange), entrecoupés de portions de familles variables *shell* (vert) ou *cloud* (bleu). L'exploration et l'analyse de ces régions variables (encadré de la figure II.1.2, par exemple) est particulièrement intéressante, puisque c'est dans ces zones que l'on va retrouver les régions de plasticité génomique.

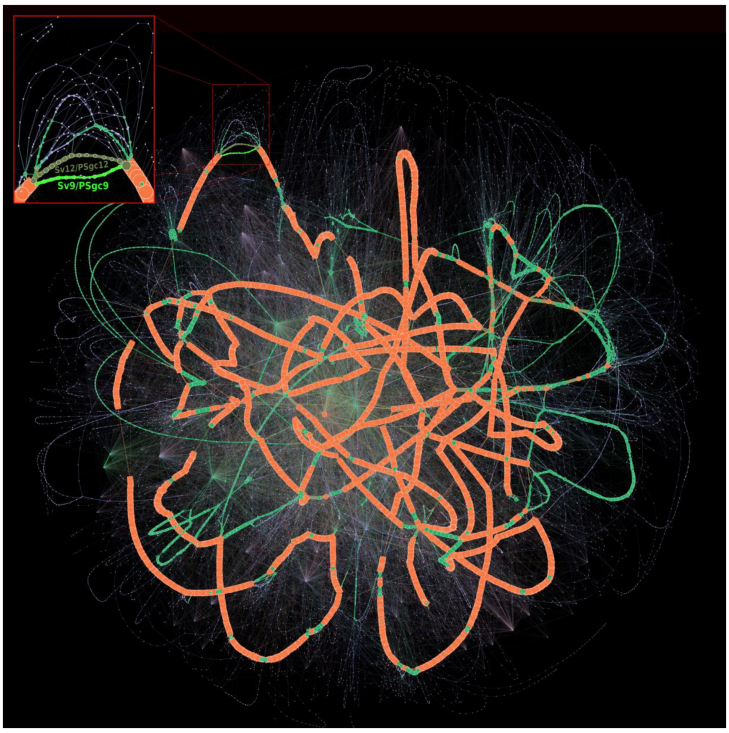


Figure II.1.2 – **Graphe de pangénome de *Acinetobacter baumannii*.** Le pangénome est construit avec PPanGGOLiN à partir de 3117 génomes de *A. baumannii*. Les arêtes reliant les familles *persistent*, *shell* et *cloud* sont respectivement colorées en orange, vert et bleu. Les connexions entre familles de gènes appartenant à différentes partitions sont représentées par des couleurs mélangées. Pour améliorer la lisibilité, les familles comptant moins de 20 gènes ne sont pas affichées, bien qu'elles représentent 84,68 % des nœuds (principalement des familles à un seul gène). Un encadré dans le coin supérieur gauche zoome sur une région ramifiée où plusieurs chemins alternatifs *shell* et *cloud* sont présents dans l'espèce. Cette région est impliquée dans la biosynthèse du principal antigène polysaccharidique de *A. baumannii*. Les deux chemins les plus fréquents (Sv12/PSgc12 et Sv9/PSgc9) sont mis en avant en kaki et vert fluorescent.

5. Le critère ICL correspond à un critère BIC (*Bayesian Information Criterion*). Le BIC estime la vraisemblance du modèle à partir du nombre d'observations dans l'échantillon et du nombre de paramètres libres du modèle. L'ICL ajoute une penalité basé sur l'entropie moyenne estimé.

1.2 . La méthode PanRGP

Le logiciel PPanGGOLiN intègre une méthode pour l'identification des régions de plasticité génomique et la prédiction des spots d'insertion, la méthode panRGP (Bazin *et al.*, 2020). Cette méthode repose sur l'utilisation du graphe de pangénome partitionné pour ne pas avoir à comparer l'ensemble des génomes (ce qui est déjà fait dans la construction du pangénome) et donc d'être plus efficace que les outils de génomique classique.

1.2.1 . Identification des régions de plasticité génomique

Les régions de plasticité génomique sont des objets génomiques : la méthode de prédiction décrite dans la figure II.1.3 est appliquée à chaque génome du pangénome, sur lesquels on a projeté les partitions du pangénome.

Pour chaque gène du génome, on calcule un score s_g de manière séquentielle le long des contigs qui est égal au score du gène précédent auquel on applique une pénalité ou une prime en fonction de la partition du gène. Si le gène est *shell* ou *cloud*, on applique une prime correspondant à la somme de 2 constantes, v qui favorise le gène variable et ϵ pour égaliser les résultats quelque soit le sens de lecture du contig. Si le gène est *persistent*, on applique une pénalité p. Pour pénaliser la succession de gènes *persistent*, la pénalité est exponentiellement proportionnelle au nombre (n) de gènes *persistent* successifs précédents (p^n).

Si le génome est linéaire, le premier gène de chaque contig aura un score de 0. Dans le cas des séquences circulaires, un premier gène est choisi et un score initial de 0 lui est attribué, puis l'algorithme assigne un score à tous les autres gènes. À la fin du contig, le gène avec un score de 0 est réévalué. Ce nouveau score sert de base pour exécuter une nouvelle boucle de calcul de score qui s'arrête dès qu'un gène obtient un score de 0 ou jusqu'à atteindre le dernier gène du contig.

Pour détecter les RGPs, l'algorithme parcourt chaque contig à la recherche du gène ayant le score le plus élevé, à condition qu'il dépasse le seuil s_{min} (fixé par défaut à 4). Ce gène marque la fin initiale de la RGP. Ensuite, les gènes situés en amont sont ajoutés progressivement jusqu'à rencontrer un gène dont le score est nul. La région est alors considérée comme une RGP uniquement si elle contient un nombre de nucléotide supérieur au seuil minimal attendu l_{min} (fixé par défaut à 3 000 bp). Enfin, le score de la RGP est recalculé en repartant du dernier gène détecté, en appliquant la même méthode de calcul que précédemment.

1.2.2 . Prédiction des spots d'insertion

Les spots correspondent à des régions où de nombreux éléments se sont insérés au cours de l'évolution, et donc ce sont des régions fortement variables. Aux extrémités de chaque RGP, on sélectionne un nombre *c* de gènes persistants non multigéniques, convertis en familles de gènes pour rendre l'algorithme indépendant de l'orientation.

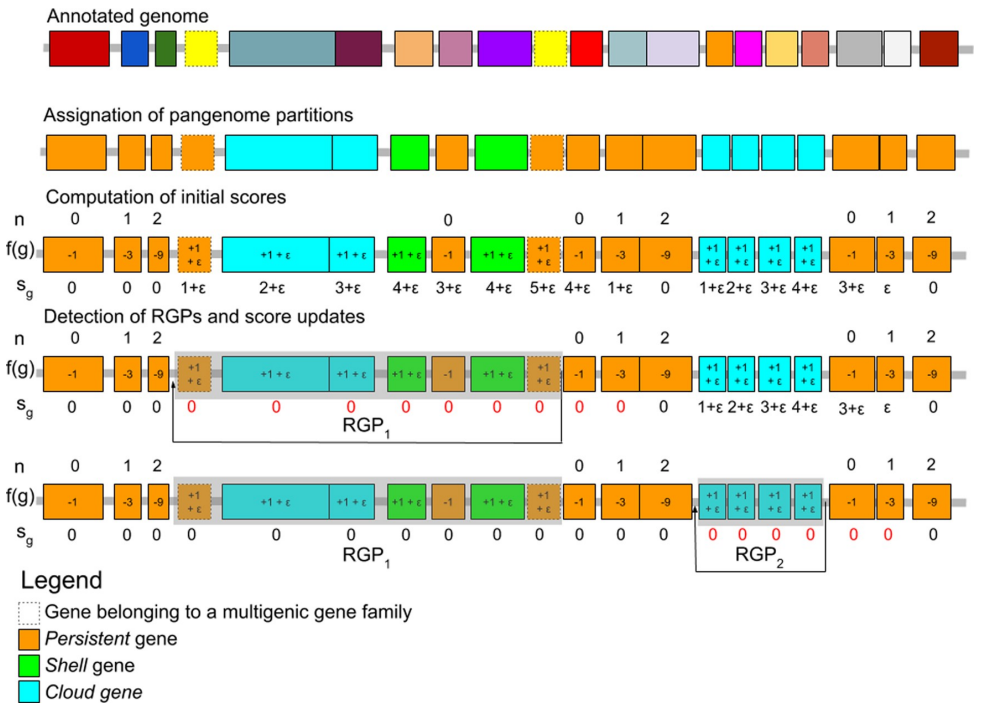


Figure II.1.3 – **PanRGP : vue d'ensemble de la méthode de détection des RGP.** Les boîtes représentent les gènes codant des protéines : les couleurs orange, vert et bleu correspondent respectivement aux partitions *persistent*, *shell* et *cloud*. Les boîtes en pointillés signalent les gènes appartenant à des familles multigéniques. Dans cet exemple, deux RGP sont détectées : RGP1 avec un score de 5 et RGP2 avec un score de 4. La valeur n associée à chaque gène correspond au nombre de gènes persistants consécutifs en aval. Les valeurs $f(g)$ représentent le résultat d'une fonction utilisée pour calculer le score de chaque gène (s_g). Ici, les paramètres par défaut p et v sont fixés respectivement à 3 et 1.

Un graphe $G(V, E)$ est construit, où chaque nœud représente les bordures d'une RGP et chaque arête relie deux nœuds ayant des ensembles de familles de gènes similaires. Deux bords sont considérés comme proches si leurs *e* premières familles sont identiques ou si leurs ensembles ordonnés se chevauchent d'au moins *o* familles. Lorsque tous les bords de deux RGPs correspondent, une arête est ajoutée, et les composantes connexes du graphe définissent les spots, auxquels sont associées les RGPs correspondantes.

Les paramètres par défaut sont $c = 3$, $e = 1$, $o = 2$. Chaque spot est évalué selon plusieurs métriques, comme le nombre de RGPs et de familles de gènes. Les RGPs sans c gènes persistants consécutifs ne sont pas prises en compte, car elles sont soit incomplètes (bords de contigs), soit correspondent à des plasmides.

1.3 . La méthode PanModule

Dans les génomes, et notamment dans les GIs et les spots, des groupes de gènes sont supposés avoir suivi la même histoire évolutive. Ces ensembles de gènes, cooccurrents et colocalisés, sont appelés des **modules conservés**. La méthode panModule (Bazin *et al.*, 2021), intégrée à PPanGGOLiN, permet de détecter ces modules depuis le graphe de pangénome partitionné.

Tout d'abord, un pangénome est reconstruit à partir des génomes, mais celui-ci intègre, en plus des arêtes de voisinage directes, des arêtes entre les familles séparées dans les génomes d'un espace intergénique inférieur à t gènes. Lorsque $t > 1$, cela équivaut à appliquer une **fermeture transitive** partielle sur les graphes de génomes, ce qui permet de relier des familles même si leurs gènes ne sont pas directement adjacents. Cette approche est particulièrement utile lorsqu'un module génomique est interrompu par l'insertion d'un gène (comme une séquence d'insertion) ou lorsque des gènes ont été perdus à la suite d'un événement de délétion ou de pseudogénisation.

Les arêtes vont ensuite être filtrées selon deux **coefficients de similarité de Jaccard** :

$$J(v_i, e_{i,j}) = \frac{w_{e_{i,j}}}{w_{v_i}}, \quad J(v_j, e_{i,j}) = \frac{w_{e_{i,j}}}{w_{v_j}} \tag{1.1}$$

où :

- w_{v_i} et w_{v_j} correspondent au nombre de gènes associés aux familles v_i et v_j, respectivement.
- $w_{e_{i,j}}$ représente le nombre de paires de gènes ayant servi à créer l'arête $e_{i,j}$ entre les nœuds v_i et v_j.

Un seuil s est défini comme la similarité minimale de Jaccard nécessaire pour considérer une arête comme appartenant à un module. Si les deux coefficients de Jaccard vérifient :

$$J(v_i, e_{i,j}) > s \quad \text{et} \quad J(v_j, e_{i,j}) > s \tag{1.2}$$

alors l'arête est conservée ; sinon, elle est supprimée du graphe. De plus, les nœuds correspondants à des familles de gènes présentes dans moins de m génomes sont également retirés.

Après cette phase de filtrage, les **composantes connexes** du graphe sont extraites à l'aide d'un algorithme de **parcours en largeur (BFS)** modifié. Une composante est considérée comme un **module prédit** si elle contient au moins trois nœuds appartenant aux familles *shell*, *cloud* ou multigéniques persistantes, selon la classification PPanGGOLiN. En revanche, les modules constitués de **familles persistantes non multigéniques** ne sont pas pris en compte. Ces familles correspondent généralement à des régions synténiques conservées dans la majorité des génomes étudiés, avec peu ou pas d'événements de réarrangement.

Les modules prédits peuvent ensuite être associés aux autres analyses de PPanGGOLiN, en identifiant sur quelles RGPs et dans quels spots sont retrouvés les modules.

2 - Évolution de PPanGGOLiN : présentation de la version 2

PPanGGOLiN est un outil dont le développement a commencé (sur GitHub) en 2018. Avec plus de 2000 commits [1] actuellement, et 7 ans de développement, le code a connu de nombreuses évolutions. De plus, ces modifications sont l'œuvre de plusieurs développeurs qui se sont succédés et qui ont collaboré activement au projet.

Le développement de nouvelles méthodes, l'ajout de nouveaux outils, ou encore simplement le débogage devenait de plus en plus lourd. De plus, dans le cadre de ma thèse, PPanGGOLiN allait être utilisé comme outil et certaines fonctionnalités devaient être améliorées. C'est avec cet objectif en tête que Jean Mainguy (ingénieur au LABGeM), Guillaume Gautreau (chercheur INRAE), Adelme Bazin (ingénieur de recherche), Alexandra Calteau (chercheuse CEA), David Vallenet (directeur de recherche CEA) et moi-même avons développé et proposé en janvier 2024 une version 2 de PPanGGOLiN. Cette nouvelle version contient de nouvelles fonctionnalités pour l'analyse des pangénomes, mais aussi de nombreuses améliorations techniques.

Dans ce chapitre, nous reviendrons sur les changements majeurs apportés dans la version 2 de PPanGGOLiN. D'autres changements plus anecdotiques ne seront pas abordés. Néanmoins, une des améliorations concerne l'écriture des notes de version qui sont plus détaillées. Toutes les modifications sont alors visibles dans ces notes sur le GitHub de PPanGGOLiN `https://github.com/labgem/PPanGGOLiN/releases`.

2.1 . Nouvelles fonctionnalités et amélioration méthodologique

2.1.1 . Developpement de nouvelles méthodes d'analyse

2.1.1.1 . Recherche de contexte génomique

a. Définition

Le contexte génomique (*Genomic Context*, GC) désigne l'organisation spécifique des gènes au sein d'un génome. Durant l'évolution, cette organisation va subir des pressions de sélection. Si elle se maintient conservée, on peut postuler que les gènes du GC sont impliqués dans un même processus biologique. C'est pourquoi, la recherche de GCs conservés dans les génomes est utilisée pour prédire la fonction de gènes inconnus en les associant à d'autres dont la fonction est connue. C'est le principe du coupable par association (Aravind, 2000). Rechercher un GC (ou un sous-ensemble de ce dernier) dans les génomes permet ainsi d'identifier des processus biologiques et des dérivés dans les génomes, comme le fait antiSMASH (Medema *et al.*, 2011), en détectant spécifiquement les clusters de gènes biosynthétiques (BGCs).

1. ajout, suppression ou modification de code validé et marqué dans le système de gestion de version (ici Git)

Une des méthodes de recherche de GC dans les génomes repose sur le voisinage des gènes dans les séquences. On considère des gènes comme fonctionnellement liés s'ils sont régulièrement retrouvés à proximité les uns des autres dans divers génomes, même sans être directement adjacents. Ce type de signal permet de détecter des associations fonctionnelles conservées, même entre des gènes non homologues.

Cette approche est particulièrement intéressante dans le cadre des graphes de pangénome, comme ceux produits par PPanGGOLiN. Le graphe de pangénome intègre déjà l'information sur le voisinage direct des gènes à travers l'ensemble des génomes étudiés. Cela permet d'extraire efficacement le contexte génomique directement depuis le graphe. De plus, cette approche permet de capturer toute la diversité génomique d'un ensemble d'organismes : non seulement les gènes conservés dans toutes les souches, mais aussi les gènes accessoires, présents uniquement dans certaines d'entre elles. Ainsi, l'extraction du contexte génomique à partir d'un pangénome permet de mieux refléter la diversité biologique réelle tout en accélérant la prédiction fonctionnelle des gènes.

b. Méthode de recherche de contexte génomique

L'algorithme de prédiction (figure II.2.1) s'inspire de celui proposé dans panModule (Bazin *et al.*, 2021). Cependant, contrairement à ce dernier, qui vise à extraire l'ensemble des contextes génomiques conservés, notre approche se focalise sur l'identification précise des contextes associés à un ensemble de protéines cibles. L'objectif est d'extraire, au sein du pangénome, les familles de gènes conservées dans le contexte de ces protéines.

La première étape consiste à aligner les gènes cibles (*target*) avec les familles de gènes du pangénome. Cet alignement est effectué à l'aide de MMSeqs2 (Steinegger et Söding, 2017), avec un seuil de 80 % d'identité et 80 % de couverture entre les séquences protéiques des gènes et celles des familles. Les familles correspondant aux séquences alignées sont alors étiquetées comme *target* (représentées en bleu et orange dans la figure II.2.1).

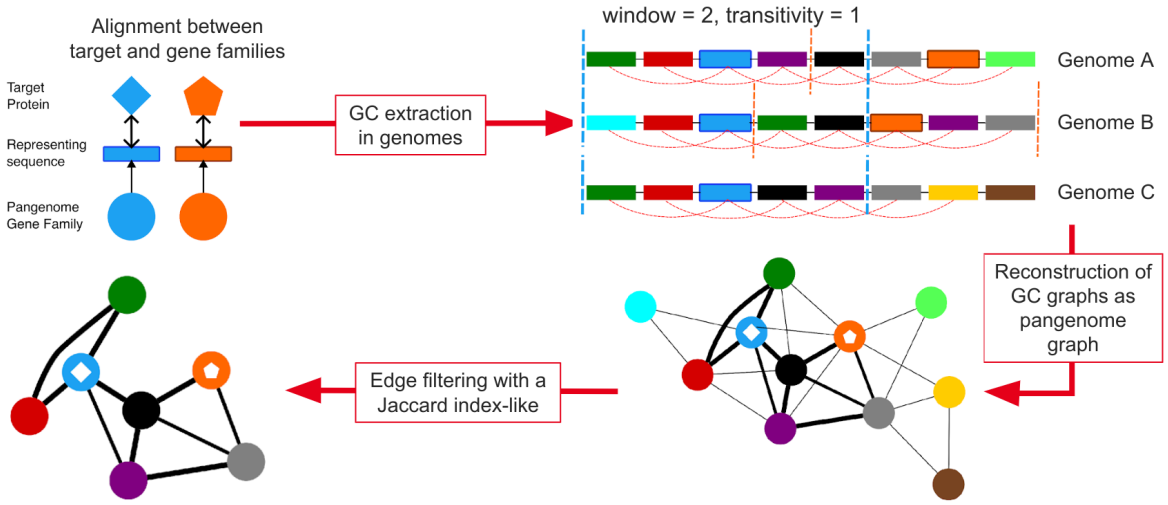


Figure II.2.1 – **Méthode de recherche du contexte génomique dans un graphe de pangénome.**

À partir de ces familles *target*, nous reconstruisons un sous-graphe du pangénome. Ce sous-graphe intègre non seulement les arêtes de voisinage direct, mais aussi des arêtes de transitivité reliant des familles situées à une distance t. Ainsi, l'algorithme capture non seulement les relations directes, mais aussi celles entre familles avoisinantes. Par exemple, dans la figure II.2.1, avec $t = 1$, des arêtes de transitivité sont créées entre le gène bleu et les gènes vert et noir, ainsi qu'entre le gène violet et les gènes rouge et gris.

Pour limiter la taille du sous-graphe, un paramètre *window* est introduit. Il contrôle le nombre de gènes adjacents — de part et d'autre de la *target* — pris en compte pour la création des arêtes de transitivité.

Le sous-graphe obtenu forme alors une composante connexe intégrant l'ensemble des relations de voisinage jusqu'à la distance t. Un filtre de Jaccard (cf. équation 1.1) est ensuite appliqué pour conserver uniquement les arêtes les plus pertinentes.

Enfin, nous extrayons toutes les composantes connexes contenant au moins une famille *target*, représentant ainsi les contextes génomiques conservés autour des protéines cibles.

2.1.1.2 . Metadonnées

Dans le cadre de l'ajout de nouvelles fonctionnalités à PPanGGOLiN, une première avancée que j'ai développée concerne l'ajout de **métadonnées** à l'ensemble des éléments du pangénome, incluant les gènes, contigs, génomes, familles, arêtes, RGPs, spots et modules. Le format attendu des métadonnées est particulièrement flexible, l'utilisateur fournit un fichier tabulé, ne nécessitant au minimum que l'identifiant de l'objet concerné dans le pangénome. L'utilisateur peut ainsi associer des métadonnées de tout type, sans restriction de format. De plus, chaque métadonnée est liée à une source spécifique, ce qui permet à un même objet d'en contenir plusieurs, issues de différentes sources d'annotation. Pour assurer une gestion optimisée et performante, ces métadonnées sont directement enregistrées dans le fichier HDF5 du pangénome. Bien qu'elles ne soient pas directement exploitées pour l'analyse du pangénome, elles sont intégrées aux sorties générées par PPanGGOLiN afin de faciliter l'interprétation et l'exploration des résultats.

2.1.1.3 . Projection

Pour faciliter l'exploration des résultats, une **nouvelle sortie de visualisation des données** a été développée en collaboration avec Jean Mainguy. Celle-ci permet de générer, pour chaque génome, un fichier JSON compatible avec Proksee (Grant *et al.*, 2023). Grâce à cette fonctionnalité, il est désormais possible de visualiser un génome, sous forme circulaire (figure II.2.2), où les éléments du pangénome ont été intégrés, notamment les parties persistantes et variables, ainsi que les RGPs, spots et modules. Proksee permet également de lancer des analyses supplémentaires sur les génomes, comme CARD (Alcock *et al.*, 2023) pour annoter les gènes de résistance aux antibiotiques, CRISPRCasFinder (Couvin *et al.*, 2018) pour la prédiction des systèmes CRISPR-Cas, ou encore Phigaro (Starikova *et al.*, 2020) qui permet d'identifier les régions de prophages.

L'intégration de cette nouvelle sortie est d'autant plus intéressante dans une autre nouvelle fonctionnalité de PPanGGOLiN : la **projection** des résultats du pangénome sur un génome externe. En effet, il est désormais possible de comparer un nouveau génome au pangénome déjà calculé. Pour cela, les gènes du génome externe sont alignés aux séquences référentes des familles de gènes en utilisant MMSeqs2 (Steinegger et Söding, 2017). Les gènes dont l'alignement dépasse un seuil défini d'identité et de couverture sont alors associés à une famille existante, tandis que les autres gènes forment chacun une nouvelle famille (singleton) attribuée à la partition du *cloud*. Ainsi, l'ensemble des gènes du génome externe se voit assigner une partition, ce qui permet d'appliquer les algorithmes de prédiction des RGPs et d'associer les spots et les modules au génome étudié.

Enfin, cette fonctionnalité de projection prend également en compte les métadonnées, assurant ainsi la transmission des informations du pangénome aux gènes du génome externe. Cette intégration renforce la cohérence des analyses en permettant d'exploiter les annotations des utilisateurs pour une meilleure interprétation des résultats.

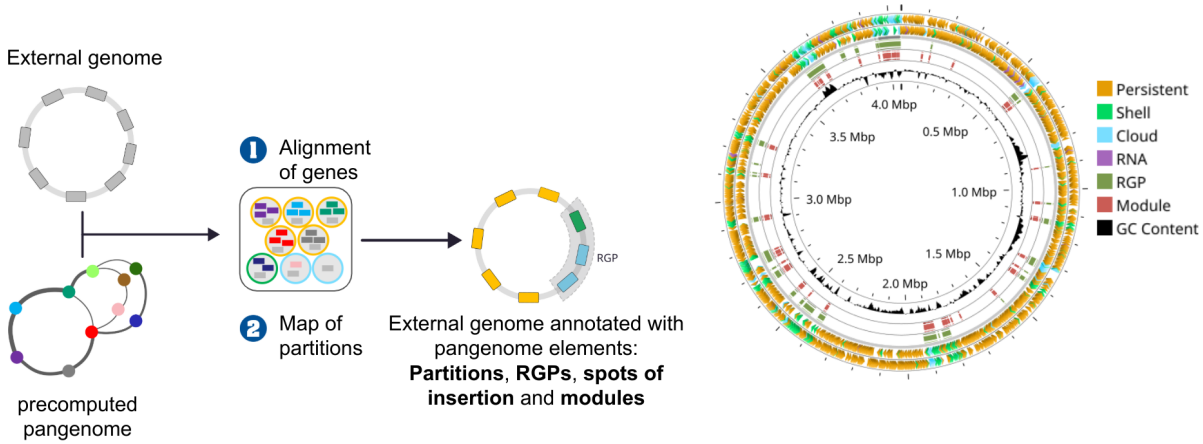


Figure II.2.2 – **Principe de fonctionnement de la méthode de projection.** Le génome sur lequel le pangénome est projeté, est un génome de la même espèce que ceux qui ont servi à construire le pangénome. Sur la droite, un exemple de sortie disponible dans PPanGGOLiN et généré dans les projections, le Proksee (Grant *et al.*, 2023) d'un génome de *A. baumannii* où les résultats du pangénome ont été projeté.

2.1.1.4 . Analyse comparée des RGPs

Une nouvelle fonctionnalité a été intégrée à PPanGGOLiN pour permettre l'analyse comparative des RGPs entre plusieurs génomes d'un même pangénome. Désormais, il est possible de regrouper (clusteriser) les RGPs similaires (figure II.2.3a). Deux RGPs sont considérés comme partageant un contenu commun si leurs gènes appartiennent aux mêmes familles de gènes. À partir de cette définition, un score de correspondance du répertoire génique (*Gene Repertoire Relatedness*, GRR) est calculé, soit sur la RGP la plus petite, soit sur la plus grande :

$$
\begin{aligned}
GRR_{min} &= \frac{\text{nombre de gènes partagés}}{\text{nombre de gènes de la plus petite RGP}} \\
GRR_{max} &= \frac{\text{nombre de gènes partagés}}{\text{nombre de gènes de la plus grande RGP}}
\end{aligned}
\tag{2.1}
$$

Le processus de clustering repose sur une modélisation par graphe. Chaque RGP est représentée sous forme d'un nœud, et une arête est ajoutée entre deux nœuds si leur GRR dépasse un seuil défini (0,8 par défaut). Ainsi, chaque composante connexe du graphe correspond à un ensemble de RGPs partageant un même répertoire génique.

Les résultats du clustering sont disponibles soit sous forme d'un fichier tabulé récapitulant les regroupements obtenus, soit dans des outils de visualisation de graphe comme Gephi (Bastian *et al.*, 2009) ou Cytoscape (Shannon *et al.*, 2003).

L'intégration des métadonnées prend de nouveau tout son sens. Il est possible d'annoter et de colorer les graphes en fonction des métadonnées associées aux gènes du pangénome, ce qui facilite l'interprétation biologique des clusters obtenus. Un exemple d'application est illustré en figure II.2.3b, où le clustering des RGPs du pangénome de *A. baumannii* a été réalisé. Les gènes ont été annotés avec CARD (Alcock *et al.*, 2023) pour identifier les éléments liés à la résistance aux antibiotiques. Deux clusters ont été extraits en guise d'exemple. Le cluster de gauche correspond à un ensemble de RGPs localisés dans un même spot (13). Parmi eux, trois RGPs sont associées à des **gènes de résistance aux antibiotiques** (nœuds en forme de triangle). Le cluster de droite regroupe des RGPs répartis sur huit spots distincts, révélant ainsi une mobilité plus variée dans les génomes. Contrairement au premier cluster, ces RGPs ne sont pas directement associées à des gènes de résistance. Cependant, d'autres métadonnées, par exemple des annotations métaboliques, pourraient être intégrées pour identifier des fonctions communes entre ces spots.

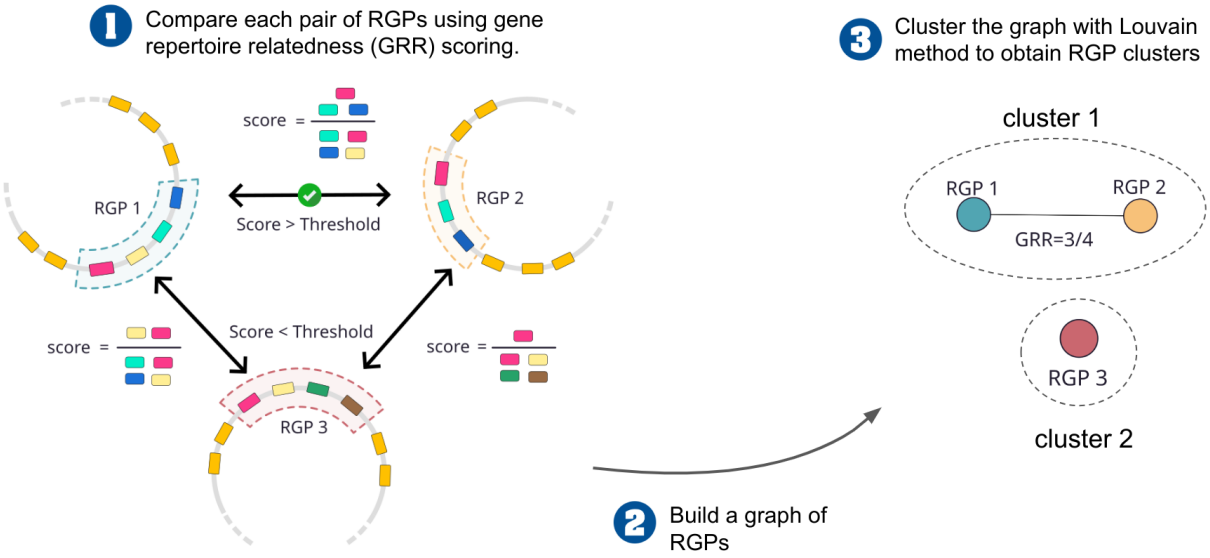
2.1.2 . Amélioration des procédures d'analyses

2.1.2.1 . Lecture des fichiers d'annotation

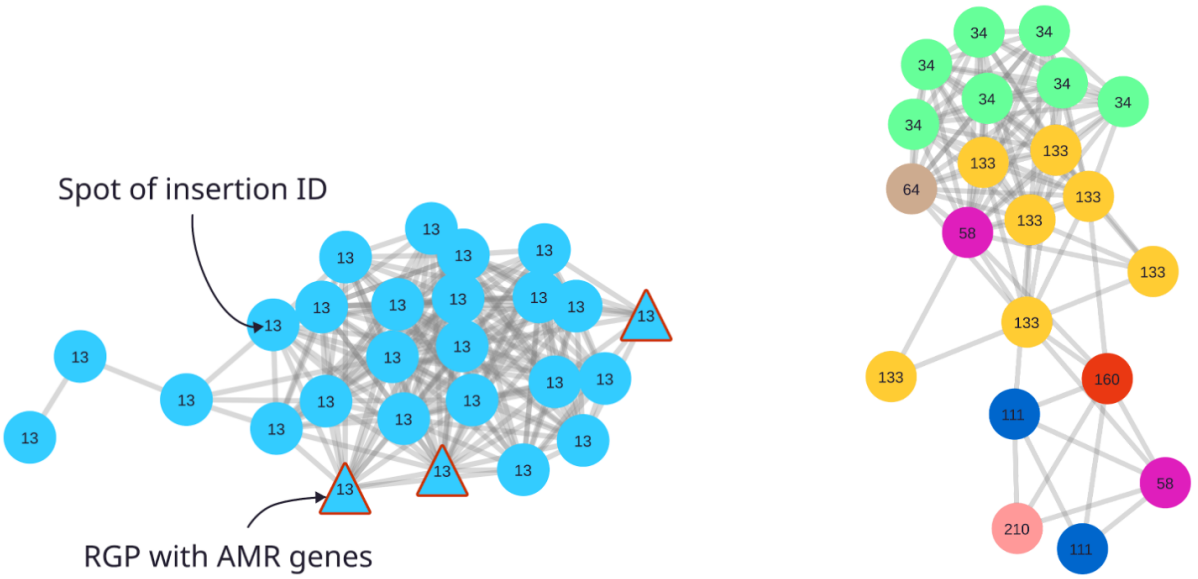
PPanGGOLiN permet la construction d'un graphe de pangénome à partir de séquences nucléiques et de génomes préalablement annotés. Les fichiers de génomes annotés (aux formats GFF ou GBFF) contiennent déjà les gènes ainsi que leurs coordonnées (contig, position, brin…). Ces fichiers peuvent également inclure des informations sur la fonction des gènes, des métadonnées relatives aux génomes et aux contigs, ainsi que la séquence des gènes ou des contigs eux-mêmes. Jusqu'à récemment, une partie de ces informations était ignorée par PPanGGOLiN. Désormais, elles sont lues et stockées dans le fichier de pangénome sous forme de métadonnées.

En nous intéressant à l'annotation fonctionnelle du pangénome, nous avons observé des divergences entre les séquences issues des fichiers d'annotation et celles correspondant aux gènes et aux familles de gènes. Nous avons notamment constaté que certaines coordonnées de gènes présentaient une complexité particulière et correspondaient à des événements biologiques spécifiques, tels que des décalages du cadre de lecture (*frameshift*). Un ensemble de cas présentant des coordonnées atypiques a ainsi été identifié et est désormais correctement pris en charge dans PPanGGOLiN.

Ces améliorations ont conduit à une meilleure construction des familles de gènes, en corrigeant des séquences auparavant incorrectement tronquées ou décalées. Par ailleurs, elles ont permis d'améliorer la réécriture des séquences dans les fichiers de sortie, notamment ceux destinés à Proksee (Grant *et al.*, 2023), et, par conséquent, la qualité des annotations fournies par les outils intégrés à cette plateforme.



(a) Méthode de clusterisation des RGPs



(b) Application du clustering des RGPs sur le pangénome de *A. baumannii* avec des métadonnées de résistance aux anitbiotiques (AMR)

Figure II.2.3 – **Clustering des RGPs.**

2.1.2.2 . Exécution de PPanGGOLiN via un fichier de configuration

PPanGGOLiN intègre désormais la possibilité de générer un fichier de configuration. Ce fichier contient l'ensemble des commandes de PPanGGOLiN pouvant être exécutées ainsi que les paramètres spécifiques à chaque commande et les globaux. A partir de ce fichier, PPanGGOLiN peut lancer une analyse *de novo* sans préciser les paramètres de la ligne de commande. Ce fichier offre ainsi une alternative plus souple et organisée à l'exécution classique en ligne de commande. Il présente aussi plusieurs avantages. (*i*) Une exécution entièrement paramétrable des workflows. Jusqu'à présent, l'exécution de PPanGGOLiN dans un workflow reposait sur un nombre restreint de paramètres en ligne de commande. Cette limitation visait à éviter une surcharge des lignes de commande ainsi que d'éventuels conflits de nommage entre paramètres. Grâce aux fichiers de configuration, il devient possible de paramétrer entièrement un workflow, en définissant de manière explicite toutes les options requises. (*ii*) Une intégration facilitée dans les pipelines d'analyse. L'utilisation de fichiers de configuration simplifie considérablement l'intégration de PPanGGOLiN dans des pipelines d'analyse. En effet, l'exécution d'outils en ligne de commande au sein de pipelines automatisés (souvent via des scripts Bash) peut poser plusieurs difficultés, telles que : une mauvaise interprétation des types de paramètres (par exemple, un entier lu comme une chaîne de caractères), des erreurs dans la gestion des chemins de fichiers, des conflits entre options spécifiques. L'emploi d'un fichier de configuration permet d'éliminer ces problèmes en standardisant et sécurisant la transmission des paramètres. Cette approche a notamment été adoptée lors de l'intégration de PPanGGOLiN dans MicroScope (Vallenet *et al.*, 2020).

L'ajout des fichiers de configuration s'inscrit également dans les principes FAIR [2]. En particulier, ils renforcent la reproductibilité des analyses. PPanGGOLiN permet ainsi de générer, à partir d'un pangénome, un fichier de configuration contenant toutes les options ayant permis sa construction. Cela garantit que, pour un même jeu de données, les analyses restent strictement reproductibles, favorisant ainsi une science plus ouverte et plus éthique. Cette avancée est particulièrement bénéfique dans un contexte de publication scientifique, où la transparence et la reproductibilité des analyses sont essentielles.

2. Les principes FAIR visent à garantir que les données et les logiciels scientifiques soient faciles à retrouver (Findable), accessibles (Accessible), interopérables (Interoperable) et réutilisables (Reusable). Bien que ces principes aient été initialement conçus pour les données et le *Big Data*, ils s'appliquent également aux outils bioinformatiques.

2.2 . Optimisation technique

Au-delà de l'ajout de nouvelles fonctionnalités, la version 2 de PPanGGOLiN a bénéficié de nombreuses améliorations visant à optimiser son efficacité, son ergonomie et sa maintenabilité. Parmi celles-ci, les optimisations techniques ont joué un rôle clé en réduisant la taille des fichiers générés, les temps de lecture ainsi que la mémoire utilisée.

2.2.1 . Amélioration de l'efficacité de PPanGGOLiN

L'une des améliorations concerne l'optimisation du processus d'annotation des génomes. À cette fin, l'exécution de Prodigal (Hyatt *et al.*, 2010) a été remplacée par l'intégration de Pyrodigal (Larralde, 2022). Ce changement présente deux avantages principaux. Premièrement, la réduction des entrées/sorties (I/O) et l'amélioration de la gestion de la mémoire. Contrairement à l'exécution de Prodigal, qui nécessitait de générer des fichiers intermédiaires contenant les résultats d'annotation, Pyrodigal retourne directement les annotations sous forme d'objets Python utilisables par PPanGGOLiN. Cette approche permet de ne pas avoir à écrire et lire des fichiers intermédiaires pour chaque génome, réduisant ainsi l'empreinte mémoire et améliorant l'efficacité des opérations d'I/O. Deuxièmement, Pyrodigal intègre plusieurs améliorations au moteur de Prodigal, notamment une optimisation du calcul du score de connexion. Ce score est utilisé pour évaluer la probabilité d'une transition entre deux codons en fonction de divers critères (cadre de lecture, type de codon – initiation ou stop –, orientation du brin…). Comme l'a souligné Larralde (Larralde, 2022), le calcul du score de connexion sur de longues séquences peut être coûteux en temps de calcul. Pyrodigal améliore ce processus en implémentant un filtrage heuristique, qui permet d'ignorer les connexions clairement invalides dès le départ. De plus, Pyrodigal exploite les fonctionnalités SIMD (Single Instruction, Multiple Data) des CPU modernes pour traiter plusieurs nœuds de calcul simultanément. Cela permet de traiter 8 nœuds avec les fonctionnalités NEON et SSE2, et 16 nœuds avec AVX2 (figure II.2.4). Ces optimisations réduisent le temps de calcul et améliorent significativement l'efficacité globale de la prédiction des gènes dans PPanGGOLiN.

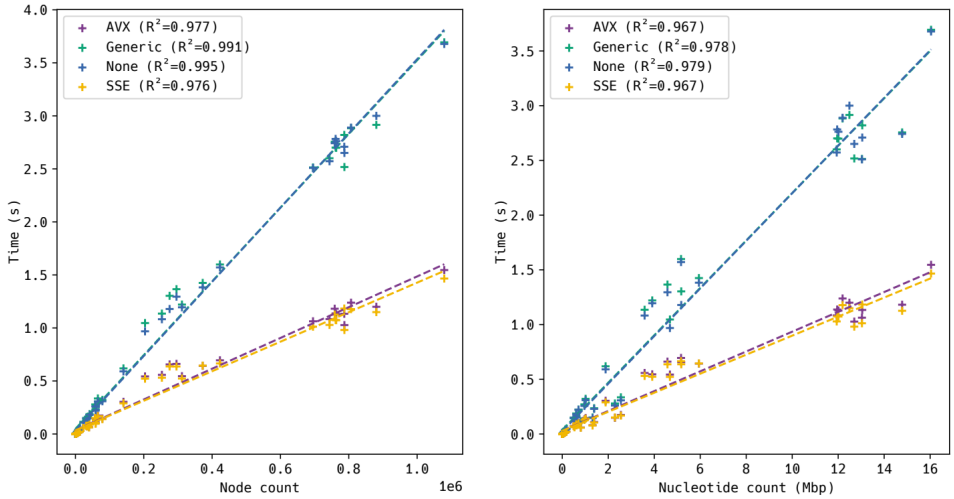


Figure II.2.4 – **Évaluation des performances de calcul des scores de connexion.** L'évaluation est réalisé avec différents backends SIMD pour le filtre heuristique (SSE2 ou AVX2), avec un backend générique (*Generic*) ou sans activer le filtre (*None*). Chaque séquence a été traitée 10 fois sur un processeur *i7-10710U* à 1,10 GHz. Extrait de (Larralde, 2022)

Toujours dans l'objectif d'amélioration de l'efficacité de PPanGGOLiN, les fonctions de lecture ont été revues, afin de réduire le temps de chargement des données et l'utilisation de la mémoire. Pour atteindre cet objectif, plusieurs modifications ont été apportées à la manière dont les objets du pangénome sont interconnectés, afin de limiter le chargement de données inutiles. Dans PPanGGOLiN, la structure hiérarchique des objets implique, par exemple qu'un gène appartient à un contig, lequel est lui-même rattaché à un génome. Or, dans certaines commandes, le chargement des gènes entraînait systématiquement celui des contigs et des génomes, augmentant ainsi le temps d'exécution. Pour pallier ce problème, j'ai réorganisé ces dépendances et réécrit plusieurs fonctions de lecture afin de rendre l'accès aux données plus sélectif et plus efficace.

L'une des améliorations les plus significatives concerne la lecture des spots. Une réécriture de cette fonction a permis une réduction drastique du temps de lecture. Sur un pangénome de 3 083 génomes de *E. coli*, le temps de lecture de 2 036 spots est passé de 25 minutes (dans un temps total de lecture du pangénome de 36 minutes) à seulement 3,5 secondes (pour un temps total réduit à 9 minutes et 38 secondes) [3].

Dans cette même optique, des améliorations ont également été apportées à l'écriture des séquences (nucléotidiques et protéiques). Dans les versions précédentes de PPanGGOLiN, pour faciliter le filtrage des séquences demandées par l'utilisateur, l'ensemble du pangénome et de ses objets était chargé en mémoire, ce qui entraînait une consommation importante de ressources. Désormais, les séquences sont directement extraites du fichier HDF5, évitant ainsi la recréation systématique des objets du pangénome et réduisant significativement l'utilisation de la mémoire.

Avec l'augmentation continue du nombre de génomes disponibles, les pangénomes générés par PPanGGOLiN sont devenus de plus en plus volumineux. Malgré une structure de stockage compacte, reposant sur le package PyTable (Team, 2002), la taille des fichiers HDF5 a considérablement augmenté, en raison du nombre croissant de génomes et des nouvelles fonctionnalités enrichissant les données stockées. Pour remédier à ce problème, l'architecture interne du fichier HDF5 a été revue afin d'éliminer les redondances, en particulier dans l'annotation des gènes et de leur séquence. En effet, entre différents génomes, les gènes peuvent partager des caractéristiques communes telles que leur position, leurs coordonnées ou encore leur orientation sur le brin. Or, ces informations étaient systématiquement dupliquées dans l'ancienne structure. Afin de réduire cette redondance, une table de référence a été mise en place pour centraliser ces informations et leur attribuer un identifiant unique, à la manière d'une base de données relationnelle. Chaque gène peut ainsi être associé à son annotation optimisée, sans nécessiter la répétition de ses caractéristiques dans l'ensemble du pangénome. Cette optimisation a permis une réduction significative de la taille des fichiers HDF5. Comme illustré en figure II.2.5, la taille des pangénomes a été divisée par 3,5 pour un ensemble de 1 000 génomes, et par 6,5 pour un pangénome de 2 500 génomes. Au-delà du gain en espace de stockage, cette amélioration contribue également à l'accélération des temps de lecture. En effet, elle participe à la lecture non systématique des informations liées aux gènes lors du chargement du pangénome, ce qui allège le processus et améliore les performances globales de l'outil.

Ces optimisations rendent PPanGGOLiN plus performant, plus rapide et plus adapté aux analyses pangénomiques de grande échelle, tout en maintenant une gestion efficace des ressources.

3. Cette amélioration est particulièrement visible sur les pangénomes de grande taille contenant un nombre important de spots, ce qui explique pourquoi le problème n'avait pas été détecté auparavant.

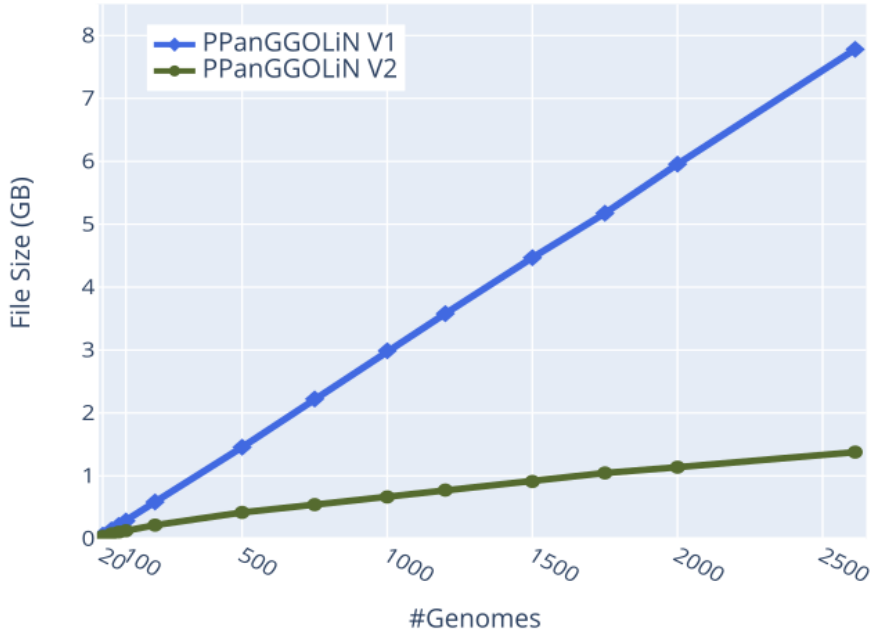


Figure II.2.5 – **Comparaison des tailles des fichiers de pangénome entre la version 1 et 2 de PPanGGOLiN.**

2.2.2 . Optimisation du code : lisibilité, maintenance, tests et processus de mise à jour

En tant que logiciel open source, le code de PPanGGOLiN est accessible à tous, conformément aux principes de la science ouverte. Ce choix présente plusieurs avantages : il permet à la communauté d'utilisateurs de signaler d'éventuels comportements inattendus (problèmes de performance, bugs, *etc.*), mais aussi de contribuer directement à l'amélioration du logiciel en proposant des corrections ou des optimisations. Afin d'assurer la pérennité et la maintenabilité du projet, nous avons entrepris une révision complète du code, avec pour objectif principal de le rendre plus lisible, homogène et facile à maintenir pour les développeurs actuels et futurs.

2.2.2.1 . Mise en place de bonnes pratiques de développement

La première étape du reformatage a consisté à aligner PPanGGOLiN avec les versions maintenues et corrigées de Python. Les packages les plus utilisés suivent également cette logique. En mettant à jour le code pour être compatible avec les versions récentes de Python, nous bénéficions des dernières mises à jour des packages, ainsi que des optimisations et corrections apportées par le langage lui-même. Ainsi, nous sommes passés de Python 3.8 à Python 3.12.

Pour améliorer la lisibilité et garantir une cohérence au sein du code, nous avons introduit un ensemble de règles et de processus applicables pour tous les développeurs. La première étape a été l'adoption des bonnes pratiques de codage en Python, en suivant les recommandations des PEP (*Python Enhancement Proposals*). L'application de ces directives présente plusieurs bénéfices. Premièrement, le code est plus structuré et li-

sible. Par exemple, les conventions de nommage permettent d'identifier rapidement la nature des variables et objets :

- Nom des classes en CamelCase (*e.g.* GeneCluster)
- Constantes en majuscules (*e.g.* MAX_ITERATIONS)
- Fonctions et méthodes en snake_case (*e.g.* compute_similarity_score)

Cette homogénéité facilite la lecture et la compréhension du code par tous les contributeurs. Deuxièmement, les règles PEP permettent de mettre en place des micro-optimisations pour de meilleures performances. Par exemple, en privilégiant l'utilisation de générateurs plutôt que des listes temporaires, ce qui permet d'optimiser l'utilisation de la mémoire.

Afin de faciliter l'application de ces règles et d'assurer leur respect de manière systématique, nous avons intégré le package python Black (`https://github.com/psf/black`) pour le formatage automatique du code. Lors de chaque mise à jour du code sur GitHub, Black est exécuté automatiquement pour reformater le code selon les standards des PEP. Cette automatisation garantit une cohérence stylistique, réduit la charge de travail des développeurs et simplifie la gestion des contributions extérieures.

Grâce à ces améliorations, le code de PPanGGOLiN devient plus clair, plus performant et plus facile à maintenir, assurant ainsi une plus grande longévité au projet et une meilleure collaboration au sein de la communauté open source.

2.2.2.2 . Maintenance et test du code

PPanGGOLiN est avant tout un projet de recherche scientifique, ce qui implique des mises à jour fréquentes pour intégrer de nouvelles fonctionnalités et améliorer ses performances. Toutefois, ces modifications peuvent introduire des bogues susceptibles d'altérer la fiabilité des analyses. Afin de garantir la stabilité et la robustesse du logiciel, nous avons grandement amélioré la stratégie de tests automatisés, couvrant différents niveaux de validation.

a. Tests unitaires : validation des fonctionnalités élémentaires

Les premiers tests mis en place sont des tests unitaires, qui permettent de vérifier le comportement individuel des fonctions. Ces tests s'assurent que chaque fonction produit bien le résultat attendu, et qu'elle génère les erreurs appropriées en cas d'entrée invalide. Actuellement, l'ensemble des classes sont testées et ainsi qu'une grande partie des fonctions.

b. Tests d'intégration : validation des interactions entre fonctions

Nous avons également mis en place des tests d'intégration, visant à évaluer le comportement global du logiciel lorsque plusieurs fonctions interagissent. En effet, une fonction peut répondre comme attendu de manière isolée tout en produisant des comportements imprévus lorsqu'elle est combinée à d'autres modules. Ces tests permettent donc de garantir la cohérence de l'ensemble du programme. Cependant, leur mise en œuvre reste complexe en raison des nombreuses interactions entre les fonctions dans PPanGGOLiN, ce qui limite la couverture de cette approche à une partie restreinte du code.

c. Tests fonctionnels : validation du comportement des commandes

Enfin, nous avons introduit des tests fonctionnels, qui vérifient le comportement des commandes complètes. Pour cela, des fichiers de résultats de référence ont été préchargés, permettant de comparer automatiquement les nouvelles sorties avec les résultats attendus. Cette approche garantit que les évolutions du code n'altèrent pas l'exactitude des analyses produites par PPanGGOLiN. Toutes les commandes de PPanGGOLiN sont testées en prenant en compte tous les paramètres.

Lors d'une mise à jour du code, un workflow automatique est exécuté et permet de vérifier que le code est bien fonctionnel. Grâce à cette infrastructure de tests, le code est désormais plus robuste, limitant ainsi l'introduction de bogues imprévus et assurant la fiabilité des résultats scientifiques générés par le logiciel.

2.2.2.3 . Amélioration de l'expérience utilisateur

PPanGGOLiN est un logiciel conçu pour les microbiologistes, dont l'expertise principale n'est pas nécessairement l'informatique. Il est donc essentiel de garantir une expérience utilisateur fluide, en facilitant aussi bien l'installation que l'utilisation du logiciel.

Un premier effort a été porté sur la réécriture complète de la documentation, afin de la rendre plus claire, plus structurée et mieux adaptée aux besoins des utilisateurs. Désormais, elle comprend :

- Une section d'installation détaillant plusieurs méthodes adaptées à différents environnements,
- Un guide de prise en main rapide pour permettre aux utilisateurs d'exécuter rapidement PPanGGOLiN et ses principaux workflows,
- Des sections approfondies décrivant en détail chaque commande et analyse réalisable,
- Un guide dédié aux développeurs, recensant les bonnes pratiques et les processus de développement spécifiques à PPanGGOLiN.

De plus, la documentation est maintenant disponible sur ReadTheDocs, ce qui la rend plus accessible, mieux référencée et conforme aux principes FAIR.

Un autre aspect clé de l'amélioration de l'expérience utilisateur concerne la gestion des erreurs. Afin de rendre les messages d'erreur plus explicites et plus informatifs, nous avons entrepris une révision complète de leur génération. Les erreurs liées à une mauvaise utilisation par l'utilisateur sont maintenant décrites de manière plus claire et pédagogique. Les erreurs techniques, destinées aux développeurs, sont quant à elles plus précises, facilitant ainsi l'identification rapide de l'origine du problème. De plus, un plus large éventail de messages a été introduit pour couvrir davantage de cas d'erreur et améliorer la gestion des exceptions.

Grâce à ces améliorations, PPanGGOLiN devient plus intuitif pour les microbiologistes et plus facile à maintenir pour les développeurs, garantissant ainsi une expérience utilisateur optimisée.

3 - Application à l'étude de la dégradation du D-Apiose

Le D-Apiose est un sucre de type pentose, principalement retrouvé dans la paroi cellulaire des plantes vasculaires, mais également présent en tant que métabolite secondaire (Pičmanová et Møller, 2016). Chez les bactéries, plusieurs voies de dégradation du D-Apiose ont été identifiées (Carter *et al.*, 2018). Parmi celles-ci, la voie de la transcétolase non oxydante (figure II.3.1) constitue un mécanisme clé. Cette voie permet la conversion du D-Apiose en D-xylulose 5-phosphate, un métabolite intermédiaire central impliqué dans de nombreuses voies métaboliques essentielles. La voie de la transcétolase non oxydante a été détectée chez diverses bactéries du sol, notamment *Actinobacillus succinogenes* et *Bacteroides vulgatus*, ainsi que chez plusieurs espèces du genre *Pectobacterium*.

En collaboration avec Guilhem Royer (AP-HP) et Erick Denamur (IAME), le LABGeM a identifié une voie alternative chez *Escherichia coli*, où l'isomérase de la première réaction est remplacée par une succession de deux oxydoréductases. Cette voie a été initialement identifiée spécifiquement dans les *Sequence Types* (ST) 131 et 14 de *E. coli*, des souches pathogènes connues pour leur multirésistance aux antibiotiques et leur implication dans les bactériémies [1] (Schembri *et al.*, 2015; de Korne-Elenbaas *et al.*, 2023). Ces souches colonisent principalement le tube digestif, où la capacité à dégrader le D-Apiose pourrait conférer un avantage sélectif, améliorant ainsi le *fitness* de ces pathogènes.

Dans ce cadre, j'ai contribué au projet en explorant les pangénomes afin d'identifier le contexte génomique associé à la fois à la voie connue (transcétolase non oxydante) et à la voie alternative de dégradation du D-Apiose.

3.1 . Recherche du contexte génomique chez les procaryotes

J'ai commencé par rechercher la voie de dégradation du D-Apiose, à la fois sous sa forme connue (transcétolase non oxydante) et sous sa forme alternative, dans les espèces procaryotes. Pour cela, j'ai analysé le contexte génomique associé aux six protéines clés impliquées dans ces voies. Ces protéines incluent : 3 enzymes communes aux deux voies, les deux sous-unités de la transcétolase et la kinase ; une enzyme spécifique à la voie connue l'isomérase ; 2 enzymes spécifiques à la voie alternative, les oxydoréductases.

L'analyse du contexte génomique a été réalisée sur un ensemble de 1 429 pangénomes d'espèces, générés à l'aide de l'outil PPanGGOLiN [2]. Ces pangénomes ont été construits à partir de 152 717 génomes issus de la base de données RefSeq (Pruitt *et al.*, 2007), en utilisant la taxonomie GTDB (Parks *et al.*, 2018) pour organiser les génomes par espèce (version RefSeq/GTDB : 220).

1. Infection bactérienne présente dans le sang.
2. Les pangénomes ont été élaborés dans le cadre du projet PanGBank, visant à constituer une base de données de pangénomes d'espèces.

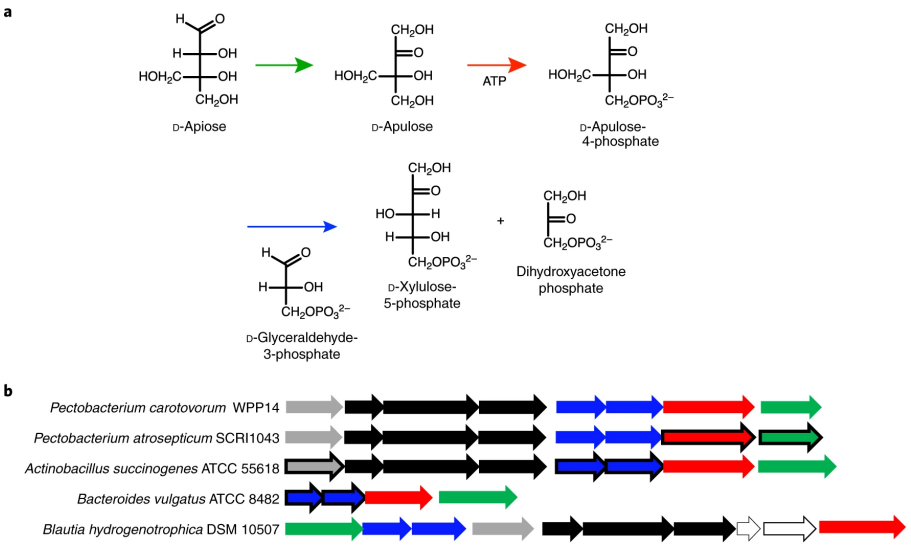


Figure II.3.1 – **Voie de dégradation du D-Apiose par la transcétolase non oxydante.** (a) Métabolites et réactions de la voie. La première réaction (en vert) correspond à l'isomérisation du D-Apiose en D-Apulose, catalysée par la D-Apiose isomérase. Le D-Apulose est ensuite phosphorylé en D-Apulose 4-phosphate (en rouge) par l'action d'une kinase. Enfin, la dernière transformation (en bleu) est catalysée par la transcétolase, qui permet la production du D-xylulose 5-phosphate. (b) Contexte génomique de la voie de dégradation chez cinq espèces bactériennes. Les gènes impliqués sont représentés par des flèches colorées, dont la couleur correspond à l'enzyme codée. Les gènes gris codent des SBP (protéines de liaison spécifiques) impliqués dans la reconnaissance du D-Apiose, tandis que les gènes noirs codent des composants d'un système de transport ABC. Extrait de (Carter *et al.*, 2018).

Le GC de dégradation du D-Apiose a été identifié dans 125 espèces, réparties sur 43 genres, 17 familles, 15 ordres, 7 classes et 4 phylums (figure II.3.2). Parmi ces familles, les **Enterobacteriaceae** sont les plus représentées (66 espèces), suivies des **Rhizobiaceae** (23 espèces) et des **Pseudomonadaceae** (12 espèces). Cette répartition suggère que la voie de dégradation du D-Apiose est relativement fréquente chez les bactéries. L'analyse de la distribution des données au sein des partitions du pangénome révèle que cette voie est principalement associée à un contexte *persistent*, avec 28 genres sur 43 présentant une majorité de familles classées dans cette même partition.

Parmi les contextes identifiés, seules 7 espèces présentent la voie connue de dégradation du D-Apiose. Six d'entre elles appartiennent au genre *Pectobacterium* : *P. atrosepticum*, *P. brasiliense*, *P. parmentieri*, *P. carotovorum*, *P. versatile* et *P. polare*. La septième espèce, *Novosphingobium capsulatum*, appartient à la famille des *Sphingomonadaceae* et est représentée en bleu-vert clair sur la figure II.3.2. Contrairement aux espèces du genre *Pectobacterium*, où le contexte génomique est retrouvé dans la partition *persistent*, celui de *N. capsulatum* se situe dans la partition *shell*, ce qui pourrait suggérer un transfert horizontal de gènes (HGT) entre ces espèces.

Par ailleurs, 3 espèces possèdent une version hybride entre la voie connue et la voie alternative, caractérisée par la présence simultanée de l'isomérase et des 2 oxydoréductases. Deux d'entre elles appartiennent à la famille des *Sphingomonadaceae*, à savoir *Sphingobium yanoikuyae* et *Sphingomonas koreensis*, tandis que la troisième, *Klebsiella aerogenes*, appartient aux *Enterobacteriaceae*. La présence de voies hybrides dans certaines espèces pourrait refléter une adaptation évolutive conférant une plus grande flexibilité métabolique en fonction des conditions environnementales.

Enfin, des contextes partiels ont été détectés dans 13 espèces. Dix d'entre elles appartiennent à la famille des *Enterobacteriaceae*, incluant *Salmonella diarizonae*, *Atlantibacter hermannii*, *Yersinia enterocolitica*, *Leclercia adecarboxylata*, *Citrobacter youngae*, *Yersinia bercovieri*, *Kosakonia radicincitans*, *Yersinia mollaretii*, *Yersinia massiliensis* et *Yersinia intermedia*. Deux autres espèces, *Paracidovorax avenae* et *Paracidovorax citrulli*, appartiennent à la famille des *Burkholderiaceae*, tandis que *Clostridioides difficile* représente la famille des *Peptostreptococcaceae*. L'existence de ces contextes partiels pourrait être attribuée à la présence d'autres voies métaboliques alternatives, comme proposé par Carter *et al.* (Carter *et al.*, 2018), ou être liée à la spécificité de l'isomérase au genre *Pectobacterium*. L'analyse de la base de données UniProt (The UniProt Consortium, 2025) suggère en effet que cette enzyme présente une faible similarité avec celles d'autres organismes, ce qui pourrait expliquer l'absence apparente de la voie connue dans certaines espèces, alors qu'un homologue fonctionnel pourrait exister.

Des recherches complémentaires seront nécessaires pour mieux comprendre les implications fonctionnelles de ces variations et explorer les mécanismes évolutifs sous-jacents.

J'ai ensuite recentré mes analyses sur la famille des Enterobacteriaceae, où la voie de dégradation du D-Apiose avait été initialement identifiée (Carter *et al.*, 2018) et à laquelle *Escherichia coli* appartient. Comme l'illustre la figure II.3.3, le contexte génomique est principalement retrouvé dans la partition *persistent* (16 espèces), suivie de la partition *shell* (8 espèces) et enfin de la partition *cloud* (5 espèces). Ces espèces appartiennent à plusieurs genres bactériens, notamment Klebsiella, Citrobacter, Serratia et Escherichia. La majorité d'entre elles sont connues pour être des pathogènes humains, partageant un habitat commun : le tube digestif. D'un point de vue phylogénétique et taxonomique, ces bactéries sont étroitement apparentées à des espèces vivant dans les sols, où la dégradation du D-Apiose issu des plantes constitue un avantage sélectif. C'est notamment le cas de *Pantoea vagans*, une espèce isolée à partir de feuilles d'eucalyptus (Brady *et al.*, 2009). La présence de la voie alternative dans la partie variable du pangénome suggère une acquisition par transfert horizontal, au cours de laquelle cette nouvelle voie métabolique aurait été transférée des bactéries vivant dans les écosystèmes du sol vers celles colonisant le microbiote intestinal et gastrique.

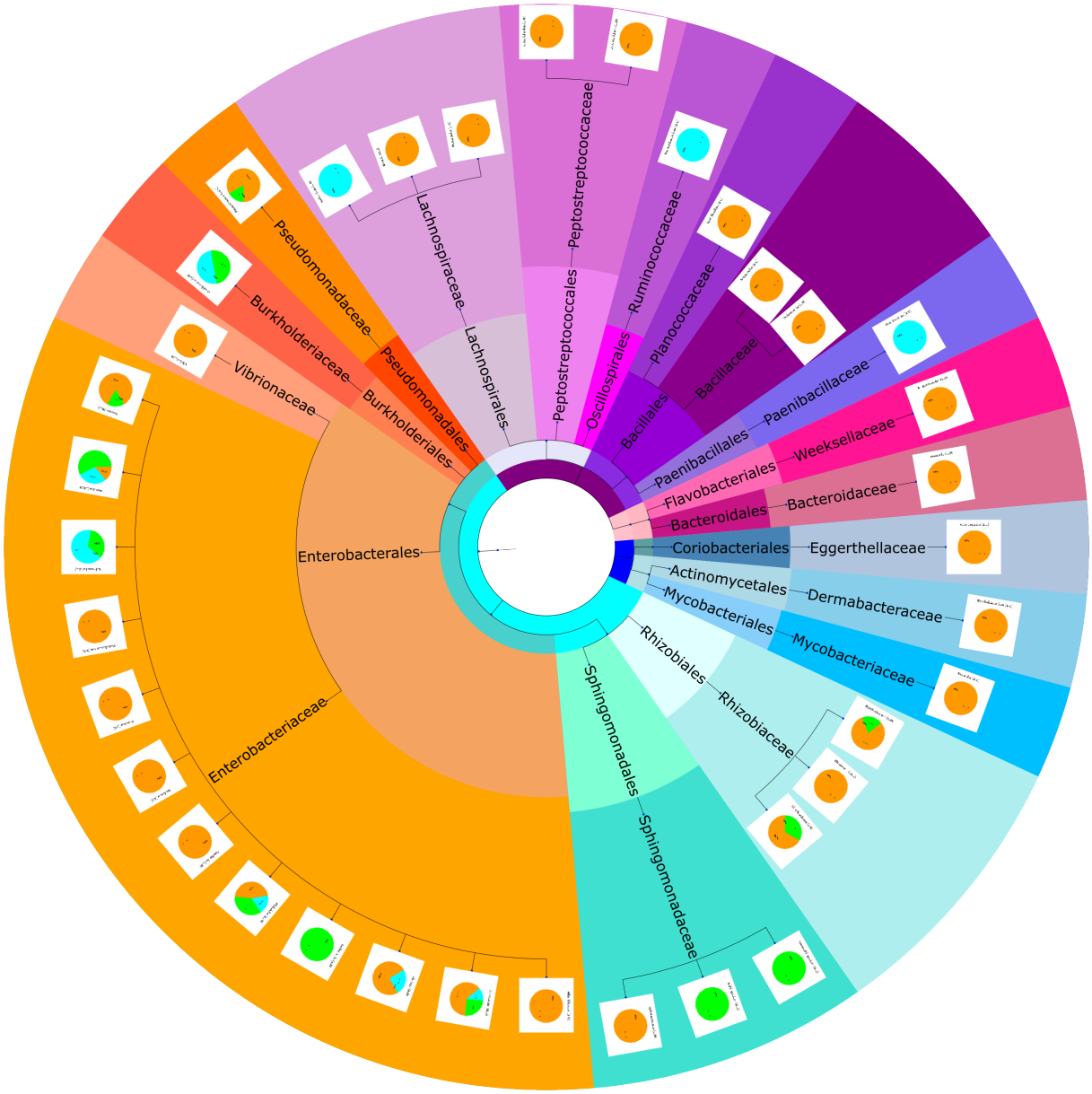


Figure II.3.2 – **Arbre taxonomique des procaryotes étiqueté par la présence de la voie de dégradation du D-Apiose.** Si une branche existe alors un GC a été détecté sinon l'embranchement n'est pas créé. Les feuilles de l'arbre représente le niveau taxonomique du genre. Les *pie chart* en bout de branche, représente la proportion de GC trouvé dans chaque partition du pangénome, indépendamment de la forme de la voie (connue, alternative, hybride ou partielle).

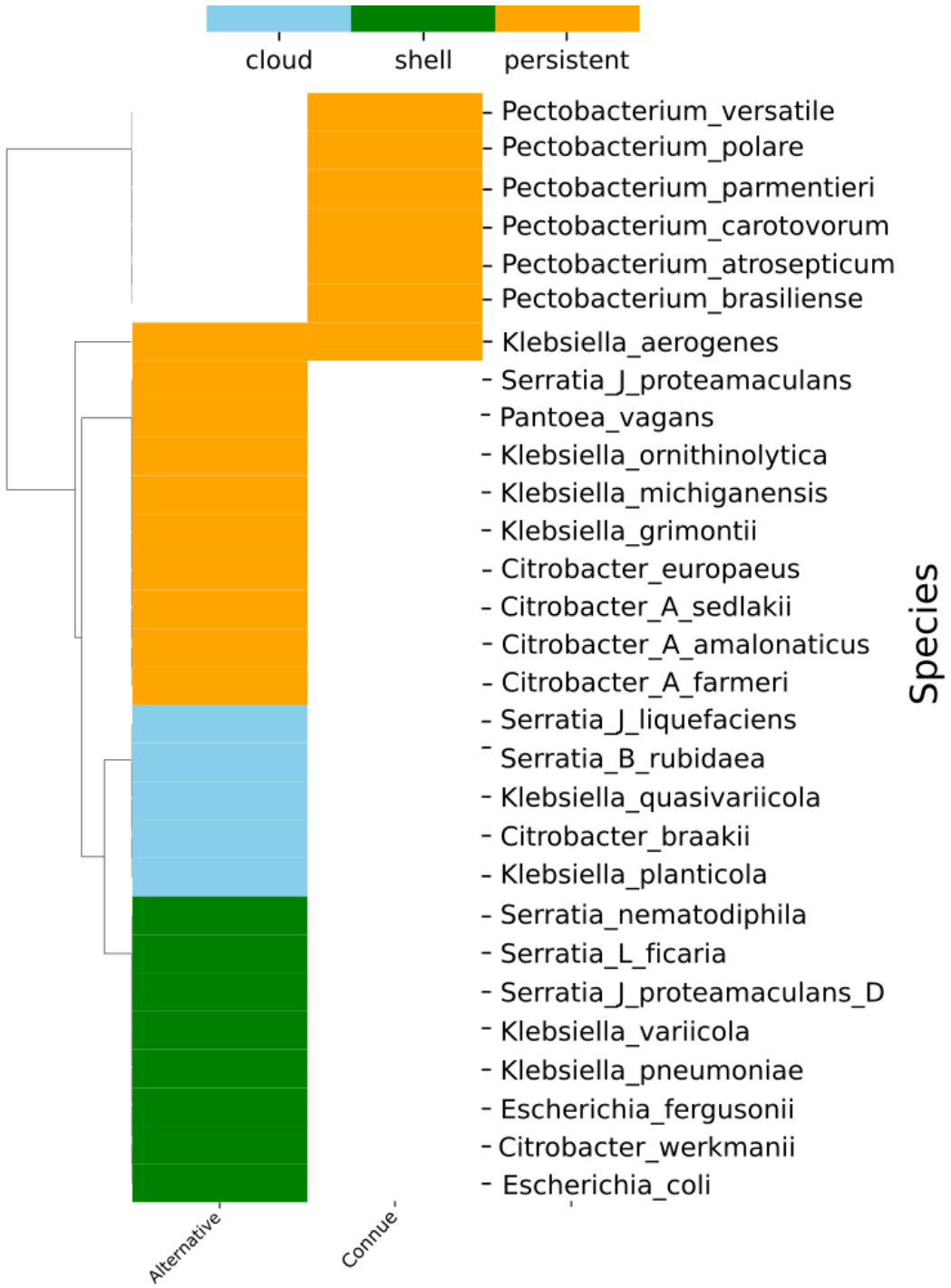


Figure II.3.3 – **Prédiction du contexte de dégradation du D-apiose dans les pangénomes des Enterobacteriaceae.**

3.2 . Analyse du pangénome de *Escherichia coli*

Par la suite, je me suis concentré sur l'espèce *Escherichia coli*, dans laquelle la voie de dégradation du D-Apiose impliquant deux oxydoréductases avait été identifiée. Pour cela, j'ai exploité le pangénome construit précédemment à partir des bases de données RefSeq et GTDB (Pruitt *et al.*, 2007; Parks *et al.*, 2018), dont la composition est indiquée dans le tableau II.3.1.

Le pangénome de *E. coli* (figure II.3.4) est majoritairement composé de familles variables (points bleus), avec des chemins *persistent* (en orange) correspondant aux régions conservées. Au sein de ces chemins, on retrouve des régions variables, constituées d'éléments *shell* (verts) et *cloud*. Ces zones correspondent généralement à des spots d'insertion, où sont localisées des régions génomiques plastiques (RGPs). C'est dans ces régions variables que j'ai recherché le contexte génomique de la voie de dégradation du D-Apiose.

Génomes	**Gènes**	**Familles**
2006	9334727	57444
Persistent	***Shell***	***Cloud***
3167	7960	46317
RGPs	**Spots**	**Modules**
164573	1968	2089

Table II.3.1 – **Composition du pangénome de *E. coli*.** Le pangénome a été généré avec PPanGGOLiN et utilise les génomes de RefSeq (Pruitt *et al.*, 2007) en suivant la taxonomie de GTDB (Parks *et al.*, 2018)

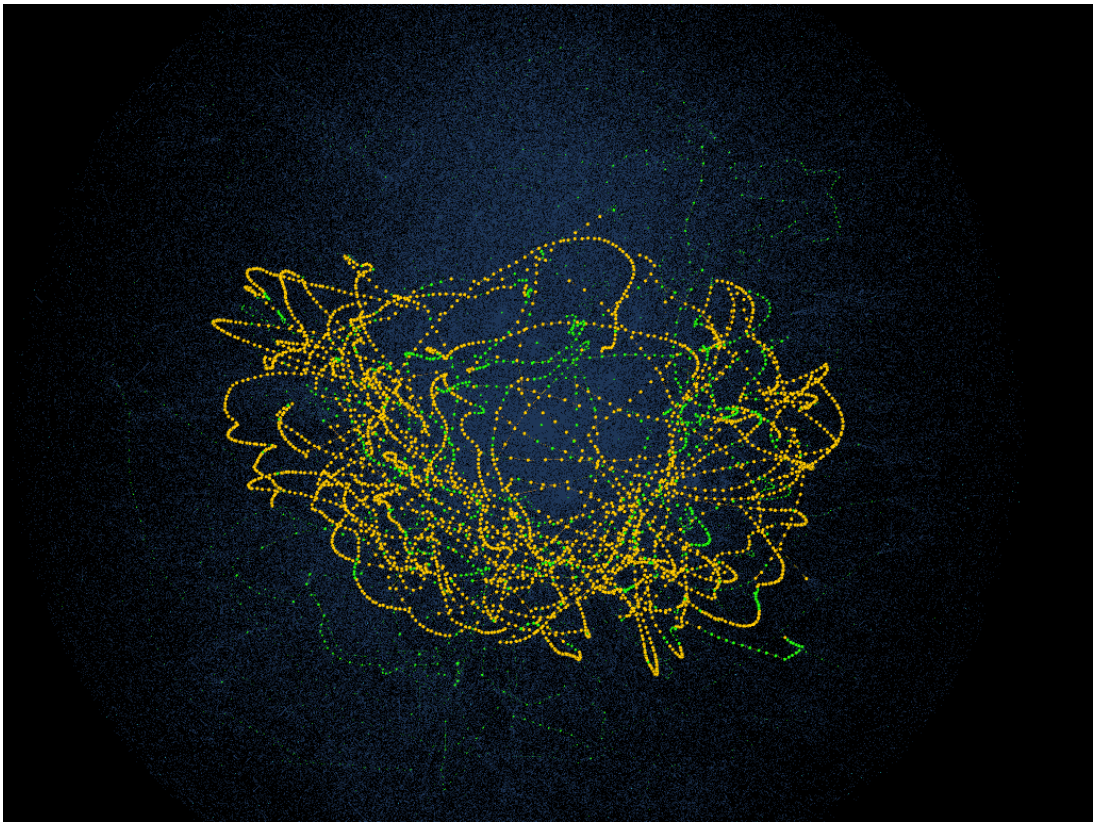


Figure II.3.4 – **Graphe de pangénome de *E. coli*.** Le graphe est visualisé avec Gephi (Bastian *et al.*, 2009). L'algorithme de spatialisation utilisé est *Force Atlas 2* avec une gravité forte et une échelle à 5 000.

L'analyse a révélé que le contexte génomique associé à cette voie est localisé dans le spot 181 du pangénome (figure II.3.5). Ce spot d'insertion est identifié dans 62 génomes, chacun contenant une seule RGP. Parmi ces génomes, 31 possèdent le contexte de dégradation du D-Apiose. Comme illustré dans la figure II.3.5, ce contexte est également associé au module 752, qui est spécifique de la voie métabolique.

L'analyse de la composition en familles du module met également en évidence la conservation de plusieurs familles de gènes qui, bien que ne jouant pas un rôle enzymatique direct, sont essentielles au fonctionnement de la voie. On retrouve notamment plusieurs familles codant des transporteurs ABC, impliqués dans la capture et le transport du D-Apiose, ainsi qu'une lipoprotéine et un facteur de transcription. Ces éléments avaient déjà été partiellement décrits par Carter *et al.* (Carter *et al.*, 2018). Toutefois, l'association systématique de ces gènes au contexte génomique prédit dans le pangénome confirme leur rôle étroit dans l'opérabilité de la voie métabolique.

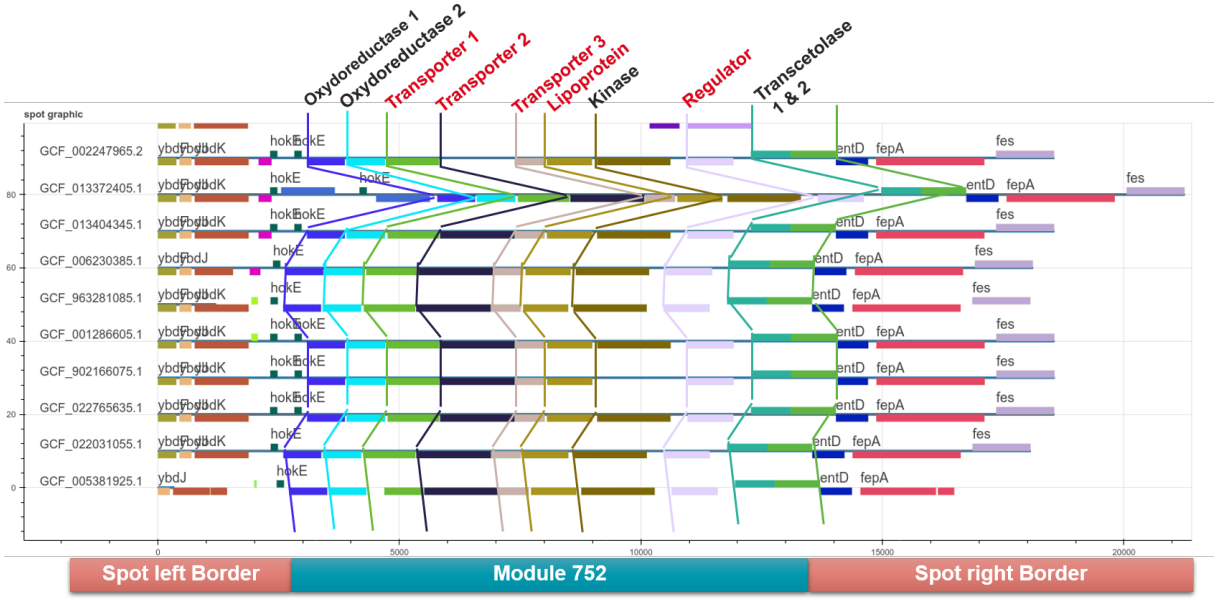


Figure II.3.5 – **Visualisation du spot 181 dans les génomes de *E. coli*.** Les gènes sont représentés par des rectangles. La couleur des gènes représente la famille à laquelle appartient le gène.

L'identification du contexte génomique de la voie de dégradation du D-Apiose dans le spot 181 du pangénome de *E. coli* constitue un indice supplémentaire en faveur d'une acquisition récente par transfert horizontal au sein de certaines lignées de *E. coli*.

3.3 . Identification et annotation de la voie de dégradation dans une nouvelle souche : BVN-ST131

En juin 2024, une nouvelle souche du type ST131 a été isolée par Van Nieuwenhuyse *et al.* (2024) (Van Nieuwenhuyse *et al.*, 2024). Les auteurs ont réussi à séquencer et à obtenir le génome complet de cette souche. Dans ce contexte, j'ai cherché à identifier la présence de la voie alternative de dégradation du D-apiose au sein de cette souche et à l'associer au spot et au module précédemment détectés.

Pour ce faire, j'ai projeté le pangénome de *E. coli* sur le génome de la souche BVN-ST131 (figure II.3.6a). J'ai ensuite recherché la présence du **spot 181** et du **module 752**, lesquels sont associés au contexte dans le pangénome. L'exploration via la carte Proksee (Grant *et al.*, 2023) a permis d'identifier une région génomique présentant le spot et le module (figure II.3.6b).

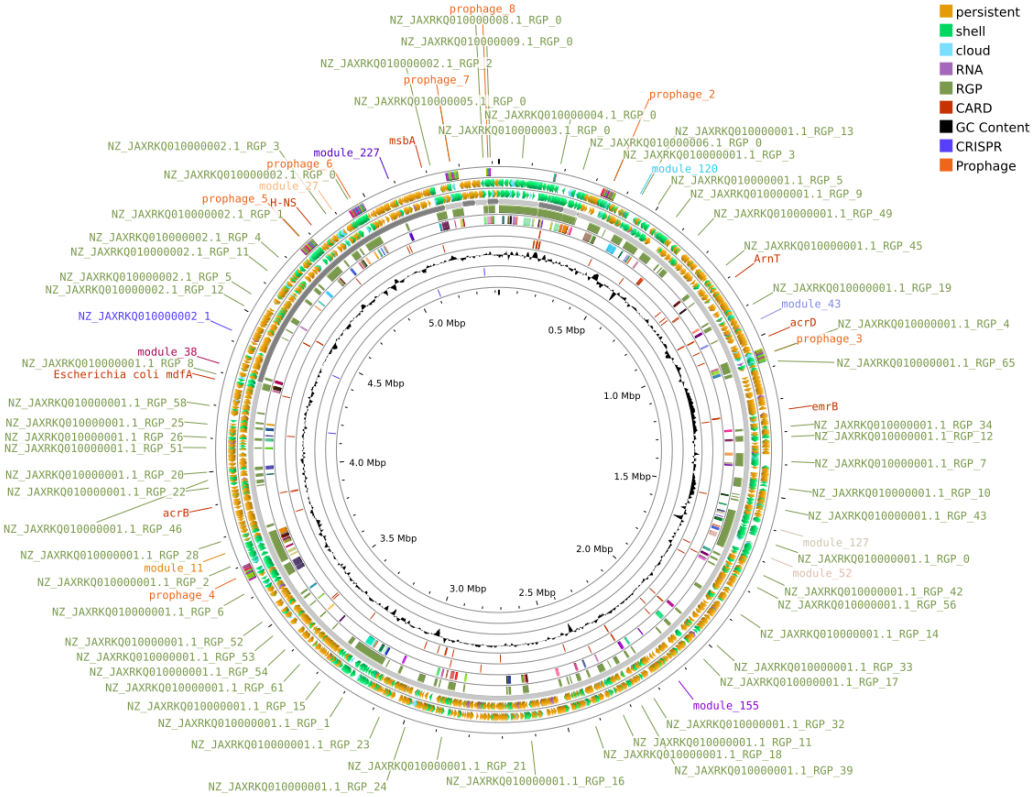
À partir de l'outil Proksee, j'ai procédé à l'alignement des protéines de la voie alternative de dégradation avec le génome de la souche, en utilisant BLAST (Altschul *et al.*, 1990), afin d'associer chaque gène identifié à une fonction spécifique (cercle externe en vert). Cette analyse a confirmé la présence de la voie alternative, incluant les 2 oxydoréductases. Par ailleurs, une annotation complémentaire des gènes restants a été réalisée à l'aide de Bakta (Schwengers *et al.*, 2021), également via Proksee, ce qui a permis de retrouver les annotations précédemment identifiées dans le spot 181 (*ABC transporter*, régulateur de transcription…).

L'identification de cette voie de dégradation dans une nouvelle souche du type ST131 confirme sa conservation au sein de ce groupe. Ces résultats vont dans le sens d'un rôle fonctionnel important dans l'adaptation et le métabolisme de ces souches, justifiant ainsi des investigations supplémentaires sur son impact physiologique et évolutif.

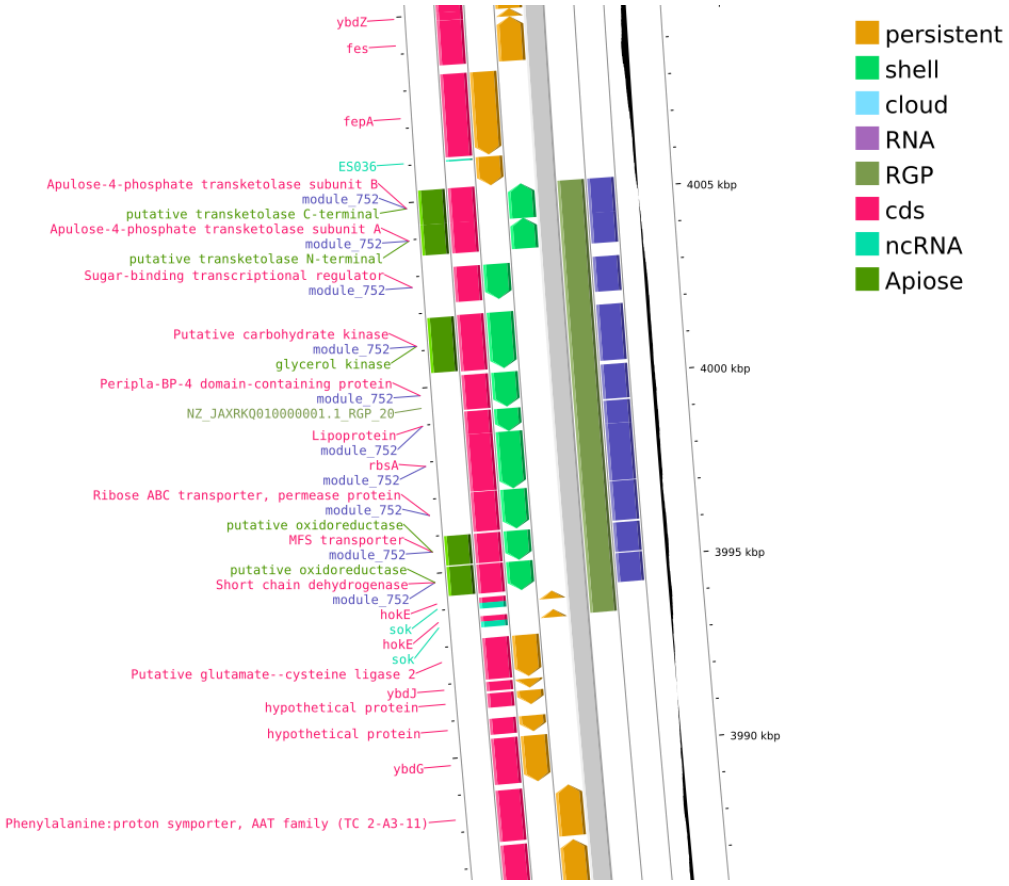
Afin de compléter notre analyse, j'ai utilisé les outils CARD (Alcock *et al.*, 2023) et Phigaro (Starikova *et al.*, 2020) afin d'identifier respectivement les gènes de résistance aux antibiotiques et les régions prophages présentes dans le génome de la souche étudiée.

L'analyse a révélé la présence d'une RGP associée au gène *mdtM*, conférant une résistance aux fluoroquinolones. Cette RGP a été retrouvée dans le **spot 99**, une région présente dans 1351 génomes, soit 67 % du pangénome, ce qui suggère qu'il s'agit d'un hotspot d'insertion. Par ailleurs, une seconde RGP, d'une taille de 88 kpb, a été identifiée. Celle-ci contient 11 gènes de résistance à divers antibiotiques : les diaminopyrimidines, les sulfamides, les aminoglycosides, les macrolides et les tétracyclines. Bien que cette RGP ne soit pas associée à un spot, les familles correspondantes aux gènes de résistance se retrouvent dans les **modules 104** et **275**, détectés respectivement dans 645 et 252 génomes. De plus, une région prophage, colocalisée avec ces deux modules, a été identifiée et caractérisée par la présence de gènes codant des intégrases et des endonucléases.

Ces résultats mettent en évidence l'intérêt de l'identification des RGP et des régions prophages dans l'étude de la dispersion des gènes de résistance aux antibiotiques au sein du pangénome de *E. coli*. L'association de ces éléments génétiques mobiles à des hotspots d'insertion et à des modules spécifiques suggère un rôle clé dans l'évolution et l'adaptation de ces souches, justifiant ainsi des analyses complémentaires sur leur impact fonctionnel et épidémiologique.



(a) Génome circulaire de la souche BVN-ST131 de *E. coli* ST131



(b) Spot 181 dans la souche BVN-ST131 de *E. coli* ST131

Figure II.3.6 – **Projection du pangénome de *E. coli* sur le génome de la souche BVN-ST131.** (a)

4 - Conclusion et perspectives

4.1 . PPanGGOLiN : bilan de la version 2.0

Depuis son lancement en 2020, PPanGGOLiN s'est imposé comme un outil pour les analyses pangénomiques, comptabilisant plus de 170 citations, auquel il faut ajouter 37 citations pour panRGP. Il offre une approche innovante et efficace de construction et de partitionnement des graphes de pangénomes. Le logiciel PPanGGOLiN a été intégré dans la plateforme MicroScope (Vallenet *et al.*, 2020), pour fournir un niveau d'information pangénomique dans les résultats d'annotation des génomes procaryotes. Il est également disponible sur la plateforme Galaxy France, maintenue par l'Institut Français de Bioinformatique (IFB). Son utilisation dépasse désormais le monde académique, avec les entreprises privées : EVOTEC et SYNGENTA, qui utilisent PPanGGOLiN pour leurs projets de R&D. The Carpentries (`https://carpentries.org/`), une entreprise dédiée à la formation en développement informatique et en data science, propose PPanGGOLiN dans un cours en ligne dédié à la pangénomique procaryote (`https://github.com/paumayell/pangenomics`).

Cette reconnaissance s'est concrétisée le 29 novembre 2023, lorsque PPanGGOLiN a reçu le **Prix "science ouverte du logiciel libre de la recherche", "espoir" de la catégorie 'Scientifique et technique'**[1], décerné par le Ministère de l'Enseignement supérieur et de la Recherche. Cette distinction souligne non seulement la qualité scientifique et technique du logiciel, mais aussi son engagement envers les principes de la science ouverte.

Avec l'arrivée de la version 2.0, PPanGGOLiN intègre des améliorations méthodologiques, techniques et ergonomiques visant à renforcer son efficacité, sa robustesse et son accessibilité.

D'un point de vue méthodologique, plusieurs nouvelles fonctionnalités ont été intégrées pour enrichir et affiner l'analyse des pangénomes. L'ajout de métadonnées permet désormais d'annoter l'ensemble des éléments du pangénome (gènes, contigs, génomes, familles, arêtes, RGPs, spots et modules), améliorant ainsi la contextualisation et l'exploration des résultats. La projection des résultats du pangénome sur des génomes externes, ouvre la voie à une analyse comparative plus fine, facilitant l'intégration des pangénomes avec de nouvelles données génomiques. Par ailleurs, la clusterisation des RGPs apporte une nouvelle perspective en permettant de regrouper les régions de plasticité génomique en fonction de leur contenu en gènes, offrant ainsi un moyen de caractériser les dynamiques évolutives au sein d'un pangénome.

D'un point de vue technique, plusieurs optimisations ont considérablement amélioré les performances et l'efficacité de PPanGGOLiN. L'intégration de Pyrodigal (Larralde, 2022) en remplacement de Prodigal (Hyatt *et al.*, 2010) pour l'annotation des génomes a permis de réduire la consommation mémoire et d'accélérer le traitement des données en évitant la création de fichiers intermédiaires. La réorganisation des fonctions de lecture et l'optimisation du format HDF5 ont permis de réduire la taille des fichiers et d'accélérer les temps de chargement, rendant ainsi l'outil plus efficace et mieux adapté aux analyses à grande échelle.

1. ouvrirlascience.fr/remise-des-prix-science-ouverte-du-logiciel-libre-de-la-recherche-2023/

En termes de maintenabilité et de développement, des efforts importants ont été réalisés pour garantir la pérennité de PPanGGOLiN en tant que logiciel open source. L'adoption des bonnes pratiques de développement en Python (PEP) a permis d'améliorer la lisibilité et l'homogénéité du code, facilitant ainsi les contributions de nouveaux développeurs. L'automatisation du formatage du code avec Black (`https://github.com/psf/black`) et la mise en place d'une infrastructure de tests robustes (tests unitaires, tests d'intégration, tests fonctionnels) garantissent la stabilité du logiciel et minimisent l'introduction de bogues imprévus lors des mises à jour.

Enfin, une attention particulière a été portée à l'expérience utilisateur, un aspect essentiel pour un outil principalement destiné aux microbiologistes. La documentation a été entièrement réécrite, rendant son contenu plus accessible, structuré et pédagogique, et elle est désormais disponible sur le site ReadTheDocs. La gestion des messages d'erreur a également été améliorée, avec des messages plus explicites, facilitant aussi bien la correction des erreurs par les utilisateurs que le débogage par les développeurs. L'intégration de fichiers de configuration pour l'exécution des workflows simplifie l'utilisation de PPanGGOLiN dans des pipelines d'analyse complexes et renforce la reproductibilité des analyses, en accord avec les principes FAIR (Findable, Accessible, Interoperable, Reusable).

L'ensemble des développements et améliorations ont pu être présentés dans plusieurs conférences, sous forme de *flash talk* à local pangénome 2023 (Mainguy *et al.*, 2023) et de démonstration à JOBIM 2024. Un article présentant la version 2 de PPanGGOLiN est également en cours de rédaction et sera prochainement soumis.

4.2 . Évolution de PPanGGOLiN : vers une version 3.0 ?

L'évolution de PPanGGOLiN ne s'arrête pas avec cette version 2.0. Plusieurs axes de développement sont envisagés pour renforcer encore davantage ses capacités analytiques, améliorer ses performances et étendre ses possibilités d'utilisation.

L'implémentation actuelle de PPanGGOLiN repose sur Python et C, mais l'utilisation d'un autre langage pourrait permettre d'améliorer encore ses performances, notamment pour la gestion de grands volumes de données. Lors de la réécriture du code, une grande partie des variables ont été typées, ce qui est une étape préliminaire importante à un passage de Python vers Cython, un langage intermédiaire entre C et Python. L'intégration d'autres langages comme Rust pour la parallélisation ou Julia pour l'optimisation des calculs pourrait également être envisagée pour certaines parties critiques du code. Une de ces parties serait notamment celle en C qui exécute l'algorithme NEM et qui n'a pas été revue.

Avec la version 2, PPanGGOLiN permet désormais de projeter le pangénome sur un génome externe. Pour aller encore plus loin, nous voudrions ajouter la possibilité d'intégrer de nouveaux génomes dans un pangénome déjà existant, sans avoir à le reconstruire entièrement. Cette fonctionnalité permettrait une mise à jour progressive du pangénome à mesure que de nouvelles données sont disponibles, notamment dans le cadre de la création d'une base de données de pangénomes.

Cette base de données de pangénome est d'ailleurs en développement au LABGeM, sous le nom de PanGBank. Pour faciliter l'accès au pangénome et pour donner encore plus d'intérêt à la commande de projection, Jean Mainguy en train de développer une API (interface de programmation d'application ou *application programming interface* en anglais) qui permettrait de télécharger un pangénome depuis PanGBank. Cela accélérerait les analyses et faciliterait la standardisation des jeux de données pangénomiques.

Une autre avancée méthodologique majeure serait de représenter et stocker les séquences du pangénome sous forme de graphe de variants, plutôt que comme des séquences linéaires. Cette approche permettrait de grandement diminuer la taille des fichiers de pangénome.

Pour terminer, les améliorations présentées dans la version 2.0 de PPanGGOLiN ont été étroitement pensées pour l'intégration de PPanGGOLiN dans PANORAMA (cf. partie III). PANORAMA intègre des méthodes pour la comparaison de pangénomes et l'utilisation de la recherche de contextes génomiques pour identifier des systèmes biologiques conservés ou variables. Ces approches se basent sur le graphe de pangénome partitionné de PPanGGOLiN et pourraient permettre de mieux comprendre la dynamique évolutive des génomes microbiens et la diversité métabolique qu'ils contiennent.

CHAPITRE III DE LA GÉNOMIQUE COMPARÉE À LA PANGÉNOMIQUE COMPARÉE

Avec l'augmentation du nombre de génomes disponibles, les approches traditionnelles basées sur l'analyse de génomes individuels montrent leurs limites. Le concept du pangénome s'est peu à peu imposé et la construction de graphes est de plus en plus répandue pour étudier leur diversité génétique. Il est désormais possible de générer un grand nombre de pangénomes, offrant pour chaque espèce une vision complète de la variabilité génétique. La comparaison de ces pangénomes offre alors l'opportunité d'identifier leurs similarités et spécificités respectives, en considérant simultanément l'ensemble des gènes.

Dans cette perspective, j'ai développé PANORAMA, un outil dédié à l'intégration de méthodes de pangénomique comparée, facilitant ainsi l'analyse systématique des variations génétiques inter-pangénomiques.

1 - PANORAMA : un nouvel outil pour la pangénomique comparée

1.1 . Prédiction des systèmes biologiques dans les pangénomes

L'annotation des pangénomes est essentielle pour donner du sens aux analyses pangénomiques, que ce soit le partitionnement, la recherche de structures (comme les modules) ou des arbres phylogénétiques. Certains outils, tels que Panaroo (Tonkin-Hill *et al.*, 2020) et PanTools (Sheikhizadeh *et al.*, 2016), offrent la possibilité d'importer des annotations externes directement dans le graphe de pangénome. D'autres, comme PanGraph (Noll *et al.*, 2023), intègrent des méthodes d'alignement des éléments du pangénome (gènes ou familles de gènes) sur des bases de données fonctionnelles. PPanGGOLiN, quant à lui, intègre ces deux approches en ajoutant des métadonnées à tous les éléments du pangénome et en proposant une commande permettant d'aligner les séquences pangénomiques sur une base de données externe.

Toutefois, ces approches se limitent à l'annotation des gènes ou des familles de gènes, à l'exception des métadonnées intégrées dans PPanGGOLiN. À ce jour, aucune méthode ne permet, à notre connaissance, d'identifier des structures fonctionnelles plus complexes, telles que des systèmes biologiques, dans le graphe de pangénome.

La prédiction de systèmes biologiques dans les données génomiques, en particulier chez les procaryotes, repose sur un large éventail d'outils et de méthodes (*cf.* section 3.2). Cependant, ces approches sont centrées sur le génome individuel et ne prennent pas en compte l'ensemble de la diversité génétique de l'espèce. Or, intégrer cette diversité est crucial pour mieux comprendre l'évolution et le rôle fonctionnel de ces systèmes.

Afin de pallier cette limitation, j'ai développé PANORAMA, un outil de pangénomique conçu pour prédire des systèmes biologiques dans les graphes de pangénome générés avec PPanGGOLiN. Cette méthode repose sur 2 points clés : (*i*) Des modèles, similaires à ceux de MacSyFinder (Abby *et al.*, 2014), définissant des règles de présence/absence des gènes et leur organisation en synténie ;(*ii*) une adaptation de la méthode de détection des contextes génomiques que j'ai développée dans PPanGGOLiN.

1.2 . Comparaison des pangénomes

La majorité des études pangénomiques se concentrent sur la diversité génétique au sein d'une espèce ou d'un environnement donné. Bien que certaines recherches explorent le pangénome à des rangs taxonomiques supérieurs (Moulana *et al.*, 2020), les études comparant plusieurs pangénomes pour analyser la diversité entre différentes espèces procaryotes restent rares.

Parmi les quelques travaux existants, C. Hyun *et al.* (Hyun *et al.*, 2022) ont proposé une analyse comparative du pangénome de 12 espèces bactériennes pathogènes. Toutefois, leur approche ne repose pas sur le graphe de pangénome, mais sur la présence/absence des familles de gènes homologues dans les génomes. Leur étude se limite à la comparaison de certaines métriques associées aux pangénomes (telles que l'ouverture ou la loi de Heaps) ainsi qu'à l'annotation des familles de gènes.

À ce jour, cette analyse manuelle et spécifique à un jeu de données particulier semble être la seule existante qui compare des pangénomes. Aucun outil pangénomique ne permet encore une comparaison automatique et non spécifique de plusieurs graphes de pangénomes afin d'identifier des structures conservées ou spécifiques.

Dans cette optique, j'ai intégré dans PANORAMA de nouvelles méthodes exploitant le graphe de pangénome ainsi que la composition en familles de gènes de structures telles que les spots, les modules ou les systèmes. Ces méthodes permettent de rechercher des éléments conservés entre plusieurs pangénomes. À notre connaissance, PANORAMA est le premier outil offrant une comparaison automatisée de graphes de pangénome, ouvrant ainsi la voie à une meilleure compréhension de la diversité métabolique et de la dynamique évolutive des génomes procaryotes.

2 - Article : PANORAMA

Panorama: A robust pangenome-based method for predicting and comparing biological systems across species

Jérôme Arnoux[1,*], Jean Mainguy[1], Laura Bry[1], Quentin Fernandez de Grado[1], Vallenet David[1], Alexandra Calteau[1]

[1] LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS

*jarnoux@genoscope.cns.fr

Abstract

Over the last decade, the expansion in the number of prokaryotic genomes available has profoundly transformed the study of genetic diversity, evolution and ecological adaptation. However, traditional approaches based on the analysis of individual genomes are showing their limitations when faced with the sheer volume of data. To overcome these limitations, the concept of the pangenome has emerged, offering an overview of genetic diversity and evolutionary dynamics within a species. In this study, we present PANORAMA, an innovative pangenomic tool designed to exploit pangenome graphs and enable interspecific comparisons to explore genomic diversity. Based on the PPanGGOLiN software suite, PANORAMA incorporates advanced methods for annotating macromolecular systems at the pangenome scale and for comparative analysis of spots of insertion between different pangenomes. We illustrate the application of PANORAMA to a Pseudomonas aeruginosa dataset, evaluating its performance against reference tools such as PADLOC and DefenseFinder. The analysis was then extended to a wider set including four Enterobacteriaceae species, demonstrating PANORAMA's ability to annotate, compare and explore the diversity and distribution of antiphage defence systems beyond the species level. This work provides a new resource for the comparative study of bacterial genomes and highlights the relevance of genome-wide approaches for deciphering the evolutionary dynamics and ecological significance of bacterial defense repertoires.

Keywords: Pangenome, Comparative genomics, anti-phage defense systems, Comparative analysis, Pseudomonas aeruginosa, Enterobacteriaceae, Bioinformatics.

Introduction

The rapid expansion of bacterial genome sequencing over the past decade has provided unprecedented opportunities to study the genetic diversity, evolution, and ecological adaptation of microbial species [1, 2]. For a significant number of species, sequences of hundreds or even thousands of strains are now available. While this wealth of information offers immense potential for discovery, it also presents significant challenges, as traditional genome-centric approaches, which focus on individual genomes, are becoming increasingly inadequate for managing and interpreting such large-scale datasets. To address these limitations, the concept of pangenome has emerged as a powerful tool. It encompasses the entire gene repertoire of a species, including core genes present in all strains and accessory genes found in only a subset, and provides a holistic view of genetic diversity and evolution within a species [3]. Pangenomics has significantly transformed microbial genomics by providing a comprehensive framework for understanding genetic diversity and functional capabilities across microbial species [4]. This approach allows researchers to investigate not only the genome of a single strain but the complete gene repertoire within a species or group of strains, the pangenome, thereby enhancing insights into microbial evolution and adaptation.

Owing to the small size of their genomes and the large number of sequences available, particularly for species of clinical interest, pangenomic analysis of microbial genomes has benefited from the early development of tools, facilitating pangenome analysis, offering visualization, comparison, and partitioning of genomic data[5]. Among these tools, **PPanGGOLiN** stands out for its unique approach to analyze pangenomes by partitioning them with a statistical algorithm [6]. PPanGGOLiN represents genomic data as a pangenome graph at the gene family level, with nodes representing homologous gene families and edges capturing their genetic contiguity, enabling the compression of information from thousands of genomes while preserving the chromosomal organization of genes. A statistical model is applied to partition gene families in persistent genome (i.e. gene families found in nearly all genomes) and variable genome, which includes the shell and cloud components corresponding to intermediate- and low-frequency gene families, respectively. PPanGGOLiN includes additional methods for the identification of Regions of Genomic Plasticity (RGPs) and their spot of insertion (panRGP method) [7] and their fine description in conserved modules (panModule method) [8], which have demonstrated their utility in identifying genomic islands and provide helpful insights into the genomic adaptability and evolution of bacteria. Despite these advances, a significant challenge remains: detecting and comparing complex macromolecular systems at the pangenome scale. In microbial genomes, the genes responsible for macromolecular systems are usually arranged in a highly structured manner, typically clustered into one or several operons composed of functionally related genes. These clusters encode coordinated systems that play essential roles in microbial life. Among them are **secretion systems**, which allow the transport of proteins and other molecules across membranes to interact with the environment or host organisms; **defense systems**, which protect the cell from foreign genetic elements; and **metabolic pathways**, which organize enzymatic reactions to efficiently produce, transform, or degrade biological molecules. Understanding the organization and diversity of these systems is key to decoding the functional capabilities of microbial genomes. Several tools have been developed to detect macromolecular systems at the genome scale, including MacSyFinder [9], PADLOC [10], and Defense Finder [11], the latter two being specialized in identifying **bacterial anti-phage immune systems**. These tools are highly effective when applied to individual genomes; however, they are not designed to detect complex systems at the pangenome level, nor to enable systematic comparisons across large genomic datasets. Tools capable of building, comparing, and functionally annotating pangenome at the scale of thousands of genomes across multiple species remain limited. Some pangenomic tools allow the functional annotation of the pangenome by incorporating results from annotation tools, as do PanGGOLiN [6], Panaroo[12], or PanTOOLs[13], or also by aligning the pangenome to a sequence database as do PPanGGOLiN[6] or PanGraph[14]. None of these tools allows searching directly into the pangenome, but only to annotate with already known results. To date, no tools are available to construct, compare, and functionally annotate pangenomes at the scale of thousands of genomes from multiple species.

Here, we introduce **PANORAMA**, a powerful computational tool designed to harness bacterial

pangenome graphs from large genomic datasets and enable comparisons across species to explore genomic diversity. Built on the PPanGGOLiN software suite, PANORAMA incorporates advanced methods for reconstructing and analyzing pangenome graphs. It offers several key features, including the ability to compare genomic contexts between pangenomes and annotate macromolecular systems at the pangenome scale. Functional annotation of biological systems is performed directly on the graph structures using rule-based models, making it possible to map and analyze complex genomic features without relying on linear genome representations. To illustrate the versatility of our approach, we focused on the comparative analysis and annotation of bacterial defense systems. Bacteria have evolved a remarkably diverse array of defense mechanisms against phages and other mobile genetic elements. These range from well-characterized systems, such as restriction-modification (R-M) systems and CRISPR-Cas complexes [15], to more recently discovered and less understood systems like BREX [16], DISARM [17], and retron-based defense systems [18]. To date, over 150 systems have been described, unveiling an unsuspected diversity of molecular mechanisms [19]. This diversity is not only taxonomically widespread but also highly dynamic; defense systems are often associated with mobile genetic elements or genomic islands and can vary extensively in composition, organization, and presence both within and between closely related species [11, 20]. Despite this complexity, large-scale comparative studies of bacterial defense systems are still scarce. Pangenome-level analyses hold great promise for revealing patterns of co-occurrence, horizontal gene transfer, evolutionary innovation, and defense strategies that are specific to particular species or lineages. In this study, we present the methodology behind PANORAMA, a graph-based pangenomic framework designed to address these challenges. We first demonstrate its application on a comprehensive dataset of Pseudomonas aeruginosa genomes and compare its performance to established genome-scale tools such as PADLOC and DefenseFinder. We then extend this analysis to a broader dataset comprising four enterobacterial species, showcasing PANORAMA's capacity to annotate, compare, and explore the distribution and diversity of phage defense systems across entire genera. Overall, this work provides a powerful new resource for the comparative study of bacterial genomes and highlights the value of pangenomic approaches in revealing the evolutionary dynamics and ecological significance of bacterial defense repertoires.

1 Results and discussion

1.1 Overview of PANORAMA

To predict macromolecular systems, PANORAMA employs rule-based models similar to those used in MacSyFinder [21]. However, instead of applying these models to individual genomes, PANORAMA operates on the pangenome graph structure of PPanGGOLiN. The rules rely on the presence/absence of specific functions predicted from pangenome gene families, incorporating constraints on their genomic organization (i.e. gene colocalization). Functional annotation of gene families of the pangenome graph is performed through alignments with HMM protein profiles [22] defined for each macromolecular system. The genomic contiguity of gene families potentially involved in a system is then assessed on the pangenome graph by applying transitive closure and edge filtering. At the end, the predicted systems consist of sets of colocalized gene families from the pangenome graph, supplemented with information on their classification within the *persistent*, *shell*, or *cloud* genome, as well as their association with RGPs, modules, and spots of integration. Systems are also projected onto the genomes to determine their presence and gene content in each strain. An additional functionality of PANORAMA is its ability to compare pangenomes, identifying similar systems and insertion spots across species. Based on a set of predicted spots or systems in several pangenomes, PANORAMA computes a Gene Family Repertoire Relatedness score for each pair of elements by detecting shared gene families. It then applies a community clustering algorithm to group similar systems or insertion spots into clusters. More details about PANORAMA methods are provided in the Materials and Methods section.

PANORAMA is available as an open-source software written in Python and designed for easy installation to facilitate broader adoption by the community (https://github.com/labgem/PANORAMA). Software commands are organized into two main workflows (Fig. 1). The PanSystem workflow begins by annotating gene families of the pangenome graph using the specified HMM library, then applies system prediction rules from the model repository. Gene family annotations can also be performed externally and provided to PANORAMA by the user in a Tab Separated Values (TSV) file. The PanCompare workflow performs comparative analyses of two or more pangenomes, including gene family clustering and system/spot comparisons. Both workflows generate textual outputs (TSV files), graph-based representations (in GEXF or GraphML formats, compatible with Gephi or Cytoscape software for visualization), and figures to summarize results. Functional annotations and predicted systems are saved in the pangenome's HDF5 file, allowing further analyses. Additional utilities are provided to automatically convert system models and HMM libraries into the PANORAMA format, with support for models from MacSyFinder[9], DefenseFinder[11], CasFinder[23] and PADLOC[10]. Models are stored in JavaScript Object Notation (JSON) format with a flexible and easy-to-understand grammar, enabling users to customize or create new models.

1.2 System prediction benchmark

A set of 941 complete genomes of *Pseudomonas aeruginosa* was used to evaluate PANORAMA's defense system predictions against the reference tools, Defense Finder (including CasFinder models) and PADLOC. Although these two tools use a similar approach to predict defense systems, they differ in the number of models (281 in Defense Finder vs. 385 in PADLOC) and in the parameters used for predicting functions from HMM alignments, as well as for applying presence/absence and colocalization rules. Thanks to its generic system representation, PANORAMA is compatible with both tools and was run using their respective system models and HMMs after format conversion (i.e., PanSystem workflow).

To conduct this benchmark, we assessed whether PANORAMA correctly assigned pangenome gene families to the appropriate systems, based on the results from Defense Finder or PADLOC. As expected, we obtained highly similar results, achieving an F1-score of 99.11% (recall: 99.31%, precision: 98.91%) using PADLOC as a reference and 98.50% (recall: 99.71%, precision: 97.32%) with Defense Finder. As shown in Fig. 2a, a substantial number of families are shared exclusively between PANORAMA and either DefenseFinder (653 families) or PADLOC (879 families), while only 985 families are common.

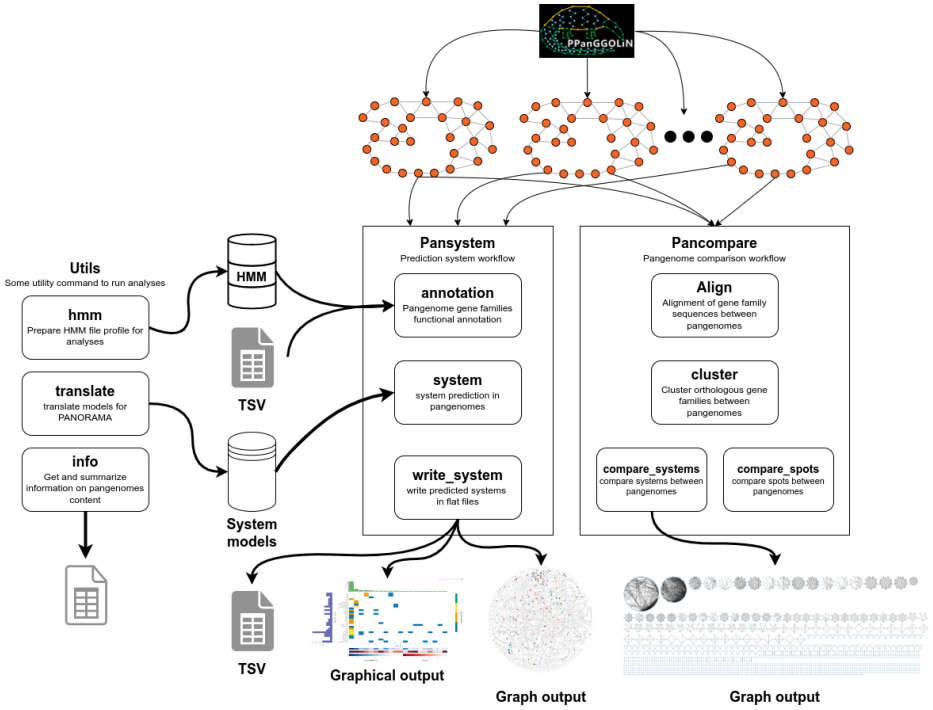


Fig. 1. PANORAMA software overview.Each rounded box represents a possible software command integrated into workflows depicted in square boxes. PANORAMA is organized into two main workflows: *Pansystem*, which focuses on system prediction and annotation within pangenomes, and *Pancompare*, which handles comparative analyses of pangenomes, including gene family alignment, clustering, and system/spot comparisons. Utility tools, such as HMM preparation and information summarization, facilitate data input and model translation. The software outputs include TSV files, graphical visualizations, and graph-based representations for comparing systems and pangenomes.

This highlights the complementary nature of these tools and demonstrates that the pangenome analysis provided by PANORAMA is highly valuable for reconciling their results. Besides, PANORAMA missed only a small number of systems, 25 for Defense Finder and 37 for PADLOC (Table 1). These cases correspond to missing annotations in the pangenome gene families. Unlike PADLOC and Defense Finder, which analyze each sequence individually for every genome, PANORAMA relies on the protein sequence of the family's representative gene for HMM alignments. Consequently, in rare instances, the alignment of the representative sequence falls just below the defined threshold, even though other sequences in the family have hits above it. Conversely, PANORAMA can identify additional results for certain genomes whose gene sequence alignment with the HMM is less conserved than with the representative sequence. A last point concerns the systems predicted only by PANORAMA (i.e., 147 with Defense Finder models and 44 with the ones of PADLOC). Although such a high number was not initially expected, a detailed analysis revealed that some of these additional systems arise because PANORAMA's genomic context search, based on colocalization rules in the pangenome graph, is less stringent than that of PADLOC and DefenseFinder. This relaxed approach allows PANORAMA to detect additional genes that are consistently conserved alongside known system components (see 'Materials and Methods' for further explanation). Other additional predictions stem from differences in HMM annotation due to the use of representative sequences for families, as mentioned earlier, but also from a bug in Defense Finder that affected only RM and Retron systems, where alignment thresholds were not properly handled by the software. This bug has since been corrected by the authors, though it was not tested in this study. Other discrepancies arise because PANORAMA allows multiple associations between gene families and systems, whereas PADLOC and Defense Finder link each gene to only one system. This is especially noticeable for systems with closely related models; for example, PANORAMA often predicts multiple Retron, RM, or Lamassu systems when other tools identify only one (Fig. 2b). One potential improvement would be to include *forbidden* families in the models to facilitate their differentiation, or to implement a scoring function, as done in MacSyFinder, to identify the best system. PANORAMA may also predict additional systems at the genome level for which the colocalization rule is not respected. These systems, flagged as *split*, have their genes separated, possibly due to rearrangements, insertion events, or assembly breaks occurring in a subset of genomes where they are predicted.

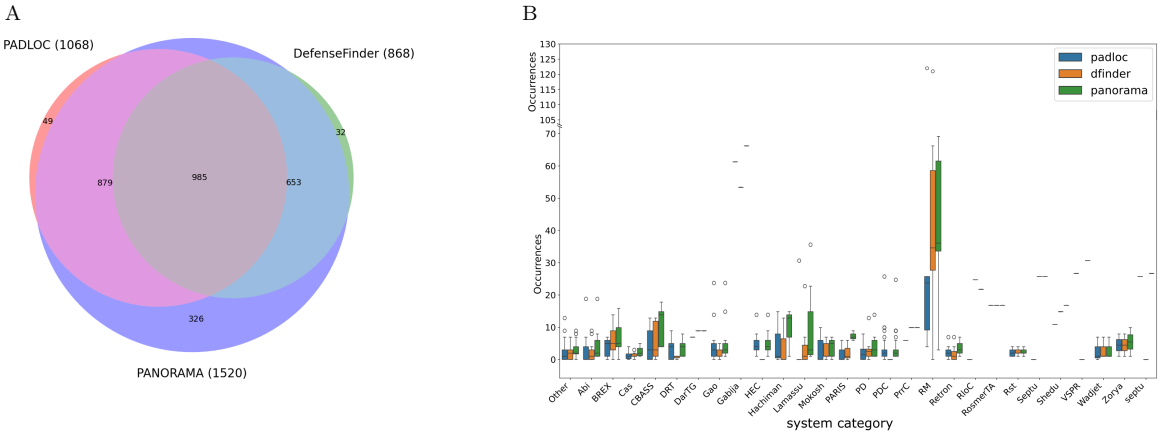


Fig. 2. Comparison of PANORAMA, PADLOC, and DefenseFinder system predictions at the pangenome level. PADLOC and Defense Finder predictions at the genome level were unified at the pangenome level by converting sets of system genes to sets of gene families. (A) Venn diagram illustrating the overlap of system families predicted by the three tools. (B) Boxplots displaying the distribution of system counts for each category, as predicted by the different tools.

Table 1. System prediction comparison between PANORAMA, PADLOC and DefenseFinder

Method 1 / Method 2	Common M1	Common M2	% common	Specific M1	Specific M2
PADLOC (1068) / DFinder (868)	502	506	52,07 %	566	362
PADLOC (1068) / PANORAMA (1064)	1031	1020	96,20 %	37	44
DFinder (868) / PANORAMA (976)	843	829	90,67 %	25	147

The tools were also evaluated for execution performance by measuring their runtime, CPU time and memory usage on a Linux server with 36 CPU cores (Table 2). Since PADLOC and Defense Finder are not designed to handle multiple genomes or parallelize computations, individual commands were executed for each genome, distributing the workload across all available CPU cores. PANORAMA significantly outperforms the other tools in runtime, being 3 to 10 times faster, while exhibiting similar memory usage. This efficiency is achieved by analyzing gene families rather than individual genes, which reduces computational overhead. Additionally, PANORAMA utilizes pyHMMER [24] for HMM alignments, optimizing the workflow by minimizing I/O operations and enabling on-the-fly result filtering. In contrast, other tools use the HMMER software directly [22], which necessitates post-processing steps.

Table 2. Benchmark results.

Tool (version)	Database (version)	#Systems predicted	Run Time (h)	CPU Time (h)	Peak Memory (GB)
DefenseFinder (1.2.2)	Defense-finder-models (1.2.4) & CasFinder (3.1.0)	881	1.55	29.02	6.84
PANORAMA (1.0.0)		976	0.51	0.82	11.22
PADLOC (2.0.0)	PADLOC-DB (2.0.0)	1090	2.58	88.72	9.24
PANORAMA (1.0.0)		1064	0.24	0.78	13.05

1.3 *Pseudomonas aeruginosa* defense system analysis

1.3.1 System prediction and analysis

Using the same set of *P. aeruginosa* genomes as for the benchmark, PANORAMA's defense system predictions with Defense Finder models were analyzed in greater detail. PANORAMA identified a total of 976 systems in the pangenome from 154 distinct models, with restriction-modification (RM) systems being the most abundant (Fig. 3). RM are present in 84% of genomes, with nearly 300 systems detected, of which type I systems are the most common, occurring 130 times. The Gabija, CRISPR-Cas, and CBASS system categories follow as the next most prevalent defense systems in genomes, each with a presence rate above 40%. These observations corroborate the study of Johnson *et al.* [25]. At the pangenome level, some system categories are highly prevalent across genomes but are represented by only a few distinct systems. For example, CRISPR-Cas systems appear in 48% of genomes, yet only 13 distinct systems are identified in the pangenome. This is further highlighted by a Shannon entropy calculation, which measures the compositional diversity of system categories (Fig. 3c). For CRISPR-Cas systems, the entropy is 1.35, indicating a high degree of conservation in their gene family composition across genomes. Among the most prevalent system categories, others show notable diversity, including RM, Gabija, and RloC. The RloC systems, for example, consist of 22 distinct systems spread across 35% of genomes, with a Shannon entropy of 21.58, indicating considerable variability in their gene family composition. These predictions are consistent with recent studies and highlight the remarkable diversity of anti-phage immune systems in prokaryotes [11].

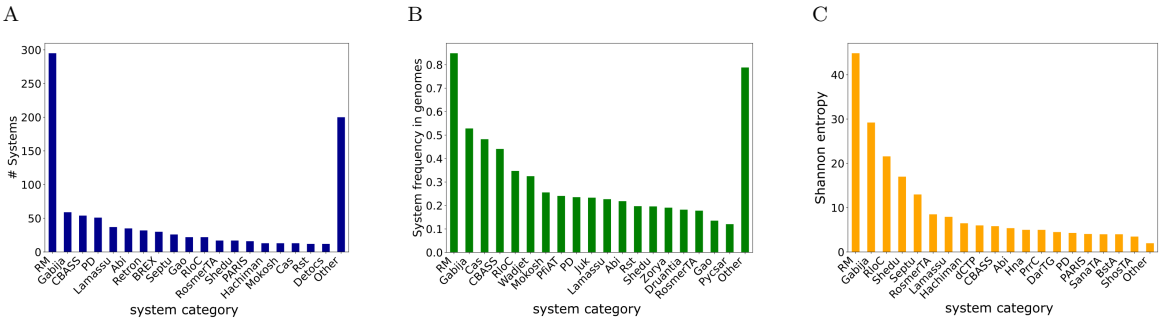


Fig. 3. System prediction metrics in *P. aeruginosa*. Systems are grouped by categories on the x-axis and ordered by decreasing values. Only the 19 highest-value system categories are displayed, with others grouped under the "Other" category. (A) Number of systems found for each category in the pangenome. (B) Relative frequency of system categories in genomes. (C) Shannon entropy of system categories.

1.3.2 Defense islands and spots of insertion

PANORAMA systems can be analyzed in conjunction with additional information extracted from the PPanGGOLiN pangenome graph, particularly concerning their association with the variable genome and their localization within RGPs and insertion spots predicted by panRGP [7]. This enables the identification of defense islands (i.e., variable regions enriched with defense systems) and their hotspots (i.e., frequently occurring insertion sites of defense islands in genomes).

Most defense systems are predicted within the variable (*shell* or *cloud*) genome of *P. aeruginosa* and are located in spots. PANORAMA identified 247 spots containing at least one defense system, representing 25% of all pangenome spots. Among them, 4 spots (7, 6, 45, 69) have a high frequency (¿25%) and exhibit the highest number of defense systems predicted at the pangenome level (¿60 systems) (Fig. 4). Notably, spots 7 and 6 are the most diverse, harboring 238 and 162 associated systems, respectively Fig. 5. They are mostly composed of RM systems (51% and 56%) but also exhibit a broad diversity of other categories, including BREX (6%), Gabija (5% and 4%), PD (4% in spot 7) and CBASS (4% in spot 6). These two spots were previously identified using a non-automated approach in the study by Johnson *et al.* [25] as core defense hotspots in *P. aeruginosa*, where they were designated CDHS-1 and CDHS-2. This further highlights the value and reliability of PANORAMA in automatically detecting defense islands and their insertion spots. Using PANORAMA, we also identified two additional defense hotspots (spots 45 and 69). Spot 45 contains 110 systems and stands out as the most balanced in terms of system categories; it is also the only hotspot with a notable presence of PARIS systems (8%). Spot 69, like spots 7 and 6, is dominated by RM systems, with PrrC systems specifically represented at 7%. Although less frequently observed across genomes (¡20%), spots 61 and 1 display highly diversified system content, comprising 72 and 65 systems, respectively. Both are also rich in RM systems, with Mokosh particularly represented in spot 1 and CBASS and PD systems notably present in spot 61 (8%). Finally, spots 4 and 9 are relatively frequent across genomes (30%) but contain few distinct systems, 31 and 10, respectively. These results highlight the potential of PANORAMA to provide a comprehensive landscape of defense systems in a species, enabling pangenome-scale analysis and the identification of defense islands with their hotspots of integration.

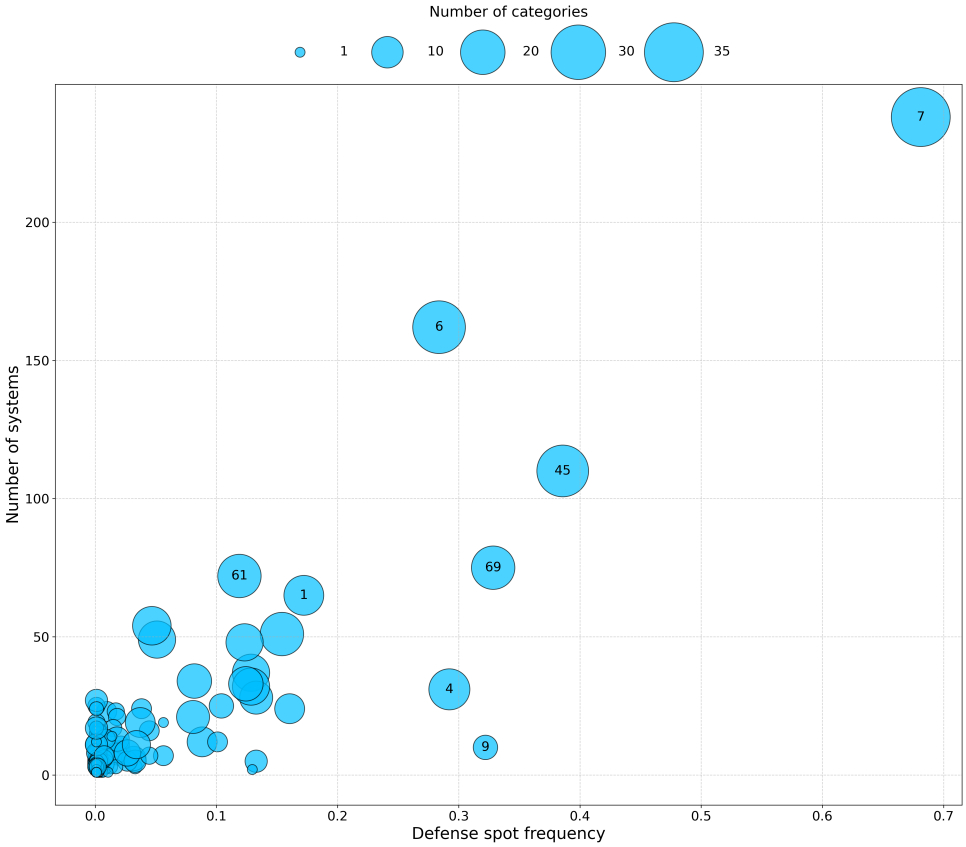


Fig. 4. System diversity and defense spot frequency in *P. aeruginosa*. This bubble plot displays the distribution of defense spots identified by PANORAMA, based on their frequency in genomes (x-axis) and the total number of defense systems identified within each spot at the pangenome level (y-axis). The size of the bubbles is proportional to the number of distinct system categories represented in each spot (legend at top shows scale).

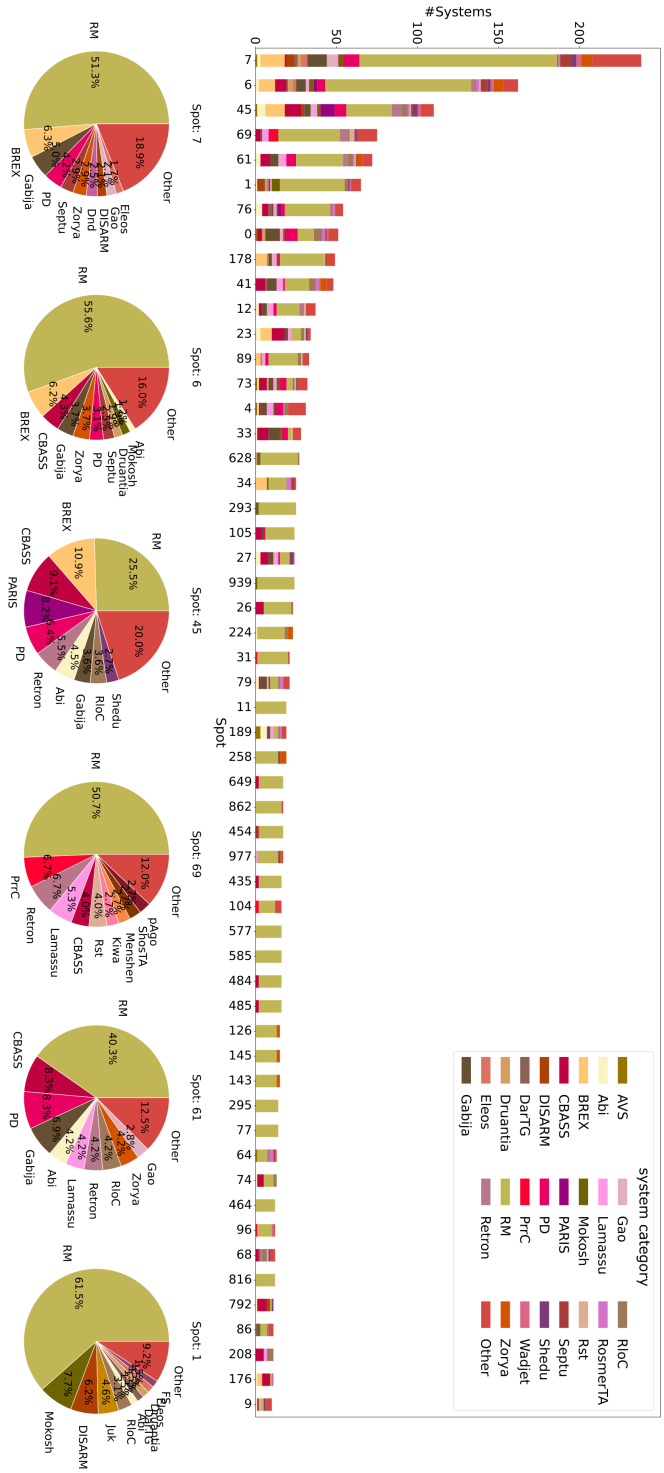


Fig. 5. Defense systems within insertion spots of the *P. aeruginosa* pangenome. The bar plot (top) displays the number of predicted defense systems in the *P. aeruginosa* pangenome for each insertion spot. Only spots with at least 10 systems are displayed. The pie charts (bottom) illustrate the system category composition for the six major insertion spots.

1.4 Pangenome comparison of Enterobacteriaceae defense arsenal

To demonstrate the comparative functionalities of PANORAMA (i.e., PanCompare workflow), the pangenomes of four Enterobacteriaceae species were analyzed for defense system and spot prediction. This dataset represent more than 6,000 genomes from *Citrobacter freundii*, *Salmonella enterica*, *Klebsiella pneumoniae* and *Escherichia coli* species (Table 3). The distribution of systems between species and their association with conserved spots were studied. Defense systems were predicted using Defense Finder models.

1.4.1 Defense system distribution in the four species

A total of 351, 461, 1005 and 1448 defense systems were predicted from the pangenomes of *Citrobacter freundii*, *Salmonella enterica*, *Klebsiella pneumoniae* and *Escherichia coli*, respectively. In addition to textual outputs, PANORAMA automatically generates a heatmap that displays system occurrences across the compared pangenomes (see Fig. S1). Among the different categories, RM systems are the most abundant in the four species, accounting for 30% to 45% of the systems found. Following RM systems, PD systems are the next most prevalent, representing 5% to 7% of the systems within Enterobacteriaceae (Fig. 6). Other notable system categories, such as Retron, CBASS, and Abi, are also relatively abundant across all pangenomes. Next, we evaluated the species-specificity of each system category by computing enrichment factors (Fig. S2). Our findings indicate that certain categories, Abi and Dnd in *S. enterica*, Juk and pAgo in *K. pneumoniae*, and Bunzi and RADAR in *C. freundii*, exhibit enrichment factors above 3, highlighting their preferential association with these species. Tiamat systems are only found in *S. enterica* and *C. freundii*. Interestingly, *E. coli* does not show any systems in higher abundance compared to other species, a sign of a more balanced diversity in its defense mechanisms.

1.4.2 Identification of conserved spots

With PANORAMA, we searched for similar spots based on their related gene families, using a gene family repertoire relatedness (GFRR, see Section subsection 3.2) threshold of at least 60%. This threshold guarantees at least one similar gene family on each side of the border. As shown in Fig. 7, we identified 99 clusters of similar spots, corresponding to 219 spots that have at least one spot with a similar bordering gene family composition with another one in an *Enterobacteriaceae* pangenomes.

E. coli is the species that shares the most spots with others (151), followed by *S. enterica* (131), then *C. freundii* (112) and *K. pneumoniae* (104). Proportionally to its number of spots, *C. freundii* is the species with the most similar spots, with 35% of its spots similar to the other pangenomes. The 2 species with the most common spots are *E. coli* and *S. enterica*, with 80 common spots. PANORAMA can then be used to highlight insertion spots at a higher taxonomic rank than the species. These could be areas of interest for research into shared evolution or exchanges between these species.

1.4.3 Spot identification and conservation

Using the comparative functionalities of PANORAMA, we identified 99 clusters of similar spots conserved in at least two of the four Enterobacteriaceae species (Fig. 7). As might be expected given their phylogenetic proximity, *E. coli* and *S. enterica* share the most common spots (i.e., 34 spot clusters, 25 of which are found only in these two species). About half of the spot clusters (n=47) are associated with a defense system in at least one species, comprising a total of 520 distinct defense systems among the 3,265 identified in the four species pangenomes. Of these, there is only one spot cluster (cluster 58) conserved in all pangenomes containing 34 defense systems. *E. coli* and *S. enterica* harbor the highest number of defense systems (263 systems) across their specific spot clusters (7 clusters). One of these clusters (spot cluster 409) includes spot 86 in *S. enterica* and spot 175 in *E. coli*. These spots rank sixth in terms of the number of defense systems, with 35 in *S. enterica* and 218 in *E. coli*, and can therefore be considered as potential defense hotspots. Analyzing their composition reveals notable similarities (Fig. S3). Both spots are mainly composed of RM systems (55% in *S. enterica*, 85% in *E. coli*), followed by BREX, CBASS,

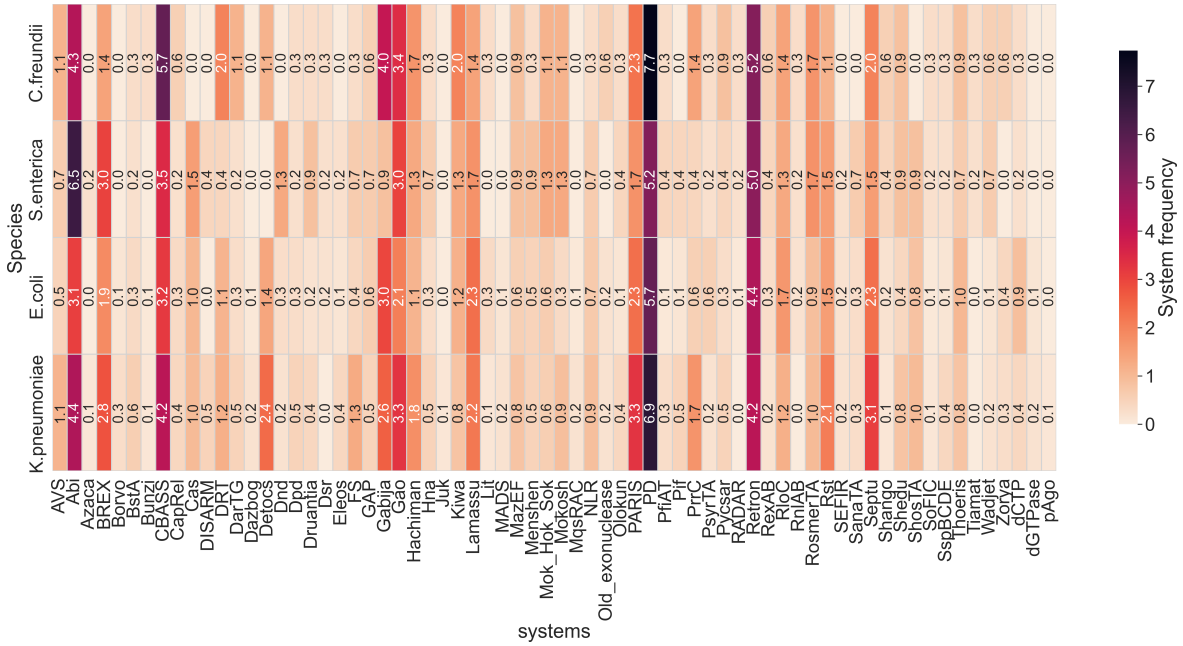


Fig. 6. Relative frequency of system categories in Enterobacteriaceae pangenomes. The relative frequencies are expressed as a percentage. RM system category was removed to get a clearer view.

and PrrC, which are generally not phage-specific. These findings support the hypothesis that defense systems may have been exchanged between the two species from this conserved hotspot. Examining the system category diversity of other spot clusters (Fig. 8), many clusters are predominantly composed of RM systems. Some clusters are dedicated to a single system category, such as BstA in cluster 2320, while others, like the previously mentioned cluster 58, are more diverse, bringing together systems from all four studied species.

Beyond their illustrative purpose, the analyses presented here highlight the ability of PANORAMA's comparative functionalities to identify conserved defense islands across species, providing valuable insights into the evolution of defense systems and their mechanism of acquisition.

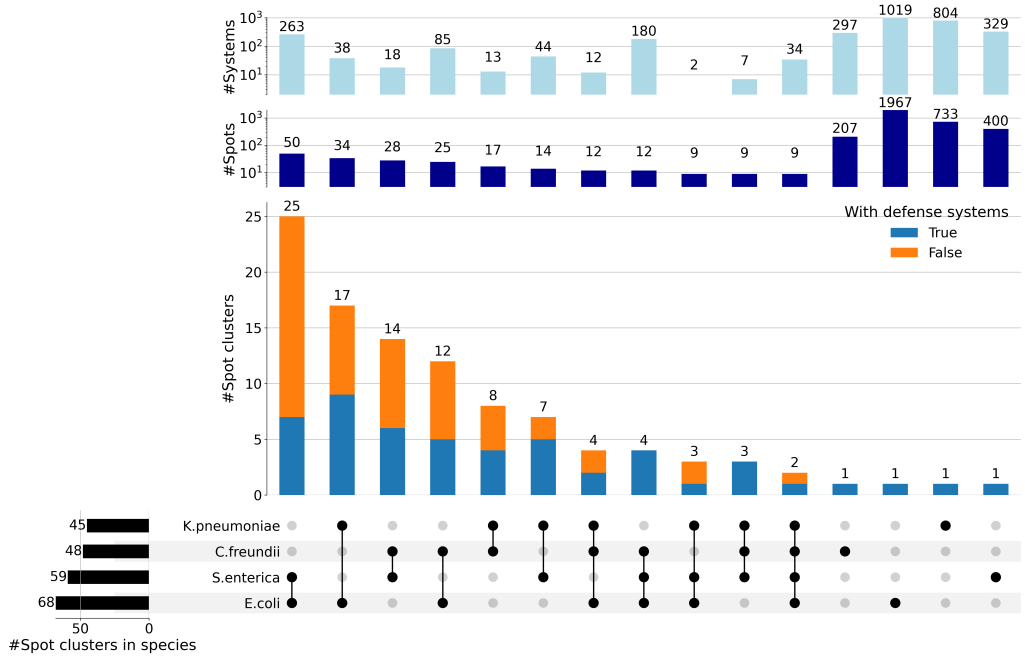


Fig. 7. Sharing of spot clusters across four species and their association with defense systems. The UpSet plot shows the number of spot clusters shared across the four compared species with stacked bars to indicate whether they contain defense systems (in blue) or not (in orange). The two top bar plots represent spot and defense system abundance metrics on a logarithmic scale.

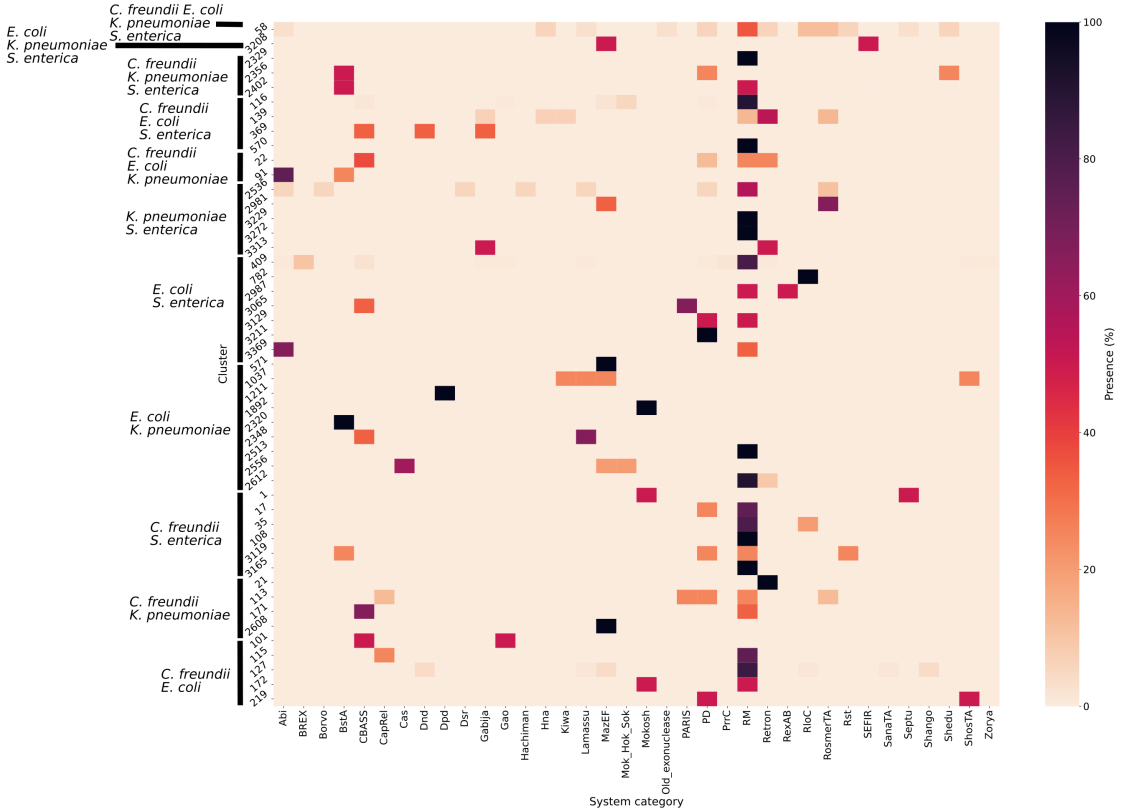


Fig. 8. Relative frequency of system categories in each spot cluster. The heatmap presents the relative frequency (%) of system categories per spot cluster. Species associated with each spot cluster are listed next to the cluster identifiers.

2 Conclusion

In this work, we introduced PANORAMA, a novel open-source framework for the prediction and analysis of macromolecular systems across prokaryotic pangenomes. Unlike existing tools that operate on individual genomes, PANORAMA leverages the pangenome graph structure provided by PPanGGOLiN to perform system-level analyses at both species and interspecies scales. It utilizes rule-based models inspired by those in MacSyFinder, but adapted to pangenomes, enabling the flexible and accurate identification of diverse functional systems, such as defense mechanisms.

Benchmarking against established tools like DefenseFinder and PADLOC demonstrated that PANORAMA achieves comparable prediction accuracy while reducing computational demands, making it particularly well-suited for large-scale analyses involving hundreds or thousands of genomes.

Beyond its performance, PANORAMA introduces key methodological innovations. Notably, its graph-based approach enables the detection of conserved genomic contexts by identifying gene families that consistently co-occur across multiple genomes. This strategy allows for the robust identification of evolutionarily maintained system components, even when their genomic proximity is disrupted in some genomes by rearrangements, insertions, or assembly fragmentation. As a result, PANORAMA is more flexible than conventional genome-based tools to detect atypical genomic architectures.

Applied to hundreds of genomes of a species, PANORAMA can quickly identify the different systems and their occurrence in individual genomes. A notable strength lies in its ability to associate systems with regions of genomic plasticity and insertion hotspots, allowing the identification of genomic islands enriched with systems and their preferred integration sites. In *P. aeruginosa*, four major hotspots were identified, including two that correspond to previously published core defense hotspots.

To further illustrate the comparative functionality of PANORAMA, we analyzed the defense repertoires of over 6,000 genomes from four Enterobacteriaceae species. PANORAMA revealed both conserved and species-specific system categories and was able to cluster insertion spots based on shared gene family content. This enabled the identification of conserved defense islands across phylogenetically related species, offering insights into the evolutionary conservation of these systems.

Altogether, PANORAMA provides a robust and extensible framework for system detection and comparative analysis at the pangenome level. Its flexible modeling format, compatibility with existing system model databases, and integrative workflows make it a valuable tool for microbiologists, bioinformaticians, and evolutionary biologists interested in understanding the distribution, function, and evolution of macromolecular systems. As the number of available genomes continues to grow, tools like PANORAMA will become increasingly critical in unveiling the complex architectures of microbial defense and other macromolecular systems.

Looking ahead, we plan to incorporate additional rule-based approaches to predict metabolic modules, which consist of the detection of genomic contexts encoding enzymes involved in the same pathway, leveraging pathway definitions from comprehensive databases such as KEGG [26] or MetaCyc [27]. We also aim to develop specific rules for detecting clusters of Carbohydrate-Active Enzymes (CAZymes) involved in polysaccharide biosynthesis and degradation. Furthermore, the pangenome graph framework provided by PANORAMA offers a promising avenue to explore genomic context and uncover novel systems. By incorporating fuzzy functional predictors, based on structural similarity or protein language models, we hope to detect new systems made of functionally analogous components co-localized within the pangenome graph.

3 Materials and Methods

3.1 Pangenome system detection workflow

To predict macromolecular systems, PANORAMA employs rule-based models that analyze the presence/absence of specific functions predicted from pangenome gene families while considering constraints on their genomic organization. The PanSystem workflow begins with gene family annotation using HMM libraries, followed by the identification and evaluation of genomic contexts within the pangenome graph based on system model rules (Fig. 9). Finally, the predicted systems at the pangenome level are mapped onto individual genomes to assess their presence and gene content.

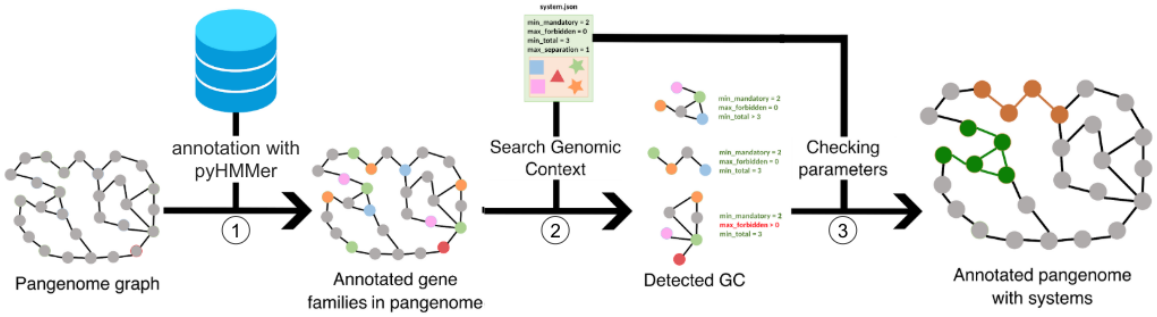


Fig. 9. PANORAMA PanSystem workflow. The detection workflow for macromolecular systems consists of multiple sequential steps: (1) gene family annotation using HMM profiles, (2) detection of genomic contexts from the pangenome graph containing annotated families, (3) checking model rules, (4) projection of systems on genomes.

3.1.1 System modeling

The Models used by PANORAMA for predicting macromolecular systems are similar to those of MacSyFinder [9] but differ in some aspects (Fig. 10). The primary components of system Models are Families (i.e. isofunctional protein families) instead of genes, and an additional hierarchical level is introduced to represent Functional Units. A Functional Unit is defined as a set of Families that work together to perform a necessary function for the system. Several Functional Units may be required for a system to operate effectively. This new level provides a more detailed and accurate description of systems, allowing the use of distinct rules for presence/absence and genomic organization both for the Families of a Functional Unit and between Functional Units. In PANORAMA, we also introduce the concept of canonical Models, which represent a more relaxed definition of systems. In PADLOC [10], these models are identified by the keyword "other" in their name. While these systems may not have been experimentally validated and might not be functional, their prediction can be valuable for identifying new systems or potential variants. During system detection, priority is given to non-canonical Models. A canonical Model is predicted only if its Families are not already associated with a non-canonical Model.

For presence/absence constraints, each Functional Unit and Family are categorized as *mandatory*, *accessory*, *neutral*, or *forbidden*. *Mandatory* elements are essential for the system. *Accessory* components are dispensable. They contribute to the system but may not be identified in all system variants due to rapid evolution or the absence of homology. *Forbidden* elements are incompatible with system functionality. They can help differentiate systems with shared components or distinguish inhibited systems. *Neutral* components are considered as associated functions but are not used to assess system predictions. However, they can serve as intermediaries in genomic context analysis by linking mandatory or accessory elements. To predict a system, quorum rules on component presence/absence are defined by two parameters: *minimum_mandatory* (the minimum number of mandatory elements required) and *minimum_total* (the minimum number of both mandatory and accessory elements). These parameters are specified at two levels: at the Model level, to assess the presence of Functional Units, and at the Functional Unit level, to evaluate predicted Families. Constraints on the genomic organization of a

system are governed by a *transitivity* parameter, which defines the maximum genomic distance between gene families in the pangenome graph corresponding to the system components (see 'Pangenomic context extraction' section).

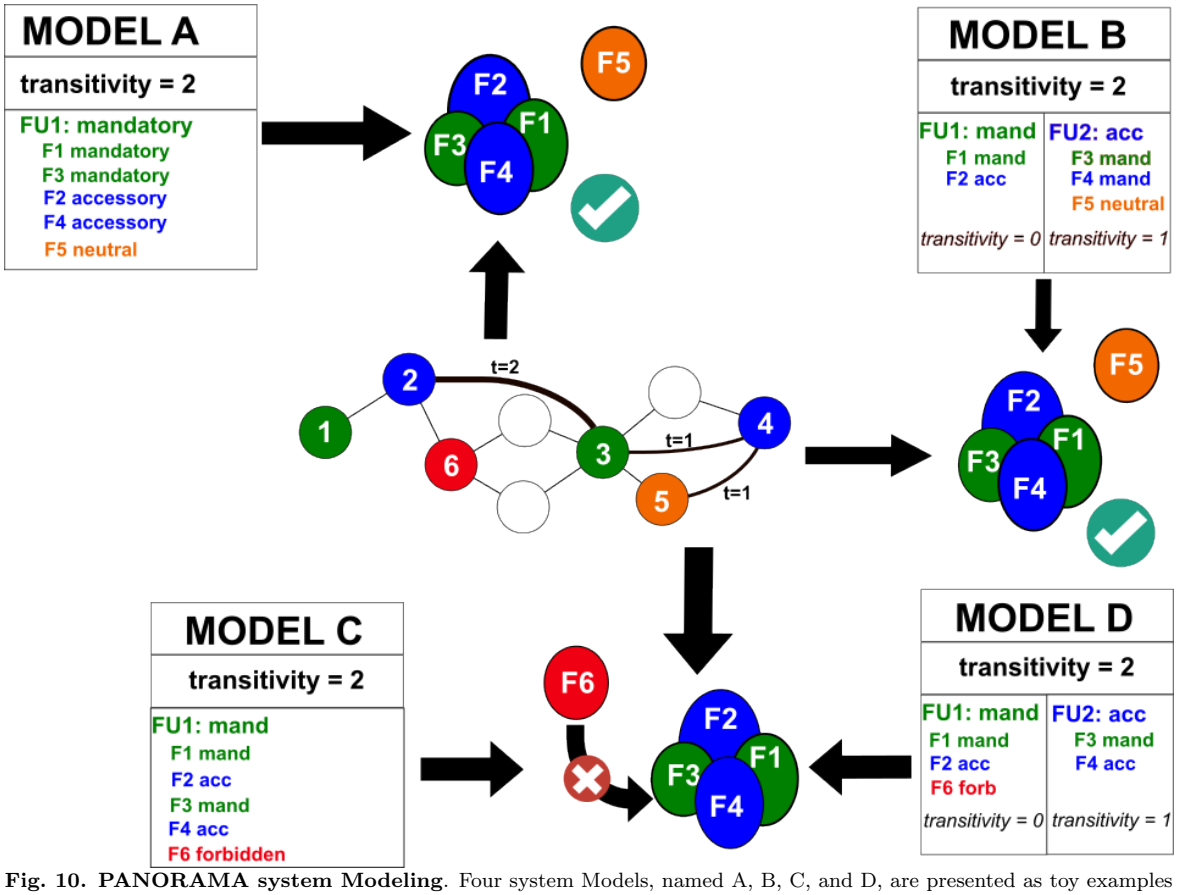


Fig. 10. PANORAMA system Modeling. Four system Models, named A, B, C, and D, are presented as toy examples to illustrate model rules. At the center, the corresponding genomic context extracted from the pangenome graph is shown. Nodes representing target families (i.e., model families) are labeled with Family numbers, and edges between them are bold, with dotted lines if they are from a transitive closure ($t >= 1$). Only transitivity edges relevant for model evaluation are represented. Functional Units (FU) and Families (F) are color-coded by their category: green for *mandatory*, blue for *accessory*, orange for *neutral*, and red for *forbidden*. Only models A (one FU) and B (two FU) are predicted. Others (C and D) are not predicted, as their extracted genomic context contains a *forbidden* Family.

3.1.2 Functional annotation of pangenome gene families

To annotate the pangenome graph, PANORAMA utilizes HMM libraries in which each HMM represents a specific Family defined in the system Models. Protein sequences of the family's representative genes are aligned to the HMM profiles using the hmmsearch method from the pyHMMER Python library [24]. PANORAMA also offers the ability to use the consensus protein sequence of gene families. It is determined by performing a multiple sequence alignment (MSA) excluding fragmented genes and then computing the consensus sequence with pyHMMER. To validate the alignments, a metadata file can be linked to an HMM library, specifying thresholds for each HMM based on various criteria, such as alignment coverage on the sequence or profile, score, e-value, and independent e-value. Alternatively, a global threshold can be applied to all HMMs. An option is also available to keep only the n best hits. Gene family annotation can also be performed externally using alternative prediction methods. In this case, a TSV file can be provided to describe the functions associated with each gene family.

3.1.3 Pangenome context extraction

For each system, connected components between gene families corresponding to Model Families (hereafter called target families) are searched for in the pangenome graph. According to the model rules on genomic colocalization of system components (i.e. the *transitivity* parameter), additional edges are added to the pangenome graph between families separated by less than t genes in the corresponding genomes. This corresponds to a partial transitive closure on the pangenome graph, enabling the connection of two target families even if their genes are not directly adjacent in the genomes. Next, a Jaccard index-based criterion (Equation 1) is computed on edges to evaluate the genomic context conservation (i.e. synteny conservation). For two families a and b connected by an edge $e_{a,b}$, the index $J_{a,b}$ is defined as the ratio of the number of genomes where the edge $e_{a,b}$ exists to the number of genomes in which at least one of the two gene families is present ($|\text{gen}_a \cup \text{gen}_b|$). To enhance the detection of systems present in a limited number of genomes, this index is computed locally, considering only genomes that contain the target families of the connected component rather than all the genomes of the pangenome. Edges having a Jaccard index below a defined threshold ($J_{a,b} < 0.8$ by default) are removed. The connected components with their remaining nodes are then evaluated in terms of presence/absence rules to validate system prediction. For each connected component, this process, which combines edge filtering and presence/absence rule validation, is applied iteratively, starting with the largest set of target families observed in a genome and then exploring other sets of target families, from largest to smallest, if they are not already included in a predicted system.

$$J_{a,b} = \frac{|\text{gen}_{e_{a,b}}|}{|\text{gen}_a \cup \text{gen}_b|} \quad (1)$$

Finally, the connected components corresponding to predicted systems at the pangenome level are mapped back onto individual genomes. Their occurrence in individual strains is assessed based on presence/absence and colocalization rules. They are classified in 3 distinct categories: *strict*, *extended* or *split*. A system is flagged as *strict* if the *transitivity* parameter is respected between genes belonging to the system. If additional genes of the connected component are found between those encoding the system, it is classified as *extended*. Indeed, applying transitive closure to the pangenome graph enables the detection of additional genes conserved with those of the system. This provides a more relaxed definition of colocalization rules than the strict intergenic space constraints employed by MacSyFinder or PADLOC. Lastly, *split* systems are those where system genes are separated by non-member genes of the connected component detected at the pangenome level or are located on different contigs. Such fragmentation can occur in a subset of genomes due to rearrangements, insertion events (e.g., insertion sequences), or assembly breaks.

3.2 Pangenome comparison workflow

The PanCompare workflow of PANORAMA enables the comparison of pangenomes and the identification of similar elements across species, such as macromolecular systems or spots of insertion. These elements are represented as sets of gene families. The first step consists in clustering all the gene families from the different pangenomes to identify groups of homologous gene families. Then, a Gene Family Repertoire Relatedness score for each pair of elements is computed. Finally, a community clustering algorithm is applied to identify clusters of elements sharing similar gene family content.

3.2.1 Gene families clustering

From all pangenomes to compare, representative protein sequences of gene families are extracted and clustered using MMSeqs2 cluster command [28] to obtain groups of homologous gene families (GGFs) sharing at least 50% of amino acid identity (–min-seq-id parameter) with an alignment coverage of 80% (-c parameter) by default. The following additional parameters are used: *–max-seqs* 400 *–min-ungapped-score* 1 *–kmer-per-seq* 80 *–alignment-mode* 2 *–cluster-mode* 1.

3.2.2 Gene Family Repertoire Relatedness score

To compare the family content of two pangenome elements, two Gene Family Repertoire Relatedness (GFRR) scores are computed using GGFs. For two pangenome elements, a and b, with their GFG content denoted as GFG_a and GFG_b, the minimal GFRR score, $minGFRR_{a,b}$, is defined as the ratio of the number of common GGFs to the minimum number of GGFs between a and b (Equation 2). Similarly, the maximal GFRR score, $maxGFRR_{a,b}$, is computed using the maximum number of GGFs (Equation 3).

$$minGFRR_{a,b} = \frac{|\text{GFG}_a \cap \text{GFG}_b|}{\min(|\text{GFG}_a|, |\text{GFG}_b|)} \tag{2}$$

$$maxGFRR_{a,b} = \frac{|\text{GFG}_a \cap \text{GFG}_b|}{\max(|\text{GFG}_a|, |\text{GFG}_b|)} \tag{3}$$

GFRR score computation can be applied to predicted macromolecular systems or spots of integrations. For spots, the two sets of gene families corresponding to spot borders are merged to compute the GFRR scores.

3.2.3 Pangenome element clustering

To identify similar elements (i.e. spots of insertion or systems) between pangenomes, GFRR scores are computed for each pair of elements. A graph is then constructed, where nodes represent pangenome elements, and edges are weighted by their corresponding GFRR scores. Edges are filtered according to a GFRR threshold ($minGFRR >= 0.6$ by default) to ensure strong similarity between connected elements. Finally, a Louvain algorithm [29], using NetworkX library implementation with GFRR weight on edges, is applied to the graph to identify non-overlapping communities corresponding to groups of similar elements between pangenomes. This functionality was used to identify conserved spots of insertion between species containing defense systems. A system is associated with a spot if all of its gene families are part of RGPs found within the boundaries of that spot.

3.3 Data, benchmark, and metrics

3.3.1 Genomic data and defense system prediction

Complete genomes of five species were downloaded from NCBI RefSeq [30] and analyzed for defense system prediction using Defense Finder (v1.2.2 with 239 models of v1.2.4 + 43 CasFinder models of v3.1.0), PADLOC (v2.0.0, 385 models of v2.0.0) and PANORAMA (v1.0.0). The dataset includes *Pseudomonas aeruginosa* (941 genomes) and four well-studied Enterobacteriaceae species: *Citrobacter freundii* (79 genomes), *Escherichia coli* (3,083 genomes), *Klebsiella pneumoniae* (1,659 genomes) and *Salmonella enterica* (1,380 genomes). Before running PANORAMA, the pangenomes of the five species were obtained using PPanGGOLiN v2.1.2 with default parameters and keeping the original RefSeq annotations. Pangenome metrics, including the number of families categorized as persistent, shell or cloud genomes, as well as the number of predicted RGPs and spots of insertion, are summarized in Table 3. These metrics are also available through the info command of PANORAMA. Models from Defense Finder and PADLOC (n=667), along with their associated HMMs (n=6,272), were converted in PANORAMA format using its internal conversion utility.

Table 3. Pangenomes content

Species	Genomes	Genes	Families	Edges	Persistent	Shell	Cloud	RGPs	Spots
C. freundii	79	385611	16865	24936	3780	2954	10131	3863	254
E. coli	3083	14447407	44278	140743	3018	7174	34086	240163	2036
K. pneumoniae	1659	8851686	32685	75073	4155	5085	23445	67678	778
P. aeruginosa	941	5811655	34058	61841	4925	6768	22365	44935	984
S. enterica	1380	6222501	23941	46535	3474	4080	16387	65920	458

3.3.2 Benchmark protocol

The *P. aeruginosa* genome dataset was used to evaluate PANORAMA's defense system predictions against the reference tools, Defense Finder v1.2.2 and PADLOC v2.0.0, using their respective system models and HMMs. To ensure consistency in comparison, the predictions from Defense Finder and PADLOC, originally consisting of gene sets associated with defense systems, were mapped to pangenome gene families by linking each identified gene to its corresponding family. Then, we define a True Positive (TP) as a gene family assigned to the same defense system by both PANORAMA and the reference tool. Gene families associated with a system by the reference tool but not predicted by PANORAMA are classified as False Negatives (FN). Gene families assigned to a different system or not predicted by the reference tool are considered False Positives (FP). To assess the performance of PANORAMA, we computed precision, recall, and F1-score using the following equations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4a}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4b}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4c}$$

The benchmark was conducted on a Linux server equipped with two Intel® Xeon® Gold 6150 processors (36 cores), 376 GiB of DDR4 RAM, running CentOS v7.9.2009 with kernel 3.10.0 and Python 3.10. Since Defense Finder and PADLOC cannot process multiple genomes simultaneously and parallelize computations, individual commands were executed for each genome, distributing the workload across the 36 available cores without concurrency. PANORAMA was run with the –threads parameter set to 36 to enable parallel computation. To evaluate execution performance, total run time, CPU time, and peak memory usage were measured for each tool. For consistency in comparison, the pangenome construction step made by PPanGGOLiN was not included in PANORAMA execution metrics as pangenomes are inputs of PANORAMA, like genomes for PADLOC and Defense Finder.

3.3.3 System category diversity

To assess the compositional diversity of systems within a given category (e.g., Restriction Modification, Gabija, CRISPR-Cas) in a pangenome, a Shannon entropy is calculated as follows:

$$H(C, p) = -\sum_{i \in C} P(sm_i) \log_2 P(sm_i) \tag{5}$$

Where:

- $H(C, p)$: the Shannon entropy of a system category C in a pangenome p.
- $P(sm_i)$: probability of a system model sm_i of C to occur in p, calculated for each system model as the ratio of the number of gene families of p associated with sm_i to the total number of gene families across all systems of C.

The higher the entropy, the more diversified the gene families that make up the systems in a given category.

3.3.4 System category enrichment

To determine the specificity of a system category in a set of pangenomes, an enrichment factor is computed as follows:

$$EF(C,p) = \frac{f_r(C,p)}{f_r(C,P)} \tag{6}$$

Where:

- C: a system category.
- P: a collection of pangenomes p.
- $f_r(C,p)$: the relative frequency of the system category C in the pangenome p, calculated as the ratio of the number of systems of C predicted in p to the total number of systems in p.
- $f_r(C,P)$: the relative frequency of the system category C in P, calculated as the ratio of the number of systems of C predicted in all $p \in P$ to the total number of systems across all $p \in P$.

An enrichment factor above 1 indicates that a system category appears n times more in a specific pangenome than in the entire collection of pangenomes.

Data availability

Acknowledgments

References

1. Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. en. *Functional & Integrative Genomics* **15.** Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 2 Publisher: Springer Berlin Heidelberg, 141–161. ISSN: 1438-7948. https://link.springer.com/article/10.1007/s10142-015-0433-4 (2025) (Mar. 2015).
2. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. en. *Nature Biotechnology* **36.** Publisher: Nature Publishing Group, 996–1004. ISSN: 1546-1696. https://www.nature.com/articles/nbt.4229 (2025) (Nov. 2018).
3. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". eng. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 13950–13955. ISSN: 0027-8424 (Sept. 2005).
4. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Current Opinion in Microbiology. Host–microbe interactions: bacteria • Genomics* **23,** 148–154. ISSN: 1369-5274. https://www.sciencedirect.com/science/article/pii/S1369527414001830 (2025) (Feb. 2015).
5. The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics* **19,** 118–135. ISSN: 1477-4054. https://doi.org/10.1093/bib/bbw089 (2024) (Jan. 2018).
6. Gautreau, G. *et al.* PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. en. *PLOS Computational Biology* **16.** Publisher: Public Library of Science, e1007732. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007732 (2022) (Mar. 2020).
7. Bazin, A., Gautreau, G., Médigue, C., Vallenet, D. & Calteau, A. panRGP: a pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics* **36,** i651–i658. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btaa792 (2022) (Dec. 2020).
8. Bazin, A., Medigue, C., Vallenet, D. & Calteau, A. *panModule: detecting conserved modules in the variable regions of a pangenome graph* en. Tech. rep. Section: New Results Type: article (bioRxiv, Dec. 2021), 2021.12.06.471380. https://www.biorxiv.org/content/10.1101/2021.12.06.471380v1 (2022).

9. Néron, B. *et al.* MacSyFinder v2: Improved modelling and search engine to identify molecular systems in genomes. fr. *Peer Community Journal* **3.** ISSN: 2804-3871. https://peercommunityjournal.org/articles/10.24072/pcjournal.250/ (2024) (2023).
10. Payne, L. J. *et al.* Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Research* **49,** 10868–10878. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkab883 (2022) (Nov. 2021).
11. Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. en. *Nature Communications* **13.** Publisher: Nature Publishing Group, 2561. ISSN: 2041-1723. https://www.nature.com/articles/s41467-022-30269-9 (2024) (May 2022).
12. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* **21,** 180. ISSN: 1474-760X. https://doi.org/10.1186/s13059-020-02090-4 (2024) (July 2020).
13. Jonkheer, E. M. *et al.* PanTools v3: functional annotation, classification and phylogenomics. *Bioinformatics* **38,** 4403–4405. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btac506 (2024) (Sept. 2022).
14. Noll, N., Molari, M., Shaw, L. P. & Neher, R. A. PanGraph: scalable bacterial pan-genome graph construction. *Microbial Genomics* **9.** Publisher: Microbiology Society, 001034. ISSN: 2057-5858. https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001034 (2024) (2023).
15. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. en. *Nature Reviews Microbiology* **18.** Publisher: Nature Publishing Group, 67–83. ISSN: 1740-1534. https://www.nature.com/articles/s41579-019-0299-x (2025) (Feb. 2020).
16. Goldfarb, T. *et al.* BREX is a novel phage resistance system widespread in microbial genomes. *The EMBO Journal* **34.** Num Pages: 183 Publisher: John Wiley & Sons, Ltd, 169–183. ISSN: 0261-4189. https://www.embopress.org/doi/full/10.15252/embj.201489455 (2025) (Jan. 2015).
17. Ofir, G. *et al.* DISARM is a widespread bacterial defence system with broad anti-phage activities. en. *Nature Microbiology* **3.** Publisher: Nature Publishing Group, 90–98. ISSN: 2058-5276. https://www.nature.com/articles/s41564-017-0051-0 (2025) (Jan. 2018).
18. Millman, A. *et al.* Bacterial Retrons Function In Anti-Phage Defense. eng. *Cell* **183,** 1551–1561.e12. ISSN: 1097-4172 (Dec. 2020).
19. Georjon, H. & Bernheim, A. The highly diverse antiphage defence systems of bacteria. en. *Nature Reviews Microbiology* **21.** Publisher: Nature Publishing Group, 686–700. ISSN: 1740-1534. https://www.nature.com/articles/s41579-023-00934-x (2025) (Oct. 2023).
20. Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolutionary Genomics of Defense Systems in Archaea and Bacteria*. en. *Annual Review of Microbiology* **71.** Publisher: Annual Reviews, 233–261. ISSN: 0066-4227, 1545-3251. https://www.annualreviews.org/content/journals/10.1146/annurev-micro-090816-093830 (2025) (Sept. 2017).
21. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. en. *PLOS ONE* **9.** Publisher: Public Library of Science, e110726. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0110726 (2022) (Oct. 2014).
22. Eddy, S. R. Accelerated Profile HMM Searches. en. *PLOS Computational Biology* **7.** Publisher: Public Library of Science, e1002195. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002195 (2024) (Oct. 2011).
23. Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research* **46,** W246–W251. ISSN: 0305-1048. https://doi.org/10.1093/nar/gky425 (2024) (July 2018).
24. Larralde, M. & Zeller, G. PyHMMER: a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics* **39,** btad214. ISSN: 1367-4811. https://doi.org/10.1093/bioinformatics/btad214 (2024) (May 2023).

3 - Conclusion et perspectives

3.1 . Prédiction de systèmes biologiques

Le développement de PANORAMA représente une nouvelle avancée pour l'analyse des pangénomes procaryotes, en proposant une approche pour l'identification des systèmes biologiques. Contrairement aux méthodes traditionnelles qui reposent principalement sur l'annotation de gènes individuels, PANORAMA exploite directement le graphe de pangénome, prenant ainsi en compte la diversité globale des génomes. La méthode de prédiction repose sur l'utilisation d'une base de données HMM et d'un ensemble de modèles spécifiques, permettant d'identifier et de contextualiser les systèmes biologiques.

En exploitant les pangénomes générés par PPanGGOLiN, il devient possible d'associer ces systèmes aux RGPs, aux spots et aux modules, ce qui enrichit l'annotation réciproque de ces éléments. L'utilisation du pangénome permet également d'identifier des systèmes autrement inaccessibles aux méthodes classiques, notamment en permettant l'annotation des fragments de gènes à partir des familles. De plus, cette approche permet la détection de systèmes "fractionnés", *i.e.* des systèmes dont les composants ne sont pas colocalisés dans un même génome, en raison d'insertions génomiques ou de leur position en bordure de contigs.

Dans la modélisation des systèmes, j'ai introduit un nouveau niveau de description, en regroupant les familles de gènes en unités fonctionnelles, ce qui améliore la caractérisation des modèles. PANORAMA ne possède pas de modèle propre, mais il permet de traduire des modèles issus de plusieurs bases de données existantes, telles que PADLOC, DefenseFinder et MacSyModels. Toutefois, la conversion entre ces bases n'est pas triviale : certaines différences dans les paramètres peuvent rendre la correspondance complexe, voire impossible. Néanmoins, la grammaire choisie et la lecture des modèles sont suffisamment simples et flexibles pour permettre l'écriture manuelle de modèles, et donc l'intégration de bases de données de systèmes propres et spécifiques.

Nous prévoyons d'étendre cette approche à d'autres bases, comme les modules métaboliques de la base de données KEGG (Kanehisa *et al.*, 2025). Ces modèles sont représentés sous une **forme normale disjonctive** (FND), qui utilise uniquement les opérateurs logiques *et*, *ou* et *non* (ce dernier ne pouvant s'appliquer qu'à un élément isolé). Si cette grammaire était standardisée à toutes les BD de systèmes, cela faciliterait les conversions entre bases et permettrait l'automatisation des traductions. De plus, elle ouvrirait la possibilité d'optimiser les expressions en appliquant des algorithmes classiques de minimisation, rendant ainsi la recherche de systèmes plus efficace. Toutefois, la FND présente certaines limites : elle ne prend pas en compte des paramètres quantitatifs comme la transitivité ou les règles de *quorum* et, dans le cas de modèles complexes, elle peut générer des expressions de taille exponentielle, difficiles à interpréter.

J'ai pu tester PANORAMA sur plusieurs bases de données, en particulier sur les modèles de défense contre les phages, où la méthode a démontré sa robustesse et son efficacité. Cependant, son application à d'autres modèles, comme ceux de conjugaison de MacSyFinder, a révélé certaines limites. Par exemple, une valeur élevée de transitivité (500 dans le modèle *T4SS_typeB* des plasmides) entraîne un goulot d'étranglement algorithmique. Cette difficulté provient d'une recherche de contexte extrêmement complexe, générant un graphe avec un nombre d'arêtes considérable, nécessitant ensuite un filtrage intensif. L'algorithme actuel n'est pas conçu pour traiter de telles valeurs. Une solution envisageable serait de distinguer les familles en unités fonctionnelles distinctes et d'introduire un mot-clé spécifique qui ne reconstruirait pas directement le contexte entre ces unités, mais se limiterait à rechercher un chemin d'une taille correspondant à la transitivité spécifiée.

3.2 . Comparaison de pangénomes

PANORAMA ouvre également la voie aux approches de pangénomique comparée. En recherchant et comparant des structures identifiées dans les graphes de pangénome, PANORAMA peut extraire les éléments conservés ou spécifiques à certains groupes d'organismes. Il devient alors possible d'identifier des liens évolutifs, de détecter des modules fonctionnels partagés et d'étudier l'émergence ou la disparition de systèmes biologiques, en prenant en compte toute la diversité génomique. Cette capacité à comparer les graphes de pangénomes offre un nouveau cadre d'analyse pour mieux comprendre l'évolution et l'organisation des génomes chez les procaryotes.

Pour ce faire, notre méthode repose sur un score (GFRR) permettant d'évaluer la similarité en contenu en famille de gènes entre les éléments des pangénomes. Ces éléments doivent au préalable être détectés au sein du pangénome afin d'assurer la robustesse de la comparaison. L'identification de structures conservées, *i.e.* des ensembles de familles de gènes partageant une organisation pangénomique similaire, reste un défi en cours d'exploration. Une piste repose sur l'application de méthodes de *machine learning*. En s'appuyant sur des modèles entraînés sur des structures connues, il serait possible de détecter et de classifier de nouvelles régions d'intérêt, ouvrant ainsi la voie à la prédiction de structures conservées inédites entre différentes espèces.

L'ensemble des développements méthodologique de PANORAMA ont pu être présentés dans plusieurs conférences, sous forme de *talk* et de poster (*cf.* Annexes A). Un article présentant PANORAMA sera prochainement soumis.

CHAPITRE IV BASE DE DONNÉES DE GRAPHE DE PANGÉNOMES

1 - Intégration de pangénomes dans une base de données orientée graphe

L'analyse des génomes procaryotes à travers les pangénomes ouvre de nouvelles perspectives et permet une approche renouvelée de l'étude des génomes et de leur évolution. Il existe quelques bases de données de pangénomes, comme Edgar (Blom *et al.*, 2009) ou panKB (Sun *et al.*, 2025), mais elles ne fournissent que des résultats d'analyses prédéfinies, sans possibilité d'extraire les pangénomes ni d'ajuster les paramètres pour réaliser de nouvelles études. De plus, l'exploration des pangénomes sous forme de graphe demeure inaccessible pour ces outils. À notre connaissance, il n'existe aucune base de données qui centralise plusieurs graphes de pangénomes tout en offrant des outils de diffusion, de visualisation et d'interrogation interactives. Pourtant, une telle BD de pangénomes constituerait une solution potentielle pour la gestion et la distribution des génomes, un enjeu d'autant plus crucial face à l'augmentation exponentielle du nombre de génomes disponibles.

Un pangénome pouvant être représenté sous forme de graphe, il est donc naturel d'adopter une base de données reposant sur une architecture de graphe pour structurer et interroger ces données. L'intérêt d'une base de données orientée graphe réside dans sa capacité à gérer efficacement ces structures non relationnelles, grâce à des systèmes de gestion optimisés et des langages de requête adaptés. L'outil PanTools (Sheikhizadeh *et al.*, 2016) utilise notamment le système de BD de graphe Neo4j `https://neo4j.com/` pour stocker un pangénome sous forme de graphe de De Bruijn, mais aussi pour analyser et visualiser le pangénome. C'est dans cette optique que nous avons développé une solution d'intégration des pangénomes dans une base de données orientée graphe, en collaboration avec des chercheurs spécialisés dans ce type de base de données.

Dans le cadre de l'édition 2022 du hackathon D4GEN [1], nous avons proposé un *challenge* ayant pour objectif d'améliorer l'efficacité du chargement des pangénomes dans les analyses comparatives réalisées avec PANORAMA. Une première solution a pu être développée par notre équipe, composée de Lucas Gruda et Sullian Le Bozec (étudiants à Télécom SudParis), Guillaume Gautreau (MaIAGE, INRAE et développeur de PPanGGOLiN), Stefania Dumbrava (SAMOVAR, Institut Polytechnique de Paris, Télécom SudParis, ENSIIE), spécialiste en bases de données et moi-même. Lors de ce challenge, notre principal défi a été l'intégration de plusieurs pangénomes dans une base de données orientée graphe, en utilisant le système Neo4j et le langage de requête Cypher [2]. Pour ce faire, nous avons d'abord réfléchi et défini un schéma pour la base de données, puis nous avons élaboré un modèle de données cohérent entre les pangénomes de PPanGGOLiN, les liens de similarités entre les familles de gènes et la structure de la base. Une fois les données intégrées, le chargement à l'ouverture de la base de données devenait pratiquement instantané. Enfin, nous avons également défini des requêtes permettant d'identifier les modules partagés entre les pangénomes, facilitant ainsi leur exploration et leur comparaison au sein de la base de données.

1. Le hackathon D4GEN est un challenge durant lequel des chercheurs et des entrepreneurs constituent une équipe avec des étudiants pour résoudre un problème de génomique ou de biotechnologie en 48 heures.
2. Ce projet nous a valu la troisième place du hackathon.

Encouragés par ces résultats, nous avons poursuivi nos travaux en collaboration avec Guillaume Gautreau, Stefania Dumbrava et Angela Bonifati (LIRIS, Université Lyon 1) experte en base de données. Nous avons commencé par améliorer la méthode d'intégration pour ajouter l'ensemble des éléments d'un pangénome sous forme de nœuds (gènes, familles de gènes, RGP, spots et modules) et leurs relations sous forme d'arêtes (appartenance d'un gène à une famille, voisinage entre familles, similarité entre familles issues de différents pangénomes...). Au moment de la conception de la méthode d'intégration, il n'existait aucun package d'intégration automatique dans Neo4J, j'ai donc repris plusieurs packages non propriétaires, conçus pour des applications spécifiques, afin d'intégrer les pangénomes dans la BD Neo4j.

Nous avons ensuite développé un workflow d'analyse présentant un ensemble de requêtes pour explorer les relations entre pangénomes et extraire des résultats d'intérêt, tels que le nombre de familles de gènes partagées entre différents pangénomes.

Nous avons appliqué ce workflow en tant que preuve de concept (*proof of concept* (POC)) afin d'évaluer sa pertinence pour l'identification des modules d'antibiorésistance partagés entre différentes espèces procaryotes. Pour cela, nous avons utilisé l'outil **PPanGGOLiN** pour construire un ensemble de **10 pangénomes** représentant les espèces du groupe **ESKAPE**, un groupe de bactéries connu pour sa résistance aux antibiotiques et son implication dans les infections nosocomiales. Les espèces sélectionnées pour l'étude ainsi que le contenu de leur pangénome sont décrits dans le tableau IV.1.1.

Pangénomes	# de gènes	# de génomes	# de familles	# d' arêtes	# de RGPs	# de spots	# de modules	taille du fichier (MB)
Acinetobacter baumannii	1 044 515	285	14 400	30 147	9 764	364	609	616
Enterobacter bugandensis	526 062	118	18 143	23 734	3 424	326	250	212
Enterobacter cloacae	651 827	137	22 953	32 270	6 083	292	526	358
Enterobacter hormaechei	739 490	159	18 166	29 798	5744	280	742	415
Enterobacter kobei	705 811	150	20 836	29 311	5 740	181	535	386
Enterobacter roggenkampii	978 031	210	26 080	40 459	8 807	319	712	537
Enterococcus faecium	570 257	207	7 889	18 627	6 195	189	318	301
Klebsiella pneumoniae	3 100 409	600	29 139	61 865	25 014	529	1 167	1 800
Pseudomonas aeruginosa	1 892 646	313	23 699	42 084	10 706	543	909	1200
Staphylococcus aureus	1 686 977	638	7 017	18 047	11 869	268	203	991

Table IV.1.1 – Description des données pangénomiques intégrées dans la base de données graphe.

Les familles de gènes ont été annotées avec la base de données CARD (Alcock *et al.*, 2023) pour rechercher des fonctions associées à la résistance aux antibiotiques. Grâce à notre approche, nous avons pu identifier des modules partagés impliqués dans la résistance aux antibiotiques au sein de ces espèces.

Ce travail, que j'ai eu l'opportunité de présenter lors du workshop SeaGraph de l'*IEEE International Conference on Data Engineering* (ICDE) 2024 (cf. Annexe A), a démontré l'intérêt et la faisabilité du stockage des graphes de pangénome dans une base de données optimisée pour cette structure.

2 - Article : Integrating Complex Pangenome Graphs

2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW) | 979-8-3503-8403-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDEW61823.2024.00052

Integrating Complex Pangenome Graphs

Jérôme Arnoux[1] Angela Bonifati[2] Alexandra Calteau[1] Stefania Dumbrava[3] Guillaume Gautreau[4]

[1]*Genoscope/LABGeM - CEA, CNRS, Paris Saclay University* [2]*IUF, CNRS LIRIS, Lyon 1 University*
[3]*SAMOVAR/Inst. Poltech de Paris, ENSIIE* [4]*MaIAGE, Université Paris-Saclay, INRAE*

***Abstract*—Graph databases are increasingly used to handle complex data pipelines, in which interconnected data is exploited for visualization and analytics. We propose a novel method, PanGraph-DB, for performing complex inter-pangenomic analysis within a graph database. As a case study, we focus on the antibiotic resistance in sequenced genomes. Over the past decade, the volumes of genomic data stored in public databases have grown exponentially, to the point of hindering comparative genomics algorithms. We show that, due to the nature of genomic data, graph databases enable accurate data and metadata analysis, visualization, and comparison across diverse genomes in the pangenomic approach. Families of graph-encoded pangenomes can then be integrated under a common mediated graph schema. The graph data integration allows to visualize and compare several pangenomes, as well as to analyze AntiMicrobial Resistance (AMR) gene niches through a combination of graph queries, whose performance and scalability we study.**

I. Introduction

Graphs are ubiquitous in several applications that rely on interconnected data to represent, explore, predict, and explain real- and digital-world phenomena. In the near future, graph ecosystems are expected to handle complex data pipelines, ranging from data pre-processing, querying, and analysis, to advanced processing, through learning and inference [1]. To optimize for performance and accuracy, such complex data pipelines need to be purposed for the particular tasks they target. In this paper, we focus on devising a custom methodology for enabling comparative genomics on pangenome graphs.

Typical analyses in comparative genomics often rely on a reference-centric approach to grasp species diversity, based only on several genomes. This reference genome, however, fails to provide sufficient coverage. A trivial solution would be to pairwise compare all the known genomes, but this would lead to a combinatorial explosion. The pangenomic approach overrides these limitations, by combining all the genomes, including the reference ones, in a unified data structure. This can be represented using various formalisms, *e.g.*, sets, Multiple Sequence Alignments, Sequence graphs, and De Bruijn graphs, as reviewed in [2]. Among these, microbial pangenome graphs increasingly rely on nodes, corresponding to clusters of similar genes (families), linked by edges, indicating their genomic neighborhood in various genomes [3], [4], [5], [6].

An open problem in pangenomics is how to compare several pangenome graphs straightforwardly. In particular, deciphering transfers of genetic information between species (pangenomes), such as AMR genes, raises many critical issues. The first hurdle is the size of the combined graphs, of the order of millions of nodes, requiring custom solutions for storage and efficient computation. Second, querying the graphs, to find similar modules for instance, can be difficult, in terms of both algorithmic and computational complexity.

We address these challenges in a practical system, by importing pangenome families in a unified property graph, under a mediated schema. The schema helps domain experts understand and explore the multi-pangenome graph and formulate graph queries that facilitate complex bioinformatics tasks, such as AMR analyses. Our method leverages the Neo4j system [7] and our queries are expressed in its native openCypher [8] language. These can, however, be equivalently encoded in any graph query language, including the future GQL [9] standard. We establish the *scalability* of the approach when varying the number of pangenome graphs, and its *efficiency* on custom AMR queries. Overall, *our work shows how to solve a complex domain-specific task, which has been considered unfeasible in classical genomics, by designing a dedicated graph processing pipeline*. To the best of our knowledge, *ours is the first work that uses graph databases for the real-world use-case of efficient and scalable multi-pangenome processing*. Our promising first results, obtained in the context of investigating the AMR of various pangenomes, open the perspective of employing this methodology in further applications. We make our PanGraph-DB artifact [10] publicly available.

Related Work. Graph databases have gained rapid adoption in the life sciences, as surveyed in [11] and as witnessed by creation of various datasets, *i.e.*, BioRDF for linked open data, GeneOntology for gene taxonomies, GProfile - for metabolism information, KEGG for gene and genome information, ChEBI for chemical entities, and the Genomic Data Model [12] for omics data. PanTools [13] is the work closest to ours, as it also uses the Neo4j graph database for comparative genomics. Their technique, though, hinges on a De Bruijn Graph and facilitates the integration of eukaryotic pangenomes. Compared to our approach, De Bruijn Graphs are at a lower level of granularity, which is subject to high variability and makes it challenging to scalably interpret and analyze functional and structural patterns across microbial species.

Tertiary data analysis has been carried out with Genomic Data Model (GDM) and GenoMetric Query Language (GMQL), to enable scientists and bioinformaticians to focus on the biological questions and the design of their experimental studies, instead of implementing the computational pipelines across different formats [12]. They focus, however, on genomic regions, and their comparisons are implemented as joins in the GMQL queries. Their query language is not

979-8-3503-8403-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDEW61823.2024.00052

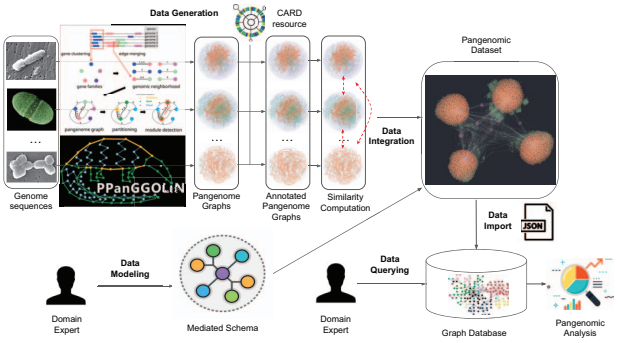


Fig. 1: Graph database driven pangenomic analysis.

graph-oriented and is thus not applicable to comparisons of families of pangenomic graphs and AMR identification.

II. Methodology

Our methodology aims to facilitate efficient comparative pangenomics. As such, we constructed the PanGraph-DB system, whose pipeline (see Figure 1) is capable of operating on pangenome graphs, computed by the PPanGGOLiN [5] framework, and of leveraging graph databases for integrated analyses by domain experts. Our approach is *system agnostic* and can be reproduced with any graph database whose data model and querying capabilities are comparable to those of the Neo4j system we employ. Given the rich *property graph data model* supported by Neo4j, domain experts can further enrich such pangenome graphs with custom properties. We illustrate this technique with a complex analysis aimed at inspecting AMR patterns in a multi-pangenome setting. Domain experts can, however, adapt and extend it to various other applications.

Data generation. Our pipeline (see Figure 1) takes as input complete ESKAPE genomes [14], from the NCBI GenBank database [15]. These are then processed by PPanGGOLiN, which performs gene clustering, edge merging, and statistical partitioning to compute pangenome graphs. Using these, we can highlight Regions of Genomic Plasticity (RGP) that are relevant to our analyses and that correspond to genomic spots (hotspots) onto which AMR genes can be integrated. We compute the information regarding *RGPs* using PPanGGOLiN's PanRGP method [16] and we further connect co-occurring and co-located gene families. Indeed, these might potentially be involved in a common biological process, as is commonly the case for AMR genes. Hence, we regroup them into structures called *modules*, with PPanGGOLiN's panModule method [17].

Gene families from these pangenome graphs are enriched with Comprehensive Antibiotic Resistance Database (CARD) annotations, in order to identify known AMR genes [18]. Also, various similarity levels between gene families of different pangenomes are computed, through dedicated alignment methods [19]. Finally, we obtain multiple CARD-annotated pangenome graphs, incorporating additional information regarding the partitions, RGPs, and modules the genomes relate to. This information, however, is not explicit and cannot be directly queried. Moreover, the graphs are largely disconnected, except for the similarity annotations between gene families.

Data modeling. As genomic analyses typically require a holistic view, encompassing information stored in all of these individual datasets, it is important to integrate them into a single graph. This is especially challenging, as the pangenome data lacks structural information and explicit labeling. To address this, we first construct a unifying schema that will shape the integrated multi-pangenome instance to be imported and analyzed in the Neo4j database. A key requirement for inter-pangenome analyses is having a data model that is expressive enough to capture multi-pangenome properties. Hence, for our AMR task, we need to enrich the previously computed datasets with further metadata, as follows. Each pangenome corresponds to a particular species, carries a mandatory name and unique identifier, and is linked to all the gene families it comprises. Note that all families must be associated with exactly one partition (persistent, shell, or cloud). To model the possible connections between families, we need to first determine whether they are part of the same pangenome. Intra-pangenome families can be marked as neighbors and, based on this, neighborhood weights can be computed, by counting their number of genomes. Inter-pangenome families can only be linked through similarity relations that can be characterized by a percentage of identity and a percentage of coverage. To facilitate the analyses, for each module, we explicitly store information regarding its gene families. Next, for each gene, we record its name, its *start* and *stop* position on the DNA sequence (*contig*), as well as its *RGPs*, which we connect to named *spots*, if they are co-located in the pangenome graph.

Data processing. The data import methodology closely follows that of the CovidGraph framework [20]. First, we create a Python dictionary for every pangenome. This has a hierarchical structure, whose parent is the pangenome itself and whose leaves are the genes. As such, we can use the *dict2graph* package [21] to create relationships between nodes, load properties for nodes and relations, as well as automatically index and merge all nodes and relationships into a graph. Note that all these tasks are parallelized. Next, we sequentially load the similarities of edges, by inspecting the alignment result table and extracting pairs of families with identity and coverage greater than 30% and 80%. Finally, we create edges between family nodes with the graphio package [22]. The expert user can then visualize, explore, and inspect the data through graph queries that can extract complex patterns [23].

III. Integrated Pangenomics

To efficiently perform genomic analyses on our graph dataset of connected pangenomes, we leverage the Neo4j *graph database*. This natively stores data as a *property graph*, *i.e.*, a directed, multi-labeled multi-graph with key/value properties attached to nodes and edges. To facilitate data integration, we design a *mediated schema* (Figure 2) that captures the integrated dataset structure. `Pangenome` nodes are connected to their genomic `Family` nodes and to neighboring and similar nodes of the same label in `Modules`. `Family`

Fig. 2: Mediated Schema.

nodes contain antibiotic resistance annotations. These nodes are classified as belonging to persistent, shell, and cloud `Partitions`, depending on whether their `Genes` are prevalent in all, some, or only a few corresponding genomes. Finally, `Genes` are part of `Contigs`, associated to particular `Genomes`, and can be part of `RGPs`, located in `Spots`.

Antibiotic Resistance Analysis. We illustrate the utility of graph databases for AMR identification, a key genomics task. Grasping the evolution of such resistance profiles across pangenomes is crucial to understanding how these elements spread between species. To conduct the analysis, the domain expert has to first comprehend the scale of such AMR profiles within the various pangenomes. She can extract aggregated information from the graph and identify the pangenomes with the highest number of relevant biological features at the intra-pangenomic level. Thus, she can express top-k queries [10] to focus on information about pangenomes (Q1, Q2, Q4), RGPs and Spots (Q3), as well as modules (Q5) that contain the most CARD-annotated families and top-k species in terms of modules (Q6). These queries allow gaining AMR insights about a species. However, to determine evolutionary patterns, one needs to analyze multiple pangenomes. Hence, a further step is to consider pangenome pairs and extract families that have *similar* AMR profiles. This can be encoded (Q7) through a main query that recovers, for a pangenome, its families and their modules, and two correlated sub-queries that further extract, for each previously identified family, all other similar families, from other pangenomes. Note that one has to explicitly export the variables reused in the correlated queries.

Q7. *Return the names of similar inter-pangenomic families and of the partitions and pangenomes they belong to.*

```
MATCH (p1:Pangenome)<-[:IS_IN_PANGENOME]-
(f1:Family)-[:HAS_PARTITION]->(s1:Partition)
WITH p1, f1, s1
MATCH (f1)-[:IS_SIMILAR]-(f2:Family)-[:
    HAS_PARTITION]->(s2)
WITH p1, f1, s1, s2, f2
MATCH (p2:Pangenome)<-[:IS_IN_PANGENOME]-(f2)
WHERE p1.name <> p2.name
RETURN p1.name, f1.name, s1.partition,
p2.name, f2.name, s2.partition
```

The query below is key to the analysis, as the information regarding `Spot` nodes allows highlighting RGP hotspots, *i.e.*, common insertion sites for Horizontal Gene Transfers.

Q8. *Identify similar inter-pangenomic AMR families where at least one has RGP-related genes, part of hotspots.*

Note that in the Cypher encoding, we explicitly name the graph patterns to filter intermediate results iteratively, using path-level reachability constraints. As such, we first specify the pattern allowing to identify families that have RGP-related genes and extract the graph objects connected to these, in particular the `Spot` information. We then refine the results with a path condition aimed at identifying pairs of similar families, within different pangenomes, and extract further information regarding their `Partition`. We also discard results wherein families do have not any AMR annotation.

```
MATCH
a=(p:Partition)<-[HAS_PARTITION]-(f1:Family)
-[:IS_IN_FAMILY]-(g:Gene)-[:IS_IN_RGP]-(r:RGP)
-[:IS_IN_SPOT]-(s:Spot)
WITH a,p,f1,g,r,s
MATCH b=(p1:Pangenome)-[:IS_IN_PANGENOME]-(f1)
-[:IS_SIMILAR]-(f2:Family)
-[:IS_IN_PANGENOME]-(p2:Pangenome)
WHERE f1.annotation IS NOT NULL AND z1<>z2
WITH a,b,p,f1,g,r,s,p1,zp2,f2
MATCH c=(f2)-[:HAS_PARTITION]-(p2)
RETURN a,b,c
```

Next, we can precisely compare the neighborhoods (contexts) of modules across pangenomes with the below query.

Q9. *Return the names and count of similar families from different pangenomes and the names of their modules.*

```
MATCH (m1:Module)<-[:IS_IN_MODULE]-(f1:Family)
      -[:IS_IN_PANGENOME]->(p1:Pangenome)
WITH f1, m1, p1
MATCH (p2:Pangenome)<-[:IS_IN_PANGENOME]-(f2:
    Family)-[:IS_SIMILAR]-(f1)
WITH f1, m1, p1, f2, p2
MATCH (f2)-[:IS_IN_MODULE]-(m2:Module)
WHERE f1.annotation IS NOT NULL AND p1 <> p2
RETURN p1, p2, m1, m2, f1, f2
```

Interestingly, families with similar CARD annotations may even belong to different partitions, highlighting the genetic variability between species, as analyzed with the query below.

Q10. *Return the names of the top 10 pairs of similar families, whose identity and coverage metrics surpass the 0.8 threshold, their partition names, and their CARD annotations.*

```
MATCH
(f1:Family)-[:HAS_PARTITION]->(s1:Partition)
WITH f1, s1
MATCH (f1)-[r1:IS_SIMILAR]->(f2)
       -[:HAS_PARTITION]->(s2:Partition)
WHERE r1.identity >= 0.8
AND r1.coverage >= 0.8
RETURN f1.name, s1.partition, f2.name,
s2.partition, r1.identity, r1.coverage,
f1.annotation, f2.annotation
ORDER BY r1.identity DESC LIMIT 10
```

Analyzing the query results, very similar inter-pangenomic families could indicate recent inter-species gene transfers.

IV. Experimental Evaluation

We performed scalability experiments on a virtual machine, running Ubuntu (version 22.04.1 LTS), equipped with an Intel Xeon Processor (Skylake, IBRS), clocked at 2.2Ghz, 153GB of RAM, and 621GB of free hard drive space. The queries were executed in the Neo4j Community Edition (version 4.4.12) and the dataset characteristics are available in our PanGraph-DB artifact [10]. Note that data import is one of the bottlenecks faced when processing multiple pangenomes. In previous analyses, even the simultaneous data loading of several PPanGGOLiN files (one HDF5 file per pangenome) was challenging in terms of memory usage. We deem our runtimes acceptable for massive cross-pangenome analyses, as they require less than a workday to fully import the datasets. The disk usage is also sustainable, given the compactness of storing 10 pangenomes (6.8GB), corresponding to billions of genes, thousands of families and genomes, as well as expert biological information (RGPs, Spots, AMR annotations).

Quantitative Analysis. We assess the *performance* and *scalability* of our methodology on the previous complex pangenomic analyses. As such, we analyze the runtimes of evaluating queries Q1-Q10 on pangenome datasets integrating an increasing number of pangenomes (see Figure 3). We note that the query execution times range from approx. 27.07 ms. on average (Q2) to approx. 66.8 ms. on average (Q8), when considering only two pangenomes and from approx. 31.8 ms. (Q5) to approx. 88.93 ms. (Q7), when considering ten pangenomes. The minimal execution times for Q2 and Q5 can be explained by the relative simplicity of the queries, as this computes count aggregates over a basic path comprised of only two edges. The maximal execution times are recorded for queries Q7 and Q8. Both are complex-correlated queries containing several subqueries. In terms of scalability, we note that the increase in execution time is nearly linear or sublinear when progressively adding more pangenomes. Moreover, we can see that the most complex queries, *i.e.*, Q3, Q7, and Q10, which also take longest to execute, exhibit the highest variability, as witnessed by their observed standard deviation. This indicates that performances start to deteriorate when considerably increasing data volumes, as the dataset integrating ten pangenomes records performance variations of up to 40%. Dealing with such scenarios requires further optimizations that graph processing systems are expected to support in the future. Nevertheless, the pangenome sizes we consider are already well beyond what can be supported by current methods that, moreover, do not support inter-pangenome analyses at all.

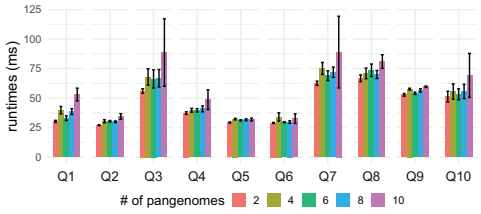


Fig. 3: Average Q1-Q10 runtimes on multiple pangenomes.

Qualitative Analysis. While resistance islands have been extensively studied in the microbiology literature [24], [25], [26], to the best of our knowledge, *ours is the first data-driven pangenomic approach to identify and compare them at scale*. Henceforth, we evaluate Q9 (Section III) on our largest dataset (ten pangenomes) and explore the results. Note that Q9 is relevant for the entire AMR pipeline, as it helps to identify similar inter-pangenome modules that contain AMR annotated genes. RGPs carrying such AMR modules may correspond to resistance islands, i.e., mobile genetic elements that contain multiple resistance genes and may be passed between species (pangenomes) through Horizontal Gene Transfers. It is thus particularly important to expose these genomic patterns to better understand their spread in ESKAPE bacteria.

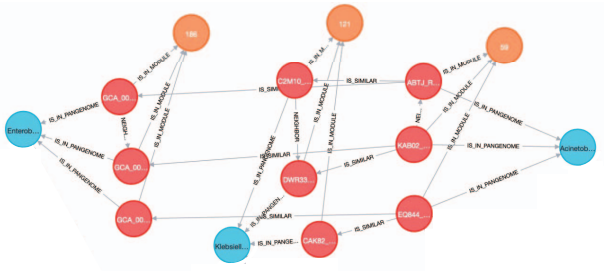


Fig. 4: Similar modules (orange) with AMR-related families (red) in distinct pangenomes (blue).

Inspecting the result graph, we can identify and extract relevant patterns. For example, between the *E. kobei*, *K. pneumoniae* and *A. baumanii* pangenomes (in blue), there are three pairs of similar gene families (in red) that are part of similar modules (in orange) in between each of the three species (see Figure 4). Having isolated these three modules, the expert user can inspect their membership in relevant resistance islands.

We notice that module identification is greatly facilitated by using a graph database, as all such pairs can be extracted with a single, declarative query. Previously, no such direct methodology for inter-pangenomic analysis and exploration was readily available. This qualitative study also revealed the importance of the underlying data model, which we will extend to explicitly represent resistance island conglomerates.

V. Conclusion and Perspectives

We have introduced the novel PanGraph-DB framework and showcased the usage of a graph database for complex pangenomic processes. By defining a mediated schema on top of several isolated families of pangenomes, we have created a unified pangenome graph that can be inspected for both intra- and inter-pangenomic analyses. We have also experimentally established the feasibility of data loading, processing, and querying with our method. Our approach is generic and can be deployed in other graph databases, as we make PanGraph-DB readily available to both the database and bioinformatics communities. We intend to optimize our pipeline to support even larger pangenomes and to extend our data model with other interesting types of genomic data, *e.g.*, metabolic pathways, defense, and virulence islands.

REFERENCES

[1] S. Sakr, A. Bonifati, H. Voigt, A. Iosup, K. Ammar, R. Angles, W. G. Aref, M. Arenas, M. Besta, P. A. Boncz, K. Daudjee, E. D. Valle, S. Dumbrava, O. Hartig, B. Haslhofer, T. Hegeman, J. Hidders, K. Hose, A. Iamnitchi, V. Kalavri, H. Kapp, W. Martens, M. T. Özsu, E. Peukert, S. Plantikow, M. Ragab, M. Ripeanu, S. Salihoglu, C. Schulz, P. Selmer, J. F. Sequeda, J. Shinavier, G. Szárnyas, R. Tommasini, A. Tumeo, A. Uta, A. L. Varbanescu, H. Wu, N. Yakovets, D. Yan, and E. Yoneki, "The future is big graphs: a community view on graph processing systems," *Communications of the ACM*, vol. 64, no. 9, pp. 62–71, 2021.

[2] Computational PanGenomics Consortium, "Computational pangenomics: status, promises and challenges," *Brief. Bioinform.*, vol. 19, no. 1, pp. 118–135, Jan. 2018.

[3] Y. Peng, S. Tang, D. Wang, H. Zhong, H. Jia, X. Cai, Z. Zhang, M. Xiao, H. Yang, J. Wang, K. Kristiansen, X. Xu, and J. Li, "MetaPGN: a pipeline for construction and graphical visualization of annotated pangenome networks," *Gigascience*, vol. 7, no. 11, Nov. 2018.

[4] S. C. Bayliss, H. A. Thorpe, N. M. Coyle, S. K. Sheppard, and E. J. Feil, "PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria," *Gigascience*, vol. 8, no. 10, Oct. 2019.

[5] G. Gautreau, A. Bazin, M. Gachet, R. Planel, L. Burlot, M. Dubois, A. Perrin, C. Médigue, A. Calteau, S. Cruveiller, C. Matias, C. Ambroise, E. P. C. Rocha, and D. Vallenet, "Ppanggolin: Depicting microbial diversity via a partitioned pangenome graph," *PLoS Comput. Biol.*, vol. 16, no. 3, 2020.

[6] G. Tonkin-Hill, N. MacAlasdair, C. Ruis, A. Weimann, G. Horesh, J. A. Lees, R. A. Gladstone, S. Lo, C. Beaudoin, R. A. Floto, S. D. W. Frost, J. Corander, S. D. Bentley, and J. Parkhill, "Producing polished prokaryotic pangenomes with the panaroo pipeline," *Genome Biology*, vol. 21, no. 1, p. 180, Jul. 2020.

[7] Neo4j, *Neo4j Graph Database*, Std., 2023. [Online]. Available: http://neo4j.org/

[8] ——, *OpenCypher*, Std., 2023. [Online]. Available: http://opencypher.org/

[9] GQL Standards Committee, *GQL*, Std., 2023. [Online]. Available: https://www.gqlstandards.org/

[10] (2023) PanGraph-DB. [Online]. Available: https://github.com/jpjarnoux/PanGraph-DB

[11] S. Timón-Reina, M. Rincón, and R. Martínez-Tomás, "An overview of graph databases and their applications in the biomedical domain," *Database J. Biol. Databases Curation*, 2021.

[12] M. Masseroli, A. Canakoglu, P. Pinoli, A. Kaitoua, A. Gulino, O. Horlova, L. Nanni, A. Bernasconi, S. Perna, E. Stamoulakatou, and S. Ceri, "Processing of big heterogeneous genomic datasets for tertiary analysis of next generation sequencing data," *Bioinformatics*, vol. 35, no. 5, pp. 729–736, 2019.

[13] E. M. Jonkheer, D.-J. M. van Workum, S. Sheikhizadeh Anari, B. Brankovics, J. R. de Haan, L. Berke, T. A. J. van der Lee, D. de Ridder, and S. Smit, "PanTools v3: functional annotation, classification and phylogenomics," *Bioinformatics*, vol. 38, no. 18, pp. 4403–4405, 07 2022. [Online]. Available: https://doi.org/10.1093/bioinformatics/btac506

[14] M. S. Mulani, E. E. Kamble, S. N. Kumkar, M. S. Tawre, and K. R. Pardesi, "Emerging strategies to combat eskape pathogens in the era of antimicrobial resistance: A review," *Frontiers in Microbiology*, vol. 10, pp. 1–24, 2019.

[15] E. W. Sayers, J. Beck, E. E. Bolton, D. Bourexis, J. R. Brister, K. Canese, D. C. Comeau, K. Funk, S. Kim, W. Klimke, A. Marchler-Bauer, M. Landrum, S. Lathrop, Z. Lu, T. L. Madden, N. O'Leary, L. Phan, S. H. Rangwala, V. A. Schneider, Y. Skripchenko, J. Wang, J. Ye, B. W. Trawick, K. D. Pruitt, and S. T. Sherry, "Database resources of the national center for biotechnology information," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D10–D17, Jan. 2021.

[16] A. Bazin, G. Gautreau, C. Médigue, D. Vallenet, and A. Calteau, "panRGP: a pangenome-based method to predict genomic islands and explore their diversity," *Bioinformatics*, vol. 36, no. Suppl_2, pp. i651–i658, Dec. 2020.

[17] A. Bazin, C. Medigue, D. Vallenet, and A. Calteau, "panmodule: detecting conserved modules in the variable regions of a pangenome graph," *bioRxiv*, 2021. [Online]. Available: https://www.biorxiv.org/content/early/2021/12/07/2021.12.06.471380

[18] B. P. Alcock, A. R. Raphenya, T. T. Y. Lau, K. K. Tsang, M. Bouchard, A. Edalatmand, W. Huynh, A.-L. V. Nguyen, A. A. Cheng, S. Liu, S. Y. Min, A. Miroshnichenko, H.-K. Tran, R. E. Werfalli, J. A. Nasir, M. Oloni, D. J. Speicher, A. Florescu, B. Singh, M. Faltyn, A. Hernandez-Koutoucheva, A. N. Sharma, E. Bordeleau, A. C. Pawlowski, H. L. Zubyk, D. Dooley, E. Griffiths, F. Maguire, G. L. Winsor, R. G. Beiko, F. S. L. Brinkman, W. W. L. Hsiao, G. V. Domselaar, and A. G. McArthur, "CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D517–D525, Jan. 2020.

[19] M. Steinegger and J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology*, vol. 35, no. 11, pp. 1026–1028, Nov. 2017.

[20] HealthECCO, "Covidgraph," https://covidgraph.org/ (visited: 07-11-2022), 2021.

[21] T. Bleimehl. (2023) dict2graph. [Online]. Available: https://git.connect.dzd-ev.de/dzdpythonmodules/dict2graph

[22] M. Preusse. (2023) Graphio. [Online]. Available: https://graphio.readthedocs.io/en/latest/

[23] A. Bonifati and S. Dumbrava, "Graph queries: From theory to practice," *SIGMOD Rec.*, vol. 47, no. 4, pp. 5–16, 2018.

[24] J. Hacker and J. B. Kaper, "Pathogenicity islands and the evolution of microbes," *The Annual Review of Microbiology*, vol. 54, pp. 641–679, 2000.

[25] O. Gal-Mor and B. B. Finlay, "Pathogenicity islands: a molecular toolbox for bacterial virulence," *Cellular Microbiology*, vol. 8, no. 11, pp. 1707–1719, Nov 2006.

[26] S. Algarni, S. C. Ricke, S. L. Foley, and J. Han, "The Dynamics of the Antimicrobial Resistance Mobilome of Salmonella enterica and Related Enteric Bacteria," *Frontiers in Microbiology*, vol. 13, p. 859854, 2022.

3 - Conclusion et perspectives

Notre *proof of concept* sur l'intégration et l'analyse de pangénomes dans des bases de données orientées graphe, ouvre la voie à de nouvelles méthodes de stockage, d'analyse et de visualisation de vastes ensembles de génomes et de pangénomes. Nous avons conçu un schéma de base de données ainsi qu'un workflow d'analyse générique, pouvant être adaptés à d'autres modèles de pangénomes. Nos tests montrent que notre approche permet des temps de réponse très rapides (de l'ordre de la milliseconde) aussi bien pour des requêtes simples que pour des analyses complexes. Bien que notre POC soit une réussite, plusieurs améliorations sont nécessaires pour en faire une véritable base de données opérationnelle dédiée à l'analyse et à la distribution des pangénomes.

Pour intégrer les pangénomes dans la base de données, j'ai développé un script Python utilisant **PPanGGOLiN** et **PANORAMA** afin d'extraire et de préparer les données. L'intégration dans Neo4j s'est appuyée sur des packages tiers, non optimisés pour nos données, entraînant des temps d'intégration relativement longs. Récemment, Neo4j a publié un package officiel, plus flexible et optimisé pour la communication avec la base. Les premiers tests, que j'ai menés, indiquent qu'il permettrait une réduction significative des temps d'intégration, ce qui constitue une piste prometteuse pour améliorer l'efficacité globale du système.

Lors d'un second hackathon D4GEN en 2023, l'intégration de **méthodes de machine learning** a été explorée dans notre workflow d'analyse. Neo4j propose plusieurs packages dédiés à l'application du ML directement sur la base de données. Nous avons testé différentes approches pour identifier des structures et des chemins pertinents dans le graphe de pangénome, mais aucun résultat probant n'a émergé. Cependant, cette première tentative nous a permis de repenser le schéma de la base de données et d'imaginer de **nouvelles métadonnées** à intégrer pour des analyses futures.

Pour construire la BD Neo4J, j'ai développé un script permettant de charger les pangénomes dans une base de données locale. Ce script avait d'abord été intégré à PANORAMA, pour sa capacité à lire plusieurs pangénomes. Toutefois, nous avons finalement choisi de développer un script indépendant, plus facile à maintenir. Ce script offre plusieurs avantages : intégration simplifiée des pangénomes dans la base, lancement automatique des analyses, indépendance vis-à-vis de l'interface Neo4j. L'objectif étant de rendre l'ensemble du processus plus accessible et automatisé, afin que chacun puisse créer une instance propre de base de données de pangénomes.

Le **LABGeM** développe une base de données de pangénomes générés par PPanGGOLiN, **PanGBank**. Ce projet repose sur une architecture de BD relationnelle **SQL**, contenant les fichiers **HDF5** pour chaque espèce et des résultats d'analyse accessibles en ligne. Une intégration avec notre POC pourrait offrir plusieurs avantages :

- Une **nouvelle manière d'organiser et de visualiser les pangénomes**
- La possibilité d'**effectuer des analyses complexes directement sur la base de données**
- Une meilleure interconnexion entre les données et les outils d'exploration

En combinant ces approches, nous pourrions développer un système de gestion des pangénomes plus robuste, interactif et performant.

CHAPITRE V CONCLUSION ET PERSPECTIVES

1 - Conclusions sur le travail de thèse

La pangénomique procaryote est un domaine en plein essor, bénéficiant de l'augmentation du nombre de génomes disponibles dans les bases de données. Bien que relativement récente dans l'histoire de la génomique, de la bioinformatique et, plus largement, de la microbiologie, l'analyse de pangénomes a déjà été largement adoptée comme outil de routine pour offrir une vision globale de la diversité génomique qu'elle permet d'explorer. Pour contribuer à ce domaine, l'objectif de ma thèse était de développer des méthodes permettant la comparaison des pangénomes en s'appuyant sur les graphes et les analyses de PPanGGOLiN, afin d'identifier des structures conservées entre différentes espèces.

Mon premier travail a ainsi consisté à concevoir une méthode de recherche de contextes génomiques dans le graphe de pangénome. Intégrée à PPanGGOLiN, cette méthode devait servir à la détection de contextes conservés à l'échelle des pangénomes. Afin de faciliter la comparaison des pangénomes, j'ai travaillé sur une nouvelle version de PPanGGOLiN, en collaboration avec Jean Mainguy, ingénieur de l'équipe, Alexandra Calteau, chercheuse au LABGeM, et David Vallenet, chef du laboratoire, ainsi qu'avec la contribution des anciens développeurs de PPanGGOLiN, Guillaume Gautreau (chercheur MaIAGE, INRAE) et Adelme Bazin (Ingénieur de recherche), également doctorants à l'époque. Cette version 2 de PPanGGOLiN apporte de nombreuses améliorations, notamment au niveau du stockage et de la lecture des données, mais aussi de nouveaux outils d'analyse, comme le *clustering* des régions de plasticité génomique. Cette version contient aussi une refonte du code selon les standards Python, une documentation renforcée ainsi qu'une restructuration du cycle de vie du logiciel. Ce travail de fond a constitué une étape clé de ma thèse, garantissant une base robuste pour la suite des développements. L'ensemble de ces contributions a été récompensé par le prix "science ouverte du logiciel libre de la recherche", "espoir" de la catégorie 'Scientifique et technique, décerné par le ministère de l'Enseignement supérieur et de la Recherche en 2023.

En parallèle du maintien et de l'amélioration de PPanGGOLiN, j'ai développé PANORAMA, le premier outil dédié à la prédiction de systèmes biologiques dans les pangénomes et à la comparaison de pangénomes. La première étape a été de concevoir la méthode de prédiction et d'élaborer les modèles associés, un travail réalisé en 2022 au cours du stage de 2e année de DUT de Laura Bry, que j'ai eu l'opportunité d'encadrer. Ces modèles offrent une représentation à la fois détaillée et flexible des systèmes biologiques, tout en restant facilement modifiables manuellement. Durant ce stage, nous avons également mis en place des méthodes de traduction des modèles de prédiction de systèmes de défense contre les phages, notamment PADLOC et DefenseFinder.

Parallèlement, j'ai poursuivi le développement de l'algorithme de prédiction de systèmes, qui a été affiné lors du stage de Quentin Fernandez De Grado en 2023, étudiant en 5e année à l'INSA de Toulouse, que j'ai co-encadré. Au cours de ce stage, nous avons testé et évalué la méthode sur des systèmes de défense, en la confrontant aux références de la littérature scientifique. Nous avons aussi associé les systèmes identifiés aux RGPs et aux spots pour rechercher d'éventuels îlots de défense. Ce travail a donné lieu à plusieurs présentations dans des conférences, sous forme de posters et de présentations orales (voir annexes A).

Dans le même temps, j'ai développé une approche dédiée à la comparaison des pangénomes, fondée sur l'analyse du contenu en familles de gènes au sein de structures déjà identifiées, telles que les RGPs, les spots, les modules et les systèmes. En calculant un score de similarité, cette méthode permet d'identifier des structures potentiellement conservées entre différentes espèces. En s'appuyant sur le pangénome, cette approche offre une comparaison globale du contenu génomique des espèces tout en intégrant l'ensemble de leur diversité.

Ces développements ont été intégrés à PANORAMA, un outil que j'ai conçu, tout en tirant parti de certains acquis de PPanGGOLiN. Il était donc nécessaire, en parallèle du développement des méthodes d'analyse, de concevoir un outil accessible à la communauté scientifique, facile à diffuser, bien documenté et conçu pour être maintenable sur le long terme.

Ce travail a donné lieu à un article qui sera prochainement soumis pour publication. L'article propose un benchmark comparatif entre PANORAMA et deux outils de référence pour la prédiction des systèmes de défense, DefenseFinder et PADLOC. Il inclut également une application sur le pangénome de *Pseudomonas aeruginosa*, où les systèmes prédits sont associés aux spots, afin d'identifier des îlots et hotspots de défense. Concernant l'aspect comparatif, j'analyse la conservation des spots entre quatre espèces de la famille des Enterobacteriaceae et les associe aux systèmes de défense afin d'identifier d'éventuels îlots de défense conservés au sein de cette famille.

Au cours de ma thèse, j'ai également contribué au développement d'une base de données orientée graphe dédiée aux pangénomes. Ce projet a débuté lors du hackathon D4GEN 2022, où, aux côtés de Guillaume Gautreau, Lucas Gruda et Sullian Le Bozec (deux étudiants de Telecom SudParis) ainsi que Stefania Dumbrava (SAMOVAR, Institut Polytechnique de Paris, Télécom SudParis, ENSIIE), nous avons remporté la troisième place. Cette initiative a donné lieu à une collaboration avec Stefania Dumbrava et Angela Bonifati (LIRIS, Université Lyon 1), toutes deux spécialistes des bases de données orientées graphe. Dans ce cadre, j'ai développé un script permettant la construction et l'intégration des pangénomes dans une base de données de ce type. Nous avons ensuite défini plusieurs requêtes visant à analyser et à comparer les pangénomes stockés. À titre d'application, nous avons intégré dix pangénomes d'ESKAPEE (un groupe d'espèces pathogènes résistantes aux antibiotiques) et effectué des recherches directes dans la base pour identifier des modules, calculés par PPanGGOLiN, similaires à ceux associés à des résistances. Ce projet a constitué une part importante de ma thèse, nécessitant l'acquisition de nouvelles compétences en bases de données orientées graphe et en langage de requête Cypher. Il a fait l'objet d'une publication et a été présenté lors du workshop SEAGRAPH de la conférence ICDE 2024, où il a suscité un vif intérêt au sein de la communauté informatique, ce type d'application et de données restant encore peu exploré.

Au cours de ma thèse, j'ai également eu l'opportunité de contribuer, de manière plus ponctuelle, à d'autres projets en lien avec la pangénomique. Ces collaborations m'ont permis d'échanger avec des chercheurs issus de disciplines variées, enrichissant ainsi ma compréhension des approches interdisciplinaires. Ces interactions ont également été l'occasion de développer ma capacité à adapter mon discours en fonction de mon public et de son niveau de connaissance.

Dans cette même perspective, j'ai eu la chance d'enseigner aux étudiants de première année du master GENIOMHE (Université Evry Val d'Essonne - Paris Saclay), dans le cadre d'un cours sur la génomique comparée et la présentation de la plateforme d'an-

notation des génomes procaryotes MicroScope. Cette expérience m’a permis d’affiner ma capacité à transmettre des concepts complexes de manière claire et accessible.

2 - Perspectives sur les méthodes développées

2.1 . Critique et amélioration possible des méthodes

2.1.1 . PPanGGOLiN et recherche de contexte génomique

La méthode développée pour la recherche du contexte génomique permet d'identifier les familles conservées entourant un ensemble de familles d'intérêt. Nos expérimentations montrent qu'elle est globalement efficace, mais qu'elle présente une sensibilité à la taille du contexte recherché ainsi qu'au paramètre de transitivité. Cette sensibilité semble être inhérente à l'algorithme de construction, qui repose sur un parcours exhaustif des génomes pour reconstituer le contexte. L'algorithme, après l'identification des familles cibles, procède à la construction du graphe de contexte et nécessite pour cela un retour aux génomes, une opération particulièrement coûteuse en termes de temps de calcul et de ressources. Ce choix méthodologique est lié à l'architecture du graphe de pangénome, qui encode uniquement le voisinage direct des gènes dans les génomes. Par conséquent, l'établissement des relations de transitivité exige un retour au niveau génomique. Un autre facteur limitant réside dans la nature multigénique de certaines familles du pangénome : plusieurs gènes appartenant à une même famille peuvent être présents au sein d'un même génome. Or, cette information n'est pas directement conservée dans le graphe, ce qui empêche de déterminer les relations de voisinage sans une analyse approfondie des génomes.

Ainsi, bien que la méthode actuelle soit bien adaptée à la recherche de contextes clairement définis et à des paramètres de transitivité modérés, son extension à la détection de systèmes complexes, notamment dans le cadre de PANORAMA, soulève déjà des défis méthodologiques. Des améliorations techniques sont envisageables, notamment la parallélisation du parcours du pangénome afin d'accélérer l'exécution de l'algorithme. Une autre approche consisterait à revoir l'étiquetage du pangénome pour y intégrer directement les informations nécessaires à la reconstruction du contexte, évitant ainsi un retour systématique aux données génomiques.

2.1.2 . PANORAMA et prédiction des systèmes

La prédiction des systèmes biologiques dans PANORAMA peut être divisée en 3 étapes : annotation fonctionnelle des familles de gènes, recherche de contextes génomiques et identification des systèmes dans le contexte.

L'étape d'annotation consiste à aligner une base de données HMM contre les séquences des familles de gènes. Cette étape est efficace et repose sur un outil externe : pyHMMER, laissant peu de place à des améliorations dans PANORAMA. Toutefois, on peut noter que les autres outils utilisent des seuils, comme la e-value, pour filtrer les résultats d'alignement. En utilisant le pangénome, le nombre de séquences par rapport à un génome peut fortement varier, ce qui modifie la e-value. Nous avons pu expérimenter dans notre benchmark des différences dans l'annotation entre les gènes et les familles de gènes à cause de tels seuils. Une amélioration possible serait de remplacer ces seuils par un critère d'alignement qui ne dépendent pas de la taille de la base de données.

Concernant la recherche de contextes génomiques et l'identification de systèmes, nous en avons déjà discuté précédemment dans le cadre de PPanGGOLiN. Toutefois, il convient d'ajouter que l'algorithme a été conçu pour être exhaustif, retournant ainsi tous les systèmes possibles. Pour améliorer son efficacité, nous pourrions envisager une réécriture de l'algorithme en intégrant des méthodes heuristiques. J'avais d'ailleurs commencé à développer des fonctions basées sur des algorithmes gloutons et hongrois, afin d'optimiser l'exécution de certaines parties du code jugées relativement lentes. En particulier, pour rechercher les systèmes rares dans les génomes, nous nous appuyons sur une approche combinatoire permettant d'identifier les systèmes potentiellement existants. Toutefois, lorsqu'elles ont été appliquées aux bases de données de systèmes de défense aux phages, ces optimisations n'ont pas significativement réduit les temps de calcul. De plus, une partie des systèmes n'était plus correctement prédite, suggérant un compromis entre rapidité et exhaustivité à explorer davantage.

Un autre point qui mériterait d'être exploré concerne la multiplication des résultats pour un même ensemble de familles. Il peut arriver que certains modèles soient proches en termes de composition et de paramètres, comme c'est le cas des systèmes de restriction-modification ou des systèmes CBASS. Avec PANORAMA, il est possible de prédire, pour un même ensemble de familles de gènes, plusieurs systèmes appartenant à la même catégorie ou non. Dans ce cas, les fichiers de sortie contiennent l'ensemble des systèmes détectés. Dans la version 2 de MacSyFinder, un calcul de score a été introduit pour pallier ce problème. Ce score repose sur la composition du système dans le génome (Néron *et al.*, 2023). Ainsi, dans MacSyFinder (sauf indication contraire), un gène ne peut appartenir qu'à un seul système, et celui ayant le meilleur score est sélectionné. À l'échelle du pangénome, un tel score pourrait toutefois s'avérer trop restrictif dans PANORAMA. Néanmoins, l'intégration d'un score fournirait une indication supplémentaire et permettrait, entre deux systèmes de la même catégorie, de filtrer les résultats et de limiter le nombre de prédictions. Si un tel score devait être mis en place, il devrait reposer sur la composition en familles du système (comme dans MacSyFinder), mais aussi prendre en compte la partition des familles. De cette manière, en ajustant des bonus et pénalités liés aux partitions, il serait possible de favoriser la prédiction des systèmes présents dans les parties variables ou conservées du pangénome.

2.2 . Perspectives et projets autour de la pangénomique

Lors de ma thèse, j'ai pu participer et discuter de plusieurs projets autour de la pangénomique.

Le premier projet auquel j'ai participé est PanGBanK. Son objectif est de constituer une base de données regroupant, pour chaque espèce, un pangénome construit avec PPanGGOLiN. De plus, cette base de données sera accessible en ligne, et des métriques ainsi que des analyses pangénomiques seront disponibles pour chaque pangénome. Une telle base constitue aujourd'hui un atout majeur pour l'ensemble des développements réalisés au cours de cette thèse, en particulier pour PANORAMA et la comparaison des pangénomes.

Parmi les autres projets auxquels j'ai pu contribuer, BlueRemediomics s'intéresse, entre autres, à la caractérisation des enzymes et des voies de biodégradation des filtres UV, ainsi qu'à la biosynthèse des exopolysaccharides (EPS) dans les bactéries marines. Mon implication a porté sur les EPS, en participant aux discussions sur l'utilisation des modèles et la recherche de systèmes dans les pangénomes. Au LABGeM, Jean Mainguy

travaille sur la définition de ces modèles. Une fois finalisés, ils permettront notamment de détecter les voies de biosynthèse des EPS dans les pangénomes, mais aussi de réaliser des analyses comparatives sur les systèmes identifiés.

3 - Perspectives sur la pangénomique et la génomique comparée

Au cours de ma thèse, j'ai développé et éprouvé des méthodes d'analyse appliquées à des jeux de données de grande envergure, en m'appuyant sur les relations phylogénétiques entre les génomes afin de construire des ensembles de données au niveau de l'espèce. L'objectif principal de cette approche est la constitution de familles de gènes présentant une similarité suffisante pour être considérées comme homologues, garantissant ainsi une certaine homogénéité fonctionnelle au sein de ces familles. Ce postulat est largement utilisé dans la construction des pangénomes, qu'ils soient basés sur des séquences génomiques ou sur des familles de gènes homologues. L'essor des études pangénomiques a permis d'étendre considérablement notre compréhension de la diversité intra-espèce, et une majorité des travaux actuels se concentrent sur l'analyse du pangénome à l'échelle d'une espèce.

Bien que le concept de pangénome ait été initialement appliqué à l'étude d'une espèce donnée, son intérêt réside dans la prise en compte et l'analyse de la diversité génomique à une échelle plus large. Cependant, peu d'études se sont jusqu'ici orientées vers une analyse pangénomique au niveau du genre, une approche pourtant prometteuse pour mieux comprendre l'évolution des génomes. Des travaux récents, tels que ceux de Jonkheer et al. (Jonkheer *et al.*, 2021), ont exploré cette voie en construisant et en analysant le pangénome du genre *Pectobacterium* à partir de 197 génomes répartis en 19 espèces, en utilisant l'outil PanTools (Sheikhizadeh *et al.*, 2016). PanTools présente l'avantage d'estimer de manière optimisée les paramètres de construction des familles de gènes homologues, permettant ainsi une analyse pangénomique plus robuste à l'échelle du genre. L'extension de ces approches constituerait une avancée majeure dans notre compréhension des dynamiques évolutives et fonctionnelles des génomes bactériens.

Un autre domaine prometteur concerne l'étude du pangénome de communautés microbiennes et son application en métagénomique. L'analyse pangénomique dans un contexte métagénomique permettrait d'obtenir une vision plus intégrative des interactions entre les espèces d'un même environnement, d'identifier des fonctions partagées et d'améliorer notre compréhension des dynamiques écologiques des génomes présents. En particulier, cette approche pourrait révéler des adaptations fonctionnelles clés ainsi que des éléments de co-évolution entre les espèces d'une même niche écologique. Par ailleurs, la métapangénomique ouvre des perspectives pour une meilleure caractérisation des microbiotes complexes, notamment en santé humaine, en agronomie et en écologie environnementale (Delmont et Eren, 2018)).

Les progrès et la démocratisation des méthodes d'apprentissage automatique pourraient largement contribuer à l'étude des pangénomes. Les algorithmes de *machine learning* permettent d'extraire des motifs complexes, de prédire des fonctions géniques et d'identifier des relations inédites entre génomes au sein d'un pangénome (Kavvas *et al.*, 2018). Par exemple, les modèles d'apprentissage profond peuvent être exploités pour classifier les gènes en fonction de leurs rôles biologiques, tandis que les approches de clustering non supervisé permettent de révéler des familles de gènes selon des critères encore inexplorés. Ces avancées méthodologiques ouvrent ainsi la voie à une compréhension plus fine de l'organisation et de la fonction des génomes microbiens.

La pangénomique demeure un domaine de recherche relativement récent, offrant encore de nombreuses perspectives d'amélioration pour affiner notre compréhension de la diversité et de l'évolution des génomes. Nous assistons aujourd'hui à une nouvelle étape dans le développement des méthodes pangénomiques. La première phase a été marquée par la définition et la construction des pangénomes, avec des outils comme PPanGGOLiN. La seconde s'est concentrée sur leur analyse et leur exploration, illustrée par des approches comme panRGP et panModule. Désormais, nous entrons dans une troisième phase, caractérisée par l'émergence de nouvelles méthodologies : la construction de pangénomes à des rangs taxonomiques plus élevés, l'essor de la métapangénomique, l'intégration de l'intelligence artificielle et, enfin, le développement de nouvelles approches comparatives, comme PANORAMA. Ces avancées méthodologiques permettront une meilleure appréhension de la dynamique évolutive des génomes et de leur rôle dans l'adaptation métabolique des organismes.

Bibliographie

Abby, S. S., Cury, J., Guglielmini, J., Néron, B., Touchon, M. et Rocha, E. P. C. (2016). Identification of protein secretion systems in bacterial genomes. *Scientific Reports*, 6(1):23080. Publisher : Nature Publishing Group.

Abby, S. S., Néron, B., Ménager, H., Touchon, M. et Rocha, E. P. C. (2014). MacSyFinder : A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLOS ONE*, 9(10):e110726. Publisher : Public Library of Science.

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A. A., Hoppenrath, M., James, T. Y., Karnkowska, A., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, D. J., Lara, E., Le Gall, L., Lynn, D. H., Mann, D. G., Massana, R., Mitchell, E. A., Morrow, C., Park, J. S., Pawlowski, J. W., Powell, M. J., Richter, D. J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F. W., Torruella, G., Youssef, N., Zlatogursky, V. et Zhang, Q. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology*, 66(1):4–119. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/jeu.12691.

Aframian, N. et Eldar, A. (2023). Abortive infection antiphage defense systems : separating mechanism and phenotype. *Trends in Microbiology*, 31(10):1003–1012.

Albers, S.-V. et Meyer, B. H. (2011). The archaeal cell envelope. *Nature Reviews Microbiology*, 9(6):414–426. Publisher : Nature Publishing Group.

Alberts, B. (1998). The Cell as a Collection of Protein Machines : Preparing the Next Generation of Molecular Biologists. *Cell*, 92(3):291–294. Publisher : Elsevier.

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A., Wlodarski, M. A., Edalatmand, A., Petkau, A., Syed, S. A., Tsang, K. K., Baker, S. J. C., Dave, M., McCarthy, M., Mukiri, K. M., Nasir, J. A., Golbon, B., Imtiaz, H., Jiang, X., Kaur, K., Kwong, M., Liang, Z. C., Niu, K. C., Shan, P., Yang, J. Y. J., Gray, K., Hoad, G. R., Jia, B., Bhando, T., Carfrae, L., Farha, M., French, S., Gordzevich, R., Rachwalski, K., Tu, M., Bordeleau, E., Dooley, D., Griffiths, E., Zubyk, H. L., Brown, E. D., Maguire, F., Beiko, R., Hsiao, W. W. L., Brinkman, F. S. L., Van Domselaar, G. et McArthur, A. G. (2023). CARD 2023 : expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research*, 51(D1):D690–D699.

Aldhebiani, A. Y. (2018). Species concept and speciation. *Saudi Journal of Biological Sciences*, 25(3):437–440.

Alkhnbashi, O. S., Costa, F., Shah, S. A., Garrett, R. A., Saunders, S. J. et Backofen, R. (2014). CRISPRstrand : predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics*, 30(17):i489–i496.

Alkhnbashi, O. S., Shah, S. A., Garrett, R. A., Saunders, S. J., Costa, F. et Backofen, R. (2016). Characterizing leader sequences of CRISPR loci. *Bioinformatics*, 32(17):i576–i585.

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Vesztrocy, A. W., Dalquen, D. A., Müller, S., Telford, M. J., Glover, N. M., Dylus, D. et Dessimoz, C. (2019). OMA standalone : orthology inference among public and custom genomes and transcriptomes. *Genome Research*, 29(7):1152–1163. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. et Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. et Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Ambroise, C., Dang, M. et Govaert, G. (1997). Clustering of Spatial Data by the EM Algorithm. *In* Soares, A., Gómez-Hernandez, J. et Froidevaux, R., éditeurs : *geoENV I — Geostatistics for Environmental Applications*, volume 9, pages 493–504. Springer Netherlands, Dordrecht. Series Title : Quantitative Geology and Geostatistics.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. et Yeh, L. L. (2004). UniProt : the Universal Protein knowledgebase. *Nucleic Acids Research*, 32(suppl_1):D115–D119.

Aravind, L. (2000). Guilt by Association : Contextual Information in Genome Analysis. *Genome Research*, 10(8):1074–1077. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab.

Argemi, X., Matelska, D., Ginalski, K., Riegel, P., Hansmann, Y., Bloom, J., Pestel-Caron, M., Dahyot, S., Lebeurre, J. et Prévost, G. (2018). Comparative genomic analysis of Staphylococcus lugdunensis shows a closed pan-genome and multiple barriers to horizontal gene transfer. *BMC genomics*, 19(1):621.

Asnicar, F., Thomas, A. M., Passerini, A., Waldron, L. et Segata, N. (2024). Machine learning for microbiologists. *Nature Reviews Microbiology*, 22(4):191–205. Publisher : Nature Publishing Group.

Avery, O. T., Macleod, C. M. et McCarty, M. (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. *The Journal of Experimental Medicine*, 79(2):137–158.

Aytan-Aktug, D., Clausen, P. T. L. C., Bortolaia, V., Aarestrup, F. M. et Lund, O. (2020). Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks. *mSystems*, 5(1):10.1128/msystems.00774–19. Publisher : American Society for Microbiology.

Aytan-Aktug, D., Nguyen, M., Clausen, P. T. L. C., Stevens, R. L., Aarestrup, F. M., Lund, O. et Davis, J. J. (2021). Predicting Antimicrobial Resistance Using Partial Genome Alignments. *mSystems*, 6(3):10.1128/msystems.00185–21. Publisher : American Society for Microbiology.

Backofen, R., Amman, F., Costa, F., Findeiß, S., Richter, A. S. et Stadler, P. F. (2014). Bioinformatics of prokaryotic RNAs. *RNA Biology*. Publisher : Taylor & Francis.

Bairoch, A., Boeckmann, B., Ferro, S. et Gasteiger, E. (2004). Swiss-Prot : Juggling between evolution and stability. *Briefings in Bioinformatics*, 5(1):39–55.

Balch, W. E., Magrum, L. J., Fox, G. E., Wolfe, R. S. et Woese, C. R. (1977). An ancient divergence among the bacteria. *Journal of Molecular Evolution*, 9(4):305–311.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. et Horvath, P. (2007). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science*, 315(5819):1709–1712. Publisher : American Association for the Advancement of Science.

Bastian, M., Heymann, S. et Jacomy, M. (2009). Gephi : An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1):361–362. Number : 1.

Batty, E. M., Chaemchuen, S., Blacksell, S., Richards, A. L., Paris, D., Bowden, R., Chan, C., Lachumanan, R., Day, N., Donnelly, P., Chen, S. et Salje, J. (2018). Long-read whole genome sequencing and comparative analysis of six strains of the human pathogen Orientia tsutsugamushi. *PLOS Neglected Tropical Diseases*, 12(6):e0006566. Publisher : Public Library of Science.

Bayliss, S. C., Thorpe, H. A., Coyle, N. M., Sheppard, S. K. et Feil, E. J. (2019). PIRATE : A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *GigaScience*, 8(10):giz119.

Bazin, A., Gautreau, G., Médigue, C., Vallenet, D. et Calteau, A. (2020). panRGP : a pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics*, 36(Supplement_2):i651–i658.

Bazin, A., Medigue, C., Vallenet, D. et Calteau, A. (2021). panModule : detecting conserved modules in the variable regions of a pangenome graph. Pages : 2021.12.06.471380 Section : New Results.

Bazinet, A. L. (2017). Pan-genome and phylogeny of Bacillus cereus sensu lato. *BMC Evolutionary Biology*, 17(1):176.

Beavogui, A., Lacroix, A., Wiart, N., Poulain, J., Delmont, T. O., Paoli, L., Wincker, P. et Oliveira, P. H. (2024). The defensome of complex bacterial communities. *Nature Communications*, 15(1):2146. Publisher : Nature Publishing Group.

Beckstette, M., Homann, R., Giegerich, R. et Kurtz, S. (2006). Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, 7(1):389.

Beier, S. et Thomson, N. R. (2022). Panakeia - a universal tool for bacterial pangenome analysis. *BMC Genomics*, 23(1):265.

Bernheim, A. et Sorek, R. (2020). The pan-immune system of bacteria : antiviral defence as a community resource. *Nature Reviews Microbiology*, 18(2):113–119. Publisher : Nature Publishing Group.

Bertani, G. et Weigle, J. J. (1953). Host controlled variation in bacterial viruses. *Journal of Bacteriology*, 65(2):113–121. Publisher : American Society for Microbiology.

Bi, D., Liu, L., Tai, C., Deng, Z., Rajakumar, K. et Ou, H.-Y. (2013). SecReT4 : a web-based bacterial type IV secretion system resource. *Nucleic Acids Research*, 41(D1):D660–D665.

Bi, D., Xu, Z., Harrison, E. M., Tai, C., Wei, Y., He, X., Jia, S., Deng, Z., Rajakumar, K. et Ou, H.-Y. (2012). ICEberg : a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Research*, 40(D1):D621–D626.

Bickhart, D. M., Kolmogorov, M., Tseng, E., Portik, D. M., Korobeynikov, A., Tolstoganov, I., Uritskiy, G., Liachko, I., Sullivan, S. T., Shin, S. B., Zorea, A., Andreu, V. P., Panke-Buisse, K., Medema, M. H., Mizrahi, I., Pevzner, P. A. et Smith, T. P. L. (2022). Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nature Biotechnology*, 40(5):711–719. Publisher : Nature Publishing Group.

Biswas, A., Fineran, P. C. et Brown, C. M. (2014). Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs. *Bioinformatics*, 30(13):1805–1813.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C. et Hugenholtz, P. (2007). CRISPR Recognition Tool (CRT) : a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, 8(1):209.

Blom, J., Albaum, S. P., Doppmeier, D., Pühler, A., Vorhölter, F.-J., Zakrzewski, M. et Goesmann, A. (2009). EDGAR : A software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics*, 10(1):154.

Bobadilla Ugarte, P., Barendse, P. et Swarts, D. C. (2023). Argonaute proteins confer immunity in all domains of life. *Current Opinion in Microbiology*, 74:102313.

Bolotin, A., Quinquis, B., Sorokin, A. et Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, 151(8):2551–2561. Publisher : Microbiology Society,.

Boniver, M., Wotquenne, P., Moutschen, M. et Rousseau, A. F. (2022). [Phage therapy, an additional strategy against multidrug-resistant bacteria]. *Revue Medicale De Liege*, 77(9):510–515.

Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., Philippon, A., Allesoe, R. L., Rebelo, A. R., Florensa, A. F., Fagelhauer, L., Chakraborty, T., Neumann, B., Werner, G., Bender, J. K., Stingl, K., Nguyen, M., Coppens, J., Xavier, B. B., Malhotra-Kumar, S., Westh, H., Pinholt, M., Anjum, M. F., Duggett, N. A., Kempf, I., Nykäsenoja, S., Olkkola, S., Wieczorek, K., Amaro, A., Clemente, L., Mossong, J., Losch, S., Ragimbeau, C., Lund, O. et Aarestrup, F. M. (2020). ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy*, 75(12):3491–3500.

Botelho, J. et Schulenburg, H. (2021). The Role of Integrative and Conjugative Elements in Antibiotic Resistance Evolution. *Trends in Microbiology*, 29(1):8–18.

Brady, C. L., Venter, S. N., Cleenwerck, I., Engelbeen, K., Vancanneyt, M., Swings, J. et Coutinho, T. A. (2009). Pantoea vagans sp. nov., Pantoea eucalypti sp. nov., Pantoea deleyi sp. nov. and Pantoea anthophila sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 59(9):2339–2345. Publisher : Microbiology Society,.

Buchfink, B., Xie, C. et Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60. Publisher : Nature Publishing Group.

Buchrieser, C., Brosch, R., Bach, S., Guiyoule, A. et Carniel, E. (1998). The high-pathogenicity island of Yersinia pseudotuberculosis can be inserted into any of the three chromosomal asn tRNA genes. *Molecular Microbiology*, 30(5):965–978. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2958.1998.01124.x.

Buck, M., Mehrshad, M. et Bertilsson, S. (2022). mOTUpan : a robust Bayesian approach to leverage metagenome-assembled genomes for core-genome estimation. *NAR Genomics and Bioinformatics*, 4(3):lqac060.

Burks, C., Fickett, J. W., Goad, W. B., Kanehisa, M., Lewitter, F. I., Rindone, W. P., Swindell, C. D., Tung, C. S. et Bilofsky, H. S. (1985). The GenBank nucleic acid sequence database. *Computer applications in the biosciences : CABIOS*, 1(4):225–233.

Campbell, A. (2003). The future of bacteriophage biology. *Nature Reviews Genetics*, 4(6): 471–477. Publisher : Nature Publishing Group.

Carroll, L. M., Larralde, M., Fleck, J. S., Ponnudurai, R., Milanese, A., Cappio, E. et Zeller, G. (2021). Accurate de novo identification of biosynthetic gene clusters with GECCO. Pages : 2021.05.03.442509 Section : New Results.

Carter, M. S., Zhang, X., Huang, H., Bouvier, J. T., Francisco, B. S., Vetting, M. W., Al-Obaidi, N., Bonanno, J. B., Ghosh, A., Zallot, R. G., Andersen, H. M., Almo, S. C. et Gerlt, J. A. (2018). Functional assignment of multiple catabolic pathways for d-apiose. *Nature Chemical Biology*, 14(7):696–705. Publisher : Nature Publishing Group.

Centers for Disease Control and Prevention (CDC) (2007). Update to CDC's sexually transmitted diseases treatment guidelines, 2006 : fluoroquinolones no longer re-

commended for treatment of gonococcal infections. *MMWR. Morbidity and mortality weekly report*, 56(14):332–336.

Chaguza, C., Yang, M., Cornick, J. E., du Plessis, M., Gladstone, R. A., Kwambana-Adams, B. A., Lo, S. W., Ebruke, C., Tonkin-Hill, G., Peno, C., Senghore, M., Obaro, S. K., Ousmane, S., Pluschke, G., Collard, J.-M., Sigaùque, B., French, N., Klugman, K. P., Heyderman, R. S., McGee, L., Antonio, M., Breiman, R. F., von Gottberg, A., Everett, D. B., Kadioglu, A. et Bentley, S. D. (2020). Bacterial genome-wide association study of hyper-virulent pneumococcal serotype 1 identifies genetic variation associated with neurotropism. *Communications Biology*, 3(1):1–12. Publisher : Nature Publishing Group.

Chang, T., Gavelis, G. S., Brown, J. M. et Stepanauskas, R. (2024). Genomic representativeness and chimerism in large collections of SAGs and MAGs of marine prokaryoplankton. *Microbiome*, 12(1):126.

Chen, J., Quiles-Puchalt, N., Chiang, Y. N., Bacigalupe, R., Fillol-Salom, A., Chee, M. S. J., Fitzgerald, J. R. et Penadés, J. R. (2018). Genome hypermobility by lateral transduction. *Science*. Publisher : American Association for the Advancement of Science.

Cheung, K.-H., Frost, H. R., Marshall, M. S., Prud'hommeaux, E., Samwald, M., Zhao, J. et Paschke, A. (2009). A journey to Semantic Web query federation in the life sciences. *BMC Bioinformatics*, 10(10):S10.

Chiang, Y. N., Penadés, J. R. et Chen, J. (2019). Genetic transduction by phages and chromosomal islands : The new and noncanonical. *PLOS Pathogens*, 15(8):e1007878. Publisher : Public Library of Science.

Chiapello, H., Bourgait, I., Sourivong, F., Heuclin, G., Gendrault-Jacquemard, A., Petit, M.-A. et El Karoui, M. (2005). Systematic determination of the mosaic structure of bacterial genomes : species backbone versus strain-specific loops. *BMC Bioinformatics*, 6(1):1–10. Number : 1 Publisher : BioMed Central.

Chun, J. et Goodfellow, M. (1995). A phylogenetic analysis of the genus Nocardia with 16S rRNA gene sequences. *International Journal of Systematic Bacteriology*, 45(2):240–245.

Chun, J. et Rainey, F. A. (2014). Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *International Journal of Systematic and Evolutionary Microbiology*, 64(Pt_2):316–324. Publisher : Microbiology Society,.

Clarke, T. H., Brinkac, L. M., Inman, J. M., Sutton, G. et Fouts, D. E. (2018). PanACEA : a bioinformatics tool for the exploration and visualization of bacterial panchromosomes. *BMC Bioinformatics*, 19(1):246.

Claverys, J.-P. et Håvarstein, L. S. (2007). Cannibalism and fratricide : mechanisms and raisons d'être. *Nature Reviews Microbiology*, 5(3):219–229. Publisher : Nature Publishing Group.

Contreras-Moreira, B. et Vinuesa, P. (2013). GET_homologues, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Applied and Environmental Microbiology*, 79(24):7696–7701. Publisher : American Society for Microbiology.

Cosentino, S. et Iwasaki, W. (2019). SonicParanoid : fast, accurate and easy orthology inference. *Bioinformatics*, 35(1):149–151.

Coulthurst, S. J. (2013). The Type VI secretion system – a widespread and versatile cell targeting system. *Research in Microbiology*, 164(6):640–654.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D. et Pourcel, C. (2018). CRISPRCasFinder, an

update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research*, 46(W1):W246–W251.

Cury, J., Touchon, M. et Rocha, E. (2017). Integrative and conjugative elements and their hosts : composition, distribution and organization. *Nucleic Acids Research*, 45(15): 8943–8956.

da Silva Filho, A. C., Raittz, R. T., Guizelini, D., De Pierri, C. R., Augusto, D. W., dos Santos-Weiss, I. C. R. et Marchaukoski, J. N. (2018). Comparative Analysis of Genomic Island Prediction Tools. *Frontiers in Genetics*, 9. Publisher : Frontiers.

Darling, A. E., Mau, B. et Perna, N. T. (2010). progressiveMauve : Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLOS ONE*, 5(6):e11147. Publisher : Public Library of Science.

Darling, A. E., Miklós, I. et Ragan, M. A. (2008). Dynamics of Genome Rearrangement in Bacterial Populations. *PLOS Genetics*, 4(7):e1000128. Publisher : Public Library of Science.

de Korne-Elenbaas, J., van der Putten, B. C. L., Boek, N. D. M., Matser, A., Schultsz, C., Bruisten, S. M. et van Dam, A. P. (2023). Putative transmission of extended-spectrum ß-lactamase-producing *Escherichia coli* among men who have sex with men in Amsterdam, the Netherlands. *International Journal of Antimicrobial Agents*, 62(1):106810.

Delmont, T. O. et Eren, A. M. (2018). Linking pangenomes and metagenomes : the Prochlorococcus metapangenome. *PeerJ*, 6:e4320. Publisher : PeerJ Inc.

Denise, R., Abby, S. S. et Rocha, E. P. C. (2019). Diversification of the type IV filament superfamily into machines for adhesion, protein secretion, DNA uptake, and motility. *PLOS Biology*, 17(7):e3000390. Publisher : Public Library of Science.

Dillingham, M. S. et Kowalczykowski, S. C. (2008). RecBCD Enzyme and the Repair of Double-Stranded DNA Breaks. *Microbiology and Molecular Biology Reviews*, 72(4): 642–671. Publisher : American Society for Microbiology.

Dimri, S. C., Indu, R., Negi, H. S., Panwar, N. et Sarda, M. (2024). Hidden Markov Model - Applications, Strengths, and Weaknesses. *In 2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*, pages 300–305.

Dobrindt, U., Hochhut, B., Hentschel, U. et Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology*, 2(5):414–424. Publisher : Nature Publishing Group.

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G. et Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, 359(6379):eaar4120. Publisher : American Association for the Advancement of Science.

Dubnau, D. et Blokesch, M. (2019). Mechanisms of DNA Uptake by Naturally Competent Bacteria. *Annual Review of Genetics*, 53(Volume 53, 2019):217–237. Publisher : Annual Reviews.

Duffin, P. M. et Seifert, H. S. (2010). DNA Uptake Sequence-Mediated Enhancement of Transformation in Neisseria gonorrhoeae Is Strain Dependent. *Journal of Bacteriology*, 192(17):4436–4444. Publisher : American Society for Microbiology.

Durant, E., Sabot, F., Conte, M. et Rouard, M. (2021). Panache : a web browser-based viewer for linearized pangenomes. *Bioinformatics*, 37(23):4556–4558.

Edgar, R. (2004). MUSCLE : multiple sequence alignment with improved accuracy and speed. *In Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.*, pages 689–690, Stanford, CA, USA. IEEE.

Edgar, R. C. (2007). PILER-CR : Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, 8(1):18.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.

Eisenstark, A. (1977). GENETIC RECOMBINATION IN BACTERIA. *Annual Review of Genetics*, 11(Volume 11, 1977):369–396. Publisher : Annual Reviews.

Emms, D. M. et Kelly, S. (2019). OrthoFinder : phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):238.

Evans, B. A. et Rozen, D. E. (2013). Significant variation in transformation frequency in Streptococcus pneumoniae. *The ISME Journal*, 7(4):791–799.

Fedrizzi, T., Meehan, C. J., Grottola, A., Giacobazzi, E., Fregni Serpini, G., Tagliazucchi, S., Fabio, A., Bettua, C., Bertorelli, R., De Sanctis, V., Rumpianesi, F., Pecorari, M., Jousson, O., Tortoli, E. et Segata, N. (2017). Genomic characterization of Nontuberculous Mycobacteria. *Scientific Reports*, 7(1):45258. Publisher : Nature Publishing Group.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G. et Ysebaert, M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA : primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507. Number : 5551 Publisher : Nature Publishing Group.

Firtina, C., Pillai, K., Kalsi, G. S., Suresh, B., Cali, D. S., Kim, J. S., Shahroodi, T., Cavlak, M. B., Lindegger, J., Alser, M., Luna, J. G., Subramoney, S. et Mutlu, O. (2024). ApHMM : Accelerating Profile Hidden Markov Models for Fast and Energy-efficient Genome Analysis. *ACM Trans. Archit. Code Optim.*, 21(1):19 :1–19 :29.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-l., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. et Venter, J. C. (1995). Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512. Publisher : American Association for the Advancement of Science.

Florensa, A. F., Kaas, R. S., Clausen, P. T. L. C., Aytan-Aktug, D. et Aarestrup, F. M. (2022). ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microbial Genomics*, 8(1):000748. Publisher : Microbiology Society,.

Flores Ramos, S., Brugger, S. D., Escapa, I. F., Skeete, C. A., Cotton, S. L., Eslami, S. M., Gao, W., Bomar, L., Tran, T. H., Jones, D. S., Minot, S., Roberts, R. J., Johnston, C. D. et Lemon, K. P. (2021). Genomic Stability and Genetic Defense Systems in Dolosigranulum pigrum, a Candidate Beneficial Bacterium from the Human Microbiome. *mSystems*, 6(5):10.1128/msystems.00425–21. Publisher : American Society for Microbiology.

Fouts, D. E., Brinkac, L., Beck, E., Inman, J. et 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 Research*, 40(22):e172.

Gaba, S., Kumari, A., Medema, M. et Kaushik, R. (2020). Pan-genome analysis and ancestral state reconstruction of class halobacteria : probability of a new super-order. *Scientific Reports*, 10(1):21205. Publisher : Nature Publishing Group.

Gao, L., Altae-Tran, H., Böhning, F., Makarova, K. S., Segel, M., Schmid-Burgk, J. L., Koob, J., Wolf, Y. I., Koonin, E. V. et Zhang, F. (2020). Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science*, 369(6507):1077–1084. Publisher : American Association for the Advancement of Science.

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., Ashbrook, D. G., Thorell, K., Rusholme-Pilcher, R. L., Liti, G., Rudbeck, E., Golicz, A. A., Nahnsen, S., Yang, Z., Mwaniki, M. N., Nobrega, F. L., Wu, Y., Chen, H., de Ligt, J., Sudmant, P. H., Huang, S., Weigel, D., Soranzo, N., Colonna, V., Williams, R. W. et Prins, P. (2024). Building pangenome graphs. *Nature Methods*, 21(11):2008–2012. Publisher : Nature Publishing Group.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B. et Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879. Publisher : Nature Publishing Group.

Gasc, A. M., Sicard, N., Claverys, J. P. et Sicard, A. M. (1980). Lack of SOS repair in *Streptococcus pneumoniae*. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 70(2):157–165.

Gautreau, G. (2020). *Conceptualisation et exploitation d'un graphe de pangénome partitionné comme représentation compacte de la diversité du répertoire génique des espèces procaryotes*. PhD Thesis.

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., Perrin, A., Médigue, C., Calteau, A., Cruveiller, S., Matias, C., Ambroise, C., Rocha, E. P. C. et Vallenet, D. (2020). PPanGGOLiN : Depicting microbial diversity via a partitioned pangenome graph. *PLOS Computational Biology*, 16(3):e1007732. Publisher : Public Library of Science.

Georjon, H. et Bernheim, A. (2023). The highly diverse antiphage defence systems of bacteria. *Nature Reviews Microbiology*, 21(10):686–700. Publisher : Nature Publishing Group.

Gibbons, H. S., Broomall, S. M., McNew, L. A., Daligault, H., Chapman, C., Bruce, D., Karavis, M., Krepps, M., McGregor, P. A., Hong, C., Park, K. H., Akmal, A., Feldman, A., Lin, J. S., Chang, W. E., Higgs, B. W., Demirev, P., Lindquist, J., Liem, A., Fochler, E., Read, T. D., Tapia, R., Johnson, S., Bishop-Lilly, K. A., Detter, C., Han, C., Sozhamannan, S., Rosenzweig, C. N. et Skowronski, E. W. (2011). Genomic Signatures of Strain Selection and Enhancement in Bacillus atrophaeus var. globigii, a Historical Biowarfare Simulant. *PLOS ONE*, 6(3):e17836. Publisher : Public Library of Science.

Glaeser, S. P. et Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and Applied Microbiology*, 38(4):237–245.

Goodman, S. D. et Scocca, J. J. (1988). Identification and arrangement of the DNA sequence recognized in specific transformation of Neisseria gonorrhoeae. *Proceedings of the National Academy of Sciences*, 85(18):6982–6986. Publisher : Proceedings of the National Academy of Sciences.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P. et Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57(1):81–91. Publisher : Microbiology Society,.

Grant, J. R., Enns, E., Marinier, E., Mandal, A., Herman, E. K., Chen, C.-Y., Graham, M., Van Domselaar, G. et Stothard, P. (2023). Proksee : in-depth characterization and visualization of bacterial genomes. *Nucleic Acids Research*, 51(W1):W484–W492.

Greener, J. G., Kandathil, S. M., Moffat, L. et Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23(1):40–55. Publisher : Nature Publishing Group.

Griffith, F. (1928). The Significance of Pneumococcal Types. *Epidemiology & Infection*, 27(2):113–159.

Grissa, I., Vergnaud, G. et Pourcel, C. (2007). CRISPRFinder : a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, 35(suppl_2):W52–W57.

Guegler, C. K. et Laub, M. T. (2021). Shutoff of host transcription triggers a toxin-antitoxin system to cleave phage RNA and abort infection. *Molecular Cell*, 81(11):2361–2373.e9. Publisher : Elsevier.

Guo, R. (2017). MongoDB's JavaScript fuzzer. *Commun. ACM*, 60(5):43–47.

Gütebier, L., Bleimehl, T., Henkel, R., Munro, J., Müller, S., Morgner, A., Laenge, J., Pachauer, A., Erdl, A., Weimar, J., Walther Langendorf, K., Vialard, V., Liebig, T., Preusse, M., Waltemath, D. et Jarasch, A. (2022). CovidGraph : a graph to fight COVID-19. *Bioinformatics*, 38(20):4843–4845.

Hackenberger, D., Imtiaz, H., Raphenya, A. R., Alcock, B. P., Poinar, H. N., Wright, G. D. et McArthur, A. G. (2024). CARPDM : cost-effective antibiotic resistome profiling of metagenomic samples using targeted enrichment. Pages : 2024.03.27.587061 Section : New Results.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R. et Goebel, W. (1990). Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extra intestinal *Escherichia coli* isolates. *Microbial Pathogenesis*, 8(3):213–225.

Haft, D. H., Selengut, J., Mongodin, E. F. et Nelson, K. E. (2005). A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. *PLOS Computational Biology*, 1(6):e60. Publisher : Public Library of Science.

Hall, J. P. J., Botelho, J., Cazares, A. et Baltrus, D. A. (2021). What makes a megaplasmid ? *Philosophical Transactions of the Royal Society B : Biological Sciences*, 377(1842): 20200472. Publisher : Royal Society.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. et Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes : a new frontier for natural products. *Chemistry & Biology*, 5(10):R245–249.

Hannigan, G. D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., Wang, R., Piizzi, G., Temesi, G., Hazuda, D. J., Woelk, C. H. et Bitton, D. A. (2019). A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*, 47(18):e110.

Harrison, P. W., Lower, R. P. J., Kim, N. K. D. et Young, J. P. W. (2010). Introducing the bacterial 'chromid' : not a chromosome, not a plasmid. *Trends in Microbiology*, 18(4): 141–148. Publisher : Elsevier.

Hauns, S., Alkhnbashi, O. S. et Backofen, R. (2024). Deepdefense : annotation of immune systems in prokaryotes using deep learning. *GigaScience*, 13:giae062.

He, S., Gao, B., Sabnis, R. et Sun, Q. (2023). Nucleic Transformer : Classifying DNA Sequences with Self-Attention and Convolutions. *ACS Synthetic Biology*, 12(11):3205–3214. Publisher : American Chemical Society.

Hennig, A., Bernhardt, J. et Nieselt, K. (2015). Pan-Tetris : an interactive visualisation for Pan-genomes. *BMC Bioinformatics*, 16(11):S3.

Her, H.-L. et Wu, Y.-W. (2018). A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains. *Bioinformatics*, 34(13):i89–i95.

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., Marschall, T., Li, H. et Paten, B. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, 42(4):663–673. Publisher : Nature Publishing Group.

Higgins, D. G. et Sharp, P. M. (1988). CLUSTAL : a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237–244.

Hinton, G. E., Osindero, S. et Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554.

Hoarfrost, A., Aptekmann, A., Farfañuk, G. et Bromberg, Y. (2022). Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nature Communications*, 13(1):2606. Publisher : Nature Publishing Group.

Hogg, J. S., Hu, F. Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J. C. et Ehrlich, G. D. (2007). Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biology*, 8(6):R103.

Holley, G. et Melsted, P. (2020). Bifrost : highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biology*, 21(1):249.

Hsu, J.-C., Hsu, C.-H., Chen, S. C. et Chung, Y. C. (2014). Correlation Aware Technique for SQL to NoSQL Transformation. *In 2014 7th International Conference on Ubi-Media Computing and Workshops*, pages 43–46, Ulaanbaatar. IEEE.

Hu, T., Chitnis, N., Monos, D. et Dinh, A. (2021). Next-generation sequencing technologies : An overview. *Human Immunology*, 82(11):801–811.

Hu, X. et Friedberg, I. (2019). SwiftOrtho : A fast, memory-efficient, multiple genome orthology classifier. *GigaScience*, 8(10):giz118.

Hu, Y., Wang, Y., Hu, X., Chao, H., Li, S., Ni, Q., Zhu, Y., Hu, Y., Zhao, Z. et Chen, M. (2024). T4SEpp : A pipeline integrating protein language models to predict bacterial type IV secreted effectors. *Computational and Structural Biotechnology Journal*, 23:801–812.

Huang, M., Liu, M., Huang, L., Wang, M., Jia, R., Zhu, D., Chen, S., Zhao, X., Zhang, S., Gao, Q., Zhang, L. et Cheng, A. (2021). The activation and limitation of the bacterial natural transformation system : The function in genome evolution and stability. *Microbiological Research*, 252:126856.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C. et Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, 1(5):1–6. Publisher : Nature Publishing Group.

Hugenholtz, P., Chuvochina, M., Oren, A., Parks, D. H. et Soo, R. M. (2021). Prokaryotic taxonomy and nomenclature in the age of big sequence data. *The ISME Journal*, 15(7):1879–1892.

Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., Jones, P., Leinonen, R., McAnulla, C., Maguire, E., Maslen, J., Mitchell, A., Nuka, G., Oisel, A., Pesseat, S., Radhakrishnan, R., Rocca-Serra, P., Scheremetjew, M., Sterk, P., Vaughan, D., Cochrane, G., Field, D. et Sansone, S.-A. (2014). EBI metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research*, 42(D1):D600–D606.

Huynen, M. A. et Bork, P. (1998). Measuring genome evolution. *Proceedings of the National Academy of Sciences*, 95(11):5849–5856. Publisher : Proceedings of the National Academy of Sciences.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W. et Hauser, L. J. (2010). Prodigal : prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119.

Hyun, J. C., Monk, J. M. et Palsson, B. O. (2022). Comparative pangenomics : analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics*, 23(1):7.

Imhoff, J. F. (2003). Phylogenetic taxonomy of the family Chlorobiaceae on the basis of 16S rRNA and fmo (Fenna-Matthews-Olson protein) gene sequences. *International Journal of Systematic and Evolutionary Microbiology*, 53(Pt 4):941–951.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. et McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226–232. Publisher : Nature Publishing Group.

Iraola, G., Forster, S. C., Kumar, N., Lehours, P., Bekal, S., García-Peña, F. J., Paolicchi, F., Morsella, C., Hotzel, H., Hsueh, P.-R., Vidal, A., Lévesque, S., Yamazaki, W., Balzan, C., Vargas, A., Piccirillo, A., Chaban, B., Hill, J. E., Betancor, L., Collado, L., Truyers, I., Midwinter, A. C., Dagi, H. T., Mégraud, F., Calleros, L., Pérez, R., Naya, H. et Lawley, T. D. (2017). Distinct Campylobacter fetus lineages adapted as livestock pathogens and human pathobionts in the intestinal microbiota. *Nature Communications*, 8(1): 1367. Publisher : Nature Publishing Group.

Jacob, F. et Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356.

Jaillard, M., Lima, L., Tournoud, M., Mahé, P., Belkum, A. v., Lacroix, V. et Jacob, L. (2018). A fast and agnostic method for bacterial genome-wide association studies : Bridging the gap between k-mers and genetic events. *PLOS Genetics*, 14(11):e1007758. Publisher : Public Library of Science.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. et Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1):5114. Publisher : Nature Publishing Group.

Jansen, R., Embden, J. D. A. v., Gaastra, W. et Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, 43(6):1565–1575. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2958.2002.02839.x.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. et Bork, P. (2008). eggNOG : automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, 36(suppl_1):D250–D254.

Johnson, C. M. et Grossman, A. D. (2015). Integrative and Conjugative Elements (ICEs) : What They Do and How They Work. *Annual Review of Genetics*, 49(Volume 49, 2015): 577–601. Publisher : Annual Reviews.

Johnston, C., Martin, B., Fichant, G., Polard, P. et Claverys, J.-P. (2014). Bacterial transformation : distribution, shared mechanisms and divergent control. *Nature Reviews Microbiology*, 12(3):181–196. Publisher : Nature Publishing Group.

Jonkheer, E. M., Brankovics, B., Houwers, I. M., van der Wolf, J. M., Bonants, P. J. M., Vreeburg, R. A. M., Bollema, R., de Haan, J. R., Berke, L., Smit, S., de Ridder, D. et van der Lee, T. A. J. (2021). The Pectobacterium pangenome, with a focus on Pectobacterium brasiliense, shows a robust core and extensive exchange of genes from a shared gene pool. *BMC Genomics*, 22(1):1–18. Number : 1 Publisher : BioMed Central.

Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. et Ishiguro-Watanabe, M. (2025). KEGG : biological systems database as a model of the real world. *Nucleic Acids Research*, 53(D1):D672–D677.

Karaolis, D. K. R., Somara, S., Maneval, D. R., Johnson, J. A. et Kaper, J. B. (1999). A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. *Nature*, 399(6734):375–379. Publisher : Nature Publishing Group.

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., Vargas, C. D., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., Pesant, S., Reynaud, E. G., Sardet, C., Sieracki, M. E., Speich, S., Velayoudon, D., Weissenbach, J., Wincker, P. et Consortium, t. T. O. (2011). A Holistic Approach to Marine Eco-Systems Biology. *PLOS Biology*, 9(10):e1001177. Publisher : Public Library of Science.

Katoh, K. et Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7 : Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., Heckmann, D., Anand, A., Yang, L., Nizet, V., Monk, J. M. et Palsson, B. O. (2018). Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. *Nature Communications*, 9(1):4306. Publisher : Nature Publishing Group.

Kel, A., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. et Wingender, E. (2003). MATCHTM : a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31(13):3576–3579.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. et Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab.

Koch, A. L. (2004). Catastrophe and What To Do About It If You Are a Bacterium : The Importance of Frameshift Mutants. *Critical Reviews in Microbiology*, 30(1):1–6. Publisher : Taylor & Francis.

Konstantinidis, K. T. (2023). Sequence-discrete species for prokaryotes and other microbes : A historical perspective and pending issues. *mLife*, 2(4):341–349. Publisher : John Wiley & Sons, Ltd.

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics*, 39(1):309–338.

Koonin, E. V. et Wolf, Y. I. (2008). Genomics of bacteria and archaea : the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, 36(21):6688–6719.

Korotkov, K. V., Sandkvist, M. et Hol, W. G. J. (2012). The type II secretion system : biogenesis, molecular architecture and mechanism. *Nature Reviews Microbiology*, 10(5):336–351. Publisher : Nature Publishing Group.

Kuhnert, P. et Korczak, B. M. (2006). Prediction of whole-genome DNA-DNA similarity, determination of G+C content and phylogenetic analysis within the family Pasteurellaceae by multilocus sequence analysis (MLSA). *Microbiology (Reading, England)*, 152(Pt 9):2537–2548.

Kulsum, U., Kapil, A., Singh, H. et Kaur, P. (2018). NGSPanPipe : A Pipeline for Pan-genome Identification in Microbial Strains from Experimental Reads. *In* Adhikari, R. et Thapa, S., éditeurs : *Infectious Diseases and Nanomedicine III*, pages 39–49, Singapore. Springer Singapore.

Lamkiewicz, K., Barf, L.-M., Sachse, K. et Hölzer, M. (2024). RIBAP : a comprehensive bacterial core genome annotation pipeline for pangenome calculation beyond the species level. *Genome Biology*, 25(1):1–21. Number : 1 Publisher : BioMed Central.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., Morgan, M. J., International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, C. f. G. R., The Sanger Centre :, Washington University Genome Sequencing Center, US DOE Joint Genome Institute :, Baylor College of Medicine Human Genome Sequencing Center :, RIKEN Genomic Sciences Center :, Genoscope and CNRS UMR-8030 :, Department of Genome Analysis, I. o. M. B., GTC Sequencing Center :, Beijing Genomics Institute/Human Genome Center :, Multimegabase Sequencing Center, T. I. f. S. B., Stanford Genome Technology Center :, University of Oklahoma's Advanced Center for Genome Technology :, Max Planck Institute for Molecular Genetics :, Cold Spring Harbor Laboratory, L. A. H. G. C., GBF—German

Research Centre for Biotechnology :, *Genome Analysis Group (listed in alphabetical order, a. i. i. l. u. o. h., Scientific management : National Human Genome Research Institute, U. N. I. o. H., Stanford Human Genome Center :, University of Washington Genome Center :, Department of Molecular Biology, K. U. S. o. M., University of Texas Southwestern Medical Center at Dallas :, Office of Science, U. D. o. E. et The Wellcome Trust : (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. Publisher : Nature Publishing Group.

Lao, J., Lacroix, T., Guédon, G., Coluzzi, C., Payot, S., Leblond-Bourget, N. et Chiapello, H. (2022). ICEscreen : a tool to detect Firmicute ICEs and IMEs, isolated or enclosed in composite structures. *NAR Genomics and Bioinformatics*, 4(4):lqac079.

Lapierre, P. et Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends in genetics : TIG*, 25(3):107–110.

Larralde, M. (2022). Pyrodigal : Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software*, 7(72): 4296.

Larralde, M. et Zeller, G. (2023). PyHMMER : a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics*, 39(5):btad214.

Laslett, D. et Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, 32(1):11–16.

Lathe, W. C., Snel, B. et Bork, P. (2000). Gene context conservation of a higher order than operons. *Trends in Biochemical Sciences*, 25(10):474–479.

Le, K. K., Whiteside, M. D., Hopkins, J. E., Gannon, V. P. J. et Laing, C. R. (2018). Spfy : an integrated graph database for real-time prediction of bacterial phenotypes and downstream comparative analyses. *Database*, 2018:bay086.

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F. et Prohaska, S. J. (2011). Proteinortho : Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics*, 12(1):124.

Lederberg, J. et Tatum, E. L. (1946). Gene Recombination in Escherichia Coli. *Nature*, 158(4016):558–558. Publisher : Nature Publishing Group.

Lederberg, J. et Tatum, E. L. (1953). Sex in Bacteria : Genetic Studies, 1945-1952. *Science*, 118(3059):169–175. Publisher : American Association for the Advancement of Science.

Lee, S. D., Kim, E. S. et Hah, Y. C. (2000). Phylogenetic analysis of the genera Pseudonocardia and Actinobispora based on 16S ribosomal DNA sequences. *FEMS microbiology letters*, 182(1):125–129.

Lescat, M., Calteau, A., Hoede, C., Barbe, V., Touchon, M., Rocha, E., Tenaillon, O., Médigue, C., Johnson, J. R. et Denamur, E. (2009). A Module Located at a Chromosomal Integration Hot Spot Is Responsible for the Multidrug Resistance of a Reference Strain from Escherichia coli Clonal Group A. *Antimicrobial Agents and Chemotherapy*, 53(6):2283–2288. Publisher : American Society for Microbiology.

Levin, B. R. (1993). The accessory genetic elements of bacteria : existence conditions and (co)evolution. *Current Opinion in Genetics & Development*, 3(6):849–854.

Li, H., Feng, X. et Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21(1):265.

Li, L., Stoeckert, C. J. et Roos, D. S. (2003). OrthoMCL : Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, 13(9):2178–2189. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab.

Li, W., Jaroszewski, L. et Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283.

Lili, L. N., Britton, N. F. et Feil, E. J. (2007). The Persistence of Parasitic Plasmids. *Genetics*, 177(1):399–405.

Lipman, D. J. et Pearson, W. R. (1985). Rapid and Sensitive Protein Similarity Searches. *Science*, 227(4693):1435–1441. Publisher : American Association for the Advancement of Science.

Ludwig, W., Weizenegger, M., Dorn, S., Andreesen, J. et Schleifer, K. H. (1990). The phylogenetic position of Peptococcus niger based on 16S rRNA sequence studies. *FEMS microbiology letters*, 59(1-2):139–143.

Luhmann, N., Holley, G. et Achtman, M. (2021). BlastFrost : fast querying of 100,000s of bacterial genomes in Bifrost graphs. *Genome Biology*, 22(1):30.

Lukjancenko, O., Thomsen, M. C., Larsen, M. V. et Ussery, D. W. (2013). PanFunPro : PAN-genome analysis based on FUNctional PROfiles.

Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M. et Konstantinidis, K. T. (2011). Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences*, 108(17):7200–7205. Publisher : Proceedings of the National Academy of Sciences.

Lyons, N. A., Kraigher, B., Stefanic, P., Mandic-Mulec, I. et Kolter, R. (2016). A Combinatorial Kin Discrimination System in Bacillus subtilis. *Current Biology*, 26(6):733–742. Publisher : Elsevier.

Maddamsetti, R., Yao, Y., Wang, T., Gao, J., Huang, V. T., Hamrick, G. S., Son, H.-I. et You, L. (2024). Duplicated antibiotic resistance genes reveal ongoing selection and horizontal gene transfer in bacteria. *Nature Communications*, 15(1):1449. Publisher : Nature Publishing Group.

Mainguy, J., Arnoux, J., Gautreau, G., Bazin, A., Vallenet, D. et Calteau, A. (2023). PPanGGOLiN V2 : technical enhancement and new features to analyze thousands of prokaryotic genomes. *In The local pangenome*, Alicante, Spain.

Makarova, K. S., Sorokin, A. V., Novichkov, P. S., Wolf, Y. I. et Koonin, E. V. (2007). Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biology Direct*, 2(1):33.

Makarova, K. S., Wolf, Y. I. et Koonin, E. V. (2013). Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research*, 41(8):4360–4377.

Makarova, K. S., Wolf, Y. I., Snir, S. et Koonin, E. V. (2011). Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *Journal of Bacteriology*, 193(21):6039–6056.

Marcus, S., Lee, H. et Schatz, M. C. (2014). SplitMEM : a graphical algorithm for pangenome analysis with suffix skips. *Bioinformatics*, 30(24):3476–3483.

Masseroli, M., Kaitoua, A., Pinoli, P. et Ceri, S. (2016). Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. *Methods*, 111:3–11.

Masseroli, M., Pinoli, P., Venco, F., Kaitoua, A., Jalili, V., Palluzzi, F., Muller, H. et Ceri, S. (2015). GenoMetric Query Language : a novel approach to large-scale genomic data management. *Bioinformatics*, 31(12):1881–1888.

Matthews, C. A., Watson-Haigh, N. S., Burton, R. A. et Sheppard, A. E. (2024). A gentle introduction to pangenomics. *Briefings in Bioinformatics*, 25(6):bbae588.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O'Brien, J. S., Pawlowski, A. C., Piddock, L. J. V., Spanogiannopoulos, P., Sutherland, A. D., Tang, I., Taylor, P. L., Thaker, M., Wang, W., Yan, M., Yu, T. et Wright, G. D. (2013). The Comprehensive Antibiotic Resistance Database. *Antimicrobial Agents and Chemotherapy*, 57(7):3348–3357. Publisher : American Society for Microbiology.

Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E. et Breitling, R. (2011). antiSMASH : rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*, 39(suppl_2):W339–W346.

Medini, D., Donati, C., Tettelin, H., Masignani, V. et Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6):589–594.

Menzel, P., Ng, K. L. et Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7(1):11257. Publisher : Nature Publishing Group.

Minkin, I., Patel, A., Kolmogorov, M., Vyahhi, N. et Pham, S. (2013). Sibelia : A Scalable and Comprehensive Synteny Block Generation Tool for Closely Related Microbial Genomes. *In Algorithms in Bioinformatics*, pages 215–229. Springer, Berlin, Heidelberg. ISSN : 1611-3349.

Minkin, I., Pham, S. et Medvedev, P. (2017). TwoPaCo : an efficient algorithm to build the compacted de Bruijn graph from many complete genomes. *Bioinformatics*, 33(24): 4024–4032.

Mirny, L. A. et Gelfand, M. S. (2002). Using orthologous and paralogous proteins to identify specificity determining residues. *Genome Biology*, 3(3):preprint0002.1.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D. et Bateman, A. (2021). Pfam : The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419.

Mitrofanov, A., Alkhnbashi, O. S., Shmakov, S. A., Makarova, K., Koonin, E. et Backofen, R. (2021). CRISPRidentify : identification of CRISPR arrays using machine learning approach. *Nucleic Acids Research*, 49(4):e20.

Moldovan, M. A. et Gelfand, M. S. (2018). Pangenomic Definition of Prokaryotic Species and the Phylogenetic Structure of Prochlorococcus spp. *Frontiers in Microbiology*, 9. Publisher : Frontiers.

Molineux, I. J. (1991). Host-parasite interactions : recent developments in the genetics of abortive phage infections. *The New Biologist*, 3(3):230–236.

Moore, W. E. C., Stackebrandt, E., Kandler, O., Colwell, R. R., Krichevsky, M. I., Truper, H. G., Murray, R. G. E., Wayne, L. G., Grimont, P. A. D., Brenner, D. J., Starr, M. P. et Moore, L. H. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic and Evolutionary Microbiology*, 37(4):463–464.

Morse, M. L., Lederberg, E. M. et Lederberg, J. (1956). TRANSDUCTION IN ESCHERICHIA COLI K-12. *Genetics*, 41(1):142–156.

Moulana, A., Anderson, R. E., Fortunato, C. S. et Huber, J. A. (2020). Selection Is a Significant Driver of Gene Gain and Loss in the Pangenome of the Bacterial Genus Sulfurovum in Geographically Distinct Deep-Sea Hydrothermal Vents. *mSystems*, 5(2):10.1128/msystems.00673–19. Publisher : American Society for Microbiology.

Nawrocki, E. P. et Eddy, S. R. (2013). Infernal 1.1 : 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935.

Needleman, S. B. et Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Noll, N., Molari, M., Shaw, L. P. et Neher, R. A. (2023). PanGraph : scalable bacterial pan-genome graph construction. *Microbial Genomics*, 9(6):001034. Publisher : Microbiology Society,.

Norri, T., Cazaux, B., Dönges, S., Valenzuela, D. et Mäkinen, V. (2021). Founder reconstruction enables scalable and seamless pangenomic analysis. *Bioinformatics*, 37(24):4611–4619.

Néron, B., Denise, R., Coluzzi, C., Touchon, M., Rocha, E. P. C. et Abby, S. S. (2023). MacSyFinder v2 : Improved modelling and search engine to identify molecular systems in genomes. *Peer Community Journal*, 3.

Ochman, H., Lawrence, J. G. et Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304. Publisher : Nature Publishing Group.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A. et Li, P. (2004). Taverna : a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054.

Oliveira, P. H. (2021). Bacterial Epigenomics : Coming of Age. *mSystems*, 6(4): 10.1128/msystems.00747–21. Publisher : American Society for Microbiology.

Oliveira, P. H., Touchon, M., Cury, J. et Rocha, E. P. C. (2017). The chromosomal organization of horizontal gene transfer in bacteria. *Nature Communications*, 8(1):841. Publisher : Nature Publishing Group.

Ozer, E. A., Allen, J. P. et Hauser, A. R. (2014). Characterization of the core and accessory genomes of Pseudomonas aeruginosa using bioinformatic tools Spine and AGEnt. *BMC Genomics*, 15(1):737.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A. et Parkhill, J. (2015). Roary : rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693.

Pallen, M. J. et Wren, B. W. (2007). Bacterial pathogenomics. *Nature*, 449(7164):835–842. Publisher : Nature Publishing Group.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A. et Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36(10):996–1004. Publisher : Nature Publishing Group.

Pascual, J., Macián, M. C., Arahal, D. R., Garay, E. et Pujalte, M. J. (2010). Multilocus sequence analysis of the central clade of the genus Vibrio by using the 16S rRNA, recA, pyrH, rpoD, gyrB, rctB and toxR genes. *International Journal of Systematic and Evolutionary Microbiology*, 60(Pt 1):154–165.

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R. et Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Mining*, 4(1):1–27. Number : 1 Publisher : BioMed Central.

Payne, L. J., Todeschini, T. C., Wu, Y., Perry, B. J., Ronson, C., Fineran, P., Nobrega, F. et Jackson, S. (2021). Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Research*, 49(19):10868–10878.

Pearson, W. R. et Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448. Publisher : Proceedings of the National Academy of Sciences.

Pedersen, T. L., Nookaew, I., Wayne Ussery, D. et Månsson, M. (2017). PanViz : interactive visualization of the structure of functionally annotated pangenomes. *Bioinformatics*, 33(7):1081–1082.

Periwal, V. et Scaria, V. (2015). Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics*, 31(1):1–9.

Perrin, A. et Rocha, E. P. C. (2021). PanACoTA : a modular tool for massive microbial comparative genomics. *NAR Genomics and Bioinformatics*, 3(1):lqaa106.

Pfeifer, E., Moura de Sousa, J. A., Touchon, M. et Rocha, E. P. C. (2021). Bacteria have numerous distinctive groups of phage–plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Research*, 49(5):2655–2673.

Pičmanová, M. et Møller, B. L. (2016). Apiose : one of nature's witty games. *Glycobiology*, 26(5):430–442.

Plaza Oñate, F., Le Chatelier, E., Almeida, M., Cervino, A. C. L., Gauthier, F., Magoulès, F., Ehrlich, S. D. et Pichaud, M. (2019). MSPminer : abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics*, 35(9): 1544–1552.

Popa, O., Hazkani-Covo, E., Landan, G., Martin, W. et Dagan, T. (2011). Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Research*, 21(4):599–609. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab.

Pruitt, K. D., Tatusova, T. et Maglott, D. R. (2007). NCBI reference sequences (RefSeq) : a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(suppl_1):D61–D65.

Qi, J., Chen, Y., Copenhaver, G. P. et Ma, H. (2014). Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. *Proceedings of the National Academy of Sciences*, 111(27):10007–10012. Publisher : Proceedings of the National Academy of Sciences.

Remm, M., Storm, C. E. et Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041–1052.

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L., Curtis, T., Escobar-Zepeda, A., Gurbich, T., Kale, V., Korobeynikov, A., Raj, S., Rogers, A., Sakharova, E., Sanchez, S., Wilkinson, D. et Finn, R. (2023). MGnify : the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 51(D1):D753–D759.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P. et Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437. Publisher : Nature Publishing Group.

Rocha, E. P. C. (2008). The organization of the bacterial genome. *Annual Review of Genetics*, 42:211–233.

Sanger, F., Nicklen, S. et Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467. Publisher : Proceedings of the National Academy of Sciences.

Schaeffer, L., Pimentel, H., Bray, N., Melsted, P. et Pachter, L. (2017). Pseudoalignment for metagenomic read assignment. *Bioinformatics*, 33(14):2082–2088.

Schembri, M. A., Zakour, N. L. B., Phan, M.-D., Forde, B. M., Stanton-Cook, M. et Beatson, S. A. (2015). Molecular Characterization of the Multidrug Resistant Escherichia coli ST131 Clone. *Pathogens*, 4(3):422–430. Number : 3 Publisher : Multidisciplinary Digital Publishing Institute.

Schleifer, K. H. (2009). Classification of *Bacteria* and *Archaea* : Past, present and future. *Systematic and Applied Microbiology*, 32(8):533–542.

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O. et Weigel, D. (2009). Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, 10(9):R98.

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D. T., Tett, A., Morrow, A. L. et Segata, N. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods*, 13(5):435–438. Publisher : Nature Publishing Group.

Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J. et Goesmann, A. (2021). Bakta : rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*, 7(11):000685. Publisher : Microbiology Society,.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. et Ideker, T. (2003). Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504.

Sharp, P. M., Shields, D. C., Wolfe, K. H. et Li, W.-H. (1989). Chromosomal Location and Evolutionary Rate Variation in Enterobacterial Genes. *Science*, 246(4931):808–810. Publisher : American Association for the Advancement of Science.

Sheikhizadeh, S., Schranz, M. E., Akdel, M., de Ridder, D. et Smit, S. (2016). PanTools : representation, storage and exploration of pan-genomic data. *Bioinformatics*, 32(17): i487–i493.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. et Zdobnov, E. M. (2015). BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Sjölander, K., Datta, R. S., Shen, Y. et Shoffner, G. M. (2011). Ortholog identification in the presence of domain architecture rearrangement. *Briefings in Bioinformatics*, 12(5):413–422.

Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. et de la Cruz, F. (2010). Mobility of Plasmids. *Microbiology and Molecular Biology Reviews*, 74(3):434–452. Publisher : American Society for Microbiology.

Smith, T. F. et Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Snipen, L., Almøy, T. et Ussery, D. W. (2009). Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics*, 10(1):385.

Snipen, L. et Liland, K. H. (2015). micropan : an R-package for microbial pan-genomics. *BMC Bioinformatics*, 16(1):79.

Srikant, S., Guegler, C. K. et Laub, M. T. (2022). The evolution of a counter-defense mechanism in a virus constrains its host range. *eLife*, 11:e79549. Publisher : eLife Sciences Publications, Ltd.

Stackebrandt, E., Kroppenstedt, R. M. et Fowler, V. J. (1983). A phylogenetic analysis of the family Dermatophilaceae. *Journal of General Microbiology*, 129(6):1831–1838.

Stamboulian, M., Guerrero, R. F., Hahn, M. W. et Radivojac, P. (2020). The ortholog conjecture revisited : the value of orthologs and paralogs in function prediction. *Bioinformatics*, 36(Supplement_1):i219–i226.

Starikova, E. V., Tikhonova, P. O., Prianichnikov, N. A., Rands, C. M., Zdobnov, E. M., Ilina, E. N. et Govorun, V. M. (2020). Phigaro : high-throughput prophage sequence annotation. *Bioinformatics*, 36(12):3882–3884.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J. et Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20(1):473.

Steinegger, M. et Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028. Publisher : Nature Publishing Group.

Steinegger, M. et Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1):2542. Publisher : Nature Publishing Group.

Stewart, G. J. et Carlson, C. A. (1986). The biology of natural transformation. *Annual review of microbiology*, 40:211–235.

Stinear, T. P., Jenkin, G. A., Johnson, P. D. et Davies, J. K. (2000). Comparative genetic analysis of Mycobacterium ulcerans and Mycobacterium marinum reveals evidence of recent divergence. *Journal of Bacteriology*, 182(22):6322–6330.

Stoye, J. (1998). Multiple sequence alignment with the divide-and-conquer method. *Gene*, 211(2):GC45–GC56.

Sun, B., Pashkova, L., Pieters, P., Harke, A., Mohite, O., Santos, A., Zielinski, D., Palsson, B. et Phaneuf, P. (2025). PanKB : An interactive microbial pangenome knowledgebase for research, biotechnological innovation, and knowledge mining. *Nucleic Acids Research*, 53(D1):D806–D818.

Sun, S., Ke, R., Hughes, D., Nilsson, M. et Andersson, D. I. (2012). Genome-Wide Detection of Spontaneous Chromosomal Rearrangements in Bacteria. *PLOS ONE*, 7(8):e42639. Publisher : Public Library of Science.

Suwanto, A. et Kaplan, S. (1989). Physical and genetic mapping of the Rhodobacter sphaeroides 2.4.1 genome : presence of two unique circular chromosomes. *Journal of Bacteriology*, 171(11):5850–5859. Publisher : American Society for Microbiology.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. et Wu, C. H. (2007). UniRef : comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. et Koonin, E. V. (2000). The COG database : a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36.

Tatusov, R. L., Koonin, E. V. et Lipman, D. J. (1997). A Genomic Perspective on Protein Families. *Science*, 278(5338):631–637. Publisher : American Association for the Advancement of Science.

Team, P. D. (2002). PyTables : Hierarchical Datasets in Python.

Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J. et Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nature Communications*, 13(1):2561. Publisher : Nature Publishing Group.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M.,

Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J. B., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R. et Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae : implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13950–13955.

Tettelin, H., Riley, D., Cattuto, C. et Medini, D. (2008). Comparative genomics : the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5):472–477.

The Computational Pan-Genomics Consortium (2018). Computational pan-genomics : status, promises and challenges. *Briefings in Bioinformatics*, 19(1):118–135.

The UniProt Consortium (2025). UniProt : the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617.

Thorpe, H. A., Bayliss, S. C., Sheppard, S. K. et Feil, E. J. (2018). Piggy : a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *GigaScience*, 7(4): giy015.

Timón-Reina, S., Rincón, M. et Martínez-Tomás, R. (2021). An overview of graph databases and their applications in the biomedical domain. *Database*, 2021:baab026.

Tonder, A. J. v., Mistry, S., Bray, J. E., Hill, D. M. C., Cody, A. J., Farmer, C. L., Klugman, K. P., Gottberg, A. v., Bentley, S. D., Parkhill, J., Jolley, K. A., Maiden, M. C. J. et Brueggemann, A. B. (2014). Defining the Estimated Core Genome of Bacterial Populations Using a Bayesian Decision Model. *PLOS Computational Biology*, 10(8):e1003788. Publisher : Public Library of Science.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D., Corander, J., Bentley, S. D. et Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, 21(1):180.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M. E., Frapy, E., Garry, L., Ghigo, J. M., Gilles, A. M., Johnson, J., Bouguénec, C. L., Lescat, M., Mangenot, S., Martinez-Jéhanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M. A., Pichon, C., Rouy, Z., Ruf, C. S., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Médigue, C., Rocha, E. P. C. et Denamur, E. (2009). Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *PLOS Genetics*, 5(1):e1000344. Publisher : Public Library of Science.

Treangen, T. J., Ondov, B. D., Koren, S. et Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15(11):524.

Tria, F. D. K. et Martin, W. F. (2021). Gene Duplications Are At Least 50 Times Less Frequent than Gene Transfers in Prokaryotic Genomes. *Genome Biology and Evolution*, 13(10):evab224.

Trucksis, M., Michalski, J., Deng, Y. K. et Kaper, J. B. (1998). The Vibrio cholerae genome contains two unique circular chromosomes. *Proceedings of the National Academy of Sciences*, 95(24):14464–14469. Publisher : Proceedings of the National Academy of Sciences.

Tse, H., Cai, J. J., Tsoi, H.-W., Lam, E. P. et Yuen, K.-Y. (2010). Natural selection retains over-represented out-of-frame stop codons against frameshift peptides in prokaryotes. *BMC Genomics*, 11(1):1–13. Number : 1 Publisher : BioMed Central.

Turner, I., Garimella, K. V., Iqbal, Z. et McVean, G. (2018). Integrating long-range connectivity information into de Bruijn graphs. *Bioinformatics*, 34(15):2556–2565.

Valentine, P. J., Shoemaker, N. B. et Salyers, A. A. (1988). Mobilization of Bacteroides plasmids by Bacteroides conjugal elements. *Journal of Bacteriology*, 170(3):1319–1324.

Vallenet, D., Calteau, A., Dubois, M., Amours, P., Bazin, A., Beuvin, M., Burlot, L., Bussell, X., Fouteau, S., Gautreau, G., Lajus, A., Langlois, J., Planel, R., Roche, D., Rollin, J., Rouy, Z., Sabatet, V. et Médigue, C. (2020). MicroScope : an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Research*, 48(D1): D579–D589.

van den Brandt, A., Jonkheer, E. M., van Workum, D.-J. M., van de Wetering, H., Smit, S. et Vilanova, A. (2024). PanVA : Pangenomic Variant Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):4895–4909. Conference Name : IEEE Transactions on Visualization and Computer Graphics.

Van Nieuwenhuyse, B., Merabishvili, M., Goeders, N., Vanneste, K., Bogaerts, B., de Jode, M., Ravau, J., Wagemans, J., Belkhir, L. et Van der Linden, D. (2024). Phage-Mediated Digestive Decolonization in a Gut-On-A-Chip Model : A Tale of Gut-Specific Bacterial Prosperity. *Viruses*, 16(7):1047. Number : 7 Publisher : Multidisciplinary Digital Publishing Institute.

Vera-Ponce de León, A., Hensen, T., Hoetzinger, M., Gupta, S., Weston, B., Johnsen, S. M., Rasmussen, J. A., Clausen, C. G., Pless, L., Veríssimo, A. R. A., Rudi, K., Snipen, L., Karlsen, C. R., Limborg, M. T., Bertilsson, S., Thiele, I., Hvidsten, T. R., Sandve, S. R., Pope, P. B. et La Rosa, S. L. (2024). Genomic and functional characterization of the Atlantic salmon gut microbiome in relation to nutrition and health. *Nature Microbiology*, 9(11):3059–3074. Publisher : Nature Publishing Group.

Vieira-Silva, S. et Rocha, E. P. C. (2010). The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. *PLOS Genetics*, 6(1):e1000808. Publisher : Public Library of Science.

Wang, H., Yang, Y., Xu, Y., Chen, Y., Zhang, W., Liu, T., Chen, G. et Wang, K. (2024a). Phage-based delivery systems : engineering, applications, and challenges in nanomedicines. *Journal of Nanobiotechnology*, 22(1):1–31. Number : 1 Publisher : BioMed Central.

Wang, M., Liu, G., Liu, M., Tai, C., Deng, Z., Song, J. et Ou, H.-Y. (2024b). ICEberg 3.0 : functional categorization and analysis of the integrative and conjugative elements in bacteria. *Nucleic Acids Research*, 52(D1):D732–D737.

Wang, Y., Wei, X., Bao, H. et Liu, S.-L. (2014). Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC Genomics*, 15(1):50.

Watanabe, T., Kure, A. et Horiike, T. (2023). OrthoPhy : A Program to Construct Ortholog Data Sets Using Taxonomic Information. *Genome Biology and Evolution*, 15(3):evad026.

Watkins, D. et Arya, D. P. (2019). Regulatory roles of small RNAs in prokaryotes : parallels and contrast with eukaryotic miRNA. *Non-coding RNA Investigation*, 3(0). Number : 0 Publisher : AME Publishing Company.

Weller, C. et Wu, M. (2015). A generation-time effect on the rate of molecular evolution in bacteria. *Evolution*, 69(3):643–652.

Wilpiszeski, R. L., Aufrecht, J. A., Retterer, S. T., Sullivan, M. B., Graham, D. E., Pierce, E. M., Zablocki, O. D., Palumbo, A. V. et Elias, D. A. (2019). Soil Aggregate Microbial Communities : Towards Understanding Microbiome Interactions at Biologically Relevant Scales. *Applied and Environmental Microbiology*, 85(14):e00324–19. Publisher : American Society for Microbiology.

Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A. et Brinkman, F. (2016). Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. *Nucleic Acids Research*, 44(D1):D646–D653.

Woese, C. R. et Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain : The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090. Publisher : Proceedings of the National Academy of Sciences.

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z. et Zhang, L. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal*, 19:6301–6314.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F. O. et Rosselló-Móra, R. (2008). The All-Species Living Tree project : A 16S rRNA-based phylogenetic tree of all sequenced type strains. *Systematic and Applied Microbiology*, 31(4):241–250.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M. et Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 67(11):2640–2644.

Zhao, Y., Sun, C., Zhao, D., Zhang, Y., You, Y., Jia, X., Yang, J., Wang, L., Wang, J., Fu, H., Kang, Y., Chen, F., Yu, J., Wu, J. et Xiao, J. (2018). PGAP-X : extension on pan-genome analysis pipeline. *BMC Genomics*, 19(1):36.

Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J. et Yu, J. (2012). PGAP : pan-genomes analysis pipeline. *Bioinformatics*, 28(3):416–418.

Zinder, N. D. et Lederberg, J. (1952). Genetic exchange in salmonella. *Journal of Bacteriology*, 64(5):679–699. Publisher : American Society for Microbiology.

ANNEXES

Table des Annexes

A Liste des publications, conférences et associés **211**

A.1 Publications 211

A.2 Présentation orale en conférence 211

A.3 Posters et associés 212

A.4 Certificats, Prix ET Bourses 212

B Poster **213**

B.1 FEMS Conference on Microbiology à Belgrade, Serbie, 2022 213

B.2 JOBIM Journées Ouvertes en Biologie, Informatique et Mathématiques à Rennes, France, 2022 214

B.3 ISMB/ECCB à Lyon, France 214

A - Liste des publications, conférences et associés

A.1 . Publications

- J. Arnoux, A. Bonifati, A. Calteau, S. Dumbrava and G. Gautreau, "Integrating Complex Pangenome Graphs," 2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW), Utrecht, Netherlands, 2024, pp. 350-354, doi : 10.1109/ICDEW61823.2024.00052.
- J. Arnoux, J. Mainguy, L. Bry, Q. Fernandez de Grado, D. Vallenet, A. Calteau, "Panorama : A robust pangenome-based method for predicting and comparing biological systems across species"

A.2 . Présentation orale en conférence

• Mai 2024	Integrating Complex Pangenome Graphs,	IEEE 40th International Conference on Data Engineering Workshops (ICDEW) à Utrecht, Pays-Bas
• Mai 2024	PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes	Journée de l'école doctorale SDSV
• Octobre 2023	PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes	The local pangenome à Alicante, Espagne
• Septembre 2023	From Genomics to Pangenomics : the path to pangenome graph	The 20th MicroScope anniversary à Évry-Courcouronnes, France
• Juillet 2023	PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes	ISMB/ECCB à Lyon, France
• Juin 2023	PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes	Mini Congrès du Genoscope à Évry-Courcouronnes, France
• Juin 2022	PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes	Mini Congrès du Genoscope à Évry-Courcouronnes, France

A.3 . Posters et associés

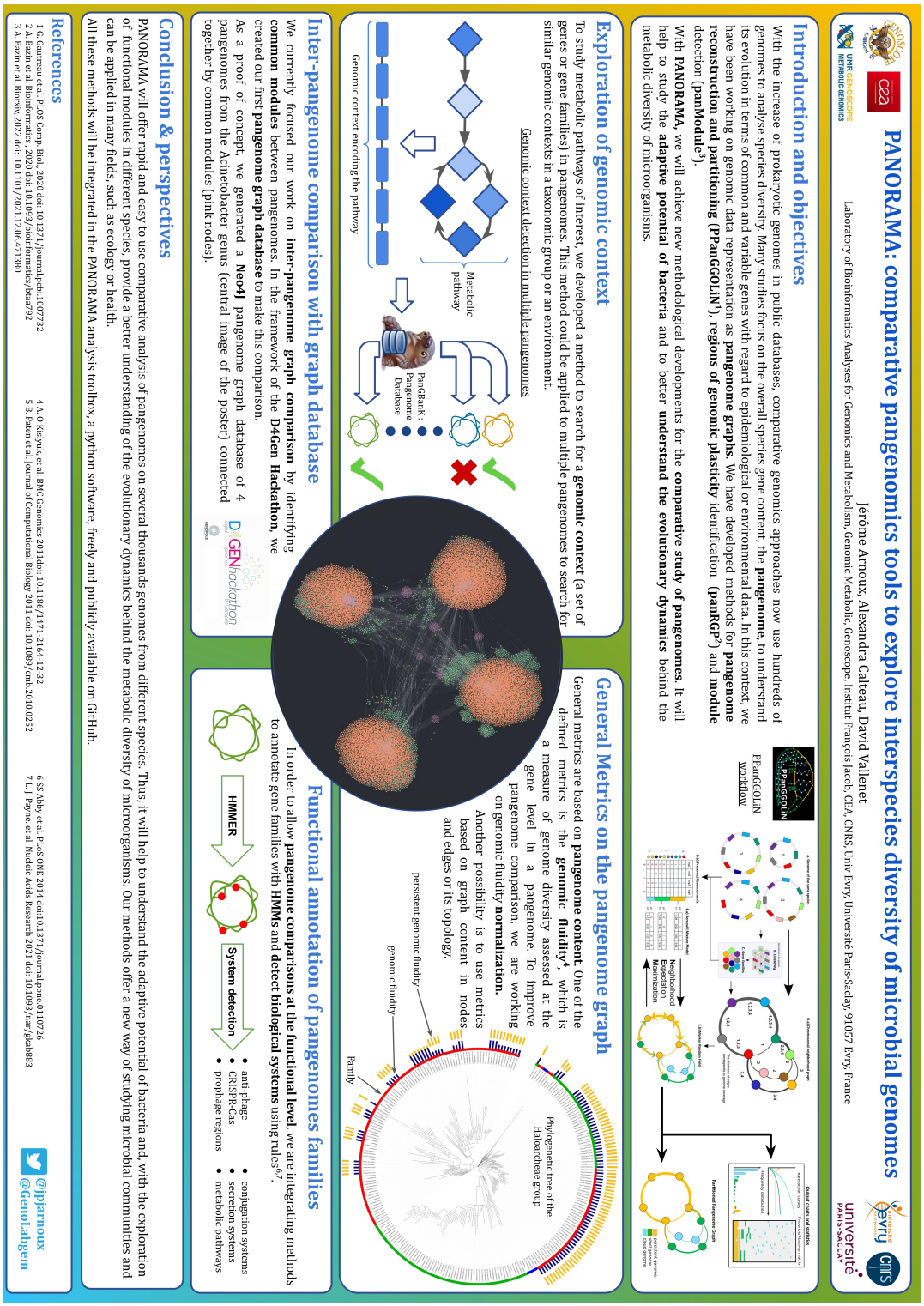
• Juillet 2023 :	PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes	ISMB/ECCB à Lyon, France
• Mai 2023 :	PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes	Journée de l'école doctorale SDSV
• Juillet 2022 :	PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes	JOBIM à Rennes, France
• Juin 2022 :	PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes	FEMS Conference on Microbiology à Belgrade, Serbie
• Mai 2022	PANORAMA : comparative pangenomics tools to explore interspecies diversity of microbial genomes	Doctoral school day SDSV

A.4 . Certificats, Prix ET Bourses

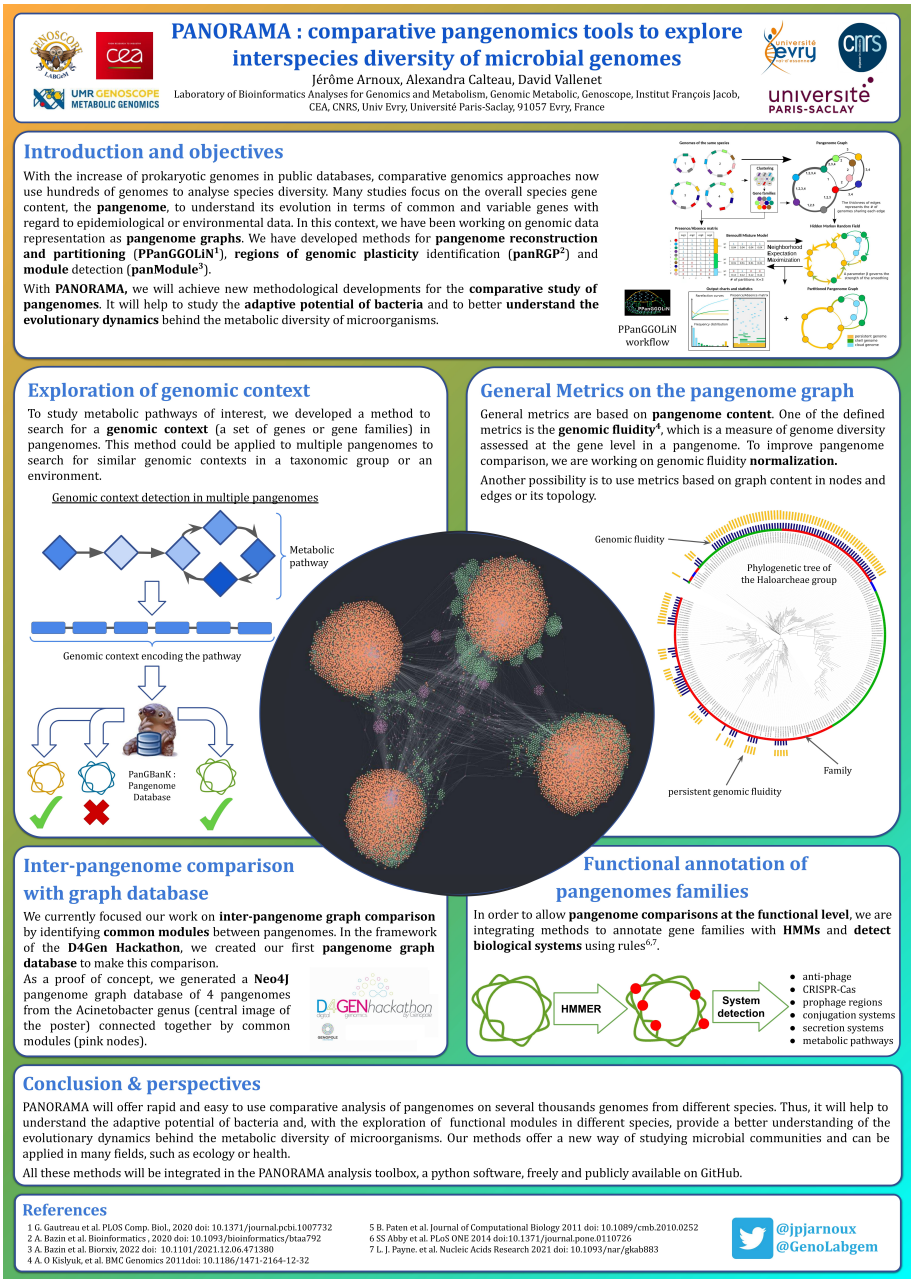
- 2023 : Prix "science ouverte du logiciel libre de la recherche", "espoir" de la catégorie 'Scientifique et technique'
- 2022 : GDR BIM Bourse de voyage pour la conférence ISMB/ECCB, Lyon
- 2022 : 3^{e} place du D4GEN Hackathon

B - Poster

B.1 . FEMS Conference on Microbiology à Belgrade, Serbie, 2022



B.2 . JOBIM Journées Ouvertes en Biologie, Informatique et Mathématiques à Rennes, France, 2022



B.3 . ISMB/ECCB à Lyon, France

PANORAMA: comparative pangenomics tools to explore interspecies diversity of microbial genomes

Jérôme Arnoux, Laura Bry, Quentin Fernandez De Grado, David Vallenet, Alexandra Calteau

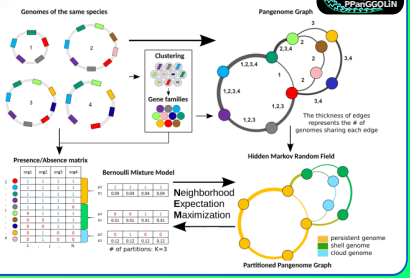
LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

OUTLINE

PANORAMA is an open-source bioinformatics toolbox, including new methodological developments for the comparative study of pangenomes. It benefits from methods for the **reconstruction and analysis of pangenome graphs**, thanks to the **PPanGGOLiN**[1] software suite. PANORAMA integrates multiple features, such as the possibility to **compare genomic context** between pangenomes or the **annotation of biological systems** at the pangenome level.

PANORAMA allows comparative analysis of pangenomes using thousands of genomes. The software is developped in python and can be install with conda. It generates many tables and graphical outputs, some of them being compatible with external bioinformatic tools.

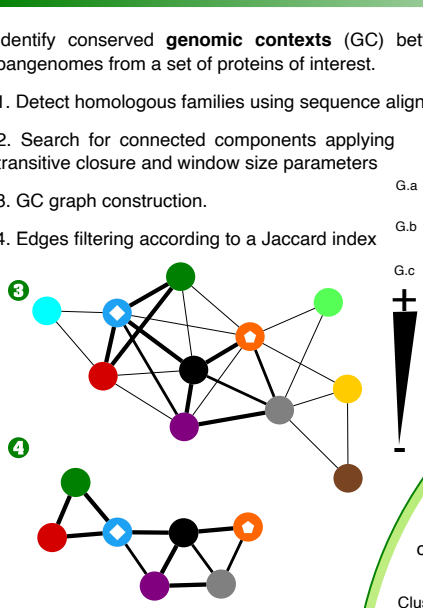
PANORAMA aims to help microbiologists understand the adaptive potential of bacteria and, through the exploration of functional modules in different species, to better understand the evolutionary dynamics behind the metabolic diversity of microorganisms.



GENOMIC CONTEXT EXPLORATION

Identify conserved **genomic contexts** (GC) between pangenomes from a set of proteins of interest.

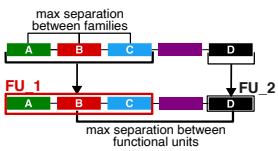
1. Detect homologous families using sequence alignment
2. Search for connected components applying transitive closure and window size parameters
3. GC graph construction.
4. Edges filtering according to a Jaccard index



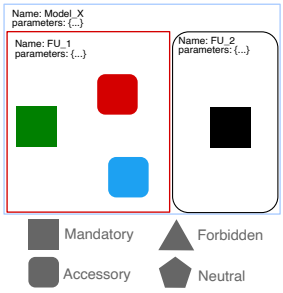
SYSTEM MODELING

PANORAMA models **biological systems** based on rules which describes gene family presence/absence and synteny conservation, similarly to MacSyFinder[4].

Model reprensation in a genome



Model reprensation schema



461models integrated:
- defense(435)
- secretion(26)

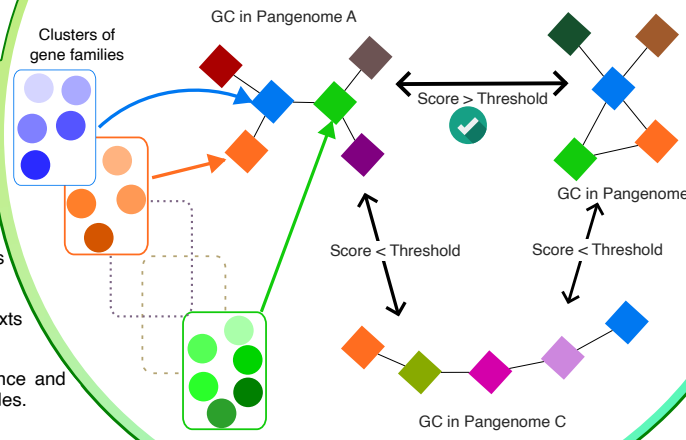
Ongoing systems:
- KEGG modules
- Polysaccharides Utilization Loci

Model structure in JSON grammar

```
{"name": "model_X",
 "parameters": {
   "max_separation": 1,
   "min_mandatory": 1,
   "min_total": 1
 },
 "func_units": [
   {"name": "FU_1",
    "type": "mandatory",
    "parameters": {
      "min_total": 2
    },
    "families": [
      {"name": "famA",
       "type": "mandatory"
      },
      {"name": "famB",
       "type": "accessory"
      },
      {"name": "famC",
       "type": "accessory",
       "parameters": {
         "max_separation": 2
       }
      }
    ]
   },
   {"name": "FU_2",
    "type": "accessory",
    "families": [
      {"name": "famD",
       "type": "mandatory"
      }
    ]
   }
 ]
}
```

COMPARATIVE PANGENOMICS ANALYSES

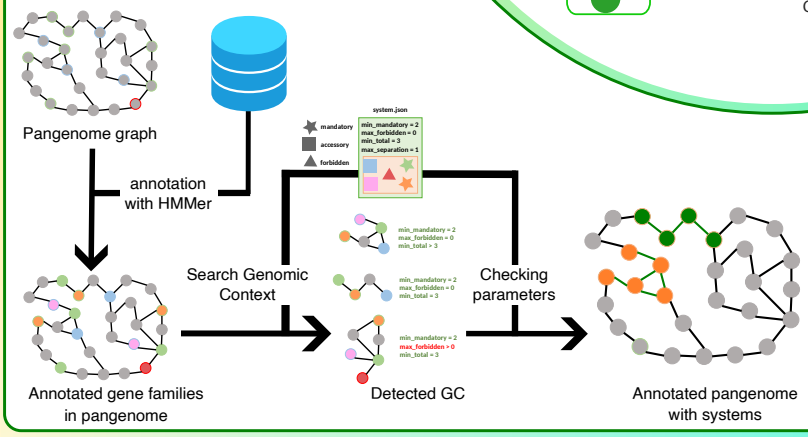
PANORAMA allows the identification of conserved GCs between pangenomes. Comparisons are based on grouping pangenome families and computing common connected components. A conservation score is calculated for each pair of GCs to obtain those that are common and may correspond to pathways or cellular processes conserved between species..



SYSTEM DETECTION

System detection in PANORAMA is based on a functional annotation of gene families, GC searching and system modeling.

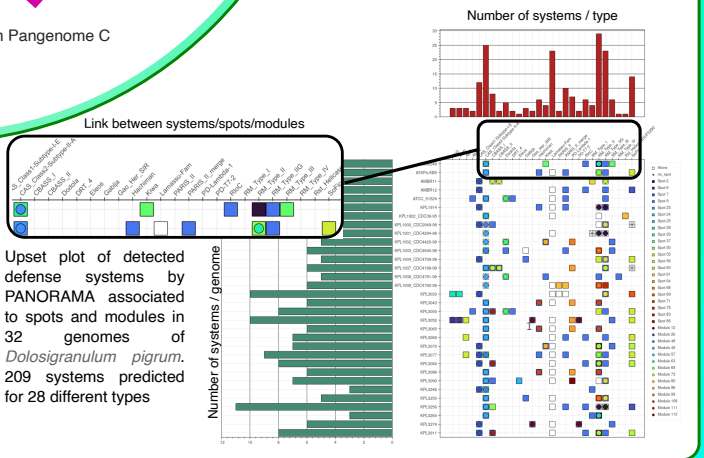
1. Annotation of gene families with HMMs associated to system-models
2. Detection of genomic contexts corresponding to the models
3. PANORAMA verifies that presence/absence and synteny parameters fit with the associated rules.



SYSTEM-ASSOCIATION

Systems detected at the pangenome level can be projected to individual genomes and associated to pangenome graph components (partition, region of genomic plasticity[2], spots[2] and modules[3]) and give them a functional annotation.

Using PANORAMA to predict defense systems allow to identify regions of genomic plasticity corresponding to anti-phage defense islands in genomes



Upset plot of detected defense systems by PANORAMA associated to spots and modules in 32 genomes of *Dolosigranulum pigrum*. 209 systems predicted for 28 different types

REFERENCES

1 G. Gautreau et al. PLOS Comp. Biol., 2020, doi: 10.1371/journal.pcbi.1007732
2 A. Bazin et al. Bioinformatics , 2020, doi: 10.1093/bioinformatics/btaa792
3 A. Bazin et al. Biorxiv, 2022, doi: 10.1101/2021.12.06.471380
4 N. Bertrand et al. Peer Community Journal, 2023, doi: 10.24072/pcjournal.250
5 F. Tesson et al. Nat Commun 2022, doi: 10.1038/s41467-022-30269-9
6 LJ. Payne et al. Nucleic Acids Res 2021, doi: 10.1093/nar/gkab883
7 M. Kanehisa et al. Nucleic Acids Res 2022, doi: 10.1093/nar/gkac963
8 JM. Grondin et al. J Bacteriol 2017, doi: 10.1128/JB.00860-16

ACKNOWLEDGMENT

cnrs GDR Groupement de recherche BIM Bioinformatique moléculaire

Contact

jarnoux at genoscope.cns.fr

@jpjarnoux

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Méthodes d'analyse comparée des pangénomes procaryotes : explorer la diversité génomique inter-espèces pour une meilleure compréhension du métabolisme

Mots clés : Bioinformatique, Microbiologie environnementale, Pangénomique, Dynamique des génomes, Îlot génomique, Systèmes de défense aux phages

Résumé : L'essor des projets de séquençage a généré plus d'un million de génomes procaryotes dans les bases publiques, nécessitant de nouvelles approches pour analyser cette masse de données. La suite logicielle PPanGGOLiN a été développée pour structurer ces informations sous forme de graphes de pangénome, permettant de compresser les données tout en conservant l'information de colocalisation des gènes. Elle intègre également des méthodes d'analyse de pangénome, panRGP, qui identifie les régions de plasticité génomique, et panModule, qui caractérise ces régions variables en sous-modules fonctionnels. Malgré ces avancées, aucune méthode ne permettait de comparer des pangénomes. Les travaux de cette thèse ont consisté à développer de nouvelles approches pour combler cette lacune. Tout d'abord, PPanGGOLiN a été enrichie par l'intégration de nouvelles méthodes, comme la recherche de contextes génomiques, et par une amélioration de son environnement logiciel. Ensuite, la méthode PANORAMA, qui se base sur les graphes de PPanGGOLiN, a été conçue pour annoter des systèmes macromoléculaires, en combinant des critères de présence/absence de fonctions et de colocalisation génomique, et pour comparer des pangénomes. Appliqué aux systèmes de défense bactériens contre les phages, PANORAMA a permis d'identifier des systèmes et des sites d'insertions conservés entre différentes espèces. Finalement, un premier prototype de base de données orientée graphe a été développé pour intégrer les données de plusieurs pangénomes afin d'exploiter au mieux leur information. Cette approche a permis d'analyser et de comparer des milliers de génomes bactériens et d'identifier des modules d'antibiorésistance communs à plusieurs espèces, mettant en lumière des mécanismes évolutifs partagés. Ces travaux ouvrent la voie à la pangénomique comparée, offrant un cadre inédit pour explorer le potentiel adaptatif des procaryotes et mieux comprendre leur dynamique évolutive. En facilitant la comparaison des pangénomes et l'identification de contextes génomiques conservés, ces développements contribuent à l'étude des interactions entre bactéries et à la caractérisation de systèmes biologiques d'intérêt.

Title : Methods for comparative analysis of prokaryotic pangenomes : exploring interspecies genomic diversity for a better understanding of metabolism

Keywords : Bioinformatics, Environmental microbiology, Pangenomics, Genome dynamics, Genomic island, Phage defense systems

Abstract : The boom in sequencing projects has generated over a million prokaryotic genomes in public databases, requiring new approaches to analyze this mass of data. The PPanGGOLiN software suite has been developed to structure this information in the form of pangenome graphs, enabling data compression while preserving gene colocalization information. It also integrates pangenome analysis methods : panRGP, which identifies regions of genomic plasticity, and panModule, which characterizes these variable regions into functional submodules. Despite these advances, there was no method for comparing pangenomes. The aim of this thesis was to develop new approaches to fill this gap. Firstly, PPanGGOLiN was enriched by integrating new methods, such as genomic context search, and by improving its software environment. Secondly, the PANORAMA method, based on PPanGGOLiN graphs, has been designed to annotate macromolecular systems, combining criteria of presence/absence of functions and genomic colocalization, and to compare pangenomes. Applied to bacterial phage defense systems, PANORAMA identified systems and insertion sites conserved between different species. Finally, a first prototype of a graph-oriented database was developed to integrate data from several pangenomes in order to make the most of their information. This approach has made it possible to analyze and compare thousands of bacterial genomes, and to identify antibiotic resistance modules common to several species, highlighting shared evolutionary mechanisms. This work paves the way for comparative pangenomics, offering a novel framework for exploring the adaptive potential of prokaryotes and better understanding their evolutionary dynamics. By facilitating the comparison of pangenomes and the identification of conserved genomic contexts, these developments contribute to the study of interactions between bacteria and the characterization of biological systems of interest.