Des génomes au métabolisme des microorganismes

David Vallenet

Université d'Évry-Val-d'Essonne

HDR

Mémoire présenté en vue de l'obtention de
l'Habilitation à Diriger des Recherches

Des génomes au métabolisme des microorganismes

David Vallenet

Soutenue le 23 juin 2020 devant le jury constitué de :

Fabien Jourdan	Directeur de Recherches, INRAE Toulouse	Rapporteur
Pierre Peterlongo	Chargé de Recherches, INRIA Rennes	Rapporteur
Jacques van Helden	Professeur, Université Aix-Marseille	Rapporteur
Hélène Chiapello	Ingénieure de Recherche, INRAE Jouy-en-Josas	Examinatrice
Frédérique Le Roux	Directrice de Recherches, Ifremer Roscoff	Examinatrice
Christophe Ambroise	Professeur, Université d'Évry-Val-d'Essonne	Examinateur

Table des matières

Préambule

Ce mémoire présente les activités de recherche que j'ai menées depuis la préparation d'un Doctorat débutée en 2002 et qui se poursuivent, depuis 2007, dans le cadre de mes fonctions de chercheur au Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme (LABGeM, sous la direction de Claudine Médigue). Le LABGeM est un laboratoire de l'Unité Mixte de Recherche "Génomique Métabolique" (UMR8030) du Genoscope (sous la direction de Patrick Wincker) qui est un département du Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) au sein de l'Institut de biologie François Jacob. L'UMR8030 est également rattachée au Centre National de la Recherche Scientifique (CNRS) et à l'Université d'Évry (membre associé de l'Université Paris Saclay).

Biologiste cellulaire de formation initiale, je me suis orienté vers une spécialisation en bioinformatique par goût pour l'informatique et la génomique. Pour cela, j'ai rejoint, en 2000, la "jeune pousse" académique initiée par Claudine Médigue sous la forme d'un laboratoire ATIGE (Actions Thématiques Incitatives de la Genopole d'Évry). Suite à un stage de Maîtrise suivi d'un Master en alternance en bioinformatique, ma formation universitaire s'est poursuivie par un Doctorat réalisé au Genoscope et dont la thèse a été soutenue en février 2007. Après l'effort considérable de séquençage du chromosome 14 humain, le Genoscope, sous l'impulsion de son directeur Jean Weissenbach, souhaitait orienter les activités de recherche du Genoscope vers la génomique environnementale pour explorer la diversité des microorganismes et des fonctions encodées par leurs gènes notamment autour des connaissances sur le métabolisme. Dans ce contexte, notre équipe avait pour objectif de concevoir des méthodes et logiciels informatiques dédiés à l'étude de génomes procaryotes. En termes d'application, *Acinetobacter baylyi* ADP1, une bactérie ayant des propriétés différentes de celles de *Bacillus subtilis* ou *Escherichia coli*, a été choisie comme nouveau modèle pour une analyse approfondie de son métabolisme. En parallèle de l'analyse de son génome, plusieurs approches expérimentales ont été initiées au sein de l'UMR8030 dont notamment la constitution d'une collection de mutants et leur phénotypage. Cette dynamique de recherche basée sur la complémentarité de l'information génomique et des approches expérimentales pour la caractérisation de processus métaboliques est ainsi au cœur de mes activités.

Portée par la dynamique des activités de service de séquençage du Genoscope et de nombreuses collaborations avec des microbiologistes désirant analyser le génome de leurs bactéries modèles, la maturité et le caractère innovant des développements informatiques réalisés nous ont amené à proposer à la communauté une plateforme d'annotation collaborative de génomes microbiens (nommée MaGe pour "Magnifying Genomes" puis MicroScope à partir de 2007). Aujourd'hui, MicroScope est utilisé par une large communauté de microbiologistes (plus de 4 700 comptes dont 65% à l'international) et a permis d'analyser plus de 14 000 génomes (plus de 1 000 citations depuis 2006). En parallèle, j'ai participé à plusieurs projets d'analyse de génomes microbiens tout en développant de nouvelles méthodes bioinformatiques au travers de l'encadrement de plusieurs thèses de Doctorat, postdocs et stages de Master.

Le projet de recherche que je conduis s'inscrit dans la compréhension des écosystèmes notamment face aux grands enjeux environnementaux. Il offre également des applications dans le biocontrôle, la bioremédiation, la valorisation de la biomasse pour la production d'énergie et la découverte de nouveaux catalyseurs pour une chimie durable. Dans cet objectif, je compte ainsi proposer de nouvelles approches méthodologiques dans l'analyse des génomes procaryotes et de leur métabolisme tout en conduisant des bioanalyses pour une caractérisation fine des espèces, fonctions et interactions présentes dans un écosystème.

Curriculum vitae

VALLENET David,
Né le 28 août 1977,
Marié, 2 enfants

Coordonnées professionnelles :
✉ CEA/Genoscope & CNRS-UMR 8030 / LABGeM
2, rue Gaston Crémieux 91037 Evry Cedex
☎ (+33) (0)1 60 87 84 53
✉ vallenet@genoscope.cns.fr

Domicile :
✉ 14 boulevard du Général Leclerc
77300 Fontainebleau

COMPÉTENCES

Bioinformatique :

- Développements de méthodes et logiciels bioinformatiques (algorithmes, systèmes d'information) pour l'analyse des génomes (annotation fonctionnelle, génomique comparée, pangénomique) et du métabolisme (prédiction de réseaux métaboliques, analyse de familles d'enzymes) des microorganismes
- Bioanalyses : valorisation de résultats expérimentaux et bioinformatiques dans le cadre de collaborations académiques et industrielles

Management :

- Encadrement de chercheurs/postdocs, ingénieurs et étudiants en thèse
- Participation à la conception et au pilotage de projets R&D académiques et industriels
- Gestion de projets et services selon les standards ISO 9001:2015 et NF X50-900:2016

DIPLÔMES UNIVERSITAIRES

- **Doctorat** spécialité Bioinformatique, Université d'Évry-Val-d'Essonne, Ecole doctorale des Génomes aux Organismes (Février 2007)
- **DESS/Master** Etude des Génomes : Outils Informatiques et Statistiques, Université de Rouen (Juin 2002)
- **Maîtrise** de Biologie Cellulaire et Physiologie, Université d'Auvergne (Juin 2000)
- **Licence** de Biologie Cellulaire et Physiologie, Université d'Auvergne (Juin 1999)

SITUATION ACTUELLE

- **Chercheur CEA** en bioinformatique depuis Mai 2007 au **Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme** (LABGeM, Dir. C. Médigue) du CEA/DRF/Genoscope & CNRS UMR8030 «Génomique métabolique» (Dir. P. Wincker)
- **Responsable scientifique de la plateforme MicroScope :**
    - Plateforme d'analyse de génomes microbiens ouverte à une communauté internationale de microbiologistes (>4700 comptes utilisateurs, >1000 citations)
    - Labellisée IBiSA depuis 2006, membre de l'Institut Français de Bioinformatique (IFB) et de France Génomique (FG)
- **Encadrement de chercheurs/postdocs, ingénieurs et étudiants en thèse** (actuellement 1 chercheur, 6 ingénieurs et 3 doctorants)

SITUATIONS PASSÉES

- **2002-2007, Thèse de Doctorat, « La plateforme MicroScope pour l'annotation de génomes microbiens. Application à l'étude de trois bactéries du genre Acinetobacter »**.
  Laboratoire de Génomique Comparative (Dir. C. Médigue), Genoscope & CNRS UMR8030
- **2000-2002, Stage de DESS, « Conception d'une base de données pour l'étude de génomes microbiens pathogènes : PathoDB. Exploration fonctionnelle et métabolique ».**
  Laboratoire de Génomique Comparative (Dir. C. Médigue), Genoscope & CNRS UMR8030 / Metabolic Explorer SA (Dir. B. Gonzalez)

EXPERTISES

- **Membre nommé de la CID51 du Comité National de la Recherche Scientifique :** membre du bureau, évaluation des chercheurs et structures, jury de concours CRCN et DR (depuis 2017)
- **Membre du comité de pilotage de la plateforme Biomics** de l'Institut Pasteur (depuis 2019)
- **Jury de concours pour des postes ITA (CNRS, INSERM et CEA)**
- **Comité et jury de thèses**
- **Reviewer de projets** (NSF, Région Aquitaine, IBISBA) **et de publications scientifiques** (Bioinformatics, BMC genomics, PLOS Comput. Biol. *etc.*)

ENSEIGNEMENTS

- **Formations continues** « Annotation and analysis of prokaryotic genomes using the MicroScope platform », Univ. d'Evry-Val-d'Essonne - Durée de 4,5 jours, 2 à 3 sessions par an.
- **Cours** « Reconstruction des réseaux métaboliques », **Master 2 ICBM** module Ingénierie Métabolique, Univ. Paris-Saclay - 6 heures par an depuis 2016
- **Cours** « Plateforme MicroScope et analyses bioinformatiques pour la microbiologie clinique », **Master 2 IMVI**, Univ. Paris Diderot – 1,5 heures par an depuis 2018
- **Module** « Les génomes procaryotes », **Master 1 GENIOMHE**, Univ. d'Evry-Val-d'Essonne – 27 heures en 2019

- **Workshop** « Metabolic network reconstruction », **école thématique ASSB**, 2 heures en 2016 et 2017

ENCADREMENTS DE TRAVAUX DE RECHERCHE

Thèses de Doctorat :

- **Gilles VIERA**, Thèse de l'Université d'Evry-Val-d'Essonne. Bourse MRT. Soutenue le 05 décembre 2011. « Reconstruction de réseaux métaboliques de souches de Escherichia coli et analyse des modèles à haut débit ». Sous la direction de : C. Médigue / Encadrement : M. Durot et D. Vallenet
- **Alexander Adam SMITH**, Thèse de l'Université d'Evry-Val-d'Essonne. Bourse Irtelis - CEA. Soutenue le 03 Février 2012. « Méthodes de reconstruction de réseaux métaboliques et de contexte génomique : application à la recherche de gènes candidats pour les activités enzymatiques orphelines de séquence ». Sous la direction de : C. Médigue / Encadrement : D. Vallenet
- **Maria SOROKINA**, Thèse de l'Université Paris-Saclay. Bourse Irtelis - CEA. Soutenue le 03 Février 2016. « Découverte et exploration des modules conservés de transformations chimiques dans le métabolisme ». Sous la direction de : C. Médigue et D. Vallenet
- **Jonathan MERCIER**, Thèse de l'Université Paris-Saclay. Financée par le laboratoire. Soutenue le 15 Mai 2017. « Logique paracohérente pour l'annotation fonctionnelle des génomes au travers de réseaux biologiques ». Sous la direction de : C. Médigue et D. Vallenet
- **Guilhem ROYER**, Thèse de l'Université Paris-Saclay depuis le 1er novembre 2016. Bourse APHP. « Génomique comparative à grande échelle de souches de Escherichia coli responsables de bactériémies chez l'Homme : implications cliniques et rôle des réseaux métaboliques ». Sous la direction de : C. Médigue, J.-W. Decousser, D. Vallenet
- **Guillaume GAUTREAU**, Thèse de l'Université Paris-Saclay depuis le 1er octobre 2016. Bourse Irtelis - CEA. Soutenance prévue le 27 Février 2020. « Conceptualisation et exploitation d'un graphe de pangénome partitionné comme représentation compacte de la diversité du répertoire génique des espèces procaryotes ». Sous la direction de : C. Médigue et D. Vallenet
- **Adelme BAZIN**, Thèse de l'Université Paris-Saclay depuis le 1er octobre 2018. Bourse CEA CFR Phare. « Analyse comparée des pangénomes : de la plasticité des génomes à la diversité métabolique du monde microbien ». Sous la direction de : C. Médigue et D. Vallenet / Encadrement : A. Calteau

Postdocs :

- **Stéfan ENGELEN**, Postdoc de 2008 à 2009 dans le cadre du projet ANR PFTV MicroScope. « Développement d'un moteur de Workflow pour le système de production de MicroScope »
- **Eugenio BELDA**, Postdoc de 2010 à 2013 dans le cadre du projet Européen FP7 MICROME. « Curation des données métaboliques de 4 génomes de référence (réactions, et associations Gene-Protein-Reaction) – développement d'outils pour la curation des réseaux métaboliques »
- **Benjamin VIART**, Postdoc de 2016 à 2018 dans le cadre du projet ANR Blue Enzyme. « Exploration du contexte génomique des enzymes, méthode NETSYN »

Production scientifique

I-Publications à comité de lecture

Depuis 2002, 69 publications (12 en premier ou dernier auteur) citées plus de 9100 fois (>750 citations par an) avec un h-index de 38, source Google scholar : https://scholar.google.fr/citations?user=rJNPLSAAAAAJ.

Publications méthodologiques :

1. Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., … **Vallenet, D.** (2019). PPanGGOLiN: Depicting microbial diversity via a Partitioned Pangenome Graph. In revision at PLoS Computational Biology. https://doi.org/10.1101/836239

2. Royer, G., Decousser, J. W., Branger, C., Dubois, M., Médigue, C., Denamur, E., & **Vallenet, D.** (2018). PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. Microbial Genomics, 4(9).

3. Mercier, J., Josso, A., Médigue, C., & **Vallenet, D.** (2018). GROOLS: reactive graph reasoning for genome annotation through biological processes. BMC Bioinformatics, 19(1), 132.

4. Sorokina, M., Medigue, C., & **Vallenet, D.** (2015). A new network representation of the metabolism to detect chemical transformation modules. BMC Bioinformatics, 16(1), 385.

5. Smith, A. A. T., Belda, E., Viari, A., Medigue, C., & **Vallenet, D.** (2012). The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. PLoS Computational Biology, 8(5), e1002540.

6. Cruveiller, S., Le Saux, J., **Vallenet, D.**, Lajus, A., Bocs, S., & Medigue, C. (2005). MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. Nucleic Acids Research, 33(suppl_2), W471–W479.

7. Bocs, S., Cruveiller, S., **Vallenet, D.**, Nuel, G., & Medigue, C. (2003). AMIGene: annotation of microbial genes. Nucleic Acids Research, 31(13), 3723–3726.

Publications sur la plateforme MicroScope :

1. **Vallenet, D.**, Calteau, A., Dubois, M., Amours, P., Bazin, A., Beuvin, M., … Médigue, C. (2019). MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. Nucleic Acids Research.

2. Médigue, C., Calteau, A., Cruveiller, S., Gachet, M., Gautreau, G., Josso, A., … **Vallenet, D.**. (2017).

MicroScope—an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data. Briefings in Bioinformatics.

3. **Vallenet, D.**, Calteau, A., Cruveiller, S., Gachet, M., Lajus, A., Josso, A., … Others. (2016). MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. Nucleic Acids Research, 45(D1), D517–D528.

4. Belda, E., **Vallenet, D.**, & Médigue, C. (2015). Accurate microbial genome annotation using an integrated and user-friendly environment for community expertise of gene functions: the microscope platform. In Hydrocarbon and Lipid Microbiology Protocols (pp. 141–169). Springer, Berlin, Heidelberg.

5. **Vallenet, D.**, Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., … Others. (2012). MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Research, 41(D1), D636–D647.

6. **Vallenet, D.**, Engelen, S., Mornico, D., Cruveiller, S., Fleury, L., Lajus, A., … Others. (2009). MicroScope: a platform for microbial genome annotation and comparative genomics. Database, 2009.

7. **Vallenet, D.**, Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., … Medigue, C. (2006). MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Research, 34(1), 53–65.

Publications d'analyse de données omiques :

1. Monteil, C. L., **Vallenet, D.**, Menguy, N., Benzerara, K., Barbe, V., Fouteau, S., … Lefevre, C. T. (2019). Ectosymbiotic bacteria at the origin of magnetoreception in a marine protist. Nature Microbiology, Vol. 4, pp. 1088–1095.

2. Desroches, M., Royer, G., Roche, D., Mercier-Darty, M., **Vallenet, D.**, Médigue, C., … Decousser, J.-W. (2018). The Odyssey of the Ancestral Escherich Strain through Culture Collections: an Example of Allopatric Diversification. mSphere, 3(1).

3. Gully, D., Czernic, P., Cruveiller, S., Mahé, F., Longin, C., **Vallenet, D.**, … Cartieaux, F. (2018). Transcriptome Profiles of Nod Factor-independent Symbiosis in the Tropical Legume Aeschynomene evenia. Scientific Reports, 8(1), 10934.

4. Borriss, R., Danchin, A., Harwood, C. R., Médigue, C., Rocha, E. P. C., Sekowska, A., & **Vallenet, D.** (2018). Bacillus subtilis, the model Gram-positive bacterium: 20 years of annotation refinement. Microbial Biotechnology, 11(1), 3–17.

5. Lassalle, F., Planel, R., Penel, S., Chapulliot, D., Barbe, V., Dubost, A., … Others. (2017). Ancestral Genome Estimation Reveals the History of Ecological Diversification in Agrobacterium. Genome Biology and Evolution, 9(12), 3413–3431.

6. Gully, D., Teulet, A., Busset, N., Nouwen, N., Fardoux, J., Rouy, Z., … Giraud, E. (2017). Complete Genome Sequence of Bradyrhizobium sp. ORS285, a Photosynthetic Strain Able To Establish Nod Factor-Dependent or Nod Factor-Independent Symbiosis with Aeschynomene Legumes. Genome Announcements, 5(30), e00421–17.

7. Blanchard, L., Guérin, P., Roche, D., Cruveiller, S., Pignol, D., **Vallenet, D.**, … Groot, A. (2017).

Conservation and diversity of the IrrE/DdrO-controlled radiation response in radiation-resistant Deinococcus bacteria. MicrobiologyOpen, 6(4).

8. Gerbore, J., Brutel, A., Lemainque, A., Mairey, B., Médigue, C., **Vallenet, D.**, … Grizard, D. (2016). Complete genome sequence of Bacillus methylotrophicus strain B25, a potential plant growth-promoting rhizobacterium. Genome Announcements, 4(2), e00058–16.

9. de Groot, A., Roche, D., Fernandez, B., Ludanyi, M., Cruveiller, S., Pignol, D., … Blanchard, L. (2014). RNA sequencing and proteogenomics reveal the importance of leaderless mRNAs in the radiation-tolerant bacterium Deinococcus deserti. Genome Biology and Evolution, 6(4), 932–948.

10. Sugawara, M., Epstein, B., Badgley, B. D., Unno, T., Xu, L., Reese, J., … Others. (2013). Comparative genomics of the core and accessory genomes of 48 Sinorhizobium strains comprising five genospecies. Genome Biology, 14(2), R17.

11. Belda, E., Sekowska, A., Le Fèvre, F., Morgat, A., Mornico, D., Ouzounis, C., … Danchin, A. (2013). An updated metabolic view of the Bacillus subtilis 168 genome. Microbiology, 159(4), 757–770.

12. Andres, J., Arsene-Ploetze, F., Barbe, V., Brochier-Armanet, C., Cleiss-Arnold, J., Coppee, J.-Y., … Others. (2013). Life in an arsenic-containing gold mine: genome and physiology of the autotrophic arsenite-oxidizing bacterium Rhizobium sp. NT-26. Genome Biology and Evolution, 5(5), 934–953.

13. Zoropogui, A., Pujic, P., Normand, P., Barbe, V., Belli, P., Graindorge, A., … Others. (2013). The Nocardia cyriacigeorgica GUH-2 genome shows ongoing adaptation of an environmental Actinobacteria to a pathogen's lifestyle. BMC Genomics, 14(1), 286.

14. Engelen, S., **Vallenet, D.**, Médigue, C., & Danchin, A. (2012). Distinct co-evolution patterns of genes associated to DNA polymerase III DnaE and PolC. BMC Genomics, 13(1), 69.

15. Alonso-Vega, P., Normand, P., Bacigalupe, R., Pujic, P., Lajus, A., **Vallenet, D.**, … Trujillo, M. E. (2012). Genome sequence of Micromonospora lupini Lupac 08, isolated from root nodules of Lupinus angustifolius. Journal of Bacteriology, 194(15), 4135–4135.

16. Zoropogui, A., Pujic, P., Normand, P., Barbe, V., Beaman, B., Beaman, L., … Others. (2012). Genome sequence of the human-and animal-pathogenic strain Nocardia cyriacigeorgica GUH-2. Journal of Bacteriology, 194(8), 2098–2099.

17. Pujol, A., Crost, E. H., Simon, G., Barbe, V., **Vallenet, D.**, Gomez, A., & Fons, M. (2011). Characterization and distribution of the gene cluster encoding RumC, an anti-C lostridium perfringens bacteriocin produced in the gut. FEMS Microbiology Ecology, 78(2), 405–415.

18. Mornico, D., Miché, L., Béna, G., Nouwen, N., Verméglio, A., **Vallenet, D.**, … Moulin, L. (2011). Comparative genomics of Aeschynomene symbionts: insights into the ecological lifestyle of nod-independent photosynthetic bradyrhizobia. Genes, 3(1), 35–61.

19. Barbe, V., Bouzon, M., Mangenot, S., Badet, B., Poulain, J., Segurens, B., … Weissenbach, J. (2011). Complete genome sequence of Streptomyces cattleya NRRL 8057, a producer of antibiotics and fluorometabolites. Journal of Bacteriology, 193(18), 5055–5056.

20. Bertin, P. N., Heinrich-Salmeron, A., Pelletier, E., Goulhen-Chollet, F., Arsène-Ploetze, F., Gallien, S., … Others. (2011). Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem

revealed by meta-and proteo-genomics. The ISME Journal, 5(11), 1735.

21. Remenant, B., de Cambiaire, J.-C., Cellier, G., Jacobs, J. M., Mangenot, S., Barbe, V., … Others. (2011). Ralstonia syzygii, the blood disease bacterium and some Asian R. solanacearum strains form a single genomic species despite divergent lifestyles. PloS One, 6(9), e24356.

22. Monot, M., Boursaux-Eude, C., Thibonnier, M., **Vallenet, D.**, Moszer, I., Medigue, C., … Dupuy, B. (2011). Reannotation of the genome sequence of Clostridium difficile strain 630. Journal of Medical Microbiology, 60(8), 1193–1199.

23. Alloisio, N., Queiroux, C., Fournier, P., Pujic, P., Normand, P., **Vallenet, D.**, … Kucho, K.-I. (2010). The Frankia alni symbiotic transcriptome. Molecular Plant-Microbe Interactions: MPMI, 23(5), 593–607.

24. Barbe, V., Cruveiller, S., Kunst, F., Lenoble, P., Meurice, G., Sekowska, A., … Others. (2009). From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later. Microbiology, 155(6), 1758–1775.

25. Jeong, H., Barbe, V., Lee, C. H., **Vallenet, D.**, Yu, D. S., Choi, S.-H., … Others. (2009). Genome sequences of Escherichia coli B strains REL606 and BL21 (DE3). Journal of Molecular Biology, 394(4), 644–652.

26. Vuilleumier, S., Chistoserdova, L., Lee, M.-C., Bringel, F., Lajus, A., Zhou, Y., … Others. (2009). Methylobacterium genome sequences: a reference blueprint to investigate microbial metabolism of C1 compounds from natural and industrial sources. PloS One, 4(5), e5584.

27. Rusniok, C., **Vallenet, D.**, Floquet, S., Ewles, H., Mouzé-Soulama, C., Brown, D., … Others. (2009). NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen Neisseria meningitidis. Genome Biology, 10(10), R110.

28. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., … Others. (2009). Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genetics, 5(1), e1000344.

29. **Vallenet, D.**, Nordmann, P., Barbe, V., Poirel, L., Mangenot, S., Bataille, E., … Others. (2008). Comparative analysis of Acinetobacters: three genomes for three lifestyles. PloS One, 3(3), e1805.

30. Muller, D., Médigue, C., Koechler, S., Barbe, V., Barakat, M., Talla, E., … Others. (2007). A tale of two oxidation states: bacterial colonization of arsenic-rich environments. PLoS Genetics, 3(4), e53.

31. Alloisio, N., Félix, S., Maréchal, J., Pujic, P., Rouy, Z., **Vallenet, D.**, … Normand, P. (2007). Frankia alni proteome under nitrogen-fixing and nitrogen-replete conditions. Physiologia Plantarum, 130(3), 440–453.

32. Normand, P., Lapierre, P., Tisa, L. S., Gogarten, J. P., Alloisio, N., Bagnarol, E., … Others. (2007). Genome characteristics of facultatively symbiotic Frankia sp. strains reflect host range and host plant biogeography. Genome Research, 17(1), 7–15.

33. Giraud, E., Moulin, L., **Vallenet, D.**, Barbe, V., Cytryn, E., Avarre, J.-C., … Others. (2007). Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia. Science, 316(5829), 1307–1312.

34. Fournier, P.-E., **Vallenet, D.**, Barbe, V., Audic, S., Ogata, H., Poirel, L., … Others. (2006). Comparative genomics of multidrug resistance in Acinetobacter baumannii. PLoS Genetics, 2(1), e7.

35. Vodovar, N., **Vallenet, D.**, Cruveiller, S., Rouy, Z., Barbe, V., Acosta, C., … Others. (2006). Complete genome sequence of the entomopathogenic and metabolically versatile soil bacterium Pseudomonas entomophila. Nature Biotechnology, 24(6), 673.

36. Strous, M., Pelletier, E., Mangenot, S., Rattei, T., Lehner, A., Taylor, M. W., … Others. (2006). Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature, 440(7085), 790.

37. Siroy, A., Cosette, P., Seyer, D., Lemaître-Guillier, C., **Vallenet, D.**, Van Dorsselaer, A., … Dé, E. (2006). Global comparison of the membrane subproteomes between a multidrug-resistant Acinetobacter baumannii strain and a reference strain. Journal of Proteome Research, 5(12), 3385–3398.

38. Siroy, A., Molle, V., Lemaître-Guillier, C., **Vallenet, D.**, Pestel-Caron, M., Cozzone, A. J., … Dé, E. (2005). Channel formation by CarO, the carbapenem resistance-associated outer membrane protein of Acinetobacter baumannii. Antimicrobial Agents and Chemotherapy, 49(12), 4876–4883.

39. Médigue, C., Krin, E., Pascal, G., Barbe, V., Bernsel, A., Bertin, P. N., … Others. (2005). Coping with cold: the genome of the versatile marine Antarctica bacterium Pseudoalteromonas haloplanktis TAC125. Genome Research, 15(10), 1325–1335.

40. Barbe, V., **Vallenet, D.**, Fonknechten, N., Kreimeyer, A., Oztas, S., Labarre, L., … Others. (2004). Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium. Nucleic Acids Research, 32(19), 5766–5779.

41. Médigue, C., Bocs, S., Labarre, L., Mathé, C., & **Vallenet, D.** (2002). L'annotation in silico des séquences génomiques-Bio-informatique (1). Médecine/sciences, 18(2), 237–250.

Publications d'analyse du métabolisme :

1. Gobet, A., Barbeyron, T., Matard-Mann, M., Magdelenat, G., **Vallenet, D.**, Duchaud, E., & Michel, G. (2018). Evolutionary Evidence of Algal Polysaccharide Degradation Acquisition by Pseudoalteromonas carrageenovora 9T to Adapt to Macroalgal Niches. Frontiers in Microbiology, Vol. 9.

2. Ficko-Blean, E., Préchoux, A., Thomas, F., Rochat, T., Larocque, R., Zhu, Y., … Others. (2017). Carrageenan catabolism is encoded by a complex regulon in marine heterotrophic bacteria. Nature Communications, 8(1), 1685.

3. Bastard, K., Perret, A., Mariage, A., Bessonnet, T., Pinet-Turpault, A., Petit, J.-L., … Others. (2017). Parallel evolution of non-homologous isofunctional enzymes in methionine biosynthesis. Nature Chemical Biology, 13(8), 858.

4. Belda, E., Van Heck, R. G. A., José Lopez-Sanchez, M., Cruveiller, S., Barbe, V., Fraser, C., … Others. (2016). The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic chassis. Environmental Microbiology, 18(10), 3403–3424.

5. Bordier, F., Stam, M., Darii, E., Tricot, S., Fossey, A., Rohault, J., … Others. (2014). Large α-aminonitrilase activity screening of nitrilase superfamily members: access to conversion and enantiospecificity by LC--MS. Journal of Molecular Catalysis. B, Enzymatic, 107, 79–88.

6. Sorokina, M., Stam, M., Médigue, C., Lespinet, O., & **Vallenet, D.** (2014). Profiling the orphan enzymes. Biology Direct, 9(1), 10.

7. Bastard, K., Smith, A. A. T., Vergne-Vaxelaire, C., Perret, A., Zaparucha, A., De Melo-Minardi, R., … Others. (2014). Revealing the hidden functional diversity of an enzyme family. Nature Chemical Biology, 10(1), 42.

8. Le Fevre, F., Mornico, D., Belda, E., **Vallenet, D.**, & Médigue, C. (2012). From automatic and expert annotation to the reconstruction of genome-scale metabolic networks and models. ModellingComplex BiologicalSystems.

9. Perret, A., Lechaplais, C., Tricot, S., Perchat, N., Vergne, C., Pellé, C., … Others. (2011). A novel acyl-CoA beta-transaminase characterized from a metagenome. PloS One, 6(8), e22918.

10. Vieira, G., Sabarly, V., Bourguignon, P.-Y., Durot, M., Le Fèvre, F., Mornico, D., … Others. (2011). Core and panmetabolism in Escherichia coli. Journal of Bacteriology, 193(6), 1461–1472.

11. Fonknechten, N., Perret, A., Perchat, N., Tricot, S., Lechaplais, C., **Vallenet, D.**, … Others. (2009). A conserved gene cluster rules anaerobic oxidative degradation of L-ornithine. Journal of Bacteriology, 191(9), 3162–3167.

12. De Berardinis, V., **Vallenet, D.**, Castelli, V., Besnard, M., Pinet, A., Cruaud, C., … Others. (2008). A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. Molecular Systems Biology, 4(1), 174.

13. Durot, M., Le Fèvre, F., de Berardinis, V., Kreimeyer, A., **Vallenet, D.**, Combe, C., … Schachter, V. (2008). Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data. BMC Systems Biology, 2(1), 85.

14. Kreimeyer, A., Perret, A., Lechaplais, C., **Vallenet, D.**, Médigue, C., Salanoubat, M., & Weissenbach, J. (2007). Identification of the last unknown genes in the fermentation pathway of lysine. The Journal of Biological Chemistry, 282(10), 7191–7197.

II-Communications orales à des conférences ou workshops

Les plus récentes depuis 2012 :

- Vallenet D. (2012) « Enzyme survey and how to find new ones », CBSO, Evry, France (Présentation orale invitée)

- Vallenet D. (2012) « Enzyme survey and how to find new ones », StarOmics symposium, Lausanne, Suisse (Présentation orale invitée)

- Vallenet D. (2014) « Enzyme survey and how to find new ones », SMPGD, Paris, France (Présentation orale invitée)

- Vallenet D., C. Médigue (2014) « Atelier annotation de génomes », Ecole thématique de Microbiologie Moléculaire, Carry Le Rouet, France (Workshop invité)

- Vallenet D. (2014) « MicroScope : an integrated platform for microbial genomics & metabolic annotation - Toward the discovery of new enzymatic activities », EBI symposium, Hinxton, UK (Présentation orale invitée)

- Mercier J. & Vallenet D. (2015) « GROOLS: Reactive Graph Reasoning for Genome Annotation », RuleML, Berlin, Allemagne (Présentation orale)

- Vallenet D. (2016) « Metabolic network reconstruction », Advances in Systems and Synthetic Biology, Evry, France (Workshop invité)

- Vallenet D. (2017) « Metabolic network reconstruction », Advances in Systems and Synthetic Biology, Lyon, France (Workshop invité)

- Vallenet D. (2018) « Genome Mining for Enzyme Discovery: the role of bioinformatics », 20 ans de Protéus, Nîmes, France (Présentation orale invitée)

- Vallenet D. (2019) « Depicting microbial species diversity via a partitioned pangenome graph », Animation Thématique pan-génomique végétale GIS Biotechnologies Vertes, Paris, France (Présentation orale invitée)

- Vallenet D. (2019) « MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic and metabolic comparative analysis », FEMS, Glasgow, UK (Présentation orale)

III-Production de logiciels

- **AMIGene**, a Web server for the prediction of protein coding regions in prokaryotic genomes, www.genoscope.cns.fr/agc/tools/amigene
- **MICheck**, a Web server to check the consistency of annotations available in public databanks, www.genoscope.cns.fr/agc/tools/micheck
- **MicroScope**, an Integrated platform for the annotation and comparative analysis of microbial genomes mage.genoscope.cns.fr/microscope. The platform integrates tools that are regularly developed at LABGeM (in total more than 25 analysis methods organized in thematic sections).
- **PlaScope**, a targeted approach to assess the plasmidome from genome assemblies at species level, github.com/labgem/PlaScope
- **GROOLS**, reactive graph reasoning for genome annotation through biological processes, github.com/Grools
- **PPanGGOLiN**, depicting microbial species diversity via a Partitioned Pangenome Graph, github.com/labgem/PPanGGOLiN
- **NetSyn**, detection of conserved genomic contexts (*i.e.* synteny conservation) among a list of protein targets, github.com/labgem/netsyn

IV-Projets financés

IV.1 Contrats de recherche académiques

Type	Titre du projet	Partenaires (PI en gras)	Thèmes des travaux	Impact	Rôle LABGeM	Début - Fin
ANR PFTV	MICROSCOPE	CEA & CNRS UMR8030 LABGeM (**C. Médigue**) & Info Genoscope (C. Scarpelli)	Développement d'une plateforme pour l'annotation et l'analyse comparative de génomes bactériens	Mise à la disposition de la communauté des microbiologistes d'un service de haut niveau et nombreuses publications collaboratives.	Responsable des développements et services proposés.	2007-2009
ANR Blanc	RARE	CNRS UMR7156 (**P. Bertin**) & CEA/CNRS UMR8030 (C. Médigue & D. LePalier) & CNRS UMR5254 (R. Duran) & CNRS UMR5569 (C. Casiot)	Reactivity of an arsenic-rich ecosystem: an integrated genomics approach	Mieux comprendre les multiples facettes d'un écosystème dont le fonctionnement est plus complexe que ne le laissait prévoir sa faible biodiversité (15 publications)	Coordinateur des bases de données, du processus d'annotation et des analyses de génomique comparative	2007-2010
ANR Blanc	COLISCOPE	INSERM U722 (**E. Denamur**) & CEA/CNRS UMR8030 (C. Médigue) & INSERM U535 (Pierre Darlu)	A sequencing project for the understanding of commensalism and virulence emergence in the *Escherichia coli/Shigella* species	Mieux comprendre les mécanismes biologiques qui participent au phénotype des souches de *E. coli*.	Annotation automatique et experte des 7 nouveaux génomes – Création et maintenance de la base de données contenant plus de 100 souches d'*E. coli* et *Shigella*	2006-2009
ANR Syscomm	METACOLI	INRA (**C. Dillmann**) & CEA/CNRS UMR8030 (C. Médigue) & INSERM (E. Denamur)	Intégration de données et modélisation de la diversité métabolique de souches commensales et virulentes d'*Escherichia coli*	Mieux comprendre les mécanismes biologiques qui participent aux phénotypes des souches de *E. coli*.	Reconstruction des réseaux métaboliques de 5 souches de *E. coli* dans 4 environnements différents	2009-2012

ANR Blanc	SESAM	IRD-RHIZO (**D. Bogusz**) & CEA/CNRS UMR 8030 (C. Médigue & D. Vallenet) & IRD LSTM (E. Giraud) & CNRS LEM (P. Normand)	Echange des signaux symbiotiques : analyse des mécanismes moléculaires les symbiotiques fixatrices d'azote Nod-indépendantes	Comprendre les mécanismes moléculaires de la symbiose nod-indépendante à l'aide d'analyses fonctionnelles	Intégration et analyses bioinformatiques des données expérimentales – Développement de la base de données et du portail Web	2011-2014
FP7	MICROME	14 partenaires Européens EBI (**P. Kersey**) Coordinateur – CEA/CNRS UMR8030 (C. Médigue) coordinateur de deux WPs	A Knowledge-Based Bioinformatics Framework for Microbial Pathway Genomics	Développement d'une ressource métabolique de référence pour les microbes à l'échelle Européenne.	Conception de l'infrastructure logicielle avec l'EBI et le SIB, responsable du WP2 « population et curation des données métaboliques » avec le SIB, et du WP7 = organiser les sessions de formation à la curation dans MICROME et les jamborées d'annotation	2010-2014
ANR Blanc	BugsInACell	CNRS ISV (**P. Mergaert**) & CEA/CNRS UMR 8030 (D. Vallenet) & IRD LSTM (E. Giraud) & CNRS EcoMic (P. Normand)	Accommodation intracellulaire des bactéries symbiotiques fixatrices d'azote	Comprendre les mécanismes d'adaptation dans la symbiose plante-bactérie pour la fixation de l'azote	Intégration et analyses bioinformatiques des données expérimentales	2014-2017
ANR	Blue Enzymes	CNRS SB Roscoff (G. Michel) & CEA/CNRS UMR 8030 (C. Médigue) & INRA IIP (Eric Duchaud)	Découverte de nouvelles enzymes pour la valorisation de la biomasse algale	Comprendre les voies de dégradations des polysaccharides d'algues par les bactéries marines.	Analyse des génomes de bactéries marines et de leur métabolisme. Développement du logiciel NetSyn pour l'analyse des contextes génomiques et d'un prédicteur de sulfatases	2015-2019
IFB-Elixir	Implementation Study	CEA/CNRS UMR 8030 (C. Médigue) & EBI (P. Kersey) &	Une ressource européenne sur le métabolisme microbien	Favoriser la standardisation et l'échanges de données avec des ressources sur le métabolisme	Export RDF de MicroScope et rédaction de spécifications	2017

		SIB (A. Bridge)		microbien dont MicroScope		
IFB	MicroCloud	CEA/CNRS UMR 8030 (**D. Vallenet**) & IFB (C. Blanchet)	Evolution de la plateforme MicroScope pour une distribution en mode SaaS	Offrir une meilleure portabilité de la plateforme sur des infrastructures Cloud.	Conduire les développements nécessaires.	2017-2018
ANR JCJC	MODAMDH	CEA/CNRS UMR 8030 (C. Vergne)	Découverte et étude d'amine déhydrogénases par approche in silico	Identifier de nouvelles familles d'enzymes	Conduire les développements et analyses bioinformatiques nécessaires.	2020-2022

IV.2 Contrats avec industriels

Introduction

Mon parcours scientifique s'inscrit dans le domaine de la bioinformatique pour l'analyse des génomes de microorganismes procaryotes et de leur métabolisme. De part leur diversité et leur capacité d'adaptation, les bactéries et archées sont un sujet d'étude passionnant aussi bien pour répondre à des questions relatives à l'origine de la vie, à l'évolution des espèces et à l'écologie que pour un intérêt médical ou biotechnologique. L'étude de leur métabolisme révèle une complexité surprenante de réactions chimiques pour la production d'énergie, de biomasse ou de métabolites secondaires. Ces réactions impliquent une multitude de molécules dont la diversité structurale est une source d'inspiration pour les chimistes dans la recherche de molécules bioactives, la chimie biomimétique et la biocatalyse.

Depuis la publication des premiers génomes complets de bactérie il y a 25 ans, les technologies de séquençage de l'ADN ont considérablement progressé et devraient permettre d'atteindre, d'ici la fin de l'année, le million de génomes procaryotes disponibles dans les banques de séquences. Pour mémoire, lorsque j'ai débuté mes travaux en bioinformatique en 2000, on ne disposait que d'une petite dizaine de génomes. Cette quantité importante de génomes bien qu'impressionnante est tout de même à relativiser de part le faible nombre d'espèces représentées (*i.e.* de l'ordre de 25 000 ce qui représente une fraction infime du nombre total d'espèces estimé entre 10^7 et 10^9). De plus, les dix espèces les plus représentées concernent uniquement des bactéries d'intérêt médical, notamment issues d'études épidémiologiques, et leurs génomes représentent près de 70% des ~500 000 génomes disponibles. Plusieurs initiatives visant à augmenter la diversité dans les bases de données de génomes ont vu le jour mais se heurtent à la difficulté des laboratoires de microbiologie à isoler et mettre en culture de nouvelles espèces.

Pour étudier la diversité génomique d'un écosystème, des approches métagénomiques ciblées (*i.e.* séquençage d'un gène marqueur conservé, généralement le gène codant la sous-unité 16S du ribosome) ou globales (*i.e.* séquençage aléatoire de l'ADN environnemental) se sont démocratisées durant ces dix dernières années. Elles permettent d'estimer la diversité en espèces tout en accédant au catalogue des gènes reconstruits à partir d'un assemblage des lectures. Plus récemment, l'amélioration des techniques d'assemblage et de classification des contigs offre la possibilité de reconstruire des milliers de génomes à partir de lectures métagénomiques (appelés MAGs pour

“Metagenome-Assembled Genome”). De même, des techniques d’amplification de l’ADN de cellules uniques permettent d’accéder aux génomes d’espèces non cultivées (appelés SAGs, pour “Single-cell Amplified Genomes”). Ces SAGs et MAGs, bien que de qualité très variable en termes de complétion et de contamination, offrent ainsi de nouvelles perspectives pour la génomique environnementale en ne considérant plus les gènes individuellement mais en travaillant sur des catalogues de génomes.

L'élucidation du métabolisme des microorganismes à partir de l’analyse de leur génome nécessite une annotation de bonne qualité qui consiste à prédire les gènes codant des protéines puis à leur assigner des fonctions enzymatiques précises. A partir de cet ensemble de fonctions prédites, un réseau métabolique est reconstruit et regroupe l’ensemble des réactions qui sont supposées exister dans la cellule. Cette étape de reconstruction suppose une conservation du métabolisme au cours de l'évolution et se base donc sur des voies métaboliques expérimentalement démontrées dans d’autres organismes ou sur des réseaux issus d’organismes modèles pour lesquels une expertise humaine a été conduite. Ce réseau peut ensuite être utilisé pour modéliser le métabolisme d’un microorganisme et prédire, par exemple, des phénotypes de croissance sur des milieux minimaux de culture. Une des grandes limites à ce type d’approche réside dans l’étape d’assignation de fonctions moléculaires aux gènes qui se base sur des relations d'homologie avec des protéines dont la fonction a été expérimentalement démontrée. Or, de nombreuses familles de protéines demeurent de fonction inconnue et, inversement, de nombreuses activités enzymatiques ne sont pas associées à une séquence d’enzyme connue.

Dans ce contexte, le premier chapitre de ce mémoire résume mes activités passées de recherche qui nous ont amené à la conception d’une plateforme d’annotation, nommée MicroScope, et à de nombreuses collaborations d’analyse de génomes de bactéries dont certaines étaient focalisées sur l’étude de leur métabolisme. MicroScope offre un ensemble d’outils pour l’annotation fonctionnelle, la génomique comparée avec notamment la détection de synténies conservées et la reconstruction des réseaux métaboliques, dans un environnement informatique intégré facilitant les expertises collaboratives entre microbiologistes. Un second volet des mes activités a porté sur le développement de méthodes bioinformatiques pour l’étude du métabolisme afin d’exploiter l’information génomique et structurale des protéines pour découvrir de nouvelles activités enzymatiques et voies métaboliques. Ce travail de recherche a été mené conjointement avec les laboratoires expérimentaux de notre UMR.

Le deuxième chapitre présente les activités de recherche en cours et leurs perspectives. Elles concernent le développement de la méthode PPanGGOLiN qui utilise une nouvelle structure de données sous la forme d’un graphe de pangénome pour représenter et partitionner l’information génomique de milliers de souches d’une même espèce. Cette méthode répond en partie aux enjeux de l’analyse de données massives en génomique et a été appliquée à l’étude de plusieurs centaines

d'espèces regroupant plus de 100 000 génomes. A partir de ce graphe de pangénome partitionné, des méthodes de prédiction d'îlots génomiques et de modules de gènes conservés sont proposées. De plus, la constitution d'une ressource de pangénomes basée sur des génomes d'isolats et des MAGs ouvre de nouvelles pistes dans le développement de méthodes utilisant des données métagénomiques pour des analyses quantitatives et fonctionnelles à l'échelle des espèces voire des souches d'un écosystème. Pour clore ce chapitre, une stratégie intégrée mêlant des méthodes computationnelles et des expérimentations est présentée. Elle consiste à identifier de nouvelles familles d'enzymes à l'aide de méthodes d'homologie lointaine, d'analogie de sites actifs et de contextes génomique, puis à explorer la diversité fonctionnelle de ces familles en réalisant des expériences de criblage d'activités enzymatiques sur des protéines représentantes.

Ce projet de recherche trouvera tout son sens au travers de collaborations avec des microbiologistes et biochimistes que je souhaite nombreuses et fructueuses. Elles contribueront à améliorer la compréhension des écosystèmes face aux grands enjeux environnementaux mais, également, à développer des applications pour le biocontrôle, la bioremédiation, la valorisation de la biomasse et la biocatalyse.

Chapitre 1 : Activités de recherche antérieures

I-De l'analyse des génomes d'Acinetobacter à la plateforme MicroScope

Biologiste cellulaire de formation initiale, je me suis orienté vers une spécialisation en bioinformatique qui s'est concrétisée par l'obtention d'un Doctorat en 2007. L'objectif premier de ma thèse était de développer des outils informatiques dédiés à l'étude de génomes microbiens et de les appliquer à l'analyse du génome d'une bactérie du sol, *Acinetobacter baylyi* ADP1. Cette analyse a été réalisée conjointement avec ma collègue biologiste, Valérie Barbe, qui préparait également un Doctorat et était en charge de l'assemblage du génome. Son annotation a révélé des caractéristiques uniques à cette bactérie en comparaison aux organismes du genre Pseudomonas [1]. Malgré un génome relativement compact (*i.e.* de taille 60 % inférieure au Pseudomonas), *A. baylyi* possède des capacités métaboliques importantes dont notamment un archipel d'îlots génomiques dédié au catabolisme d'une grande variété de composés organiques produits par le métabolisme secondaire des plantes. Suite à la publication de ce génome, une étude complémentaire a été menée sur deux bactéries du même genre, des *Acinetobacter baumannii*, dont une caractéristique principale est d'être associée à un nombre important d'infections nosocomiales. Deux articles ont ainsi été publiés. Une première analyse a montré que la souche *A. baumannii* AYE arbore un îlot génomique de 86 kb contenant 45 gènes impliqués dans la résistance aux antibiotiques et dont la présence est associée à son phénotype de multirésistance [2]. Un deuxième article a porté sur une analyse plus fine de ces génomes et a mis en évidence : (i) une conservation importante dans l'espèce *A. baumannii* des capacités métaboliques préalablement identifiées dans la souche ADP1 (ii) et un ensemble de spécificités dans le contenu en gènes qui nous a permis de formuler des hypothèses sur les capacités d'adaptation de ces bactéries à des environnements divers comme l'intestin du pou ou le milieu hospitalier [3]. Ces analyses ont été le point de départ à une meilleure compréhension des bactéries du genre Acinetobacter et se sont poursuivies au sein de l'UMR8030 par la constitution d'une collection de mutants pour la souche ADP1 réalisée par l'équipe de Véronique de Berardinis [4]. Cette collection a ensuite été utilisée pour observer des phénotypes de perte de croissance sur un grand nombre de métabolites. L'annotation du génome, associée à une ressource importante de données expérimentales, a ainsi permis

d'initier le projet Thesaurus métabolique. Ce projet visait à revisiter le métabolisme d'*A. baylyi* et à compléter les connaissances sur ses fonctions enzymatiques.

Durant ma thèse et parallèlement aux analyses de génomes du genre Acinetobacter, les activités de service de séquençage du Genoscope m'ont amené à échanger avec de nombreux collaborateurs microbiologistes. Ils portaient un intérêt particulier aux outils informatiques que nous avions développés et souhaitaient ainsi les utiliser pour analyser leurs bactéries nouvellement séquencées. C'est dans ce contexte que nous avons mis en place une plateforme d'annotation collaborative de génomes microbiens ouverte à la communauté des microbiologistes. Une première version de ce système d'information (nommé MaGe pour "Magnifying Genomes" puis MicroScope[1] à partir de 2007) a été publiée en 2006 et avait permis d'analyser et de publier une dizaine de génomes [5].

Pour replacer l'origine de la plateforme MicroScope dans le contexte scientifique de l'époque, nous disposions, au début de l'année 2000, d'un système d'annotation de génomes procaryotes appelé Imagene [6]. Ce logiciel avait été développé par Claudine Médigue qui était à l'Institut Pasteur dans l'équipe d'Antoine Danchin, François Rechenmann de l'INRIA (Institut National de Recherche en Informatique et en Automatique) et Alain Viari de l'ABI (Atelier de BioInformatique) de l'Université Paris VI. Imagene possédait une architecture très élaborée qui était basée sur un modèle objet. Au travers d'interfaces graphiques, l'utilisateur pouvait réaliser une gestion fine des données (module "Data Manager") et des analyses en tâches et sous-tâches (module "Task Manager") puis explorer les résultats sous forme de rapports textuels ou de représentations graphiques et annoter le génome (module "Result Manager"). En parallèle, nous commencions à développer des méthodes additionnelles pour l'analyse des génomes et des bases de données relationnelles pour organiser les connaissances. Stéphanie Bocs travaillait sur des méthodes de prédiction de gènes qui ont donné lieu à la publication des logiciels AMIGene [7] et MICheck [8]. Laurent Labarre développait une méthode de calcul de synténies conservées [9] et des interfaces Web de visualisation des résultats. Stéphane Descorps-Declère intégrait de nouvelles méthodes dans Imagene. Pour ma part, durant mon stage de Maîtrise et mon alternance de DESS, je réalisais des analyses de génomes et m'intéressais aux méthodes de génomique comparée et d'analyse fonctionnelle notamment pour la prédiction des réseaux métaboliques. Etant donné la complexité des développements au sein du logiciel Imagene due, notamment, à l'utilisation d'un langage de programmation dérivé de Lisp et à l'abandon de sa maintenance par les développeurs initiaux, nous avons décidé de développer un nouveau système d'annotation de génomes procaryotes : celui-ci étant basé sur des technologies Web, qui facilitent l'accès au système et favorisent l'annotation collaborative, et sur une base de données relationnelle pour la persistance des connaissances. Une première version prototype de la plateforme MicroScope a été

[1] https://mage.genoscope.cns.fr/microscope

opérationnelle dès octobre 2002 (*i.e.* début de ma thèse) pour initier l'annotation du génome de la bactérie *A. baylyi* ADP1.

En parallèle à notre initiative, un consortium public-privé regroupant l'INRIA, l'Institut Pasteur et les sociétés de biotechnologie Hybrigenics et Genome Express a vu le jour en 1999 pour développer un nouveau système d'annotation basé sur les idées originales du logiciel Imagene mais avec des technologies de programmation plus modernes et un nombre de développeurs beaucoup plus conséquent. Ce système a été nommé Genostar puis Iogma suite à sa reprise à la fin du consortium par la société Genostar fondée par François Rechenmann en 2004. Nous avons eu l'occasion de tester ce logiciel et de réaliser des développements de modules d'analyse mais étions peu convaincus de l'intérêt de son adoption car, de part sa conception (*i.e.* application lourde à installer sur chaque ordinateur client et sans système client-serveur permettant de centraliser les données), il ne permettait pas des analyses collaboratives de génomes au travers du Web. Le module GenoLink de Genostar [10], développé en partie par mon collègue Laurent Labarre qui était en thèse CIFRE avec la société Hybrigenics, amenait quant à lui une grande originalité dans la manière d'interroger des connaissances structurées dans un modèle objet par des requêtes représentées sous la forme d'un graphe. Malgré un certain succès commercial auprès de grands groupes privés, la société a fermé en 2017 et le logiciel Iogma n'a plus été maintenu.

Après avoir soutenu ma thèse et face au succès rencontré par la plateforme MicroScope, j'ai eu l'opportunité de pouvoir continuer mes activités de recherche en tant que chercheur CEA statutaire dans le laboratoire LABGeM (sous la direction de Claudine Médigue) de l'UMR 8030 au Genoscope. Concernant la plateforme MicroScope, j'exerce une activité de responsable scientifique avec l'aide de ma collègue Alexandra Calteau (chercheuse CEA) qui coordonne les activités de management par la qualité (référentiels ISO 9001:2015 et NF X50-900:2016) et de formation. Alexandra supervise, également, plusieurs travaux d'ingénieurs et de stagiaires de Master pour l'intégration de nouvelles méthodes d'analyse. Depuis décembre 2017, Mathieu Dubois nous a rejoint en tant qu'ingénieur de recherche CNRS et gère les évolutions technologiques de la plateforme avec l'aide de trois autres ingénieurs permanents (Aurélie Génin-Lajus, Zoé Rouy et David Roche) et d'une technicienne (Stéphanie Fouteau) qui s'occupent en plus du suivi des projets. Je ne citerai pas ici le nom de la vingtaine d'ingénieurs (contractuels pour la plupart) et de stagiaires qui ont grandement participé à l'évolution de la plateforme mais je les en remercie.

MicroScope offre des outils efficaces d'analyses de génomes procaryotes en combinant notamment des méthodes de contextes génomiques comme la détection de synténies conservées et un processus de reconstruction des réseaux métaboliques des organismes étudiés. Les résultats d'analyses sont modélisés et intégrés dans une base de données nommée PkGDB (pour "Prokaryotic Genome DataBase") qui donne un accès rapide et complet à l'information via une interface utilisateur Web nommée MaGe. L'interprétation

humaine des résultats est ainsi facilitée et des hypothèses sur la ou les fonctions des gènes étudiés peuvent être rapidement formulées par un raisonnement plus global sur des processus biologiques comme les voies métaboliques. Dans l'optique de maintenir le système au niveau de l'état de l'art de l'analyse de génomes procaryotes et d'innover en proposant de nouvelles fonctionnalités, plusieurs évolutions importantes ont été réalisées depuis la première publication en 2006 et sont détaillées ci-dessous.

Suite à l'arrivée des nouvelles technologies de séquençage et face au nombre croissant de génomes nouvellement séquencés, nous avons dû mettre en place un système automatisé de gestion des calculs pour augmenter notre productivité en termes de capacité d'analyse de nouveaux génomes. Un gestionnaire de *Workflow* a donc été développé (dans le cadre du projet ANR PFTV MicroScope et du postdoc de Stefan Engelen). Cette innovation, accompagnée de nouvelles fonctionnalités, a donné lieu à la publication d'une deuxième version de la plateforme en 2009 [11].

Durant ces dix dernières années, un objectif particulier a été suivi pour l'intégration de nouveaux outils pour la curation et l'exploration des réseaux métaboliques qui ont été réalisés notamment dans le cadre du projet européen Microme. Nous avons également développé deux modules d'analyse de données expérimentales obtenues par séquençage pour la transcriptomique quantitative (module TAMARA) et l'étude des mutations de souches évoluées (module PALOMA). Ces deux modules ont été réalisés sous la responsabilité de Stéphane Cruveiller (ancien chercheur du LABGeM qui a quitté l'équipe en juin 2018). Une nouvelle version de la plateforme MicroScope a été publiée en 2013 [12] puis en 2017 [13].

Plus récemment, nous avons amélioré l'ergonomie de l'interface Web avec un nouveau sélecteur de génomes et des options additionnelles dans la représentation cartographique. Des outils de classification fonctionnelle, de prédiction de gènes de résistance aux antibiotiques ou impliqués dans la virulence, ainsi que de caractérisation de régions génomiques (*e.g.* clusters pour la biosynthèse de métabolites secondaires, systèmes de sécrétion et de résistance aux phages) ont été intégrés. Nous travaillons également sur l'intégration de la notion de pangénome dans le modèle de données de MicroScope (*cf.* chapitre 2 section I.7). A partir de la définition de groupes de génomes supposés appartenir à une même espèce (les "MicroScope Genome Clusters", MICGCs), nous construisons des pangénomes avec la méthode PPanGGOliN puis prédisons des régions de plasticité génomique correspondant le plus souvent à des îlots génomiques (*cf.* chapitre 2 sections I.3 et I.4). Ces nouvelles fonctionnalités sont décrites dans un article qui vient d'être publié dans l'édition spéciale sur les bases de données du journal Nucleic Acids Research [14].

Finalement, pour faire face aux nombreuses demandes d'analyse de génomes, nous explorons des solutions basées sur les technologies du Cloud pour proposer des services d'analyse MicroScope (projet MicroCloud, dans le cadre de l'infrastructure de l'Institut Français de Bioinformatique). Ces technologies permettraient de déployer des instances MicroScope à la demande pour une analyse rapide de génomes.

La plateforme MicroScope est membre de l'Institut Français de Bioinformatique (IFB) et de France Génomique (FG) qui sont deux infrastructures nationales en biologie, respectivement, pour la bioinformatique et la génomique. Aujourd'hui, MicroScope rassemble plus de 4 700 utilisateurs qui ont un compte personnalisé (35% en France et 65% à l'international). Cette large communauté a réalisé plus de 600 000 annotations expertes (~20 000 en 2019). Depuis 2002, plus de 14 000 génomes ont été analysés et nous avons actuellement un rythme d'intégration de plusieurs centaines de nouveaux génomes par mois. Cette capacité de traitement, combinée à une communauté internationale de microbiologistes réalisant des expertises, fait de MicroScope un des systèmes informatiques les plus performants pour l'analyse et l'exploration des génomes microbiens (plus de 1 000 citations depuis 2006).

II-Analyse de génomes et reconstruction de réseaux métaboliques

Un autre volet de mes activités de recherche s'inscrit dans l'étude du métabolisme des microorganismes sur la base de leur information génétique et de données expérimentales. La plateforme MicroScope a servi de support pour de nombreuses analyses de systèmes biologiques variés. Elles ont été réalisées en collaboration avec plusieurs équipes spécialisées et ont porté, par exemple, sur des études de la symbiose plante-bactérie [15–20] (projets ANR SESAM et BugsInACell), de bactéries impliquées dans la détoxification de l'arsenic [21–23] (projet ANR RARE), de bactéries marines dégradant des polysaccharides d'algues [24] (projet ANR Blue Enzymes) ou encore de bactéries symbiotiques à l'origine de la magnétoréception chez des micro-eucaryotes marins [25].

Pour d'autres projets, une reconstruction de réseaux métaboliques à partir de l'information génomique des organismes a été menée :

- Suite à l'annotation du génome d'*A. baylyi* ADP1 et parallèlement à l'analyse de phénotypes de croissance sur la banque de mutants, un modèle métabolique a été établi pour cette bactérie [26]. Ces travaux ont été réalisés dans le cadre de la thèse de Maxime Durot et ont nécessité plusieurs itérations pour améliorer les prédictions du modèle en comparaison avec les résultats expérimentaux d'essentialité des gènes et de phénotypes de croissance.

- Un re-séquençage et une mise à jour des annotations du génome de *Bacillus subtilis* ont conduit à la publication de deux articles amenant un regard nouveau sur le métabolisme de cette bactérie modèle [27,28] (dans le cadre du projet européen Microme et du postdoc d'Eugenio Belda). Récemment, un effort de curation des annotations de *B. subtilis* à l'aide de la plateforme MicroScope a permis d'enrichir considérablement les connaissances sur cette bactérie [29]. Toujours dans le contexte du projet Microme, les annotations du génome de *Pseudomonas putida* KT2440 ont également été améliorées au regard de ses capacités métaboliques pour une utilisation comme châssis pour la biologie de synthèse [30].

- Une approche métagénomique a été utilisée pour l'étude de l'ancien site minier de *Carnoulès* dont les eaux de drainage sont très polluées en composés toxiques en particulier l'arsenic [21] (projet ANR RARE). Nous avons montré que sept souches bactériennes, dont les génomes ont été reconstruits, dominent l'écosystème. Cinq d'entre elles représentent des bactéries encore non-cultivées. Une analyse statistique de ces données, combinées à des expériences de

protéomiques et de RT-PCR, a permis de construire un modèle intégré des interactions métaboliques. Plusieurs capacités métaboliques, exprimées *in situ*, ont été identifiées comme l'oxydation du fer, du soufre et de l'arsenic qui sont des mécanismes clés de la biominéralisation, et des associations syntrophiques dans le métabolisme de nutriments et de vitamines.

- Les réseaux métaboliques de 29 souches d'*Escherichia coli* ont été reconstruits et comparés pour comprendre le lien entre le pouvoir pathogène et la capacité métabolique de ces bactéries [31] (dans le cadre du projet ANR METACOLI et de la thèse de Gilles Vieira). Ces travaux ont montré que la proportion de réactions communes aux différentes souches (57%) est beaucoup plus élevée que celle des gènes communs (13%), ce qui suggère une diversité plus faible du métabolisme dans cette espèce en comparaison de la diversité des autres fonctions. De plus, la variabilité des fonctions métaboliques entre souches se situe essentiellement dans le catabolisme. Néanmoins, un faible nombre de réactions a été associé à la pathogénie ou au commensalisme. Ce travail a également servi de base à la reconstruction de modèles métaboliques à l'échelle de la cellule.

Ces différents travaux ont montré l'intérêt de l'information génomique dans l'interprétation de systèmes biologiques qui, de plus, peut être combinée à d'autres approches expérimentales à haut débit comme par exemple la protéomique ou la transcriptomique. Néanmoins, les outils informatiques de prédiction de la fonction des gènes et de reconstruction des réseaux métaboliques montrent beaucoup de limites et une étape de curation (*i.e.* expertise humaine des données) est généralement nécessaire [32,33]. Un problème de fond est : comment transférer des fonctions biologiques sur plusieurs dizaines de millions de protéines disponibles dans les banques de séquences alors qu'uniquement quelques dizaines de milliers ont des fonctions expérimentalement démontrées ? Un autre problème réside dans le faible nombre d'espèces étudiées expérimentalement et, donc, l'univers des protéines de fonction connue ne couvre qu'une infime partie de la diversité des fonctions dans le vivant. Un de nos objectifs est donc de capitaliser au maximum la curation de données expérimentales au sein de la plateforme MicroScope tout en maintenant une qualité et une cohérence dans les annotations automatiques produites.

Dans le cadre de la thèse de Jonathan Mercier, un système expert, nommé GROOLS[2], a été développé pour assister les bio-analystes dans la curation des fonctions enzymatiques des protéines [34]. GROOLS évalue la complétude et la consistance des annotations d'un génome à l'aide d'une représentation en graphe des connaissances sur les voies métaboliques. Ce logiciel utilise une logique paracohérente pour diffuser des observations (*i.e.* expectations et prédictions) au travers du graphe et ainsi alerter l'utilisateur sur des annotations incohérentes ou manquantes. En parallèle à ce projet, une

[2] https://github.com/Grools/grools-application

collaboration avec l'équipe UniProt de l'EBI ("The European Bioinformatics Institute") est en cours autour de la conception d'un système à base de règles pour l'annotation fonctionnelle des protéines, nommé UniFIRE[3].

[3] https://gitlab.ebi.ac.uk/uniprot-public/unifire

III-Exploration de nouvelles activités enzymatiques

Une seconde activité autour de la thématique d'étude du métabolisme porte sur la bioanalyse et le développement de méthodes bioinformatiques pour la découverte de nouvelles activités enzymatiques. Ce travail de recherche est mené conjointement avec les laboratoires expérimentaux de notre UMR. Nos compétences s'inscrivent dans la modélisation en base de données d'informations sur le métabolisme, l'analyse de séquences, les méthodes de contextes génomiques et métaboliques, et la modélisation structurale. Notre objectif est de tirer, au maximum, partie de l'information génomique et structurale des protéines pour découvrir de nouvelles fonctions enzymatiques et voies métaboliques, et interpréter ou proposer des expérimentations.

Comme évoqué précédemment, une difficulté majeure dans l'étude du métabolisme des organismes à partir de leur génome est l'assignation de fonctions correctes aux gènes prédits. A cela s'ajoute un problème inverse qui correspond à des activités enzymatiques caractérisées expérimentalement mais dont on ne connaît aucune séquence de protéine catalysant la réaction. Nous avons écrit une revue sur ces activités enzymatiques orphelines ("orphan enzymes") et montré que leur proportion demeure très élevée malgré l'essor des technologies de séquençage [35]. En 2013, plus de 22 % des activités enzymatiques répertoriées n'avaient pas de séquence connue dans aucun organisme et ce pourcentage passait à 50 % si on considérait toutes les réactions répertoriées dans les bases de données métaboliques.

Depuis ces douze dernières années, nous avons mené plusieurs projets démarrant par des approches bioinformatiques et donnant lieu à la validation expérimentale de nouvelles enzymes et activités. La démarche bioinformatique utilisée est principalement basée sur l'étude de contextes génomiques. Le but est d'identifier des sous-ensembles de gènes conservés dans plusieurs organismes. Cette conservation, qui peut être calculée à l'aide de la co-localisation sur le chromosome (opéron ou synténie conservée) ou de profils de présence/absence de gènes (profils phylogénétiques), est un bon indicateur permettant de déduire qu'un groupe de gènes peut participer à un même processus biologique. Différentes méthodes de contexte génomique ont été intégrées dans la plateforme MicroScope [36,37]. Parallèlement au projet *Cloaca maxima* de métagénomique des bassins de la station d'épuration des eaux usées d'Evry, nous avons réalisé plusieurs études sur des voies métaboliques de fermentation qui ont conduit à la découverte de nouvelles enzymes :

- La voie de fermentation de la Lysine est connue depuis les années cinquante mais, jusqu'à notre étude publiée en 2007 [38], trois étapes enzymatiques de cette voie métabolique demeuraient

orphelines de séquences de protéines. Nous avons ainsi identifié, dans les données de séquence de *Cloaca maxima*, trois gènes (*kdd*, *kce*, et *kal*) qui codent, respectivement, une L-erythro-3,5-diaminohexanoate déhydrogenase, une enzyme de clivage du 3-keto-5-aminohexanoate et une 3-aminobutyryl-CoA ammonia lyase. Des études complémentaires de génomique comparée ont montré que 12 bactéries, dont le génome était disponible, possèdent ces gènes et sont donc à même de fermenter la lysine.

- Une étude similaire a été réalisée sur la voie de dégradation oxydative de la L-ornithine en anaérobie qui a été caractérisée plus de 70 ans auparavant. Quatre gènes ont ainsi été identifiés par des méthodes de contexte génomique et les activités enzymatiques des protéines correspondantes ont été validées expérimentalement [39].

- L'analyse du métagénome de *Cloaca maxima* a notamment permis de reconstruire le génome complet d'une bactérie non cultivable de la nouvelle division candidate WWE1 [40]. L'analyse du génome de cette bactérie, nommée Candidatus *Cloacamonas acidaminovorans*, a montré qu'elle utilise une voie alternative de fermentation de la lysine par l'intermédiaire d'une nouvelle activité enzymatique : une acyl-CoA beta-transaminase qui a été caractérisée expérimentalement [41].

Face à ces différents succès, nous avons décidé d'automatiser cette démarche visant à détecter des gènes candidats pour des activités enzymatiques orphelines de séquence. La méthode bioinformatique CanOE (pour "Candidate genes for Orphan Enzymes") a ainsi été développée dans le cadre de la thèse d'Alexander Smith [42]. CanOE combine simultanément la recherche de contextes génomiques et métaboliques conservés dans plusieurs organismes. Ces unités de conservation sont calculées en utilisant un algorithme sur les graphes et sont nommées des métabolons [9]. La méthode CanOE a été appliquée sur plus d'un millier de génomes procaryotes et a permis d'identifier des gènes candidats pour 70 activités enzymatiques orphelines. Nous avons tenté de valider expérimentalement une de ces prédictions dans le cadre de la dégradation en anaérobie de l'allantoïne dans *E. coli* K-12 mais sans succès. Cette voie métabolique n'est probablement plus active dans cette souche : l'opéron correspondant contient un pseudogène (*i.e.* gène *ylbE*). Néanmoins, cet opéron est retrouvé conservé dans d'autres organismes distants dans la phylogénie (*e.g.* des Bacillus et Clostridium) ; des expérimentations nouvelles sur ces bactéries pourraient être envisagées.

Actuellement, plus de 1000 activités enzymatiques sont orphelines de séquence dans les bases de données publiques. Parmi ces activités, CanOE détecte un faible nombre de candidats. Une limitation importante réside dans la difficulté d'ancrer ces activités dans un contexte métabolique

informatif. En effet, la majorité des activités orphelines n'ont pas de voisinage métabolique avec des enzymes connues [35].

Dans le cadre de la thèse de Maria Sorokina, nous avons travaillé sur une nouvelle représentation informatique du métabolisme. Le graphe de réactions est transformé en un graphe de transformations chimiques en regroupant dans un même nœud toutes les réactions réalisant une même transformation. De plus, une probabilité de transition d'une transformation à une autre est calculée à partir de la topologie initiale du graphe de réactions. Cette représentation plus "relâchée" du contexte métabolique a ainsi permis de détecter des modules de transformations chimiques conservés dans le métabolisme [43].

Une seconde facette de cet axe de recherche est l'exploration de la diversité enzymatique de familles de protéines afin de détecter des nouvelles activités enzymatiques impliquées dans de nouvelles voies métaboliques ou ayant un intérêt en biocatalyse. Cette démarche a été initiée à partir des résultats obtenus sur la fermentation de la lysine [38]. La protéine Kce a été identifiée comme catalysant le clivage du 3-keto-5-aminohexanoate. Cette protéine appartient à une famille Pfam qui était de fonction inconnue [44] (Pfam PF05853, DUF849) et dont la grande majorité des organismes, possédant une protéine de cette famille, ne sont pas des fermenteurs de lysine. De plus, un alignement multiple des séquences de cette famille confirme une conservation de résidus clés dans le site actif pour le mécanisme réactionnel. Ces constatations nous ont amené à émettre l'hypothèse que la réaction serait conservée dans la famille mais s'appliquerait *in-vivo* sur d'autres composés chimiques impliqués dans d'autres voies métaboliques que la fermentation de lysine. Ce projet BKACE (pour "β-Keto Acid Cleavage Enzymes") a été publié dans la revue Nature Chemical Biology [45] et a consisté en une analyse intégrée couplant des approches bioinformatiques et expérimentales :

- La famille (725 protéines) a été partitionnée en 32 sous-groupes supposés iso-fonctionnels par des méthodes bioinformatiques combinant la phylogénie, la conservation de synténie et la classification des sites actifs par modélisation structurale [46].

- Un criblage enzymatique a ensuite été conduit sur des représentants de chaque sous-groupe et contre un panel de β-keto acides : 124 protéines exprimées ont été testées pour une activité enzymatique sur 16 composés chimiques différents. Un certain nombre de ces activités a été validé biochimiquement et 14 nouvelles activités ont été mises en évidence.

- Les résultats ont été finalement interprétés à la lumière des structures des protéines et de leur contexte génomique. 7 groupes ont été définis et nous avons proposé des rôles *in-vivo* dans 4 contextes métaboliques différents.

Des stratégies de type CanOE, étendue aux modules de transformations chimiques, et BKACE, combinant les contextes génomiques et l'analyse des sites actifs, devraient gagner en synergie par une généralisation des méthodes et leur application à d'autres familles de protéines (*cf.* chapitre 2 section II.2). La détection de gènes candidats pour des activités enzymatiques orphelines permettra de découvrir de nouvelles familles d'enzymes qui, à leur tour, pourront être explorées dans leur diversité fonctionnelle. Certaines de ces nouvelles fonctions s'intégreront dans des voies métaboliques nouvelles dont des activités seront orphelines de séquences.

Chapitre 2 : Projets de recherche

I-Analyse comparée des pangénomes : de la plasticité des génomes à la diversité des écosystèmes

I.1 Motivation de l'approche

Ces vingt dernières années ont vu l'explosion des projets de séquençage conduisant à un déluge de plusieurs centaines de milliers de génomes disponibles dans les banques de séquences de l'INSDC ("International Nucleotide Sequence Database Collaboration"). La base de données GenBank du NCBI comporte ainsi près de 500 000 génomes complets procaryotes et, étant donné la croissance actuelle, le million de génomes devrait être atteint pour la fin de l'année 2020. Ce rythme effréné pose des questions sur la capacité des grands centres de l'INSDC à continuer de fournir un tel service à la communauté pour des raisons techniques mais également financières. Certaines bases de données ont fait le choix de ne plus être exhaustives. C'est le cas, par exemple, de UniProtKB qui est une ressource collectant les séquences de protéines issues pour la plupart des projets génomes [47]. Depuis avril 2018, les protéomes (*i.e.* l'ensemble des séquences de protéines issues de l'annotation d'un génome) dits redondants ne sont plus intégrés dans cette base[4]. Ce choix purement technique permet ainsi de maintenir le service mais ne fait que repousser le problème d'accumulation de données inhérent à ces ressources qui, pour la plupart, n'ont pas changé fondamentalement de modèle de données depuis leur création dans les années 1980.

Les avancées technologiques dans les équipements de séquençage de l'ADN, notamment en termes de débit pour la technologie Illumina, et une baisse des coûts participent grandement à cette explosion de données. En avril 2014, la société Oxford Nanopore Technologies (ONT) a proposé aux laboratoires de tester une nouvelle technologie de séquençage utilisant des nanopores constitués de protéines transmembranaires modifiées (*i.e.* des porines). Ce type de séquenceur, contrairement aux autres technologies, n'est pas basé sur une étape de polymérisation ni d'étiquetage chimique des

[4] https://www.uniprot.org/help/proteome_redundancy

nucléotides mais est directement capable de lire une molécule d'ADN ou d'ARN : une molécule qui traverse le pore entraîne une variation du courant électrique qui est traduite en séquence de nucléotides. Les avantages de cette technologie sont multiples : (i) faible coût d'équipement et donc accessible à un grand nombre de laboratoires (ii) longueur de lectures importantes (plusieurs dizaines de kilobases contre 300 bases au maximum pour les séquenceurs Illumina) (iii) appareillage portable pour une analyse in situ (iv) acquisition des données en temps continu permettant des analyses en flux. Malgré des avancées significatives, deux contraintes dans l'utilisation de cette technologie subsistent : (i) la quantité d'ADN nécessaire est importante pour obtenir un bon rendement (ii) le taux d'erreur des lectures est de l'ordre de 10%.

Bénéficiant de ces avancées dans les technologies de séquençage, les projets de génomique environnementale sont également en croissance en nombre et en quantité de données générées durant ces dix dernières années. Ils consistent au séquençage d'échantillons d'ADN ou d'ARN prélevés dans différents écosystèmes. Les approches ciblées (*i.e.* séquençage d'un gène marqueur amplifié ou région, encore appelées métabarcoding) pour l'identification et la quantification des espèces présentes sont maintenant très souvent complétées par du séquençage aléatoire dit métagénomique. Les lectures obtenues sont assemblées en contigs pour prédire l'ensemble des gènes présents dans un écosystème. A partir de ces gènes, une assignation taxonomique peut être réalisée mais également une prédiction de la fonction des gènes qui est un point de départ pour identifier les processus biologiques d'importance dans l'écosystème étudié. Ainsi, des catalogues de plusieurs millions de gènes ont été constitués au travers de consortia internationaux notamment pour l'étude du microbiote humain [48] ou des océans [49]. Plus récemment, grâce aux progrès dans l'assemblage et la classification ("binning") des contigs ou lectures, des méthodes permettent de regrouper les contigs en ensembles homogènes en termes de couverture en lectures et de composition en nucléotides [50–52]. Ainsi, chaque ensemble est supposé regrouper des contigs d'un même génome (appelé MAG pour "Metagenome-Assembled Genome"). Ce processus a été systématiquement appliqué en combinant les données de plusieurs études métagénomiques et, au travers de trois analyses indépendantes, un catalogue unifié de 280 000 génomes du microbiote intestinal humain a été constitué [53]. A cela s'ajoute un nombre croissant de génomes d'espèces non cultivées obtenus par des techniques d'amplification de l'ADN de cellules uniques (des SAGs, pour "Single-cell Amplified Genomes") [54]. Ces SAGs et MAGs offrent ainsi de nouvelles perspectives pour la génomique environnementale en ne considérant plus les gènes individuellement mais en travaillant sur des catalogues de génomes pour des analyses quantitatives et fonctionnelles plus fines à l'échelle des espèces voire des souches d'intérêt dans un écosystème. Il est à noter que les assemblages des SAGs et MAGs sont souvent incomplets notamment au niveau de régions de composition atypique en nucléotides et ils comportent

potentiellement des problèmes de contigs chimériques ou de contaminations intra- ou inter-espèces [55]. A ce jour, il n'existe pas encore de processus standardisé de contrôle qualité pour évaluer ces assemblages et ils ne sont que très rarement soumis aux banques de séquences participant ainsi au problème de décentralisation des données en génomique.

Le traitement de cette masse de données impose un changement de paradigme dans la représentation des connaissances et dans les algorithmes utilisés. Les études de génomique comparée sont maintenant basées sur l'analyse de plusieurs centaines, voire milliers de souches d'une même espèce. Les comparaisons de génomes deux à deux sont donc de plus en plus difficilement envisageables et les approches basées sur des génomes de référence sont limitées car, par définition, elles ne permettent pas de capturer la variabilité qui n'est pas présente dans les références choisies. Ainsi, nous travaillons depuis quelques années sur une nouvelle représentation des données génomiques en utilisant le concept de pangénome, celle-ci servant de référence exhaustive qui compresse l'information de milliers de génomes dans une seule structure de donnée tout en conservant l'information de voisinage génomique des gènes. Sur cette structure en graphe, des algorithmes (*e.g.* partitionnement, recherche de chemins ou de composantes connexes) peuvent être développés pour, par exemple, classifier les gènes d'un pangénome, comparer des génomes d'une même espèce ou de de différentes espèces et analyser des données métagénomiques à l'échelle de l'espèce et des souches.

I.2 Introduction à la notion de pangénome

En microbiologie, l'origine du concept de pangénome est généralement attribuée à deux études de 2005 : Medini *et al.* [56] et Tettelin *et al.* [57]. Un pangénome désigne l'union de tous les gènes (ou séquences génomiques) présents dans un groupe de génomes provenant le plus souvent de la même espèce. Dans un pangénome, sont généralement distingués deux sous-ensembles : les gènes cœurs et accessoires, deux concepts introduits avant celui de pangénome notamment par Campbell *et al.* [58]. Les gènes cœurs sont ceux conservés dans toutes les souches et constituent donc le patrimoine génétique d'une espèce qui est maintenu dans les populations par évolution verticale. Les gènes accessoires ne sont retrouvés que dans une sous-partie des souches et sont majoritairement issus d'événements de transferts horizontaux et non de duplications [59]. Bien que considérées auparavant comme rares et ayant peu d'impacts dans l'évolution des procaryotes [60,61], les premières analyses de génomique comparée au début des années 2000 ont montré que les gènes accessoires sont cruciaux pour comprendre la capacité d'adaptation des microorganismes [62]. En effet, ils constituent un répertoire très étendu de gènes qui peuvent conférer divers traits expliquant les différences phénotypiques observées au sein d'une espèce. Pour calculer un pangénome à l'échelle des gènes, une

première étape consiste à regrouper l'ensemble des gènes d'une espèce en familles de gènes homologues. Pour cela plusieurs méthodes existent et sont décrites dans cet article de revue [63]. Ainsi, les tailles du pangénome, du génome cœur et accessoire sont généralement exprimées en nombre de familles de gènes.

Avec l'augmentation du nombre de génomes étudiés par espèce, la dichotomie génome cœur et accessoire a commencé à soulever des questions. En effet, une définition stricte du génome cœur impose la présence des gènes dans tous les génomes comparés ce qui a pour conséquence que sa valeur diminue mécaniquement au fur et à mesure que des génomes sont ajoutés. Il devient alors difficile de comparer des études car les résultats obtenus sont très variables suivant le nombre de génomes considérés. Par exemple, pour l'espèce *E. coli*, une étude a estimé, avec 61 génomes, que le génome cœur possède 993 familles de gènes [64], soit bien moins que les 2 344 familles d'une analyse basée sur 17 génomes [65] ou bien que les 1 976 familles d'une autre basée sur 20 génomes [66]. Ceci est problématique car le génome cœur est censé correspondre aux éléments stables donc à la signature même des espèces et ne devrait pas décroitre aussi fortement en fonction du nombre de génomes considérés. La première raison de cette décroissance a une origine technique. En effet, la qualité de l'assemblage des génomes va fortement influencer celle de la prédiction des gènes : des gènes ne vont pas être prédits dans certains génomes à cause de trous d'assemblage ou d'erreurs dans la séquence nucléotidique comme des insertions ou délétions qui engendrent des décalages de cadre de lecture.

Une autre raison est d'origine biologique. En effet, le concept de génome cœur est à préciser. Si certains gènes sont absolument indispensables, et ce quelles que soient les circonstances, à la survie d'une cellule bactérienne comme les gènes impliqués dans la synthèse des protéines ou la réplication de l'ADN, à l'inverse, d'autres fonctions clés n'ont pas toujours besoin d'être présentes dans la cellule. Ainsi, des gènes du génome cœur peuvent être perdus par certaines souches car ils correspondent à des fonctions qui ne sont plus indispensables dans leur environnement car potentiellement inutiles ou réalisées par d'autres membres de la communauté comme suggéré dans l'hypothèse de la reine noire [67]. C'est pourquoi nous préférons le terme de génome persistant, comme proposé par les auteurs de [68] dans un contexte de biologie de synthèse, pour nommer l'ensemble des gènes conservés dans une large majorité des génomes d'une espèce. Les termes "soft core" [69], "extended core" [70,71] et "stabilome" [72] sont également utilisés pour désigner le génome persistant. Pour prendre en compte des absences possibles dans l'estimation du génome persistant, un seuil de fréquence de présence des familles de gènes dans les génomes est fixée généralement entre 90 et 99%.

Les études de génomique comparée se contentent généralement de diviser le pangénome en génome cœur et accessoire. Cette dichotomie ne prend pas en compte le fait que les fréquences de présence des familles de gènes dans un pangénome suivent une distribution ayant une forme de lettre U asymétrique (cf. Figure 1). Ceci reflète que les gènes n'ont pas les mêmes taux de gain et de perte dans les populations et, comme proposé par Koonin *et al.* [73] et modélisé par Collins *et al.* [74], un pangénome peut être divisé en trois classes :

- le génome *persistent* (ou persistant), pour les familles de gènes présentes dans presque tous les génomes (taux faibles de gain et de perte)
- le génome *shell* (ou coquille), pour les familles présentes à des fréquences intermédiaires dans l'espèce (taux modérés de gain et de perte)
- le génome *cloud* (ou nuage), pour les familles présentes à des fréquences faibles dans l'espèce (taux élevés de gain et de perte).

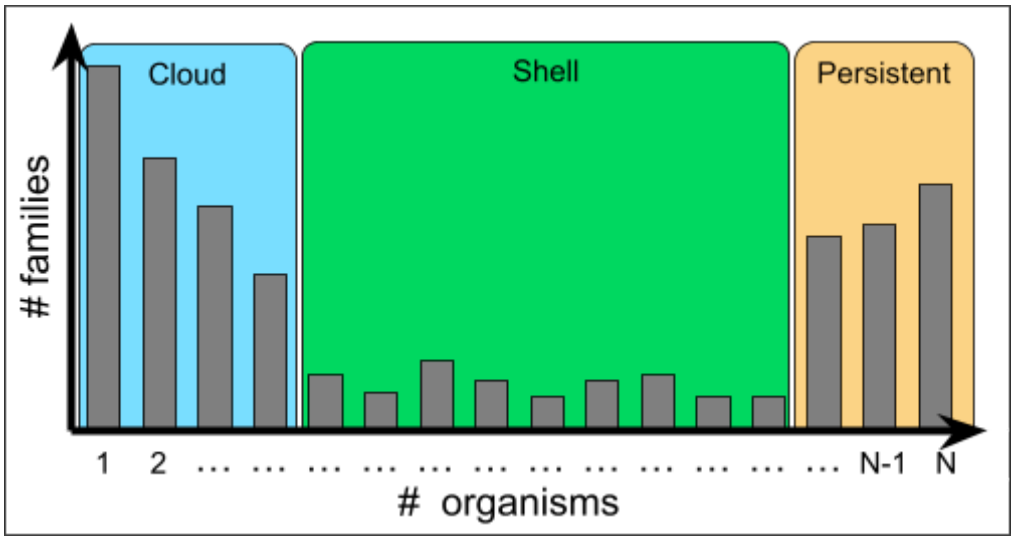

Figure 1 : Illustration de la distribution du nombre de familles de gènes présentes dans 1 à N génomes.

Ainsi, en plus du génome *persistent* qui constitue le patrimoine génétique maintenu dans l'espèce, le génome *shell* est un élément clé pour comprendre la dynamique des génomes. En effet, le *shell* reflète la manière dont l'espèce s'adapte à différents environnements via l'acquisition et le maintien de gènes dans la population, alors que les gènes du *cloud* vont correspondre à des propriétés émergentes dans l'espèce comme, par exemple, l'acquisition récente de gènes de résistance aux antibiotiques.

En aparté, le terme pangénome est également utilisé pour désigner l'union des séquences nucléotidiques des génomes d'une espèce. Diverses structures de données sont ainsi proposées pour regrouper d'une manière non redondante toutes les variations nucléotidiques et structurales présentes

dans un ensemble de génomes de référence au sein d'un graphe appelé "variation graph", "genome graph" ou "pangenome graph" [75]. Des méthodes permettent ensuite d'aligner des lectures de séquençage sur ces graphes d'une manière plus sensible que sur les génomes individuellement et ainsi améliorer la détection de variants nucléotidiques. D'autres méthodes utilisent ces graphes pour faire de la génomique d'association (*i.e.* associer la présence/absence de gènes ou mutations à des traits phénotypiques) [76].

I.3 La méthode PPanGGOLiN : graphe de pangénome partitionné

Dans le cadre de son Doctorat, Guillaume Gautreau a conçu et développé une méthode nommée PPanGGOLiN[5] (pour "Partitioned PanGenome Graph Of Linked Neighbors") avec l'aide d'Adelme Bazin (un doctorant de l'équipe travaillant sur des approches de génomique comparée à l'échelle des pangénomes). PPanGGOLiN représente un pangénome sous la forme d'un graphe où les nœuds correspondent à des familles de gènes homologues et les arêtes à des relations de contiguïté génomique de ces familles dans les génomes (*i.e.* deux familles sont connectées dans le graphe si elles partagent une paire de gènes voisins dans au moins un génome). Les arêtes sont étiquetées avec les identifiants de génomes présentant le voisinage. Cette structure de données permet ainsi de compacter l'information de milliers de génomes en utilisant des familles d'homologues tout en gardant l'information de contexte génomique de ces gènes. En effet, l'information de co-localisation des gènes est d'une importance primordiale dans l'étude de l'évolution des génomes procaryotes pour plusieurs raisons : (i) environ deux tiers des gènes sont organisés en opérons correspondant à des unités polycistroniques où les gènes sont co-transcrits puis traduits dans la même échelle de temps et d'espace dans la cellule [77] (ii) des gènes co-localisés sur le génome sont souvent liés d'un point de vue fonctionnel et participent à un même processus biologique [78] (iii) les gènes du génome *persistent* ont tendance à partager des organisations conservées dans les génomes [79] (iv) les gènes issus de transferts horizontaux (i.e. les gènes du génome *shell* et *cloud*) s'insèrent dans des régions préférentielles (*i.e.* points chauds d'insertions) [80].

A partir de ce graphe de pangénome, un apprentissage statistique est réalisé pour classifier les familles de gènes suivant la trichotomie génome *persistent*, *shell* et *cloud*. Le choix de la méthode de classification résulte d'une collaboration avec deux statisticiens : Catherine Matias (Sorbonne Université) et Christophe Ambroise (Université d'Evry-Val-d'Essonne). La méthode retenue se nomme NEM (pour "Neighboring Expectation-Maximization algorithm") et a été initialement proposée dans le cadre de l'analyse d'images [81,82]. Elle combine un modèle basé sur un mélange de

[5] https://github.com/labgem/PPanGGOLiN

K distributions de Bernoulli multivariées permettant de partitionner les familles de gènes suivant leur vecteur de présence/absence avec un critère de régularité spatiale dans le graphe de pangénome basé sur un champ aléatoire de Markov caché (Markov Random Field : MRF). Le MRF favorise que deux familles voisines dans le graphe de pangénome soient classifiées dans la même partie pour les raisons biologiques évoquées précédemment. Après maximisation de la vraisemblance du modèle de mélange de Bernoulli multivarié, contraint à être spatialement régulier, on obtient alors le graphe de pangénome partitionné. Enfin, le nombre de parties à détecter (K) peut être fourni par l'utilisateur ou déterminé par l'algorithme.

PPanGGOLiN a été appliqué d'une manière systématique sur l'intégralité des espèces procaryotes représentées par plus de 15 génomes dans la banque GenBank du NCBI. Ainsi, 439 espèces (136 287 génomes) ont été analysées. Nous avons montré que l'estimation du génome *persistent* réalisée par PPanGGOLiN est bien plus stable et indépendante de l'échantillonnage des génomes que la méthode classique (*i.e.* appelée "soft-core") basée sur un seuil de fréquence de présence des familles de gènes (*i.e.* ≥95% de présence). De plus, cette étude ouvre de nouvelles pistes pour comprendre l'importance du génome *shell* dans la dynamique des génomes au sein d'une espèce et dans l'adaptation à divers environnements. Il est à noter d'une manière un peu surprenante que la proportion de gènes *shell* par génome n'est pas corrélée avec la taille des génomes. Les espèces possédant le plus de gènes *shell* ont généralement un *shell* très hétérogène qui est classifié par PPanGGOLiN en plusieurs parties et reflète une potentielle structuration de l'espèce en sous-populations se spécialisant dans des environnements différents. Pour terminer, nous avons évalué l'utilisation de PPanGGOLiN pour l'analyse d'espèces du microbiote humain représentées par des MAGs [83]. Ces génomes sont potentiellement très incomplets car issus de l'assemblage de données métagénomiques. L'estimation du génome *persistent* des MAGs semble correcte en comparaison avec des données de génomes complets d'isolats. De plus, PPanGGOLiN a détecté des familles de gènes *shell* spécifiques aux MAGs qui pourraient être importantes pour l'adaptation des espèces dans le microbiome.

Ces travaux font l'objet d'un article qui est en cours de révision dans la revue PLOS Computational Biology et dont la pré-publication est incluse dans ce mémoire.

PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph

Guillaume Gautreau[1], Adelme Bazin[1], Mathieu Gachet[1], Rémi Planel[1], Laura Burlot[1], Mathieu Dubois[1], Amandine Perrin[2,5], Claudine Médigue[1], Alexandra Calteau[1], Stéphane Cruveiller[1], Catherine Matias[3], Christophe Ambroise[4], Eduardo PC Rocha[2], David Vallenet[1]

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

[2] Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris, France

[3] Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Université de Paris, Centre National de la Recherche Scientifique, Paris, France

[4] Laboratoire de Mathématiques et Modélisation d'Evry, UMR CNRS 8071, Université d'Évry Val d'Essonne, Évry, France

[5] Sorbonne Université, Collège doctoral, Paris, France

Cette pré-publication a été déposée sur le site d'archives bioRxiv et est disponible à cette adresse avec les données additionnelles : https://doi.org/10.1101/836239

PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph

Guillaume Gautreau[1], Adelme Bazin[1], Mathieu Gachet[1], Rémi Planel[1¤a], Laura Burlot[1], Mathieu Dubois[1], Amandine Perrin[2,5], Claudine Médigue[1], Alexandra Calteau[1], Stéphane Cruveiller[1¤b], Catherine Matias[3], Christophe Ambroise[4], Eduardo PC Rocha[2], David Vallenet[1*]

1 LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS, Evry, France
2 Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris, France
3 Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Université de Paris, Centre National de la Recherche Scientifique, Paris, France
4 Laboratoire de Mathématiques et Modélisation d'Evry, UMR CNRS 8071, Université d'Evry Val d'Essonne, Evry, France
5 Sorbonne Université, Collège doctoral, Paris, France

¤a Current Address: Hub de Bioinformatique et Biostatistique - Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, Paris, France ¤b Current Address: PathoQuest SAS, BioPark – bâtiment B, 11 rue Watt, 75013 Paris, France
* vallenet@genoscope.cns.fr

Abstract

The use of comparative genomics for functional, evolutionary, and epidemiological studies requires methods to classify gene families in terms of occurrence in a given species. These methods usually lack multivariate statistical models to infer the partitions and the optimal number of classes and don't account for genome organization. We introduce a graph structure to model pangenomes in which nodes represent gene families and edges represent genomic neighborhood. Our method, named PPanGGOLiN, partitions nodes using an Expectation-Maximization algorithm based on multivariate Bernoulli Mixture Model coupled with a Markov Random Field. This approach takes into account the topology of the graph and the presence/absence of genes in pangenomes to classify gene families into persistent, cloud, and one or several shell partitions. By analyzing the partitioned pangenome graphs of isolate genomes from 439 species and metagenome-assembled genomes from 78 species, we demonstrate that our method is effective in estimating the persistent genome. Interestingly, it shows that the shell genome is a key element to understand genome dynamics, presumably because it reflects how genes present at intermediate frequencies drive adaptation of species, and its proportion in genomes is independent of genome size. The graph-based approach proposed by PPanGGOLiN is useful to depict the overall genomic diversity of thousands of strains in a compact structure and provides an effective basis for very large scale comparative genomics. The software is freely available at `https://github.com/labgem/PPanGGOLiN`.

Author summary

Microorganisms have the greatest biodiversity and evolutionary history on earth. At the genomic level, it is reflected by a highly variable gene content even among organisms from the same species which explains the ability of microbes to be pathogenic or to grow in specific environments. We developed a new method called PPanGGOLiN which accurately represent the genomic diversity of a species (i.e. its pangenome) using a compact graph structure. Based on this pangenome graph, we classify genes by a statistical method according to their occurrence in the genomes. This method allowed us to build pangenomes even for uncultivated species at an unprecedented scale. We applied our method on all available genomes in databanks in order to depict the overall diversity of hundreds of species. Overall, our work enables microbiologists to explore and visualize pangenomes alike a subway map.

Introduction

The analyses of the gene repertoire diversity of species - their pangenome - have many applications in functional, evolutionary, and epidemiological studies [1, 2]. The core genome is defined as the set of genes shared by all the genomes of a taxonomic unit (generally a species) whereas the accessory (or variable) genome contains genes that are only present in some genomes. The latter is crucial to understand bacterial adaptation as it contains a large repertoire of genes that may confer distinct traits and explain many of the phenotypic differences across species. Most of these genes are acquired by horizontal gene transfer (HGT) [3]. This usual dichotomy between core and accessory genomes does not consider the diverse ranges of gene frequencies in a pangenome. The main problem in using a strict definition of the core genome is that its size decreases as more genomes are added to the analysis [4] due to gene loss events and technical artifacts (i.e. sequencing, assembly or annotation issues). As a consequence, it was proposed in the field of synthetic biology to focus on persistent genes, i.e. those conserved in a large majority of genomes [5]. The persistent genome is also called the soft core [6], the extended core [7, 8] or the stabilome [9]. These definitions advocate for the use of a threshold frequency of a gene family within a species above which it is considered as *de facto* core gene. Persistent gene families are usually defined as those present in a range comprised between 90% [10] and 99% [11] of the strains in the species. This approach addresses some problems of the original definition of core genome but requires the setting of an appropriate threshold. The gene frequency distribution in pangenomes is extensively documented [7, 8, 12–16]. Due to the variation in the rates of gene loss and gain of genes, the gene frequencies tend to show an asymmetric U-shaped distribution regardless of the phylogenetic level and the clade considered (with the exception of few species having non-homogeneous distributions as described in [17]). Thereby, as proposed by [12] and formally modeled by [14], the pangenome can be split into 3 classes: (1) persistent genome, for the gene families present in almost all genomes; (2) shell genome, for gene families present at intermediate frequencies in the species; (3) cloud genome, for gene families present at low frequency in the species.

The study of pangenomes in microbiology now relies on the comparison of hundreds to thousands of genomes of a single species. The analysis of this massive amount of data raises computational and algorithmic challenges that can be tackled because genomes within a species have many homologous genes and it is possible to design new compact ways of representing and manipulating this information. As suggested by [18], a consensus representation of multiple genomes would provide a better analytical framework than using individual reference genomes. Among others, this proposition has led to a paradigm shift from the usual linear representation of reference genomes to a

representation as variation graphs (also named "genome graphs" or "pangenome graphs") bringing together all the different known variations as multiple alternative paths. Methods [19–21] have been developed aiming at factorizing pangenomes at the genome sequence-level to capture all the nucleotide variations in a graph that enables variant calling and improves the sensitivity of the read mapping (summarized in [22]).

The method presented here, named PPanGGOLiN (Partitioned PanGenome Graph Of Linked Neighbors), introduces a new representation of the gene repertoire variation as a graph, where each node represents a family of homologous genes and each edge indicates a relation of genetic contiguity. PPanGGOLiN fills the gap between the standard pangenomic approach (that uses a set of independent and isolated gene families) and sequence-level pangenome graph (as reviewed in [23]). The interest of a gene-level graph compared to a sequence graph is that it provides a much more compact structure in clades where gene gains and losses are the major drivers of adaptation. This comes at the cost of disregarding polymorphism in genes and ignoring variation in intergenic regions and introns. However, the genomes of prokaryotes have very small intergenic regions and are almost devoid of introns justifying a focus on the variation of gene repertoires [12], which can be complemented by analysis of intergenic and intragenic polymorphism. PPanGGOLiN uses a new statistical model to classify gene families into persistent, cloud, and one or several shell partitions. To the best of our knowledge three statistical methods are available to partition a pangenome. Two of them use probabilistic models that partition dichotomously the pangenome only into core and accessory components [26, 27]. Conversely, the method proposed and implemented by Snipen *et al.* [24, 25] (micropan R package) classifies a pangenome in K partitions using a Binomial Mixture Model relying on gene family frequencies. Unlike these three methods, PPanGGOLiN is not based on frequencies but combines both the patterns of occurrence of gene families and the pangenome graph topology to perform the classification. In the following sections we present an overview of the method, an illustration of a pangenome graph and then the partitioning of a large set of prokaryotic species from GenBank. We evaluate the relevance of the persistent genome computed by PPanGGOLiN in comparison to the classical soft core genome. Next, we illustrate the importance of the shell structure and dynamics in the study of the evolution of microbial genomes. Finally, we compare GenBank results to the ones obtained with Metagenome-Assembled Genomes (MAGs) to validate the use of PPanGGOLiN for metagenomic applications.

Results and discussion

Overview of the PPanGGOLiN method

PPanGGOLiN builds pangenomes for large sets of prokaryotic genomes (i.e. several thousands) through a graphical model and a statistical method to classify gene families into three classes: persistent, cloud, and one or several shell partitions. It uses as input a set of annotated genomes with their coding regions classified in homologous gene families. As depicted in Fig 1, PPanGGOLiN integrates information on protein-coding genes and their genomic neighborhood to build a graph where each node is a gene family and each edge is a relation of genetic contiguity (two families are linked in the graph if they contain genes that are neighbors in the genomes). Thanks to this graphical model, the structure of the pangenome is resilient to fragmented assemblies: an assembly gap in one genome can be offset by information from other genomes, thus maintaining the link in the graph. To partition this graph, we established a statistical model taking into consideration that persistent genes share conserved genomic organizations along genomes (i.e. synteny conservation) [28] and that horizontally

transferred genes (i.e. shell and cloud genes) tend to insert preferentially in a few chromosomal regions (hotspots) [29]. Thereby, PPanGGOLiN favors two gene families that are consistent neighbors in the graph to be more likely classified in the same partition. This is achieved by a hidden Markov Random Field (MRF) whose network is given by the pangenome graph. In parallel, the pangenome is also represented as a binary Presence/Absence (P/A) matrix where the rows correspond to gene families and the columns to genomes. Values are 1 for the presence of at least one member of the gene family and 0 otherwise. This P/A matrix is modeled by a multivariate Bernoulli Mixture Model (BMM). Its parameters are estimated via an Expectation-Maximization (EM) algorithm taking into account the constraints imposed by the MRF. Each gene family is then associated to its closest partition according to the BMM. This results in a partitioned pangenome graph made of nodes that are classified as either persistent, shell or cloud. The strength of the MRF constraints increases according to a parameter called β (if $\beta = 0$, the effect of the MRF is disabled and the partitioning only relies on the P/A matrix) and it depends on the weight of the edges of the pangenome graph which represents the number of gene pairs sharing the neighborhood. Another originality of our method is that, even if the number of partitions (K) is estimated to be equal to 3 (persistent, shell, cloud) in most cases (see 'Analyses of the most represented species in databanks' section), more partitions can be used if the pangenome matrix contains several contrasted patterns of P/A. These additional partitions are considered to belong to the shell genome and reflect a heterogeneous structure of the shell (see 'Shell structure and dynamics' section).

Illustration of a Partitioned Pangenome Graph depicting the *Acinetobacter baumannii* species

We computed the pangenome of 3 117 *Acinetobacter baumannii* genomes from GenBank using PPanGGOLiN. For the persistent, shell and cloud genomes, we obtained 3 084, 1 529 and 64 833 gene families, respectively. If we compare our results with those of Chan *et al.* study [18], the size of the persistent genome predicted by PPanGGOLiN is included in their soft core estimation ranging from 2 833 (95% of presence) to 3 126 (75% of presence) gene families using 249 *A. baumannii* genomes. On the partitioned pangenome graph built with PPanGGOLiN (Fig 2), the gene families classified as persistent (orange nodes) correspond to the conserved paths that are interrupted by many islands composed of shell (green nodes) and cloud genomes (blue nodes). These islands appear to be frequently inserted in hotspots of the persistent genome thus pinpointing regions of genome plasticity. The average node degree within the same partition is 2.80 for the persistent genome while the shell genome has a higher average degree (3.95, P=5.0e-6 with bilateral unpaired 2-sample Student's t test) and the cloud a lower one (1.97, P=3.3e-40 with the same test). The shell genome is the most diversified in terms of network topology with many interconnections between families reflecting a mosaic composition of regions from different HGT events [29]. The major part of the cloud has a shell-like graph topology with a large connected component containing 60% of the nodes. In addition, the cloud also contains isolated components that are nearly linear (3 606 components having on average 4.25 nodes) and singletons (10 575 nodes), presumably because it includes very recently acquired genetic material. Finally, large families of mobile genes, mostly transposable elements, can be easily detected because they constitute hubs (i.e. highly connected nodes) in the graph. They vary rapidly their genetic neighborhoods and can be found in multiple loci.

As an example of the more detailed analysis that can be done using the graph, a zoom on a region containing the genes required for the synthesis of capsular polysaccharides is highlighted in Fig 2. *A. baumannii* strains are involved in numerous

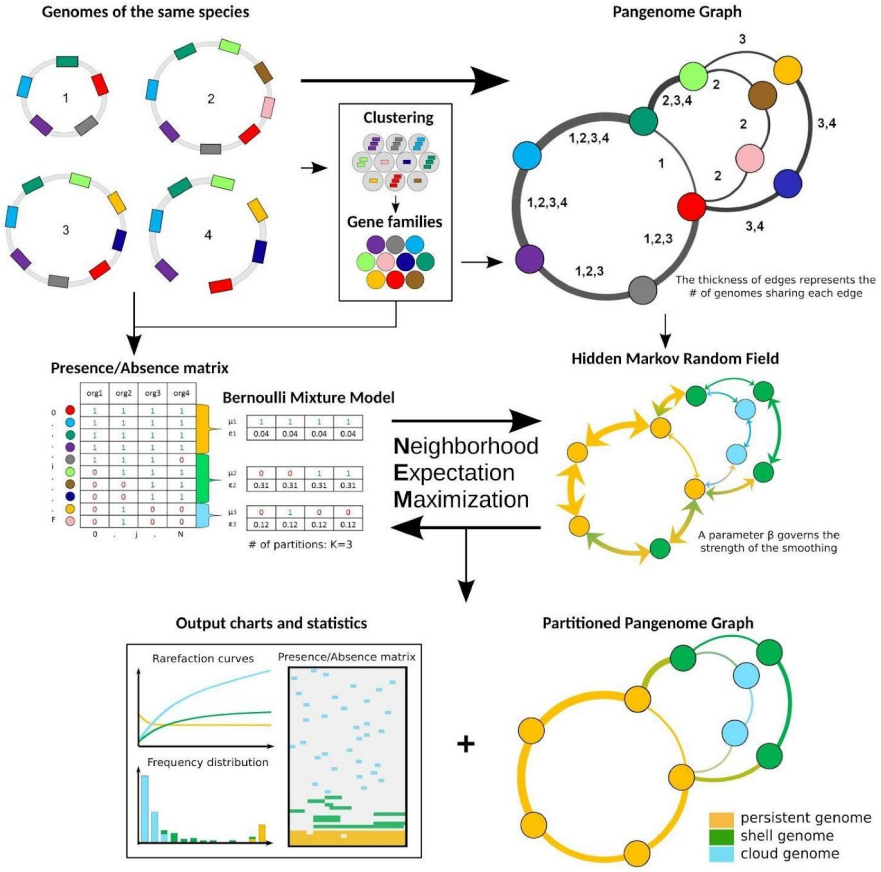

Fig 1. Flowchart of PPanGGOLiN on a toy example of 4 genomes. The method requires annotated genomes of the same species with their genes clustered into homologous gene families. Annotations and gene families can be predicted by PPanGGOLiN or directly provided by the user. Based on these inputs, a pangenome graph is built by merging homologous genes and their genomic links. Nodes represent gene families and edges represent genomic neighborhood. The edges are labeled by identifiers of genomes sharing the same gene neighborhood. In parallel, gene families are encoded as a presence/absence matrix that indicates for each family whether or not it is present in the genomes. The pangenome is then divided into K partitions ($K = 3$ in this example) by estimating the best partitioning parameters through an Expectation-Maximization algorithm. The method involves the maximization of the likelihood of a multivariate Bernoulli Mixture Model taking into account the constraint of a Markov Random Field (MRF). The MRF network is given by the pangenome graph and it favors two neighbors to be more likely classified in the same partition. At the end of this iterative process, PPanGGOLiN returns a partitioned pangenome graph where persistent, shell and cloud partitions are overlaid on the neighborhood graph. In addition, many tables, charts and statistics are provided by the software. The number of partitions (K) can either be provided by the user or determined by the algorithm.

nosocomial infections and their capsule plays key roles in the overall fitness and pathogenicity. Indeed, it protects the bacteria against environmental stresses, host immune responses and can confer resistance to some antimicrobial compounds [30]. Over one hundred distinct capsule types and their corresponding genomic organization have been reported in *A. baumannii* [31]. A zoom on this region of the graph shows a wide variety of combinations of genes for the synthesis of capsular polysaccharides. Based on the *A. baumannii* 3 117 genomes available in GenBank, we detected 229 different paths, sharing many common portions, but only a few are conserved in the species (only 24 paths are covered by more than 10 genomes). Among them, two alternative shell paths seem to be particularly conserved (from the *gnaA* to the *weeH* genes in the figure 3 of [31]). Based on the nomenclature of [31], one (colored in khaki green in the Fig 2) corresponds to the serovar called PSgc12, contains 14 gene families of the shell genome and is fully conserved in 581 genomes. The other (colored in fluo green in the Fig 2) corresponds to the serovar PSgc9 (equivalent to PSgc7), contains 11 gene families of the shell genome and is fully conserved in 408 genomes. This analysis illustrates how the partitioned pangenome graph of PPanGGOLiN can be useful to study the plasticity of genomic regions. Thanks to its compact structure in which genes are grouped into families while preserving their genomic neighborhood information, it summarizes the diversity of thousands of genomes in a single picture and allows effective exploration of the different paths among regions or genes of interest.

Analyses of the most represented species in databanks

We used PPanGGOLiN to analyze all prokaryotic species of GenBank for which at least 15 genomes were available. This is the minimal number of genomes we recommend to ensure a relevant partitioning. The quality of the genomes was evaluated before their integration in the graph to avoid especially taxonomic assignation errors and contamination that can have a major impact on the analysis of pangenomes (see Methods). This resulted in a dataset of 439 species pangenomes, whose metrics are available in S10 File. We focused our analysis on the 88 species containing at least 100 genomes (Fig 3). This data was used for in-depth analysis of persistent and shell genomes (see the two next sections). Proteobacteria, Firmicutes and Actinobacteria are the most represented phyla in this dataset and comprise a variety of species, genome sizes and environments. In contrast, Spirochaetes, Bacteroidetes and Chlamydiae phyla are represented by only one or two species (*Leptospira interrogans*, *Bacteroides fragilis*, *Flavobacterium psychrophilum* and *Chlamydia trachomatis*). For each species, we computed the median and interquartile range of persistent, shell and cloud families in the genomes. As expected, we observed a large variation in the range of these values: from pathogens with reduced genomes such as *Bordetella pertussis* or *C. trachomatis* which contain only a small fraction of variable gene families (less than ≈5% of shell and cloud genomes) to commensal or environmental bacteria such as *Bifidobacterium longum* and *Burkholderia cenocepacia* whose shell represents more than ≈35% of the genome. Furthermore, for a few species the number of estimated partitions (K) is greater than 3 (11 out of 88 species), especially for those with a higher fraction of shell genome. Hence, our method provides a statistical justification for the use of three partitions as a default in pangenome analyses, while indicating that species with large shell content might be best modeled using more partitions (see 'Shell structure and dynamics' section).

Estimation of the persistent genome in comparison to the soft core approach

To demonstrate the added value of PPanGGOLiN, we compared our statistical method to a classical approach where persistent genes are those present in at least 95% of the

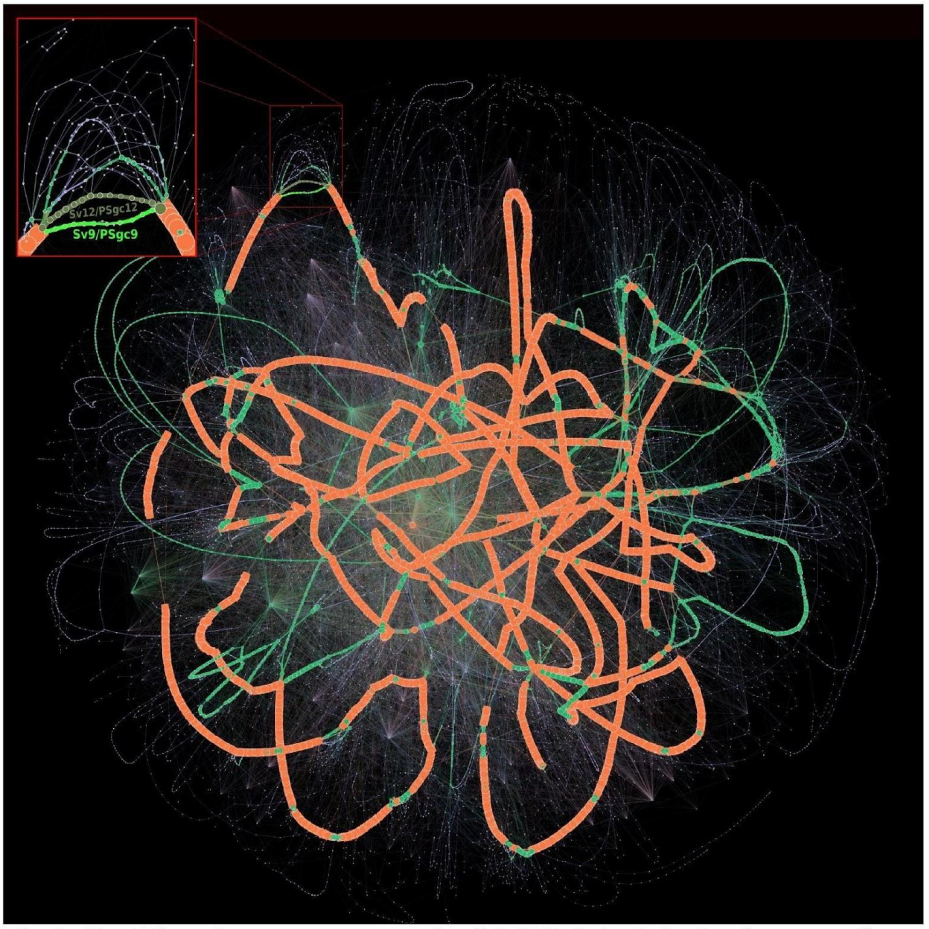

Fig 2. Partitioned pangenome graph of 3 117 *Acinetobacter baumannii* genomes. This partitioned pangenome graph of PPanGGOLiN displays the overall genomic diversity of 3 117 *Acinetobacter baumannii* strains from GenBank. Edges correspond to genomic colocalization and nodes correspond to gene families. The thickness of the edges is proportional to the number of genomes sharing that link. The size of the nodes is proportional to the total number of genes in each family. The edges between persistent, shell and cloud nodes are colored in orange, green and blue, respectively. Nodes are colored in the same way. The edges between gene families belonging to different partitions are shown in mixed colors. For visualization purposes, gene families with less than 20 genes are not shown on this figure although they comprise 84.68% of the nodes (families mostly composed of a single gene). The frame in the upper left corner shows a zoom on a branching region where multiple alternative shell and cloud paths are present in the species. This region is involved in the synthesis of the major polysaccharide antigen of *A. baumannii*. The two most frequent paths (Sv12/PSgc12 and Sv9/PSgc9) are highlighted in khaki and fluo green. The Gephi software (`https://gephi.org`) [32] with the ForceAtlas2 algorithm [33] was used to compute the graph layout with the following parameters: Scaling=8000, Stronger Gravity=True, Gravity=4.0, Edge Weight influence=1.3.

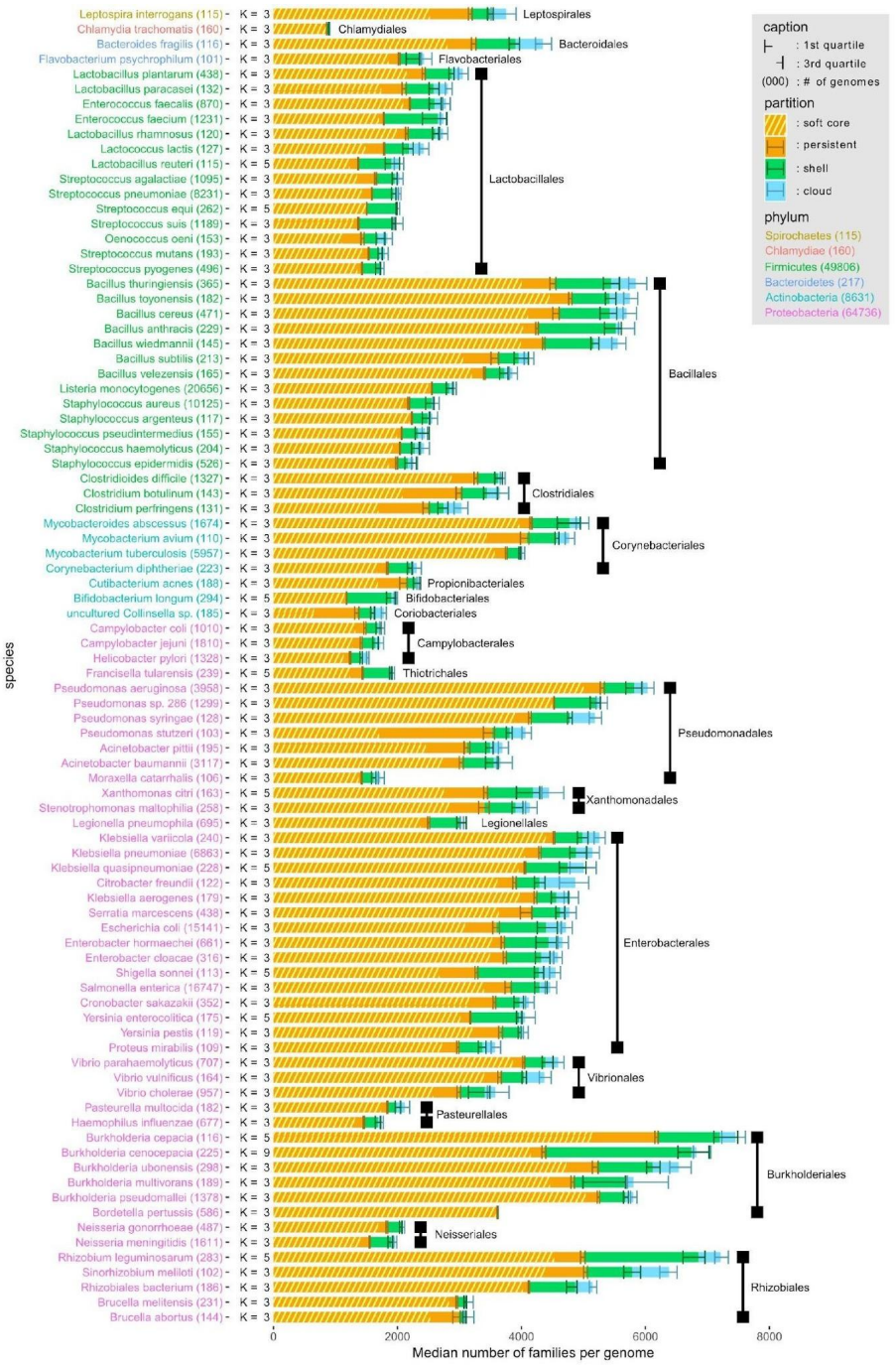

Fig 3. Distribution of PPanGGOLiN partitions in the genomes of the most represented species in GenBank. Each horizontal bar shows the median number of gene families per genome among the different PPanGGOLiN partitions (persistent, shell and cloud) in the 88 most represented species in GenBank (having at least 100 genomes). The error bars represent the interquartile ranges. Hatched areas on the persistent genome bars show the median number of gene families for the soft core ($\geqslant$95% of presence). The species names are colored according to their phylum and sorted by taxonomic order and then by decreasing cumulative bar size. Next to the species names, the number of genomes is indicated in brackets and the number of partitions (K) that was automatically determined by PPanGGOLiN is also shown.

genomes (generally called the soft core approach). Indeed, this threshold is very often used in pangenomic studies probably because it is the default parameter in Roary [35] which is to date the most cited software to build bacterial pangenomes. In the 88 studied species, the number of persistent gene families is greater than or equal to the soft core with an average of 11% (SD=9%) of additional families (see Fig 3 and S10 File). Furthermore, persistent gene families include those of the soft core with the exception of very few gene families (12 families in total for all studied species). The gene family frequencies in each of the 88 pangenomes are available in S1 Fig. For four species, *Pseudomonas stutzeri*, *Clostridium perfringens*, *Clostridium botulinum* and *Colinsella sp.*, the size of the soft core genome is unexpectedly small and represents less than 55% of the genomes whereas it is above 75% for the PPanGGOLiN persistent. For the first three species, this could be due to sampling effects and species heterogeneity. For the last one (*Colinsella sp.*), this could be explained by the fact that the species is made of incomplete genomes from metagenomes (i.e. MAGs) that were submitted as complete genomes in GenBank.

For an in-depth comparison of these approaches, we performed multiple resamplings of the genome dataset for each species in order to measure the variability of the pangenome metrics and the impact of genome sampling according to an increasing number of genomes considered in the analyses (hereafter called rarefaction curves) (see Methods and S2 Fig as an example for *Lactobacillus plantarum*). These rarefaction curves indicate whether the number of families tends to stabilize, increase or decrease. To this end, the curves were fit with the Heaps' law where γ represents the growth tendency [34] (hereafter called γ-tendency). The persistent component of a pangenome is supposed to stabilize after the inclusion of a certain number of genomes, which means it has a γ-tendency close to 0. In addition, interquartile range (IQR) areas along the rarefaction curves were computed to estimate the variability of the predictions in relation to the sampling. Small IQR areas mean that the predictions are stable and resilient to sampling. Using these metrics, the PPanGGOLiN predictions of the persistent genome were evaluated in comparison to the soft core approach.

We observed that the γ-tendency of the PPanGGOLiN persistent is closer to 0 than that of the soft core approach (mean of absolute γ-tendency=9.1e-3 versus 2.5e-2, P=1.5e-9 with one-sided paired 2-sample Student's t-test) with a lower standard deviation error too (mean=5.3e-04 versus 2.1e-03, P=9.5e-11 with one-sided paired 2-sample Student's t-test) (see Fig 4 and S10 File). A major problem of the soft core approach is that the γ-tendency is high for many species (32 species have a γ-tendency above 0.025), suggesting that the size of the persistent genome is not stabilized and tends to be underestimated. Besides, the IQR area of the PPanGGOLiN prediction is far below the one of the soft core genome (mean=4906.6 versus 11645.9, P=8.9e-07 with unilateral paired 2-sample Student's t test). It can be partially explained because the threshold used in the soft core method induces a 'stair-step effect' along the rarefaction curves depending on the number of genomes sampled. This is illustrated on S2 Fig showing a step every 20 genomes (i.e. corresponding to $20 = \frac{100}{100-95}$ where 95% is the threshold of presence used) on the soft core curve of *L. plantarum*. We found a total of 20 species having atypical values of γ-tendency (absolute value above 0.05) and/or IQR area (above 15 000) for the soft core and only 2 species for the persistent genome of PPanGGOLiN, which are *Bacillus anthracis* and *Burkholderia cenocepacia*. For *B. cenocepacia*, it could be explained by the high heterogeneity of its shell (see next section), which is made of several partitions and complicates its distinction from the persistent genome during the process of partitioning. For *Bacillus anthracis*, the source of variability to define the persistent genome is a result of an incorrect taxonomic assignation in GenBank of about 17% of the genomes that are, according to the Genome Taxonomy DataBase (GTDB) [36], actually *B. cereus* or *B. thuringiensis*. This

issue was not detected by our taxonomy control procedure because these species are at the boundary of the conspecific genomic distance threshold used (see Methods). Some of persistent gene families of *bona fide B. anthracis* may therefore shift between persistent or shell partitions depending on the resampling. Excluding these misclassified genomes, we predicted a larger persistent genome than the one of the initial full set of genomes (about a thousand gene families more) with a γ-tendency much closer to 0 (-0.017 versus a γ-tendency of 0.036 for the soft core genome) and a lower IQR area (8367.0 vs 32167.1). Altogether, these results suggest that our approach provides a more robust partitioning of gene families in the persistent genome than the use of arbitrary thresholds. Indeed, the statistical method behind PPanGGOLiN uses directly the information of the gene family P/A whereas the soft core is based only on frequency values. PPanGGOLiN can then classify families with similar frequencies in different partitions by distinguishing them according to their pattern of P/A in the matrix and their genomic neighborhood. The main drawback of using family frequency to partition pangenomes is that even if it was possible to determine the best threshold for each species it would still not take into account that some persistent gene families may have atypically low frequency. This may be due to high gene losses in the population or technical reasons like belonging to a genomic region that is difficult to assemble (i.e. genes that are missing or fragmented in draft genome assemblies).

Shell structure and dynamics

Two types of pangenome evolution dynamics are generally distinguished: open pangenomes and closed ones [1, 2, 34]. From rarefaction curves, the dynamics of pangenomes can be assessed using the γ-tendency of a Heaps' law (see Methods) fitting. A low γ-tendency means a rather closed pangenome whereas a higher γ-tendency means a rather open pangenome. A closed pangenome rigorously means a stabilized pangenome and we found no species obeying this strict criterion (that is to say $\gamma = 0$). This suggests that instead of using binary classifications for pangenomes, it is more useful to quantify the degree of openness of pangenomes given the flux of horizontal gene transfer and gene loss [7]. We computed rarefaction curves for the 88 studied species and determined the γ-tendency for different pangenome components (see S10 File and S3 Fig). The distribution of γ values of the PPanGGOLiN shell genome shows a greater amplitude of values than the other components of the pangenome such as the whole pangenome or the accessory component. This indicates that the main differences in terms of genome dynamics between species seem to reside in the shell genome.

As expected, we found a positive correlation (Spearman's ρ=0.46, P=8.2e-06) between the total number of shell gene families in a species and the γ-tendency of the shell (S4 Fig). This means that species with high γ-tendency do accumulate genes that are maintained and exchanged in the population at relatively low frequencies, suggesting they may be locally adaptive. More surprisingly, although one could expect that larger genomes have a larger fraction of variable gene repertoires, the fraction of shell and cloud genes per genome does not correlate with the genome size (Spearman's ρ=0.007, P=0.95, Fig 5). The results remain qualitatively similar when analyzing the shell or the cloud separately (see S5 Fig and S6 Fig). During this analysis, we noticed that, among host-associated bacteria with relatively small genomes (between ≈2000 and ≈3000 genes), three species (*Bifidobacterium longum*, *Enterococcus faecium* and *Streptococcus suis*) have a high fraction of shell genes ($> 28\%$) but low shell γ-tendency. Two of them (*B. longum* and *E. faecium*) are found in the gut of mammals and the third (*S. suis*) in the upper respiratory tract of pigs. They differ from other host-associated species in our dataset that are mainly human pathogens (e.g. bacteria of the genus *Corynebacterium*, *Neisseria*, *Streptococcus*, *Staphylococcus*) and have a low fraction of shell genes ($< 20\%$). It is possible that these three species have specialized in their ecological niches while

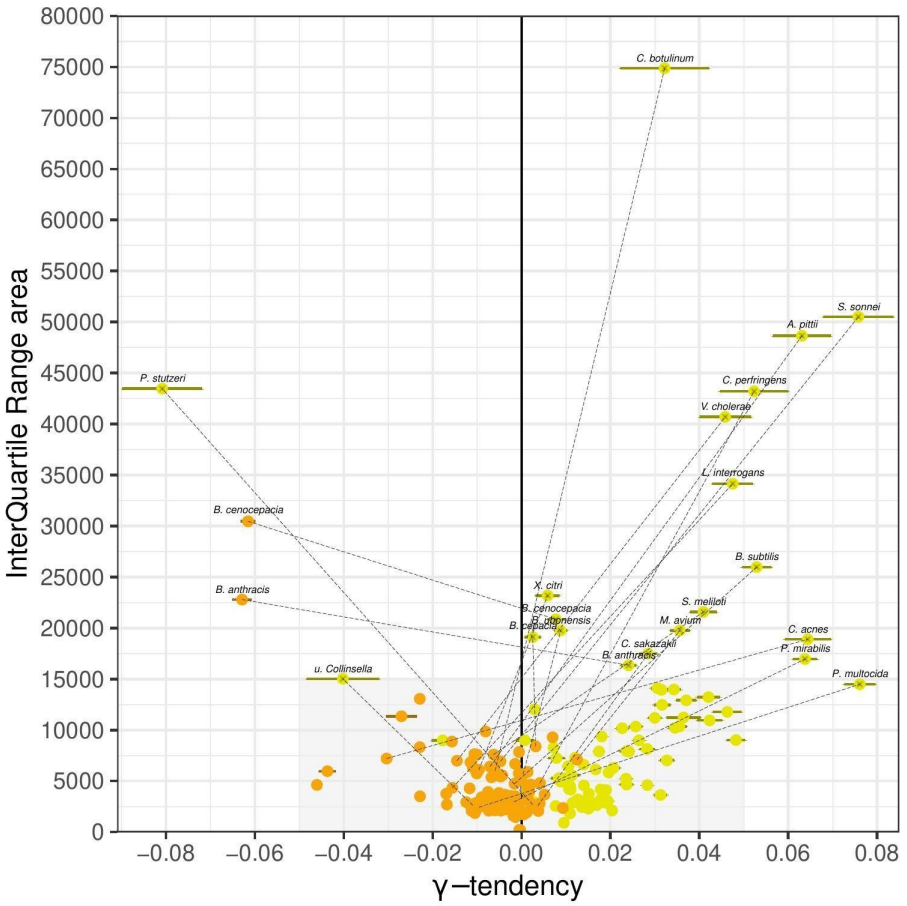

Fig 4. γ-tendencies and IQR areas of the persistent and the soft core rarefaction curves. Each of the 88 most abundant species in GenBank are represented by two points: orange points correspond to the PPanGGOLiN persistent values and yellow points to the ones of the soft core ($\geqslant$95% of presence). A dashed line connects the 2 points if either the soft core or the persistent values are not in the range of the grey area ($-0.05 \leqslant \gamma \leqslant 0.05$ and $0 \leqslant IQR_{\text{area}} \leqslant 15000$). The colored horizontal bars show the standard errors of the fitting of rarefaction curves via the Heaps' law.

maintaining a large and stable pool of shell genes for their adaptation to environmental stress. Further analysis would be required to confirm this hypothesis.

We then investigated the importance of the phylogeny of the species on the patterns of P/A of the shell gene families (shell structure). To this end, Spearman's rank correlations were computed between a Jaccard distance matrix generated on the basis of patterns of P/A of the shell gene families and a genomic distance obtained by Mash pairwise comparisons between genomes [37]. Mash distances were shown to be a good estimate of evolutionary distances for closely related genomes [38]. This correlation was examined in relation to the fraction of gene families that are part of the shell genome for each species (Fig 6). We observed that species with a high fraction of shell ($> 20\%$ of their genome) have a shell structure that is mainly explained by the species phylogeny (i.e. shell P/A are highly correlated with genomic distances, Spearman's $\rho > 0.75$). In addition, PPanGGOLiN predicts a number of partitions (K) for these species often greater than 3. Hence, their shell is more heterogeneous between subclades and becomes structured in several partitions whereas for species with a single shell partition the shell is less structured, possibly indicating many gene exchanges between strains from different lineages. Among the nine species with a large shell genome (excluding *B. anthracis* due to taxonomic assignation errors), only two of them (*Shigella sonnei* and *Lactobacillus reuteri*) showed a relatively low correlation of their shell structure with the phylogeny (Fig 6). For *S. sonnei*, this could be explained by a high number of gene losses in the shell of this species that result from convergent gene loss mediated by insertion sequences (preprint: [39]). For *L. reuteri*, these bacteria colonize the gastrointestinal tract of a wide variety of vertebrate species and have diversified into distinct phylogenetic clades that reflect the host where the strains were isolated, but not their geographical provenance [40]. As illustrated in S7 Fig, the shell of *L. reuteri* shows patterns of P/A that are only partially explained by the species phylogeny. Indeed, we observed clusters of families present across strains from distinct lineages that could contain factors for adaptation to the same host. In contrast, the shell structure of *B. longum* strongly depends on phylogenetic distances showing a clear delineation of adult and infant strains that have specialized into two subspecies (see S8 Fig).

We would like to stress the importance of the shell in the study of the evolutionary dynamics of bacteria. The shell content reflects the adaptive capacities of species through the acquisition of new genes that are maintained in the population. We found that the proportion of shell genes does not increase with the genome size. Instead, the shell accounts for a large fraction of the genomes of species when it is structured in several partitions. We can assume that those species are made of non-homogeneous subclades harboring specific shell genes which contribute to the specialization of the latter. Finally, it could be of interest to associate phenotypes to patterns of shell families that co-occur in different lineages independently of the phylogeny.

Analysis of Metagenome-Assembled Genomes in comparison with isolate genomes

The graph approach should make our tool robust to gaps in genome data, making it a useful tool to analyze pangenomes obtained from MAGs. To test this hypothesis, we built the pangenomes of the Species-level Genome Bins (SGBs, clusters of MAGs that span a 5% genetic diversity and are assumed to belong to the same species) from the recent paper of Pasolli *et al.* [41]. This study agglomerated and consistently built 4 930 SGBs (154 723 MAGs) from 13 studies focussed on the composition of the human microbiome. We skipped the quality control step (already performed by the authors), and computed the pangenomes following the procedure we used for the GenBank species. The only parameter which differs is the K value which is set to 3 as the

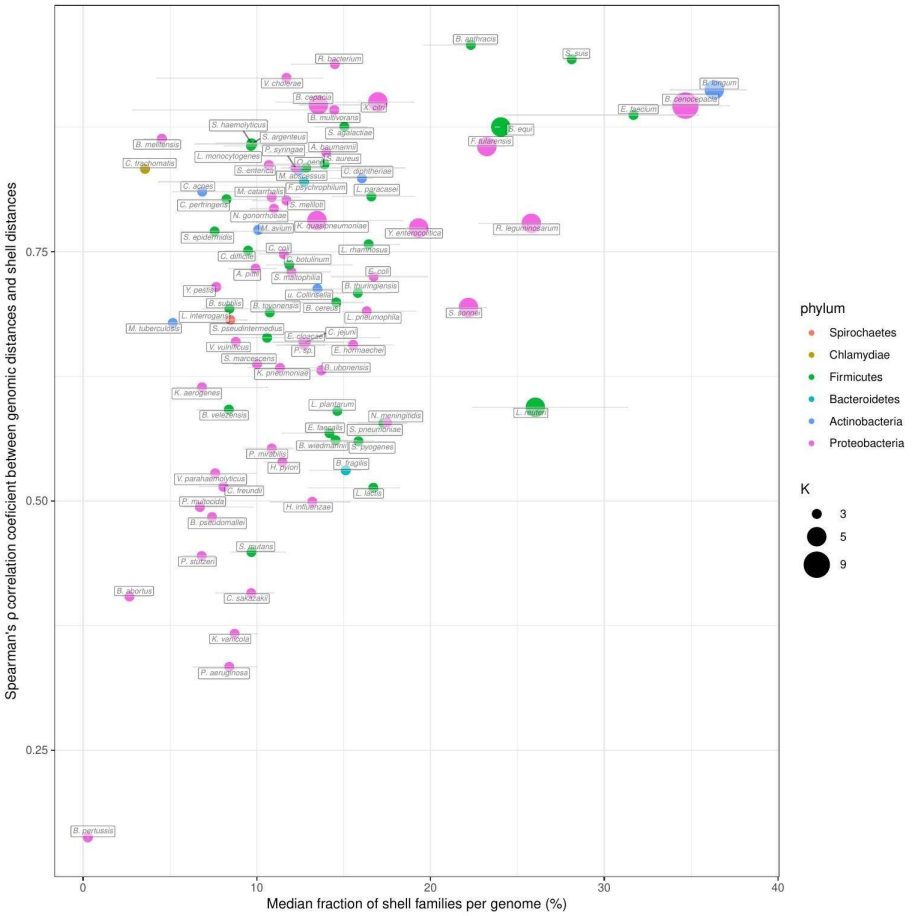

Fig 6. Spearman's ρ correlation coefficients between the shell genome presence/absence patterns and the MASH genomic distances compared with the shell fraction per genome. The results for the 88 most abundant species in GenBank are represented. The error bars show the interquartile ranges of the shell fraction. The points are colored by phylum and their size corresponds to the number of partitions (K) used.

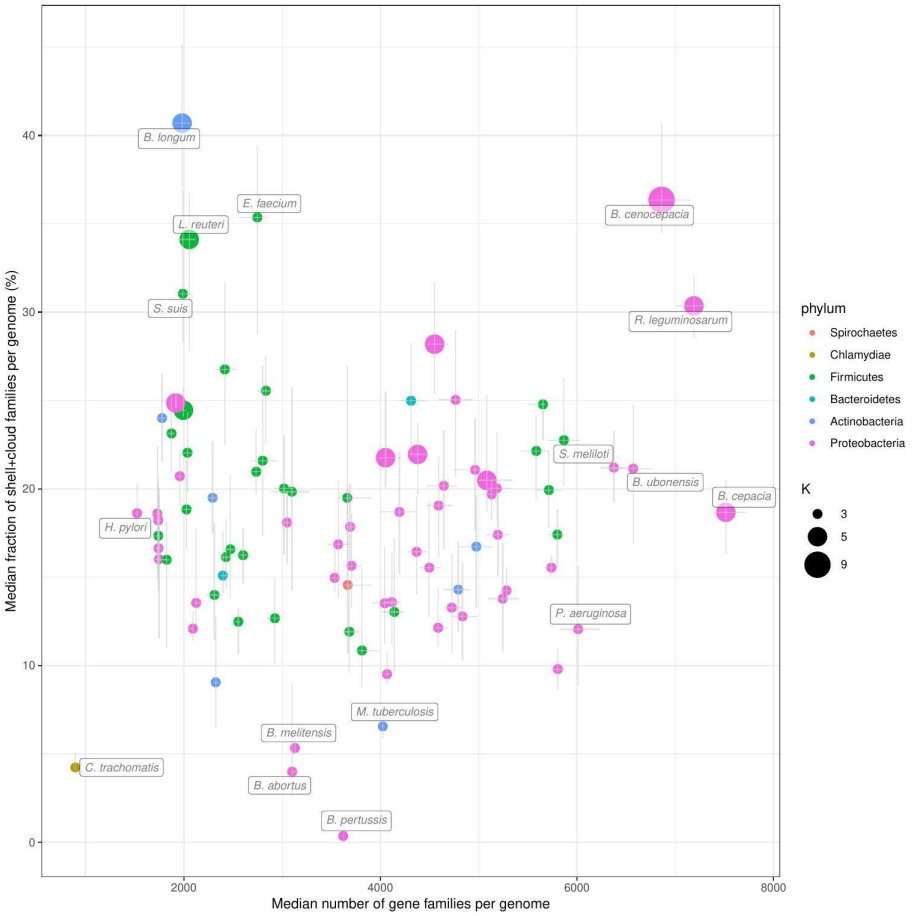

Fig 5. Fraction of the variable (shell + cloud) families per genome compared to the number of gene families. The results for the 88 most abundant species in GenBank are represented. The error bars show the interquartile ranges of the two variables. The points are colored by phylum and their size corresponds to the number of partitions (K) used.

detection of several shell partitions is difficult for MAGs because of their incompleteness. To make the comparison with GenBank species, SGBs were grouped according to their estimated species taxonomy (provided by the supplementary table S4 of [41]). In this table, we noticed potential errors in the taxonomic assignation of two species (*Blautia obeum* and *Chlamydia trachomatis* corresponding to SGBs 4844 and 6877, respectively) and thus excluded them from the analysis. Keeping the same constraint as previously, only species with at least 15 genomes in both MAGs and GenBank were used for the comparison. A total of just 78 species (corresponding to 151 SGBs) could be analyzed as a lot of microbiome species are laborious to cultivate and thus less represented in databanks (see S11 File). Then, we compared the MAG pangenome partitions predicted by PPanGGOLiN with those obtained with GenBank genomes. To perform this, we aligned MAG and GenBank families for each species and computed the percentage of common families for each partition (see Methods and S11 File).

We observed that the size of the estimated persistent genome of MAGs is similar to the one of GenBank genomes for most species (Fig 7). In 55 out of the 78 species, the absolute fold change of persistent size is less than 1.2 and ≈90% (SD=5%) of its content is common between MAGs and GenBank genomes. The 23 other species with more important differences showed smaller persistent genomes with only 60% (SD=15%) of the persistent genome of GenBank being found in MAGs. For these species, the PPanGGOLiN method missed a fraction of the persistent genome due to the incompleteness of MAGs. Indeed in such cases, the missing gene families are mostly classified in the shell of the MAGs which contains 32% (SD=11%) of the GenBank persistent families. Nevertheless, 89% (SD=9%) of the MAG persistent families match the GenBank ones, meaning that PPanGGOLiN correctly assigned persistent families for MAGs even if the persistent genome of these 23 species is incomplete. However, two species, *Bifidobacterium longum* and *Faecalibacterium prausnitzii*, have less than 75% of their MAG persistent families in common with GenBank ones. For *B. longum*, this could be explained by the fact that the MAGs were obtained mostly from human adult samples while this species in databanks are from a broader host range (infants and pigs). It means that the MAG persistent might contain additional genes related to host-specificity. As a matter of fact, 412 gene families from the MAG persistent (25% of the total MAG persistent) are found in the GenBank shell which supports our hypothesis. For *F. prausnitzii*, the differences might be explained by a poor estimation of the persistent using GenBank data due to the low number of considered genomes (17 genomes versus 4232 MAGs). As expected, the soft core (based on the usual threshold of 95% presence) is unrealistically low in the MAG species with only ≈98 gene families on average and only 4 species out of 78 having more than 500 families classified in the soft core (see S11 File). Hence, the soft core approach is not well adapted to the analysis of MAGs. Furthermore, using lower thresholds of presence is not adequate because defining a unique threshold for all the families misses the heterogeneity of gene family presence in MAGs.

To explore the diversity within the pangenome of each species, we compared the shell of GenBank genomes and MAGs for the 55 ones with similar persistent genomes. Interestingly, we observed for all the 55 species only a partial overlap between the MAGs and GenBank shells (see S9 Fig). Indeed, as the MAGs are obtained only from a specific environment (i.e. the human microbiome), the diversity of GenBank is not fully captured by MAGs. It is especially the case for most of the Firmicutes and Proteobacteria. Conversely, most of the MAGs of Bacteroidetes phylum cover more than half of GenBank diversity while containing a large fraction of shell genes that are lacking in the shell of isolate genomes (i.e. less than 45% of the families are represented in the shell of GenBank). As already reported by Pasolli *et al.* [41], this confirms that the MAGs considerably improve the estimate of the genetic diversity of Bacteroidetes

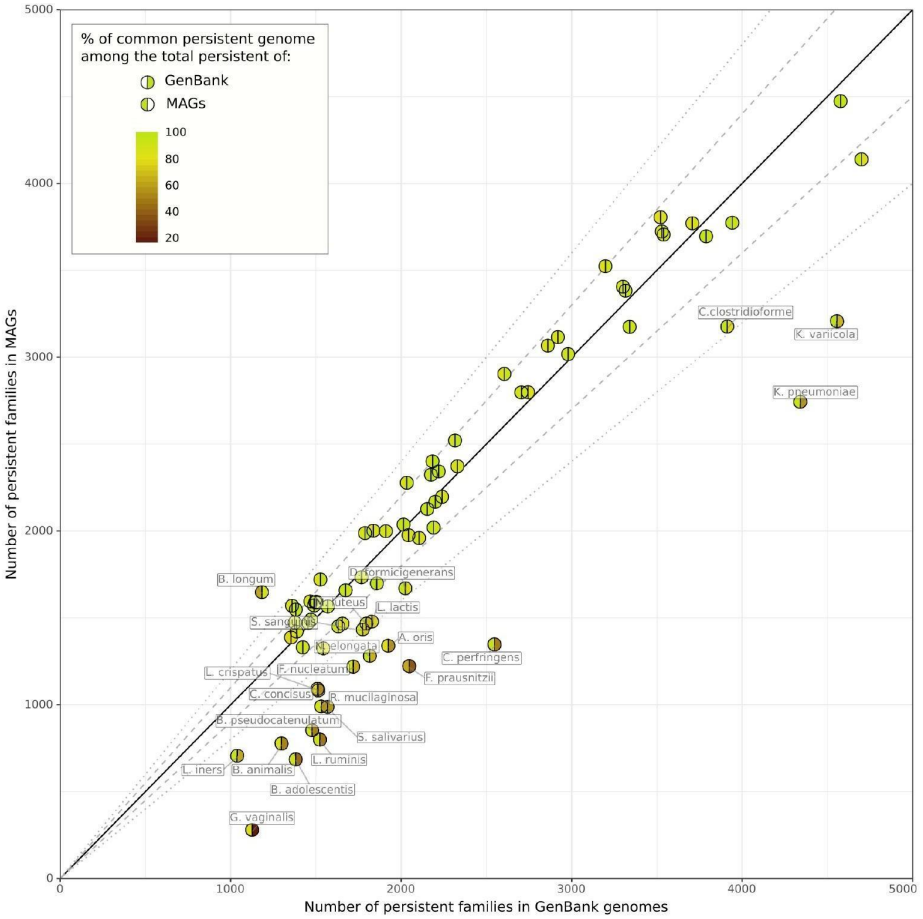

Fig 7. Illustration of the persistent genome overlaps between GenBank genomes and MAGs. Results for 78 species are represented. The colors of the hemispheres provide the percentage of common persistent gene families among the total persistent of MAGs (left hemisphere) or GenBank genomes (right hemisphere). The solid, dashed and dotted lines indicate the identity, a fold change of 1.1 and a fold change of 1.2 between the persistent genome sizes.

which are key players in the gut microbiome.

In summary, we have shown that PPanGGOLiN is able to provide an estimation of the persistent genome even using MAGs, which may miss significant numbers of genes and be contaminated by fragments from other genomes. This is especially the case for the accessory genome because its assembly coverage and nucleotide composition generally differ from those of the persistent genome making the binning of these regions more difficult. Nevertheless, PPanGGOLiN is able to find shell gene families in MAGs bringing new genes that may be important for species adaptation in the microbiome. Hence, it enables further analyses, even for uncultured species lacking reference genomes, such as the reconstruction of the core metabolism from the persistent genome to predict culture media or the study of the landscape of horizontally transferred genes within species.

Conclusion

We have presented here the PPanGGOLiN method that enables the partitioning of pangenomes in persistent, shell and cloud genomes using a gene family graph approach. This compact structure is useful to depict the overall genomic diversity of thousands of strains highlighting variable paths made of shell and cloud genes within the persistent backbone. The statistical model behind PPanGGOLiN makes a more robust estimation of the persistent genome in comparison to classical approaches based on gene family frequencies in isolate genomes and also in MAGs. The definition of shell partitions based on statistical criteria allowed us to understand genome dynamics within species. We observed different patterns of shell with regard to phylogeny that may suggest different adaptive paths for the diversification of the species. It should be stressed that genome sampling is one of the main limitations of pangenome studies and can therefore influence PPanGGOLiN partitioning especially for the shell genome. An improvement in the method could be to normalize the data to remove sampling bias. But as suggested by Brockhurst *et al.* [42], this issue should first be examined from a biological perspective by collecting and analyzing genomes from ecologically coherent microbial populations or ecotypes.

Future applications of PPanGGOLiN could include the prediction of genomics islands within the shell and cloud genomes. A first version of this application (Bazin *et al.*, in preparation) is already integrated in the MicroScope genome analysis platform [43]. Next, it would be interesting to determine the architecture of these variable regions by predicting conserved gene modules using information on the occurrence of families and their genomic neighborhood in the pangenome graph. Regarding metagenomics, pangenome graphs of PPanGGOLiN could be used as a reference (i.e. instead of individual genomes) for species quantification by mapping short or long reads on the graph to compute the coverage of the persistent genome. Indeed, each gene families of the partitioned pangenome graph could embed a variation graph as an alignment template [19]. Moreover, coverage variation in the shell or cloud genomes could allow the detection of strain-specific paths in the graph that are signatures of distinctive traits within microbiotes.

To conclude, the graph-based approach proposed by PPanGGOLiN provides an effective basis for very large scale comparative genomics and we hope that drawing genomes on rails like a subway map may help biologists navigate the great diversity of microbial life.

Materials and Methods

To explain the partitioning of pangenomes, we first need to describe the method based on the P/A matrix only (BinEM) and then the method built upon it that uses the pangenome graph to improve the partitioning (NEM).

Modeling the P/A matrix via a Multivariate Bernoulli Mixture Model

PPanGGOLiN aims to classify patterns of P/A of gene families into K partitions ($K \in \mathbb{N}; K \geqslant 3$). Input data consists of a binary matrix X in which a x_{ij} entry is 1 if family i is present in a genome j and 0 otherwise (Fig 1) where $1 \leqslant i \leqslant F$ in each of the F gene families and $1 \leqslant j \leqslant N$ in each of the N genomes. A first approach for partitioning the data relies on a multivariate Bernoulli Mixture Model (BMM) estimated through the Expectation-Maximization (EM) algorithm [44] (named the BinEM method). The number of partitions K may be greater than 3 (persistent, shell and cloud) due to the possible presence of antagonist P/A patterns among the different strains of a species. Therefore, two of the partitions will correspond to the persistent and cloud genome and a number of $K-2$ partitions will correspond to the shell genome. The value of K can be either provided by the user or determined automatically (see next section).

In the BMM, the matrix comprises data vectors $X_i = (x_{ij})_{1 \leqslant j \leqslant N}$ describing P/A of families, which are assumed to be independent and identically distributed with a mixture distribution given by:

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

where $\pi = (\pi_1, \ldots, \pi_k, \ldots, \pi_K)$ denotes the mixing proportions satisfying $\pi_k \in [0, 1]$; $(\sum_{k=1}^{K} \pi_k) = 1$ and where π_k is the unknown proportion of gene families belonging to the k^{th} partition. Moreover, $\mu_k = (\mu_{kj})_{1 \leqslant j \leqslant N} \in \{0; 1\}^N$ are the centroid vectors of P/A of the k^{th} partition representing the most probable binary states and $\epsilon_k = (\epsilon_{kj})_{1 \leqslant j \leqslant N} \in [0, \frac{1}{2}]^N$ are the unknown vectors of dispersion around μ_k. The default values of the dispersion vector ϵ_k associated to each centroid vector μ_k are constrained to be identical for all the ϵ_{kj} of a specific k partition (for all the genomes of a specific partition) in order to avoid over-fitting but it is possible to release this constraint. The parameters of this model, as well as corresponding partitions, are estimated by the EM algorithm. To speed up the computation of the EM algorithm, a heuristic is used to initialize the BMM parameters in order to converge to a relevant partitioning using fewer EM-steps. This heuristic consists in setting π_k with equiprobable proportions equal to $1/K$ while the ϵ_{kj} and μ_{kj} parameters are initialized triangularly.

Given $s = 1/\lceil K/2 \rceil$, the triangular initialization consists of:

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

An interesting consequence of this initialization is that the persistent genome will be the first partition ($k = 1$) while the cloud genome will correspond to the last partition ($k = K$). This particular initialization solves the classical label switching problem in our context.

Partitioning of the P/A matrix

To perform the partitioning of the P/A matrix, each gene family i must be allocated to a single partition. The variables $\{Z_i\}_{1 \leqslant i \leqslant F}$ with a state space $\{1, \ldots, K\}$ indicate the partition to which each gene family i belongs. Therefore, once the NEM parameters are optimized, the method automatically assigns the gene families to their most probable partition z_i according to the model if their estimated posterior probability is above 0.5. If no partition can be assigned in this way, then the gene family is assigned to the shell (partition with intermediate frequency).

Selection of the optimal number of partitions (K)

To determine the optimal K, named $\hat{K}$, the algorithm runs multiple partitionings with increasing values of K. After a few steps of the EM algorithm (10 steps by default), the Integrated Completed Likelihood (ICL) [45] is computed for each K. The ICL corresponds to the Bayesian Information Criterion (BIC) [46] penalized by the estimated mean entropy and is calculated as:

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

and

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

where $\log \mathbb{P}_K(X \mid \theta)$ is the data log-likelihood under a multivariate BMM with K partitions and $\theta = (\{\pi_k\}_{1 \leqslant k \leqslant K}, \{\mu_{kj}\}_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant N}, \{\epsilon_{kj}\}_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant N})$. This log-likelihood can be calculated as follows:

$$\log \mathbb{P}_K(X \mid \theta) = \sum_{i=1}^{F} \log \left(\sum_{k=1}^{K} \pi_q \prod_{j=1}^{N} \epsilon_{kj}^{|x_{ij} - \mu_{kj}|} (1 - \epsilon_{kj})^{1 - |x_{ij} - \mu_{kj}|} \right)$$

Moreover, $\hat{\theta}$ is the maximum likelihood estimator (approximated through the BinEM algorithm) and $dim(K)$ is the dimension of the parameter space for this model. Here, $dim(K) = K(N + 2)$ if the dispersion vector ϵ_k associated to each centroid vector μ_k is constrained to be identical for all the ϵ_{kj} of a specific k partition and $dim(K) = K(2N + 1)$ if the dispersion vector ϵ_k is free. Relying on this criterion, the best number of partitions is selected as $\hat{K} = \arg\min_K((1 - \delta_{ICL})ICL(K))$ where δ_{ICL} is a sufficiently small margin to avoid choosing a too high K value that would provide no significant gain compared to a lower value of K (by default $\delta_{ICL} = 0.05 \times (max(ICL) - min(ICL))$).

Generation of the pangenome graph

PPanGGOLiN uses a graph-based representation to store and visualize pangenomes. In this graph, the nodes correspond to gene families and the edges to genetic contiguity (i.e. genes that are direct neighbors in a genome). Two nodes are connected if the corresponding gene families contain at least one pair of genes that are adjacent in a genome. Edges are labeled with the corresponding genome identifiers and weighted by the proportion of genomes sharing that link. This process results in a pangenome graph (see Fig 2 as an example).

Formally, a pangenome graph $G = (V, E)$ is a graph having a set of vertices $V = \{(v_i)_{(1 \leqslant i \leqslant F)}\}$ where F is the number of gene families in the pangenome associated with a set of edges $E = \{e_{i \sim i'}\} = \{(v_i, v_{i'})\}, v_i \in V$, $v_{i'} \in V$ where the couple of vertices

$(v_i, v_{i'})$ are gene families having their genes $(v_{i,j}, v_{i',j})$ adjacent on the genome j and where the function $countNeighboringGenes(v_i, v_{i'})$ counts the adjacency occurrences in the N genomes. Each edge $\{e_{i\sim i'}\}$ has a weight $w_{i\sim i'}$ where $w_{i\sim i'} = \frac{1}{N}\sum_{j=1}^{N} countNeighboringGenes(v_{i,j}, v_{i',j})$.

Partitioning via Neighboring Expectation-Maximization

From the graph previously described, the neighborhood information of the gene families is used to improve the partitioning results. Indeed, the BinEM approach described above is extended by combining the P/A matrix X with the pangenome graph G. This relies on a hidden Markov Random Field (MRF) model whose graph structure is given by G. In this model, each node belongs to some unobserved (hidden) partitions which are distributed among gene families according to a MRF which favors two neighbors to be more likely classified in the same partition. Conditional on this hidden structure, the binary vectors of P/A are independent and follow a multivariate Bernoulli distribution with proportion vectors depending on the associated partition. This approach is called NEM, as it relies on the Neighboring Expectation-Maximization algorithm [47–49]. As such, NEM tends to smooth the partitioning by grouping gene families that have a weighted majority of neighbors belonging to the same partition. The previously introduced latent variables $\{Z_i\}_{1\leqslant i\leqslant F}$, that indicate the partition to which each gene family belongs are now distributed according to a MRF. More precisely, they have the following Gibbs distribution:

$$\mathbb{P}(\{Z_i\}_{1\leqslant i\leqslant F}) = W_\beta^{-1} \exp\Big(\sum_{i=1}^{F}\sum_{k=1}^{K} \pi_k 1_{Z_i=k} + \beta \frac{F}{\sum_{i\sim i'} w_{i\sim i'}} \sum_{i\sim i'} w_{i\sim i'} 1_{Z_i=Z_{i'}}\Big)$$

where 1_A is the indicator function of event A and the second sum concerns every pair $(i \sim i')$ of neighbor gene families. The parameter $\beta \geq 0$ corresponds to the coefficient of spatial regularity. The $\frac{F}{\sum_{i\sim i'}^{E} w_{i\sim i'}}$ is a corrector term ensuring that the strength of the spatial smoothing is balanced regardless of the number of gene families. Indeed when the number of genomes (N) increases, the number of gene families (F) tends to be higher than the sum of the edge weights. Finally,

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

is a normalizing constant. Note that W_β cannot be computed, due to a large number of possible configurations. The degree of dependence between elements is controlled by the parameter β. Neighboring elements will be more inclined to belong to the same group with a higher value of this parameter. Here, the data vectors $(X_i)_{1\leqslant i\leqslant F}$ are not independent anymore. However, conditional on the latent groups $(Z_i)_{1\leqslant i\leqslant F}$, they are independent and follow the multivariate Bernoulli distribution:

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

Many different techniques may be used to approximate the maximum likelihood estimator in the hidden MRF. NEM relies on a mean-field approximation for the distribution of the latent random variables $Z_{i1\leqslant i\leqslant F}$ conditional on the observations. It should be noted that the optimal number of partitions (K) is not determined automatically using NEM and is therefore first estimated using the BinEM approach.

Issues resulting from high-dimensional statistics and parallelization

As plenty of statistical approaches, NEM is not adapted to high dimensional settings (i.e. whenever the condition $F >> N$ is not satisfied). This can occur in pangenomics as the discovery rate of new families in the pangenome slightly decreases when new genomes are added. Mathematical solutions to this problem seem to exist [50–52] for example via the weighting of genomes (based on their respective contribution to the pangenome diversity) or via sparse partitioning methods. An improvement of NEM should include these solutions and could be a perspective of this work.

Pangenome software must be designed to scale up to thousands of genomes. NEM scales quadratically with the number of genomes and is hard to parallelize. Thus, it leads to intensive computations when thousands of genomes are included in the analysis.

Our solution to the mentioned issues is to sample the genomes in chunks and to perform multiple partitioning in parallel. Each family must be involved in at least $N_{total}/N_{samples}$ samplings and will be partitioned only if it is classified in the same partition in at least 50% of the samplings where it is present (absolute majority). If some families do not respect this condition, we continue sampling until all gene families have been partitioned. Chunks have to be large enough to be representative; therefore a size of at least 500 genomes is advised.

Analysis of isolate genomes and Metagenome-Assembled Genomes

To obtain the set of isolate genomes to be analyzed, we downloaded all archaeal and bacterial genomes (220 561 genomes) of the GenBank database at the date of the 17th of April 2019. We removed genome assemblies that do not respect quality control criteria defined by GenBank. They correspond to entries with an assembly status flag different from "status=latest" in the "assembly_status.txt" files. In addition, genomes were discarded if they had more than 1000 contigs or a L90 > 100. These filters allowed us to exclude poor quality assemblies, some of which may correspond to contaminated genomes and others to incomplete ones. For each species (identified by its NCBI species taxid), a pairwise genomic distance matrix was computed using Mash (version 2.0) [37]. To avoid redundancy, if several genomes are at a Mash distance < 0.0001, only one was kept (the one having the lowest number of contigs). A single linkage clustering using SiLiX (version 1.2.11) [54] was then performed on the adjacency graph of the Mash distance matrix considering only distances below or equal to 0.06. This Mash distance corresponds to a 94% Average Nucleotide Identity (ANI) cutoff which is a usual value to define species [55]. Genomes that were not in the largest connected component were discarded to remove potential taxonomic assignation errors. Only species having at least 15 remaining genomes were then considered for the analysis. The list of all the GenBank assembly accessions used after filtering is available in S12 File. This dataset consists of 439 species encompassing 136 287 genomes (see S10 File). MAGs from the Pasolli *et al.* study [41] were downloaded from `http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html`. In this dataset, the genomes are already grouped in Species Genome Bins. These SGBs do not exactly match the GenBank taxonomy. Thus, SGBs assigned with the same species name (column "estimated taxonomy" in the supplementary table S4 of [41]) were merged to allow comparison with GenBank. SGBs that do not have a taxonomy assigned at the species level were not considered. A total of 583 species encompassing 698 SGBs and 71 766 MAGs were analyzed but only MAGs from 78 species were finally compared to GenBank genomes. To avoid introducing a bias in our analysis due to heterogeneous gene calling, GenBank annotations were not considered as they were obtained using a

variety of annotation workflows. Genomes from GenBank and Pasolli *et al* were consistently annotated using the procedure implemented in PPanGGOLiN. Prodigal (version 2.6.2) [56] is used to detect the coding genes (CDS). tRNA and tmRNA genes are predicted using Aragorn (version 1.2.38) [57] whereas the rRNA are detected using Infernal (version 1.1.2) [58] with HMM models from Rfam [59]. In the case of overlaps between a RNA and a CDS, the overlapping CDS are discarded. Homologous gene families were determined using MMseqs2 (version 8-fac81) [60] with the following parameters: coverage=80% with cov-mode=0, minimal amino acid sequence identity=80% and cluster-mode=0 corresponding to the Greedy Set Cover clustering mode. PPanGGOLiN partitioning was executed on each species using the NEM approach with a parameter $\beta = 2.5$. The nodes having a degree above 10 (which is the default parameter) were not considered to smooth the partitioning via the MRF. The number of partitions (K) was determined automatically for each NCBI species using a $\delta_{ICL} = 0.05$ and iterating between 3 and 20 for the possible values of K. K was fixed at 3 for the MAG analysis. The partitioning was done using chunks of 500 genomes when there were more than 500 genomes in a species. To compare PPanGGOLiN results between MAGs and GenBank genomes for each species, the representative sequences of each MAG gene family (extracted using the mmseqs2 subcommand: "result2repseq") were aligned (using mmseqs2 "search") on those of GenBank genomes. If the best hit of the query had a sequence identity $> 80\%$ and a coverage $> 80\%$ of the target, the 2 corresponding gene families of each dataset were associated.

Rarefaction curves

To represent the pangenome evolution according to the number of sequenced genomes, a multiple resampling approach was used. For each species with at least 100 genomes, 8 rarefaction curves showing the evolution of the pangenome and the persistent, shell, cloud, soft core, soft accessory, exact core and exact accessory components were computed for sample sizes of 1 to 100 genomes randomly drawn from the set of all genomes of the species. Each sample size was analyzed using 30 different samples. For each sample, the number of partitions K is automatically determined between 3 and the K obtained on all the genomes of the species. A non-linear Least Squares Regression was performed to fit the rarefaction curves with Heaps' law $F = \kappa N^{\gamma}$ where F is the number of gene families, N the number of genomes, γ the tendency of the evolution and κ a proportional factor [34]. Subset sizes ≤ 15 were not used for the fitting as they are sometimes too variable to ensure a good fitting. The function "scipy.optimize.curve_fit" of the Python scipy package (version 1.0.0), based on the Levenberg-Marquardt algorithm, was used to fit the rarefaction curves. For each subset size, the median and quartiles were calculated to obtain a ribbon of interquartile ranges (IQR) along the rarefaction curves. We call the area of this ribbon the IQR area (see S2 Fig as an example).

PPanGGOLiN software implementation

PPanGGOLiN was designed to be a software suite performing the annotation of the genomic sequences, building the gene families and the pangenome graph before partitioning it. Users can also provide their own annotations (GFF3 or GBFF format) and gene families. The application stores its data in a compressed HDF5 file but can also return the graph in GEXF or JSON formats and the P/A matrix with the partitioning in CSV or Rtab files (similarly to the ones provided by Roary [35]). It also generates several illustrative figures, some of which are presented in the article. PPanGGOLiN was developed in the Python 3 and C languages and is intended to be easily installable on Linux and Mac OS systems via a BioConda package [61] (see

`https://bioconda.github.io/recipes/ppanggolin/README.html`). The code is also freely available on the GitHub website at the following address: `https://github.com/labgem/PPanGGOLiN`.

Supporting information

S1 Fig. Density distributions of the gene family frequencies of each partition. Results for the 88 most abundant species in GenBank are represented in addition with a global distribution of the gene family frequencies from all the species. Density values of the cloud genome above 100 (y-axis) were trimmed for visualization purpose. The dashed yellow vertical bars indicate the threshold of frequency ($\geqslant$95%) used to delimit the soft core genome.

S2 Fig. Evolution of the persistent, shell, soft core and exact core metrics of Lactobacillus plantarum compared to the number of genomes. The rarefaction curves represent the evolution of the partition sizes as a function of an increasing number of genomes in random subsets of genomes. Plain lines connect the medians while colored areas represent the interquartile ranges. A regression curve (bold dashed line) is drawn fitting all the points of each partition by the Heaps' law ($F = \kappa N^{\gamma}$). The total area of the interquartile ranges (IQR) is indicated for each partition.

S3 Fig. Density distributions of the Heaps' law γ-tendencies. These γ-tendencies were obtained by fitting a Heaps' law on rarefaction curves between subset sizes of 15 to 100 genomes in the 88 most abundant species in GenBank. The exact core median and exact accessory are not shown.

S4 Fig. Shell γ-tendency compared to the total number of shell families normalized by the median number of gene families per genome in each species. Results for the 88 most abundant species in GenBank are represented. The points are colored by phylum and their size corresponds to the number of partitions (K) used.

S5 Fig. Fraction of shell families per genome compared to the number of gene families. Results for the 88 most abundant species in GenBank are represented. The points are colored by phylum and their size corresponds to the number of partitions (K) used.

S6 Fig. Fraction of cloud families per genome compared to the number of gene families. Results for the 88 most abundant species in GenBank are represented. The points are colored by phylum and their size corresponds to the number of partitions (K) used.

S7 Fig. Presence/Absence matrix of the shell genome of *L. reuteri* ordered by a Neighbor Joining tree based on the MASH distances. The leaves of the tree are colored by host or origin. This information was obtained from the metadata in GenBank files (host and isolation source qualifiers).

S8 Fig. Presence/Absence matrix of the shell genome of *B. longum* ordered by a Neighbor Joining tree based on the MASH distances. The leaves of the tree are colored by species clusters defined by the GTDB database (release R04-RS89), namely (*B. infantis* or *B. longum*). "NA" values corresponds to genomes not available in GTDB.

S9 Fig. Illustration of the shell genome overlaps between MAGs or GenBank of 55 species. The x-axis represents the percentage of common shell of the GenBank shell while the y-axis corresponds to the percentage of common shell of the MAGs shell. Diamonds and squares represent MAGs and GenBank genomes, respectively. They are colored by phylum and their size indicates the number of genomes.

S10 File. Table compiling all the metrics obtained from the pangenomes of the 439 GenBank species. This is a CSV file.

S11 File Table compiling all the metrics obtained from the comparison of PPanGGOLiN results between MAGs and GenBank genomes in 78 species. This is a CSV file.

S12 File. List of GenBank assembly accessions for the 439 studied species. This is a TSV file where each line correspond to all the GenBank assembly accession used in this study for each 'species id' in the NCBI taxonomy.

Acknowledgments

We acknowledge Alexandre Renaux and Jonathan Mercier for their preliminary insights on pangenome graphs. We thank Mélanie Buy for drawing the PPanGGOLiN logo. Finally, we thank Guilhem Royer, Valentin Sabatet, Johan Rollin, Mohammed-Amin Madoui, Tom Delmont, Nicolas Pons and Pierre Peterlongo for all their advice along this work.

References

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci USA. 2005;102(39):13950–13955.
2. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005;15(6):589–594.
3. Treangen TJ, Rocha EPC. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLOS Genetics. 2011;7(1):1–12. doi:10.1371/journal.pgen.1001284.
4. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced Escherichia coli genomes. Microb Ecol. 2010;60(4):708–720.
5. Acevedo-Rocha CG, Fang G, Schmidt M, Ussery DW, Danchin A. From essential to persistent genes: a functional approach to constructing synthetic life. Trends Genet. 2013;29(5):273–279.

6. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 2013;79(24):7696–7701.

7. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. Trends Genet. 2009;25(3):107–110.

8. Bolotin E, Hershberg R. Horizontally Acquired Genes Are Often Shared between Closely Related Bacterial Species. Front Microbiol. 2017;8:1536.

9. Vesth T, Wassenaar TM, Hallin PF, Snipen L, Lagesen K, Ussery DW. On the Origins of a Vibrio Species. Microbial Ecology. 2010;59(1):1–13. doi:10.1007/s00248-009-9596-7.

10. Periwal V, Patowary A, Vellarikkal SK, Gupta A, Singh M, Mittal A, et al. Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of Mycobacterium tuberculosis pangenome. PLoS ONE. 2015;10(4):e0122979.

11. Livingstone PG, Morphew RM, Whitworth DE. Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets. Front Microbiol. 2018;9:3187.

12. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008;36(21):6688–6719.

13. Baumdicker F, Hess WR, Pfaffelhuber P. The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol. 2012;4(4):443–456.

14. Collins RE, Higgs PG. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol Biol Evol. 2012;29(11):3413–3425.

15. Lobkovsky AE, Wolf YI, Koonin EV. Gene frequency distributions reject a neutral model of genome evolution. Genome Biology and Evolution. 2013;.

16. Bolotin E, Hershberg R. Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. Genome Biol Evol. 2015;7(8):2173–2187.

17. Moldovan MA, Gelfand MS. Pangenomic Definition of Prokaryotic Species and the Phylogenetic Structure of Prochlorococcus spp. Frontiers in Microbiology. 2018;9:428. doi:10.3389/fmicb.2018.00428.

18. Chan AP, Sutton G, DePew J, Krishnakumar R, Choi Y, Huang XZ, et al. A novel method of consensus pan- chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii. Genome Biol. 2015;16:143.

19. Garrison E, Siren J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36(9):875–879.

20. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–915.

21. Rakocevic G, Semenyuk V, Lee WP, Spencer J, Browning J, Johnson IJ, et al. Fast and accurate genomic analyses using genome graphs. Nat Genet. 2019;51(2):354–362.

22. Consortium TCPG. Computational pan-genomics: status, promises and challenges. Brief Bioinformatics. 2016;.

23. Zekic T, Holley G, Stoye J. Pan-Genome Storage and Analysis Techniques. Methods Mol Biol. 2018;1704:29–53.

24. Snipen L, Almøy T, Ussery DW. Microbial comparative pan-genomics using binomial mixture models. BMC Genomics. 2009;10:385.

25. Snipen L, Liland KH micropan: an R-package for microbial pan-genomics BMC Bioinformatics. 2015;16:79.

26. van Tonder AJ, Mistry S, Bray JE, Hill DM, Cody AJ, Farmer CL, et al. Defining the estimated core genome of bacterial populations using a Bayesian decision model. PLoS Comput Biol. 2014;10(8):e1003788.

27. Gumiere T, Meyer K, Burns AR, Gumiere SJ, Bohannan BJM, Andreote FD. A probabilistic model to identify the core microbial community. bioRxiv. 2018;doi:10.1101/491183.

28. Fang G, Rocha EP, Danchin A. Persistence drives gene clustering in bacterial genomes. BMC Genomics. 2008;9(1):4.

29. Oliveira PH, Touchon M, Cury J, Rocha EPC. The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun. 2017;8(1):841.

30. Singh JK, Adams FG, Brown MH. Diversity and Function of Capsular Polysaccharide in Acinetobacter baumannii. Front Microbiol. 2018;9:3301.

31. Hu D, Liu B, Dijkshoorn L, Wang L, Reeves PR. Diversity in the major polysaccharide antigen of Acinetobacter baumannii assessed by DNA sequencing, and development of a molecular serotyping scheme. PLoS ONE. 2013;8(7):e70329.

32. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks; 2009.Available from: `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`.

33. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLOS ONE. 2014;9(6):1–12. doi:10.1371/journal.pone.0098679.

34. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008;11(5):472–477.

35. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691–3693.

36. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36(10):996–1004.

37. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016;17(1):132. doi:10.1186/s13059-016-0997-x.

38. Criscuolo A. A fast alignment-free bioinformatics procedure to infer accurate distance-based phylogenetic trees from genome assemblies. Research Ideas and Outcomes. 2019;5:e36178. doi:10.3897/rio.5.e36178.

39. Hawkey J, Monk JM, Billman-Jacobe H, Palsson B, Holt KE. Impact of insertion sequences on convergent evolution of Shigella species. bioRxiv. 2019;doi:10.1101/680777.

40. Oh PL, Benson AK, Peterson DA, Patil PB, Moriyama EN, Roos S, et al. Diversification of the gut symbiont Lactobacillus reuteri as a result of host-driven evolution. ISME J. 2010;4(3):377–387.

41. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019;176(3):649–662.

42. Brockhurst MA, Harrison E, James PJ, Richards T, McNally A, MacLean C The Ecology and Evolution of Pangenomes Current Biology. 2019;29(20):1094–1103.

43. Vallenet D, Calteau A, Dubois M, Amours P, Bazin A, Beuvin M, et al. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. Nucleic Acids Research. 2019;doi:10.1093/nar/gkz926.

44. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B. 1977;39(1):1–38.

45. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000;22(7):719–725. doi:10.1109/34.865189.

46. Schwarz G. Estimating the Dimension of a Model. Ann Statist. 1978;6(2):461–464. doi:10.1214/aos/1176344136.

47. Ambroise C, Dang M, Govaert G. Clustering of Spatial Data by the EM Algorithm. In: Soares A, Gómez-Hernandez J, Froidevaux R, editors. geoENV I — Geostatistics for Environmental Applications. Dordrecht: Springer Netherlands; 1997. p. 493–504.

48. Ambroise C, Govaert G. Convergence of an EM-type algorithm for spatial clustering. Pattern Recognition Letters. 1998;19(10):919 – 927. doi:https://doi.org/10.1016/S0167-8655(98)00076-2.

49. Dang M, Govaert G. Spatial Fuzzy Clustering using EM and Markov Random Fields. In: InternationalJournal of System Research and Information Science; 1998. p. 183–202.

50. Bouguila N. On multivariate binary data clustering and feature weighting. Computational Statistics and Data Analysis. 2010;54(1):120 – 134. doi:https://doi.org/10.1016/j.csda.2009.07.013.

51. Yamamoto M, Hayashi K. Clustering of multivariate binary data with dimension reduction via L1-regularized likelihood maximization. Pattern Recognition. 2015;48(12):3959–3968. doi:10.1016/j.patcog.2015.05.026.

52. Śmieja M, Hajto K, Tabor J. Efficient mixture model for clustering of sparse high dimensional binary data. Data Mining and Knowledge Discovery. 2019;doi:10.1007/s10618-019-00635-1.

53. Laing CR, Whiteside MD, Gannon VPJ. Pan-genome Analyses of the Species Salmonella enterica, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar. Front Microbiol. 2017;8:1345.

54. Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics. 2011;12:116.

55. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA. 2005;102(7):2567–2572.

56. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.

57. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32(1):11–16.

58. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–2935.

59. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018;46(D1):D335–D342.

60. Steinegger M, Soeding J. Sensitive protein sequence searching for the analysis of massive data sets. Nat Biotech. 2017;doi:doi:10.1038/nbt.3988.

61. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature methods. 2018;15(7):475–476.

I.4 Prédiction des îlots génomiques à partir du graphe de pangénome

De part sa structure compacte et l'information de voisinage génomique qu'il contient, le graphe de pangénome de PPanGGOLiN peut être utilisé pour des approches de génomique comparée. Dans cette optique, Adelme Bazin a débuté une thèse en octobre 2018 que je co-encadre avec ma collègue du LABGeM, Alexandra Calteau.

Un des premiers objectifs est d'étudier les régions variables dans le graphe qui correspondent à des chemins de gènes du *shell* ou du *cloud* interrompant le squelette du génome *persistent*. Ces régions de plasticité génomique (RGP) correspondent, notamment, à des îlots génomiques ("Genomic Islands", GI) impliqués dans la pathogénie, la résistance aux antibiotiques ou, encore, à des îlots d'adaptation au mode de vie de l'organisme comme le saprophytisme ou la symbiose [84]. En effet, les gènes obtenus par transferts horizontaux constituent la principale source de diversité des pangénomes [59] et s'insèrent préférentiellement dans certains loci des génomes appelés points chauds d'intégration ("hot spot") [80]. Pour identifier ces GIs dans les génomes procaryotes, il existe trois catégories d'approches : (i) les méthodes détectant les différences en termes de composition en nucléotides (ii) les méthodes fondées sur la génomique comparée (iii) des méthodes hybrides intégrant les deux approches précédentes [85]. Par ailleurs, ces méthodes affinent parfois leurs résultats en recherchant des séquences répétées ou des gènes de fonction spécifique (*e.g.* séquences d'insertion, nommées IS, ou des séquences d'ARNt) qui se trouvent souvent en bordure des îlots génomiques. Les méthodes utilisant le différentiel de composition nucléotidique sont relativement précises pour détecter des transferts horizontaux récents mais fonctionnent moins bien quand le transfert est ancien. En effet, les gènes transférés tendent à s'homogénéiser au cours du temps avec la composition native de l'espèce en accumulant des mutations [86]. De plus, un îlot ne sera pas détecté si la région transférée est issue d'un organisme dont le génome a une composition similaire à celle du génome receveur. En ce qui concerne les méthodes basées sur la génomique comparée, elles repèrent des gènes ayant une distribution phylogénétique différente de celle à laquelle on pourrait s'attendre dans l'hypothèse où les gènes seraient hérités verticalement (*i.e.* des gènes présents dans une souche mais absents dans d'autres souches de la même espèce). Ces méthodes sont plus sensibles et précises que celles basées sur la composition. Cependant, leurs résultats sont très influencés par le choix au préalable de génomes de références qui ne doivent pas être ni trop proches ni trop distants du génome à analyser. De plus, elles utilisent souvent des comparaisons de génomes deux à deux en les alignant ou en comparant leur contenu en gènes, ce qui les rendent difficilement extensibles à l'analyse de plusieurs milliers de génomes.

La méthode que nous développons, nommée panRGP, est capable de détecter des RGPs dans les génomes à partir du partitionnement effectué par PPanGGOLiN. Ces régions contiennent les GIs mais également des plasmides et potentiellement des régions perdues dans un sous-ensemble de souches suite à des évènements de réduction de génomes qui sont anciens dans l'évolution de l'espèce. A partir d'un pangénome reconstruit pour une espèce, la méthode panRGP projette les résultats du partitionnement sur les génomes afin d'associer chaque gène à une partie du pangénome (génome *persistent*, *shell* ou *cloud*). Elle parcourt ensuite chaque contig en attribuant un score aux gènes qui correspond à la somme du poids du gène avec le score du gène précédent. Le poids pour les gènes *shell* et *cloud* est de +1 et pour les gènes *persistent* de $-(3^n)$ permettant de pénaliser fortement l'insertion de plusieurs gènes *persistent* (n étant le nombre de gènes *persistent* consécutifs). Le score d'un gène est borné à 0 pour sa valeur minimale. Cet algorithme bien que parcourant les contigs du début à la fin est symétrique (*i.e.* il produit les mêmes résultats si le parcours est réalisé dans l'autre sens). L'objectif est d'extraire les régions les plus grandes contenant une majorité de gènes du *shell* ou du *cloud* potentiellement interrompus par quelques gènes *persistent*. Ainsi, la méthode extrait les RGPs de score maximum qui correspond au score du gène à la fin d'une RGP, le début d'une RGP étant déterminé en parcourant en sens inverse les gènes jusqu'à rencontrer un gène de score nul.

Pour évaluer les résultats de panRGP en comparaison des autres méthodes de détection de GIs, nous avons réalisé une évaluation sur un jeu de données constitué de six génomes complets ayant des annotations expertisées de GIs [87]. Les auteurs ont inspecté les régions correspondant effectivement à des îlots génomiques (régions positives) et déterminé celles qui n'en sont pas (régions négatives). La fiabilité des différentes méthodes (cf. Table 1) a été évaluée en calculant la sensibilité, la spécificité, la précision et le score F1. L'unité de mesure est en nombre de nucléotides. Comme attendu les méthodes basées sur la génomique comparée obtiennent des résultats nettement meilleurs que les approches basées sur la composition en nucléotides. La méthode panRGP se classe en tête avec celle de xenoGI [88] ce qui montre l'intérêt d'utiliser un pangénome de référence pour prédire les GIs. La méthode xenoGI est par contre très gourmande en ressource de calcul. Les auteurs indiquent que l'analyse de 40 souches demande 20 heures de calcul sur 50 cœurs (en utilisant 500 Go de mémoire vive). À l'inverse, la méthode panRGP est pour sa part quasiment instantanée et peut donc s'appliquer sur des milliers de génomes une fois le pangénome obtenu par PPanGGOLiN. À titre d'exemple, la méthode PPanGGOLiN utilisée sur 1000 génomes de l'espèce *Salmonella enterica* requiert un temps de calcul d'environ 45 minutes sur 16 cœurs et 12 Go de mémoire vive.

Méthode	Score F1	Précision	Spécificité	Sensibilité	Approche
panRGP	**0,932**	**0,931**	**1,0**	0,884	comparée (pangénome)
xenoGI	0,917	0,905	0,935	**0,924**	comparée
islandviewer4	0,791	0,817	0,998	0,669	hybride
IslandCafe	0,715	0,752	**1,0**	0,574	compositionnelle
GI-Cluster	0,743	0,761	0,87	0,714	compositionnelle
PredictBias	0,805	0,788	0,856	0,771	compositionnelle
IslandPath-DIMOB	0,636	0,702	0,998	0,479	compositionnelle
SIGI-CFR	0,52	0,687	0,993	0,434	compositionnelle
AlienHunter	0,642	0,705	0,753	0,57	compositionnelle
SIGI-HMM	0,444	0,591	0,817	0,325	compositionnelle
ZislandExplorer	0,278	0,513	0,833	0,18	compositionnelle

Table 1 : Evaluation des résultats de panRGP en comparaison d'autres méthodes de prédiction d'îlots génomiques.

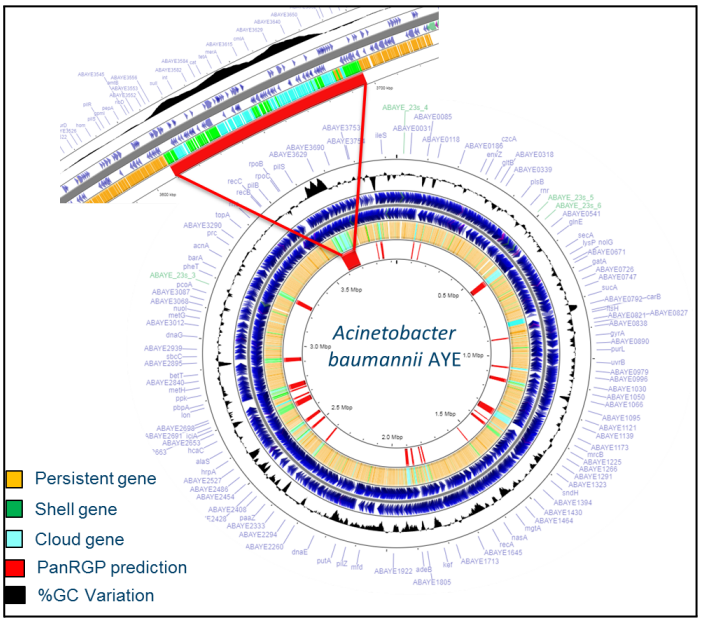

Figure 2 : Îlots génomiques prédits par panRGP pour la souche *A. baumannii* AYE.

Les gènes sont représentés en bleu et les prédictions de panRGP en rouge. Le partitionnement de PPanGGOLiN est également indiqué (en orange pour le *persistent*, vert pour le *shell* et bleu pour le *cloud*) ainsi que la variation du pourcentage en GC par rapport à la moyenne du génome. Un zoom est réalisé sur un îlot génomique de 86 kb contenant 45 gènes impliqués dans la résistance aux antibiotiques. Cette région est constituée majoritairement de gènes *cloud* et *shell* et de quelques gènes *persistent* dans sa partie terminale.

Pour illustrer les prédictions de panRGP, la Figure 2 présente les résultats obtenu pour *Acinetobacter baumannii* AYE, une souche qui a été impliquée dans des infections nosocomiales en France en 2001 [89]. L'analyse de son génome nous a permis d'identifier un îlot génomique de 86 kb contenant 45 gènes impliqués dans la résistance aux antibiotiques [2,3]. Cet îlot a été retrouvé en intégralité par panRGP.

La méthode panRGP fonctionne sur les génomes complets mais aussi sur les génomes dont l'assemblage est en plusieurs contigs. Pour ces derniers, certaines RGPs ne seront pas bornées par des gènes *persistent* et seront donc considérées comme partielles (*i.e.* l'intégralité du contig correspondra à une RGP). En effet, les GIs contiennent assez fréquemment des familles de gènes répétées (*e.g.* des gènes de transposase) qui ont pour conséquence d'interrompre l'assemblage. Sur les RGP complètes, une méthode additionnelle de panRGP consiste à détecter les groupes de gènes *persistent* bornant les RGPs pour définir des sites d'insertion ("spots"). Pour autoriser d'éventuelles pertes de gènes *persistent* qui peuvent avoir lieu pendant l'événement d'insertion d'une région dans un génome, panRGP compare des paires de n-uplets de gènes (paramétré à 3) pour regrouper les régions flanquantes en *spots* en autorisant des recouvrements partiels de k gènes (paramétré à 2). Ces *spots* ont un intérêt dans l'étude de l'évolution des espèces, par exemple, en déterminant les points chauds d'intégration ("hotspots", sites d'intégration fréquents au sein des souches et contenant une importante variabilité de gènes) et leur dynamique en termes de diversité génétique (turnover et imbrication) [80].

Cette méthode illustre l'intérêt du graphe de pangénome comme structure de données dans la conception d'algorithmes pour la génomique comparée pouvant passer à l'échelle pour l'analyse de plusieurs milliers de génomes. Nous avons pour objectif de présenter ce travail à la conférence ECCB ("European Conferences on Computational Biology") 2020 dont les actes sont publiés dans le journal Bioinformatics. La méthode panRGP est d'ores et déjà intégrée dans la plateforme MicroScope [14] et donc accessible aux microbiologistes.

I.5 Détection de modules conservés dans les îlots génomiques

La thèse d'Adelme Bazin se poursuivra par le développement d'une méthode de détection de modules au sein des îlots génomiques. En effet, lorsque l'on compare des GIs de plusieurs souches s'étant insérés dans un même *hotspot*, on observe à la fois une grande variabilité dans le contenu et l'organisation des gènes mais également des conservations de sous-groupes de gènes co-localisés que l'on nomme module (*cf.* Figure 3 issue d'un article décrivant des modules conservés dans le *hotspot* de

l'ARNt *leuX* chez *E. coli* [90]). Ces modules ont un intérêt d'une part pour comprendre l'origine évolutive des GIs qui résultent souvent de plusieurs événements d'insertion et, également, d'un point de vue fonctionnel car les gènes d'un module sont supposés participer à un même processus biologique.

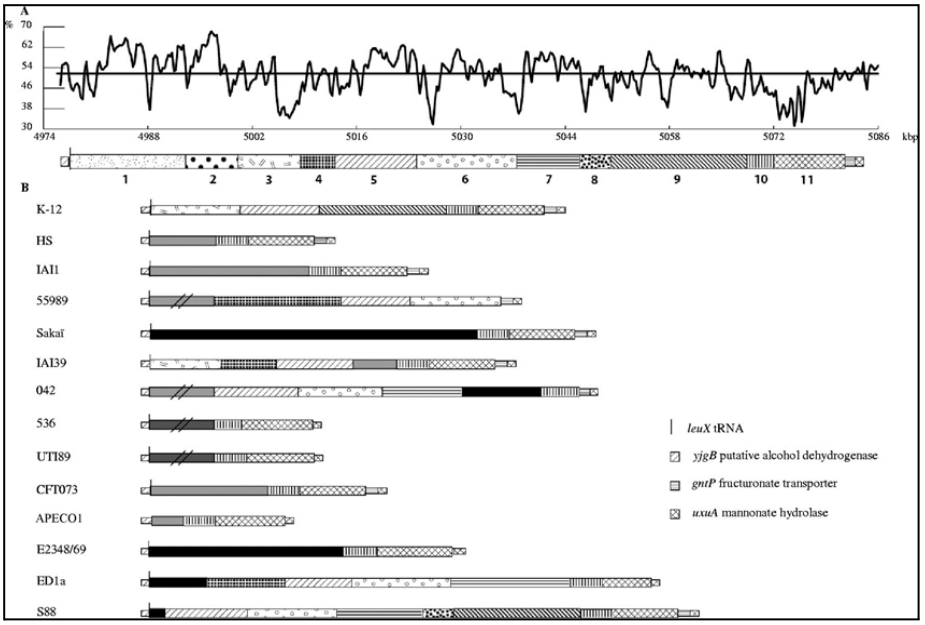

Figure 3 : Îlots génomiques et modules conservés dans le *hotspot* de l'ARNt *leuX* chez 15 souches *E. coli*.
Cette figure est issue de l'article [90] (figure 4). (A) Organisation de l'îlot de la souche UMN026 composé de 11 modules représentés par des rectangles décorés. La courbe représente le pourcentage en GC calculé sur une fenêtre de 500-pb et la barre horizontale indique la moyenne en GC du génome (50,7%). (B) Organisation de l'îlot dans 14 autres souches. Cet îlot est borné par les gènes *persistent yjgB* et ARNt *leuX* (absent dans APECO1 and 042) en 5' et les gènes *gntP* (absent dans 5 souches) et *uxuA* en 3'. Les modules en noir représentent des régions spécifiques aux souches et les modules en gris sont retrouvés dans d'autres souches mais pas dans cet îlot. Ces modules ont été déterminés par une expertise manuelle au travers de l'analyse des résultats de conservation de synténie disponibles dans la plateforme MicroScope [5].

Pour détecter ces modules à partir des résultats de panRGP, le cadre informatique qui a été choisi est celui de la recherche d'ensembles d'items fréquents ("frequent itemset mining"). Les ensembles sont les RGPs prédites par panRGP dans tous les génomes d'une espèce et les items les familles de gènes homologues contenues dans ces RGPs. On note ici que la co-localisation des familles au sein des RGPs n'est pas imposée car les ensembles ne sont pas ordonnés. Ce choix simplifie grandement la difficulté du problème informatique et, en pratique, a peu de conséquences car des familles conservées au sein de plusieurs RGPs sont très souvent contiguës sur les génomes. Pour autoriser des conservations partielles (*i.e.* absence de quelques familles dans certaines RGPs

constituant un module), nous allons rechercher des ensembles maximaux d'items pour lesquels la similitude de leur couverture (*i.e.* les ensembles de transactions qui les contiennent) dépasse un seuil spécifié de similarité. Nous avons choisi l'indice de Jaccard pour mesurer la similitude entre les ensembles. Une solution exacte à ce problème existe et a été implémentée et optimisée [91]. Les premiers résultats obtenus sont encourageants et l'implémentation de l'algorithme passe à l'échelle (*e.g.* analyse de 3 117 génomes contenant 99 751 RGPs et 46 875 familles en 12 minutes avec un indice de Jaccard seuil ≥ 0.95). Seules les RGPs complètes (*i.e.* bornées par des gènes *persistent*) sont considérées pour éviter des biais dus aux artefacts d'assemblage qui introduisent de fausses absences de familles quand la RGP recouvre tout un contig et donc est potentiellement incomplète. De plus, pour limiter la redondance, les RGPs contenant le même ensemble de familles sont fusionnées avant l'analyse. On note tout de même un effet important du seuil de l'indice de Jaccard choisi générant pour des seuils élevés de nombreux modules de petite taille et, inversement, pour des valeurs basses de seuil. Des améliorations sont donc à envisager pour une meilleure définition des modules. Une piste possible serait de faire varier le seuil de l'indice de Jaccard puis de réconcilier localement les modules en recherchant des communautés dans un graphe où les nœuds sont les familles et les arêtes représentent l'appartenance à un même module pondérée par l'indice de Jaccard. La prise en compte d'une distance phylogénétique serait également à considérer pour donner plus de poids aux modules conservés entre des souches éloignées et ainsi pénaliser des conservations dans des îlots hérités verticalement.

Une étape de validation des modules prédits sera ensuite réalisée soit à partir de quelques RGPs expertisées manuellement (*e.g.* le *hotspot leuX* chez *E. coli*) ou d'une manière plus globale en déterminant si les modules regroupent des familles d'un même processus cellulaire (*e.g.* enzymes d'une même voie métabolique, protéines d'un système de sécrétion, gènes d'un prophage). De plus, une comparaison avec des résultats plus globaux issus de méthodes de contexte génomique (*i.e.* appliquées sur des génomes de différentes espèces et pour tous les gènes de leur génome), comme ceux de la base de données STRING [92], permettront d'évaluer si des familles d'un même module ont des scores d'association plus élevés que des familles appartenant à différents modules d'une même RGP.

I.6 Pangénomique comparée à l'échelle d'une espèce ou d'un écosystème

Le graphe de pangénome partitionné de PPanGGOLiN ainsi que les prédictions de RGPs et de modules ouvrent de nouvelles voies dans l'analyse comparée de génomes d'une même espèce ou de plusieurs espèces dans un écosystème pour réaliser, par exemple, des études d'association génotype-phénotype (appelées GWAS pour "Genome Wide Association Studies").

I.6.1 La ressource panGBank

Tout d'abord, il nous paraît nécessaire de constituer une base de données exhaustive de pangénomes de référence à partir des centaines de milliers de génomes d'isolats ou de MAGs disponibles dans les banques de séquences. Une première version de cette ressource, nommée panGBank, a été établie pour la publication de la méthode PPanGGOLiN et nous a amené à identifier plusieurs points à améliorer.

Un premier point concerne l'assignation taxonomique des génomes indiquée dans les banques comme GenBank. En effet, malgré l'effort de curation notamment au travers du projet RefSeq [93], il subsiste de nombreux problèmes dans la définition des espèces procaryotes. Ainsi, une initiative relativement récente a pour objectif de réviser cette taxonomie au travers d'une analyse globale de la phylogénie à l'échelle des génomes [94]. Cette ressource, nommée GTDB, a permis de reclassifier 58% des ~100 000 génomes analysés notamment en définissant de nouvelles espèces. De plus, l'outil GTDB-Tk, développé par les mêmes auteurs, permet de réaliser une assignation taxonomique pour des génomes nouvellement séquencés [95]. Pour la ressource panGBank, nous avons donc décidé d'utiliser la taxonomie de GTDB et de développer une méthode rapide d'assignation taxonomique basée sur une estimation de l'ANI ("Average Nucleotide Identity") avec les génomes de référence de GTDB en utilisant le logiciel MASH [96] (*i.e.* l'outil GTDB-Tk étant trop lent pour assigner la taxonomie de centaines de milliers de génomes). Plusieurs études ont en effet montré que des similarités génomiques de type ANI étaient de bonnes métriques pour classifier des génomes au niveau espèce [97]. Parallèlement, une évaluation de la complétion et contamination des génomes serait nécessaire pour s'assurer de la bonne qualité des pangénomes de référence, même si la méthode PPanGGOLiN est résiliente à ces problèmes dans l'évaluation du génome *persistent* et *shell*. Pour cela, une méthode, nommée CheckM, a été développée par les mêmes auteurs de GTDB-Tk mais elle est également trop lente pour être appliquée sur la volumétrie de génomes que nous souhaitons analyser [98]. Ainsi, nous envisageons de filtrer les génomes ayant des proportions aberrantes de *persistent*, *shell* et *cloud* suite à un premier partitionnement par PPanGGOLiN. Un traitement

particulier pour les génomes de MAGs sera également à définir surtout pour les espèces représentées majoritairement par ce type de génomes.

Un deuxième point concerne la construction des familles de gènes homologues. Pour cela, nous utilisons la méthode MMseqs2 [99]. Elle permet d'aligner des séquences de protéines environ 1000 à 100 000 fois plus rapidement que BLASTp avec un niveau de sensibilité comparable et donc peut s'appliquer sur des millions de séquences, ce qui est indispensable pour pouvoir constituer des pangénomes contenant des dizaines de milliers de génomes. MMseqs2 recherche des couples de k-mers étant séparés par un même nombre de résidus entre les séquences requêtes et cibles puis les alignent via un algorithme classique de Smith-Waterman [100]. Nous réalisons des comparaisons des séquences en acides aminés pour gagner en sensibilité en étant tolérant aux mutations synonymes. Les seuils d'alignement retenus pour constituer les familles sont de 80% d'identité et de 80% de couverture d'alignement sur les deux séquences. Les séquences sont ensuite regroupées en familles en utilisant l'algorithme de regroupement glouton par couverture d'ensemble ("Greedy Set cover") implémenté dans MMseqs2. Bien que satisfait de cette approche, nous souhaitons tout de même l'améliorer car nous avons constaté que de nombreuses familles du génome *cloud* (souvent des singletons) s'alignent partiellement (couverture < 80%) avec des familles du *persistent*. Ces gènes correspondent en fait à des fragments issus de CDSs prédites en bordure de contigs, d'erreurs de séquençage ou de pseudogènes et ont pour conséquence une surestimation du *cloud* dans le pangénome (i.e. ces fragments représentent en moyenne 25% des familles du génome *cloud* et uniquement 3% du *shell*). L'identification de ces pseudogènes est importante d'un point de vue fonction cellulaire et évolution [101,102]. En termes de topologie dans le graphe de pangénome, les familles de fragments forment fréquemment de petits chemins de 1 à 3 nœuds qui se branchent sur des chemins de gènes *persistent* et sont communément appelées les "écailles du PPanGGOLiN". Pour pallier ce problème, nous étudions une solution visant à appliquer un deuxième filtre après la constitution des familles. Il consiste à ré-aligner les séquences représentatives de chaque famille avec un taux de couverture de 80% mais uniquement sur la plus petite des séquences et, ainsi, repérer des familles du *cloud* (et éventuellement du *shell*) de plus faible effectif dont leurs séquences sont incluses dans d'autres familles du *persistent* ou du *shell* d'effectif plus important. Ces familles de fragments seraient alors fusionnées avec leur famille originelle tout en étiquetant leurs gènes comme fragments. Un inconvénient de cette approche est que les fragments de gènes ne s'alignent pas forcément correctement car leur traduction peut être erronée en raison de la présence de mutations ou d'erreurs de nucléotide décalant la phase de lecture. Des approches, comme celle proposée par les méthodes PEPPA [103] (*i.e.* où les séquences représentatives des familles sont alignées sur les génomes) ou Panaroo [104] (*i.e.* qui utilise la topologie d'un graphe de pangénome), sont donc également à étudier.

La ressource panGBank devra être mise à jour régulièrement et nécessitera donc le développement d'un *workflow* supportant l'ajout incrémental de nouveaux génomes dans les pangénomes de chaque espèce. Une API ("Application Programming Interface") sera également à mettre en place permettant à la communauté de requêter la ressource et de télécharger les pangénomes d'intérêt pour des analyses locales. Dans un second temps, une interface Web permettant de naviguer dans les pangénomes sera développée. Un premier prototype a été conçu par Rémi Planel, un ancien ingénieur de l'équipe (cf. Figure 4).

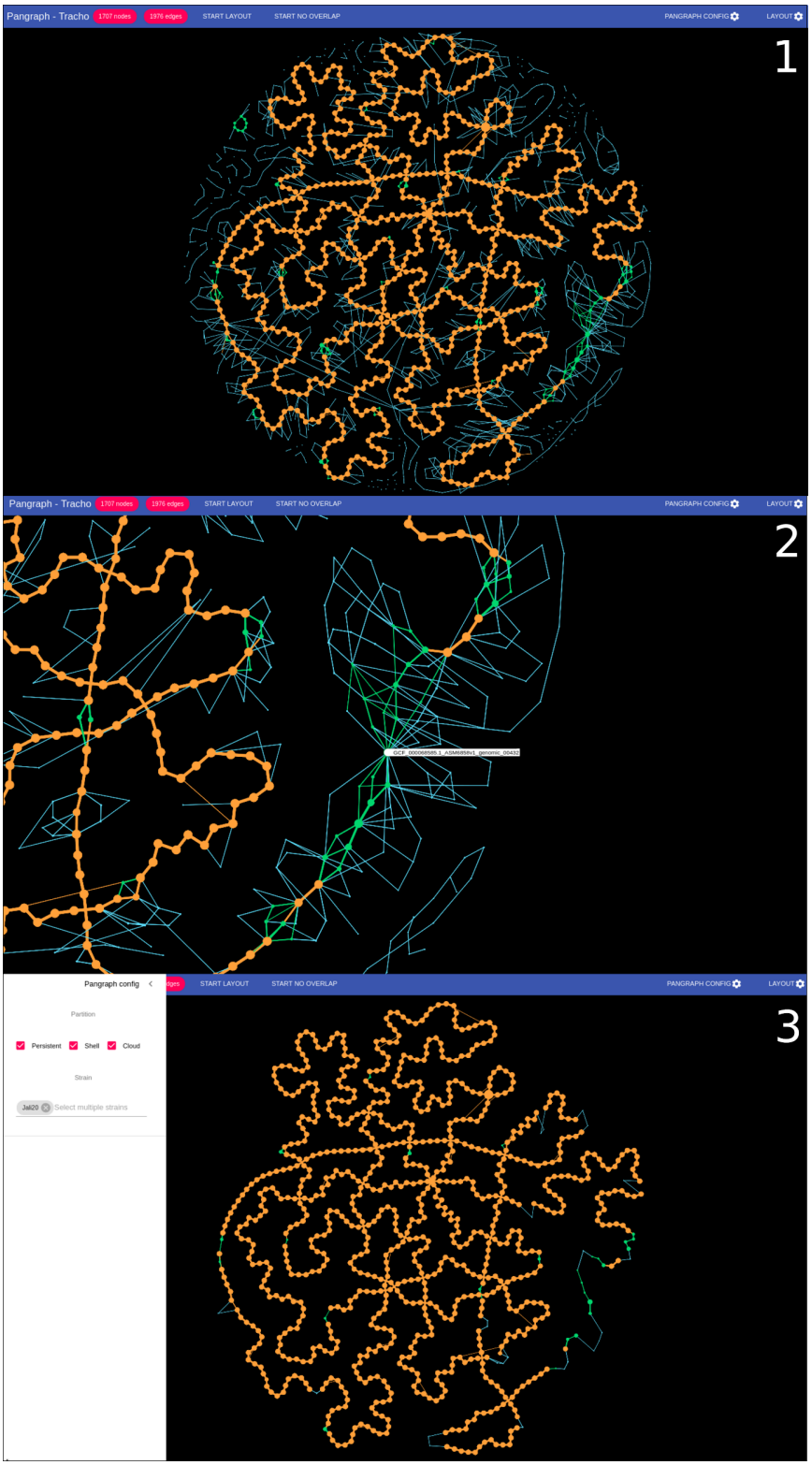

Figure 4 : Prototype de l'interface Web de panGBank représentant le pangénome de *Chlamydia trachomatis*.
(1) Visualisation de l'intégralité du pangénome. (2) Zoom sur une famille d'intérêt. (3) Visualisation du pangénome d'un sous-ensemble de souches (seule la souche Jali20 est sélectionnée dans l'exemple).

I.6.2 Génomique d'association : développements futurs et cas d'études

Comme évoqué précédemment, nous souhaitons utiliser les graphes de pangénomes pour explorer la diversité génomique des espèces procaryotes autour de diverses questions biologiques. Nous conduirons ainsi des analyses bioinformatiques sur des espèces d'intérêt pour identifier des déterminants génétiques d'importance pour un phénotype (ou un paramètre environnemental) ou sur des écosystèmes pour une caractérisation fine des espèces, fonctions et interactions présentes.

A l'échelle d'une espèce, le concept de corréler un phénotype avec des variations génomiques est maintenant applicable sur des milliers de génomes [105]. Par contre, l'obtention de phénotypes consistants sur des milliers de souches demeure difficile à mettre en œuvre du fait des contraintes expérimentales. Ces données phénotypiques sont donc encore rares et concernent, généralement, des phénotypes de croissance sur des milieux minimaux (*i.e.* en faisant varier la source de carbone ou d'azote) ou de résistance à des antibiotiques. Des facteurs environnementaux peuvent être également utilisés à la place des phénotypes et correspondent le plus souvent à des paramètres d'habitat (*i.e.* lieu ou hôte d'isolement d'une souche) ou à des données cliniques. Les deux principales difficultés dans ce type d'analyse sont d'avoir un échantillonnage non biaisé (*i.e.* une distribution équilibrée des échantillons entre les phénotypes) et de tenir compte de la stratification de la population qui est généralement importante pour les espèces bactériennes. Plusieurs méthodes ont ainsi été développées utilisant, pour certaines, un modèle linéaire mixte, un positionnement multidimensionnel ("MultiDimensional Scaling", MDS) ou une régression phylogénétique pour contrôler les liens de proximité au sein des échantillons en capturant la structure fine des populations [106–108]. En entrée de ces méthodes, plusieurs types de données peuvent être utilisés de manière indépendante pour représenter la diversité génomique : (i) des présences/absences de familles de gènes (ii) des variations nucléotidiques ("Single Nucleotide Polymorphisms", SNPs) à l'image de ce qui est fait en génétique humaine [109] (iii) des comptages de k-mers. Plus récemment, la méthode DBGWAS [76] représente un pangénome au niveau des séquences sous la forme d'un graphe de De Bruijn compacté (cDBG) et associe les nœuds du graphe, appelés unitigs, à des phénotypes avec le même type de méthodes statistiques que celles évoquées précédemment. L'avantage de DBGWAS est donc d'éviter la redondance des méthodes basées sur les k-mers et d'offrir une représentation en sous-graphes de régions génomiques contenant des unitigs associés au phénotype. Ces sous-graphes permettent, ainsi, de repérer plusieurs profils de mutations pour un même gène ou des régions variables contenant des gènes spécifiques à certaines souches. Un inconvénient de DBGWAS réside dans le fait que plusieurs unitigs peuvent au final correspondre à plusieurs variants alléliques d'un même gène dont seule la présence est d'importance pour le phénotype observé ; il en est de même pour une région génomique

qui va se retrouver découpée en de multiples chemins si elle contient du polymorphisme. Chacun de ces unitigs sera considéré indépendamment dans le test d'association entraînant une perte de puissance statistique.

Ainsi, au travers du graphe de pangénome de PPanGGOLiN, nous souhaitons conduire des études d'associations tout d'abord au niveau de la présence/absence des familles de gènes puis à l'échelle des variations nucléotidiques si aucun gène ou région génomique ne semble associé au phénotype. Dans ce dernier cas, nous envisageons de construire localement pour chaque famille de gènes un graphe de variants (*i.e.* en utilisant, par exemple, le logiciel vg [110]) qui facilitera l'obtention d'un catalogue de SNPs car la méthode ne nécessite pas de génomes de référence. Concernant l'analyse de phénotypes complexes dits multifactoriels (ou de facteurs environnementaux), plusieurs chemins évolutifs dans l'espèce peuvent leur être associés correspondant potentiellement à des combinaisons différentes de modules au sein des îlots génomiques. Ainsi, il serait intéressant d'agréger les présences de gènes d'un même module dans la méthode statistique pour gagner en puissance (*cf.* technique de "burden testing" utilisée pour l'identification de variants rares en génétique humaine [111]). D'autres améliorations de méthodes d'association seraient d'explorer d'une manière indépendante et/ou complémentaire les différentes parties du pangénome (*i.e. persistent*, *shell* et *cloud*) car elles diffèrent beaucoup dans leur histoire évolutive.

Dans le cadre de la thèse de Guilhem Royer, nous analysons les génomes d'environ 1400 souches d'*E. coli*. Ces souches ont été isolées de bactériémies (en 2005 [112] et en 2017 du projet Septicoli en cours de publication, respectivement 350 et 545 génomes), d'infections ou de colonisations pulmonaires en réanimation (210 génomes) [113], ainsi que de portage rectale chez des patients sains (280 génomes) [114]. Ainsi, nous explorons leur pangénome à la recherche de déterminants bactériens associés à des données cliniques particulières (e.g. porte d'entrée des bactériémies ou sévérité des infections). Les études réalisées jusqu'alors se sont bien souvent limitées à la recherche de facteurs de virulence spécifiques et n'ont pas permis d'identifier de facteur génétique unique. En effet, pour l'heure seul le statut de l'hôte (*e.g.* âge, comorbidités) apparaît lié à la sévérité des infections [112]. C'est pourquoi un focus particulier sera fait sur l'analyse du réseau métabolique de l'espèce reconstruit directement à partir du pangénome dans le but d'identifier des fonctions enzymatiques potentiellement importantes dans la pathologie humaine et impliquées, par exemple, dans le catabolisme ou la biosynthèse de métabolites secondaires. Cette étude du pan-métabolisme d'*E. coli* poursuivra les travaux de thèse de Gilles Vieira [31].

A l'échelle d'un écosystème, nous envisageons de construire un graphe de pangénome pour chaque espèce présente à partir de génomes d'isolats et de MAGs. Nous développerons ensuite une stratégie d'alignement des lectures métagénomiques sur les familles de gènes. Cette stratégie devra

être rapide tout en gardant une spécificité importante. Une approche, similaire à celle proposée dans le logiciel MetAlign [115], pourrait être utilisée. Une première étape consistera à identifier les espèces présentes à partir d'un algorithme de type MinHash permettant de compresser les k-mers représentatifs des familles de gènes *persistent* de chaque pangénome et des lectures pour les comparer et obtenir un indice d'inclusion ("containment index", méthode décrite dans [116]). Une deuxième étape consistera à aligner les lectures sur toutes les familles des pangénomes des espèces identifiées en utilisant potentiellement des techniques d'indexation de k-mers qui auraient été identifiés comme représentatifs et spécifiques d'une famille. A la fin de cette étape, chaque famille sera associée à une abondance correspondant au nombre de lectures qui s'alignent sur la famille. La distribution de ces données de comptage sur l'intégralité des familles du *persistent* (*i.e.* plusieurs milliers de gènes marqueurs) devrait ainsi permettre de quantifier finement l'abondance de l'espèce. Ensuite, les données de comptage des familles du génome *shell* et *cloud* pourraient servir à déterminer le nombre de souches présentes pour chaque espèce et leur contenu en gènes variables. Une méthode statistique de déconvolution basée sur un modèle de mélange guidé par la topologie du graphe de pangénome serait à imaginer en s'inspirant, par exemple, de la méthode metaMix [117].

La principale limite de cette approche basée sur des pangénomes de référence est d'avoir un catalogue de génomes suffisamment représentatif de la diversité de l'écosystème étudié. Les progrès dans les techniques d'assemblage et de classification ("binning") permettent maintenant de reconstruire des milliers de génomes (MAGs) à partir de lectures métagénomiques. Ainsi, un catalogue unifié de de 280 000 génomes du microbiote intestinal humain a été constitué [53] et plus de 87% des lectures métagénomiques [83] s'alignent sur ces génomes de référence ce qui est bon indicateur pour la faisabilité de l'approche proposée. Pour améliorer ces catalogues de génomes, notamment sur leur contenu en *shell* et *cloud*, les graphes de pangénome de PPanGGOLiN pourraient servir de support pour classifier et aligner partiellement des contigs ou des lectures longues (PacBio ou ONT) sur le graphe du *persistent* et ainsi enrichir les pangénomes de nouveaux chemins.

Ces données de comptage sur les familles de gènes et l'identification des différentes souches présentes serviront d'entrées pour des méthodes de métagénomique d'association permettant d'identifier des espèces, souches et gènes dont l'abondance est corrélée avec des facteurs environnementaux. En parallèle, la dynamique des transferts horizontaux intra- et inter-espèces pourra être étudiée au travers des îlots génomiques et des modules prédits. Finalement, en suivant l'air du temps, des méthodes d'apprentissage de type réseaux de neurones convolutifs seraient applicables en considérant le graphe de pangénome comme une image de trois couleurs (représentant le *persistent*, *shell* et *cloud*) où le voisinage des pixels correspond à celui des familles dans le graphe et leur intensité aux valeurs d'abondance dans chaque échantillon métagénomique. Ce raccourci imagé n'est pas si trivial car les réseaux convolutifs traitent classiquement en entrée des signaux définis

spatialement par des grilles cartésiennes (typiquement des sons et des images) or ce n'est pas le cas pour les graphes, un nœud pouvant avoir de multiples voisins. L'application des réseaux convolutifs sur des structures en graphe est un sujet de recherche encore très actif et avec de nombreuses applications [118]. De plus, pour appliquer ce type de méthodes d'apprentissage, de nombreux échantillons métagénomiques sont nécessaires et doivent être associés à des labels correspondant à des facteurs environnementaux. Le projet du million de microbiomes humains[6] devrait aller dans ce sens.

I.7 Représentation pangénomique dans la plateforme MicroScope

La plateforme MicroScope contient actuellement plus de 12 000 génomes. Un atout de la plateforme sont les fonctionnalités de génomique comparée dont notamment le calcul de synténies conservées. Pour réaliser cela, nous comparons les protéomes deux à deux et dans les deux sens via le programme BLASTp ce qui fait un total de plus 100 millions de comparaisons réalisées. Même si le temps d'exécution pour réaliser ces comparaisons reste surmontable car les calculs sont faits au fur et à mesure suivant l'intégration de nouveaux génomes, le stockage et l'indexation des résultats deviennent problématiques en termes de volumétrie (*i.e.* plus de 45 To de données indexées dans une base de données relationnelle). Une projection simple supposant que l'espace disque utilisé augmente linéairement avec le nombre de comparaisons nous amènerait à plus de 100 Po de données pour la comparaison de 500 000 génomes ce qui n'est pas envisageable étant donné les technologies de stockage sur le marché et notre budget. Ainsi, nous devons rapidement envisager de nouvelles solutions.

La représentation pangénomique proposée par PPanGGOliN a donc pour objectif d'être intégrée dans le modèle de données de la plateforme et permettra de réaliser non plus des comparaisons de génomes deux à deux, mais des comparaisons génomes contre pangénomes diminuant ainsi grandement le nombre de comparaisons à réaliser. Une étape préalable était de définir des groupes de génomes supposés appartenir à une même espèce, nommées MICGCs (pour "MicroScope Genome Clusters") [14]. Les MICGCs sont calculés à partir de distances génomiques déterminées avec le logiciel MASH [96] suivi d'une recherche de communauté dans un graphe pondéré par ces distances avec la méthode Louvain [119]. Pour chaque MICGC, un graphe de pangénome est ensuite construit et mis à jour si de nouveaux génomes intégrés appartiennent au MICGC. Actuellement, nous travaillons sur la constitution de trois sets de familles de gènes homologues qui sont définis suivant des valeurs seuils de similarité en acides aminées : 80%, 50% et 30%. Ces familles permettront d'établir des relations de similarité entre les gènes des génomes et les

[6] https://en.mgitech.cn/news/96/

familles de gènes des pangénomes. A partir de ces relations, des synténies conservées seront recherchées entre les génomes et les pangénomes. L'algorithme de recherche de synténie utilisé jusqu'à maintenant [9] devra sans doute être adapté. Une des difficultés majeures réside dans la constitution des familles qui dans l'idéal devraient avoir le même niveau de sensibilité que les alignements BLASTp des protéines deux à deux. De plus, nous devons prendre en compte les événements de fusions et fissions de gènes (*e.g.* cas des pseudogènes, *cf.* section I.6.1) qui sont très fréquents et correspondent à des alignements partiels. Différentes stratégies reposant sur les résultats du logiciel MMseqs2 de constitution de familles [99] sont en cours d'évaluation.

Une fois cette stratégie établie, son intégration dans MicroScope demandera plusieurs modifications dans le modèle de données, la représentation des résultats (*e.g.* cartes de synténie) et dans d'autres outils utilisant les résultats de comparaison de génomes (e.g. recherche de profils phylogénétiques). Une conséquence majeure sera de ne plus fournir à l'utilisateur de valeurs précises d'identité de séquence entre paires d'homologue ni de positions d'alignement. Néanmoins, un calcul dynamique d'alignement deux à deux pourra être proposé. Ces évolutions sont essentielles pour maintenir le service MicroScope face à l'afflux de données. A plus long terme, l'utilisation de pangénomes comme objet de référence pour la navigation et l'exploration dans MicroScope sera à réfléchir (*cf.* prototype de l'interface Web de panGBank, Figure 4).

II-Découverte de nouvelles familles d'enzymes et exploration de leur diversité fonctionnelle

II.1 Motivation de l'approche

Les microorganismes, au regard de leur histoire évolutive qui s'étend sur plusieurs milliards d'années et de leur capacité d'adaptation à divers environnements pour certains extrêmes, offrent un univers immense mais encore peu exploré de fonctions métaboliques. Les enzymes orchestrent cette diversité chimique du vivant en catalysant les réactions avec une efficacité et une sélectivité qui sont remarquables. Il est à noter que de nombreuses enzymes sont capables de catalyser d'autres réactions en plus de celles physiologiquement connues [120]. En effet, deux types de promiscuité peuvent être définis pour les enzymes : la promiscuité de substrats (*i.e.* catalyse de la même transformation mais sur d'autres substrats avec une plus ou moins bonne efficacité) ou la promiscuité de réactions (*i.e.* catalyse de transformations différentes). Le rôle combiné de la promiscuité des enzymes et des transferts horizontaux de gènes est d'ailleurs supposé jouer un rôle important dans l'évolution du métabolisme des procaryotes [121].

De nombreuses bases de données regroupent des connaissances, issues de la bibliographie, sur les réactions, les enzymes et les voies métaboliques. En recoupant les informations contenues dans ces ressources, on constate qu'une proportion importante d'activités enzymatiques sont orphelines de séquences de protéines identifiées comme catalysant ces réactions (appelées "orphan enzymes"). Cette proportion a été estimée à 22% (n=1143) dans notre étude de 2013 [35] basée sur le contenu de la ressource Enzyme[7], qui attribue un numéro EC à chaque nouvelle activité expérimentalement démontrée, et sur les annotations des protéines d'UniProtKB [47]. Une mise à jour de cette étude, réalisée en mai 2018, montre que le pourcentage d'activités orphelines reste stable autour de 20% (n=1248) mais que le nombre d'activités a augmenté de plus de 1000 entrées EC (5096 en 2013 contre 6176 en 2018). Cette augmentation est due pour les trois-quarts à des activités qui avaient été caractérisées avant 2013 mais qui étaient non assignées à un numéro EC, et pour le quart restant à des activités découvertes à partir de 2013. Ceci illustre l'important retard que les bases de données en biologie peuvent avoir dans des connaissances pourtant primordiales car basées sur des preuves expérimentales. Quand on observe la distribution du nombre d'activités enzymatiques nouvellement découvertes au cours du temps (Figure 5, courbe rouge), deux périodes se dessinent : (i) une première

[7] base de donnée maintenue par l'*Enzyme Commission* de l'IUBMB ("International Union of Biochemistry and Molecular Biology"), https://www.qmul.ac.uk/sbcs/iubmb/enzyme

de 20 ans qui a débuté dans les années 1950 et correspond à l'âge d'or de la biochimie où de nombreux laboratoires menaient des activités de recherche dans ce domaine (ii) une deuxième période qui a débuté dans les années 1980 et semble se poursuivre de nos jours. Le début de cette deuxième période coïncide avec l'émergence des technologies de biologie moléculaire (*e.g.* séquençage de l'ADN, clonage) qui ont permis d'identifier de nombreuses séquences d'enzymes tout en caractérisant de nouvelles activités. Néanmoins, l'arrivée des technologies de séquençage massif au cours du XXI^{e} siècle ne semble pas accélérer la découverte de nouvelles activités. La chute observée des courbes de découverte d'activités (à partir de 2011) et de caractérisation de séquences (à partir de 2006) vient probablement du retard dans les connaissances encore présent dans les bases de données.

Face à ce constat et à l'accumulation des données de séquençage, nous avons pour objectif de faire progresser les connaissances dans le métabolisme des microorganismes en exploitant au mieux les données de génomique qui donnent accès à une diversité grandissante de séquences. Ces découvertes dans la chimie du vivant ouvriront de nombreuses applications dans différents domaines que ce soit en écologie, bioremédiation, valorisation de la biomasse ou en biocatalyse.

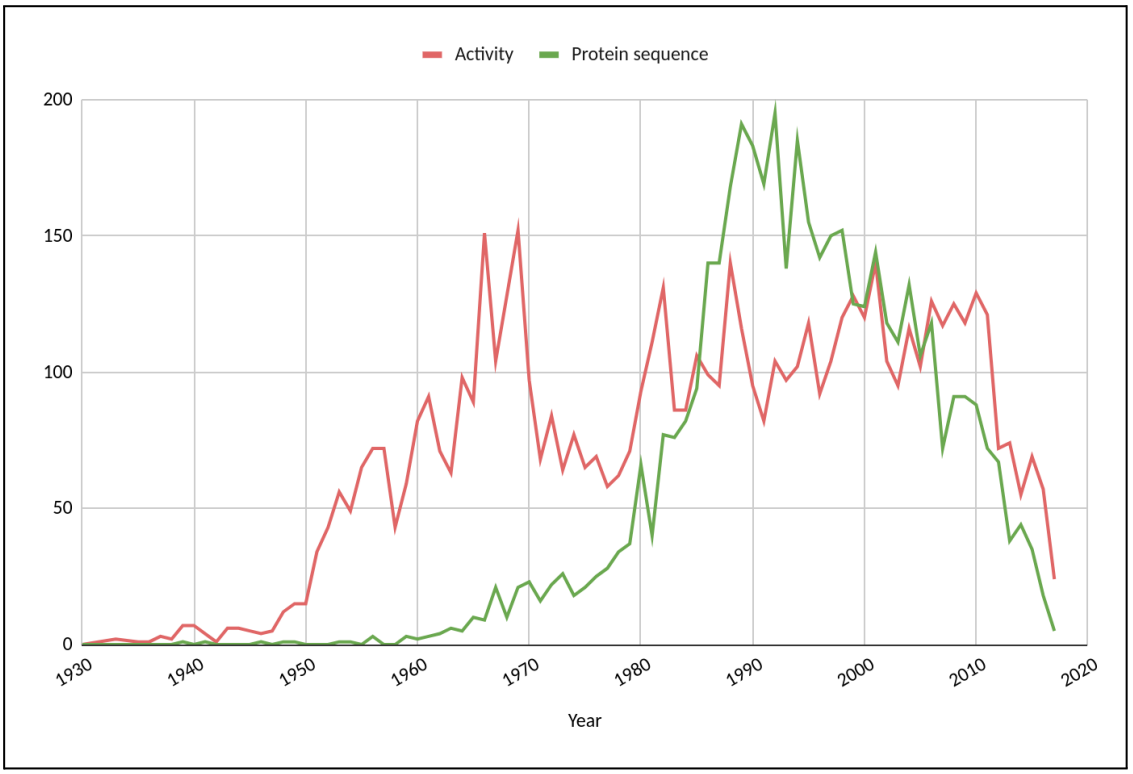

Figure 5 : Découverte d'activités enzymatiques au cours du temps.
La courbe rouge indique le nombre d'activités enzymatiques par année de découverte. Les activités enzymatiques correspondent à celles de la classification EC. La courbe verte indique le nombre d'activités enzymatiques associées pour la première fois à une séquence de protéine pour chaque année. Ces données ont été extraites des bases de données IntEnz [122] et UniProtKB [47] à la date du 15 mai 2018. L'année d'association d'une activité à une séquence de protéine est estimée à partir des données bibliographiques d'UniProtKB et correspond à l'année de publication de l'article le plus ancien associé à une protéine annotée avec le numéro EC correspondant. Seuls les articles associés à un maximum de 10 protéines sont utilisés pour ne pas considérer des publications portant sur l'analyse de grandes régions génomiques et contenant que très rarement des validations expérimentales.

II.2 Stratégie computationnelle et expérimentale intégrée

Depuis plusieurs années, nous développons une stratégie intégrée au sein de notre UMR qui combine des approches computationnelles et expérimentales pour explorer la diversité enzymatique des séquences de protéines issues des projets de génomique. Basée sur une étude pilote que nous avons publiée en 2014 [45] (*cf.* chapitre 1 section III), cette stratégie s'articule sur deux axes (*cf.* Figure 6).

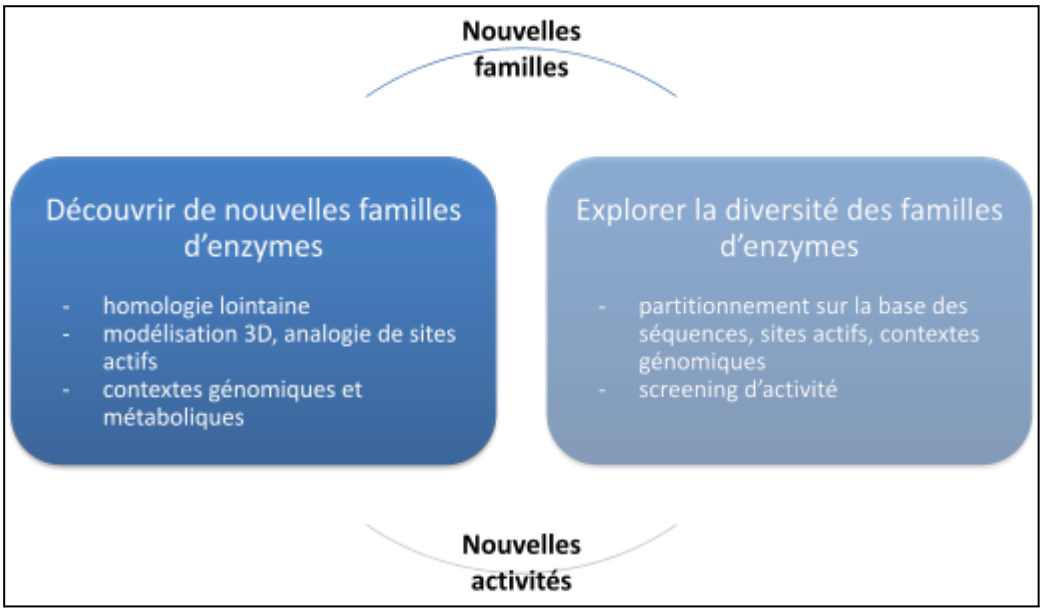

Figure 6 : Illustration d'une stratégie combinant des approches computationnelles et expérimentales pour l'exploration de la diversité fonctionnelle de familles d'enzymes.

Le premier axe consiste à identifier des familles de protéines de fonction inconnue qui seraient potentiellement de nouvelles familles d'enzymes. Pour cela, plusieurs types de méthodes sont utilisés :

- des méthodes de contexte génomique (*i.e.* co-localisation chromosomique, profils phylogénétiques, événements de fusion/fission) qui sont en partie détaillées dans cet article de revue pour des applications sur le métabolisme [123]. L'idée est d'associer des gènes de fonction inconnue à des gènes de fonction connue par des preuves indirectes et de supposer que ces gènes participent à un même processus biologique comme une voie métabolique. La méthode CanOE [42], que nous avons développée, intègre en plus la notion de contexte métabolique (*cf.* chapitre 1 section III).

- la détection d'homologues distants par des techniques de comparaison de profils de familles de protéines de type HMM ("Hidden Markov Model") qui sont plus sensibles que des

alignements deux à deux de séquences protéiques [124]. Des relations d'homologie entre des familles de fonction inconnue et des familles d'enzymes peuvent ainsi être inférées et amènent à supposer qu'elles partagent des structures et fonctions similaires.

- la recherche d'analogie de sites actifs par modélisation structurale des familles de protéines. Cette approche a été développée par ma collègue Karine Bastard du LABGeM et consiste à modéliser la structure des protéines appartenant à des familles de fonction inconnue puis de prédire les poches et résidus catalytiques à l'aide la méthode ASMC [46] (*cf.* paragraphe ci-dessous pour une description de cette méthode). L'organisation spatiale des résidus potentiellement catalytiques (i.e. appelée catalophore) est ensuite comparée avec celles d'enzymes connues pour détecter des sites actifs similaires qui permettent de supposer que le mécanisme réactionnel est conservé.

Ces différentes approches proposent ainsi des pistes sur des fonctions enzymatiques qui sont souvent à affiner au travers de bioanalyses plus poussées en utilisant, par exemple, la plateforme MicroScope [14]. L'activité supposée est ensuite validée expérimentalement sur au moins un représentant de la famille avant d'être étudiée d'une manière plus exhaustive comme décrite dans le deuxième axe de la stratégie.

Le deuxième axe de la stratégie vise à explorer la diversité fonctionnelle de familles d'enzymes. L'objectif est d'étudier la promiscuité en substrats d'une famille pour tenter de déterminer les activités physiologiques et les voies métaboliques associées. A partir d'un sous-ensemble représentatif de protéines, les activités enzymatiques sont testées sur toute une gamme de substrats au sein de la plateforme de clonage et criblage de l'UMR pilotée par Véronique de Berardinis. Le choix des substrats à tester est fait en recherchant, dans les bases de données, des composés chimiques dont la structure contient le ou les groupements fonctionnels impliqués dans le mécanisme réactionnel qui est supposé conservé dans la famille. Ce choix de métabolites est souvent limité par la disponibilité commerciale des molécules qui peut être complétée par une synthèse réalisée par l'équipe de chimie (dirigée par Anne Zaparucha) de notre UMR. Pour sélectionner les protéines représentantes à tester, plusieurs méthodes bioinformatiques sont appliquées :

- un réseau de similarité des protéines de la famille est calculé à partir d'alignements des séquences deux à deux. Des composantes connexes de ce réseau sont ensuite extraites suivant un seuil de score d'alignement et des algorithmes de détection de communautés. Des groupes de protéines similaires en séquence sont ainsi constitués. Des méthodes de classification

hiérarchique ou une phylogénie des protéines de la famille peuvent être également considérées.

- la méthode NetSyn[8] regroupe les protéines suivant leur similarité de contexte génomique. Cette méthode est décrite dans l'article de l'étude pilote [45] et sera prochainement soumise à publication avec quelques améliorations notamment sur le calcul du score de conservation de synténie et le temps d'exécution. Brièvement, NetSyn extrait les régions génomiques des gènes correspondant aux protéines de la famille sur une fenêtre contenant n gènes voisins (généralement 5) de part et d'autre des gènes d'intérêt. Des synténies conservées sont ensuite recherchées entre chaque paire de régions génomiques [9]. Puis, un graphe est construit où les nœuds représentent les protéines d'intérêt et les arêtes une conservation de synténie pondérée par un score qui dépend du nombre de gènes en synténie. Un partitionnement de ce graphe est finalement réalisé par des algorithmes de détection de communautés pour définir des groupes de protéines ayant des contextes génomiques similaires.

- la méthode ASMC [46], développée par ma collègue Karine Bastard, réalise dans une première étape un alignement structural des protéines de la famille dont leur structure a été modélisée par homologie (*i.e.* cela nécessite qu'au moins un membre de la famille ait une structure caractérisée expérimentalement). La ou les poches catalytiques potentielles sont ensuite détectées à l'aide du logiciel Fpocket [125]. Une classification hiérarchique des poches est alors réalisée à partir d'un alignement structural de leurs résidus pour définir des groupes de protéines ayant des poches similaires. Au sein de ces poches, ASMC va détecter des résidus conservés dans toute la famille qui correspondent potentiellement aux acides aminés impliqués dans le mécanisme réactionnel (*i.e.* résidus catalytiques) et, également, des résidus spécifiques à chaque groupe de poches qui sont potentiellement impliqués dans la stabilisation des ligands et donc dans la spécificité de substrats.

Au travers de ces méthodes de classification des protéines suivant leur similarité de séquences, de poches catalytiques et de contextes génomiques, nous obtenons ainsi plusieurs partitionnements de la famille. Pour réconcilier ces résultats, une méthode de type *clustering ensemble* [126] est appliquée pour obtenir une classification consensus des protéines d'une famille en groupes supposés iso-fonctionnels. Au moins deux représentants de chaque groupe sont ensuite sélectionnés (*i.e.* suivant des critères de facilité de clonage) pour cribler leurs activités sur la gamme de substrats choisis.

[8] https://github.com/labgem/netsyn

Les résultats de criblage sont ensuite interprétés par : (i) des analyses de modélisation moléculaire impliquant le plus souvent du *docking* de ligands pour identifier les interactions jouant un rôle dans le mécanismes réactionnel et la spécificité de substrat et, ainsi, affiner la classification structurale des poches catalytiques de la famille (ii) des analyses de contexte génomique se focalisant sur la fonction des gènes voisins dans les groupes NetSyn pour essayer d'élucider les activités physiologiques et les voies métaboliques impliquées.

II.3 Perspectives méthodologiques et d'analyses

Dans l'objectif d'améliorer la stratégie présentée ci-dessus et de l'appliquer sur un grand nombre de familles de protéines, plusieurs améliorations méthodologiques et cadres d'analyse peuvent être proposés :

- la constitution d'une base de connaissance sur des familles d'enzymes comprenant des validations expérimentales est indispensable, et servira de socle notamment pour les méthodes de recherche d'homologues lointains et d'analogie de sites actifs. Elle regroupera les informations de plusieurs bases de données concernant notamment : les domaines et familles de protéines, les structures tridimensionnelles, les activités enzymatiques et les réactions et voies métaboliques associées. Cette base de connaissance sera régulièrement enrichie par des données expérimentales produites dans notre UMR ou décrites dans la littérature.

- les méthodes de recherche d'homologues lointains et d'analogie de sites actifs génèrent un nombre potentiellement important de résultats dont la spécificité est difficilement appréciable sans analyses complémentaires. Des résultats de contextes génomiques et métaboliques peuvent ainsi être utiles pour sélectionner parmi les activités prédites celles qui sont les plus pertinentes. La méthode CanOE montrant trop de limites dues à la difficulté d'ancrer des enzymes candidates dans un contexte génomique et métabolique informatif, le graphe de transformations chimiques que nous avons précédemment défini servirait de base pour déterminer des contextes métaboliques plus génériques (*cf.* chapitre 1 section III) [43]. Ainsi, la cohérence globale des prédictions d'activités dans une région génomique pourrait être vérifiée en sélectionnant le ou les ensembles d'activités qui correspondent aux chemins les plus probables dans le graphe de transformations chimiques.

- les avancées en métagénomique autour de la constitution de catalogues de gènes et de la reconstruction de génomes (MAGs) de divers écosystèmes offrent un pan encore inexploré de séquences de protéines. De nouvelles approches, permettant de directement assembler des séquences de protéines sans passer par l'étape de reconstruction de contigs, commencent à voir le jour dont, notamment, la méthode Plass qui a permis aux auteurs d'obtenir, sur des échantillons métagénomiques complexes, 2 à 10 fois plus de séquences de protéines que des approches d'assemblage classiques (*e.g.* 2 milliards de protéines à partir de 640 échantillons du sol) [127]. Ces milliards de séquences permettent ainsi d'augmenter considérablement la diversité des familles de protéines mais, également, d'en constituer de nouvelles qui ne sont pas encore représentées dans les bases de données généralistes contenant majoritairement des données sur des organismes cultivés en laboratoire. Cette volumétrie de nouvelles séquences implique de repenser les méthodes d'analyse présentées ci-dessus notamment pour leur passage à l'échelle.

A court terme, une partie de ces avancées méthodologiques sera réalisée dans le cadre du projet MODAMDH (ANR JCJC de Carine Vergne, collègue chimiste de l'UMR) qui vise à cribler la biodiversité, notamment issue des données de métagénomique, à la recherche de nouvelles séquences d'amines déshydrogénases pour des applications en biocatalyse (*cf.* Figure 7). Pour ce projet, les travaux en bioinformatique seront réalisés dans notre laboratoire notamment via le recrutement d'un postdoc en modélisation moléculaire.

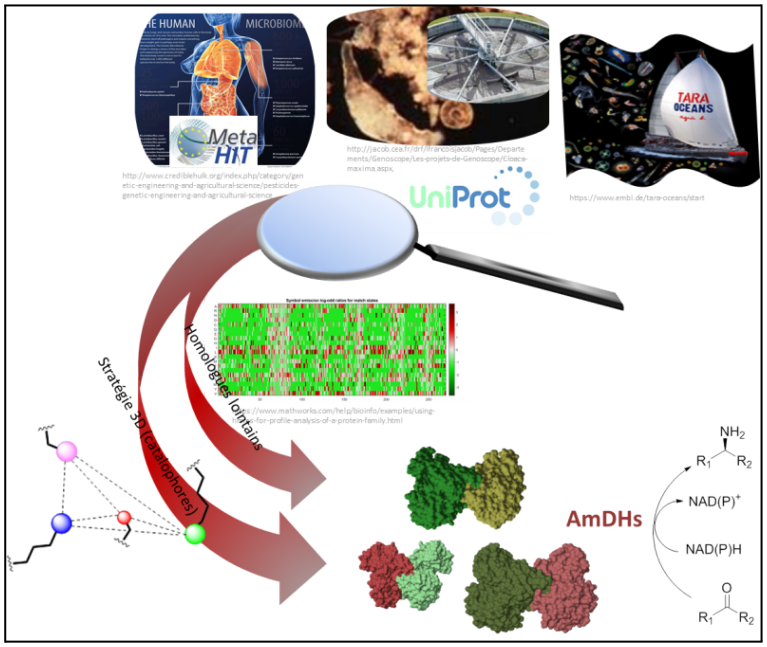

Figure 7 : Illustration du projet MODAMDH, "*In silico* approach for amine dehydrogenase discovery".

Plusieurs autres pistes seraient à explorer dans l'objectif de parfaire nos connaissances du métabolisme des microorganismes.

Tout d'abord, les techniques expérimentales que nous utilisons pour le criblage d'activités enzymatiques testent unitairement par puits une protéine en confrontation d'un substrat et nécessitent de présupposer de la réaction. Des approches non ciblées à haut débit pourraient être envisagées dont, notamment, celle décrite dans la publication de Sévin *et al.* [128] qui confronte individuellement des protéines surexprimées ou purifiées à des centaines de métabolites issus d'un extrait cellulaire de microorganisme (*i.e.* celui d'*E. coli* dans l'étude citée). Grâce à la précision des techniques de spectrométrie de masse en métabolomique, des accumulations et déplétions de métabolites peuvent être observées donnant ainsi des indications sur l'activité enzymatique de la protéine testée. En plus de ce type d'analyse, l'apport de la métabolomique est d'une importance majeure car elle révèle un univers complexe de métabolites à l'échelle d'une cellule dont on ne connaît pas l'origine métabolique (*i.e.* les activités enzymatiques qui les produisent) ni très souvent leur structure chimique. La caractérisation de ces métabolites orphelins est un sujet de recherche à part entière qui nécessite des méthodes chémoinformatiques [129] et des expérimentations analytiques poussées [130].

Une étude systématique des voies métaboliques supposées universellement conservées dans le vivant car essentielles au métabolisme de base des cellules (*e.g.* les voies de biosynthèse des acides aminés, de cofacteurs, de nucléosides ou les voies de production d'énergie) serait intéressante à conduire à l'échelle des centaines de milliers de génomes disponibles qui représentent des espèces de plus en plus diversifiées au sein de l'arbre de la vie. Ce type d'analyse a un intérêt dans la compréhension de l'évolution du métabolisme et peut révéler des surprises comme en témoigne notre étude sur le métabolisme de la méthionine qui a montré une double convergence fonctionnelle pour deux familles d'enzymes impliquées dans une même étape de la voie de biosynthèse [131]. La méthode GROOLS [34] (*cf.* chapitre 1 section II) pourrait être systématiquement appliquée pour valider les annotations des génomes au regard de nos connaissances actuelles sur ces voies. Les incohérences détectées nous amèneraient à envisager l'existence de voies ou familles d'enzymes alternatives ou, plus simplement, à améliorer la sensibilité et spécificité des prédicteurs de fonctions des gènes.

Une dernière piste à envisager serait d'analyser des voies métaboliques d'importance à l'échelle d'écosystèmes. Des analyses métagénomiques et pangénomiques (cf. section I) seraient conduites pour déterminer la présence ou non de voies métaboliques dans un écosystème, puis pour

identifier et quantifier les espèces et familles d'enzymes impliquées. Des prédictions partielles indiqueraient qu'une voie alternative est à élucider, et des abondances variables seraient le signe d'une possible coopération métabolique entre plusieurs espèces. Ces voies métaboliques avec leurs abondances et espèces associées seraient ensuite à analyser en association avec des facteurs environnementaux.

Conclusion et perspectives

Ce mémoire clôture 20 années d'expériences trépidantes en bioinformatique. Ma recherche s'articule au travers de nombreux échanges avec des collègues mathématiciens, informaticiens, chimistes et biologistes mêlant ainsi une interdisciplinarité forte qui est primordiale pour répondre aux nouveaux défis en biologie, sans oublier le rôle des ingénieurs qui sont des moteurs essentiels dans la conception et la mise en oeuvre des méthodes et analyses.

L'analyse de données métagénomiques par une approche pangénomique associée à une stratégie hybride combinant des méthodes et expérimentations pour l'exploration du métabolisme devraient gagner en synergie. Elles offriront de nouvelles pistes pour l'étude d'écosystèmes en termes de processus métaboliques et d'interactions entre microorganismes. Les progrès technologiques dans le séquençage par nanopore de longues lectures permettront d'assembler davantage de génomes à partir d'échantillons environnementaux et donc d'enrichir les pangénomes de référence en nombre de souches et d'espèces représentées. L'identification et la quantification de métabolites par des tests biochimiques ou des analyses métabolomiques seraient, également, un atout majeur pour identifier des processus métaboliques d'importance dans un écosystème.

Comme évoqué précédemment dans ce manuscrit, l'acquisition de connaissances dans la diversité génomique des écosystèmes et le métabolisme des microorganismes offre de nombreuses applications concernant le biocontrôle, la bioremédiation, la valorisation de la biomasse et la biocatalyse. Dans l'optique de repousser les limites de la biodiversité, des recherches en biologie de synthèse s'accentuent depuis une dizaine d'années. Leur but est de conceptualiser la fabrication de systèmes biologiques pour un objectif défini qui n'existe pas naturellement ou est peu efficace. Cela peut se faire à l'échelle d'une protéine pour la conception *de novo* d'une séquence en acides aminés correspondant, par exemple, à une enzyme ayant l'activité souhaitée ou, encore, à l'échelle d'une cellule en modifiant les voies métaboliques d'un microorganisme pour, par exemple, fixer le CO_2 atmosphérique. Des connaissances pointues sur le métabolisme de l'organisme châssis utilisé sont nécessaires pour mener à bien ces conceptions ainsi que le développement de méthodes pour les améliorer. Ces méthodes sont basées, pour certaines, sur de l'apprentissage par renforcement utilisant des simulations ou des résultats expérimentaux comme données d'entraînement [132,133]. D'autres approches que l'on pourrait qualifier d'apprentissage naturel consistent à faire de l'évolution dirigée à long terme de populations de cellules soumises à des pressions de sélection pour obtenir le phénotype

attendu. Cette piste est explorée au Genoscope notamment via des expérimentations utilisant un automate de culture continue [134], nommé GM3, pour lesquelles nous réalisons les analyses bioinformatiques pour identifier et comprendre les mutations à l'origine du phénotype.

Concernant la plateforme MicroScope, le service rendu à la communauté est indéniable et apporte une certaine satisfaction à l'équipe. Nous nous efforçons d'innover en proposant de nouvelles fonctionnalités comme, par exemple, l'intégration récente de la méthode panRGP pour la prédiction d'îlots génomiques à partir de pangénomes. Néanmoins, maintenir un tel service avec des ressources humaines précaires et des besoins croissants en technicité face à l'afflux de données ne se fait pas avec la plus grande sérénité. La solution de financement des plateformes académiques proposée par les tutelles est de facturer le service sous forme de prestations ce qui est sans doute audible et applicable pour certains types d'activité mais pose la question des moyens à mettre en oeuvre et des coûts additionnels que cela engendre, sans parler de l'impact négatif qu'aurait la restriction d'accès à un service d'analyse bioinformatique à l'ère de l'*open data*. Un appel aux dons, comme cela est pratiqué par certains instituts de renom ou par des projets logiciels du monde de l'*open source*, serait peut être plus efficace.

Pour conclure, je compare souvent le métier de chercheur à celui d'un sportif de haut niveau combinant des épreuves individuelles et collectives ainsi que des marathons et des sprints. Puis un jour, on devient entraîneur pour transmettre notre savoir et notre passion mais sans promesse de médailles, quoique certains chercheurs en reçoivent, ni de salaire au niveau de celui d'un joueur de pétanque professionnel pour ne pas parler de football. Mais peut-être qu'un jour viendra où le star-système prôné par certains aura trouvé toute sa place avec des concours du meilleur chercheur diffusés en live sur YouTube pour décrocher un poste tant mérité, en espérant que d'ici là les conséquences d'une politique de sélection dite darwinienne n'aient pas complètement décimé l'écosystème de la recherche. Une dernière petite phrase pour parfaire mon élocution en préparation de la soutenance et mettre un point final à ce mémoire : un chercheur sachant chercher sans chercher de l'argent est un chercheur heureux.

Bibliographie

1. Barbe V, Vallenet D, Fonknechten N, Kreimeyer A, Oztas S, Labarre L, et al. Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium. Nucleic Acids Res. 2004;32: 5766–5779.

2. Fournier P-E, Vallenet D, Barbe V, Audic S, Ogata H, Poirel L, et al. Comparative genomics of multidrug resistance in Acinetobacter baumannii. PLoS Genet. 2006;2: e7.

3. Vallenet D, Nordmann P, Barbe V, Poirel L, Mangenot S, Bataille E, et al. Comparative analysis of Acinetobacters: three genomes for three lifestyles. PLoS One. 2008;3: e1805.

4. De Berardinis V, Vallenet D, Castelli V, Besnard M, Pinet A, Cruaud C, et al. A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. Mol Syst Biol. 2008;4: 174.

5. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, et al. MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res. 2006;34: 53–65.

6. Médigue C, Rechenmann F, Danchin A, Viari A. Imagene: an integrated computer environment for sequence annotation and analysis. Bioinformatics. 1999;15: 2–15. doi:10.1093/bioinformatics/15.1.2

7. Bocs S, Cruveiller S, Vallenet D, Nuel G, Médigue C. AMIGene: Annotation of MIcrobial Genes. Nucleic Acids Res. 2003;31: 3723–3726. doi:10.1093/nar/gkg590

8. Cruveiller S, Le Saux J, Vallenet D, Lajus A, Bocs S, Médigue C. MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. Nucleic Acids Res. 2005;33: W471–9. doi:10.1093/nar/gki498

9. Boyer F, Morgat A, Labarre L, Pothier J, Viari A. Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. Bioinformatics. 2005;21: 4209–4215. doi:10.1093/bioinformatics/bti711

10. Durand P, Labarre L, Meil A, Divo J-L, Vandenbrouck Y, Viari A, et al. GenoLink: a graph-based querying and browsing system for investigating the function of genes and proteins. BMC Bioinformatics. 2006;7: 21. doi:10.1186/1471-2105-7-21

11. Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, Lajus A, et al. MicroScope: a platform for microbial genome annotation and comparative genomics. Database. 2009;2009.

12. Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S, Lajus A, et al. MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Res. 2012;41: D636–D647.

13. Vallenet D, Calteau A, Cruveiller S, Gachet M, Lajus A, Josso A, et al. MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. Nucleic Acids Res. 2016;45: D517–D528.

14. Vallenet D, Calteau A, Dubois M, Amours P, Bazin A, Beuvin M, et al. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. Nucleic Acids Res. 2020;48: D579–D589. doi:10.1093/nar/gkz926

15. Alloisio N, Queiroux C, Fournier P, Pujic P, Normand P, Vallenet D, et al. The Frankia alni symbiotic transcriptome. Mol Plant Microbe Interact. 2010;23: 593–607.

16. Bonaldi K, Gourion B, Fardoux J, Hannibal L, Cartieaux F, Boursot M, et al. Large-scale transposon

mutagenesis of photosynthetic Bradyrhizobium sp. strain ORS278 reveals new genetic loci putatively important for nod-independent symbiosis with Aeschynomene indica. Mol Plant Microbe Interact. 2010;23: 760–770.

17. Giraud E, Moulin L, Vallenet D, Barbe V, Cytryn E, Avarre J-C, et al. Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia. Science. 2007;316: 1307–1312.

18. Mornico D, Miché L, Béna G, Nouwen N, Verméglio A, Vallenet D, et al. Comparative genomics of Aeschynomene symbionts: insights into the ecological lifestyle of nod-independent photosynthetic bradyrhizobia. Genes . 2011;3: 35–61.

19. Normand P, Lapierre P, Tisa LS, Gogarten JP, Alloisio N, Bagnarol E, et al. Genome characteristics of facultatively symbiotic Frankia sp. strains reflect host range and host plant biogeography. Genome Res. 2007;17: 7–15.

20. Sugawara M, Epstein B, Badgley BD, Unno T, Xu L, Reese J, et al. Comparative genomics of the core and accessory genomes of 48 Sinorhizobium strains comprising five genospecies. Genome Biol. 2013;14: R17. doi:10.1186/gb-2013-14-2-r17

21. Bertin PN, Heinrich-Salmeron A, Pelletier E, Goulhen-Chollet F, Arsène-Ploetze F, Gallien S, et al. Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta-and proteo-genomics. ISME J. 2011;5: 1735.

22. Andres J, Arsène-Ploetze F, Barbe V, Brochier-Armanet C, Cleiss-Arnold J, Coppée J-Y, et al. Life in an arsenic-containing gold mine: genome and physiology of the autotrophic arsenite-oxidizing bacterium rhizobium sp. NT-26. Genome Biol Evol. 2013;5: 934–953. doi:10.1093/gbe/evt061

23. Muller D, Médigue C, Koechler S, Barbe V, Barakat M, Talla E, et al. A tale of two oxidation states: bacterial colonization of arsenic-rich environments. PLoS Genet. 2007;3: e53.

24. Ficko-Blean E, Préchoux A, Thomas F, Rochat T, Larocque R, Zhu Y, et al. Carrageenan catabolism is encoded by a complex regulon in marine heterotrophic bacteria. Nat Commun. 2017;8: 1685. doi:10.1038/s41467-017-01832-6

25. Monteil CL, Vallenet D, Menguy N, Benzerara K, Barbe V, Fouteau S, et al. Ectosymbiotic bacteria at the origin of magnetoreception in a marine protist. Nat Microbiol. 2019;4: 1088–1095. doi:10.1038/s41564-019-0432-7

26. Durot M, Le Fèvre F, de Berardinis V, Kreimeyer A, Vallenet D, Combe C, et al. Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data. BMC Syst Biol. 2008;2: 85.

27. Barbe V, Cruveiller S, Kunst F, Lenoble P, Meurice G, Sekowska A, et al. From a consortium sequence to a unified sequence: the Bacillus subtilis 168 reference genome a decade later. Microbiology. 2009;155: 1758–1775.

28. Belda E, Sekowska A, Le Fèvre F, Morgat A, Mornico D, Ouzounis C, et al. An updated metabolic view of the Bacillus subtilis 168 genome. Microbiology. 2013;159: 757–770.

29. Borriss R, Danchin A, Harwood CR, Médigue C, Rocha EPC, Sekowska A, et al. Bacillus subtilis, the model Gram-positive bacterium: 20 years of annotation refinement. Microb Biotechnol. 2018;11: 3–17.

30. Belda E, van Heck RGA, José Lopez-Sanchez M, Cruveiller S, Barbe V, Fraser C, et al. The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic chassis. Environ Microbiol. 2016;18: 3403–3424. doi:10.1111/1462-2920.13230

31. Vieira G, Sabarly V, Bourguignon P-Y, Durot M, Le Fèvre F, Mornico D, et al. Core and panmetabolism in Escherichia coli. J Bacteriol. 2011;193: 1461–1472.

32. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of

computational protein function prediction. Nat Methods. 2013;10: 221–227. doi:10.1038/nmeth.2340

33. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009;5: e1000605. doi:10.1371/journal.pcbi.1000605

34. Mercier J, Josso A, Médigue C, Vallenet D. GROOLS: reactive graph reasoning for genome annotation through biological processes. BMC Bioinformatics. 2018;19: 132. doi:10.1186/s12859-018-2126-1

35. Sorokina M, Stam M, Médigue C, Lespinet O, Vallenet D. Profiling the orphan enzymes. Biol Direct. 2014;9: 10.

36. Engelen S, Vallenet D, Médigue C, Danchin A. Distinct co-evolution patterns of genes associated to DNA polymerase III DnaE and PolC. BMC Genomics. 2012;13: 69.

37. Belda E, Vallenet D, Médigue C. Accurate Microbial Genome Annotation Using an Integrated and User-Friendly Environment for Community Expertise of Gene Functions: The MicroScope Platform. In: McGenity TJ, Timmis KN, Nogales B, editors. Hydrocarbon and Lipid Microbiology Protocols. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. pp. 141–169. doi:10.1007/8623_2015_179

38. Kreimeyer A, Perret A, Lechaplais C, Vallenet D, Médigue C, Salanoubat M, et al. Identification of the last unknown genes in the fermentation pathway of lysine. J Biol Chem. 2007;282: 7191–7197.

39. Fonknechten N, Perret A, Perchat N, Tricot S, Lechaplais C, Vallenet D, et al. A conserved gene cluster rules anaerobic oxidative degradation of L-ornithine. J Bacteriol. 2009;191: 3162–3167.

40. Pelletier E, Kreimeyer A, Bocs S, Rouy Z, Gyapay G, Chouari R, et al. “Candidatus Cloacamonas acidaminovorans”: genome sequence reconstruction provides a first glimpse of a new bacterial division. J Bacteriol. 2008;190: 2572–2579. doi:10.1128/JB.01248-07

41. Perret A, Lechaplais C, Tricot S, Perchat N, Vergne C, Pellé C, et al. A novel acyl-CoA beta-transaminase characterized from a metagenome. PLoS One. 2011;6: e22918.

42. Smith AAT, Belda E, Viari A, Medigue C, Vallenet D. The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. PLoS Comput Biol. 2012;8: e1002540.

43. Sorokina M, Medigue C, Vallenet D. A new network representation of the metabolism to detect chemical transformation modules. BMC Bioinformatics. 2015;16: 385.

44. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44: D279–85. doi:10.1093/nar/gkv1344

45. Bastard K, Smith AAT, Vergne-Vaxelaire C, Perret A, Zaparucha A, De Melo-Minardi R, et al. Revealing the hidden functional diversity of an enzyme family. Nat Chem Biol. 2014;10: 42.

46. de Melo-Minardi RC, Bastard K, Artiguenave F. Identification of subfamily-specific sites based on active sites modeling and clustering. Bioinformatics. 2010;26: 3075–3082. doi:10.1093/bioinformatics/btq595

47. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47: D506–D515. doi:10.1093/nar/gky1049

48. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464: 59–65. doi:10.1038/nature08821

49. Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, et al. A holistic approach to marine eco-systems biology. PLoS Biol. 2011;9: e1001177. doi:10.1371/journal.pbio.1001177

50. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7: e7359.

doi:10.7717/peerj.7359

51. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11: 1144–1146. doi:10.1038/nmeth.3103

52. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3: e1319. doi:10.7717/peerj.1319

53. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome. Microbiology. bioRxiv; 2019. p. 48.

54. Lasken RS, McLean JS. Recent advances in genomic DNA sequencing of microbial species from single cells. Nat Rev Genet. 2014;15: 577–584. doi:10.1038/nrg3785

55. Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and Complete Genomes from Metagenomes. Microbiology. bioRxiv; 2019. p. 4.

56. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005;15: 589–594. doi:10.1016/j.gde.2005.09.006

57. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A. 2005;102: 13950–13955. doi:10.1073/pnas.0506758102

58. Campbell A. Evolutionary significance of accessory DNA elements in bacteria. Annu Rev Microbiol. 1981;35: 55–83. doi:10.1146/annurev.mi.35.100181.000415

59. Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 2011;7: e1001284. doi:10.1371/journal.pgen.1001284

60. Cohan FM. Genetic exchange and evolutionary divergence in prokaryotes. Trends Ecol Evol. 1994;9: 175–180. doi:10.1016/0169-5347(94)90081-7

61. Levin BR. Periodic selection, infectious gene exchange and the genetic structure of E. coli populations. Genetics. 1981;99: 1–23. Available: https://www.ncbi.nlm.nih.gov/pubmed/7042453

62. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol. 2001;55: 709–742. doi:10.1146/annurev.micro.55.1.709

63. Zekic T, Holley G, Stoye J. Pan-Genome Storage and Analysis Techniques. Methods Mol Biol. 2018;1704: 29–53. doi:10.1007/978-1-4939-7463-4_2

64. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced Escherichia coli genomes. Microb Ecol. 2010;60: 708–720. doi:10.1007/s00248-010-9717-3

65. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol. 2008;190: 6881–6893. doi:10.1128/JB.00619-08

66. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009;5: e1000344. doi:10.1371/journal.pgen.1000344

67. Morris JJ, Lenski RE, Zinser ER. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio. 2012;3. doi:10.1128/mBio.00036-12

68. Acevedo-Rocha CG, Fang G, Schmidt M, Ussery DW, Danchin A. From essential to persistent genes: a functional approach to constructing synthetic life. Trends Genet. 2013;29: 273–279. doi:10.1016/j.tig.2012.11.001

69. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and

robust microbial pangenome analysis. Appl Environ Microbiol. 2013;79: 7696–7701. doi:10.1128/AEM.02411-13

70. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. Trends Genet. 2009;25: 107–110. doi:10.1016/j.tig.2008.12.004

71. Bolotin E, Hershberg R. Horizontally Acquired Genes Are Often Shared between Closely Related Bacterial Species. Front Microbiol. 2017;8: 1536. doi:10.3389/fmicb.2017.01536

72. Vesth T, Wassenaar TM, Hallin PF, Snipen L, Lagesen K, Ussery DW. On the origins of a Vibrio species. Microb Ecol. 2010;59: 1–13. doi:10.1007/s00248-009-9596-7

73. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008;36: 6688–6719. doi:10.1093/nar/gkn668

74. Collins RE, Higgs PG. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol Biol Evol. 2012;29: 3413–3425. doi:10.1093/molbev/mss163

75. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. Brief Bioinform. 2018;19: 118–135. doi:10.1093/bib/bbw089

76. Jaillard M, Lima L, Tournoud M, Mahé P, van Belkum A, Lacroix V, et al. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. PLoS Genet. 2018;14: e1007758. doi:10.1371/journal.pgen.1007758

77. Touchon M, Rocha EPC. Coevolution of the Organization and Structure of Prokaryotic Genomes. Cold Spring Harb Perspect Biol. 2016;8: a018168. doi:10.1101/cshperspect.a018168

78. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 1999;96: 2896–2901. doi:10.1073/pnas.96.6.2896

79. Fang G, Rocha EPC, Danchin A. Persistence drives gene clustering in bacterial genomes. BMC Genomics. 2008;9: 4. doi:10.1186/1471-2164-9-4

80. Oliveira PH, Touchon M, Cury J, Rocha EPC. The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun. 2017;8: 841. doi:10.1038/s41467-017-00808-w

81. Ambroise C, Dang M, Govaert G. Clustering of Spatial Data by the EM Algorithm. geoENV I — Geostatistics for Environmental Applications. 1997. pp. 493–504. doi:10.1007/978-94-017-1675-8_40

82. Mo Dang GG. Spatial Fuzzy Clustering using EM and Markov Random Fields. InternationalJournal of System Research and Information Science. 1998. Available: http://citeseerx.ist.psu.edu/viewdoc/similar?doi=10.1.1.22.9501&type=ab

83. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019;176: 649–662.e20. doi:10.1016/j.cell.2019.01.001

84. Hacker J, Carniel E. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep. 2001;2: 376–381. doi:10.1093/embo-reports/kve097

85. Bertelli C, Tilley KE, Brinkman FSL. Microbial genomic island discovery, visualization and analysis. Brief Bioinform. 2019;20: 1685–1698. doi:10.1093/bib/bby042

86. Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol. 1997;44: 383–397. doi:10.1007/pl00006158

87. Bertelli C, Brinkman FSL. Improved genomic island predictions with IslandPath-DIMOB. Bioinformatics. 2018;34: 2161–2167. doi:10.1093/bioinformatics/bty095

88. Bush EC, Clark AE, DeRanek CA, Eng A, Forman J, Heath K, et al. xenoGI: reconstructing the history of

genomic island insertions in clades of closely related bacteria. BMC Bioinformatics. 2018;19: 32. doi:10.1186/s12859-018-2038-0

89. Poirel L, Menuteau O, Agoli N, Cattoen C, Nordmann P. Outbreak of extended-spectrum beta-lactamase VEB-1-producing isolates of Acinetobacter baumannii in a French hospital. J Clin Microbiol. 2003;41: 3542–3547. doi:10.1128/jcm.41.8.3542-3547.2003

90. Lescat M, Calteau A, Hoede C, Barbe V, Touchon M, Rocha E, et al. 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. Antimicrob Agents Chemother. 2009;53: 2283–2288. doi:10.1128/AAC.00123-09

91. Segond M, Borgelt C. Item Set Mining Based on Cover Similarity. Advances in Knowledge Discovery and Data Mining. 2011. pp. 493–505. doi:10.1007/978-3-642-20847-8_41

92. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47: D607–D613. doi:10.1093/nar/gky1131

93. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44: D733–45. doi:10.1093/nar/gkv1189

94. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36: 996–1004. doi:10.1038/nbt.4229

95. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019. doi:10.1093/bioinformatics/btz848

96. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17: 132. doi:10.1186/s13059-016-0997-x

97. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9: 5114. doi:10.1038/s41467-018-07641-9

98. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25: 1043–1055. doi:10.1101/gr.186072.114

99. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35: 1026–1028. doi:10.1038/nbt.3988

100. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147: 195–197. doi:10.1016/0022-2836(81)90087-5

101. Goodhead I, Darby AC. Taking the pseudo out of pseudogenes. Curr Opin Microbiol. 2015;23: 102–109. doi:10.1016/j.mib.2014.11.012

102. Koonin EV. Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. BMC Biol. 2016;14: 114. doi:10.1186/s12915-016-0338-2

103. Zhou Z, Achtman M. Accurate reconstruction of bacterial pan- and core- genomes with PEPPA. Bioinformatics. bioRxiv; 2020. p. 511.

104. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline. Genomics. bioRxiv; 2020.

105. Dutilh BE, Backus L, Edwards RA, Wels M, Bayjanov JR, van Hijum SAFT. Explaining microbial phenotypes on a genomic scale: GWAS for microbes. Brief Funct Genomics. 2013;12: 366–380.

doi:10.1093/bfgp/elt008

106. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016;1: 16041. doi:10.1038/nmicrobiol.2016.41

107. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics. 2018;34: 4310–4312. doi:10.1093/bioinformatics/bty539

108. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016;17: 238. doi:10.1186/s13059-016-1108-8

109. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2017;18: 41–50. doi:10.1038/nrg.2016.132

110. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36: 875–879. doi:10.1038/nbt.4227

111. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010;34: 188–193. doi:10.1002/gepi.20450

112. Lefort A, Panhard X, Clermont O, Woerther P-L, Branger C, Mentré F, et al. Host factors and portal of entry outweigh bacterial determinants to predict the severity of Escherichia coli bacteremia. J Clin Microbiol. 2011;49: 777–783. doi:10.1128/JCM.01902-10

113. La Combe B, Bleibtreu A, Messika J, Fernandes R, Clermont O, Branger C, et al. Decreased susceptibility to chlorhexidine affects a quarter of Escherichia coli isolates responsible for pneumonia in ICU patients. Intensive Care Med. 2018;44: 531–533. doi:10.1007/s00134-018-5061-8

114. Massot M, Daubié A-S, Clermont O, Jauréguy F, Couffignal C, Dahbi G, et al. Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of Escherichia coli from community subjects in the Paris area in 2010 and evolution over 30 years. Microbiology. 2016;162: 642–650. doi:10.1099/mic.0.000242

115. LaPierre N, Alser M, Eskin E, Koslicki D, Mangul S. Metalign: Efficient alignment-based metagenomic profiling via containment min hash. Bioinformatics. bioRxiv; 2020. p. 97.

116. Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB, et al. Mash Screen: high-throughput sequence containment estimation for genome discovery. Genome Biol. 2019;20: 232. doi:10.1186/s13059-019-1841-x

117. Morfopoulou S, Plagnol V. Bayesian mixture analysis for metagenomic community profiling. Bioinformatics. 2015;31: 2930–2938. doi:10.1093/bioinformatics/btv317

118. Zhang S, Tong H, Xu J, Maciejewski R. Graph convolutional networks: a comprehensive review. Comput Soc Netw. 2019;6: 626. doi:10.1186/s40649-019-0069-y

119. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008. p. P10008. doi:10.1088/1742-5468/2008/10/p10008

120. Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem. 2010;79: 471–505. doi:10.1146/annurev-biochem-030409-143718

121. Glasner ME, Truong DP, Morse BC. How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation. FEBS J. 2019. doi:10.1111/febs.15185

122. Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, et al. IntEnz, the

integrated relational enzyme database. Nucleic Acids Res. 2004;32: D434–7. doi:10.1093/nar/gkh119

123. Hanson AD, Pribat A, Waller JC, Crécy-Lagard V de. “Unknown” proteins and “orphan” enzymes: the missing half of the engineering parts list – and how to find it. Biochem J. 2010;425: 1–11. doi:10.1042/BJ20091328

124. Chen J, Guo M, Wang X, Liu B. A comprehensive review and comparison of different computational methods for protein remote homology detection. Brief Bioinform. 2018;19: 231–244. doi:10.1093/bib/bbw108

125. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10: 168. doi:10.1186/1471-2105-10-168

126. Alqurashi T, Wang W. Clustering ensemble method. Int J Mach Learn & Cyber. 2019;10: 1227–1246. doi:10.1007/s13042-017-0756-7

127. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods. 2019;16: 603–606. doi:10.1038/s41592-019-0437-4

128. Sévin DC, Fuhrer T, Zamboni N, Sauer U. Nontargeted in vitro metabolomics for high-throughput identification of novel enzymes in Escherichia coli. Nat Methods. 2017;14: 187–194. doi:10.1038/nmeth.4103

129. Tsugawa H. Advances in computational metabolomics and databases deepen the understanding of metabolisms. Curr Opin Biotechnol. 2018;54: 10–17. doi:10.1016/j.copbio.2018.01.008

130. Dias DA, Jones OAH, Beale DJ, Boughton BA, Benheim D, Kouremenos KA, et al. Current and Future Perspectives on the Structural Identification of Small Molecules in Biological Systems. Metabolites. 2016;6. doi:10.3390/metabo6040046

131. Bastard K, Perret A, Mariage A, Bessonnet T, Pinet-Turpault A, Petit J-L, et al. Parallel evolution of non-homologous isofunctional enzymes in methionine biosynthesis. Nat Chem Biol. 2017;13: 858.

132. Koch M, Duigou T, Faulon J-L. Reinforcement Learning for Bioretrosynthesis. ACS Synth Biol. 2020;9: 157–168. doi:10.1021/acssynbio.9b00447

133. Schreck JS, Coley CW, Bishop KJM. Learning Retrosynthetic Planning through Simulated Experience. ACS Cent Sci. 2019;5: 970–981. doi:10.1021/acscentsci.9b00055

134. Marlière P, Patrouix J, Döring V, Herdewijn P, Tricot S, Cruveiller S, et al. Chemical evolution of a bacterium’s genome. Angew Chem Int Ed Engl. 2011;50: 7109–7114. doi:10.1002/anie.201100535