Elaboration of innovative single-cell genomics approach and application to soil bacterial communities

Solène Mauger

COLLEGE ECOLOGIE
DOCTORAL GEOSCIENCES
BRETAGNE AGRONOMIE ALIMENTATION

Université de Rennes

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES

ECOLE DOCTORALE N° 600
Ecologie, Géosciences, Agronomie, Alimentation
Spécialité : *Ecologie et évolution*

Par

Solène MAUGER

Elaboration of an innovative single-cell genomics approach and application to soil bacterial communities

Thèse présentée et soutenue à Rennes le 22 novembre 2023

Unité de recherche : ECOBIO - UMR6553 – Université de Rennes, Cellenion - Lyon

Rapporteurs avant soutenance :

Lucie BITTNER MC, ISYEB Paris
Graeme NICOL DR CNRS, EEA/Ampère Lyon

Composition du Jury :

Président :	Christophe MOUGEL	DR, INRAE Rennes
Examinateurs :	Stéphane HACQUARD	Research group leader, Institut Max Planck Cologne
	Christophe MOUGEL	DR, INRAE Rennes
Rapporteurs :	Lucie BITTNER	MC, ISYEB Paris
	Graeme NICOL	DR CNRS, EEA/Ampère Lyon
Dir. de thèse :	Philippe VANDENKOORNHUYSE	PR, ECOBIO Rennes
Co-encadrement :	Cécile MONARD	CR CNRS, ECOBIO Rennes

Invité(s)
Cécile Thion Senior scientist, Cellenion Lyon

RESUME EN FRANÇAIS

Les microorganismes regroupent les organismes microscopiques uni- ou pluri-cellulaires, qu'ils soient eucaryotes ou procaryotes. Ces organismes sont présents dans de nombreux écosystèmes où leurs fonctions sont diverses et indispensables au maintient des équilibres métaboliques et biogéochimiques de leur environnement. Les microorganismes ont co-évolués avec les organismes structurellement plus complexes (*i.e.* les animaux et les végétaux) et permettent d'étendre les fonctions de leur hôte, leur offrant une plus grande résilience face au aléas environnementaux. Malgré leur prévalence et les intérêts écologiques qu'ils offrent, les microorganismes demeurent méconnus en termes de diversité génétique, de fonctions, d'interactions et d'évolution. Les bactéries, des organismes procaryotes unicellulaires, présentent une grande diversité génétique entre les espèces mais aussi au sein des espèces. Leur évolution réside sur l'apparition continuelle de mutants formant des populations et ainsi de la diversité intra-spécifique qui permet la sélection des lignées les plus adaptées aux changements environnementaux et le maintien de la structure des communautés bactériennes. Afin de comprendre les dynamiques évolutives des bactéries il est donc nécessaire de capter l'information à l'échelle de leur matériel génétique pour retranscrire leur diversité ainsi que les paramètres environnementaux qui régissent les interactions et les besoins métaboliques des bactéries. Le choix des outils pour étudier les microorganismes est guidé par l'échelle d'étude désirée. La culture de souches bactériennes en laboratoire permet d'étudier les métabolismes et interactions dans des milieux simplifiés. A l'inverse, le séquençage de l'ADN bactérien environnemental (metagénomique) offre un aperçu de la diversité microbienne en milieu naturel. L'utilisation des approches de séquençage de cellules unique (ou single-cell omics) a récemment émergé et pourrait permettre d'étudier les génomes bactériens individuels dans leur milieu naturel, leur métabolisme et les possibles interactions intra- et inter-espèces dans un contexte évolutif, permettant l'accès à l'échelle populationnelle et donc à la possible exploration de questionnements en écologie microbienne actuellement hypothétiques. L'application de cette technique nécessite l'utilisation de matériel dédié à l'isolement des cellules bactériennes, de la lyse de leur paroi cellulaire puis de l'amplification de leur génome avant de procéder à des méthodes de préparation de librairies de séquençage. Ces étapes présentent de nombreux biais ainsi qu'un

coût non négligeable. Cette approche est fortement soumise aux risques de contamination et à ce jour permet d'étudier quelques centaines cellules bactériennes dont les génomes récupérés demeurent grandement partiels. Ces limites en termes de coût, de biais moléculaires, et de manque de représentativité de la réelle diversité des communautés bactériennes ne permettent pas une application globale de cette technique au service de l'écologie microbienne. Un article détaillant les applications possibles et les limites du séquençage de cellules bactériennes uniques a été publié dans le journal Trends in Ecology and Evolution durant ce projet de thèse.

Dans ce contexte, mon travail de thèse vise à i) développer un protocole de préparation de librairie sur cellules uniques applicable aux bactéries en limitant au maximum les biais associés à cette approche, ii) valider le protocole développé sur des souches bactériennes avec génomes complets référencés et proposer une procédure de décontamination des données automatisée et iii) appliquer le protocole à un échantillon environnemental afin de tester son efficacité à répondre à des questionnements en écologie microbienne et le comparer à des approches de métagénomique.

Le protocole de préparation de librairies de cellules uniques a été développé autour de l'instrument d'isolement de cellules cellenONE, permettant une distribution précise des bactéries détectées visuellement par la caméra de l'appareil dans le support souhaité. Afin de réduire les coûts et les risques de contamination, le protocole a été imaginé pour être utilisé dans des volumes réactionnels grandement réduits par rapport aux solutions classiques de préparations de librairies (de l'ordre du nanolitre versus microlitre sur le marché). Pour ce faire, chaque étape du protocole en amont de l'identification des ADN cellulaire - permettant leur regroupement dans de grands volumes et ainsi leur purification – ont été élaborées pour être compatibles entre elle sans besoin nécessaire de purification et réalisable en volumes réactionnels réduits. Premièrement, la lyse cellulaire utilisée provient d'une publication de Stepanauskas et al., 2017. Sa composition alcaline a permis de décomposer la paroi bactérienne sans endommager l'ADN et n'a pas inhibé la réaction moléculaire à suivre : l'amplification du génome. Pour cette étape, la Multiple Displacement Amplification (MDA) a été utilisée et les volumes réduits à partir du kit REPLI-g advanced DNA single cell kit de Qiagen. Cette étape a permis d'obtenir de l'ADN en grande quantité à une taille autour de 10 000 pb pour chaque cellule isolée. Une fois amplifié, les brins d'ADN ont été fragmentés à 250 pb grâce au kit QIAseq FX DNA Library Kit de Qiagen. Ensuite, des adaptateurs d'amorces

de séquençage personnalisés ont été ajoutés à l'extrémité de chaque brin contenant un identifiant et résultant en une identification unique de l'ADN par groupe de 96 cellules. Ces groupes de 96 cellules ont ensuite été rassemblés afin d'y ajouter les amorces avec index de séquençage Illumina Nextera XT résultant en un autre niveau d'identification, cette fois de l'échantillon. Les contrôles qualité effectués à chaque étape étant validés (taille des fragments, rendement, validation par PCR), le protocole a pu être appliqué à des souches bactériennes connues pour évaluer son efficacité.

Le protocole a été appliqué aux souches *Pseudomonas fluorescens* et *Staphylococcus epidermidis.* Les données générées ont été comparées aux génomes de référence pour évaluer la quantité et la qualité de l'information récupérée. Les résultats montrent qu'une part réduite du génome original est conservée, ce biais venant probablement de la méthode d'amplification du génome (MDA) produisant une couverture du génome non uniforme. Un pipeline de décontamination a été développé et a en revanche démontrer la faible quantité de contaminants introduits dans les échantillons. Les contaminants retrouvés provenaient du manipulateur ou des kits de préparation des échantillons, ils ont donc été éliminé des données finales. Afin de valider ce pipeline de décontamination automatique des données, un jeu de données publié de séquençage de cellules uniques provenant d'échantillons marins a été utilisé. Sur ces données, des contaminants ont été identifiés par le pipeline et ont permis de souligner le manque de considération de ces contaminations dans les outils d'évaluation de la qualité des génomes assemblés via séquençage de cellules uniques (*e.g.* Check M). Cette partie a démontré le bon fonctionnement de protocole de préparation des échantillons précédemment développé et a souligné l'importance de l'effort nécessaire autour de la décontamination universelle de ce type de données afin de limiter des erreurs introduites dans les génomes de référence.

Le protocole de préparation de librairies de cellules uniques a ensuite été appliqué à un échantillon environnemental afin d'évaluer sa capacité à retranscrire les informations nécessaires à la réflexion autour de questionnements poussés en écologie microbienne tels que l'effet de l'acidité du sol sur la structure des communautés microbiennes. Notamment, la taille des assemblages, la diversité des communautés bactériennes retrouvées, la présence de gènes marqueurs et la possibilité de réaliser une phylogénie à partir des échantillons ont été testés. L'échantillon utilisé provenait du site expérimental de Craibstone en Ecosse, où un gradient d'acidité contrôlé est appliqué sur des parcelles agricoles. Ici, des échantillons de sol

provenant du traitement pH 4.5, 6 et 7.5 ont été utilisés. En parallèle du protocole développé, les échantillons ont également été traité via des techniques de métagénomique à titre de comparaison. La diversité bactérienne retrouvé via les SAGs (Single Amplified Genomes) était assez divergente de celle retrouvée via les MAGs (Métagenome Assembled Genomes). La diversité via les SAGs était plus importante que les MAGs. De plus, les assemblages SAGs présentaient moins de petites séquences que les MAGs, indiquant une importante présence de séquences contaminantes dans les MAGs. Les SAGs permettent de différencier l'ADN cible de l'ADN contaminant environnemental, permettant l'obtention d'assemblages de séquences plus propres. En revanche, ces assemblages demeurent de petite taille et nécessite d'être améliorer pour une exploitation optimale de la présence de gènes marqueurs et de la diversité intra-spécifique. Le placement de ces assemblages dans un arbre phylogénétique a montré la présence de certains individus appartenant à des familles bactériennes sous représentées. La plupart des échantillons SAGs et MAGs n'étaient cependant pas affilié à des familles bactériennes, témoignant de l'assemblage de trop faible qualité.

Ce travail a démontré la capacité du séquençage de l'ADN de bactéries uniques à retranscrire avec une plus grande précision que la métagénomique la composition des communautés bactériennes. Cette approche reste cependant complexe et nécessite des équipements et un personnel spécialisé. Afin d'améliorer la qualité des assemblages, il est nécessaire d'optimiser les techniques moléculaires, notamment l'amplification du génome, afin de limiter la contamination introduite, les erreurs d'amplification et d'uniformiser la représentation du génome. Grâce à ces amélioration, l'annotation des génomes sera rendue plus systématique et des approches de modélisation de l'interaction et l'évolution des métabolismes pourra être envisagée. Le futur de l'écologie microbienne réside dans la combinaison des approches techniques, quelles soient axées sur la génomique, la transcriptomique, la cultivation ou la modélisation, afin d'obtenir une image complète et précise des processus évolutifs dans lesquels les microorganismes sont impliqués.

Acknowledgements

Au début de ma thèse, quelqu'un m'a dit « une thèse, ça se fait tout(e) seul(e) ». Après ces presque 4 années de préparation au doctorat, je n'ai jamais eu l'impression d'être seule. Au contraire, aujourd'hui j'ai beaucoup de personnes à remercier qui m'ont accompagnée de près ou de loin pendant mon travail, et dans ma vie.

D'abord, mes encadrants :

Ça commence avec Philippe, mon « mentor » depuis le master, qui m'a transmis sa passion pour les petites choses du sol. Merci de m'avoir donné l'opportunité de travailler à tes côtés sur ce sujet « qui donne à rêver ». Merci pour tes conseils, ton soutien, ta patience et ton optimise inébranlable.

Cécile M., on s'est rencontré le jour où je vous disais être partante pour cette thèse. Merci de m'avoir fait confiance, sans me connaître. Merci pour ta bienveillance, ta compréhension et ton calme dans toutes les situations.

Cécile T., tu n'as pas eu le choix non plus de m'avoir moi à encadrer (décidément…), mais je me souviens de tes mots après mon premier passage à Lyon « je suis contente que ce soit toi ». J'ai beaucoup beaucoup appris à tes côtés, beaucoup rigolé aussi. Merci pour nos conversations sans fin pendant les purifs dans les premiers locaux de Cellenion, devant le cellenONE ou en visio. Merci pour ta franchise, ton dynamisme et tes TOC de corrections.

Cellenion,

Merci à Guilhem d'avoir été également moteur de ce projet et de m'avoir fait intégrer Cellenion. J'y ai beaucoup appris et rencontré de très belles personnes malgré mes petites apparitions à Lyon. Léna, Johannes, Solène, merci pour les temps passés au labo, à recommencer les manips et designs d'adaptateurs en tout genre. A tous les autres aussi, pour votre soutien technique et moral mais aussi pour les sorties du soir et du weekend (heureusement pas tout le temps au Hopper). A Laura aussi, que je n'ai jamais rencontré « en vrai » pour l'instant mais dont la présence a été très importante pour moi durant cette dernière année.

Ecobio,

Merci aux ingé des plateformes Ecogeno, PEM et Ecochim de m'avoir beaucoup appris et d'avoir répondu à mes nombreuses questions. Surtout Ecogeno, je ne saurais pas compter le nombre de fois où je suis arrivée dans votre bureau pour vous solliciter… Sophie et Romain, merci pour votre patience, votre disponibilité et votre bonne humeur. Antoine, tu as eu la lourde tâche de m'aider pour mes manips (déjà, désolée..) et tu les as rendu bien plus sympathiques à coups de gonflage de gants et memes en tout genre. Merci pour ton soutien pendant ces périodes d'essais de MDA en cellenCHIP qui n'en finissait plus (et d'avoir partagé les joies de cellenONE).

Merci au staff, en particulier Jordan, Bertrand et Isabelle qui essayent de répondre à nos demandes urgentes au mieux parmi d'autres demandes urgentes.

Merci à Cécile GR, Ning, Alexis et Achim pour vos retours constructifs sur mon travail afin de m'aiguiller au mieux.

Un coucou aux doctorants de « l'ancienne » promo: Claire, Lucie, Marine, Léa et Victor. J'ai commencé ma thèse avec vous, même si vous avez essayé de m'en dissuader. Finalement je ne regrette pas et surtout je garderai longtemps en mémoire les danses et acrobaties propres au bureau 112.

Merci aux doctorants (aka les nains) avec qui j'ai partagé cette aventure : je crois que peu de personnes peuvent se réjouir d'avoir trouvé un groupe d'amis comme le notre pendant leur thèse. Merci pour les pauses du midi à refaire le monde dans le canapé, pour les débats, les encouragements, les moments de compassions, les rigolades. Sans oublier nos weekends à thème type jardinage et plage et bien sûr les soirées raclette-mojitos. Ces moments m'ont réellement permis d'avancer dans les moments les plus difficiles de ma thèse (à peu près tout le temps, donc). Petite mention spéciale pour Léa <3.

A ma famille, qui à ce jour ne comprends peut-être toujours pas ce que je fais, mais qui n'a jamais douté de mes capacités. Me ressourcer auprès de vous a été indispensable.

Et enfin Sam, mon pilier, sans qui je ne sais pas dans quel état ce manuscrit et moi-même aurions fini. Tu mérites un petit bout de ce diplôme pour avoir vécu la thèse à travers moi…

Ce manuscrit est pour vous, avec vous.

Table of contents

Preamble

All sciences start with curiosity.

From the very beginning, Men have evolved in collaboration with Nature. Their inherent willpower to understand their environment has been the source of their evolution. From learning to start a fire, to farming, and mastering minerals, great steps in natural sciences have always been initiated by curiosity. Through observation of the detailed ecological processes around us, forms of life be it big or small, were decrypted, classified, and exploited. Knowledge has become the power of humankind, compensating for the lack of physical abilities our bodies present compared to other animals. We are not the strongest, we do not run the fastest, nor can we hear sound from kilometres away or navigate across the globe using magnetic fields. However, what we can do better than any other species is think. We have used our amassed knowledge to protect, conserve, and anticipate our fate but also the fate of our ecosystem and its inhabitants. Nothing resists Man. In our quest for power, we have often disregarded Nature in order to dominate, creating technologies and economic benefits along the way. Along with our ideas and thoughts, we have expanded, often believing we could control everything, and that life had little to no secrets left for us to uncover.

That all changed when we discovered microbes.

"Microbes have been here since life began, almost 4 billion years ago. They created the system that we live in, and they sustain it."

"We may not see them, but they're running the show."

Dan Buckley

Understanding microbes

I- Microbiology through Centuries

1.1- The first big steps

The discovery of microscopic organisms dates back to the 17th century when Antoni Leeuwenhoek and Robert Hooke observed bacteria (baptized “animalcules” at the time) and fungi under simple microscopes for the first time (Gest, 2004). Two centuries later, the cultivation of bacteria on artificial media was initiated by Louis Pasteur’s work (Bonnet et al., 2020) and largely contributed to the development of microbiology in human health research. Fifty years ago, the first DNA sequencing technology was developed by Sanger (Sanger et al., 1977) giving the opportunity to microbiologists to get a more precise census of microbial abundance and diversity in various ecosystems. The genome of *Haemophilus influenzae,* identified as responsible for the flu disease in 1892 by Richard Pfeiffer (Pfeiffer, 1892), was the first to be completely sequenced from the bacterial realm in 1995 by Craig Venter’s group (Fleischmann et al., 1995). Through the centuries, scientists discovered inconceivable bacterial functions such as quorum sensing, bioluminescence, or metal reduction (Myers & Nealson, 1988; K. H. Nealson & Hastings, 1979; K. H. Nealson et al., 1970). Known for these discoveries, Kenneth Nelson experienced rejections from editors with the argument: “Bacteria do not do this”, twice. Just like these editors, did Antoni Leeuwenhoek have any idea that the “animalcules” he observed would have such abilities? Still today, we might still be very naïve concerning the real potential of prokaryotes.

1.2- Incorporating microbes in the Tree of Life

To help conceptualize life organization and evolution, scientists attempted to arrange organisms around interconnected branches through centuries, with one of the first famous trees proposed by Haeckel in 1866 (Figure 1a). The first Tree of Life proposing three main domains as we know them today was published a century later (Woese & Fox, 1977) (Figure 1b). It was built based on molecular data which was a little revolutionary at the time but generated some mistrust regarding this study and difficulties to evaluate its veracity.

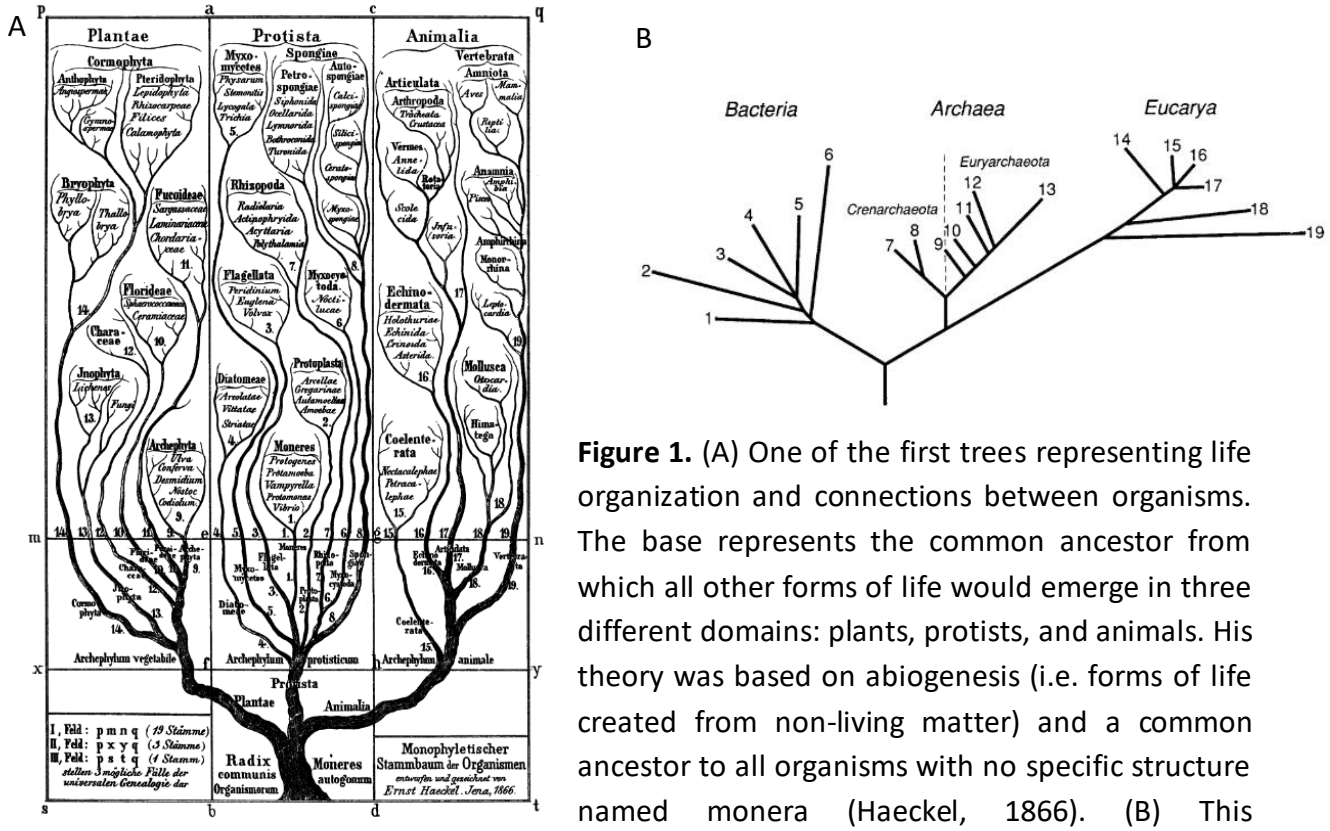


Figure 1. (A) One of the first trees representing life organization and connections between organisms. The base represents the common ancestor from which all other forms of life would emerge in three different domains: plants, protists, and animals. His theory was based on abiogenesis (i.e. forms of life created from non-living matter) and a common ancestor to all organisms with no specific structure named monera (Haeckel, 1866). (B) This classification was based on molecular data and proposed three groups with distinguished cell types: Bacteria, Archaea and Eucarya. Here, life is evolving from all branches and was first initiated by an ancestor of prokaryotes (figure from Pace et al., 2012).

Nonetheless, the outcomes of this work raised major questions regarding the origin of life and the evolutionary history of all organisms. The origin of life and its evolution were largely debated and generated various hypotheses based on morphology where microbes could not

easily be integrated and distinguished. The use of sequencing technology in phylogeny allowed the scientific community to measure the immense reservoir of diversity that microorganisms represent and is today the main way of incorporating new branches within the clades of Bacteria and Archaea (Hug et al., 2016; Tahon et al., 2021; D. Wu et al., 2009), Figure 2. Recent studies confirmed that eukaryotes evolved from archaea 2 billion years ago (Guy & Ettema, 2011; Kelly et al., 2011; T. A. Williams et al., 2013), putting Bacteria and Archaea realms as the basis of all forms of life when they appeared 3.2 billion years ago. The organization of phylogeny is evolving every year with our ability to look deeper and broader into the genetic signatures of newly discovered organisms (Hug et al., 2016; Parks et al., 2018). Our perception of prokaryotic phylogeny is especially concerned with rearrangements due to their complexity in size, interactions, diversity, and dependence on advanced molecular tools, which are constantly being improved. These microorganisms at the basis of life might still be the most mysterious organisms whilst being essential for all other forms of life. For these reasons, microbiology is a very dynamic domain, with significant implications in biological and ecological research fields.

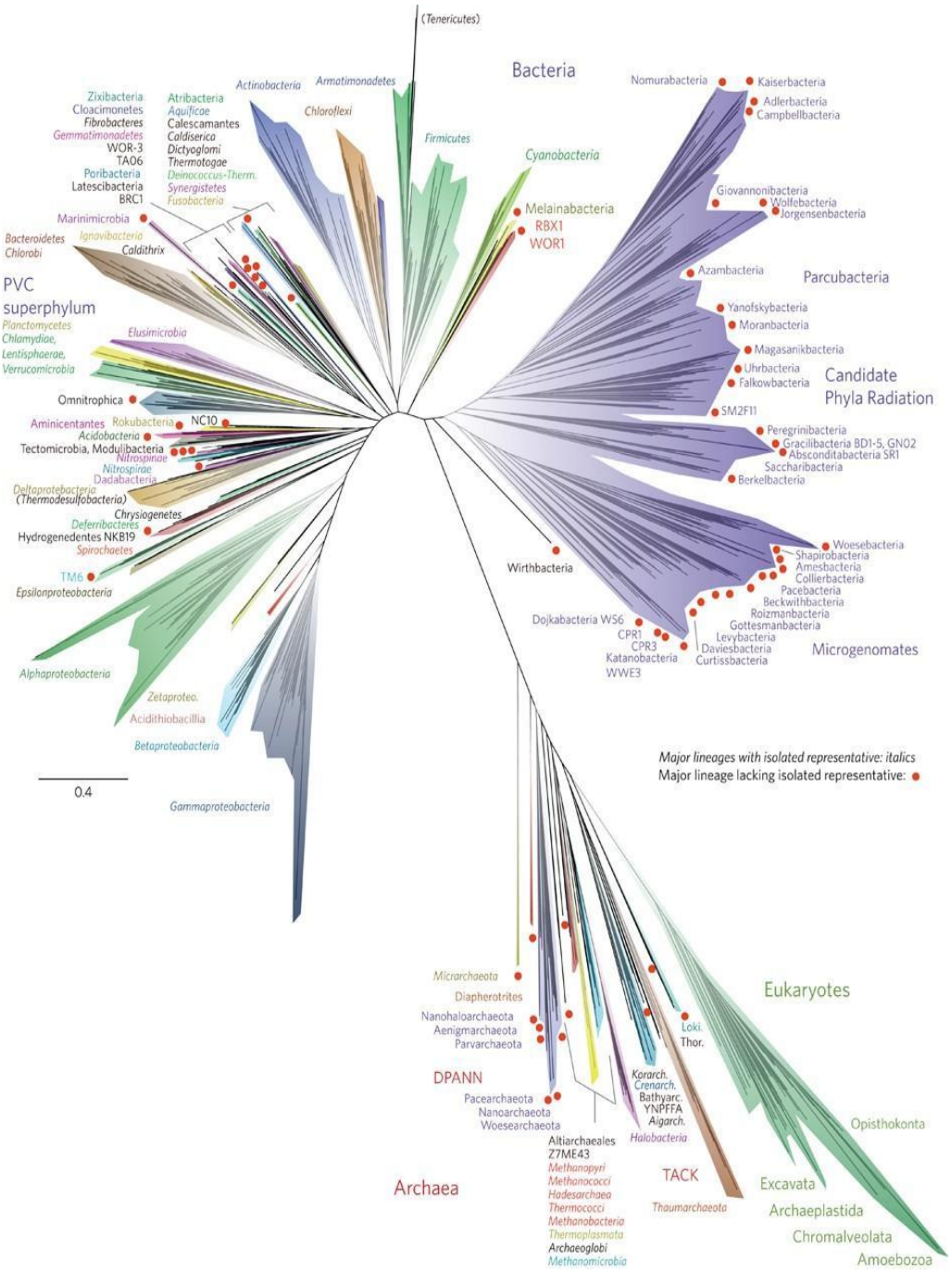


Figure 2. Recent representation of the Tree of Life by Hug et al. in 2016. The tree was revolutionized by using innovative approaches such as single-cell omics. The branches without isolated representatives are marked with a red dot and testify of the necessity of molecular tools developments for prokaryotes phylogeny construction.

1.3- (Micro)biology, not without ecology

We, multicellular eukaryotes, are the result of millions of years of evolution initiated by bacteria and archaea (Guy & Ettema, 2011) via successive symbiosis between microbes (Douglas, 2014). Our evolutionary history is closely related to microbes and remains this way still today. Not only do we carry bacterial genes, but we also are the home for billions of bacteria within our tissues (Savage, 2003; Sender et al., 2016) like all plants and animals. In our bodies, microbial cells are evaluated to be more abundant than human cells (Savage, 2003). If only a few multicellular eukaryotes do not present a microbiota (Hammer et al., 2017, 2019), most of them highly rely on their microbial symbionts for multiple aspects of their life. The implication of microbes in animal and plant health and resilience is acknowledged to be essential for metabolism functioning or stress tolerance (Hoang et al., 2021; Houwenhuyse et al., 2021). This association is a partnership, a symbiosis, where the host sees its evolutionary potential extended by the presence and functions of its microbiota (Henry et al., 2021). Such collaboration questions the individual concept in biology (Gilbert et al., 2012) and testifies to the necessity to consider microbiology in light of ecological dynamics.

Every host or environment in which bacteria are present can turn either incredibly strong and resilient or dramatically ill. Hosts' fate is so closely related to their microbiota that the holobiont concept which has been richly discussed (Moran & Sloan, 2015; Theis et al., 2016) is now commonly applied in microbial ecology (Bordenstein & Theis, 2015; Hassani et al., 2018; Vandenkoornhuyse et al., 2015). Genes of the host and its microbiota are forming the hologenome, the functions provided by both entities will determine the fitness of the holobiont and its evolutionary trajectory (Figure 3). Therefore, both hosts and microbes will evolve together. Historically, the plant-microbes association coincides with the colonization of terrestrial lands by plants, 450 million years ago (Knack et al., 2015; B. Wang et al., 2010), suggesting the stability and durability of this partnership.

Microorganisms support major functions for their host, for instance in plants, where the microbiota is essential for nutrient and water uptake from the soil to increase plant productivity (Van Der Heijden et al., 2016), water retention from the leaf (Raddadi et al., 2018), or for pathogen protection (Mendes et al., 2011; Ritpitakphong et al., 2016; Vannier et al., 2019). There is a rising interest in considering microbes in plant breeding and productivity in agriculture (Compant et al., 2019; Gopal & Gupta, 2016). Similar patterns of dependencies are observed in animals, particularly detailed in humans, where bacteria enhance food digestion and nutrient absorption, protect against external pathogens, and affect the psychological state of their host (Fung et al., 2017; Kamada et al., 2013; Nicholson et al., 2012; Sharon et al., 2014). These microbes together work in harmony with their host but can easily turn into a heavy companion to carry. A destabilization of the microbiota equilibrium is characteristic of some diseases or chronic disorders and decreases the resilience capacities of the host (Martinez-Medina et al., 2014; Rigottier-Gois, 2013). The power of microbes on their host can be such that it can control and extend their ecological niche (Hoang et al., 2021; Schönknecht et al., 2013) and behavior (Archie & Theis, 2011; Yuval, 2017).

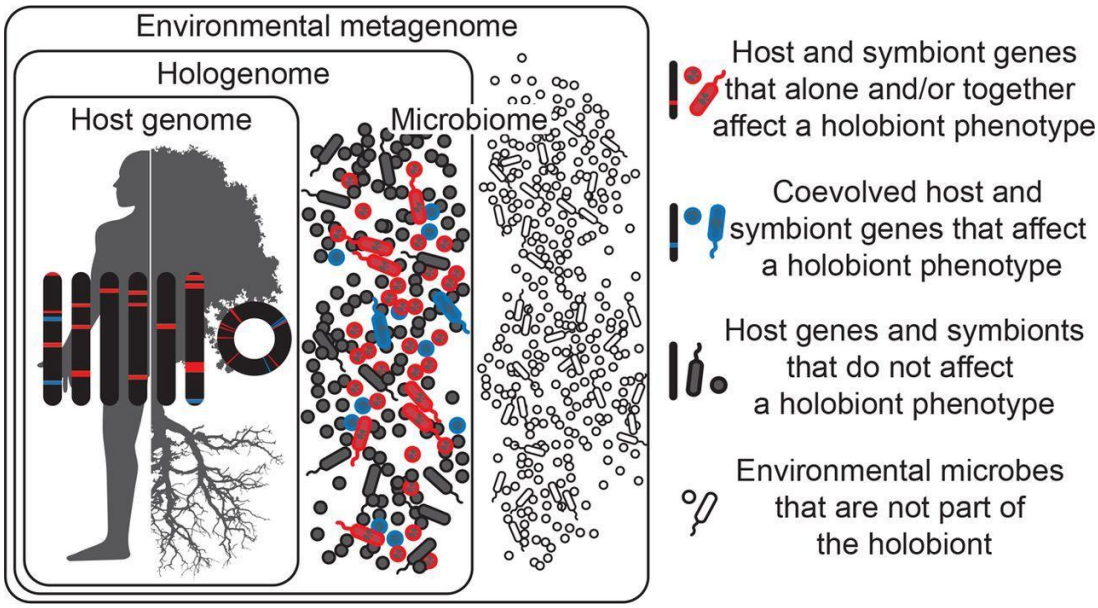


Figure 3. The holobiont concept representation, containing the host and its symbionts genes (Theis et al., 2016).

The absence of individuality is also applicable within microbiota, as each member depends on the other, alike or not, for their growth and evolution. A major factor for bacterial growth is quorum sensing which allows members of a community to coordinate and communicate (Enomoto et al., 2017; Kenneth H. Nealson & Hastings, 2006). A microbiota, as a unit, presents its own rules of interactions and organization where gene flows between members are very common via Horizontal Gene Transfer (HGT) and vary according to the species and environmental conditions (Dagan et al., 2008; Polz et al., 2013). Many environmental parameters influence community composition and structure, modifying the expressed functions within the microbial community and therefore the types of interactions involved. The disruption of the community is buffered by its structure, often involving keystone taxa serving as pilar to metabolic networks (Banerjee et al., 2018; Tang et al., 2022). A complex system of interdependencies exists and is extremely challenging to characterize within natural microbial communities. The observation and characterization of interaction types have been induced between strains in lab experiments (Carlström et al., 2019; Mee et al., 2014), enabling the elaboration and exploration of novel ecological theories. The organization of the community around a common supply of metabolites (commonly called “common goods”) has been theorized by the Black Queen Hypothesis after the observation of metabolic dependencies and auxotrophies (i.e. the incapacity of an organism to synthesize a compound that is necessary to its metabolism) in the marine ecosystem (Morris et al., 2012) and is a concrete example of the bonded evolutionary trajectory of microbes (D’Souza et al., 2018; Douglas, 2020; Estrela et al., 2016). The composition of the common goods is specific to a community and forms niches with very stable microbial structures (Pascual-García et al., 2020).The availability of nutrients in environments is indeed determining the rise of specific bacterial interactions and dependencies, be it cooperation or competition (Mataigne et al., 2022).

The microbial world is complex and contains very different types of organisms. Bacteria, fungi, archaea, and viruses interact and regulate themselves by being in competition on many occasions. Viruses infect bacteria, archaea, and micro-eukaryotes and as a consequence influence the community composition, interactions, and nutrient cycling in which these microorganisms are involved (Weitz & Wilhelm, 2012). This regulation is largely understudied

for technical and methodological reasons even though recent improvements in these aspects are being made (Cristinelli & Ciuffi, 2018; Smith et al., 2022). Bacteria and fungi also interact in the rhizosphere where they compete in nitrogen cycling (Tatsumi et al., 2020), and soil colonization (X. Li et al., 2020). Within microbes, we will only focus on bacteria in the rest of this thesis work, keeping in mind that these interactions would be also judicious to implement in future work.

At a broader scale, bacteria are essential in biogeochemical cycles such as organic matter degradation (Rousk & Bengtson, 2014), depollution, bioremediation (Deng et al., 2019), and nitrogen cycling where they support major functions (Prosser et al., 2020). From cell biology to community structure, bacteria influence and are influenced by their ecosystem and therefore cannot be fully pictured without it. As a result, microbes can only be fully understood in the light of their ecological condition, i.e. their direct neighbourhood and environments in which they live. Integrating such ecological parameters in microbial studies is essential to upgrade their accuracy but is particularly challenging. Efforts in the field are concentrated on associating and upgrading tools and methodologies to reach this goal of a higher level of representativity of natural ecological processes involving microbes (Ross & Whiteley, 2020).

II- Methodology in microbial ecology

The study of microbes has been evolving with and thanks to technical advances. The outputs in data and interpretation are highly dependent on the scale of the study and tools employed so one can never detach the observations made from the methodology that led to the results. With the complexification and diversification of tools in microbiology, scientists are assembling clues from different standpoints and contexts to extend the knowledge of this Kingdom. However, this can also lead to a lack of uniformity and representativity of the techniques employed and outputs (Abellan-Schneyder et al., 2021), creating vast debates among microbiologists questioning whether the chosen procedures are truly adequate in microbiology research.

2.1- Tools diversification

Cultivation is the oldest and the cheapest way of directly studying microbes. It relies on finding the adequate growing parameters (e.g. medium, temperature, agitation, time..) for each strain. The growth medium contains nutrients necessary for bacterial growth, which are not easily identified for most strains and are not systematically sufficient. Indeed, some bacteria grow better in the presence of other strains in co-cultures or with complex chemical components, for instance, plant exudates (Dhungana et al., 2023). The limits given to this approach are many, the most common being that most bacteria remain uncultured (Lloyd et al., 2018). Some known bacteria have not been grown in the lab yet, whether they require complex or unknown parameters to grow, but most bacteria cultivability has simply not been tested. Recent omics studies testify of the lack of isolate representative for most taxa (Hug et al., 2016; Lloyd et al., 2018; Steen et al., 2019) but also of the gap between cultivated and wild strains' genomes (Baker & Dick, 2013). Extrapolating to the animal kingdom, this is equivalent to raising conclusions on a wild animal's behavior based on a domesticated one's. Therefore, making inferences on wild populations from isolated strains is highly biased. Currently, most of the knowledge on Archaea and Bacteria relies on well-studied isolates or reconstructed genomes of uncultured cells, which is providing information on phenotypes but not on the physiology and metabolic status of the cells. It is essential to achieve the cultivation of more strains to understand their cell functioning and biology and increase the catalogue to which uncultivated bacteria will be compared. This is why cultivation is still a growing topic with a complexification of approaches (Lewis & Ettema, 2019; Lewis et al., 2021).

In the mid-70s, access to bacterial DNA gave the possibility to study complex environmental microbial communities. This culture-independent approach was initiated by the sequencing of 16s rRNA gene which is a highly conserved region. This sequence is ubiquitous in prokaryotes and presents quantifiable variations allowing the classification of organisms. The use of amplicon sequencing has been a huge step forward in taxonomic surveys and phylogeny in many ecosystems (Wilson & Blitchington, 1996; Woese & Fox, 1977; Yarza et al., 2014). Cheap and easy to prepare, amplicon sequencing is now routine in microbial ecology to appreciate

bacterial diversity in various environments. This approach is mainly used for community composition description but lacks robustness when trying to access diversity at low organizational levels (Ellegaard & Engel, 2016; Van Rossum et al., 2020).

Shotgun sequencing gives access to potentially more gene markers to reveal the functional potential of microbes. Its use is now common but requires a higher effort of sequencing and is, therefore, more expensive than amplicon sequencing. Many bioinformatic tools have been developed to cope with the complex datasets generated with this approach and aimed at assessing taxonomic and functional diversity within species (Brown, 2015; Crits-Christoph et al., 2020; Sangwan et al., 2016). In practice, genome assemblies are incomplete and sequencing errors are not systematically separated from SNP variations between individuals. Reconstruction of genomes via metagenome-assembled genomes (MAGs) presents major limitations such as chimer generations and low-quality assemblies (Alneberg et al., 2018; Bowers, Kyrpides, et al., 2017; Shaiber & Eren, 2019). Therefore, there is only a small chance to describe intra-species diversity with accuracy (See Chapter I). Shotgun and amplicon sequencing is mainly applied to bulk genomic DNA extracted from the bacterial matrix, capturing external DNA as well as dead and dormant cells. This external DNA represents a consequent amount of material compared to the targeted DNA from living microbial cells, even though it does not seem to impact the proportions in microbial diversity (Courtois et al., 2001), suggesting the low quality of the contaminant DNA captured via direct extraction from samples. Genomes are reconstructed from short-read sequencing, even though long-fragment sequencing is more and more frequent (e.g. PacBio and Oxford Nanopore sequencing technologies) (Jain et al., 2016; Lu et al., 2016), with the risk of mixing sequences from different individuals, populations, or species. Moreover, the gap between the subject of the study and the database used as a reference greatly influences the results for SNP calling or data interpretation (Breitwieser et al., 2018; Bush et al., 2020). Molecular techniques enable us to avoid cultivation limitations, but bring other inconveniences previously stated mostly related to sample preparation and data treatment. Moreover, as most microbes remain unknown, there is no direct way to verify the accuracy of the extraction and cell lysis process, the true universality of the used primers, or the extent of external microbial contamination of the samples. Moreover, as mentioned earlier,

environmental data are interpreted based on a catalogue of cultivated strain references to serve as a basis from which to compare sequences. The information contained in databases is the limit to which we can identify strains and genes. For microbes, the rate of mutations is fast and precautions must be taken when extrapolating in-lab observations and gene expression towards wild strains. Despite the major advances in molecular tools, most microbes and their coding potential remain unknown and have been referred to as "Microbial Dark Matter" (MDM) (Bernard et al., 2018; Marcy et al., 2007). The investigation of the MDM is ongoing and multiple genes and strains are discovered each year (Escudeiro et al., 2022; Rinke et al., 2013; T. J. Williams et al., 2022).

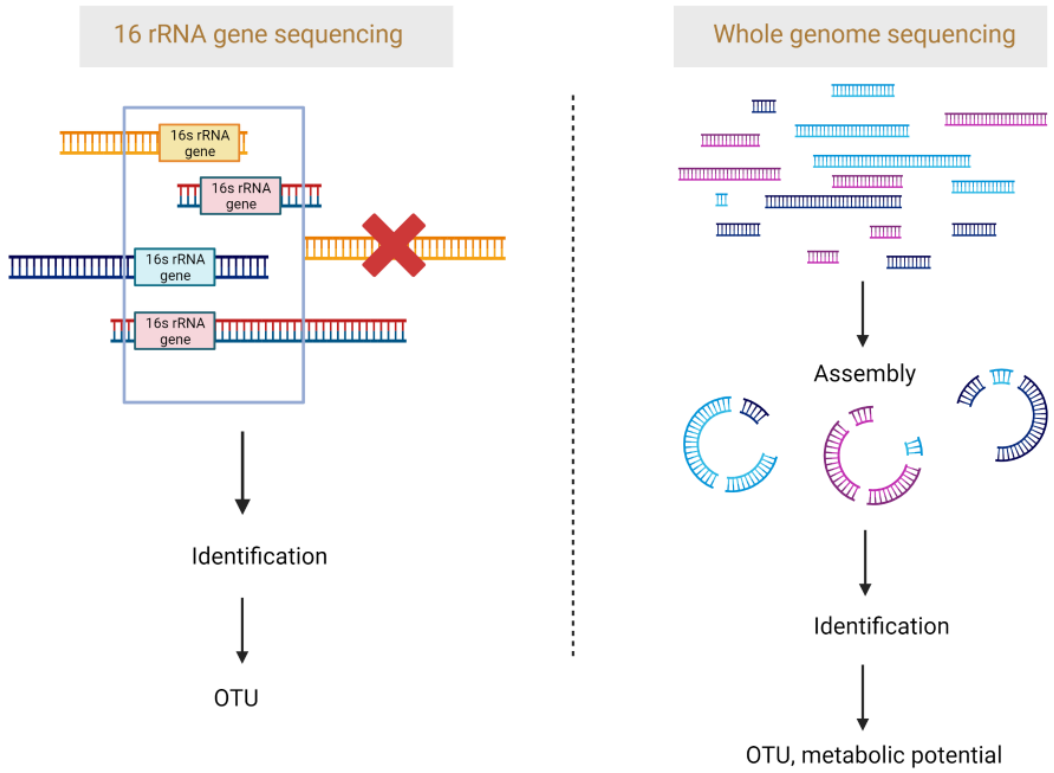


Figure 4. Principles of 16rRNA gene and whole genome sequencings. The sequencing of 16rRNA gene only considers this conserved gene for taxonomic identification of Operational Taxonomic Units (OTUs). For whole genome sequencing, DNA fragments of different bacteria can be assembled to reconstruct the core genomes of the community.

How to interpret genomes is a front-of-science question. Modelling is a way to analyse microbial networks from genome annotations and offers the possibility to understand complex interactions, which is very limited from molecular studies alone and available transcriptomics information in databases. Metabolic networks (or genome-scale metabolic models (GEMs)) allow the prediction of compounds produced by a corpus of genes, the functions of the microbe, and therefore its possibilities in terms of interactions with other members of the community. Built with mathematical designs, the use of models enables the study of many levels of interactions from simple nutrient exchange to complex networks of a microbial community. The more complex the analysed network, the more difficult it is to integrate the concerned metrics (Antoniewicz, 2020). The integration of top-down and bottom-up approaches is the best practice for microbial network analysis according to specialists (Lawson et al., 2019). As microbes live either in structured or well-mixed environments with direct and indirect effects on their fitness, the identification of input variables is very challenging (Gorter et al., 2020). Modelling is highly dependent on experimentations, that allow the identification of the variables which to build the models from, and can be seen as a hypothesis and ecological theories generator (Martinez-Rabert et al., 2023). The outcomes are highly dependent on -omics data annotations and completeness, whose quality is not always satisfying (Vandenkoornhuyse et al., 2010). However, modelling remains the most accurate approach compared to molecular approaches alone to put light on metabolisms and fine-scale interactions of bacterial strains and communities. Piece by piece, the reconstitution of the microbial world is evolving through the lens of our methodologies and technical tools with their associated limitations and biases (Figure 5). There is a complementarity in all methods used in microbiology from which, by combining the results, we should get a well-advanced picture of microbial organization and functioning (Stubbendieck et al., 2016). However, the combination of knowledge requires elaborating standardized methodologies and vocabulary, from sample collection to data treatment (Bokulich et al., 2020; A. Tripathi et al., 2018; Van Rossum et al., 2020). A combination of methods serving a common purpose started to emerge and scientists called for standardized procedures and systematic benchmarking (Meisner et al., 2022; Rainey & Quistad, 2020).

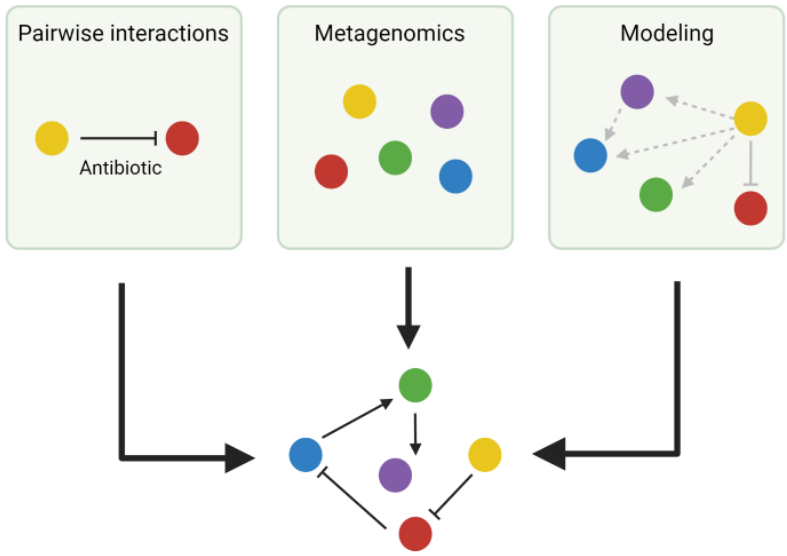


Figure 5. Outcomes of different approaches in microbial ecology, from which the results can be integrated to obtain a better view of microbes functioning. Re-drawn from Stubbendieck et al., 2016.

2.2- Getting closer to representativity

2.2.1- Hypothesis-driven over descriptive studies?

The microbial world comprehension is still mainly descriptive, first via microscopes and cultivation, and kept evolving with sequencing data. Hypothesis-driven studies are less common in microbiology even though they aim at causalities and physiological processes understanding which are scarce in the field. Descriptive and deductive studies do not serve the same purpose in science, and their use does raise questions of methodology: some are concerned about the interest and scientific accuracy of employing descriptive procedures. Rather than responding to a scientific procedure, descriptive studies would mostly assess technical questions without giving clues on the physiological mechanism behind them (Prosser, 2020, 2022; Prosser et al., 2007). The questions about methodology in microbial ecology are many (Prosser & Martiny, 2020; A. Tripathi et al., 2018): Do we need a big amount of data prior to testing the hypothesis

or raising conclusions on microbial organization and functioning? Is description relevant in itself? Is microbiology currently suited for hypothesis-driven studies?

Testing ecological questions requires precisely determining the subject of study and the parameters tested, implying a perfect understanding of which parameters are playing a role in the studied ecological process. Omitting important variables might totally change not only the results interpretation but also the hypothesis generation via models that require the list of elements to integrate for a particular ecological phenomenon. However, the overwhelming "unknowns" composing microbes in diversity, abundances, genes, functions, interactions, and ecological niches, make it difficult to target specific questions. Most of the ecological and environmental parameters involving microbes are still unclear, limiting the ecosystems to which the hypothesis can be formulated. Some variables might be missed and not included in deductive studies, resulting in biased or incomplete observations (A. Tripathi et al., 2018). Microbiology needs aggregated scientific knowledge for hypothesis-driven questions to be posed correctly (A. Tripathi et al., 2018). However, the aggregation of data based on observations needs to be standardized, which is currently not systematic. The consideration of scales is superficial and results in a lack of reproducibility (Ladau & Eloe-Fadrosh, 2019; Prosser & Raaijmakers, 2020). Most studies describing bacterial communities do so at very large scales, not being identical between different studies and removing the information where community structure and interactions happen (e.g. 0.1-1mm for terrestrial habitats) (Cordero & Datta, 2016; Nunan et al., 2020; Prosser & Raaijmakers, 2020). The temporal scale is also neglected or not fitted to the observed process, and dynamic patterns are most often overlooked. Overall, giving more importance to these scales is needed to identify which parameters possess a scale-dependent effect on microbes (Ladau & Eloe-Fadrosh, 2019). Being able to consider short vs long-term effects on microbes, their phylogeny, and their habitat structure is fundamental to considering well-built hypotheses driven by theories. Therefore, description still has its place in microbiology to furnish the basic knowledge of community content and variables to consider, if used properly. There are also examples of descriptive studies presenting deductive power. An historical example is the work of John Snow who understood how cholera disease was carried

by simply making a map of deaths related to the disease locations and revealing that it was carried by waters (Paneth et al., 1998).

Beyond the methodology employed, capturing fine-scale processes in nature is very limited due to our technical approaches and tools. To preserve such fine standpoints for observations and hypothesis testing procedures, synthetic communities were manipulated in laboratories (Ciccarese et al., 2020; Vannier et al., 2019). This strategy is useful to understand physiological mechanisms but only with a very limited selected strain. This represents two major problems: i) natural communities are far too complex in structure and interactions to fully predict their evolution and functioning only based on such experiments and ii) only strains that can be grown in vitro can be used, representing a limited fraction of the overall microbial diversity, which prevent the extrapolation of results towards most microbes.

2.2.2- Considering populations as units

To address questions of bacterial community evolution in a changing environment, the units of selection and as discussed earlier, the scales, must be chosen carefully. One important unit in microbiology and all life sciences is populations (i.e. variants of the same species) which are more relevant and accurate than the species concept when it comes to microbes (García-García et al., 2019; Niccum et al., 2020). Species are indeed a blurry concept in microbiology and the level of organization should be considered as gradients rather than boxes with a constant flow of genes inherited from vertical or horizontal transfers (de la Cruz & Davies, 2000; Van Rossum et al., 2020). Van Rossum et al. (2020) proposed a classification of terminology based on the proportion of identical nucleotides (Average Nucleotide Identity, ANI) to distinguish genomes, strains, subspecies, and species (Figure 6).

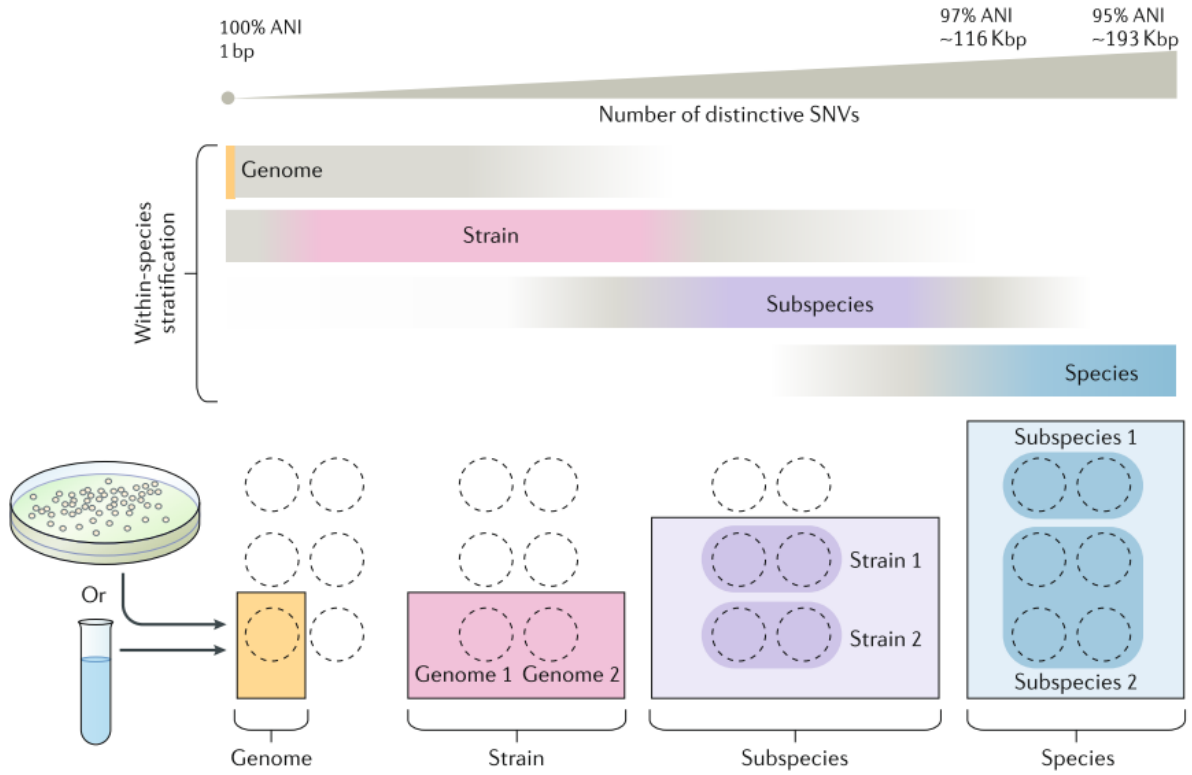


Figure 6. Van Rossum et al 2020. Terminology to employ based on SNV of within-species diversity. Each strain contains different genomes, subspecies contain different strains and species are composed by subspecies.

Populations (i.e. individuals from the same species living in the same area) represent a massive yet barely known reservoir of diversity for bacteria which is primordial to uncover if we hope to fully understand evolutionary dynamics within bacterial communities. Accessing the population levels of the bacterial realm would refine ecological concepts that are today understood in the light of global community scale or via in-vitro experiments focusing on a few strains only.

Recent studies succeeded in highlighting the role of newly theorized “keystone taxa” in the community structure and stability (Banerjee et al., 2018; Carlström et al., 2019; Tang et al., 2022). While there is no universal definition for these keystone entities, Banerjee et al. (2018) proposed to consider them as highly connected taxa that strongly influence the structure and functioning of the microbiome, shifting when the keystone taxa are removed (Banerjee et al.,

2018). There are multiple examples of such organizations in the microbial world, mostly discovered via network analysis (Banerjee et al., 2018). To understand the causality of strain interactions, in-lab testing is necessary to validate the correlations highlighted via network analysis. Multiple studies have shown the impact of the absence of single strains on the overall community (Carlström et al., 2019; Tang et al., 2022), helping to understand the functioning of such structures. The presence and proportions of keystone taxa are dependent on the nutrient content composition; therefore we can expect a shift in this community structure in major environmental conditions modifications (P. Wang et al., 2022; C. Wu et al., 2023). The complexity of such networks in nature is such that their reconstruction in artificial conditions cannot be achieved.

Taxa with relatively less direct impacts on the community structure compared to keystone species have been reported and are classified as “rare taxa”, acting in the background of cell interactions. Rare taxa are often overlooked in molecular studies, even though the correlation between abundance and key role in the community is not systematic (Shade et al., 2014). They have fundamental functions in the community, such as a buffering role in fluctuating environments to prevent other members from perishing. Their presence increases the connectivity between cells and strengthens the stability of the community (Jousset et al., 2017; Shade et al., 2014). Rarity can be driven by local and temporal environmental conditions or biotic factors and can testify to the narrow ecological niche of the species or trade-off in stressful conditions (Jousset et al., 2017). The rare status of a species might therefore fluctuate between communities and habitat conditions. Populations are very prone to this constant fluctuation in numbers, as they represent the inner diversity reservoir of species and therefore can be at the origin of rare taxa emergence. They ensure lineage continuity via very diverse sets of gene versions to increase the success of adaptation. Intra-species diversity can also rise after an environmental modification as a response to stress by encouraging mutations and gene flows (Figure 7(Davis & Isberg, 2016)). Microbial populations ensure the persistence within a microbial community under changing environmental conditions (García-García et al., 2019).

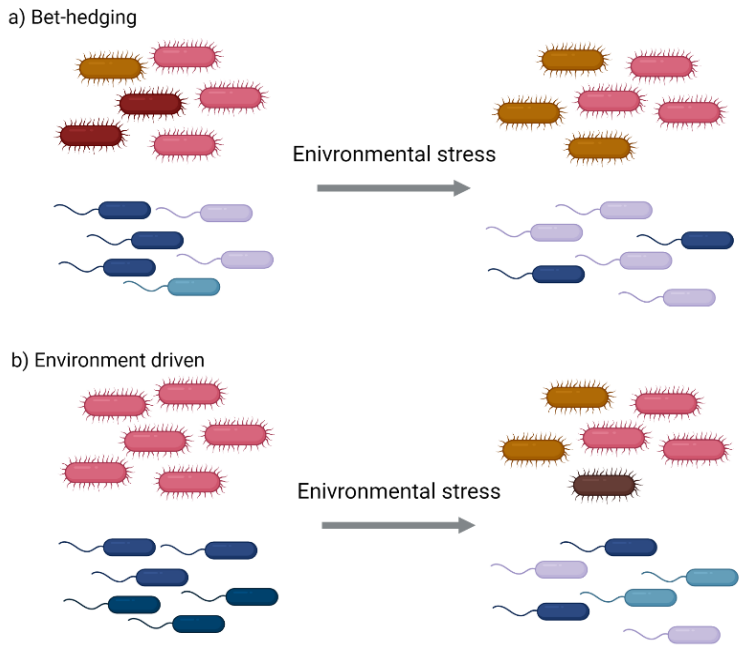


Figure 7. Modified from Davis & Isberg 2016. Generation of intra-species diversity via a) bet-hedging where diversity is altered by environmental stress, only the fittest populations will survive and spread and b) driven by environmental change where the rise of intra-species diversity rises after environmental stress, resulting in more phenotypes. The colors represent the phenotype of each cell and the shape (with or without a flagellum) represents each genome or species.

Populations represent an overlooked reservoir of diversity with fundamental roles in community functioning. While most of the interactions are made between species, some can be identified below the species scale (Baishya et al., 2021; J. Wang et al., 2021). The Black Queen Hypothesis model has been elaborated to explain metabolic dependencies between species (Morris et al., 2012), but is perfectly applicable to microbial populations as well (Mas et al., 2016). The reservoir of rare subspecies is a stock of important functions with potentially significant leaky compounds produced for the community. Different phenotypes can be found within species, and between strains. This has been particularly studied in the host health context, where many species can present pathogenic and commensal strains (e.g. *Escherichia*

coli (Leimbach et al., 2013) and *Bacteroides fragilis* (Pierce & Bernstein, 2016)). This phenotype variability is also observable in nutrient cycling (Neuenschwander et al., 2017), between seasons (Garcia et al., 2018), in drug response (Maier et al., 2018), or nitrogen fixation (Triplett & Sadowsky, 2003). Variability can also appear within populations, forming subpopulations, via preferential HGT or stochastic events (Davis & Isberg, 2016). Exploration of species intra-diversity in genomes and phenotypes is in its infancy and is fundamental for microbial evolution, interactions, and community structure comprehension.

Identifying and studying microbial populations is currently complex and limited by our technical methods. The genetic variations between species variants are difficult to highlight for most molecular approaches and bioinformatic tools. The few nucleotide variants separating populations are not easily distinguished from sequencing errors, it becomes however easier with the increase in sequencing depth (Van Rossum et al., 2020). Moreover, considering the huge diversity of microbes in ecosystems and the risk of external contamination, identifying populations requires sophisticated and robust protocols from DNA extraction to data analysis (See Chapter I). To help with this, research groups have developed bioinformatic tools (Arevalo et al., 2019; Brown, 2015; Crits-Christoph et al., 2020; Sangwan et al., 2016) to try to identify populations from metagenomic data. Traditional Metagenome Assembled Genomes (MAGs) are useful for uncultivated strains study but are, by definition, very likely to produce chimeric genomes (i.e. variations at the population level cannot be integrated within MAGs up to date and see also Chapter I) for complex microbial communities especially. Populations are very dynamic and therefore highly dependent on environmental fluctuation. These elements should therefore be included when aiming at population scale comprehension. So far, attempts to reach the microbial population levels with precision have been made via *in vitro* experiments in very controlled systems and modeling (Mas et al., 2016; Sanchez & Gore, 2013). New methods are being developed to solve this issue, notably single-cell omics which are becoming increasingly popular in microbiology. Promising to overcome major limitations regarding population investigation via metagenomics, this technique has already been applied to multiple environmental samples (Berube et al., 2018b; Pachiadaki et al., 2019a; Stepanauskas et al.,

2017; Zheng et al., 2022) but is still in development and presents limitations related to cost, contamination, and low throughput (Gawad et al., 2016a).

After centuries of discoveries and technological improvements, the study of microbial populations is the next challenge in microbial ecology in terms of methodology, molecular analyses, and data interpretation.

Context and objectives

The study of bacterial populations and communities from natural habitats has been accelerated in recent years with single-cell omics approaches, conducted via various methodologies, suited for the type of sample analysed and the goal of the study. There are still many limitations to single-cell genomics, notably the low genome coverage, the high cost of materials, and the contamination of samples from diverse sources.

This work aimed to counteract these limitations to broaden the possibilities of single-cell genomics application to microbes. This has been done by following four large questions:

- What is the current state of single-cell omics for microbes, and how can it be improved theoretically?
- In practice, what is the optimum strategy for single-cell genomic application on bacteria for limited contamination and high throughput with technical tools currently available?
- How should the data be handled for the purest and largest genome production?
- How well can we recover bacterial information from single-amplified genomes (SAGs) and what are their advantages compared to traditional metagenomics on environmental samples to respond to ecological questions?

A detailed synthesis of uses of single-cell omics and their possible applications for microbes is presented in Chapter One and valorised as a published article in Trends in Ecology and Evolution journal. Steps for sample preparation, potential applications in microbial ecology, and suggestions for robustness improvements of single-cell omics on microbes are discussed.

From this inventory, I developed a protocol for single-cell library preparation applied to an environmental sample that would be cost-efficient, easy to use, and adapt. I engaged most of my time in the development of the protocol after selecting the different approaches to be tested for sample preparation. From exciting molecular reagents and published techniques, I evaluated the efficiency and compatibility of each step, from cell isolation to library

preparation. Each step was first tested individually in bulk, then combined, and an attempt for miniaturisation using Cellenion technologies is discussed. The protocol elaboration is detailed in chapter two, with most of the tests and quality control procedures. This development resulted in experiments to validate the genome recovery on referenced strains. To assess data quality and improve SAGs recovered from our samples, we worked on developing an automated pipeline to clean single-cell genomic data. The results of this experiment and pipeline application are presented in an article in Chapter Two which will soon be submitted to the Nature Methods journal.

Finally, the developed workflow from cell isolation to bioinformatic treatment was applied to soil bacteria communities, a fraction of the microbial world with many implications for plant health, biogeochemical cycles, and ecosystem functioning but which remains blurry to our knowledge. The soil microbiota is very complex and most of its composition and functioning remain to be discovered. We chose to focus on the soil acidity parameter which is known to greatly influence microbiota composition and structure greatly but with little knowledge about the sub-species responses, gene distribution, and phylogenetic diversity. We worked on soil samples from the Craibstone experiment that have been well characterized and have been used as a model for decades to study the pH effect on soil biotic and abiotic properties. Additional steps were incorporated in the protocol for the soil samples treatment to extract the cells from the soil matrix. I also used metagenomics and mini-metagenomics on the same samples to compare the approaches' outputs and their capacity to describe natural bacterial communities. The results of this experiment are detailed in Chapter Three.

With this work, I aimed to overcome some of the limitations of single-cell omics on microbes to propose a less biased and cost-effective approach for future improved applications of this method. I discussed ways to improve sample preparation and the implications of this work for microbial ecology. More broadly, this work opens up new avenues of research, and questions our habits in microbiology for sample sampling, data analysis, and storage which, independently of technical limits, solely relies upon proper scientific methodologies that should be systematically applied.

Chapter I

Single-cell omics applied to microbiology

Bacterial cells operate just like a multicellular organism, with metabolic functions, costs, and trade-offs. Bacteria are metabolic machinery on their own and will highly respond to environmental constraints to keep functioning. For this reason, bacterial evolution involves a high rate of mutation and frequent horizontal gene transfers to be able to quickly adapt to various environments. Just like herbivores and carnivores regulate each other, bacterial populations emerge, decay, and interact to conserve the stability of the community. Any external intervention, be it biotic or abiotic, will re-arrange the proportion of each cell type. Comprehending populations from the cell's perspective is the major goal of single-cell omics and their implication in microbial ecology is potentially immense.

Trends in
Ecology & Evolution

Review

Contribution of single-cell omics to microbial ecology

S. Mauger,[1,2,*] C. Monard,[1] C. Thion,[2] and P. Vandenkoornhuyse[1,*]

Micro-organisms play key roles in various ecosystems, but many of their func- tions and interactions remain undefined. To investigate the ecological relevance of microbial communities, new molecular tools are being developed. Among them, single-cell omics assessing genetic diversity at the population and com- munity levels and linking each individual cell to its functions is gaining interest in microbial ecology. By giving access to a wider range of ecological scales (from individual to community) than culture-based approaches and meta- omics, single-cell omics can contribute not only to micro-organisms' genomic and functional identification but also to the testing of concepts in ecology. Here, we discuss the contribution of single-cell omics to possible breakthroughs in concepts and knowledge on microbial ecosystems and ecoevolutionary processes.

Highlights

Microbes are involved in many ecosystems but remain understudied, mostly due to technical limitations.

Application of omics technologies to microbes involves changes in the scales of ecological studies.

Single-cell omics offer the opportunity to study microbes at a finer scale than meta-omics tools.

Single microbial cell omics enable the assessment and exploration of the dynamics of genetic changes from individuals to higher levels of ecological complexity.

Ecological scales

Interactions between organisms take place at all organizational levels, from molecules to communities and within or between species, and shape ecosystem dynamics. Ecological interactions are difficult to understand due to the number of biotic and abiotic parameters involved. Assembling knowledge at various ecological scales and from different standpoints is therefore crucial in the study of ecological and evolutionary processes. This is particularly true in the case of microbes, in which individuals can be seen as metabolic units involved in complex metabolic networks at much higher ecological scales (e.g., [1,2]). Therefore, accessing genetic and metabolic information of microbes is a necessary step to understand ecosystem functioning. In microbial metabolic units, 'small' changes in genomes and metabolic pathways may have significant impact on the microbial community organization and hence on ecosystems. Deciphering processes at large ecological scale therefore requires observation of fine ecological scale (i.e., at the individual cell level), which is the biggest challenge of current environmental microbiology [3] (Figure 1A).

A fundamental level of organization in ecology is the species. However, due to gene flow between cells that increases with ecological overlap and genetic similarity [4], the microbial species con- cept and thus also populations (see Glossary) are not clearly defined entities. Interactions and diversity at the population level (i.e., between individual cells of the same population) (Box 1) are still obscure because they are not often analyzed in environmental microbiology. Given the natural mutation rate in bacteria (~10^{-7} substitutions per nucleotide; e.g., [5]), even a single colony con- tains genetic variations (i.e., variants within a cell population). The population has been suggested to be more relevant than the species level for microbes [6,7], and species usually contain genetically divergent microorganisms. Considering the hierarchical levels of ecology, populations are keys to assessing genetic structure within species and, over time, changes therein. They thereby provide insights into ecoevolutionary processes and advance our understanding of microbiota composition dynamics (Box 1).

[1]Université de Rennes 1, CNRS, UMR 6553 ECOBIO, 35042 Rennes, France [2]Cellenion, 60 Avenue Rockefeller, 69008 Lyon, France

*Correspondence: solene.mauger@univ-rennes1.fr (S. Mauger) and philippe.vandenkoornhuyse@univ-rennes1.fr (P. Vandenkoornhuyse).

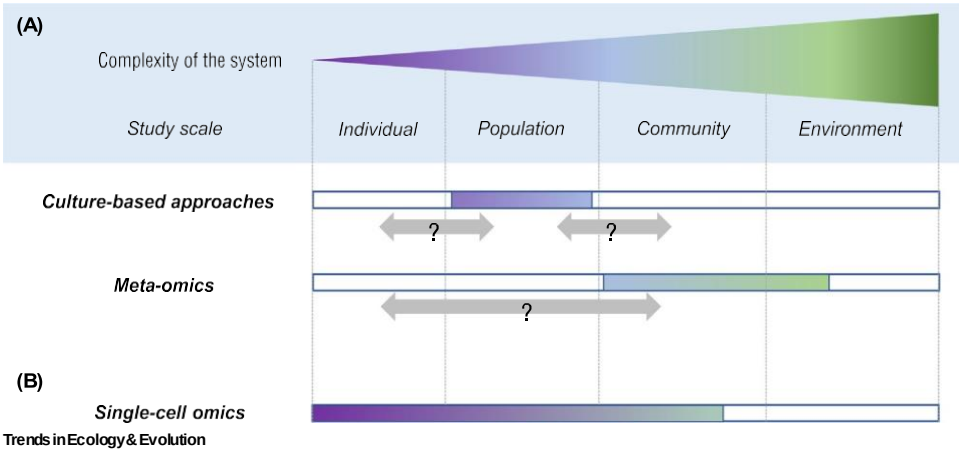


Figure 1. Gradient of ecological complexity, study scales, and associated approaches. Although other higher levels of ecology can be used, individual–population–community–environment scales describe much of the subject of ecology. A) The culture-based approaches aim at studying microbial populations or a very limited number of strains; therefore, the outcoming data cannot be fully informative about higher levels of ecology. On the contrary, meta-omics approaches cover a range from the community to environmental scales of microbial ecology and do not provide finer information on the ecological gradient. In both cases, the individual scale is unattainable while being at the basis of ecological processes. B) Single-cell omics cover the scale from the individual microbe to the community from the same environmental sample, which allows one to connect the outcoming information of each ecological scale.

Our understanding of the microbial world and its ecological roles is still very limited [8]. Understanding the functions played by microbial cells in a complex community remains a frontier in microbial ecology. Beyond the technical limits that microbiology is facing, the information gathered from culture-based studies or from natural ecosystems can be difficult to interpret (Figure 1A). Laboratory experiments attempt to reproduce optimal ecological conditions for microbes by selecting from among the many biotic and environmental parameters [6] in order to understand specific processes such as trait trade-offs [9], interactions between strains [10], the production of metabolites [11], or genome evolution [12]. Extrapolating observations obtained *in vitro*, at restricted scales, to higher ecological scales such as natural communities and ecosystems requires particular attention. Conversely, observations made from environmental samples, including microbial community composition, diversity, or global functions, are less specific and represent an average of the microbial community. Ideally, we want to get the most information out of each level of approach (i.e., precise interactions and genetic dynamics from culture-based studies coupled with global function and diversity of a community with meta-omics). However, our

Glossary

Metagenome-assembled genomes (MAGs): *in silico* reconstruction of an artificial microbial genome obtained from one or multiple binned metagenomes that represent the core genome of the population.

Meta-omics: group of molecular biology technologies, extensively used to access unculturable organisms, by studying the bulk pool of biomolecules from environmental samples to reveal genomes (metagenomics), transcriptomes (metatranscriptomics), proteomes (metaproteomics), and metabolites (metabolomics).

Niche complementarity/ partitioning: ecological concept describing how species differential specialization in different combinations of resource uses and functions allows them to coexist in the same environment.

Populations: applied to bacteria in natural communities for individuals with identical or different genomes from the same species gathered in a specific environment or sample. An isolated micro-organism culture also comprises a population.

Box 1. Microbial population

In ecology, populations are individuals belonging to the same species living in the same environment, although the definition varies with different viewpoints [81]. Microbial populations represent a unit of diversity and selection. Within these populations, diversity can be either genetic or phenotypic. The diversity within a population to some extent buffers an environmental stress because existing variants are able to survive the stress and/or allow rapid phenotype switching (e.g., Bet-Hedging [31,82]), but positive selection of new variants can also be induced by the stress. This organizational level is therefore a key to understanding genetic structure; haplotype fitness; and the dynamics of ecological interactions, including associations of microbial species, symbioses, host–pathogen interactions, and ecosystem functioning, resilience, and stability. For instance, resistance to antibiotics can vary within populations [83], and the virulence pathogens can vary across subpopulations [84]. Genetic diversity and ecological features such as niches can vary between lineages [85], so that subpopulations are able to coexist through niche diversity. To capture the total genetic and phenotypic diversity and get a holistic view of populations, the scale must thus be tuned down to the individuals that compose the population in a given sample [84]. Otherwise, applying the current bacterial species concept to make population-level inferences may lead to false or partial interpretation of ecological phenomena.

ignorance of intricacies and interactions of ecological scales with one another still jeopardizes the assembly of the resulting information to answer specific ecological questions in environmental microbiology. Moreover, the uncertainty of community composition and the complexity of microbial interactions [2] make it even more difficult to target specific scientific hypotheses on natural communities and to choose the appropriate tools.

Specific tools for specific ecological questions

Like in any new field of exploration, ecological patterns within microbial communities are first observed and described but poorly understood [13], testifying to the enormous lack of knowledge concerning microbes [14]. The use of DNA- and RNA-based methods to study natural microbial communities has demonstrated the existence of a prodigious wealth of micro-organisms that remained unsuspected some years ago (e.g., [15]). Among meta-omics techniques, metagenomics and metatranscriptomics are the most widely used methods to explore microbiota. These techniques enabled a breakthrough in our understanding of microbial phylogenetic relationships [16], species diversity and abundance [17], metabolic abilities [18], and functional diversification [19]. The development and use of metagenome-assembled genomes (MAGs) led to discoveries that advanced our understanding of bacterial life and modified our perception of the tree of life [15,19]. Some studies attempted to reconstitute population-level genomes from metagenomes. For example, Crits-Christoph *et al*. [20] investigated genetic variation within populations of highly abundant soil bacteria by studying MAGs and observed spatial differentiation of alleles. However, inferring population features from meta-omics data remains limited, especially when genomes are inaccurately or incompletely reconstructed from short sequence fragments (Figure 2, Key figure). The use of MAGs becomes challenging when microbial richness and diversity within a community are high and taxa are phylogenetically close [21]. During genome assembly, stitching of fragments from different individual genomes and/or contaminant DNA can occur, creating chimeras that are irrelevant for the study of populations. In this case, the approach would necessarily conceal a considerable proportion of molecular diversity [22]. In addition, the molecular biology and bioinformatics methods used in meta-omics approaches are varied and based on different criteria and as- sumptions in the absence of a consensus, leading to contrasting results and interpretations [23,24]. Overall, it might be difficult to directly link the detected functions to their original microbial cell from meta-omics data, thereby limiting the identification of signaling pathways and trade-offs in gene regulation. Meta-omics approaches proved useful in describing communities using large-scale sampling and have made it possible to answer questions related to community composition and its associated global functions but not to fully understand the mechanisms underlying these patterns. Nevertheless, bioinformatics research has developed algorithms aiming to identify genetic variations in microbial populations: Vertically and horizontally inherited genes can be differentiated, and, from population-specific sweeps, SNPs can be detected (e.g., DiscoSNP, PopCOGenT) [4,25].

To complement meta-omics data, modeling approaches are used to explore microbial interactions and fluxes of metabolites and to reconstruct ecological networks in complex microbiomes [26–29]. These approaches provide a possible explanation and scenarios of interactions in natural communities, but they sometimes end in contradictions with culture-based experiments [1]. Indeed, models can predict a certain kind of interaction (e.g., cross-feeding) that is not verified or proven wrong in an experimental setup due to the oversight of key parameters such as growing conditions, space, and (very often) time [1]. They also rely on the co-occurrence of phenomena, which is more associated with correlations than cause–consequences relationships. Culture-dependent approaches may help reach the population level in simple community compositions through controlled and simplified laboratory-scale experiments [30] and can be effective for testing patterns observed in meta-omics studies and deconstructing mechanistic hypotheses (e.g., interactions, metabolism) (e.g., [31]).

Key figure

Microbial communities observed through meta-omics versus single-cell omics

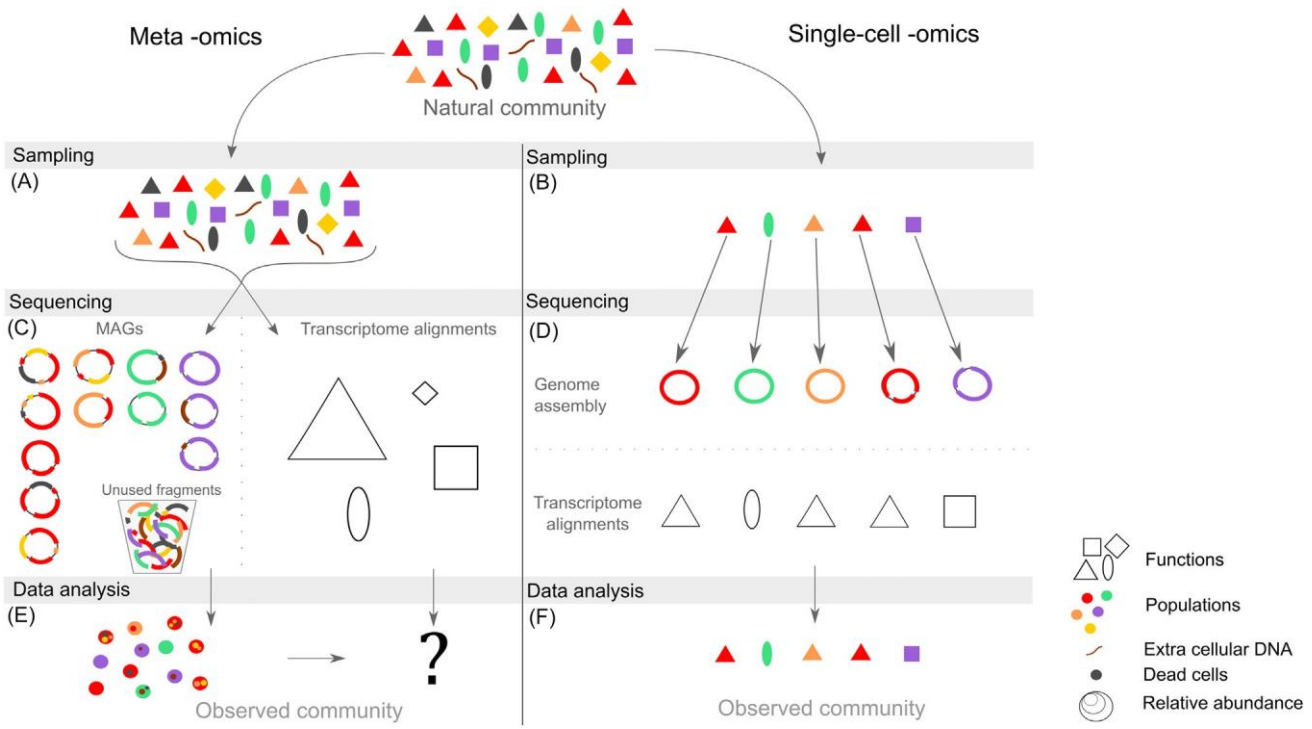


Figure 2. The key steps of meta-omics (left) and single-cell omics (right) approaches are shown, resulting in contrasting representations of a natural microbial community. Different genomes are represented by different colors: Red, orange, and yellow show genomes of close relatives (i.e., intrapopulation genomic variants). Dead cells are shown in gray and extracellular or host DNA in brown. Different functions are represented by different symbols (triangles, squares, diamonds, or ovals). (A) After meta- omics sampling, the cell proportions are maintained, but transient DNA and dead cells are not filtered. (B) In single-cell omics, a smaller proportion of the community is sampled, and dead cells can be excluded. (C) In meta-omics, the unit sequenced is the complete extracted sample. Metagenome-assembled genomes (MAGs) are partial and include chimeras (i.e., unreal collages of closely related genomes, dead cells, and extracellular material). Meta-transcriptomic analysis yields averaged relative abundances (represented by the size of the symbol) of functions within the sample. (D) By contrast, in single-cell omics, each cell is a sequenced unit and can be associated with its genome and/or transcriptome. (E) The community observed through meta-omics is representative of the composition of the whole community but not of the associated genes and functions. (F) With genome and transcriptome information from single cells, the observed community is undersampled but is closer to the natural community: If the sampling scale is appropriate, rare populations and functions are more likely to be detected.

The lack of information on individual cells and mostly on populations (i.e., both functions and phylogeny) using existing methods limits our understanding of observed processes. Many studies aim at unraveling microbiotic diversity, primarily in plants, soil, water, and animal bodies, but few explain associated community assembly and evolutionary mechanisms [32–34]. In this context, microbiolo- gists and ecologists are searching for other technical possibilities or approaches, such as single-cell omics, to complete the knowledge provided by current methods.

The alternative scope of single-cell omics

Single-cell whole-genome sequencing (scWGS) and single-cell transcriptomics [single-cell RNA sequencing (scRNAseq)] were first developed for eukaryotic cells and used in cancer research,

revealing both intrapopulation genetic diversity and heterogeneous genome expression. As in cancer research, where differentiation in space of the genome expression among cells has been observed [35], a pioneer paper on *Pseudomonas aeruginosa* biofilm using a fluorescence- based approach (i.e., parallel sequential fluorescence *in situ* hybridization) revealed a differentiation in space and time of cell expression [36]. Because spatial single-cell microbial approaches could allow the understanding of the drivers and mechanisms leading to the self-organization of these microbial structures, new developments are expected to expand in health science and many other fields of microbial ecology research. Single-cell approaches are promising candidates for microbial studies because they provide a complementary view to metagenomics and metatranscriptomics that have different strengths but also weaknesses (Figure 2).

Single-cell omics technologies require additional steps to prepare a sample for sequencing as compared with meta-omics techniques, especially with regard to cell isolation, for which different technical options are available [37,38]. Once the cells are lysed, DNA and RNA content from a sin- gle cell is in the femtogram scale for bacteria (i.e., 1000-fold less than in animal cells). Preparing the sequencing library, which typically requires nanogram ranges of material, will need an ultraefficient prior amplification step [e.g., multiple displacement amplification (MDA), the most widely used approach for bacteria] [39].

Single-cell approaches enable accurate access to genomic and transcriptomic information for each cell, so that the assembled cell information is highly representative of the original population (Figure 2). This enables the identification of heterogeneity in gene assemblage, gene expression, and metabolic pathways between cells. Single-cell transcriptomic and genomic information provides a link between phylogeny and functional traits and reveals the physiological status of an individual cell at the time of sampling. This is particularly important, considering that the individual gene expression profiles of genetically close cells may differ. What is more, some cell isolation tools, such as automated image-based isolation devices (cellenONE, Cellenion; and ICELL8, Takara Bio), make it possible to select cells on the basis of their integrity, their physiology, and/or their functional markers and to minimize contamination by the host or extracellular DNA. This is very promising for microbiology to, for instance, select active cells in the studied sample at the time of sampling and reveal which of them are taking part in the community productivity.

A seminal paper on single-cell microbial genome analysis was published in 2005 [40] and paved the way for further improvement of single-cell omics, notably on the amplification method (here MDA) and lysis buffer. Recent studies using single-cell omics have improved our understanding of intraspecific diversification and metabolic capacities at a limited scale [41,42]. Assessing the true individual cell gene assemblage and expression using single-cell omics will make it possible to study the hitherto unexplored microbial population level and the functioning of a given microbiome by linking the different ecological scales (Figure 1B). Indeed, single-cell omics enable access to information additional to that in culture-based and meta-omics studies (the single indi- vidual ecological scale) while also, from the environmental sample, giving information on the population and community interactions. To a broader extent, this will enable better access to ecoevolutionary pressures and evolutionary processes within microbial communities.

Applications of single-cell omics in microbial ecology

Single-cell omics provide information at the cell level by changing the camera angle when studying environmental communities and can contribute to microbiology and ecology at many levels by exploring microbial diversity or microbial interactions.

The tremendous diversity of single-microbe genomes

Observations of microbiota can complement/validate the diversity observed by meta-omics on fungi [43], human samples [8], and marine viruses [44] and can resolve cryptic bacterial species, which currently mainly rely on cultivable strains [45–47]. Single-microbe omics therefore contrib- ute to the microbial inventory,

which is still in its infancy in many ecosystems [48]. The use of single-cell–based approaches could demonstrate the existence of the discovered sequence-based lineages and discover possible new branches. Isolating cells from environmental samples can also cast light on rare organisms that might be obscured in millions of genome fragments using meta-omics. These rare organisms are considered to play key roles in community dynamics because they overproportionally contribute to the functions of the microbial community in fluctuating environments [49,50]. Although most studies that apply this approach use a limited number of cells, a recent survey of the marine microbiome recovered no less than 12 000 genomes from single cells [51], revealing a high degree of uniqueness and limited clonality in the analyzed samples of seawater and providing evidence for the ecological roles of uncultured microbial groups. Single-cell genomics, by looking at individual genomes instead of core genomes from meta-omics, from natural microbial communities represent an unprecedented opportunity to complete the identification and classification of microbes. This is a key step sometimes missing in environmental microbiology [52]: knowing what to look at and why to formulate hypotheses in ecology and better understanding processes involving microbes.

Ecological and evolutionary hypothesis testing using single-cell omics

Single-cell genomics and transcriptomics approaches therefore help to answer the questions 'What are these microbes?' and 'What are they doing (or capable of doing)?'. They also help to understand why and how observed patterns happen. Linking environmental and community parameters to individual gene expression and bacterial interactions enables a mechanistic understanding of underlying biotic and abiotic conditions to patterns. This represents an opportunity to explore multiple ecological theories and hypotheses, notably on interactions of microbes at many levels: within the community, with external microbes (i.e., viruses), and with their host. One of the hottest topics in microbial ecology is the link between diversity and function, including the productivity of the ecosystem that relies on the niche partitioning theory. This hypothesis states that species co-existence is enabled by species specialization in different available resources (or combinations of resources), thereby reducing interspecific competition [53] but likely modifying the microbial population structure [54]. Specialization in specific resources raises many questions concerning microbial interactions through the exchange of metabolites [55], loss of traits [56], and genome reduction [57]. The Black Queen Hypothesis (BQH), one of the ecological theories that conceptualized this phenomenon, states that microbial community assembly and complexity are at least partially determined by functional dependencies resulting from gene loss(es). Testing this hypothesis requires using environmental samples to evaluate functional redundancy within communities, the expression and distribution of the functions between interacting (micro-)organisms, and the impact of genotypic interactions on these functions. At the community level, it is impossible to access this information through meta-omics because BQH evolution is supposed to also occur at the population level [53,54], which is undetectable by most of the meta-omics tools used so far. Ecoevolutionary processes can so far be explicitly assessed only through ecological models [58] or from dedicated *in vitro* experiments [42] that require deciding which gene and which organisms to look at. Single-cell approaches can help to investigate such hypotheses and other theories in ecology and highlight patterns of interactions that shape microbial communities.

A breach to viral host ecology

The diversity of microbial communities is also influenced by the viral infections to which they are exposed, which is particularly difficult to evaluate in nature [59]. A recent study assessed viral infections in the ocean thanks to single-cell genomics by identifying virus sequences in uncultured protist cells [44]. Such interactions are widespread but generally missing in current microbial ecology analyses. Among other roles, viruses are known to (i) control microbial community dynamics and drive microbial host evolution [60,61] and (ii) impact ecosystem changes and biogeochemical cycles [59,62]. Single- cell omics can help us understand the roles played by viruses in microbial populations by making it possible to assess both the prevalence of prophage sequences and possible lateral gene transfers [63,64], which would reveal preferential association of specific viruses with specific bacteria.

Single-cell insights into bigger-scale interactions

Generally speaking, single-cell omics can highlight diverse levels of interaction in natural habitats either between the microbes that compose a microbiota or between the microbes/microbiota and their host. Despite the great number of studies on the composition of human microbiota, very little is known about interactions between the microbes and with their host [65]. The single- cell study of host–pathogen interactions paves the way for understanding infectious processes through microbiota dynamics, metabolic capacities, and host resistance [66]. Dysbiosis is often shown to display a higher β-diversity interpreted as a higher stochasticity in the microbiota assembly [67]. Among other explanations, this apparent higher stochasticity may be the result of drastic pathogen-induced changes in the habitat (i.e., transitory state in the microbiota dynamics), modification of a component of the microbiota caused by genetic change(s), and/or a functional modification expressing a modified phenotype that leads to disequilibrium in the microbiota community. With the aim of understanding how a disorder of the microbiota leads to disease or the reverse (i.e., how a disease can modify the composition of a microbiota), single-cell microbiota analyses of genomes and transcriptomes would help better define the characteristics of dysbiosis (i.e., dysbiosis mechanisms). Using single-cell genomics makes it possible to address hypotheses related to changes in bacterial populations, whereas microbial single-cell transcriptomics may be more appropriate to decrypt the functions of microbes, metabolic abilities, and cellular states [68]. These two strategies are necessary to understand how microbial interactions occur within communities as well as their possible impact on the ecosystem [51,69].

Microbial single-cell omics are expected to improve understanding of not solely functional interactions but also the underlying evolutionary processes. Microbial single-cell omics should also promote a shift in standpoint from observation to interpretation and also offer new opportunities to test macroecological theories on microbes. The development of microbial single-cell omics will have high impacts on our understanding of microbial communities in many environments, such as (i) in freshwater and marine ecosystems, to define the interaction of bacteria and phytoplankton through the exchange of metabolites and to test links to blooms [70]; (ii) in soils, to better assess the provision of services by plant microbiota, including nutrient and water uptake and protection against pathogens [71]; and (iii) in plant, human, and animal health, to better decipher how dysbiosis could be a cause or consequence of a disease. However, it cannot be ignored that the wide development of single-cell omics applied to micro-organisms is subject to technical limits.

Limits of single-cell omics on microbes

The current limited number of cells studied in the published papers questions the representativeness of the analyses. Considering the number of microbial cells contained in a given environmental sample, one can wonder how many cells need to be isolated to cover the diversity of a sample, from hundreds [41] to thousands [51]. The limited number of analyzed cells is mostly due to technical problems and the cost of such experiments. It has to be emphasized that the current use of single-cell omics for microbes must solve many technical obstacles (Box 2), reviewed in [72,73], such as cell isolation, lysis, and a biased amplification step. The structure of the microbial cell wall is complex, and, unlike animal cells, they do not break easily. The diversity of cell wall composition across phylogeny and physiological status (e.g., peptidoglycan layers, spores, capsules) makes it challenging to find a universal lysis method able to breach each cell without dam- aging its content or inhibiting enzymatic reactions downstream. Different protocols have been used in recent microbial single-cell–based studies, using either heat, temperature shocks, sonication, enzymes, detergents, or combinations of these [74,75].

Box 2. Single microbial cell omics approaches are technically challenging

Crucial but solvable issues (Table I) should be addressed at each step of single-microbe approaches (Figure I). First, microbial cells need to be properly isolated from complex environmental samples (1). The quality of this step is critical because it will directly impact the following applications, whether it concerns culturomics (2) or molecular analyses. The latter demands a prior lysis of microbial cells once isolated (3), which might appear simple but was and is still a major padlock in microbiology. Once the molecular material is available, whole-genome (4) and whole-transcriptome (5) sequencing can be performed, which both require particular attention to aspects listed in Table I. For each of these five steps, solutions are proposed in Table I to solve the associated challenges. However, the combination of the solutions represents an additional complication because the combined solutions might not be employable within the same protocol. The reduction of reaction volumes down to 'nanovolumes' is likely the most sensible solution, limiting contamination probability and allowing high throughput and cost reduction. Overall, single-cell omics applied to microbes need to focus on three guidelines: representativeness from molecules to samples, compatibility between the steps of sample preparation, and care throughout the process.

Challenge	Possible solution
1. Isolation of single microbial cells	
Community representative sampling	High throughput
Cell isolation from complex matrices, such as soil, sediments, host tissue, feces, mucus, among others	Sonication, filtration, density gradient centrifugation
Exhaustive/targeted labeling/detection	Fluorescence, antibodies
Maintenance of axenic conditions	(As) clean (as possible) room + Ⓥ[a]
2. Culturomics experiments	
Maximum viability/cultivability	Gentle cell handling, liquid dispensing
Choice of culture conditions	High-throughput media screening + Ⓥ
Assessment of monoclonality (culture purity)	Microscopy, targeted sequencing
3. Microbial cell lysis/permeabilization	
Efficiency across phylogeny	(Ultra-)sonication, thermal shock, heat, enzymolysis, detergents, among others
Preservation of DNA/RNA quality and quantity	Gentle procedure, avoid purification
Minimum contamination from reagents and prevention of subsequent steps	Physical rather than biological/chemical + Ⓥ
4. Single-microbe WGS	
Superefficient (100 000– to 1 million–fold) amplification (1–10 f. DNA per prokaryotic cell)	Ⓥ, Molecular crowding, linear amplification (e.g., *in vitro* transcription)
Even/broad coverage, high fidelity, and no chimera creation	Minimum number of PCR cycles and/or primary template amplification (e.g., *in vitro* transcription)
Minimum contamination from reagents	Minimum reagent amount + Ⓥ
Cost reduction	Cell barcoding for multiplexing + Ⓥ
Bioinformatics	Dedicated tools for cell demultiplexing, monoclonality test, and so forth
5. Single-microbe RNA sequencing	
Same as WGS (up to 100 f. RNA per prokaryotic cell)	Ⓥ, Molecular crowding
Amplification bias	Unique molecular identifier
No polyadenylation tail on prokaryotic mRNA	RNA polyadenylation tailing, random priming, ribosomal RNA targeted depletion

[a]Ⓥ,

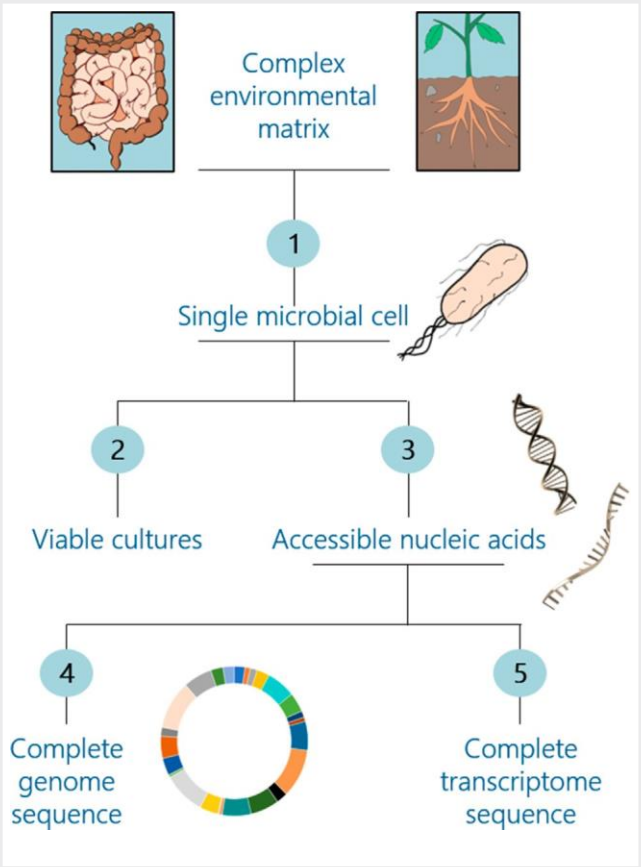


Figure I. Typical steps of single microbial cell omics approaches.

Outstanding questions

To what extent can a single technical approach realistically reflect complex natural processes?

How many cells need to be isolated from a natural environment to accurately represent the population and/or community from which they originate? And how can this be assessed?

What criteria should be used to determine the scale of sampling in natural environments?

How can we use information obtained by single-cell omics to gain a better un- derstanding of ecosystem functioning?

How can we make the application of proteomics, metabolomics, and multi- omics approaches to single microbes more realistic?

In addition to technical issues, one can wonder how to be certain that cells are isolated and lysed equally and not preferentially, depending on their physical/physiological status. The amplification step, usually made via MDA, has been reported to be imprecise concerning the genome amplification uniformity, even though it presents a better genome coverage than other approaches [i.e., multiple annealing and looping–based amplification cycles (MALBAC) [39]]. This amplification step is highly relevant as it was suggested to be the cause of incomplete reconstruction of single amplified genomes [76], although solutions are being developed [77]. As the price of library preparation represents most of the cost of these new single-cell omics for microorganisms, reduction of reaction volumes in 'nanolibraries' should be very cost-effective (Box 2). The probability of contamination decreases with the miniaturization of the reagent volumes [78] and associated robotics. Working in nanovolumes seems to be a convenient solution to solve multiple problems; however, this introduces new volume-related challenges such as pipetting or sample purification.

For these reasons, single-cell omics have sometimes been used in combination with meta-omics to combine the possibility of fine-scale analysis with high throughput [76,79]. It also represents a good opportunity to validate multiple aspects of single-cell omics: (i) the isolation and lysis universality, (ii) the sample preparation (genome amplification and library preparation for sequencing), (iii) the lack of contamination, and (iv) the representativeness of the sample covered by single-cell omics (see Outstanding questions).

Concluding remarks

Soon, single-cell omics applied to micro-organisms could become a gold standard in microbial ecology thanks to the knowledge produced by focusing on individual genomes and transcriptomes and, possibly, individual proteomes and metabolomes. Today, technical problems prevent the testing of broad ecological hypotheses. Generalizing ecological single-cell studies on microbes requires the development of robust high-throughput techniques with a high cost-effectiveness ratio (see Outstanding questions). A knowledge upshot is expected in microbial interactions and ecoevolutionary boundaries through the enabling of mechanistic characterization of deterministic populations and community assembly processes. Currently, the use of metagenomics and single-cell genomics in the same study appears to be the best solution, combining the strengths of the two approaches: (i) high throughput and α/β diversity and (ii) fine-scale analysis by scWGS and/or scRNAseq [76,80]. Ideally, one would not overinterpret meta-omics data and rather would use those data to build hypotheses based on mechanisms, which can be tested using single-cell approaches. Approaches that will allow more accurate assessment of microbial genome diversity and genome functioning within complex microbiota are impatiently awaited. Still, the future of microbial single-cell omics will likely fuel a new perception of the world of micro-organisms.

Acknowledgments

We thank ANR LABCOM 'Microscale-Lab' for funding.

Declaration of interests

The authors have no interests to declare.

References

1. Coyte, K.Z. and Rakoff-Nahoum, S. (2019) Understanding competition and cooperation within the mammalian gut microbiome. *Curr. Biol.* 29, 538–544
2. Pacheco, A.R. and Segrè, D. (2019) A multidimensional perspective on microbial interactions. *FEMS Microbiol. Lett.* 366, fnz125
3. Dini-Andreote, F. *et al.* (2020) Towards meaningful scales in ecosystem microbiome research. *Environ. Microbiol.* 23, 1–4
4. Vanlnsberghe, D. *et al.* (2020) How can microbial population genomics inform community ecology? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375, 20190253
5. Westra, E.R. *et al.* (2017) Mechanisms and consequences of diversity-generating immune strategies. *Nat. Rev. Immunol.* 17, 719–728
6. García-García, N. *et al.* (2019) Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. *ISME J.* 13, 2969–2983
7. Niccum, B.A. *et al.* (2020) Strain-level diversity impacts cheese rind microbiome assembly and function. *mSystems* 5, e00149-20
8. Marcy, Y. *et al.* (2007) Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 mi- crobes from the human mouth. *Proc. Natl. Acad. Sci. U. S. A.* 104, 11889–11894
9. Moreno-Gámez, S. *et al.* (2020) Wide lag time distributions break a trade-off between reproduction and survival in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 117, 18729–18736
10. Ratzke, C. *et al.* (2020) Strength of species interactions deter- mines biodiversity and stability in microbial communities. *Nat. Ecol. Evol.* 4, 376–383
11. Rütting, T. *et al.* (2021) The contribution of ammonia-oxidizing archaea and bacteria to gross nitrification under different sub- strate availability. *Soil Biol. Biochem.* 160, 108353
12. Hall, J.P. *et al.* (2018) Competitive species interactions constrain abiotic adaptation in a bacterial soil community. *Evol. Lett.* 2, 580–589
13. Tripathi, A. *et al.* (2018) Are microbiome studies ready for hypothesis-driven research? *Curr. Opin. Microbiol.* 44, 61–69
14. Lloyd, K.G. *et al.* (2018) Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *mSystems* 3, e00055-18
15. Hug, L.A. *et al.* (2016) A new view of the tree of life. *Nat. Microbiol.* 1, 16048
16. Alteio, L.V. *et al.* (2020) Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil. *mSystems* 5, e00768-19
17. Daims, H. *et al.* (2015) Complete nitrification by *Nitrospira* bacteria. *Nature* 528, 504
18. Cernava, T. *et al.* (2017) Deciphering functional diversification within the lichen microbiota by meta-omics. *Microbiome* 5, 82
19. Parks, D.H. *et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2, 1533–1542
20. Crits-Christoph, A. *et al.* (2020) Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. *ISME J.* 14, 1834–1846
21. Nayfach, S. *et al.* (2019) New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568, 505–510
22. Nelson, W.C. *et al.* (2020) Biases in genome reconstruction from metagenomic data. *PeerJ* 8, e10119
23. Li, L. *et al.* (2015) Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent. *J. Virol. Methods* 213, 139–146
24. Mande, S.S. *et al.* (2012) Classification of metagenomic sequences: methods and challenges. *Brief. Bioinform.* 13, 669–681
25. Uricaru, R. *et al.* (2015) Reference-free detection of isolated SNPs. *Nucleic Acids Res.* 43, e11
26. Antoniewicz, M.R. (2020) A guide to deciphering microbial inter- actions and metabolic fluxes in microbiome communities. *Curr. Opin. Biotechnol.* 64, 230–237

27. Belcour, A. *et al*. (2020) Metage2Metabo, microbiota-scale met- abolic complementarity for the identification of key species. *eLife* 9, e61968
28. Gorter, F.A. *et al*. (2020) Understanding the evolution of interspe- cies interactions in microbial communities. *Philos. Trans. R. Soc. Lond. B Biol. Sci*. 375, 20190256
29. Perez-Garcia, O. *et al*. (2016) Metabolic network modeling of microbial interactions in natural and engineered environmental systems. *Front. Microbiol*. 7, 673
30. Morin, M. *et al*. (2018) Changes in the genetic requirements for microbial interactions with increasing community complexity. *eLife* 7, e37072
31. Beaumont, H.J.E. *et al*. (2009) Experimental evolution of bet hedging. *Nature* 462, 90–93
32. Kiers, E.T. *et al*. (2011) Reciprocal rewards stabilize cooperation in the mycorrhizal symbiosis. *Science* 333, 880–882
33. Werner, G.D. and Kiers, E.T. (2015) Partner selection in the mycorrhizal mutualism. *New Phytol*. 205, 1437–1442
34. Amin, S.A. *et al*. (2015) Interaction and signaling between a cosmopolitan phytoplankton and associated bacteria. *Nature* 522, 98–101
35. Alon, S. *et al*. (2021) Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science* 371, eaax2656
36. Dar, D. *et al*. (2021) Spatial transcriptomics of planktonic and sessile bacterial populations at single-cell resolution. *Science* 373, eabi4882
37. Gawad, C. *et al*. (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet*. 17, 175
38. Woyke, T. *et al*. (2017) The trajectory of microbial single-cell sequencing. *Nat. Methods* 14, 1045
39. Chen, Z. *et al*. (2017) Tools for genomic and transcriptomic analysis of microbes at single-cell level. *Front. Microbiol*. 8, 1831
40. Raghunathan, A. *et al*. (2005) Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol*. 71, 3342–3347
41. Engel, P. *et al*. (2014) Hidden diversity in honey bee gut symbionts detected by single-cell genomics. *PLoS Genet*. 10, e1004596
42. Hennon, G.M. *et al*. (2018) The impact of elevated CO_2 on *Prochlorococcus* and microbial interactions with 'helper' bacte- rium *Alteromonas*. *ISME J*. 12, 520–531
43. Ahrendt, S.R. *et al*. (2018) Leveraging single-cell genomics to ex- pand the fungal tree of life. *Nat. Microbiol*. 3, 1417–1428
44. Castillo, Y.M. *et al*. (2019) Assessing the viral content of uncul- tured picoeukaryotes in the global ocean by single cell geno- mics. *Mol. Ecol*. 28, 4272–4289
45. Hahn, M.W. *et al*. (2016) Complete ecological isolation and cryp- tic diversity in *Polynucleobacter* bacteria not resolved by 16S rRNA gene sequences. *ISME J*. 10, 1642–1655
46. Hoetzinger, M. *et al*. (2019) *Polynucleobacter paneuropaeus* sp. nov., characterized by six strains isolated from freshwater lakes located along a 3000 km north-south cross-section across Europe. *Int. J. Syst. Evol. Microbiol*. 69, 203–213
47. Pitt, A. *et al*. (2019) *Aquirufa antheringensis* gen nov., sp nov. and *Aquirufa nivaliisiivae* sp nov., representing a new genus of widespread freshwater bacteria. *Int. J. Syst. Evol. Microbiol*. 69, 2739–2749
48. Bernard, G. *et al*. (2018) Microbial dark matter investigations: how microbial studies transform biological knowledge and em- pirically sketch a logic of scientific discovery. *Genome Biol. Evol*. 10, 707–715
49. Shade, A. *et al*. (2014) Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio* 5, 1371–1385
50. Xu, Q. *et al*. (2021) Rare bacterial assembly in soils is mainly driven by deterministic processes. *Microb. Ecol*. Published online April 1, 2021. https://doi.org/10.1007/s00248-021-01741-8
51. Pachiadaki, M.G. *et al*. (2019) Charting the complexity of the marine microbiome through single-cell genomics. *Cell* 179, 1623–1635.e11
52. Prosser, J.I. (2020) Putting science back into microbial ecology: a question of approach. *Philos. Trans. R. Soc. Lond. B Biol. Sci*. 375, 20190240
53. Mas, A. *et al*. (2020) Reflections on the predictability of evolution: towards a conceptual framework. *iScience* 23, 101736
54. Mas, A. *et al*. (2016) Beyond the Black Queen Hypothesis. *ISME J*. 10, 2085–2091
55. Zelezniak, A. *et al*. (2015) Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl. Acad. Sci. U. S. A*. 112, 6449–6454
56. Ellers, J. *et al*. (2012) Ecological interactions drive evolutionary loss of traits. *Ecol. Lett*. 15, 1071–1082
57. Morris, J.J. *et al*. (2012) The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio* 3, e00036- 12
58. Wang, M. *et al*. (2020) Selfishness driving reductive evolution shapes interdependent patterns in spatially structured microbial communities. *ISME J*. 15, 1387–1401
59. Weitz, J.S. and Wilhelm, S.W. (2012) Ocean viruses and their ef- fects on microbial communities and biogeochemical cycles. *F1000 Biol. Rep*. 4, 17
60. Frada, M. *et al*. (2008) The 'Cheshire Cat' escape strategy of the coccolithophore *Emiliania huxleyi* in response to viral infection. *Proc. Natl. Acad. Sci. U. S. A*. 105, 15944–15949
61. Scanlan, P.D. and Buckling, A. (2012) Co-evolution with lytic phage selects for the mucoid phenotype of *Pseudomonas fluorescens* SBW25. *ISME J*. 6, 1148–1158
62. Rousk, J. and Bengtson, P. (2014) Microbial regulation of global biogeochemical cycles. *Front. Microbiol*. 5, 103
63. Roux, S. *et al*. (2014) Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell-and meta-genomics. *eLife* 3, e03125
64. Labonté, J.M. *et al*. (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J*. 9, 2386–2399
65. Hassani, M.A. *et al*. (2018) Microbial interactions within the plant holobiont. *Microbiome* 6, 58
66. Mills, E. and Avraham, R. (2017) Breaking the population barrier by single cell analysis: one host against one pathogen. *Curr. Opin. Microbiol*. 36, 69–75
67. Zaneveld, J.R. *et al*. (2017) Stress and stability: Applying the Anna Karenina principle to animal microbiomes. *Nat. Microbiol*. 2, 17121
68. Hatzenpichler, R. *et al*. (2020) Next-generation physiology approaches to study microbiome function at single cell level. *Nat. Rev. Microbiol*. 18, 241–256
69. McNulty, R. *et al*. (2021) Droplet-based single cell RNA sequencing of bacteria identifies known and previously unseen cellular states. *bioRxiv* Published online March 16, 2021. https://doi.org/10.1101/2021.03.10.434868
70. Landa, M. *et al*. (2016) Shifts in bacterial community composition associated with increased carbon cycling in a mosaic of phytoplankton blooms. *ISME J*. 10, 39–50
71. Vandenkoornhuyse, P. *et al*. (2015) The importance of the microbiome of the plant holobiont. *New Phytol*. 206, 1196–1206
72. Kaster, A.K. and Sobol, M.S. (2020) Microbial single-cell omics: the crux of the matter. *Appl. Microbiol. Biotechnol*. 104, 8209–8220
73. Clingenpeel, S. *et al*. (2015) Reconstructing each cell's genome within complex microbial communities – dream or reality? *Front. Microbiol*. 5, 771
74. He, J. *et al*. (2016) Improved lysis of single bacterial cells by a modified alkaline-thermal shock procedure. *Biotechniques* 60, 129–135
75. Liu, Y. *et al*. (2018) The development of an effective bacterial single-cell lysis method suitable for whole genome amplification in microfluidic platforms. *Micromachines* 9, 367
76. Alneberg, J. *et al*. (2018) Genomes from uncultivated prokary- otes: a comparison of metagenome-assembled and single- amplified genomes. *Microbiome* 6, 173
77. Kogawa, M. *et al*. (2018) Obtaining high-quality draft genomes from uncultured microbes by cleaning and co-assembly of single-cell amplified genomes. *Sci. Rep*. 8, 2059
78. Blainey, P.C. (2013) The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol. Rev*. 37, 407–427
79. Probst, A.J. *et al*. (2018) Differential depth distribution of micro- bial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nat. Microbiol*. 3, 328–336
80. Waples, R.S. and Gaggiotti, O. (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol. Ecol*. 15, 1419–1439

81. Veening, J.-W. *et al*. (2008) Bistability, epigenetics, and bet-hedging in bacteria. *Annu. Rev. Microbiol*. 62, 193–210

82. Chua, S.L. *et al*. (2016) Selective labelling and eradication of -tolerant bacterial populations in *Pseudomonas aeruginosa* biofilms. *Nat. Commun*. 7, 10750

83. Davis, K.M. and Isberg, R.R. (2016) Defining heterogeneity within bacterial populations via single cell approaches. *Bioessays* 38, 782–790

84. Garcia, S.L. *et al*. (2018) Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. *ISME J*. 12, 742–755

Chapter II

Elaboration and validation of the single-cell workflow protocol

Unlike most molecular approaches, there is no universal solution on the market for single-cell genomics sample preparation for microbes because of the existence of major technical padlocks to be broken such as costs, contamination, genome amplification, and single-cell isolation(Mauger et al., 2022). From cell isolation to sequencing, the possibilities for technical designs are many, with their own limitations and biases. The best approach is the one that covers the study requirements, whilst optimizing the extent of data interpretation and being suited to the single-cell isolation device.

I- The procedure to elaborate the protocol

To elaborate the workflow, research of available procedures was made to gather different sample preparation approaches. A list of pros and cons for each method was made based on literature and in collaboration with molecular biologists from Cellenion, from which the approaches were then selected. The different chosen steps were individually tested to set the adequate parameters and their compatibility with the other molecular reactions, from cell lysis to library preparation for Whole Genome Sequencing.

1.1- Current state of single-cell genomic technique for bacteria

While some single-cell studies focus on the 16s rRNA gene fraction to identify community members (Nishikawa et al., 2022), most of them attempt to recover whole genomes from bacteria to respond to more populational-oriented questions (Arikawa et al., 2021; Garcia et al., 2018; Ghylin et al., 2014; Hosokawa et al., 2022; López-Escardó et al., 2017; Zheng et al., 2022) . As the choice of strategies for sample preparation is determined by the scientific

question and biological model, the following argumentation on preferable techniques will be led by representativity and costs independently of specific study requirements other questions might bring. There is however a general pattern to follow for such an experiment detailed in Figure 1, each step is detailed in the next paragraphs. Here, we aimed at recovering genomes from environmental bacteria, therefore our methodological research has been oriented to respond to this purpose.

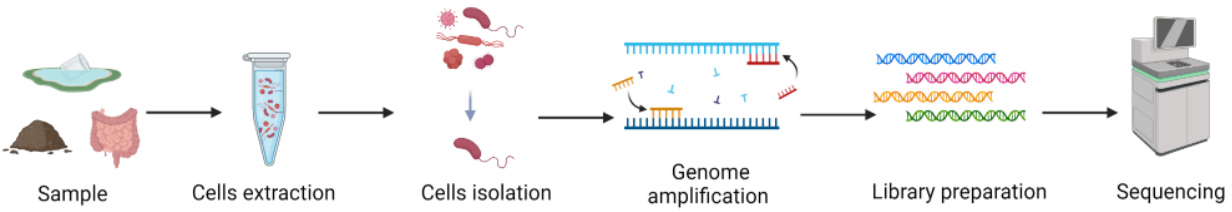


Figure 1. Main steps for single-cell whole genome sequencing (scWGS) sample preparation on bacteria.

1.1.1- Cell isolation

The success of single-cell omics first relies on our ability to isolate cells from other members of the community. For this purpose, diverse techniques are available such as image-based isolation tools, microfluidics, flow cytometry, and optical tweezers (Woyke et al., 2017). Isolation of single bacteria is challenging due to their small size (0.2 to 8 µm), diversity in shape, and the tendency for specific taxa to aggregate (e.g. *Coccus* and *Bacillus* genera). Fluorescent-activated cell sorting (FACS) based on fluorescence and cell morphometry detection using flow cytometry is the most frequently used for cell isolation (Figure 2). This technique presents high throughput and requires the labelling of the cells with fluorescent reagents such as DNA dyes (e.g. SYTO dyes). FACS however does not handle well changes in particle morphologies, applies a high stress on the cells, and does not offer image validation of cell isolation. Other approaches such as microfluidics allow the performance of molecular reactions following the encapsulation of cells, then its lysis, in circulating nano- or pico-litre volume droplets(Liu et al., 2019; Yu et al., 2017)(Figure 2). This effectively reduces the reagent cost and the contamination risk of FACS

approaches. Hence, these droplet-based methods (X. Zhang et al., 2019) have proven their efficiency in high single-cell throughput with very complex technologies(Lan et al., 2017). They however do not offer high flexibility for customized miniaturized library preparation (Woyke et al., 2017) or cell visual observation. Furthermore, whether a droplet contains only one cell relies on chances ruled by Poisson law, so that the vast majority of droplets are empty, and doublets also are generated without the possibility of assessing single-cell accuracy. To avoid such technical limitations, we have chosen an image-based isolation approach, the cellenONE (Cellenion, France) (presented in detail in paragraph 1.2). The cellenONE gives the opportunity to personalize the isolation parameters precisely: with or without fluorescence, isolation of cells with specific size and elongation range, isolation of one or multiple cells, but also to handle liquids in nanoliters for molecular reactions miniaturization. Its piezo-acoustic capillary technology for drop generation is also very gentle and offers a high rate of cell integrity (Busley et al., 2023; Coker et al., 2022). Fluorescent live/dead labelling of the bacterial cell was developed in parallel with the library preparation workflow and is presented in Box 1.

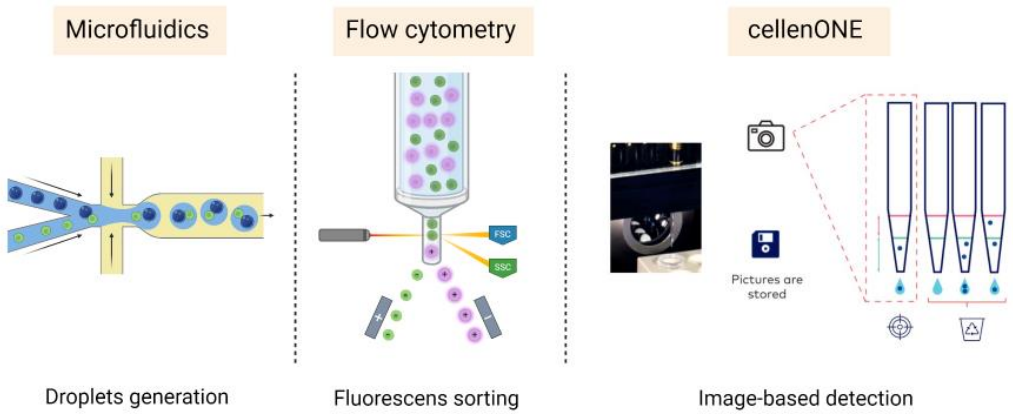


Figure 2. Single-cell isolation tools. Microfluidics and flow cytometers such as FACS are the most common on the market, despite their price and limitations. The cellenONE offers an image-based detection of the cells as well as liquid handling for personalized sample preparation.

1.1.2- Cell lysis and DNA preparation

There are a few different bacterial membrane lysis solutions available on the market for genomics DNA extraction, but they cannot be used for single-cell omics applications because of strong constraints in the lysis protocol adaptation. The challenge is to efficiently and equally break the cell wall and membrane of various types of bacteria of different compositions (Figure 3) without purifying the sample before subsequent molecular biology reactions, which would remove the few femtograms of DNA released from a single cell. Therefore, the lysis reagent mix nature and volume must be compatible with subsequent reactions. There is also no certainty that all taxa will be sensitive to the lysis and no direct way of measuring its efficiency at breaking the wall of all bacteria, as most of them remain unknown. However, the efficiency of each lysis strategy can be compared. There are a few references using lysis buffers for specific microbial cells (e.g. Cyanobacteria (Hall et al., 2013) or soil bacteria (Stepanauskas et al., 2017)), and frequently, the use of a combination of alkaline and temperature treatment is used (He et al., 2016; Liu et al., 2018; Stepanauskas et al., 2017). Precaution must be kept regarding DNA integrity that can be affected by some lysis strategies such as sonication (Fykse et al., 2003).

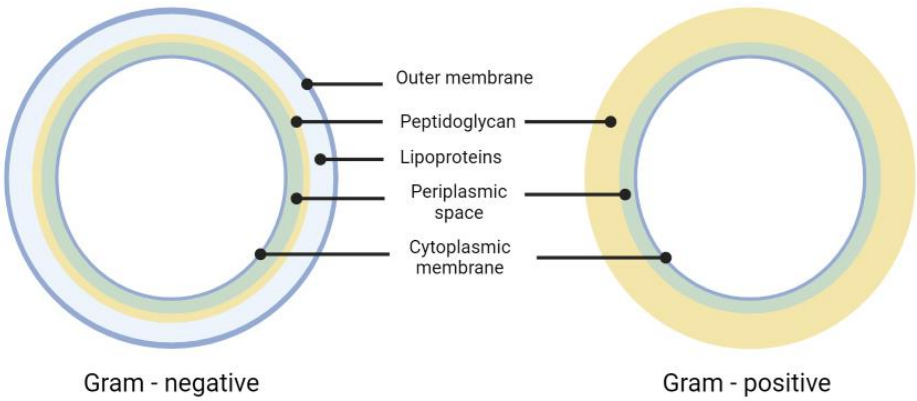


Figure 3. Cell wall of gram-positive and gram-negative bacteria as classical examples of differences in composition and structure.

With only femtograms of DNA per bacterial cell, a step of genome amplification is virtually mandatory. The traditional Polymerase Chain Reaction (PCR) is generally avoided here

to limit the intrusion of amplification errors. Specific amplification procedures are employed for this purpose, linear, semi-linear, or exponential genomic amplifications (Gawad et al., 2016). Commonly, this amplification is done by Multiple Displacement Amplification (MDA) (Figure 4) which can be used via commercialized kits and enzymes. This isothermal and exponential genome amplification with random primers is extremely efficient but lacks coverage uniformity and can over-represent some parts of the genome (Gawad et al., 2016; Woyke et al., 2017). MDA can also introduce errors, and exponentially amplify these errors, although at a lower rate than PCR. Alternatives based on this method have emerged to limit its bias such as X-WGA (Stepanauskas et al., 2017) or Primary Template-directed Amplification (PTA) (Gonzalez-Pena et al., 2021). Very different approaches are also used such as transposon tagmentation and in vitro transcription (IVT) strategies (Chen et al., 2017; Yin et al., 2019), offering linear amplification, and using transposons. Nevertheless, MDA remains the most efficient workflow in terms of yield, genome coverage, low chimera generation rates, and hands-on time, for single-cell DNA amplification (Estévez-Gómez et al., 2018) and its coverage uniformity can be improved by working in lower volumes (Fu et al., 2015; Nishikawa et al., 2015). After this amplification step, the samples contain enough DNA for library preparation (i.e. nanograms per microliter).

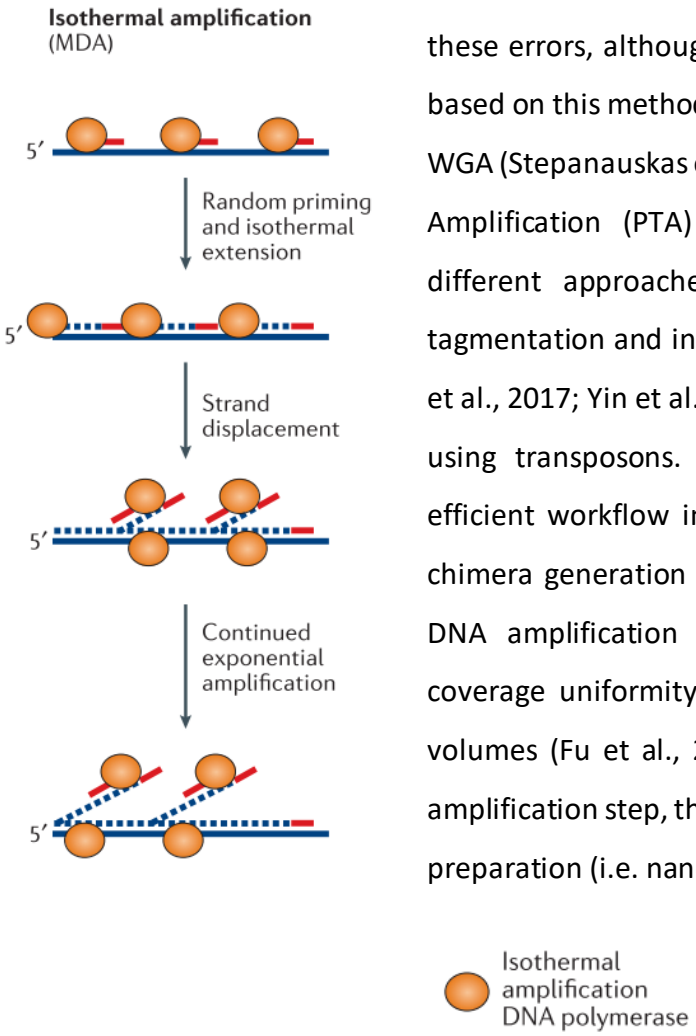


Figure 4. Multiple Displacement Amplification (MDA) using random primers and isothermal exponential amplification with a Phi29 high fidelity polymerase (Figure from Gawad et al 2016).

1.1.3- Sequencing and cost

Sequencing technologies have evolved and propose various possibilities that can be chosen from simple or double indexing, short or long fragments, with different sequencing depths. The goal of single-cell omics is to uncover the understudied diversity of bacteria at both inter- and infra-species levels; one would want to increase the sequencing depth to maximize both SNP detection and genome reconstruction. Increasing the sequencing depth also increases the cost and/or limits the number of samples (i.e. the number of cells here) that can be sequenced. To maximize the number of samples, the use of indexes identifying cells and samples must also be adapted. On the market, library preparation kits generally propose to tag 96 samples (e.g. QIAseq FX DNA Library CDI Kit, QIAGEN), up to 384 (NEBNext® Multiplex Oligos, NEB).

1.1.4- The problematics

Traditional library preparation procedures must be modified for single-cell omics application on microbes to increase the number of cells to be sequenced and get closer to bacterial diversity representativity. Three main actions can be taken to lead this adaptation: 1- Lower the reaction volumes to i) lower the contamination risk, ii) lower the costs, and iii) include more cells into the study. 2-Tag cells and samples in a way that would not limit the study to 384 cells per sequencing and 3- Increase the universality of all steps, mainly those of cell isolation and lysis which are highly dependent on the matrix from which the sample is taken, with specific cell types.

The first objective of this project was to elaborate a single-cell genomic protocol from cell isolation to library preparation applicable to environmental samples, and more specifically, to soil samples. No such complete solution is available on the market, therefore we combined and adapted existing techniques after a complete prospection of available options at the time, presented below.

1.2- The cellenONE and cellenCHIPs

Our protocol was built around two technical solutions proposed by Cellenion: the cellenONE and the cellenCHIP. The cellenONE is an automated isolation and dispensing platform for cells and liquid handling, initially developed and mainly used for eukaryotic cells (Funnel et al 2022, Salehi et al 2021). The particles are detected via automatic visual analysis with a magnifying camera. The samples are handled by a glass capillary (Piezo Dispense Capillary – PDC) able to produce drops with high precision from 150 to 600 pL and are compatible with particles of wide-range sizes (0.5 to 80 µm). The PDC positioning is also highly accurate with a precision of 25 µm in space on the x-, y-, and z-axis. One of the strengths of the cellenONE is the shooting of images of each drop generated, that can be saved (all of them or only those of isolation events). The cell characteristics (size, circularity, fluorescence intensity) are also saved within a table. Fluorescent channels can be activated, and the isolation parameters are customizable as well as the plate in which the distribution is done. The cellenONE does not rely on the Poisson distribution for single-cell isolation and distinguishes between events of isolation and events of drop recycling. There are two areas along the PDC, the ejection zone that corresponds to the next dispensed drop and the sedimentation zone that represents the safety zone which can be manually set (Figure 2). Isolation events will occur only in the situation where one single particle is detected in the ejection zone, and none is present in the sedimentation zone. In any other case, the drop is discarded or may be recovered in a recovery vial if the sample is precious (Figure 2). Pictures of isolation events can be stored for further manual verification of the isolation accuracy.

The cellenCHIP (Figure 5) is a consumable from Cellenion developed for miniaturized omics applications giving the possibility to drastically lower the cost of single-cell sample preparation. The miniaturisation procedure is explained in Box 2. The chip contains 384 wells with working volumes from 50 to 500 nL. Its size is equivalent to a microscope slide and is made of optically clear polypropylene. Up to eight cellenCHIPs can be placed in the cellenONE at the same time for sample processing, for a total of 3072 wells. The use of the cellenCHIP offered a possibility for miniaturized sample preparation for single-cell genomics on microbes but no such

experiments on microbes were previously tested in this chip. Therefore, requirements regarding the cellenONE and the cellenCHIP were considered for the protocol elaboration.

Figure 5. The cellenCHIP design. Image from Cellenion.

These products offer many possibilities but also some restrictions. The requirements for PDCs and cellenCHIP uses are listed in Table 1. Despite its advantages, each reaction happening in the cellenCHIP implies the addition of reagents in each of its wells, followed by sealing and centrifugation to ensure liquid disposal at the bottom of the wells. Therefore, each step represents a long handling time and risk of contamination. The cellenCHIP usage should be as short as possible, and samples should be pooled early in the workflow thus early single-cell barcoding. The barcodes are known and unique sequences of nucleic acids present on primer adapters. Ideally, the barcodes should be present in the wells of the cellenCHIP prior to the sample preparation to avoid washes of the PDC between the distribution of different barcodes. Therefore, these barcodes should not be used as a template before necessary by other enzymes and should also not be degraded. Because the diameter of the cellenCHIP wells is smaller than that of the PDC, no sample uptake can be performed. The design of the cellenCHIP and miniaturized volumes do not allow purification of the samples, demanding high compatibility of each step as well as no sensitivity of enzymes to molecular reaction residues that would generate its inhibition. The protocol should not contain any PCRs within the cellenCHIP but only isothermal reactions due to the limited thermal conductivity of the chip which was still in development at this stage. Lastly, the reagents necessary for molecular reactions or cell lysis should be usable by the PDC, with a limited viscosity.

Table 1. List of major requirements to consider for the PDC and cellenCHIP usage.

PDC	cellenCHIP
Washes between liquids	No purification
Avoid viscous reagents	No PCRs
	No sample uptake

1.3- Combining the most efficient approaches.

1.3.1- Scenarios

The protocol was built to offer efficient and universal lysis and genome amplification while keeping the contamination of external DNA to a minimum. To limit these contaminant DNA as well as costs while increasing the possible number of cells studied, we decided to aim toward a miniaturized library preparation (Box 2), therefore the technical and molecular choices of the protocol were made to answer miniaturisation requirements. Enzymatic and chemical reaction compatibility was a crucial issue, given the impossibility of undertaking purification steps in the cellenCHIP. The number of steps should be reduced as much as possible to limit the risk of contamination from reagents but also from manipulation. All potential cross-compatibilities must be tested beforehand: these tests were conducted in regular volumes (microliters) in tubes and on multiple cells or gDNA (later referred to as “bulk” tests) before miniaturisation to compare the robustness of molecular reactions in different volumes.

We identified strategies for library preparation involving either transposons and in vitro transcription (Chen et al., 2017) or MDA (e.g., Stepanauskas et al., 2017) for genome amplification. For both, details of possible reagents/kit were listed as well as possible steps for barcode addition. We placed the MDA approaches in priority for multiple reasons: (i) we had a possibility to adapt existing kits from QIAGEN, and (ii) MDA is the most efficient way known of amplifying bacterial genomes with relatively robust enzymes that might be less damaged by the remaining lysis buffer than IVT strategies using transposons. The MDA strategies also required fewer reactions in the cellenCHIP and overall quicker sample preparation.

In a second time, we identified the possible moments for barcode additions (Figure 6). As developed above, the barcodes should be added as early as possible to allow sample gathering early in the workflow, giving the possibility to purify between the next steps and manipulate fewer samples. For adapter addition in the cellenCHIP prior to the workflow execution, their design should forbid any other enzyme to use them before the adapter ligation step. We imagined multiple solutions such as using blocking bases by modifying the 3' end of the oligos but only tested a design with hybrid adapters that would be composed of one DNA and one RNA strand, which would prevent the polymerase from using them as templates (Figure 7). However, the tests in bulk were not conclusive (not shown here) and we later focused only on DNA-DNA adapter designs.

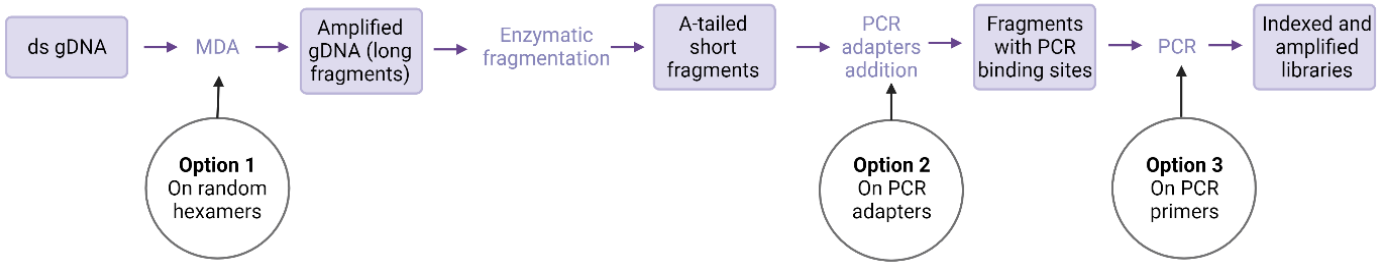


Figure 6. Workflow to be tested with the possible implementations of the barcodes to tag each cell individually.

Another theoretical possibility explored was the addition of the barcodes during the MDA by using transformed random primers containing the barcodes (Figure 6, Option 1). This would have required 3' end modification of the designed oligomers, to prevent the 3—>5' exonuclease activity of the Phi29 polymerase, the enzyme that catalyzes the MDA reaction. We quickly realized that we would obtain very long fragments (up to 30 Kbp after the MDA) with barcodes that would very likely be fragmented during the next steps, therefore losing the barcode information for many fragments. The Cellenion team later explored this approach unsuccessfully with many different barcoded hexamers designs but any personalization of the hexamers seemed to stop the MDA from working. We therefore chose to work on adding the

barcodes during the PCR adapters ligation step (Figure 6, Option 2). The option 3 was not tested due to the incapacity of performing PCRs in the cellenCHIP as previously explained.

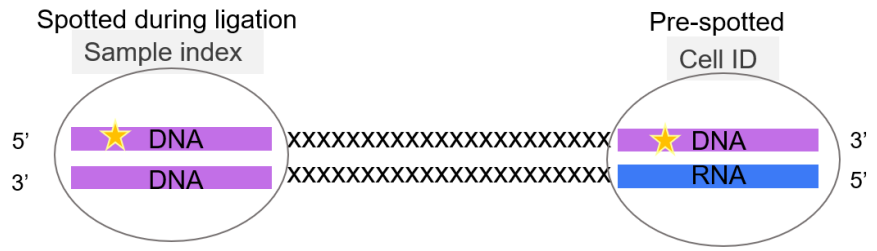


Figure 7. Design of DNA-RNA hybrid adapters. The hybris part would be pre-spotted in the cellenCHIP by Cellenion prior to sample preparation and the DNA-DNA part during the ligation step. The unique cell barcode would have been contained on the hybrid part and the unique sample barcode on the DNA-DNA part.

1.3.2- The lysis

There are multiple options for bacterial cell lysis detailed in the literature: heat, heat shock, alkaline treatment, chemical lysis, enzymatic digestion, or a combination of these (Islam et al. 2017). The use of specific compounds enables to break down the different layers of the bacterial wall (Islam et al. 2017). The validation of the lysis strategy was done by a 16srRNA gene qPCR using QuantiFast SYBR® Green PCR Kit with the 341F (5'-CCTACGGGAGGCAGCAG-3') and 534R (5'-ATTACCGCGGCTGCTGGCA-3') primers used at a final concentration of 1 µM. The qPCR cycle was set as follows: 95°C-5min, (95°C-10s, 60°C-30s, 72°C-30s) x40, with a final melting curve (95°C-5s, 65°C-1min, 97°C-continuous). To prevent cells from breaking with the PCR temperatures, the samples were centrifuged at 10 000 g for 10 minutes and only the supernatant was used to quantify the DNA released by the lysis. Prior to these tests, for each of the three strains tested (*Escherichia coli*, *Bacillus subtilis,* and *Micrococcus luteus*), the bacterial cell count corresponding to the OD600 was estimated by flow cytometry (BD Accuri C6) quantification allowing the control of the number of cells in input for the lysis tests directly by OD600 measurements.

For our application, we first hoped to achieve universal lysis with the least reagents in input as possible to avoid downstream reaction inhibition and limit external contamination. I started to test some temperature treatments in bulk on gram-negative and -positive strains: *E.coli*, *B.subtilis,* and *M.luteus*, with either stable temperatures or heat shocks (Table 2). None of these treatments solely based on temperature increased the available DNA in the suspension compared to the control.

Table 2. Temperatures and incubation times of different lysis tests on gram-positive and -negative bacteria strains.

Control	70°C	90°C	Heat-shock 1	Heat-shock 2	Heat-shock 3	Heat-shock 4
Room temperature 5 min	10s/30s/1min/5min	2s/5s/10s/30s/1min	-80°C 4min + 65°C 2min	-20°C 2 min +65°C 2 min	-20°C 2 min +65°C 2 min X3 cycles	Ice 2min +65°C 2min

From this point we decided to work on lysis buffers, either directly taken from published papers or with some variations to obtain eleven different combinations (Table 3). The first buffer was found in Liu et al. (2018), Buffer 9 in Stepanauskas et al. (2017), and Buffer 11 in Kang et al. (2015). The compatibility of the lysis buffer with the 16s qPCR was first tested as a control so that it would not be disturbed by the different reagents and therefore could be kept as a lysis efficiency validation. The qPCR was run with or without the buffers presented in Table 3 with different concentrations of DNA or water (Negative control). However, by looking at the average cycle threshold of each sample (Table 4), the lysis buffer components disturbed or fully inhibited the 16s qPCR except B7 and B8, preventing the validation of the lysis efficiency via qPCR for the lysis tests with these buffers.

Table 3. Lysis buffers to be tested on multiple strains.

	Buffer 1	Buffer 2	Buffer 3	Buffer 4	Buffer 5	Buffer 6	Buffer 7	Buffer 8	Buffer 9	Buffer 10	Buffer 11
Achromopeptidase			0,5 mg / ml		0,5 mg / ml						
Lysozyme	2 mg /ml	0,5 mg / ml	0,5 mg / ml		0,5 mg / ml	0,5 mg / ml				2 × 10−7 U µl	
Mutanolysin				250 U/ml							
Triton X-100					0,10%	0,10%	0,10%	0,10%		0,10%	0,10%
DTT	200 mM	0,5 mM	0,5 mM	0,5 mM	0,5 mM		0,5 mM	0,5 mM	100mM	2mM	2mM
EDTA	0,5 mM	0,5 mM	0,5 mM	0,5 mM	0,5 mM	0,5 mM	0,5 mM	0,1 mM	10mM	0,2mM	0,2mM
KOH									0,4M		
Tris HCL									1M	100 mM	100 mM
KCL										200mM	200mM
RNaseOut										0,04U/uL	0,04U/uL
Incubation		37°C - 10min 80°C 5min					37°C 10min		4°C 10min	37°C - 10min 80°C 5min	

Table 4. Average cycle threshold (Ct) of triplicates from which the qPCR amplifies exponentially and validates the amplification efficiency. If not disturbed by the lysis buffers, the Ct should be similar to the control's Ct without any buffers.

	Without buffer	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11
NEG	30,605	27,69	23,10	29,76	NA	22,49	32,00	32,93	23,6	25,0	15,0
DNA 10	16,37	22,41	19,38	27,04	NA	22,75	15,79	15,70	23,7	19,6	18,4
DNA 10^{-1}	20,68	28,43	14,78	29,84	NA	22,62	20,00	19,81	23,8	24,0	22,9
DNA 10^{-2}	25,02	29,70	20,54	29,95	NA	22,49	24,24	24,00	24,9	28,9	27,7
DNA 10^{-3}	29,08	22,08	15,58	29,74	NA	22,17	28,37	28,35	24,1	34,2	32,7
DNA 10^{-4}	31,77	27,22	8,37	29,66	NA	22,78	31,51	31,81	24,8	35,0	35,0
DNA 10^{-5}	31,24	34,45	19,65	29,83	24,28	22,60	32,45	33,17	24,5	35,0	35,0

Therefore, we decided to test these buffers directly in the conditions of the wished protocol, that is just before the genome amplification, without purification. The REPLI-g Single Cell Kit (QIAGEN) was used for MDA validation following the cell lysis. I selected five buffers to be tested, this time on cells of *E.coli*, *B.subtilis*, and *M.luteus*. The commercialized lysis buffer in the REPLI-g Single Cell Kit (QIAGEN) was used as a positive control, Buffers 2,7, 9, and 10 were chosen to make a prior selection of the type of lysis buffer: with or without enzyme (Lysozyme), detergent (Triton), alkaline treatment (KOH) or a combination of some of these elements. Optimal working temperatures were applied for each buffer (Table 4). Sample preparation was made in volumes respecting the initial protocol of the REPLI-g Single Cell Kit (see paragraph 1.3.3). All genome amplification worked except with buffer 2 (Figure 8), where no DNA amplification has been measured except for one late sample.

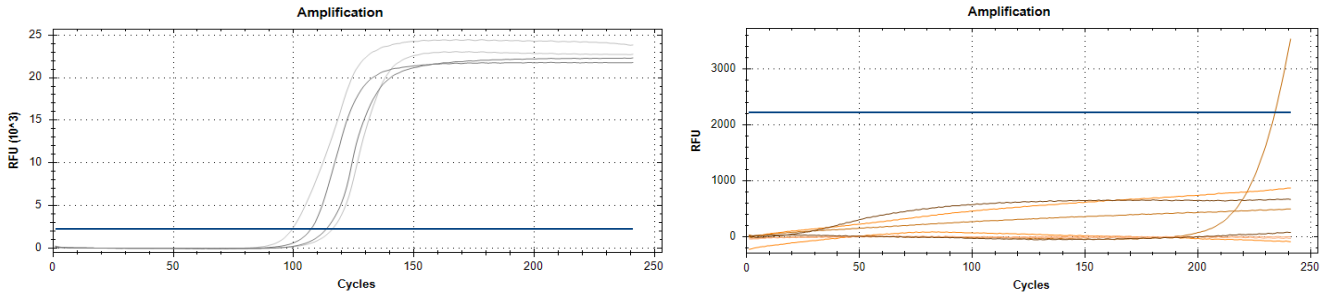


Figure 8. Real-time MDA on DNA (i.e. positive control) on the left and on cells lysed with buffer 2 on the right. The cycles represent the number of minutes, one picture of the fluorescence was taken per minute.

Additional tests revealed that the MDA polymerase was systematically inhibited by enzymes including Lysozyme, Mutanolysin, and Achromopeptidase. Negative controls (no cells nor gDNA) of the MDA were also positive (i.e. presence of DNA measured by Qubit (Invitrogen) and long fragments measured with a Fragment Analyzer or TapeStation (Agilent)). As warned by the manufacturer, the MDA can result in up to 40ng/µL yield in negative controls due to the concatemerization of random hexamers primers. Therefore, to differentiate bacterial DNA from random primer yields, the qPCR previously used for thermal lysis quality control was performed on purified MDA products. The qPCR on the positive MDAs (with buffers 7, 9, 10 and from the REPLI-g kit) demonstrated good efficiency of each lysis buffer on *E.coli* and *B.subtilis*, but only the buffer with alkaline treatment (Buffer 9) also seemed to break the cell wall of *M.luteus* (Figure 9). At this stage, buffer 9 from the Stepanauskas publication (Stepanauskas et al., 2017) was selected for further tests regarding its compatibility with the library preparation reagents. Beyond its broad efficiency demonstrated on soil samples (Stepanauskas et al., 2017) and on the strains we have tested, the choice of this lysis buffer was also adequate for the adaptation of the miniaturised protocol: it is fast and easy to prepare, the incubation is quick and can easily be done on ice or within the cellenONE platform and none of its compounds is viscous. For its usage in our protocol, we mixed the Stepanauskas buffer with saline solutions (Storage buffer), following the requirements of the MDA kits of QIAGEN. The final settings for bacterial cell lysis in bulk were: 2 µL of storage buffer (+ cells), 1,5 µL of Stepanauskas buffer (100mM DTT, 10mM

EDTA, 0,4M KOH), incubation 10 minutes at 4°C and final addition of 1,5 µL of stop solution (Tris HCL 1M, pH4).

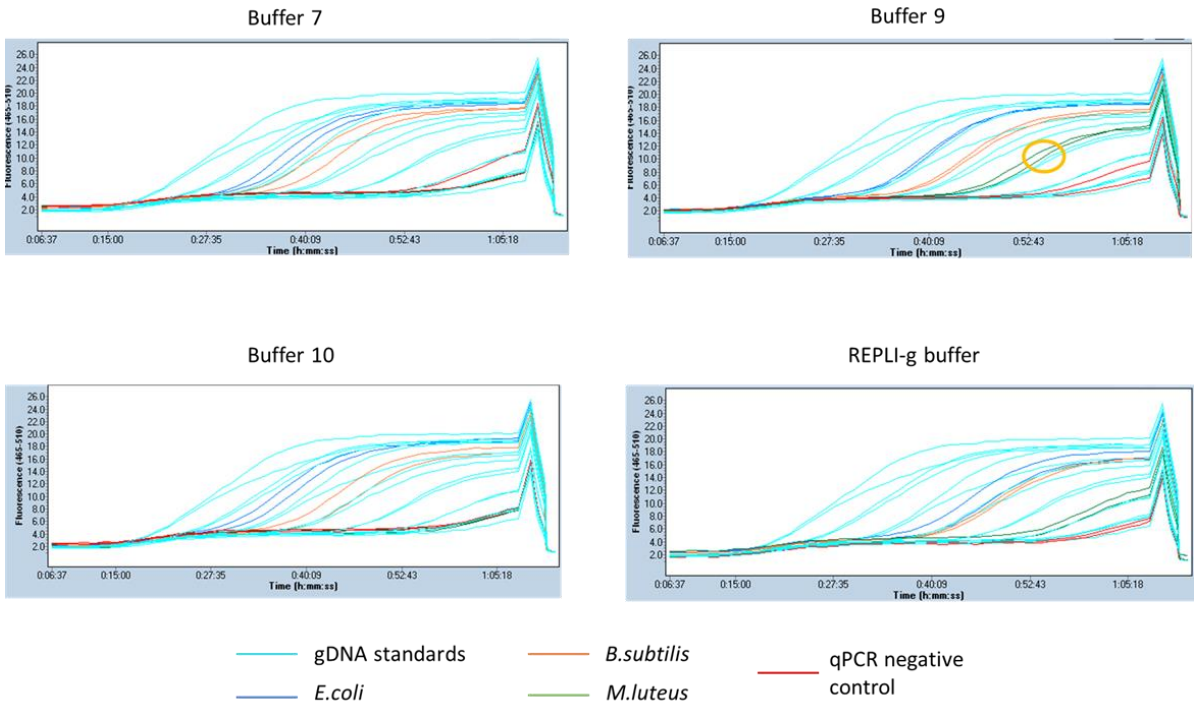


Figure 9. Validation via 16s qPCR of the MDA amplified gDNA extracted from bacterial cells after their lysis using different lysis buffers. The *M.luteus* was amplified after the cellular lysis with buffer 9 (green line circled in yellow).

1.3.3- The genome amplification

Besides, we tested two MDA kits to limit the development requirements in molecular biology. The kit from QIAGEN "REPLI-g Advanced DNA Single cell kit" is presented to be adapted for microbial application after some lysis adjustments by the manufacturer and to be more efficient than the previous REPLI-g Single Cell Kit. A kit from Bioskryb (ResolveDNA kit) proposing a PTA-MDA (Primary Template Amplification) was also tested and presents advantages regarding the control of amplification errors, as the generation of fragments occurs only via the synthesis from the initial fragment. However, the Bioskryb kit did not offer solutions for

bacterial cell lysis at the time. The genome amplification efficiency of both kits was tested following the cell lysis with either the QIAGEN or Stepanauskas lysis buffers. The Bioskryb kit was less efficient (i.e. lower amplification factor) than QIAGEN when used with the QIAGEN lysis buffer or its own.

The QIAGEN kit was therefore kept for further development. The genome amplification efficiency was quantified with fluorescence on nucleic acids, just like a qPCR, by adding 0.5 uL of Syto13 100 uM in the mix of the reactions and taking a picture every minute. For all bulk tests (i.e., MDA on gDNA or multiple cells in volumes respecting the kit recommendations), the total reaction volume was divided by two and composed as follows: 4.5 µL NFW, 14.5 µL reaction buffer, 1µL polymerase, 0.5 µL Syto 13 100 µM, for a total volume of 20 µL + 5 µL of the template (lyzed cells or gDNA). A stock of amplified samples was made using the Stepanauskas lysis buffer and the REPLI-g Advanced DNA Single cell kit, with and without final purification to serve as a template for library preparation tests (Figure 10).

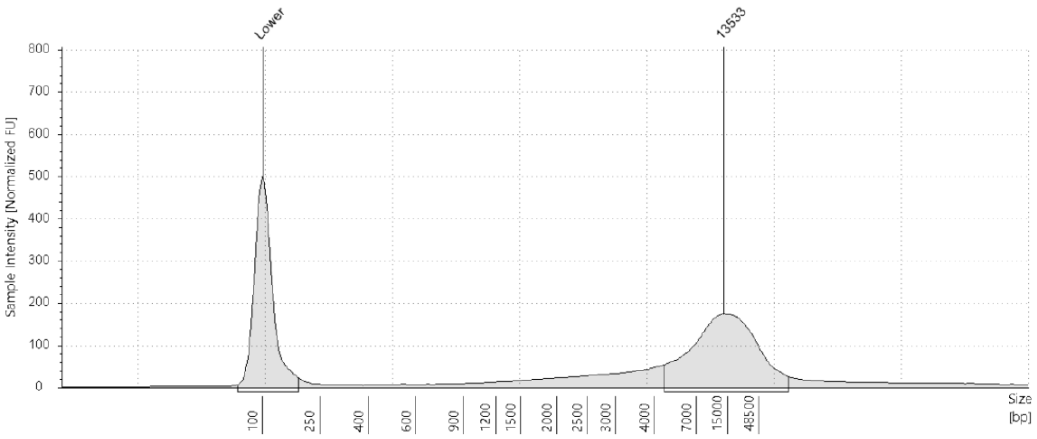


Figure 10. Fragment Analyzer profile of a MDA-amplified genome, with typical fragment size distribution between 6000 and 30000bp.

1.3.4- The fragmentation

For commercial and practical convenience, we tested two commercialized kits for library preparation: the QIAGEN® QIAseq FX DNA Library Kit and NEBNext® Ultra™ II FS DNA Library

Prep Kit from Illumina. The fragmentation step was done using a fragmentase enzyme, working at an optimal temperature of 32°C for QIAGEN and 37°C for NEB.

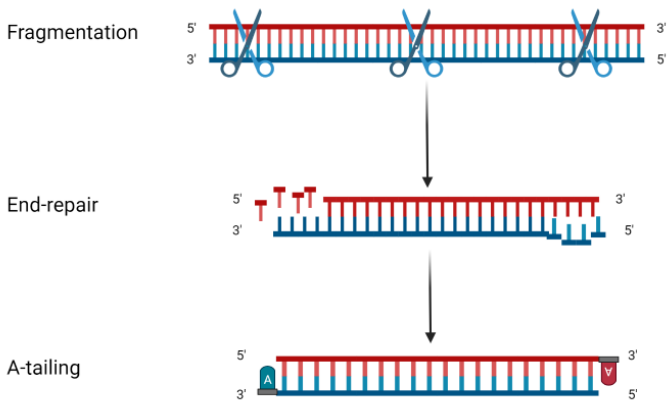


Figure 11. 3 in 1 reaction of fragmentation, end-repair and a-tailing with the QIAGEN® QIAseq FX DNA Library Kit.

The size of the fragments can be modulated by changing the concentration of DNA in input and the time of the incubation. Following the recommendations of both NEB and QIAGEN, we first evaluated the adequate parameters to obtain a satisfying fragment size. With the NEB kit and various parameters tested (i.e. DNA concentrations from 1 to 20 ng/µL and incubation time from 2 to 20 minutes), we failed to obtain a fragment size higher than 100 bp (Figure 13B) while we aimed at a minimum size of 200 bp. The results were also identical with or without prior purification of MDA-amplified DNA and w/o the Stepanauskas lysis. The same tests were done with the QIAGEN kit. While the fragmentation was not affected by the lack of prior purification of the samples and the lysis buffer, the fragment size changed with the sample concentration and incubation time (Figure 12). Final consistent settings were found after many tests and chosen for the final protocol: a final sample concentration of 2 ng/uL incubated 12 min at 32°C resulted in fragment sizes with a pic between 200 and 260 bp (Figure 13A). This reaction, besides fragmenting the DNA, also executed the end-repair and A-tailing (addition of a single adenine base in 3' position) of the fragments (Figure 11), allowing the ligation of PCR adapters with a thymine base overhang.

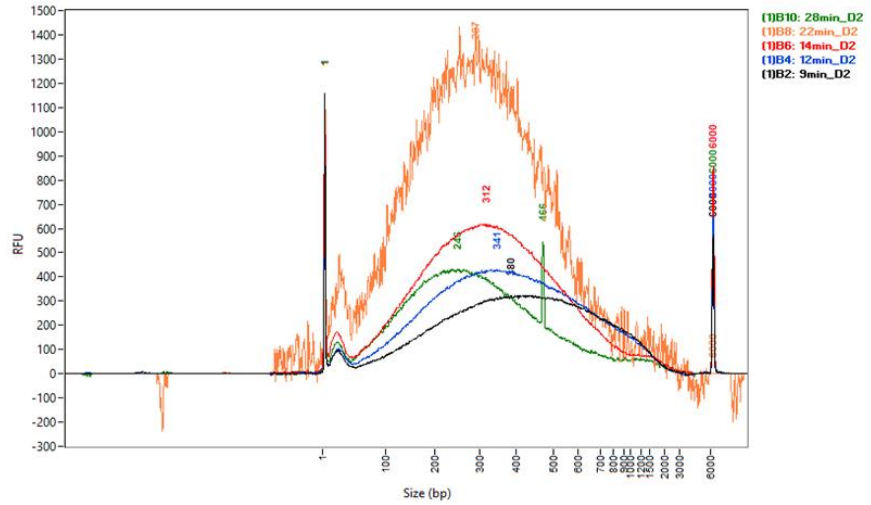


Figure 12. Shift in fragment sizes with incubation time of QIAGEN fragmentation, on a non-purified MDA sample with a final concentration of 1ng/µL.

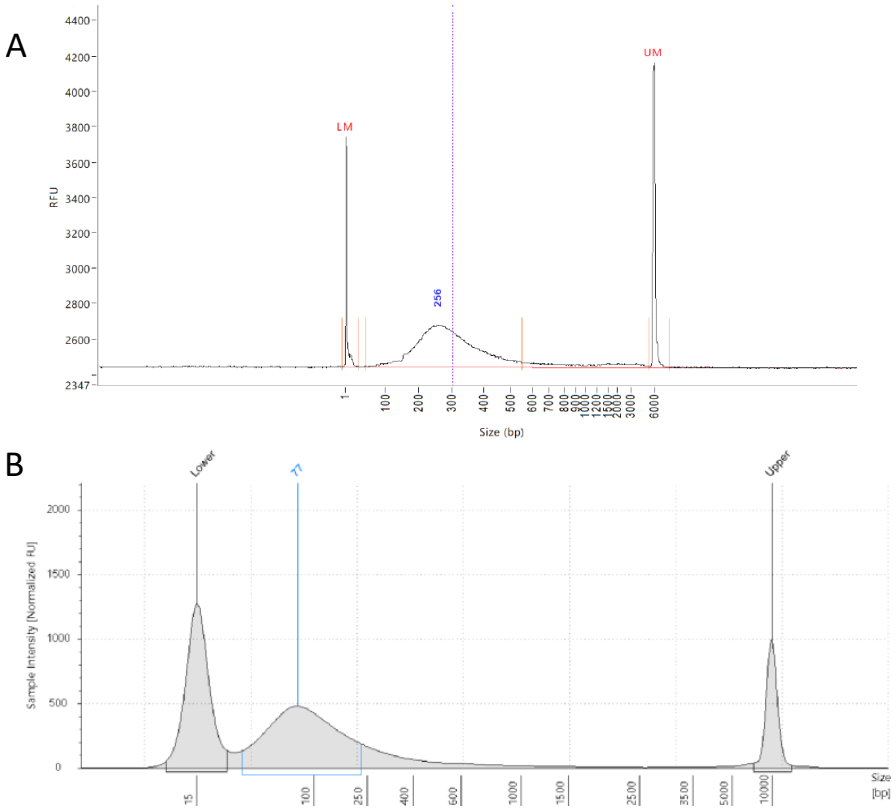


Figure 13. Fragment sizes obtained with (A) the QIAGEN® QIAseq FX DNA Library Kit with average pic at 256 bp and (B) the NEBNext® Ultra™ II FS DNA Library Prep Kit from Illumina with a pic at 77 bp.

1.3.5- The ligation of PCR primer binding site with cell barcodes and library amplification

The QIAseq FX DNA Library Kit was validated for fragmentation and therefore kept for the ligation of PCR adapters. The kit proposed up to 96 different barcodes to identify the samples. The sequences of these adapters were unknown to us, as well as the design of the PCR primers provided in the kit. To increase the number of unique barcodes and thus the multiplexing possibilities, we designed barcoded PCR adapters that should be i) able to bind to the end-repaired fragments after the fragmentation, without purification following the lysis and MDA steps, ii) containing unique barcodes, and iii) containing the binding sites for the primers of the final library amplification PCR. We worked with the primers from Nextera XT Index Kit v2 (Illumina), offering many possible combinations for indexing during the final PCR. The design of the final adapter was validated with a phosphorylation treatment in the 5' position (Figure 14 A). A combination of 12 barcodes in columns and 8 barcodes in rows allowed the unique identification of 96 different cells. Each of these cells can be gathered to receive other indexes contained on the Nextera primers (i5 and i7, Figure 14 B) identifying each pool of 96 cells (Figure 15). The success of the ligation tests in bulk was evaluated in multiple ways: i) the size of the fragments should increase by the size of the primer adapter (i.e. 80 bp). ii) a qPCR with QuantiFast SYBR® Green PCR Kit and temperature cycle and custom primers to bind to the adapters and quantify the number of fragments with adapters and iii) the final indexing PCR should be successful (i.e. increase in the size of the fragments, and in DNA concentration quantified by KAPA Library Quantification Kit).

The library preparation protocol was validated once all the compatibility tests were validated (Table 5).

Table 5. Compatibility tests verified in bulk of each step of the workflow.

Bulk validation	Lysis				
Lysis	-	Genome amplification			
Genome amplification	✓	-	Fragmentation		
Fragmentation	✓	✓	-	Ligation	
Ligation	✓	✓	✓	-	Final PCR
Final PCR	✓	✓	✓	✓	-

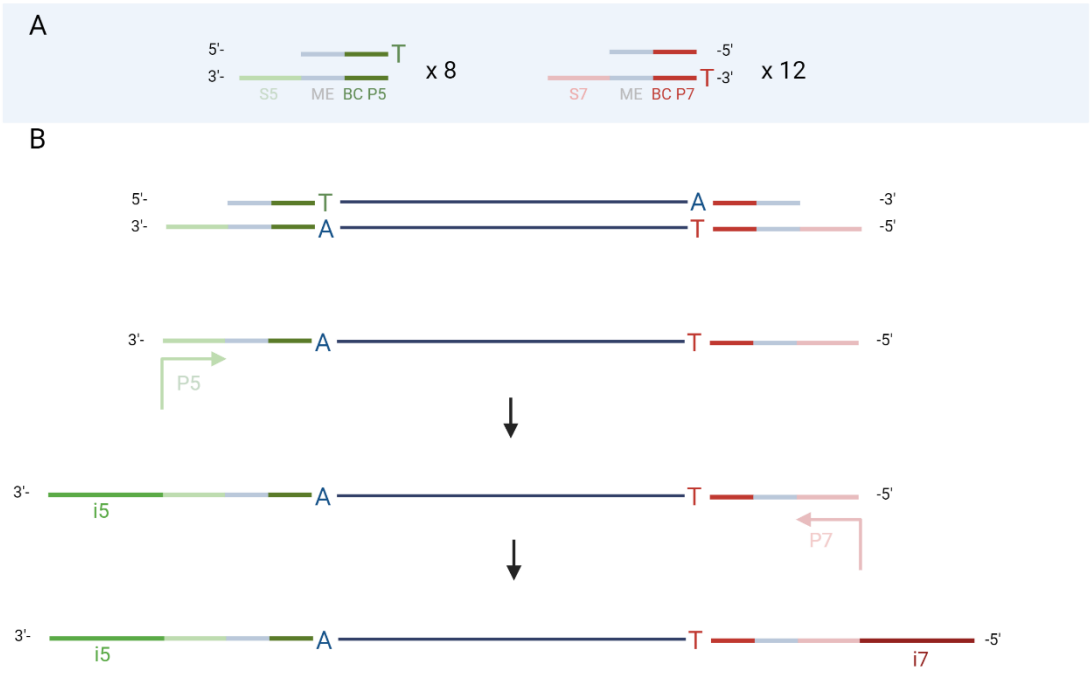


Figure 14. (A) Adapters design; each contains a barcode (BC), a mosaic end (ME), and an index PCR binding site (S5 or S7). The adapters are phosphorylated in 5' to enable their adhesion to the A-tailed fragment. (B) Library construction after adapters ligation. Each strand is amplified with the indexing PCR with primers containing the i5 and i7 indexes.

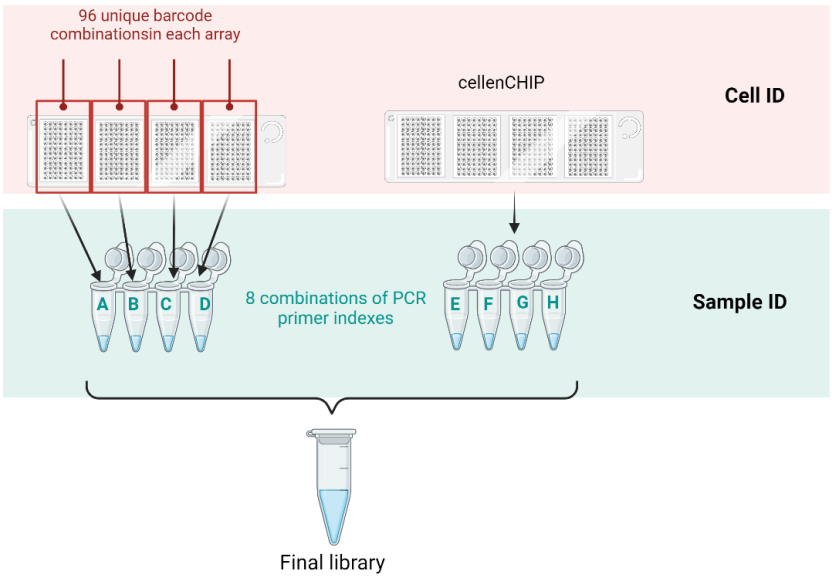


Figure 15. Sample multiplexing strategy. The cell IDs were added during the ligation step with barcoded adapters in dual indexing to tag the DNA of 96 samples. Each array of 96 samples was pooled for the indexing PCR step where specific combinations of PCR primer indexes allowed another layer of identification for each pool. When identified, the samples were pooled to compose the final library in one single tube.

The developments in nanovolumes were not satisfying and required additional work (Box 2). Further tests are needed with cellenCHIPs made of other materials (treated plastics, aluminium…). Therefore, the decision was taken to conserve the workflow which was validated in bulk, and decrease the reaction volumes to a minimum of what could be handled by hand pipetting to obtain volumes between bulk and miniaturisation tests. The protocol was executed in 384 well plates or 96 well plates depending on the step. The figure 16 summarizes the workflow setup as it was used for final experiments.

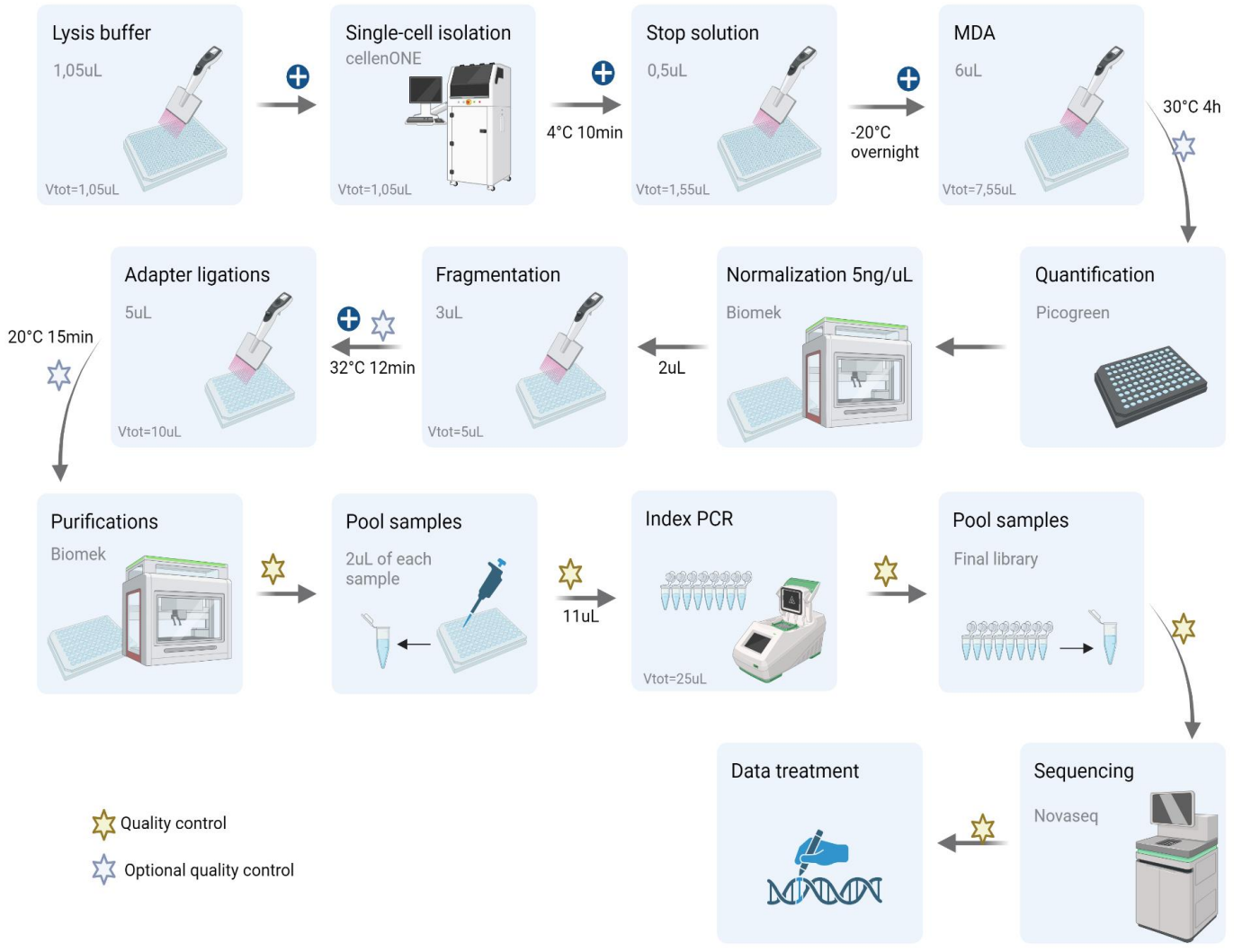


Figure 16. Final total workflow for single-bacteria whole genome sequencing

This workflow was applied to referenced strains to evaluate the genome recovery quality and to develop a bioinformatics pipeline for single-bacteria genomic data decontamination, presented in the next paragraphs of this section. I closely followed the development of this pipeline which was elaborated by two bioinformaticians of our research group. The application of this single-cell sample preparation was done on environmental samples, the results of this experiment are detailed in the next chapter of the thesis.

Box 1: Towards live/dead staining of bacterial cells using the cellenONE

For the cellenONE to isolate bacterial cells, fluorescent channels were used. Two elements constrained the choice of the fluorescent reagents: we aimed at combining labelling to discriminate live from dead cells and both labellings had to be detectable by the fluorescence channels present in the cellenONE (Figure 1.1), without overlapping. The four fluorescent channels proposed by the cellenONE gave the opportunity to test many reagents (Figure 1S, Supplementary). To summarize, all nucleic acid dyes in the green channel were perfectly adapted for bacterial isolation using the cellenONE (Syto 9, 13, and 24 for all cells and Sytox green for dead cells). However, every attempt to add a complementary dye in another channel failed: either the particles were not visible with the cellenONE ($DiBAC_4(3)$, Propidium iodide, CTC) or inconsistently labelled (Cell Tracker Deep Red, Figure 1.2). When detected, it was because the cells received a very high dye concentration. Other kits such as ViaGram Red and Sytox Red reagents were tested at Cellenion with the same outcome, possibly highlighting an incompatibility between the intensity of these red dyes emissions and the sensitivity threshold of the cellenONE red channel for those wavelengths.

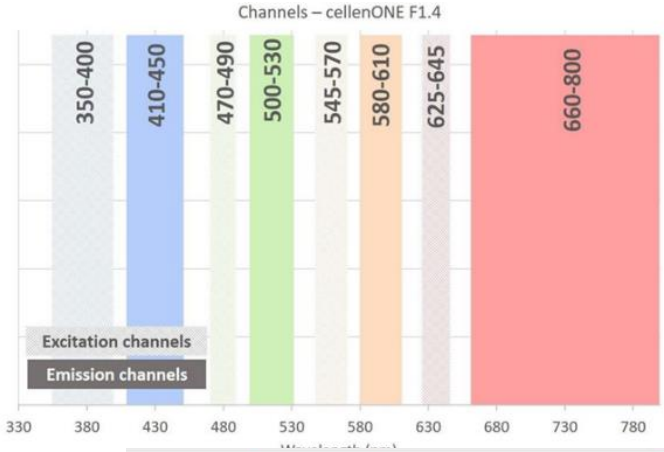


Figure 1.1. Excitation and emission channels of the 4 LEDs contained in the cellenONE.

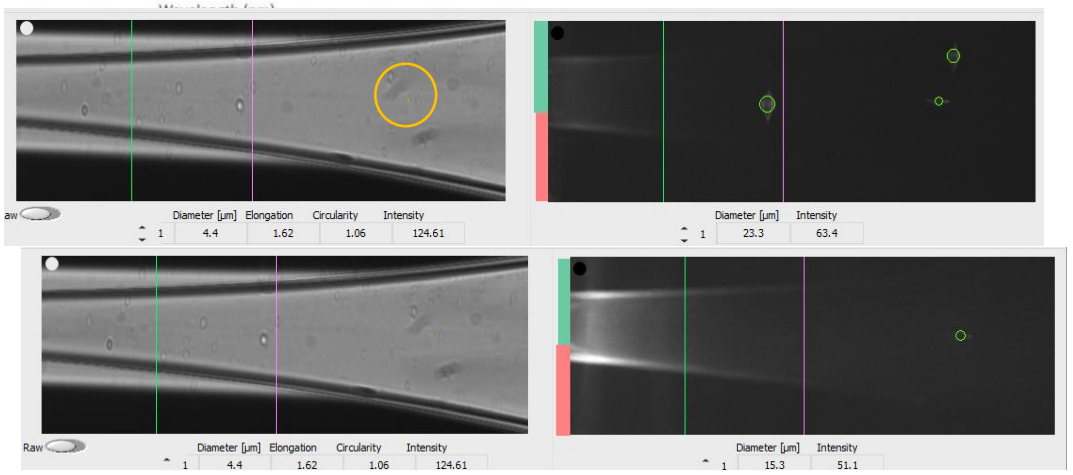


Figure 1.2. View of the PDC containing a bacterial suspension, stained with Sytox green (top image) and Cell Tracker Deep Red (bottom image). Except for one bacterium circled in yellow, no bacteria were

detected without fluorescence. While three bacteria are detected in the green channel, and supposedly dead, only one is detected in red which should contain all bacteria, dead or alive.

When testing double staining, we attempted to validate the status of the cells (alive or dead) of the fluorescence by individually growing them on pre-filled wells with a solid growth medium but did not observe better viability (with cultivability as a proxy) from cells stained as “live” or “active” than cells stained as “dead” or “inactive”. This indicated (i) toxicity of fluorescent dyes at the concentrations required by cellenONE detection, and (ii) inefficiency of the tested live/dead staining or incompatibility with the cellenONE. For my thesis work, I decided to keep only the staining with Syto 9 or 13 dyes which were the only unbiased labelling tested. The application of live/dead staining for bacteria requires more sensitive fluorescence detection by the cellenONE. Recent research at Cellenion led to the development of a new software dedicated to bacteria isolation, that enables >92% single-cell accuracy even in bright field, removing the need for a fluorescent staining step if not desired.

Box 2: Towards miniaturisation

Three steps of the protocol needed to be performed in the cellenCHIP prior to adapters ligation and cell barcoding: the lysis, the gDNA amplification, and the fragmentation, (Figure 2.1).

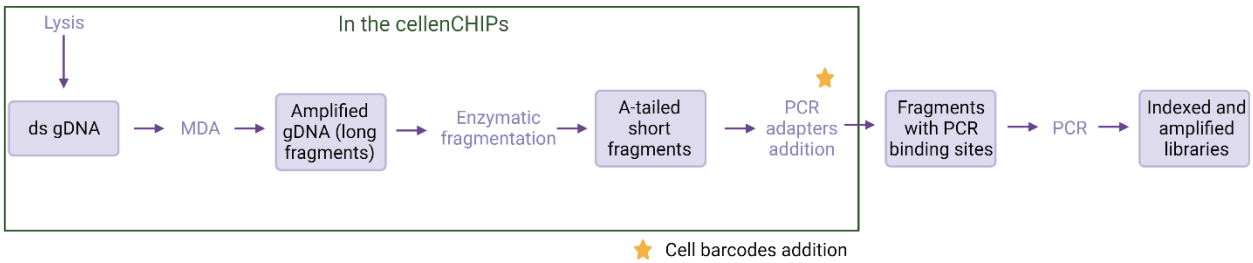


Figure 2.1. Library preparation protocol from isolated bacteria. The steps from the lysis to the PCR adapter ligation containing the cell barcodes are performed in the cellenCHIP.

For the miniaturisation of the sample preparation, the concentrations of the enzymes and buffers previously validated in bulk were conserved as a hypothetic way to avoid any possible dysfunction of molecular reactions. Moreover, some solutions were viscous, especially the ligation buffer which therefore needed to be diluted to allow the distribution with the PDC while keeping the right final concentrations. After calculations and dispensing tests, the proportions of each mix were used as follows: 21 nL of lysis buffer + 9 nL of stop solution, 120 nL of MDA mix, 90 nL of fragmentation mix, 5 nL of adapters, and 355 nL of ligation mix for a total of 600 nL. For the PDC to dispense properly, the mixes and buffers had to be free of air bubbles by spinning at 6000 rpm for approximately 20 minutes for gas removal without enzyme damage in a Labnet C1301-B centrifuge. The complete workflow demanded many manipulations and required alternations of different workstations between the cellenONE, thermal cyclers, and centrifuges. Because of the differences between the temperature set on the thermal cycler and the temperature within the cellenCHIP, each temperature was modified to reach the adequate temperature within the wells. This measure was done with thermal probes placed in the wells on the cellenCHIP during thermal cycles, with a closed lid.

Table 2.1. combinations of in-chip or in-tube steps of the workflow and the corresponding validation (green) or failure (red) of the workflow via indexing PCR.

MDA	Fragmentation	Ligation	PCR validation
Tube	cellenCHIP	cellenCHIP	
Tube	Tube	cellenCHIP	
cellenCHIP	cellenCHIP	Tube	
cellenCHIP	Tube	Tube	
Tube	Tube	Tube	
cellenCHIP	cellenCHIP	cellenCHIP	

In the same way as bulk validations, each step was individually validated in the chip before processing the total workflow, with the same quality controls. All steps resulted in similar results to bulk ones, despite the high variability of the MDA success in these low volumes (i.e. from 1 to 300 ng/µL of yield with identical settings). The steps were then combined one by one to control their compatibility (Table 5). The total workflow success was approved when the final indexing PCR was validated. The table 2.1 summarizes the results of these tests; the PCR was not working when the fragmentation was done in the cellenCHIP indicating that the library preparation was not successful. Despite the validation of fragment size, we suspect the end-repair or A-tailing step, supposedly done simultaneously with the fragmentation, to have failed in nanovolumes, which could not be spotted with the Fragment Analyzer for quality control.

Supplementary

Table S1. Summary of the staining reagents tested for bacterial cells detection with the cellenONE

Reagent(s)	Kit (Supplier)	Target	Exc/em	Rcommended consentration	Used concentration	Incubation	Strain	Dead treatment?	Observation	Conclusion	Ref
DiBAC4(3)	(Thermofisher Scientific)	Enters depolarized cells to bind to proteins	493/516 nm	1uM	1uM	37°C 20min	*E.coli*	90°C 5min or RT for a few days	Very few cells were observed.	No further investigations	Rezaeinejad, Ivanov 2020
Syto 9 and 13	(Thermofisher Scientific)	Bind to nucleic acids of all cells	485/498 nm (Syto 9) 488/509nm (Syto 13)	50nM-20uM	7,5uM - 10uM	RT 15-30min	*E.coli, P.fluorescens, M.luteus, B.subtilis, S.epidermidis*	No	Very bright, must insist on washes after the staining to limit fluoresens of the glass capillary tip	Validated, would need to be combined with a dead/dormant cell marker	Langsrud & Sundheim, 1996
Syto 9 and Propidium iodide (PI)	LIVE/DEAD™ BacLight™ Bacterial Viability Kit (Thermofisher Scientific)	Nucleic acids, all cells (Syto9) and cells with damaged membranes only (Propidium)	485/498 nm (syto9) 490/635 nm (PI)	30uM (PI) - 5uM(Syto 9)	30uM - 0.8mM (PI)	RT 15min	*E.coli, M.luteus*	Isopropyl 70%	Very bright staining in green (syto 9), nothing in red (PI) except in very high concentrations but only for a moment: it is very likely we reached toxic concentrations for the cells	Still no issue with the green staining for all cells, we need to find an alternative for live/dead staining	Defives et al. 1999
CTC (5-Cyano-2, 3-ditolyl tetrazolium chloride), Syto 24 or DAPI	BacLight™ RedoxSensor™ CTC Vitality Kit (Thermofisher Scientific)	Redox activity of live cells(CTC), nucleic acids of all cells (Syto or DAPI)	450/630 nm (CTC) 490/515 nm (Syto 24)	5mM CTC, 1uM Syto 24	5mM CTC, 10uM Syto 24	37°C 30min with CTC then RT 15min with Syto 24	*E.coli, M.luteus*	Isopropyl 70%	CTC fluorescence was not observed with the cellenONE, observed under a microscope.	The wave length for CTC is not adapted to the cellenONE, no further investigations with this e kit	Schaule et al. 1993
Sytox green	(Thermofisher Scientific)	Nucleic acids of cells with compromised membranes (=dead cells)	504/523 nm	0.5-5uM	1-10uM	RT 15-30min	*E.coli, M.luteus, Acidovorax,Paenar throbacter*	No	Good detection with high concentrations	Good for usage, need to be combine with a dye targeting something else than nucleic acid, and not in green	Lebaron et al. 1997
Cell Tracker deep red	(Thermofisher Scientific)	Amine groups present on proteins, all cells	630/650 nm	250nM-1uM	250nM-20uM	RT 15-30min	*E.coli, M.luteus, Acidovorax,Paenar throbacter*	No	Detection with high concentrations, but some cells were visible without fluorescence and not stained	Cell tracker is not fully reliable for our application	Charubin et al. 2020

II- Wet-lab preparation and automatized decontaminating procedure for single-bacteria genomics

Article in preparation for submission to Nature Methods journal

Authors: S. Mauger, Y. Sevellec, L. Carret, N. Robert, C. Monard, C. Thion, J. Bagnoli, F. Monjaret, P. Vandenkoornhuyse.

2.1- Introduction

Single-cell omics applications in microbiology started almost two decades ago (Raghunathan et al. 2005) and broadened access to bacterial interactions, population diversification, and evolutionary dynamics understanding within natural bacterial communities (Bawn et al. 2022; Davis and Isberg 2016; Kashtan et al. 2014; Labonté et al. 2015). The cell-level observations refine the microbiology theories and complement the interpretation of the observations from metagenomics approaches. Each step of single-cell omics sample preparation can be customized: cell extraction, lysis, genome amplification, library preparation, and bioinformatic treatments. The choice of the approach depends on the matrix (e.g. soil, water, tissue, pure cultures), the study's goal (whole genome sequencing or specific targeted genes) but also, and mainly, the cost. While keeping an important sequencing depth seems primordial for single-nucleotide variants (SNV) detection between genomes and should not be neglected for single-cell omics (Van Rossum et al. 2020), the decrease in cost has been concentrated on limiting the reaction volumes, which limits the contamination risks simultaneously (Blainey 2013; Mauger et al. 2022). The technologies for single-cell omics applications evolve yearly and have branched into diverse tools for cells and liquid handling (Chen, Chen, and Zhang 2017; Woyke, Doud, and Schulz 2017). Technical requirements for handling single cells and nano-volumes demand very complex and costly installations and in the case of microfluidics do not offer much flexibility for custom sample preparation (Woyke, Doud, and Schulz 2017). Flow cytometry such as fluorescence-activated cell sorting (FACS) does provide higher throughput than other cell isolation tools, but its inability to visually isolate the cells, to isolate without prior staining, to cope with diverse cell morphologies, and

to give the opportunity to work in nano-volumes leaves many padlocks for single-microbes sample preparation. Moreover, the chosen procedure for single-cell sample preparation will inevitably influence the resulting sequencing data handling.

Regardless of the chosen isolation technique, single-cell omics are still subject to contamination, they generally present very partial genome reconstruction (López-Escardó et al. 2017; Zheng et al. 2022), and low sample throughput compared to metagenomics. The sources of contamination are diverse; from biological, technical, or bioinformatic sources. The biggest argument for single-cell omics application compared to meta-omics approaches is to avoid chimera reconstruction and distinguish between similar populations of bacterial strains. The distinction of contaminant DNA from the target is, therefore, a priority to obtain the purest genomes possible. Just like sample preparation, no universal procedure is applied to single-cell omics data treatment (Alneberg et al. 2018; Anstett et al. 2023; Berube et al. 2018a; Bowers et al. 2017; Chijiiwa et al. 2020; López-Escardó et al. 2017; Nishikawa et al. 2022; Pachiadaki et al. 2019). This general lack of standardization, decontamination, and benchmarking makes it difficult to compare single-cell omics data and evaluate their robustness. The exploration of the quality of Single Amplified Genomes (SAGs) is superficial, and only a few papers are attempting to decontaminate SAGs in depth (Anstett et al. 2023; Bowers, Doud, and Woyke 2017; López-Escardó et al. 2017). The most common decontamination procedure is based on the taxonomic assignation of contigs, selected and removed from the final assembly with different tools (Anvio's (Eren et al. 2015), ProDeGe (Tennessen et al. 2015), acdc (Lux et al. 2016), MDMcleaner (Vollmers et al. 2022)). Few tools exist for contigs contamination evaluation (Cornet and Baurain 2022), but no pipeline proposes to remove it prior to reads assembly to avoid misassemblies, mostly because each dataset presents its specific challenges and contaminants related to the sample and its preparation.

Here we developed a full microbial single-cell genomic pipeline including both the single-cell genomic library preparation workflow and an innovative automated sequence decontamination pipeline. The cellenONE instrument used optically detects bacterial cells without staining needed and produces drops in the picoliter range. Its versatility allows the personalization of isolation parameters to fit the sample requirements, and objectives of the experiment, which makes it highly reliable and accurate. We modified and optimized existing

library preparation procedures from published or commercialized approaches to lower the reaction volumes and applied this protocol to referenced strains of bacteria to evaluate the quality of the genomes recovered. We developed an automatized pipeline called SINCERE DATA (SINgle-CEll REads Decontamination through Automatic Taxonomic Assignation) for single-cell genomics data treatment able to identify and delete reads classified as contaminated based both on contig coverage and their associated taxonomic affiliation. The pipeline can distinguish between coverage differences of the target DNA and contaminant DNA, present in small quantities in many samples with abnormal coverage profiles. We also applied this pipeline as an example to previously published datasets(Berube et al. 2018b) highlighting the necessity of SAGs decontamination prior to the assembly stage. We point out the precautions in sample handling for single bacteria genome sequence data production and emphasize the necessity of chasing *in silico* contaminants for data robustness and accuracy.

2.2- Materials & methods

2.2.1- Cell preparation

Two strains of bacteria (*Pseudomonas fluorescens* ATCC® 13525 and *Staphylococcus epidermidis* ATCC® 12228) were grown in nutrient broth (Merck) at 28°C for *P. fluorescens* and 37°C for *S. epidermidis*. The genomic DNA of *Bacillus subtilis* and *Micrococcus luteus* were extracted with GenElute™ Bacterial Genomic DNA Kit (Merck) and used as positive controls. For single-cell isolation, the equivalent of 10^7 cells of exponential phase bacterial cultures were pelleted for 3 minutes at 9000 rcf to remove the supernatant and resuspended in sterile PBS in equal volumes. This step was repeated once. Cell suspensions were diluted to reach an approximative final concentration of 10^5 cells per mL, diluting the cell suspension in degassed PBS beforehand by 15-min vacuum sonication in sciPURATOR (Scienion).

2.2.2- Single-bacteria isolation and lysis

The protocol was built with the inclusion of an innovative cell isolation tool, the cellenONE, allowing the isolation of bacteria based on particle automatic optical detections. Its glass Piezo Dispense Capillary (PDC) can handle liquids in small volumes (i.e. 250-800 pL per drop, with <0.25% variance) which limits the external DNA contamination risks and gives the possibility of working in nanovolumes. The cellenONE X1 BSC model is placed in a Class II

Biosafety Cabinet (BSC) to protect experimental staff from biohazards and prevent external contamination of samples. The temperature of the platform holding the samples was controlled and set at 4°C for better reagent stability. The PDC used for cell suspension handling and cell isolation was systematically sterilized between the manipulation of different strains, by successive aspiration and immersions in chlorine 0.5%, hydrogen peroxide 3% and ethanol 70% for 2 minutes each and ultrasonication (cellenONE Sterilization task). The detection and isolation parameters were based on size and elongation measured on brightfield images, thanks to the microLIFE package of the cellenONE software, dedicated to the isolation of small particles. Quality control analysis of isolation runs was performed using microLIFE viewer and single-cell accuracy (i.e., proportion of the positions where one and only one cell was dispensed) was 89 %. Positions with isolation errors (doublets or empty droplets due to false detection) were recorded for further confrontation with sequencing results.

Single cells were isolated in wells of a 384-well plate where the lysis buffer was previously distributed by hand, under a PCR flow hood. Out of 274 used wells, 123 received no cells to serve as a negative control of the sample preparation, 22 contained genomic DNA for positive controls, and 129 received single cells. The lysis buffer was chosen based on the lack of subsequent molecular reaction inhibition and its efficiency in breaking the membranes of various bacteria. The core of the buffer was taken from Stepanauskas et al. (2017) who applied an alkaline treatment to soil bacteria to break their cell membrane. The lysis buffer was composed as follows: DTT 100 mM, EDTA 10 mM, KOH 0.4 M and diluted in water prior to the distribution of 1.05 µL (0.45 µL of buffer, 0.6 µL of nuclease-free water (NFW)) in each well. Genomic DNA from *B. subtilis* and *M. luteus* was also distributed by hand as a positive control, for a final 100 pg input of gDNA in the reaction (0.2 µL in each well at 0.5 ng/µL). After the distribution, the plate was briefly centrifuged to ensure cell deposition in lysis buffer and placed for 10 min at 4°C. To stop the lysis and re-adjust the pH of this alkaline buffer, 0.5 µL of 1 M Tris HCL pH4 was added to each well, by hand. After being centrifuged, the plates were placed at -20°C overnight.

2.2.3- gDNA amplification

The next day, the plate was thawed on ice, centrifuged, and manipulated under the BSC. For the genome amplification, either the REPLI-g Advanced DNA Single-cell kit (QIAGEN) or a mix of NEB Phi 29 enzyme and reaction buffer was used. We used reagents from these two

suppliers to evaluate the potential contamination sources and viability of these approaches. For both, 6 µL of mix were added in each well consisting of 4.35 µL of reaction buffer, 1.35 µL of NFW and 0.3 µL of enzyme for the QIAGEN Kit and 0.75 µL of reaction buffer, 2.53 µL of NFW, 0.75 µL dNTP at 10 mM, 1.89 µL of random primers at 200 µM and 0.08 µL of enzyme for the NEB reagents. The MDA was performed for 4 hours (QIAGEN products) or 8 hours (NEB products) at 30°C with a final step at 65°C for 3 min in a Bio-Rad C1000 Touch Thermal Cycler. Each amplification was quantified using Picogreen (Invitrogen) reading green fluorescence with qPCR cycler (BioRad CFX96), then amplified DNA was normalized at 5 ng/µL in NFW and stored at -20°C until further use.

2.2.4- Enzymatic fragmentation

The libraries were prepared using the QIAseq FX DNA Library kit from QIAGEN or NEBNext Ultra II FS DNA Library Prep Kit from NEB. 2 µL of normalized MDA product were placed in wells of a new 384 well plate, and 3 µL of fragmentation mix were added, consisting of 1.5 µL of water, 0.5 µL of fragmentation buffer, and 1 µL of enzyme for the QIAGEN kit and 1 µL of nuclease-free water, 1 µL of fragmentation buffer and 1 µL of the enzyme (Diluted 11.6 times) for NEB reactions. The tubes and plates were kept on ice during manipulation to hold the enzyme activity. The plate was placed in a thermocycler previously set at 4°C and the fragmentation cycle started for 12 min at 32°C followed by 30 min at 65°C.

2.2.5- Adapter ligation

For this study, we designed 20 adapters (12 in columns and 8 in rows) with specific barcodes to tag and differentiate 96 cells with combinatorial barcoding. The pairs of barcodes were attributed 4 times to cover the 384 wells. Each adapter was added by hand in each well of new 96 well plates, 0.5 µL for QIAGEN and 0.2 µL for NEB wells, at 10 µM. The ligation mix (4.5 µL for QIAGEN and 5.1 µL for NEB) was added on top of the 5 µL of fragmented DNA. For QIAGEN, the mix composition was 2 µL of buffer, 1 µL of enzyme, and 1.5 µL of NFW. For NEB, the mix contained 4.6 µL of ligation mastermix, 0.2 µL of enhancer, and 0.3 µL of NFW. The plate was incubated at 20°C for 15 minutes. The ligations of each set of 96 pairs of barcodes were pooled in one single tube to be immediately purified with magnetic SPRI beads (Bagnoli et al. 2018) at 0.8x first and 1x in a second time. The samples were kept at 4°C until further use.

2.2.6- Indexing PCR and sequencing

The final PCR was made with the NEBNext® Ultra™ II Q5® Master Mix and standard Illumina index primers with double indexing (IDT-DNA). Per sample, 1.25 µL of each primer was used, 5 µL of the sample, 5 µL of NFW, and 12.5 µL of Master mix. The libraries were sequenced on a Nextseq 2000 sequencer with a P2 600 cycles flow cell (Illumina), offering an average Q30 of 88.18% and 374.47M total reads.

2.2.7- Sequence data treatment

The demultiplexing of the dual indexes was performed with cutadapt (V. 4.1) with an error rate of 15% (Martin 2011). The adapter and quality trimming were also performed using cutadapt (V. 4.1) with a minimal quality set to 15 and a minimal length set to 30 bp. The quality of the trimmed reads was evaluated with fastqc (V. 0.11.5, Andrews (2010)). From the total dataset, 30 samples were selected based on their a priori higher quality of assembly prior to decontamination measured by QUAST (V.502, Gurevich et al., 2013) and CheckM (V.1.2.0, Parks et al., 2015). The first step of contaminants detection was performed via taxonomic affiliation of reads with Kraken2 (V. 2.1.2, Wood et al., 2019) on the PlusPFP precompiled database (downloaded the 31/01/23). Sequence-reads decontamination was manually performed based on Kraken taxonomy, and the reads affiliated to Eukaryotes, Enterobacteriaceae, Salasvirus, and Bacilliceae were removed. The Enterobacteriaceae are suspected to be carried by the MDA reagents as well as the *Salasvirus* from which the Phi29 polymerase is extracted. We also suspected a cross-contamination with *Bacillus* genomic DNA in some wells that were used as a positive control, the reads and contigs of Bacilliaceae were therefore removed for all the samples. These decontamination outputs were used as a comparison for the pipeline decontamination efficiency explained below.

At the same time, an automated pipeline, SINCERE DATA (Figure 2), was developed to identify and remove contaminants based both on contigs coverage and the taxonomic affiliation of associated reads. Trimmed reads were first assembled with spades (V. 3.15.5, Bankevich et al., 2012) using the careful and single-cell option (--careful –sc). The contigs shorter than 500bp were then removed using reformat.sh from the bbtools suite (V. 39.0.0, Bushnell et al., 2017). From each assembly, the most abundant taxa were determined with Kraken2. If a given taxon was identified as the main taxon in more than 70% of the sample, it was reported by the pipeline as a potential contamination and advised to be removed from the reads prior

to the automatic decontamination. In our dataset, we detected contamination from the Bacillaceae positive control wells in some neighbour wells of other samples. These taxa reads were also removed. Potential contaminant regions in the assemblages could be detected from abnormally high coverages. The detection of these outlier regions from the mean genome coverage was made using a succession of tools: i) alignments of the reads on the assembly with bowtie2 (V. 2.5.1, Langmead & Salzberg, 2012), ii) sorting and indexing of the alignments using SAMtools (V. 1.1.7, Danecek et al., 2021) iii) evaluating the coverage with bedtools (V. 2.27.1, Quinlan & Hall, 2010). The z-score is a statistical measurement that calculates how much the standard deviation of a given value diverges from the mean of a group of values. The z-score was calculated for each nucleotide position and the regions that exceeded a z-score of 2 for more than 30bp were selected as outliers. The outliers distant from less than 100bp from each other were merged to reduce the computational burden. The outliers were extracted using SAMtools and received a taxonomic assignment. Reads from outlier regions with main taxonomic affiliation identical to the main taxonomy of this sample assembly were conserved. However, if the taxon found was different, the outlier reads affiliation was identified as a potential contaminant. The potential contaminants of each sample were collected and compared to the total collection. If a given potential contaminant was present in more than 20% of the samples and was not the dominant taxon in more than 20% of the samples then its contaminant status was validated, resulting in the suppression of all reads assigned to this taxon as well as its "children" taxa using krakentools (https://github.com/jenniferlu717/KrakenTools).

Final assemblies were performed for each decontamination strategy without the contaminated reads with spades (V. 3.15.5, Bankevich et al., 2012) using the careful and single-cell option (--careful –sc) and contigs smaller than 500 bp were also removed. To suppress the potential remaining contaminants, contigs from presumed contaminants (Eukaryota, Enterobacteriaceae, Bacilliaceae, and Salsavirus) were removed for the manual decontamination. The contaminants detected by the pipeline for the automatic decontamination had their contigs removed as well. The quality of the final assembly was estimated with QUAST (V.502, Gurevich et al., 2013) and CheckM (V.1.2.0, Parks et al., 2015), and quality reports were aggregated using multiqc (V.1.14, Ewels et al., 2016). Finally, Kraken2 was used to determine the main taxonomic assignment of the final SAGs, and .biom files were

generated using Kraken biom (V. 1.2.0, Dabdoub, 2016) for visualization of the taxonomy and data treatment. The Pavian application (Breitwieser and Salzberg 2020) was used for Kraken reports visualisation and Sankey diagram execution.

The SINCERE DATA pipeline was applied to a published dataset (Berube et al 2018), aiming at referencing uncultivated genomes of marine microbes to increase the content of the databases for these organisms. Briefly, the cell isolation was performed with FACS technology and MDA was performed in 384 wells as previously described as well as sequencing data handling (Stepanauskas et al. 2017). From the trimmed reads, assemblies were done identically to our samples. We then randomly chose samples from the published dataset within each quality category following the Genomic Standards Consortium (GSC) (Bowers et al., 2017): 8 SAGs of low quality, 17 of medium quality and 5 of high quality. We measured putative completeness and contamination rates with CheckM and used Anvi'o (Eren et al. 2015) to detect ribosomal gene presence necessary for the GSC (Bowers et al., 2017). The automated decontamination pipeline was then applied to these samples, using the trimmed reads as entry data.

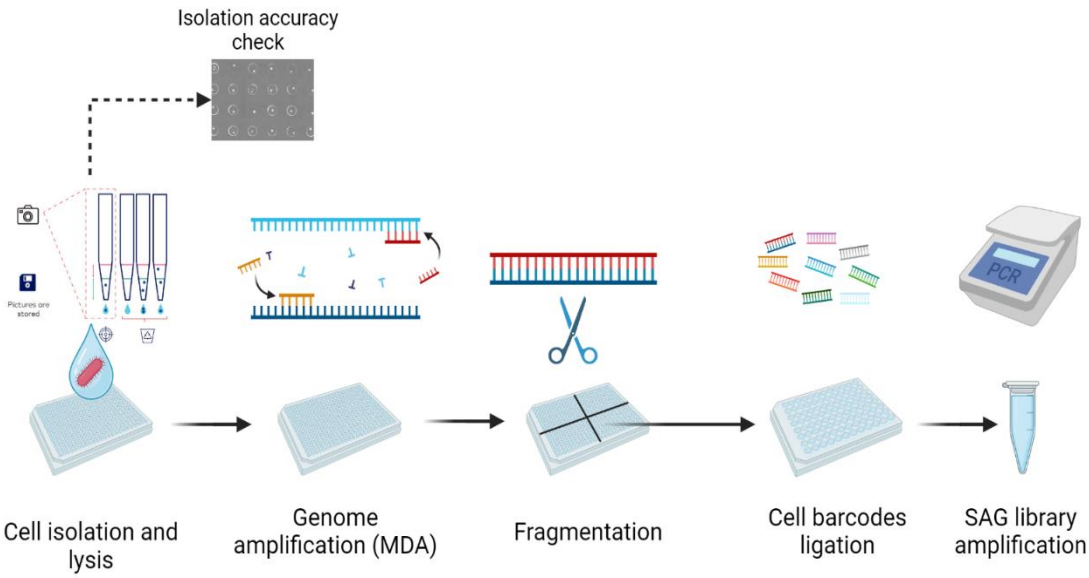


Figure 1. Main steps for sample preparation for single-cell library sequencing. The isolation of the cleaned cells was made in a 384 well-plate with the cellenONE technology (Cellenion), generating droplets of 250-600pL. The cell lysis was performed prior to the genome amplification with miniaturized Multiple Displacement Amplification (MDA). Fragments were shortened with the fragmentation step and immediately barcoded per group of 96 samples to uniquely identify each cell with designed adapters during the ligation. These 96 samples were then pooled to be PCR-amplified as one sample and received unique indexes for library multiplexing.

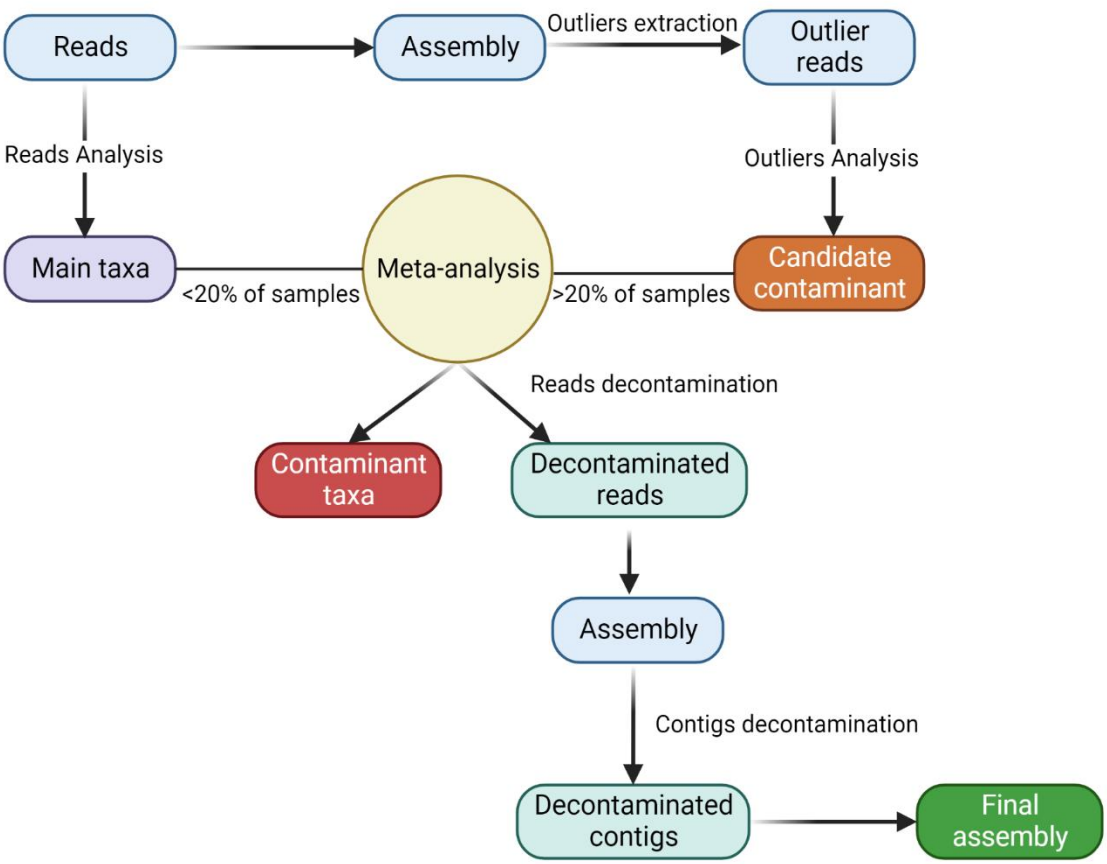


Figure 2. Functioning of the automatic decontamination pipeline SINCERE DATA. From trimmed reads, the main taxa of each sample were determined in parallel to being assembled. Reads that were affiliated to outlier regions (i.e. abnormally high coverage) on the assembly were selected for each sample and affiliated to a main taxonomy. If this taxonomy was different from the taxonomy of the total sample, this outlier was considered a "Candidate-contaminant". The candidate-contaminants of each sample were compared to the total collection of the candidate-contaminants: if they were present in more than 20% of the samples and their taxonomic affiliation corresponded to less than 20% of the main affiliation of all the samples, they were considered as definitive contaminants. The reads affiliated to these contaminants were removed from the dataset, the assembly was performed and the contigs with the same taxonomy as the previously detected contaminants were also removed.

2.3- Results

2.3.1- Decontaminations of referenced strains genomes

The comparison of sequence quality between the NEB and QIAGEN library preparation on positive and negative controls did not show any notable differences (Supplementary Figure S1) and can be used as possible alternatives.

Out of the 30 SAGs selected, 23 were assigned to *Staphylococcus epidermidis* and 7 to *Pseudomonas fluorescens*. Based on the Genomic Standards Consortium (GSC) (Bowers et al., 2017), all our SAGs were considered as low quality before and after the decontamination process (i.e. Genome completeness below 50%) (Figure 3C). By removing contaminated contigs, the assembly length inevitably dropped from a median of 172.62 Kb for the assembly with trimmed reads to 54.05 Kb with manual decontamination and 38.6 Kb with automatic decontamination (Figure 3B). The same dynamic was observed for the number of assigned contigs (Figure 3E). The measured N50 slightly increased with the two decontamination procedures (Figure 3A). The completeness of the assemblies was similar for the trimmed and automatically decontaminated samples (i.e. ~2%), and on average at 0 % for the manually decontaminated samples (Figure 3C). The contamination measured for the three conditions was however similar, close to 0 % (Figure 3D). Overall, we observed similar measured qualities between the manual and automatic decontamination except for the completeness which was at its lowest with the manual decontamination.

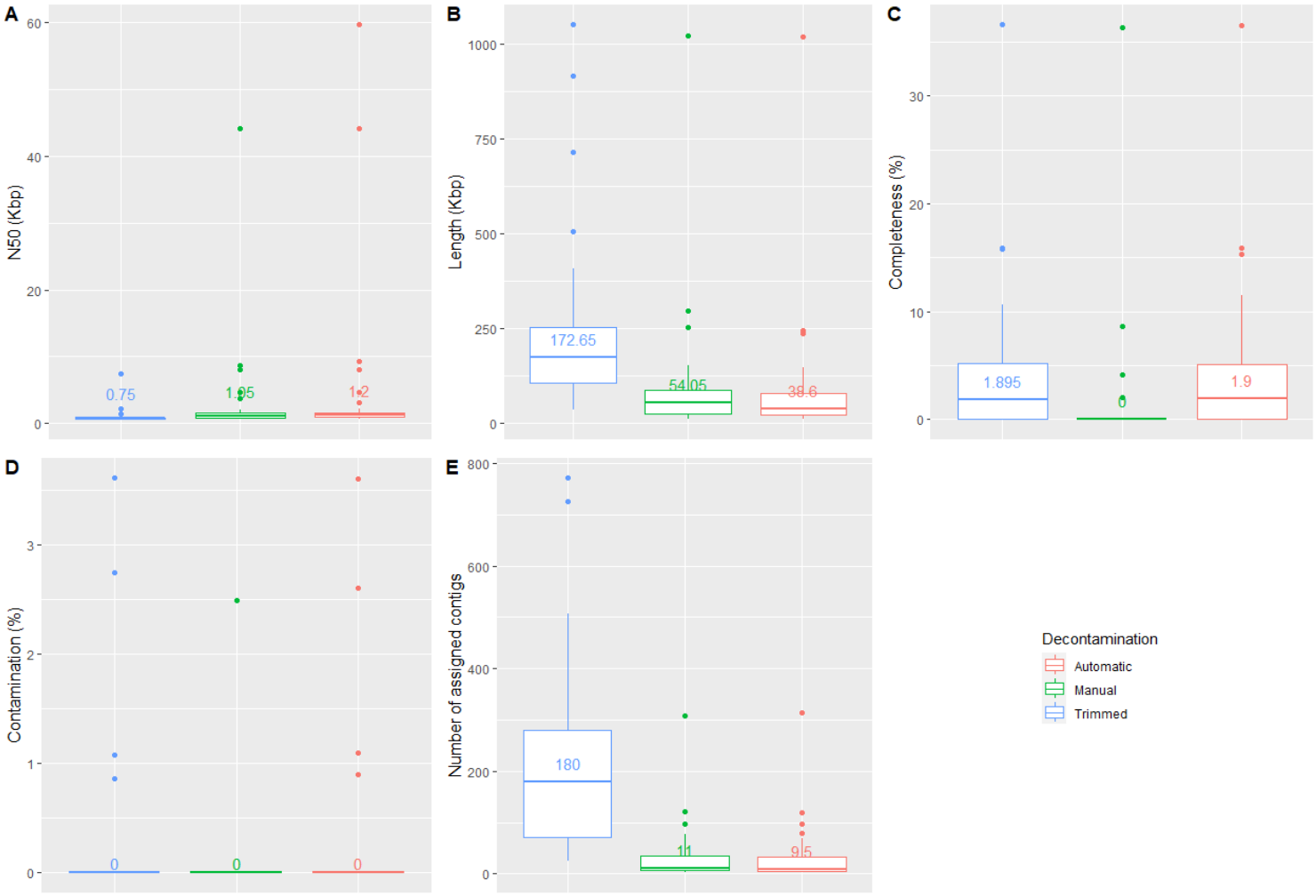


Figure 3. Sequence statistics of the 30 selected SAGs from our pure strains dataset with only trimmed reads (blue), manual decontamination (green), or automatic decontamination (red). For the manual decontamination, reads and contigs of Eukaryota, Enterobacteriaceae, Bacilliaceae, and Salsavirus taxa were removed. The median is indicated for each boxplot (A) N50 measurements in Kb correspond to the size of the smallest contig at 50% of the total length of the assembly (B) length of assemblies in Kb (C) completeness of the assembly estimated by CheckM (D) putative contamination estimates with CheckM and (E) number of taxonomically identified contigs by Kraken2 for each assembly.

Table 1. List of putative contaminants detected by the automatic decontamination pipeline SINCERE DATA in both pure strains and published (Berube et al., 2018) datasets. Their identification is based on the listing of overly covered regions along the genome of each sample (i) above a z-score (i.e. standard deviations from the mean coverage of the region) of 2 for more than 30bp and (ii) with a taxonomic identification different than the main taxa found in the sample by Kraken2. The green taxa represent the potential contaminants validated as contaminants based on the outlier ratio of a minimum of 0.2 and sample ratio of a maximum of 0.2. The total candidate contaminants list from the published dataset used (Berube et al., 2018) to test the automated SINCERE DATA contained 181 taxa.

	taxID	Taxa Name	Outliers ratio: >0.2	Sample ratio: <0.2
Pure strain dataset	**2842328**	**Salasmaviridae**	**0.33**	**0.17**
	9604	**Hominidae**	**0.23**	**0.13**
	90964	Staphylococcaceae	**0.30**	0.23
	1236	Gammaproteobacteria	0.17	**0.07**
	3650	Cucurbitaceae	0.13	**0.00**
	186817	Bacillaceae	0.10	**0.40**
	1653	Corynebacteriaceae	0.10	**0.00**
	772	Bartonellaceae	0.07	**0.00**
	91347	Enterobacterales	0.07	**0.00**
	468	Moraxellaceae	0.07	**0.00**
	19955	Ebenaceae	0.03	**0.00**
	3803	Fabaceae	0.03	**0.00**
	3503	Fagaceae	0.03	**0.00**
	41294	Nitrobacteraceae	0.03	**0.00**
	31957	Propionibacteriaceae	0.03	**0.00**
	2005525	Tannerellaceae	0.03	**0.00**
Published dataset (Berube et al. 2018)	9604	**Hominidae**	**0.50**	**0.00**
	3803	**Fabaceae**	**0.46**	**0.00**
	4479	**Poaceae**	**0.43**	**0.00**
	4144	**Oleaceae**	**0.39**	**0.00**
	186817	**Bacillaceae**	**0.39**	**0.00**
	4070	**Solanaceae**	**0.36**	**0.00**
	641	**Vibrionaceae**	**0.36**	**0.00**
	4210	**Asteraceae**	**0.36**	**0.00**
	3629	**Malvaceae**	**0.36**	**0.00**
	49546	**Flavobacteriaceae**	**0.32**	**0.00**
	31979	**Clostridiaceae**	**0.29**	**0.00**
	2808963	**Arcobacteraceae**	**0.29**	**0.00**
	3977	**Euphorbiaceae**	**0.29**	**0.00**
	3745	**Rosaceae**	**0.25**	**0.00**
	2762318	**Weeksellaceae**	**0.25**	**0.00**
	4710	**Arecaceae**	**0.21**	**0.00**
	4118	**Convolvulaceae**	**0.21**	**0.00**
	3465	**Papaveraceae**	**0.21**	**0.00**
	468	**Moraxellaceae**	**0.21**	**0.00**
	3650	**Cucurbitaceae**	**0.21**	**0.00**
	90964	**Staphylococcaceae**	**0.21**	**0.00**
	33958	**Lactobacillaceae**	**0.21**	**0.00**
	24986	Rubiaceae	0.18	0.00
	1162	Nostocaceae	0.18	0.00
	3602	Vitaceae	0.18	0.00
	72294	Campylobacteraceae	0.18	0.00
	81852	Enterococcaceae	0.18	0.00
	4671	Dioscoreaceae	0.18	0.00
	4642	Zingiberaceae	0.18	0.00
	119060	Burkholderiaceae	0.18	0.00
	...	...	...	...

For the manual decontamination of the pure strains dataset, the Eukaryotes, Enterobacteriaceae, Salasvirus, and Bacilliceae reads and contigs were removed. The Enterobacterales were also detected as a contaminant by the pipeline prior to its full execution because of their high proportion in the samples (i.e. main taxon in more than 70 % of the samples) and their reads were therefore removed for the total pipeline execution. We also manually removed the Bacilliaceae reads and contigs because of cross-contamination from these wells to their neighbours, not visible in this selection of 30 samples by the pipeline but detected in the total collection (data not shown). These four taxa were similarly detected with SINCERE DATA as potential contaminants (Table 1). A total of 16 potential contaminants (i.e. coverage pics above the z-score with different taxonomic identification than the sample from which it is extracted) were identified by the pipeline in the outliers regions, but only the Homindeae and Salasvirus taxa were confirmed as contaminants as they were present in more than 20% of the outlier regions and less than 20% of the samples.

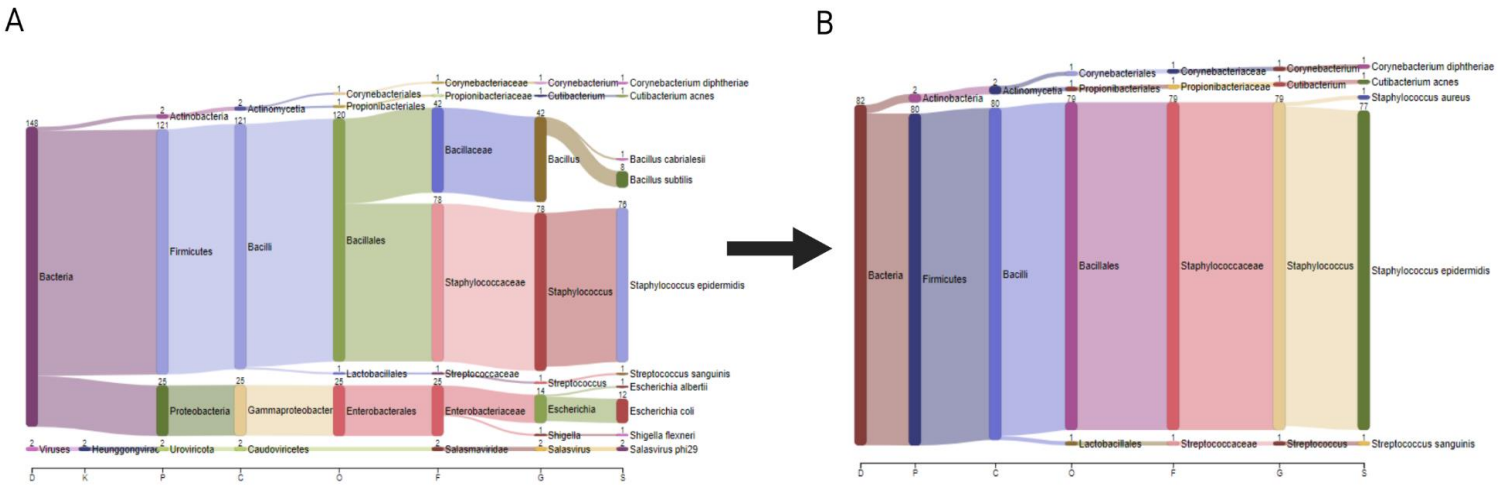


Figure 4. Sankey diagram obtained with Pavian of Kraken taxonomic assignment of contigs from one sample of the pure strains dataset before and after the automatic decontamination with the SINCERE DATA pipeline. This sample before decontamination, with trimmed reads only (A), was 140.4 Kbp long, with an N50 of 0.8 Kb, putative completeness, and a contamination rate of 1.72 % and 0 % respectively. After the automatic decontamination (B), the assembly length was 95.6 Kb, N50 was 1.2 Kp and putative-completeness and contamination were at 1.7 % and 0 %.

Despite the low putative contamination estimates by CheckM (Figure 3D), the visualisation of the sample's taxonomic taxa showed no unique affiliation of the contigs (Figure 4A). The assembly of this example sample from trimmed reads shows various taxonomy, half of them corresponding to the target DNA of *Staphylococcus epidermidis* (76 contigs out of 150), but some to the *Bacillus* genus (42 contigs), the Enterobacterales order (25 contigs), the *Salavirus phi29* (2 contigs) and Actinobacteria (2 contigs). The automated decontamination allowed the identification of the *Bacillus* genus, Enterobacterales order and the *Salasvirus phi29* as contaminants and were therefore removed from the dataset to obtain a cleaner assembly with a total of 82 contigs (Figure 4B).

2.3.2- Application to SAGs dataset from environmental samples

The decontamination pipeline was tested on a dataset from Berube et al. (2018), who recovered the SAGs from marine samples, with in theory more contaminants (i.e. from environmental DNA that could be co-isolated with the cell) than our samples from pure cultures. Here we evaluated the capacity of the pipeline to work on a reads dataset from environmental samples to improve the purity of these genomes for a more accurate representativity of uncultivated organisms.

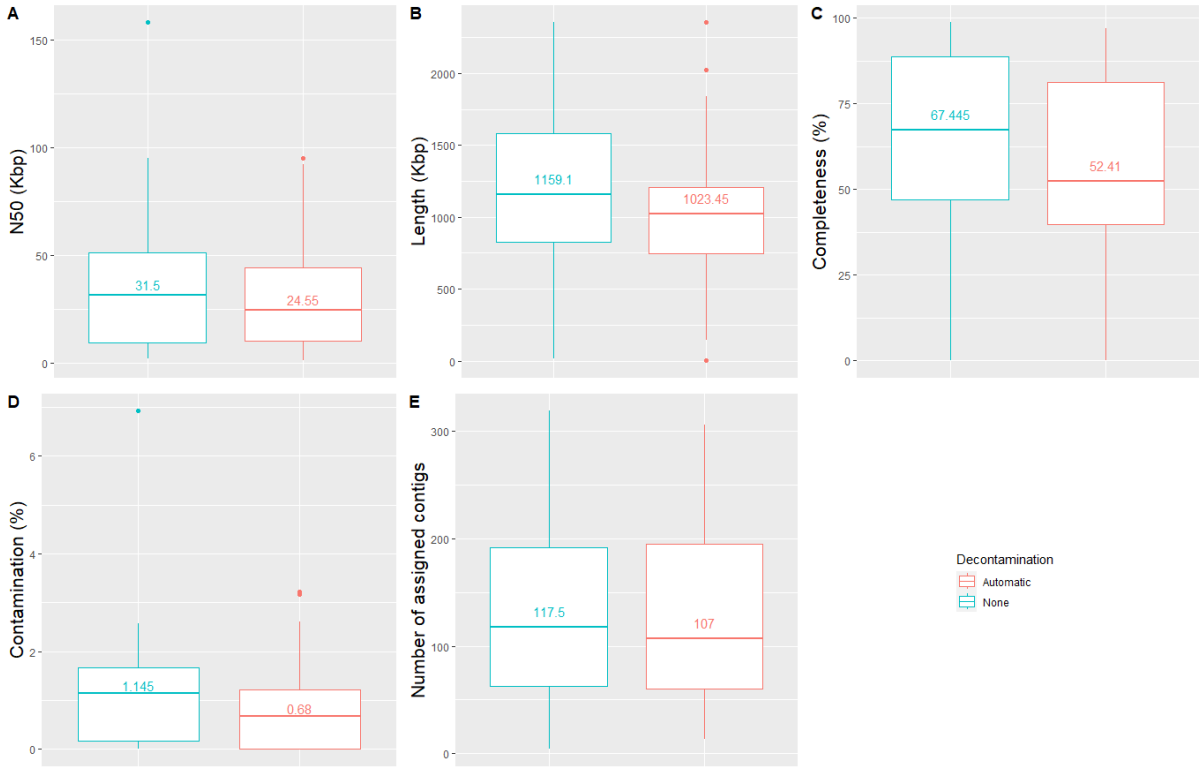


Figure 5. Sequence statistics of the 30 selected SAGs (Berube et al., 2018) without (blue), and with automatic decontamination with the SINCERE DATA pipeline (red). The median values are indicated for each boxplot. (A) The N50 measurements in Kb correspond to the size of the smallest contig at 50% of the total length of the assembly. (B) The lengths of assemblies in Kb. (C) The putative completeness of the assembly measured by CheckM. (D) putative contamination estimates determined by Check M and (E) the number of assigned contigs for each assembly.

The 30 randomly selected SAGs from a published dataset (Berube et al., 2018) presented lower N50 measures after the automatic decontamination (Figure 5A). The SINCERE DATA decontamination leads to a decrease of the median assembly length ((Figure 5B), the number of assigned contigs (Figure 5E), and the estimated contamination with CheckM (Figure 5D).

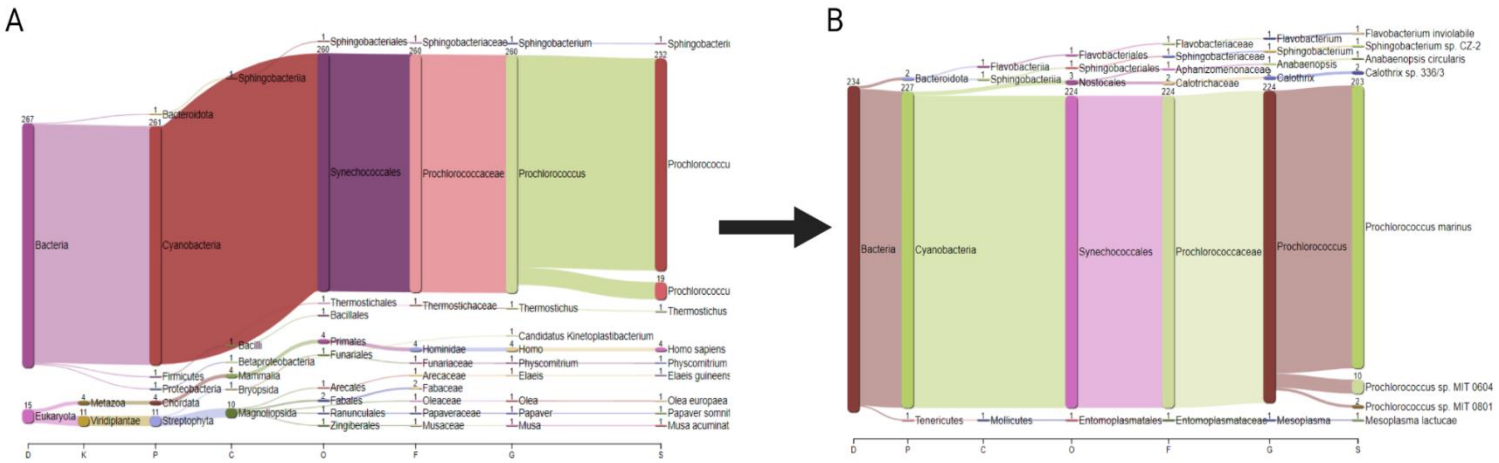


Figure 6. Sankey diagram obtained with Pavian of Kraken taxonomic assignment of contigs from the sample AG-402-K10 of the published dataset (Berube et al., 2018) before and after the automatic decontamination with the SINCERE DATA pipeline. This sample before decontamination (A) was 1644.7 Kbp long, with an N50 of 36.3 Kbp, completeness and a contamination rate of 92.45 % and 6.93 % respectively. After the automatic decontamination (B), the assembly length was 1549.2 Kbp, N50 was 29.5 Kbp and completeness and contamination were at 94.25 % and 3.17 %.

The measured completeness of this sample before the decontamination was lower than after the automatic decontamination: 92.45 % versus 94.25%. Despite this high completeness before decontamination, contaminants were detectable in the sample (Figure 6A), including 15 contigs belonging to various Eukaryota. After the removal of these contigs, the final assembly was 1.55 Mb long and contained 234 contigs from which 203 contigs were identified as *Prochlorococcus marinus* (Figure 6B).

A total of 181 potential contaminants were detected within the 30 selected SAGs (Table 1). After filtration, 22 taxa were considered as contaminants because they were present in more than 20 % of the outlier regions of all samples but their identification was identical to less than 20 % of the identity of the samples. Among them, bacteria but also eukaryotes were found, mainly plant taxa (12 out of the 22 taxa removed).

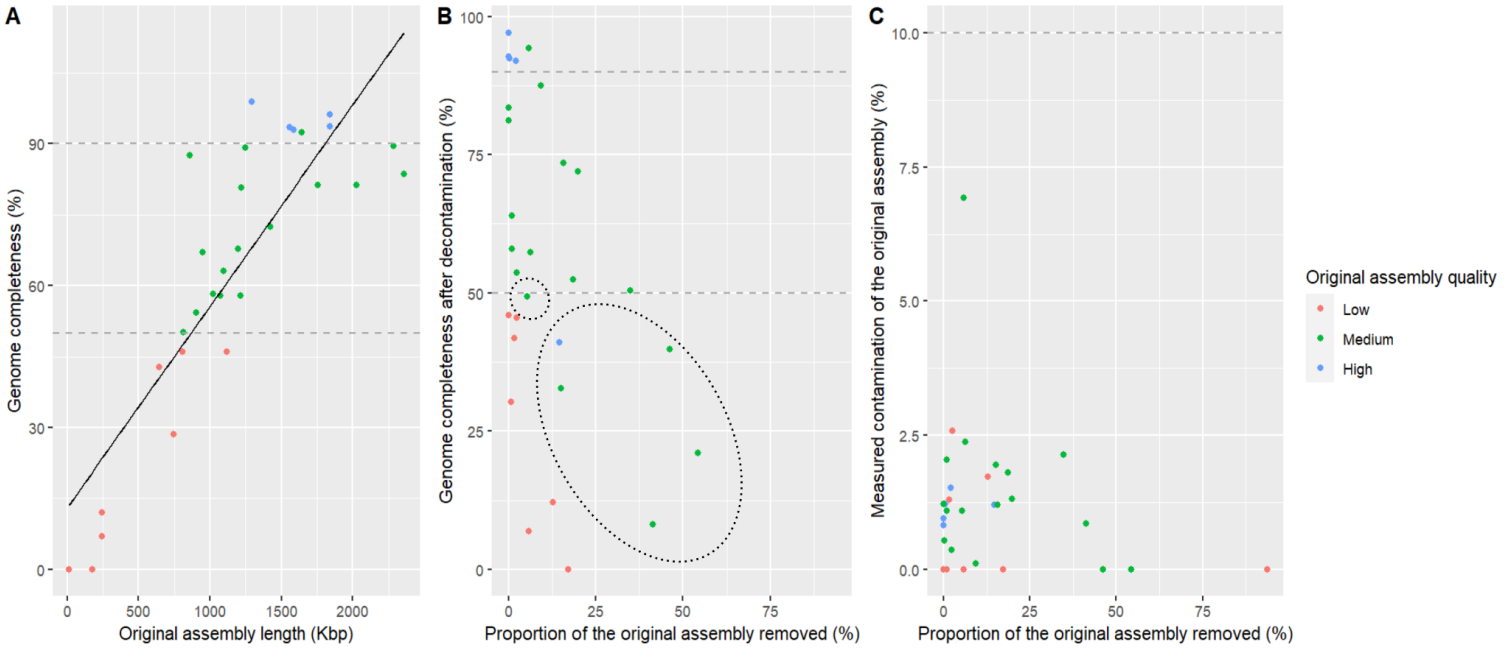


Figure 7. Evaluation of the putative completeness and contamination (measured by CheckM) correlation with the assembly length for each SAGs randomly chosen from the published dataset (Berube et al., 2018). The colours correspond to the assembly quality evaluation based on the GSC (Bowers et al., 2017): low-quality SAGs present completeness and contamination below 50 % and 10 %, medium-quality SAGs a completeness between 50 and 90% with 10% contamination maximum and high-quality SAGs have a completeness superior to 90% and contamination below 5%. Dash lines represent these thresholds. (A) Correlation between the genome completeness measured on the original assembly without decontamination and its length. (B) Correlation between the completeness measured after the automatic decontamination and the proportion of the assembly that was removed based on contamination identification with SINCERE DATA (Table 1). The regression line is indicated in black. (C) Correlation between the measured contamination (CheckM) of the original assembly and the proportion of the assembly that was removed based on contamination identification with SINCERE DATA. The one SAG in green with a completeness above 90 % in figure A was classified as medium quality because of its contamination rate that was higher than 5% (Figure C).

The completeness measurements were correlated to the assembly length of the original assembly made from the published dataset, without decontamination (Berube et al., 2018), where the biggest assemblies had the highest completeness measurements (Figure 7A). The genome completeness after the decontamination with the SINCERE DATA pipeline was not correlated to the portion of the assembly that was identified as a contaminant and removed, but above 15% of the assembly removal, the completeness of a few samples (dashed circle,

Figure 7B) changed quality categories: 5 medium quality and 1 high-quality SAGs became of low quality. These samples were highly contaminated with Eukaryota and Gammaproteobacteria taxa. The proportion of the assemblies that were removed was not correlated to the measured contamination by CheckM of the original samples (Figure 7C).

2.4- Discussion

Our protocol for single-cell genomics library preparation allowed the recovery of 30 SAGs from pure strains and highlighted the necessity for decontamination procedures of single-cell datasets. The contaminants and protocol efficiency were easier to evaluate on pure strains with referenced genomes and demonstrated the precautions that should be taken when preparing single-cell libraries for sequencing. Improvements of this protocol to limit the sample contamination should be focused on lowering the reaction volumes (Sobol 2023) and improving the quality of the genome amplification step (Stepanauskas et al., 2017). The detection of contamination from isolated cells questions the amount of external DNA present in metagenomics which are processed as targeted DNA. The use of single-cell genomics shows its capacity to distinguish between targeted and environmental and contaminant DNA for future improvements in genomic data interpretation.

Despite the various contamination sources and profiles, the SINCERE DATA pipeline was able to identify and remove contaminants while preserving the integrity of the targeted assembly. The thresholds used for the presence of the candidate contaminant sequences in the outliers and in the sample are adjustable depending on the study design. Here, both datasets aimed at specific microbial taxa with in theory little diversity in the dataset. Therefore, the threshold for the presence of taxa in outlier regions was very strict (20 %) but could be raised to 60 % in a bigger and more diverse sample. We would also expect these potential contaminants to be the dominant taxa in fewer samples from bigger datasets, the threshold could therefore be set at 10 % instead of 20 %.

The sequence statistics before the decontamination of both pure strains and published datasets (Berube et al., 2018) were very different, and qualities assessed by the GSC (Bowers et al., 2017) with Check M were lower for the assemblies from single cell of pure bacterial

cultures (Figures 3 and 5). This can be explained by multiple factors: i) The genome size of *Procholococcus* targeted by the published study is relatively small (i.e. 1.6-2.7 Mbp, Berube et al., 2018) compared to one of the two strains used in our study (*P. fluorescens genome size is* 6.7 Mbp, (Rainey, Bailey, and Thompson 1994) and 2.5 Mb for *S. epidermidis* (Galac et al. 2019)) and therefore will be measured as more complete with the same sequencing effort (Zheng et al 2022) and ii) The cyanobacterial SAGs were selected for sequencing based on the kinetic of the genome amplification (MDA or WGA-X) which was proved to be positively correlated with good genome recoveries (Stepanauskas et al., 2017). On the contrary, a random selection of MDA-amplified genomes can show very poor genome recovery rates even on well-referenced strains such as *E.coli* (Stepanauskas et al., 2017).

We have developed an automated decontamination pipeline for single-cell genomics data to clean up SAGs with at least as much precision as manual decontamination. On pure strains, the pipeline detected the same contaminant taxa that we manually removed with similar final assembly quality metrics. Moreover, no sequences from the targeted strains (i.e. *Pseudomonas fluorescens* and *Staphylococcus epiderminis*) were automatically removed, showing the ability of the procedure to detect DNA of interest despite the initial contamination detected with Kraken (Figure 4). The pipeline warned us about important proportions of Enterobacteriaceae in the samples that should be deleted, this step can be ignored if desired depending on the experimental design.

We then applied this approved automatic decontamination procedure to a published dataset, obtained from environmental samples, to validate its ubiquity. We detected potential contaminating sequences (Table 1) including for instance plants interpreted as free environmental DNA having contaminated the single cell sequence data and highlighted either the potential contaminants from sample manipulations or the taxonomic affiliation limitations of Kraken which attempts to identify sequences with the closest match, even distant (Wood et al., 2019). These sequences were most likely originating from free bacterial and eukaryotic DNA (i.e. environmental DNA) isolated together with the isolated cells. This environmental DNA sequenced simultaneously with the targeted cell was detectable and removable at both the reads level and contig level on the final assembly allowing to improve the SAGs purity. For some SAGs, this was beneficial regarding the estimated completeness and contamination measurements (Figure 6). However, the general dynamics for the

completeness decreased after the decontamination, putting into question the computed estimators of completeness and contamination by the CheckM tool for SAGs quality measurements (Bowers et al., 2017). CheckM detects broad sets of genes across bacterial and archaeal genomes and therefore is necessarily influenced by both the length of the assembly and the reference data content (Parks et al., 2015). We would, based on this evaluation, have expected to find more contaminated sequences with our pipeline in low-quality assemblies, which was not what we observed (Figure 7B). Instead, some samples classified as 'high' and 'medium' quality by CheckM before the decontamination dropped into the 'low' quality category after the decontamination with a high proportion of removed sequences (Figure 7B). This underlines that the estimated completeness with CheckM is a good quality metric under the assumption that assemblies are not chimeric. The putative contamination metric measured by CheckM is used for chimera detection, also based on marker gene detection. We found no correlation of this measured contamination with the portion of sequenced removed with our decontamination procedure (Figure 7C). The incapacity of CheckM to detect eukaryotic sequences was limiting for putative contamination assessment by this tool, as many contaminants were eukaryotic in both datasets (Figure 4, Figure 6, Table 1). The "quality" of SAGs should therefore contain a length (i.e. completeness, genome coverage…) and purity evaluation of the assemblies based on other criteria than marker genes. Other tools exist for the evaluation of genome completeness and contamination (i.e. Anvi'o (Eren et al., 2015), ProDeGe (Tennessen et al., 2016) and acdc (Lux et al., 2016)) but CheckM remains the most widely used for SAGs quality evaluation (Anstett et al. 2023; Chijiiwa et al. 2020; Nishikawa et al. 2022).

2.5- Conclusion

This work shows the necessity of deep decontamination prior to and after genome assemblies, especially on uncultivated taxa, to avoid the referencing of biased genomic information on which bioinformatic tools functioning are based. Various types of single-cell omics data are increasingly produced, with diverging strategies to handle them. Here we propose this pipeline as an additional attempt to unify and improve the quality and comparison of the single-cell datasets (Hugenholtz et al. 2021; Vollmers et al. 2022) that will

allow standardised data contamination exploration and provide cleaner genomes from the uncultivated majority. This automated decontamination, based on taxonomy and coverage of genome regions, will keep evolving with the increasing number of bacterial genome explorations in microbiology (Jiao et al. 2021; Rinke et al. 2013; Zamkovaya et al. 2021). However, bioinformatic procedures such as the one developed here should be coupled with a general improvement of single-cell omics data production to limit contamination sources.

Perspectives

A benchmarking study will be done to compare the SINCERE DATA decontamination pipeline efficiency to other decontamination procedures in the literature which are based on the manual selection of the contaminants to remove.

2.6- Supplementary

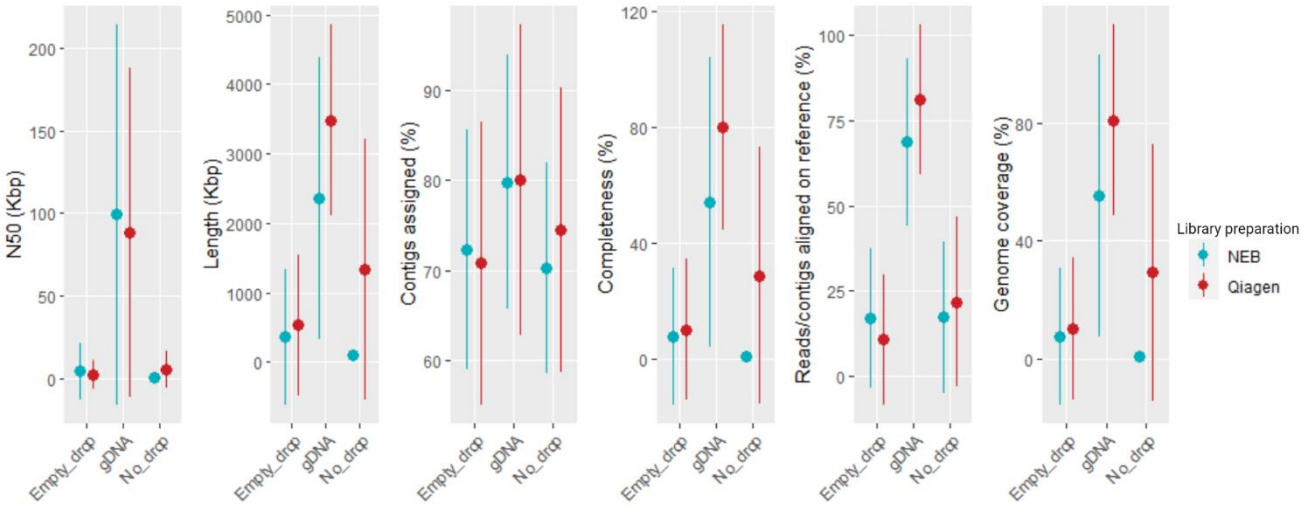


Figure S1. Sequence statistics on experimental controls: empty drop, gDNA and no drop with QIAGEN or NEB library preparation.

2.7- Bibliography

Alneberg, Johannes et al. 2018. “Genomes from Uncultivated Prokaryotes: A Comparison of Metagenome-Assembled and Single-Amplified Genomes.” *Microbiome* 6(1): 173. https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0550-0 (June 26, 2019).

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data

Anstett, Julia et al. 2023. “A Compendium of Bacterial and Archaeal Single-Cell Amplified Genomes from Oxygen Deficient Marine Waters.” *Scientific data* 10(1): 332.

Bagnoli, Johannes W. et al. 2018. “Sensitive and Powerful Single-Cell RNA Sequencing Using McSCRB-Seq.” *Nature Communications 2018 9:1* 9(1): 1–8. https://www.nature.com/articles/s41467-018-05347-6 (August 31, 2023).

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology, 19(5), 455-477.

Bawn, Matt et al. 2022. “Single-Cell Genomics Reveals Population Structures from in Vitro Evolutionary Studies of Salmonella.” *Microbial Genomics* 8(9). https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000871.

Berube, Paul M. et al. 2018a. “Data Descriptor: Single Cell Genomes of Prochlorococcus, Synechococcus, and Sympatric Microbes from Diverse Marine Environments.” *Scientific Data* 5(March): 1–11.

Blainey, Paul C. 2013. “The Future Is Now: Single-Cell Genomics of Bacteria and Archaea.” *FEMS Microbiology Reviews* 37(3): 407–27. https://academic.oup.com/femsre/article-lookup/doi/10.1111/1574-6976.12015 (February 27, 2019).

Bowers, Robert M., Devin F.R. Doud, and Tanja Woyke. 2017. “Analysis of Single-Cell Genome Sequences of Bacteria and Archaea.” *Emerging Topics in Life Sciences* 1(3): 249–55. http://www.emergtoplifesci.org/lookup/doi/10.1042/ETLS20160028 (June 28, 2019).

Bowers, Robert M et al. 2017. “Minimum Information about a Single Amplified Genome

(MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology* 35(8): 725–31. http://www.nature.com/articles/nbt.3893 (July 2, 2019).

Breitwieser, Florian P., and Steven L. Salzberg. 2020. "Pavian: Interactive Analysis of Metagenomics Data for Microbiome Studies and Pathogen Identification." *Bioinformatics* 36(4): 1303. /pmc/articles/PMC8215911/ (October 2, 2023).

Bushnell, Brian, Jonathan Rood, and Esther Singer. 2017. "BBMerge – Accurate Paired Shotgun Read Merging via Overlap." *PLOS ONE* 12(10): e0185056. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0185056 (September 22, 2023).

Chen, Zixi, Lei Chen, and Weiwen Zhang. 2017. "Tools for Genomic and Transcriptomic Analysis of Microbes at Single-Cell Level." *Frontiers in Microbiology* 8: 1831. http://journal.frontiersin.org/article/10.3389/fmicb.2017.01831/full (June 21, 2019).

Chijiiwa, Rieka et al. 2020. "Single-Cell Genomics of Uncultured Bacteria Reveals Dietary Fiber Responders in the Mouse Gut Microbiota." *Microbiome* 8(1): 5. http://www.ncbi.nlm.nih.gov/pubmed/31969191 (February 13, 2020).

Cornet, Luc, and Denis Baurain. 2022. "Contamination Detection in Genomic Data: More Is Not Enough." *Genome Biology* 23(1): 1–15. https://doi.org/10.1186/s13059-022-02619-9.

Dabdoub, SM (2016). kraken-biom: Enabling interoperative format conversion for Kraken results (Version 1.2)

Danecek, Petr et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10(2): 1–4. https://dx.doi.org/10.1093/gigascience/giab008 (September 22, 2023).

Davis, Kimberly M., and Ralph R. Isberg. 2016. "Defining Heterogeneity within Bacterial Populations via Single Cell Approaches." *BioEssays* 38(8): 782–90. http://doi.wiley.com/10.1002/bies.201500121 (November 2, 2020).

Eren, A. Murat et al. 2015. "Anvi'o: An Advanced Analysis and Visualization Platformfor 'omics Data." *PeerJ* 2015(10): e1319. https://peerj.com/articles/1319 (September 25,

2023).

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics* 32(19): 3047–48. https://dx.doi.org/10.1093/bioinformatics/btw354 (September 22, 2023).

Galac, M. R. et al. 2019. “Complete Genome Sequence of Staphylococcus Epidermidis CSF41498.” *Microbiology Resource Announcements* 8(2). /pmc/articles/PMC6328648/ (September 26, 2023).

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics (Oxford, England)* 29(8): 1072–75. https://pubmed.ncbi.nlm.nih.gov/23422339/ (September 22, 2023).

Hugenholtz, Philip et al. 2021. “Prokaryotic Taxonomy and Nomenclature in the Age of Big Sequence Data.” *ISME Journal* 15(7): 1879–92.

Jiao, Jian Yu et al. 2021. “Microbial Dark Matter Coming to Light: Challenges and Opportunities.” *National Science Review* 8(3): 1–5.

Kashtan, Nadav et al. 2014. “Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus.” *Science* 344(6182): 416–20.

Labonté, Jessica M. et al. 2015. “Single-Cell Genomics-Based Analysis of Virus-Host Interactions in Marine Surface Bacterioplankton.” *ISME Journal* 9(11): 2386–99.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods 2012 9:4* 9(4): 357–59. https://www.nature.com/articles/nmeth.1923 (September 22, 2023).

López-Escardó, David et al. 2017. “Evaluation of Single-Cell Genomics to Address Evolutionary Questions Using Three SAGs of the Choanoflagellate Monosiga Brevicollis.” *Scientific Reports* 7(1): 11025. http://www.nature.com/articles/s41598-017-11466-9 (July 2, 2019).

Lux, Markus et al. 2016. “Acdc - Automated Contamination Detection and Confidence Estimation for Single-Cell Genome Data.” *BMC Bioinformatics* 17(1): 1–11.

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1397-7 (September 26, 2023).

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17(1): 10–12. https://journal.embnet.org/index.php/embnetjournal/article/view/200/479 (September 22, 2023).

Mauger, S., C. Monard, C. Thion, and P. Vandenkoornhuyse. 2022. “Contribution of Single-Cell Omics to Microbial Ecology.” *Trends in Ecology and Evolution* 37(1): 67–78.

Nishikawa, Yohei et al. 2022. “Validation of the Application of Gel Beads-Based Single-Cell Genome Sequencing Platform to Soil and Seawater.” *ISME Communications 2022 2:1* 2(1): 1–11. https://www.nature.com/articles/s43705-022-00179-4 (December 19, 2022).

Pachiadaki, Maria G. et al. 2019. “Charting the Complexity of the Marine Microbiome through Single-Cell Genomics.” *Cell* 179(7): 1623-1635.e11.

Parks, Donovan H. et al. 2015. “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes.” *Genome Research* 25(7): 1043–55.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26(6): 841–42. https://dx.doi.org/10.1093/bioinformatics/btq033 (October 2, 2023).

Raghunathan, Arumugham et al. 2005. “Genomic DNA Amplification from a Single Bacterium.” *Applied and environmental microbiology* 71(6): 3342–47. http://www.ncbi.nlm.nih.gov/pubmed/15933038 (June 21, 2019).

Rainey, P. B., M. J. Bailey, and I. P. Thompson. 1994. “Phenotypic and Genotypic Diversity of Fluorescent Pseudomonads Isolated from Field-Grown Sugar Beet.” *Microbiology* 140(9): 2315–31. https://www.microbiologyresearch.org/content/journal/micro/10.1099/13500872-140-9-2315 (September 26, 2023).

Rinke, Christian et al. 2013. “Insights into the Phylogeny and Coding Potential of Microbial Dark Matter.” *Nature* 499(7459): 431–37.

Van Rossum, Thea, Pamela Ferretti, Oleksandr M. Maistrenko, and Peer Bork. 2020. “Diversity within Species: Interpreting Strains in Microbiomes.” *Nature Reviews Microbiology* 18(9): 491–506.

Stepanauskas, Ramunas et al. 2017. “Improved Genome Recovery and Integrated Cell-Size Analyses of Individual Uncultured Microbial Cells and Viral Particles.” *Nature Communications* 8(1): 1–10.

Tennessen, Kristin et al. 2015. “ProDeGe: A Computational Protocol for Fully Automated Decontamination of Genomes.” *The ISME Journal 2016 10:1* 10(1): 269–72. https://www.nature.com/articles/ismej2015100 (September 26, 2023).

Vollmers, John, Sandra Wiegand, Florian Lenk, and Anne Kristin Kaster. 2022. “How Clear Is Our Current View on Microbial Dark Matter? (Re-)Assessing Public MAG & SAG Datasets with MDMcleaner.” *Nucleic Acids Research* 50(13): e76–e76. https://dx.doi.org/10.1093/nar/gkac294 (July 6, 2023).

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. “Improved Metagenomic Analysis with Kraken 2.” *Genome Biology* 20(1): 1–13. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1891-0 (September 22, 2023).

Woyke, Tanja, Devin F.R. Doud, and Frederik Schulz. 2017. “The Trajectory of Microbial Single-Cell Sequencing.” *Nature Methods* 14(11): 1045–54.

Zamkovaya, Tatyana, Jamie S. Foster, Valérie de Crécy-Lagard, and Ana Conesa. 2021. “A Network Approach to Elucidate and Prioritize Microbial Dark Matter in Microbial Communities.” *ISME Journal* 15(1): 228–44.

Zheng, Wenshan et al. 2022. “High-Throughput, Single-Microbe Genomics with Strain Resolution, Applied to a Human Gut Microbiome.” *Science* 376(6597). https://www.science.org/doi/10.1126/science.abm1483.

III- Conclusion of the chapter

We elaborated and optimized a protocol for single-cell genomics library preparation on bacteria. This development demanded many accessory controls, that themselves needed to be developed for proper evaluation of the molecular steps efficiencies. Preparing customised protocols from kits that did not provide confidential information regarding the enzyme used or buffer composition was also a challenge for the compatibility of each step of the workflow. This most likely prevents the application of single-cell genomics at reasonable prices and encourages the rise of homemade sample preparation, especially for the genome amplification step (Stepanauskas et al., 2017; Zheng et al., 2022). This step is responsible for most of the single-cell genomics limitations (i.e. contamination amplification and low genome coverage) and therefore the main focus for SAGs generation improvement (Gonzalez-Pena et al., 2021; Stepanauskas et al., 2017; Woyke et al., 2011). An additional way of improvement is the miniaturisation of the molecular reactions, for reduced cost, contaminants, and biases (Nishikawa et al., 2015). This has been tested, and, while being blind to the content and exact functioning of the molecular steps of the used kits, it was difficult and it did not simply require a reduction of the volumes. Indeed, at such low volumes, liquid evaporation is emphasised, enzymatic activities can be affected, and reaction times and temperatures must be reviewed. From my experience, options to optimize genome amplification remain notably library preparation based on the kinetic of the reaction as done in the literature (Stepanauskas et al., 2017). Also, coupled information on the SAGs generated could be added by performing a 16s rRNA gene PCR on the amplified genomes, for the comparison of the taxonomic results as well as easier visualisation of unreferenced genome information (Nishikawa et al., 2022).

We succeeded in generating SAGs with this protocol from referenced strains but had difficulties evaluating the quality of our genome assemblies compared to similar studies in the literature (see Chapter II notably). We have looked into the sequences datasets in depth but no ‘gold-standard’ strategies existed in the literature. This is how the automatic decontamination pipeline idea emerged, after noticing contaminations in SAGs generated from the sequence data of environmental samples (data presented in Chapter III, which to this date have not been decontaminated with the developed decontamination pipeline). This pipeline was validated on both referenced and environmental bacteria and fills a gap in the

bioinformatic treatment of such datasets. We expect this pipeline to be very useful for other research groups but also to evolve with SAGs library preparation, just like metagenomics data handling keeps evolving even after years of application in microbiology (Marotz et al., 2018; McArdle & Kaforou, 2020; Olson et al., 2018). Ideally, we would appreciate acknowledging the rise of more genomic tools adapted to single-cell data treatment for completeness and contamination evaluations as well as taxonomic assignation. With adapted tools, the decontamination pipeline functioning and output as the one developed here could only be improved.

Chapter III

Single-cell, meta-genomics, and mini-metagenomics: complementarities and limits for soil bacteria community exploration

The very first objective of this chapter, and of this thesis, was to explore ecological theories via the lens of single-cell omics. We have had hopes to explore the Black Queen Hypothesis at first, but finally chose a more descriptive approach to the soil bacterial communities which are very vast and largely unknown. The application of the developed library preparation protocol presented in Chapter II demonstrated the effort and time necessary to recover a few dozen SAGs. This questioned the feasibility of a broad-scale ecological experiment and raised more technical questions we hoped to address first regarding the technical padlock and difficulties we would encounter with environmental samples. Therefore, as a first step prior to ecological question testing, I explore in this chapter the suitability of SAGs recovered with the library preparation protocol for ecological testing and compare them to traditional metagenomics and newly employed mini-metagenomics. The possibilities for SAGs quality investigations are many, and to this date still ongoing. This chapter gathers the first results of this exploration and will keep evolving into an article in the near future.

I- Introduction

Molecular-based studies have revolutionized the microbiology field with technologies that do not cease to evolve. Single-cell genomics (SCG) has been used for a few years in microbial ecology and has broadened our view on fine-scale cell organizations and microbial diversity (Engel et al., 2014; Kashtan et al., 2014; Pachiadaki et al., 2019b). The use of SCG allows the discovery of unreferenced genomes and highlights populational variants more efficiently than traditional metagenomics (Neuenschwander et al., 2017; Van Rossum et al., 2020). Beyond the description of microbial community diversity, the strengths of SCG have also led to major advances in microbial ecology dynamics understanding. Single-cell omics applied to microbes are providing missing pieces in microbial ecology studies such as host-symbiont evolution (Chijiiwa et al., 2020; Labonté et al., 2015), bacterial community evolutionary and interactions potential in various ecosystems (Bawn et al., 2022; Garcia et al., 2018; Roux, Hawley, Torres Beltran, et al., 2014). Rare or abundant cells with specific features can be equally studied once isolated to get rid of abundance-dependent representation of diversity. This is of great interest in microbiology, where the total bacterial diversity is estimated to range from eight hundred thousand to beyond trillions (Id et al., 2019; Locey & Lennon, 2016) but with an estimation of only 2% of this diversity possessing a reference in databases (Z. Zhang et al., 2020). Mini-metagenomics has been positioned between single-cell and meta-genomics for fine-scale observation and consists of isolating multiple cells into the same sample to be processed just like single-cell samples, but with higher throughput. This approach has enabled the discovery of novel bacterial lineages in Yellowstone National Park hot springs (Yu et al., 2017), but is also not commonly used and our ability to recover pure populational information from such datasets still needs to be verified. The current microbial investigation is not equal between taxa and biomes: some remain primarily understudied (e.g. free-living organisms) while others are routinely cultured or sequenced (e.g. endosymbionts) (Zhang et al., 2020). This results in biased reference databases towards specific taxa, which influence bioinformatic treatments that rely on these databases to decrypt the unknown. This technical padlock does not totally prevent discoveries of unreferenced bacterial strains. Yearly, the Microbial Dark Matter (MDM) is being uncovered a little bit more with - among other approaches - the help of single-cell omics technologies to furnish additional information on bacterial diversity (J. Y. Jiao et al., 2021). These undescribed organisms play a central role in

microbial network connections and along with their identification modify our perception of ecosystem functioning (Zamkovaya et al., 2021). This MDM is today mainly investigated through the lens of metagenomics which produces large datasets from environmental DNA (Wooley et al., 2010). The reconstitution of potential genomes from environmental genetic material has been developed into Metagenome Assembled Genomes (MAGs) (Tyson et al., 2004) which greatly nourished the reconstitution of bacterial phylogeny (Hug et al., 2016; Parks et al., 2018) and the incorporation of uncultivated strain genomes in databases (Escudeiro et al., 2022).

A largely studied yet still mysterious portion of the microbial biomes is the soil ecosystem. The abundance of bacteria is estimated to vary between millions and billions of cells per gram of soil (Knudsen, 2010) with as many interaction possibilities. Soil bacteria have critical roles in biogeochemical cycles (Rütting et al., 2021; Swan et al., 2011) and plant health (Hassani et al., 2018). Understanding the ecological processes in soil is complex due to the large part of unknown bacterial diversity and functions but also of the various biotic and abiotic parameters influencing bacterial communities (Isobe et al., 2020; Wilpiszeski et al., 2020). The soil pH has been identified as a major driver for microbial community structure and composition (S. Jiao & Lu, 2020; Daniel R. Lammel et al., 2018; Y. Li et al., 2018; Rousk et al., 2010; A. Tripathi et al., 2018; Wan et al., 2020; Zhalnina et al., 2015) and indirectly influences microbiota by modifying bioavailability of nutrients and plants requirements, which, in return, modulates their interactions with their microbiota (Daniel R. Lammel et al., 2018; Wan et al., 2020). Many biogeochemical processes are closely correlated with pH (Malik et al., 2018; Neina, 2019), making it difficult to find a consistent effect of pH solely on bacterial communities, and to identify taxa acidity preferences across locations (Daniel Renato Lammel et al., 2015). The majority of the studies assessing pH influence on bacterial community structures are based on the Operational Taxonomy Units (OTU) diversity investigation generated by 16rRNA gene sequencing (Barnett et al., 2020; Bartram et al., 2014; S. Jiao & Lu, 2020; Daniel R. Lammel et al., 2018; Schlatter et al., 2020; B. M. Tripathi et al., 2018; Wan et al., 2020; Q. Xu et al., 2021; Yavitt et al., 2021). Different pH preference patterns have been identified for bacterial taxa, with variations depending on the study and its location (Daniel R. Lammel et al., 2018), showing that the effect of pH on bacteria might occur at more precise levels of organization than what can be observed from OTUs. Indeed, the pH preferences of

prokaryotes seem to be phylogenetically conserved (S. Jiao & Lu, 2020; Daniel R. Lammel et al., 2018; B. M. Tripathi et al., 2018), with specific sets of genes associated with acidity niches (Gubry-Rangin et al., 2015; Ramoneda et al., 2023). This highlights the necessity of whole genome gene sets investigation to understand bacterial preferences and future adaptation to environmental changes.

Few recent studies have started to look at environmental drivers involving pH for microbial dynamics with MAGs (Garner et al., 2023; Lee et al., 2022; R. Xu et al., 2022), with the limitation that MAGs present: no access to the population levels of these communities and possible production of chimeric genomes. Because intra-species and intra-populations variations are fundamental for bacterial communities functioning, interaction and evolution (Davis & Isberg, 2016; García-García et al., 2019), the exploration of gene content to answer such ecological questions should be set at the cell level. To this date, no broad use of SAG for this purpose has been tested on environmental samples but shows to be encouraging according to tests done on a few samples (López-Escardó et al., 2017).

Here, we aimed to recover genomes from soil bacterial communities by using single-cell, mini-meta-genomics and meta-genomics as complementary and comparative approaches to evaluate their differences in genome qualities and content. We evaluated the potential of reconstituted and partial genomes to be used for ecological applications such as pH preferences among bacteria and explored if single-cell whole genome sequencing is ready for such broad applications. We sampled soil from three pH treatments of the Craibstone field experiment in Aberdeen which has been extensively studied and characterized (Aigle et al., 2020; Bartram et al., 2014; Kemp et al., 1992), from which we recovered SAGs, mini-MAGs and MAGs each from aliquots of the same prepared cells. The qualities of these assemblies were assessed, as well as their capacity to provide input regarding bacterial community patterns from a diversity, phylogenetic and gene content perspective.

II- Materials and methods

2.1- Soil samples

The soil was taken from the experimental farm of Craibstone (Aberdeen,UK), where a gradient of pH has been maintained for over 65 years (Bartram et al., 2014). Soil samples coming from

the pH 4.5; 6 and 7.5 treatments were sampled in March 2022, sieved at 2 mm and stored at 4°C for 3 months. Seven days prior to cell extraction, 500 g of soil samples were placed in a 1.5 L sterile closed glass jar at 25°C in the dark for reactivation. Every two days, the jar was shortly opened to renew the air. The soil humidity was assessed by weighing approximately 10g of soil before and after being dried at 105°C for 48 h. The humidity of soil pH 4.5 was 21.5 % and 25 % for soils from pH 6 and 7.5.

Besides, the sample pH was measured after a 4-day desiccation in the oven at 38°C. To measure the carbon and nitrogen content of each sample, 15 g of soil was dried at 55°C for a week. The dry samples were then milled at maximum speed (30 Hz) for one minute in RETSCH Mixer Mill MM 400 with 20 mm zirconium beads. Lastly, 40 mg of milled soil was wrapped in tin paper for carbon and nitrogen measurements in Vario Pyrocube ELEMENTAR analyzer.

2.2- Cell extraction

The cell extraction procedure was largely inspired by the protocol of Ouyang et al. (2021): 40 g of soil was added to 80 mL of 0.5 % Tween 20 (Thermo Scientific) in sterile PBS (Thermo Scientific) in a blender (HENDI 230718 Blender) and blended at 22,000 rpm for 3 min at 1-minute intervals, with 1-minute incubation on ice to cool the mixture down. Between each soil, the blender was carefully rinsed with pure water and 70 % ethanol as previously described (Ouyang et al., 2021).

A 20 mL subsample of the soil extract was used for cell extraction and divided into eight aliquots of 2.5 mL that were added on top of 2 mL sterile 80 % Nycodenz in sterile 5 ml tubes using a 2 mL pipet. The tubes were centrifuged at 4°C, 15,000 g for 40 minutes with slow acceleration and deceleration. The cell layers from the tubes were pooled two by two using a pipette and filtered through a 30 µM MACS® SmartStrainers filter (Miltenyi Biotec) before being centrifuged for 15 minutes with the same previous parameters. Finally, the supernatant from the four tubes was discarded, and the cells were resuspended in 100 µL of sterile PBS in two tubes for DNA extraction or in 250 µL of sterile PBS for cell sorting in the two other tubes.

2.3- gDNA extraction for metagenomics

The genomic DNA (gDNA) extraction protocol was adapted from Nicolaisen et al (2008). Cells previously resuspended in 100 µL of PBS were mixed with 250 µL of 10% CTAB - 0,7 M NaCl buffer, 250 µL of 240 mM K_2HPO_4/KH_2PO_4 pH 8.0 buffer, and 500 µL of phenol-chloroform-isoamyl alcohol (25:24:1) pH 8 in 2 mL lysis matrix tubes (MP Biomedicals). The tubes were

then quickly vortexed and shaken during 3 minutes of fast bead beating using a TissueLyser (QIAGEN). After a 10-min centrifugation at 16,000 g and 4° C, the aqueous phase was transferred to a new 2 mL tube and completed with an equivalent volume of chloroform-isoamyl alcohol (24:1). The centrifugation was repeated, and the aqueous phase was mixed with two volumes of 30% PEG 6000 - 1.6 M NaCl buffer in a new 2 mL tube. The gDNA was precipitated by the addition of 4 µL of glycogen (20 mg/mL). The solution was homogenized by inverting the tubes and placed at -20°C overnight. The next day, the samples were thawed and centrifuged at 18000 g, at 4°C for 30 minutes. The supernatant was discarded and replaced by 500 µL of ethanol 70%. A second centrifugation was performed at 18,000 g, at 4°C for 10 minutes, the supernatant was discarded, and the tubes were left open to dry for 2 min under a PCR flow hood. The gDNA was finally resuspended in 50 µL of nuclease-free water (NFW). The quality and quantity of the gDNA were assessed with a Nanodrop 1000 spectrophotometer (Thermo Fisher Scientific), and the size of the fragments was controlled with a Fragment Analyzer (Agilent).

2.4- Single-cell staining, isolation, and lysis

The 250 µL solution containing the extracted cells was mixed with 1 µL of SYTO™ 13 Green Fluorescent Nucleic Acid Stain (ThermoFisher), vortexed, and kept in the dark for 30 minutes. The stained cells were washed three times following these steps: centrifugation (9000 g, 3 min); removal of the supernatant, and resuspension of the cell pellet in 250 µL PBS. Prior to cell isolation, 1.05 µl of lysis buffer was added by hand in each well of a 384-well plate (Roche) under a flow hood. The stock of lysis buffer was composed of DTT 100 mM, EDTA 10 mM, and KOH 0.4 M with a final addition of Tris HCl 1M pH 4 (Stepanauskas et al., 2017). The lysis buffer was diluted prior to the distribution of 1.05 µL (0.45 µL of buffer, 0.6 µL of NFW). The instrument cellenONE (Cellenion) was used for cell isolation purposes. The green fluorescence channel was used to optically detect the Syto 13 fluorescence and automatically sort the cells in wells of one 384-well plate per soil, which was maintained below 5°C during cell isolation by the cellenONE temperature-control system. 380 wells were dedicated to receive one cell and 4 wells were kept empty as negative controls for the library preparation. For mini-metagenomics, not 1 cell per well but 10 cells from the soil pH 4.5 only were isolated in 48 wells of a 384-well plate. After cell isolation, the plates were centrifuged for 2 minutes at 2000 g and kept at 4°C for 10 minutes. To stop the lysis, 0.5 µL of Stop solution (TrisHCl 1M pH 4)

was added to each well. After being centrifuged, the plates were sealed and placed at -20°C overnight.

2.5- Single-cell and mini-metagenomics genome amplification

The plates were thawed on ice and opened under a PCR flow hood. The genome amplification was performed using the REPLI-g Advanced DNA Single Cell Kit (QIAGEN). The volumes were lowered to a minimum for cost optimization and contamination limitation: 6 µL of mix consisting of 1.35 µL of NFW, 4.35 µL of reaction buffer and 0.3 µl of Phi29 polymerase were added to each well. The MDA was performed for 4 hours at 30°C, with a final step at 65°C for 3 min in a Roche LightCycler 380. Each amplification was quantified using Picogreen (Invitrogen) and the green fluorescence was read using a SAFAS Xenius spectrophotometer. From these measured MDA product concentrations, DNA concentrations of the samples were normalized to 5 ng/µL in NFW for further steps using a Biomek robot (Beckman Coulter).

2.6- Library preparation and sequencing

For the single-cell and mini-metagenomics, the libraries were prepared using the QIAseq FX DNA Library Kit (QIAGEN) and the volumes were also revised. On top of 2 µL of normalized MDA-amplified DNA, 3 µL of fragmentation mix was added, consisting of 1.5 µL of NFW, 0.5 µL of fragmentation buffer and 1 µL of enzyme. The total reaction, tubes, and plates were carefully kept cold during the manipulation. The plates were placed in a pre-cooled cycler (4°C) and the fragmentation was run for 12 min at 32°C, followed by 30 min inactivation at 65°C. The ligation of primer adapters was directly done on top of the fragmentation reaction by adding 2 µL of ligation buffer, 1.5 µL of water and 1 µL of enzyme. Custom adapters were designed for maximizing sample multiplexing and used in combinatorial pairs to identify 96 cells by crossing 20 unique barcodes. In each well, 0.5 µL of each adapter (i.e. 5' and 3' sides) at 10 µM was added. After the incubation of the plate at 20°C for 15 minutes, the samples were directly purified with AmPure beads (Beckman Coulter) at 0.8x and 1x with automated liquid handler Biomek (Beckman Coulter). Purified samples with unique identification were pooled per set of 96 combinatorial barcodes. The final indexing PCR was performed with the QIAseq FX DNA Library Kit buffers and settings but with the Nextera XT Index Kit v2 primers. Per group of 96 samples, a unique set of primers was used as an additional multiplexing level. The same protocol, i.e. library preparation kit, custom ligation adapters and final PCR primers

was used for the metagenomics library preparation but with the conservation of the QIAseq FX DNA Library Kit (QIAGEN) reaction volumes.

Final quality controls of the amplified libraries were performed after a final purification with 1X AmPure beads: quantification using QuBit 4 (Thermo Fisher Scientific), fragment size using the FragmentAnalyzer (Agilent), and library quantification using KAPA Library Quantification Kit (Illumina) in a ROCHE Lightcycler 480. The sequencing of libraries was done in two rounds: the single-cell, mini-metagenomics and metagenomic libraries of pH 4.5 were sequenced on a NextSeq 2000 using a P1 150 cycles flow cell. Samples from pH 6 and 7,5 were simultaneously sequenced on a NovaSeq 6000 with an S1 300 cycles flow cell.

2.7- Data treatment and analysis

2.7.1- SAGs , mini-MAGs and MAGs generation

Reads from single-cell samples were demultiplexed based on the dual indexing with cutadapt (V. 4.1, Martin, 2011) with an error rate of 15%. Cutadapt was also used for the adapter and quality trimming with a length set to 30 bp and minimal quality set to 15. The quality of the trimmed reads was measured with fastqc (V 0.11.5, Andrews, 2010) and the contaminants detection was performed with Kraken2 (V. 2.1.2, Wood et al., 2019) via taxonomic affiliation of reads on the PlusPFP precompiled database (downloaded the 31/01/23). Based on Kraken taxonomy, the reads were manually decontaminated by removing Eukaryotes, Enterobacteriaceae and Salasvirus reads. The Enterobacteriaceae were largely present in the datasets and are suspected to be carried by the MDA reagents as well as the Salasvirus from which the Phi29 polymerase is extracted. Many reads remained unassigned by Kraken, but were kept for the genome assemblies. The assembly was performed with Spades (V. 3.15.5, Bankevich et al., 2012) using the careful and single-cell option (--careful –sc) and contigs smaller than 500 bp were removed and assemblies below 50 kbp were not studied further. The taxonomy of the remaining unassigned contigs was assessed with BLAST (Altschul et al., 1997), and were either identified as eukaryotes or gave no hits, they were therefore removed. The quality and contamination of the SAG were evaluated with QUAST (Gurevich et al., 2013) and CheckM (Parks et al., 2015).

Metagenomic and mini-metagenomic reads data were processed with Anvi'o for metagenome-assembled genomes ((mini-)MAGs) generation and quality assessment (https://astrobiomike.github.io/metagenomics/metagen_anvio). (mini-)MAGs were manually refined to lower the detected contamination when possible by inspecting contigs with different coverage and QC content than the total MAG and verifying the Blast assignment and gene annotation, if available. The taxonomy of the assembled genomes was determined using Mash (V. 2.3, Ondov et al., 2016). The SAGs with identical identification were merged with Anvi'o to produce co-assembled genomes (CAGs) (Eren et al., 2015). Anvi'o was also used to detect the presence of ribosomal genes 5S, 16S, 18S, 23S, and 28S in each assembly.

2.7.2- Phylogenetic tree

Mash (V. 2.3, Ondov et al., 2016) was used to detect the closest reference for each assembly using mash distance against a collection of prokaryotic representative genomes from the NCBI refseq collection (downloaded the 11/01/23, O'Leary et al., 2016). In addition, eukaryotes references were added based on Hug et al. (2016), for a final collection of 17,556 reference genomes. This reference collection was used in mashtree (V. 1.2.0, Katz et al., 2019) to position our SAGs and MAGs into the Tree of Life. To construct the distance matrix, 21 bp-sized kmer and a sketch size of 100,000 were used. The minimum depth was fixed to one given the low coverage (<2x) in some contigs. The phylogenetic prediction was performed using the Neighbor-Joining (NJ) algorithm. From the closest reference information for each strain, a metadata table was created including taxonomic identifiers (taxids) of the reference strain extracted from NCBI using entrez-direct (v.16.2, Kans, 2023). Those taxids were used to extract detailed taxonomy using taxonkit (V. 0.15.0, Shen & Ren, 2021) based on NCBI taxonomic data (downloaded the 07/09/2023). The tree data were visualized and annotated using Itol (https://itol.embl.de/, Letunic & Bork 2021).

III- Results

3.1- Sequencing reads taxonomy

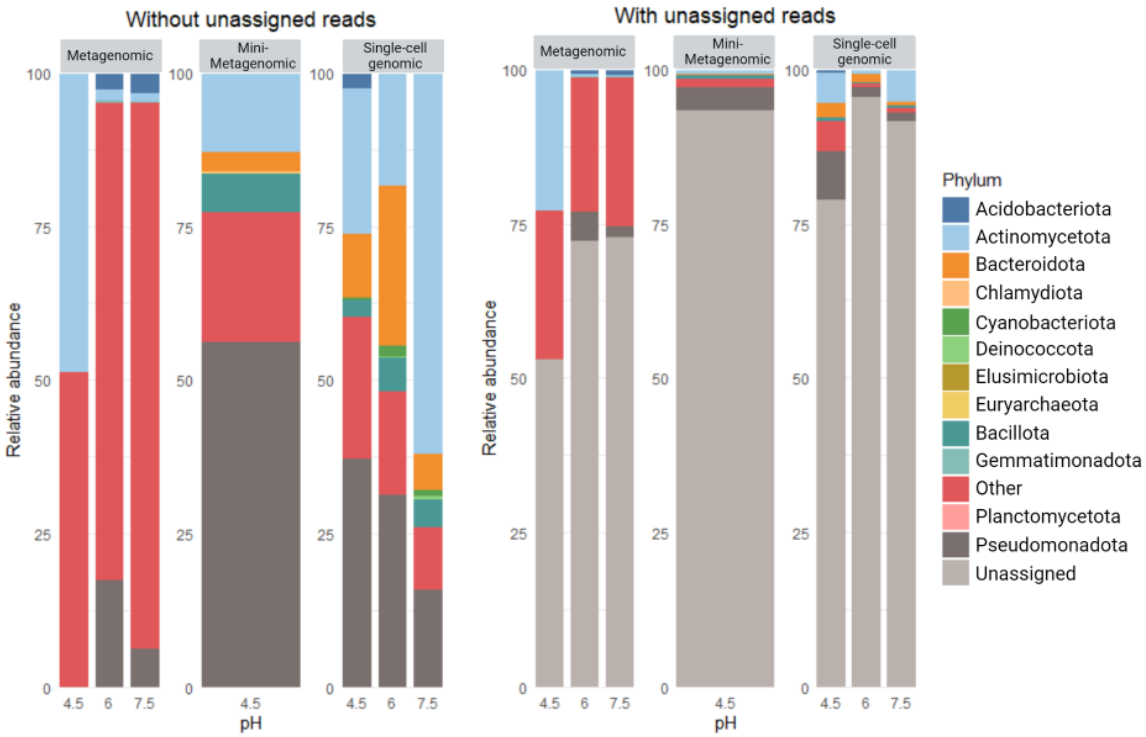


Figure 1. Relative abundance of reads per sample in single-cell, mini-metagenomics and metagenomics samples at the phylum level without or with unassigned reads. The taxonomy was determined by Kraken2. Phylum represented by less than 10,000 reads per sample were gathered in the "Other" group and contained 48 phylum in the metagenomic samples and 58 in the single-cell samples. The number of represented reads for each metagenomic sample were: 0.8 M (pH 4.5), 16.6 M (pH 6), and 14.9 M (pH7.5) whereas the single-cell samples had 40.3 M (pH 4.5), 327 M (pH 6) and 465 M (pH 7.5) and 395 M for the mini-metagenomics.

The exploration of the taxonomy assigned to the sequencing reads by Kraken2 showed a large proportion of unassigned reads for metagenomic (between 53 and 73%), mini-metagenomic (89%) and single-cell samples (between 78 and 90%), (Figure 1). The majority of assigned metagenomic reads were grouped in the "Other" phylum category, containing the phylum represented by less than 10,000 reads, but these phyla represented less than 25 % of the reads for the single-cell and mini-metagenomic samples. The second most represented phyla for the metagenomic samples were the Actinomycetota, Pseudomonadota and Acidobacteriota. These phyla were also present in the mini-metagenomic and single-cell samples but in different proportions. The mini-metagenomic and single-cell samples

presented phyla not represented in high proportions in the metagenomic samples: the Bacillota, Bacteroidota, Cyanobacteriota and some Archaea (i.e. Euryarchaeota). These proportions have however to be considered in the light of the disparities in sequencing depths (Figure 1), as well as the following results of this section.

3.2- Assembled genome qualities

After the assembly, manual reads decontamination and quality filtering, a total of 210 SAGs, 42 MAGs and 7 Mini-MAGs were recovered from the sequencing data (Table 1, Supplementary Table S1). Within MAGs, 24 presented similar contigs binning profiles and were considered common to the soil pH 6 and pH 7.5.

Table 1. Number of assembled genomes from single-cell (SAGs), metagenomic (MAGs) and mini-metagenomic (Mini_MAGs) samples. The SAGs were assembled with Spades (V. 3.15.5, Bankevich et al., 2012) and the mini(MAGs) with Anvi'o (Eren et al., 2015). The common MAGs count of MAGs between pH 6 and pH 7.5 (n=24) represents the number of MAGs presented very similar contigs profile based on Anvi'o clustering (i.e. similar coverage and GC content).

	SAG	MAGs		Mini_MAGs
Sample soil pH 4.5	93	2		7
Sample soil pH 6	56	5	24	
Sample soil pH 7.5	61	11		

The measured qualities of SAGs, Mini-MAGs and MAGs are presented in Figure 2. Disparities between the approaches were observed for all quality measurements: the SAGs had highly divergent values of N50 (from 2 to 95.3 Kb) with a median of 7.7 Kb (Figure 2A). The MAGs had a median N50 value of 1.76 Kb which was lower than the Mini-MAGs with a median of 19 Kb. The median length of SAGs (100 Kb) was 30 times lower than for the Mini-MAGs (3,298 Kb) and MAGs (2,577 Kb) (Figure 2B). The putative completeness of the assemblies measured by CheckM was ~ 50% for both MAGs and Mini-MAGs but close to 0 for the SAGs (Figure 2C), estimates to take into account with caution (See Chapter II). The estimated putative contamination was close to 0 % for SAGs, 1.4 % for mini-MAGs and 2.8 % for MAGs (Figure 2D). The number of contigs per assembly was largely superior in MAGs (1,305) compared to Mini-MAGs (440) and SAGs (33) (Figure 2E). It can be emphasized (Table 2) that the number of bacterial ribosomal gene (16S and 23S) hits found in the assemblies of SAGs is higher than in Mini-MAGs and MAGs.

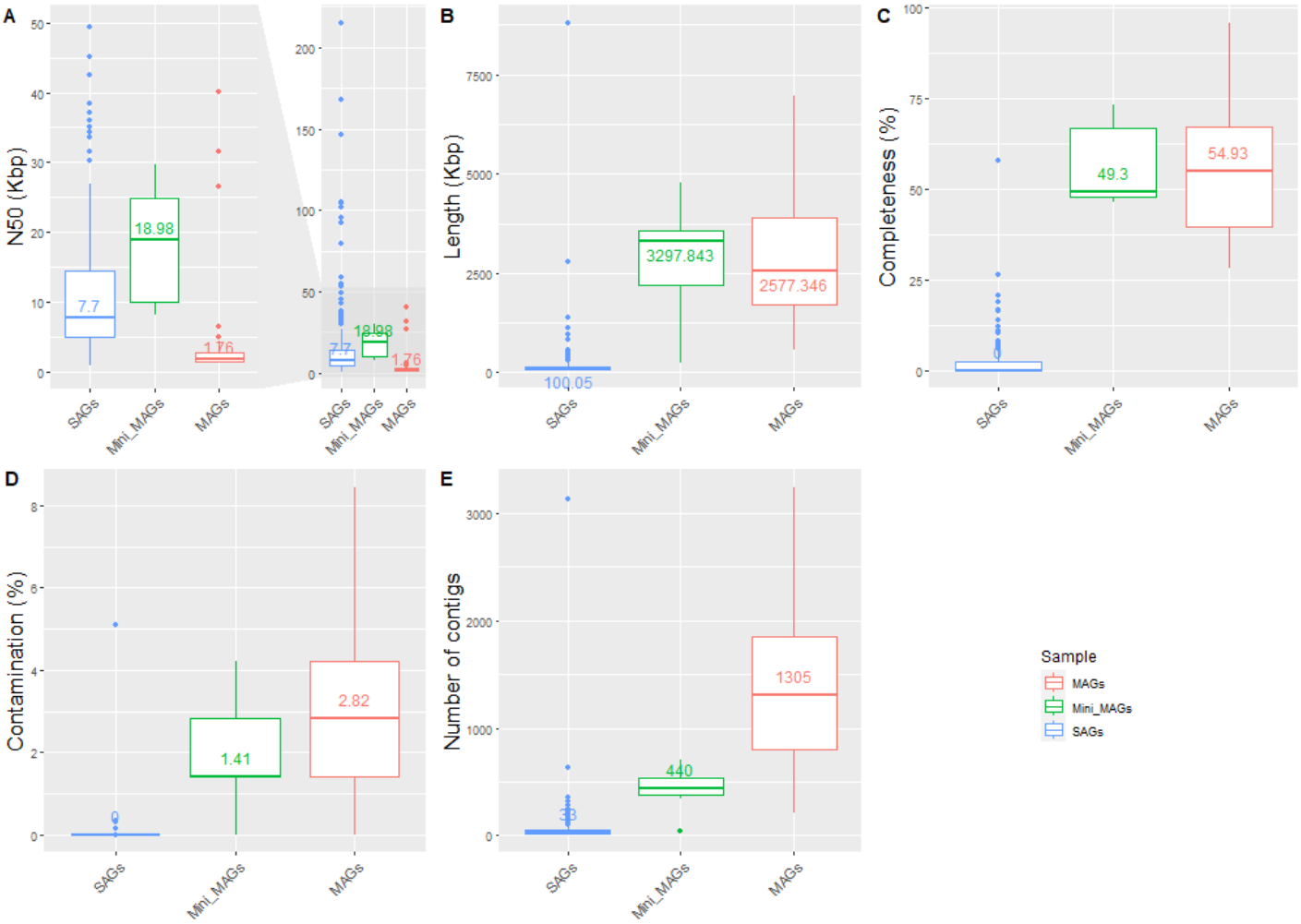


Figure 2. Sequence statistics of the recovered SAGs (blue), Mini-MAGs (green), and MAGs (red). The median is indicated for each boxplot (A) N50 measurements in Kb correspond to the size of the smallest contig needed to cover 50% of the total length of the assembly (B) sizes of assemblies in Kb (C) putative completeness estimates of the assembly by CheckM (D) putative contamination estimates by CheckM and (E) number of contigs for each assembly. Data corresponding to all samples are represented: 210 SAGs, 7 Mini_MAGs and 42 MAGs.

Table 2. Hits for bacterial (5S, 16S, 23S) and eukaryotic (18S, 28s) ribosomal gene in SAGs, Mini-MAGs, MAGs and Co-assembled Genomes from SAGs (CAGs) estimated with Anvi'o.

	SAG	MAGs	Mini_MAGs	CAGs
Ribosomal_RNA_16S	27	3	2	4
Ribosomal_RNA_18S	0	0	0	1
Ribosomal_RNA_23S	42	18	2	7
Ribosomal_RNA_28S	0	0	0	2
Ribosomal_RNA_5S	0	0	0	0

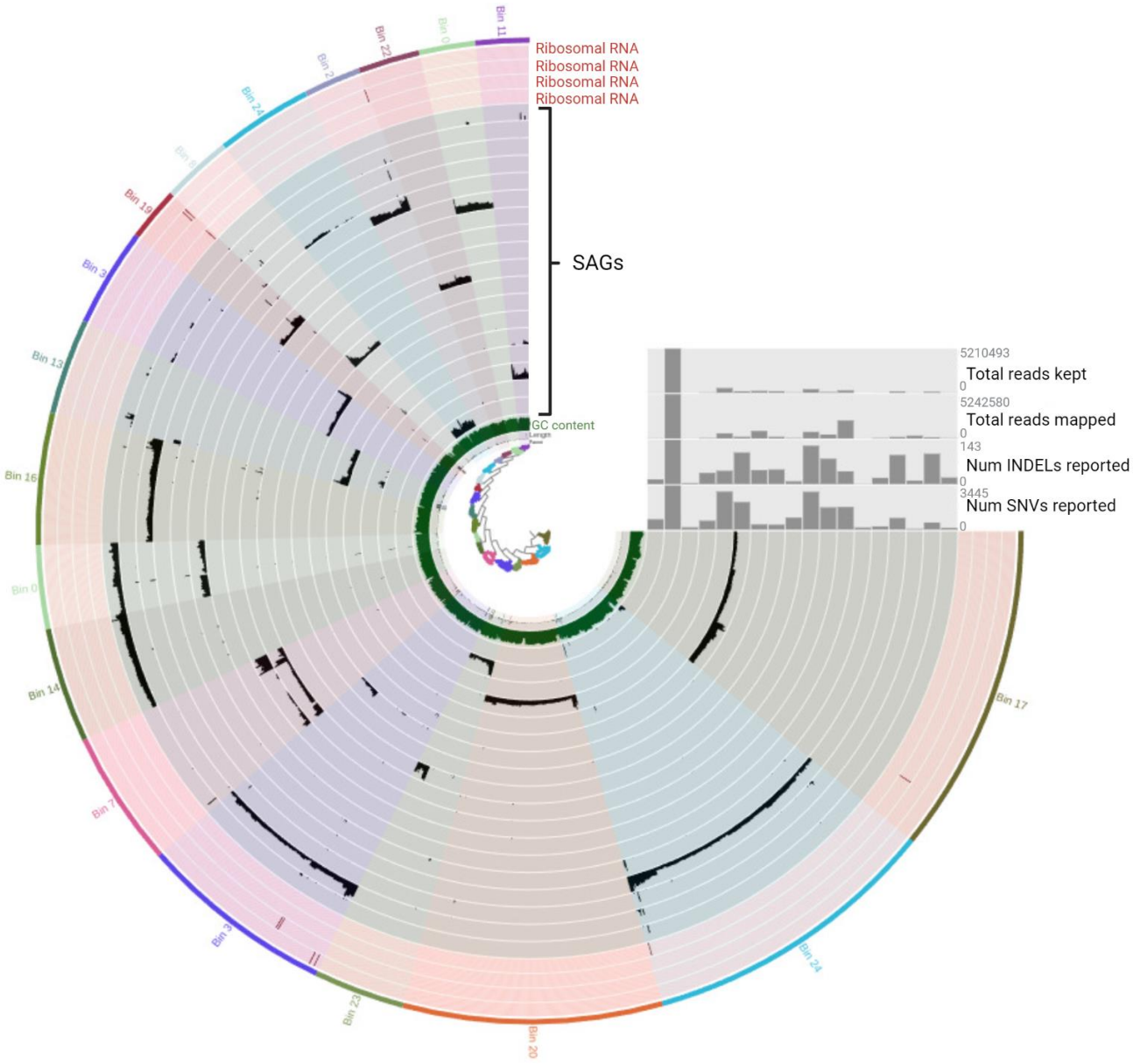


Figure 3. Co-assembly of 18 SAGs identified as *Mycobacterium* genus merged with Anvi'o. The four external layers represent hits for 16s, 18s, 23s and 28s ribosomal RNAs. The bins contain groups of contigs automatically gathered based on similar estimated QC content presented in the internal green layer and coverage shown by black signals in each sample layer. The total reads conserved, the total reads mapped, and the number of indels (i.e. insertion-deletion mutations) and SNV (Single Nucleotide Variants) for each SAGs are shown in the four histograms.

The co-assembly of 18 SAGs identified as *Mycobacterium* genus was done with Anvi'o and resulted in a 5.7 Mb assembly, containing 6958 contigs with an N50 of 2,714 b (Figure 3). The estimated putative completeness and contamination of this assembly were 56 % and 11.3 %. The bins contained collections of contigs with similar QC content and coverage, mainly from

single SAGs. Two hits were identified for 16S rRNA, 1 for 18s rRNA, 5 for 23S rRNA and 2 for 28S rRNA genes (Table 2). Another co-assembly was done on 17 *Methylobacterium* SAGs (data not shown) and resulted in a 5.2 Mb assembly with an estimated putative completeness of 45 % and 10.6 % of contamination. The assembly contained 6,471 contigs, an N50 of 2259 b, 2 hits for 16s rRNA and 2 for 23s rRNA genes. As a comparison, the SAGs with the highest completeness used for these co-assemblies were at 12.1 % for *Methylobactrium* and 13.8 % for *Mycobacterium*.

3.3- Representation of bacterial taxonomy through assembled genomes

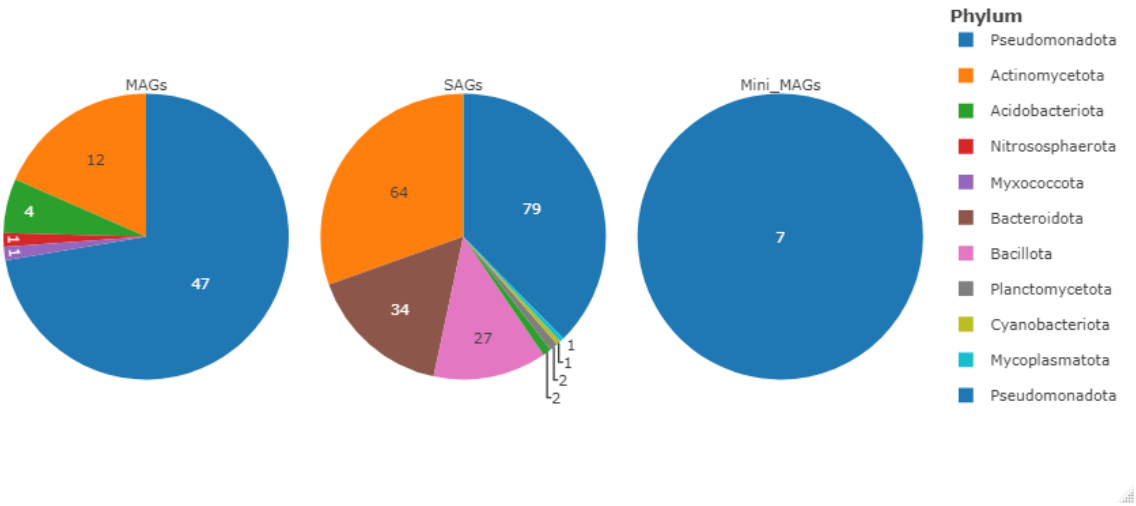


Figure 4. MAGs, SAGs and Mini_MAGs taxonomy at the phylum level, assigned with Mash.

All soil samples confounded, the community compositions represented by the three approaches were different at the phylum level (Figure 4). Only the Pseudomonadota phylum was common to all approaches, but present in different proportions, being the only phylum detected in the mini-MAGs (n=7) and representing, 71 % of the MAGS (n=47) and only 38 % of the SAGs (n=79). Actinomycetota were represented by 64 SAGs and 12 MAGs. Only a few Acidobacteriota assembled genomes were recovered, 2 SAGs and 4 MAGs. Two phyla were present only in MAGs with one representative each; the Nitrosphaerota and the Myxococcota. The phyla represented exclusively in SAGs were, by decreasing order: Bacteroidota, Bacillota, Planctomycetota, Cyanobacteria, and Mycoplasmatota.

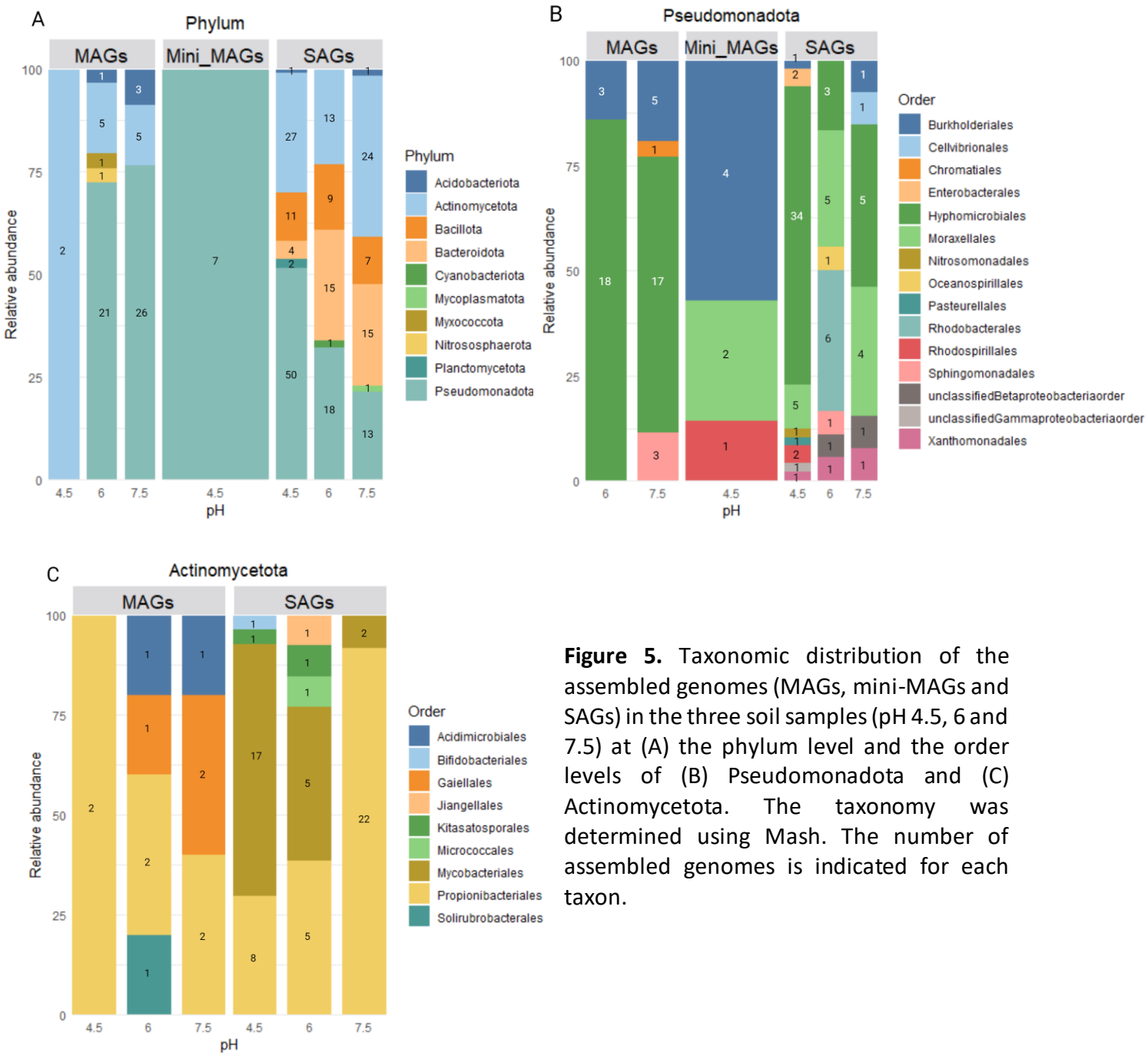


Figure 5. Taxonomic distribution of the assembled genomes (MAGs, mini-MAGs and SAGs) in the three soil samples (pH 4.5, 6 and 7.5) at (A) the phylum level and the order levels of (B) Pseudomonadota and (C) Actinomycetota. The taxonomy was determined using Mash. The number of assembled genomes is indicated for each taxon.

Table 3. Taxonomic richness measured per assembly method (SAGs, MAGs, and Mini-MAGs) and per soil sample at the phylum, class and order levels.

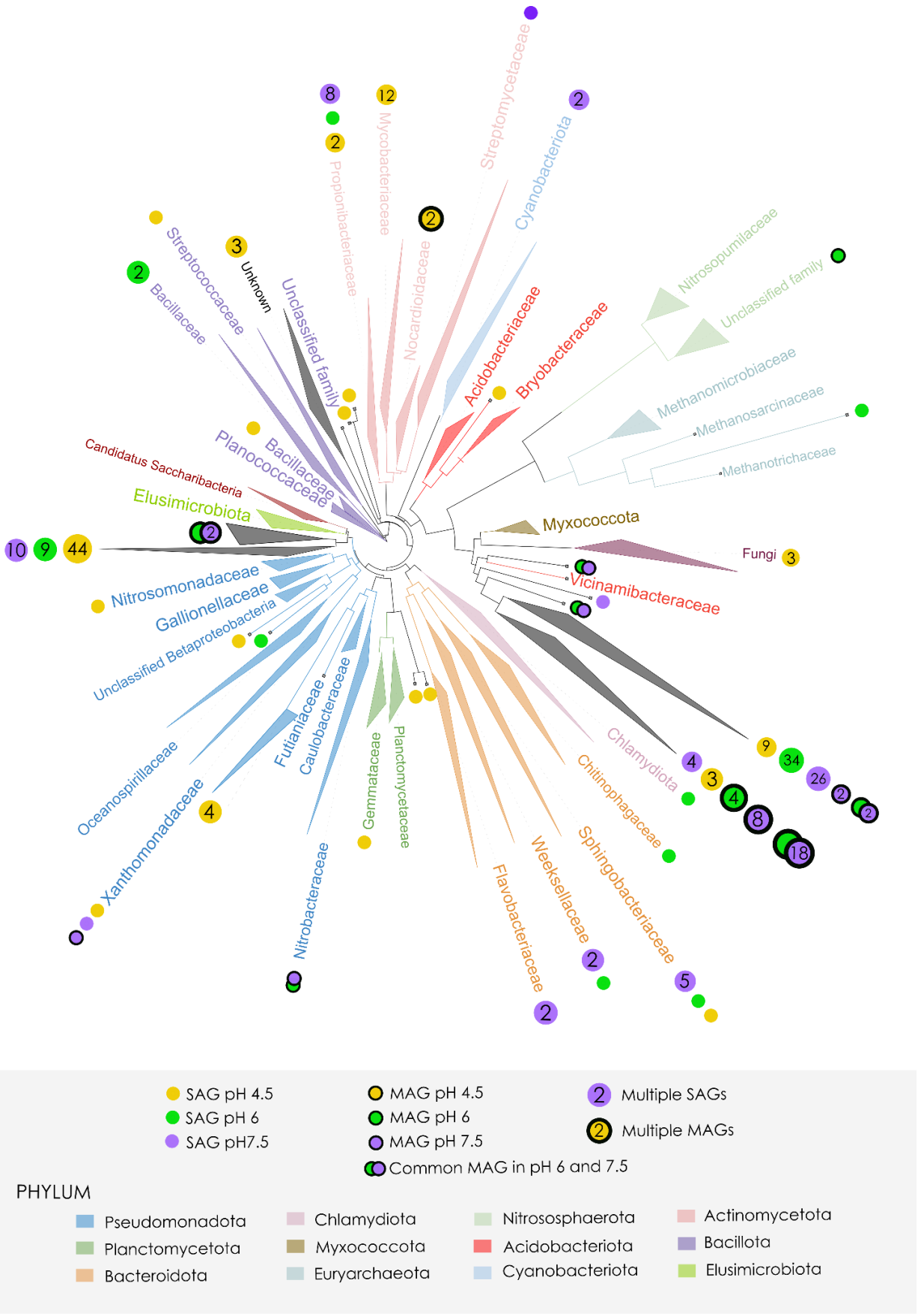
	Phylum richness			Class richness			Order richness		
	SAGs	MAGs	Mini_MAGs	SAGs	MAGs	Mini_MAGs	SAGs	MAGs	Mini_MAGs
Sample soil pH 4.5	6	1	1	10	1	3	20	1	3
Sample soil pH 6	5	5	-	11	9	-	19	9	-
Sample soil pH 7.5	6	3	-	12	6	-	15	8	-
Total richness	8	5	1	17	10	3	33	13	3

For all soil samples, the taxonomic richness represented by the SAGs was higher than the one from the MAGS at the phylum, class and order levels (Table 3). The Pseudomonadota was the main common phylum between SAGs, MAGs, and Mini-MAGs but presented different taxonomic compositions and proportions at the order level (Figure 5B). These differences were observable for each soil sample: all samples in SAGs presented higher or equal richness than those in MAGs or Mini-MAGs (Table 3) and the taxonomic composition of each sample differed depending on the assembly method (Figure 5). The same observations were made for the Actinomycetota orders between SAGs and MAGs (Figure 5C).

3.4- Phylogeny

A first version of the phylogenetic tree positioning the SAGs and MAGs into the Tree of Life is proposed in Figure 6. The major outcome of this phylogenetic tree was the clustering of many SAGs and MAGs, not systematically affiliated to identical referenced families (clades in grey, Figure 6). Additionally, the SAGs and MAGs from these clades were segregated: most of the SAGs were apart from the MAGs and SAGs from pH 4.5 samples were grouped away from SAGs from pH 6 and 7.5 samples (Figure 7). Surprisingly, 3 SAGs from soil pH 4.5 were grouped within fungi taxa. This area of the Tree contained Fungi, Archaea, some bacteria, and many unclassified samples. A few samples (7 SAGs and 1 MAG) were however individually positioned next to bacterial and archaeal families, known or under-described (i.e. named “Unclassified family”), in the Bacillota, Nitrososphaerota, Acidobacteriota, Planctomycetota, and Pseudomonadoa phylum.

Figure 6. Phylogenetic tree with position of SAGs and MAGs assemblies. The positioning of the samples was made by calculating their mash distances with 17,556 reference genomes. The collection of reference genomes and samples was positioned into the Tree using Mashtree and Mash distances (i.e. estimates of mutation rate from *k*-mer count and Poisson model of mutation) (V. 1.2.0, Katz et al., 2019). The phylogenetic clustering was performed by the Neighbor-Joining (NJ) algorithm. The colours on the nodes represent the Phylum. The abundance of SAGs and MAGs is indicated at the tips of the clades. To root the tree, the Bacillota phylum was used.


Streptomycetaceae
Cyanobacteriota
Nitrosopumilaceae
Unclassified family
Methanomicrobiaceae
Methanosarcinaceae
Methanotrichaceae
Bryobacteraceae
Acidobacteriaceae
Nocardioidaceae
Mycobacteriaceae
Propionibacteriaceae
Streptococcaceae
Bacillaceae
Unknown
Unclassified family
Bacillaceae
Planococcaceae
Candidatus Saccharibacteria
Elusimicrobiota
Myxococcota
Fungi
Vicinamibacteraceae
Nitrosomonadaceae
Gallionellaceae
Unclassified Betaproteobacteria
Oceanospirillaceae
Xanthomonadaceae
Futianiaceae
Caulobacteraceae
Gemmataceae
Planctomycetaceae
Nitrobacteraceae
Chlamydiota
Chitinophagaceae
Sphingobacteriaceae
Weeksellaceae
Flavobacteriaceae
SAG pH 4.5
SAG pH 6
SAG pH7.5
MAG pH 4.5
MAG pH 6
MAG pH 7.5
Common MAG in pH 6 and 7.5
Multiple SAGs
Multiple MAGs
PHYLUM
Pseudomonadota
Planctomycetota
Bacteroidota
Chlamydiota
Myxococcota
Euryarchaeota
Nitrososphaerota
Acidobacteriota
Cyanobacteriota
Actinomycetota
Bacillota
Elusimicrobiota

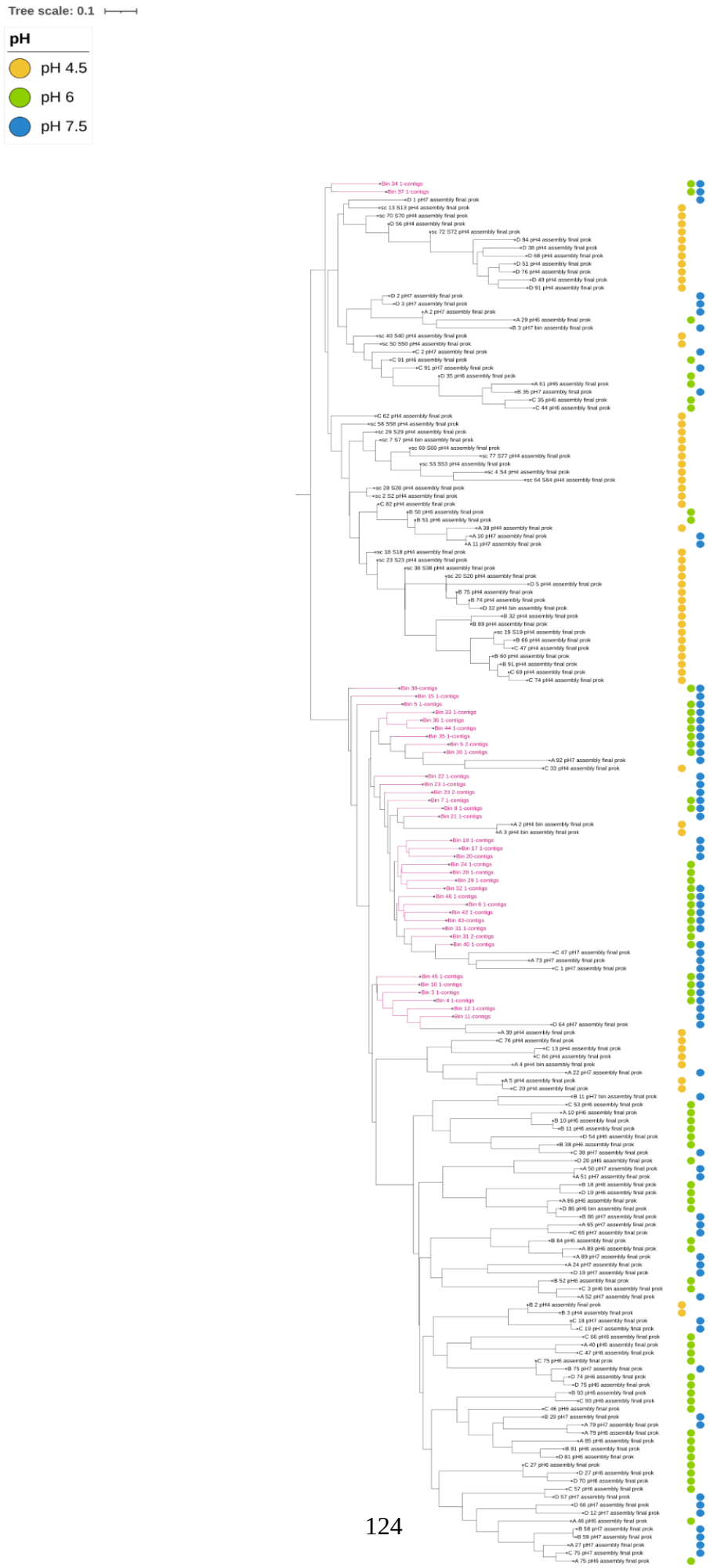

Tree scale: 0.1
pH
pH 4.5
pH 6
pH 7.5

Figure 7. Branches extracted from the phylogenetic tree (Figure 6) from the three main grey clades containing most of the SAGs and MAGs assemblies recovered from soil samples pH 4.5 (yellow), pH 6 (green) and pH 7.5 (blue). The purple nodes represent the MAGs, and the black ones represent the SAGs.

IV- Discussion

We have evaluated the potential of single-cell, mini-metagenomics and metagenomics-assembled genomes to provide clean and complete genomes from soil samples and discuss their relevance in ecological applications involving microbes. We chose soil samples from a well-studied experimental field with different pH treatments, not aiming for an ecological conclusion, but to evaluate the potential of these approaches to specifically process such samples and to imagine potential further applications.

Metagenomics is traditionally used to study environmental microbes from DNA extracted from the matrix. Here, we first extracted the cells from the soil samples before DNA extraction to limit the presence of external DNA and start the sample preparation from the same material as used for single-cell and mini-metagenomics. Therefore, except for the cell isolation, genome amplification and lower reaction volumes, single-cell and mini-metagenomic samples received the same library preparation as metagenomic samples. Prior to the putative genome assemblies, the taxonomy of the sequencing reads was examined to observe from which material the genomes would be recovered. We have detected with the Kraken2 tool a vast majority of unassigned reads (Figure 1), up to 75 % in metagenomic and up to 90% in single-cell and mini-metagenomic samples. This can be explained by i) the lack of referenced sequences in the databases (Z. Zhang et al., 2020), ii) the incorporation of residual reads during sample preparation with molecular reactions and iii) the remaining environmental DNA in the samples (López-Escardó et al., 2017). To avoid deleting relevant prokaryote information if not conserved, these reads were kept for putative genome assemblies. The following results have to be considered carefully, as the total sequencing reads for single-cell and mini-metagenomics samples were higher than metagenomics samples (i.e. 832 M and 395 M total reads for single-cell and mini-metagenomic samples versus 32 M for metagenomics). Therefore, we cannot totally exclude that the differences observed are not coming from this difference in sequencing depth. The majority (up to 80%)

of assigned metagenomic reads were grouped in the “Other” phylum category, containing the phylum represented by less than 10,000 total reads per sample (Figure 1). This category represented a maximum of 25% of the assigned reads in the single-cell and mini-metagenome samples. This shows that metagenomics might capture information about many taxa but in smaller quantities than cell-centred approaches. In these latter, the reads were distributed in more major phyla, with therefore more probability to see their genome reconstituted. This assumption was verified with the taxonomy assignation of assembled genomes for each approach, as we recovered 42 MAGs from millions of cells as a starting material, 210 SAGs from 1140 cells and 7 mini-MAGs from 48 times 10 cells with similar taxonomy proportions than the assigned reads taxonomy (Figure 4, Figure 1). Like the sequence reads taxonomy, the assembled genome taxonomy presented more richness in SAGs compared to MAGs at different taxonomic levels (Table 3). This higher richness in SAGs was true for all the samples (Figure 5). The soil sample taxonomy composition was different from one approach to another, showing that the microbial community composition interpretation might vary with our technical approaches.

The information recovered from these assemblies have to be put into the perspective of their quality estimations. We obtained more SAGs than MAGs and mini-MAGs, but are their exploitability equivalent? The assemblies from single-cell samples were smaller than mini-MAGs and MAGs (Figure 2B), as it is common (Alneberg et al., 2018). However, the bigger size of MAGs is not a guarantee of quality and purity and is likely chimeric (Alneberg et al., 2018, see Chapter II). The size of Mini-MAGs and MAGs were equivalent, but not the other measurements: the N50 value was superior for Mini-MAGs testifying of their longer contigs which were fewer in Mini-MAGs than MAGs (Figure 5). This suggests that MAGs are largely fragmented with small contigs compared to a cell-centered approach. Regarding the SAGs, the number of contigs was lower than those obtained with the two other approaches, which is coherent with their smaller size, but the N50 value was higher than MAGs as well. We put into question the legitimacy of CheckM in Chapter II to assess assembly qualities and expose its positive correlation with assembly size. Here, the patterns between assembly size and measured completeness were identical: high values for MAGs and mini-MAGs and low values for SAGs for both metrics (Figure 5), despite the a priori more fragmented composition of MAGs with very few ribosomal genes detected (Table 2). The mini-metagenomic approach

seems to allow a better assembly quality than MAGs and SAGs while processing very few cells. However, just like metagenomics, multiple cells are analysed per sample and there is no certainty that the recovered assemblies are non-chimeric. These SAGs, MAGs and Mini-MAGs were manually decontaminated only. We aim to compare these results with the automatic decontamination pipeline developed in Chapter II in order to identify contaminant taxa possibly not removed here for an improvement of the measured quality metrics. To improve the evaluation of the exploitability of the assemblies and for metabolic pathways visualisation we also plan to examine annotated genomes.

We have tested the co-assembly (CAGs) of single-cell samples with identical genus assignment (i.e. *Mycobacterium* or *Methylobacterium*) to construct larger putative genomes from purer assemblies than mini-MAGs. The CAGs lengths were of 5.6 Mb and 5.2 Mb, which are in the range of these genus genome sizes (Leducq et al., 2022; Wee et al., 2017). The completeness of the CAGs was similar to the (Mini-)MAGs, but with higher measured contamination and 2 hits for the 16S rRNA gene. This suggests a potential redundancy in the sequences used for the CAGs and would explain that the measured completeness does not reach 100 % despite the long length of the assemblies. Moreover, the presence of 18S and 28S rRNA genes suggested a potential eukaryote contamination (Table 2). The content in bacteria ribosomal genes in Mini-MAGs was inferior to the CAGs and SAGs (Table 2), suggesting that CAGs allow the regrouping of the SAGs information. They indeed have been shown to improve the protein recovery of SAGs (Mangot et al., 2017). These co-assembly results are very promising for improving assemblies of low-quality single-amplified genomes which are known to be partial and require multiple SAGs from the identical strain to reach higher genome coverage (Zheng et al., 2022). Higher quality SAGs or co-assembled genomes can be of use to evaluate the purity of MAGs and improve the contigs binning (Arikawa et al., 2021). However, these CAGs do not allow us to draw ecological conclusions on individual bacterial entities like SAGs, as they were constructed from different species, and might not be as straightforward as we expected.

The phylogenetic tree construction (Figure 6) purposes were to (i) position the SAGs and MAGs in the tree of life, (ii) detect the potential of these approaches to point out under-described bacteria and (iii) evaluate if the MAGs and SAGs quality was sufficient to be

exploited in phylogeny. The tree reconstituted with the incorporation of our samples presented part of the SAGs and MAGs with no logical taxonomic clustering measured by Mash. The phylogenetic clustering was made with mash distance calculations on marker genes, therefore the taxonomy measured and the phylogenetic results should correspond. This likely reflects the unreferenced content of our assembled genomes as previously highlighted (Nishikawa et al., 2022) and/or the low quality of the assemblages. The SAGs and MAGs were not phylogenetically mixed and were rather organized in groups (Figure 7) which supports previous findings showing that the representation of taxa is preferably done by SAGs or MAGs (Woyke et al., 2017). Despite some samples being positioned in under-described bacterial families, some were placed within fungi (Figure 6). These results potentially highlight the remaining contamination present in our samples, which would require to be automatically decontaminated with the SINCERE data protocol explained in Chapter II. The fungi clade reference genomes might have been problematic for the tree reconstruction, and will be re-calculated without any eukaryote reference to observe whether the sample's clustering was due to their unknown sequence compositions or to calculation issues of the Mash distances.

Overall, cell-centred approaches seem to present purer assemblies than MAGs. In this study, it can however be emphasized that the sizes of the SAGs were not as high as expected. Similarly, as in Chapter 2, the single-cell reads contained a diversity of sequenced DNA within each sample. This was interpreted as free environmental DNA present within the cell suspension as well as incorporated contamination during our manipulations, which necessarily limited the number of sequences allocated to the targeted DNA and reduced the assembly lengths. This environmental DNA detected in SAGs is delicate to measure and remove from metagenomic data which by definition re-create putative genomes from various sequence origins, including external cellular DNA. Therefore, this might lead to misinterpretations of diversity and functions. The fact that the metagenomic assemblies were very fragmented in this study could also be explained by the presence of this environmental DNA.

V- Conclusion

The bacterial community was taxonomically richer in the lens of SAGs. With technical improvements, single-cell genomics is suited for assessing ecological questions under some conditions: i) improve the recovered genome length by optimising the genome amplification strategy (Stepanauskas et al., 2017), ii) enhance gene databases and genome completeness for easier annotation to evolve from a descriptive to a deductive demarch iii) combine genomic with transcriptomic data to extract the expressed functions of each cell and iv) use metabolic modelling to comprehend potential interaction, niches, and dependencies within bacteria populations and communities. This, in the near future, will be accessible with more standardised and universal single-cell omics approaches and might become the preferred genomic approach of microbiologists.

VI- Supplementary

Table S1. List of all SAGs, MAGs and Mini-MAGs recovered from the soil samples with corresponding measured quality metrics and taxonomic identification.

Method	Sample	N50 (Kbp)	Number of contigs	Size (Kbp)	Completeness (%)	Contamination (%)	Phylum	Class	Order	Family
MAG	Bin_1_1	1951	844	1624960	29.58	1.41	Actinomycetota	Actinomycetes	Propionibacteriales	Nocardioidaceae
MAG	Bin_10_1	1737	1790	3090179	64.79	4.23	Actinomycetota	Actinomycetes	Propionibacteriales	Nocardioidaceae
MAG	Bin_11	1463	1359	2030792	28.17	4.23	Pseudomonadota	Betaproteobacteria	Burkholderiales	Sphaerotilaceae
MAG	Bin_12_1	1417	1784	2647979	54.93	1.41	na	na	na	na
MAG	Bin_14	1271	264	352237	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Legionellales	Legionellaceae
MAG	Bin_15_1	1199	1500	1876381	28.17	4.23	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Hyphomicrobiaceae
MAG	Bin_16	1140	259	303155	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Hyphomicrobiaceae
MAG	Bin_17_1	1431	2249	3339986	46.48	1.41	Pseudomonadota	Betaproteobacteria	Burkholderiales	Sphaerotilaceae
MAG	Bin_18_1	2791	2584	6188383	53.52	4.23	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	Vicinamibacteraceae
MAG	Bin_19_1	1489	485	736889	54.93	0.00	Pseudomonadota	Gammaproteobacteria	Chromatiales	Woeseiaceae
MAG	Bin_2	40143	206	2900965	85.92	4.23	Actinomycetota	Actinomycetes	Propionibacteriales	Nocardioidaceae
MAG	Bin_20	1935	2091	3910882	35.21	2.82	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	Vicinamibacteraceae
MAG	Bin_21_1	2395	1482	3267689	77.46	5.63	Pseudomonadota	Alphaproteobacteria	Sphingomonadales	Sphingosinicellaceae
MAG	Bin_22_1	6386	578	2362934	92.96	1.41	Pseudomonadota	Alphaproteobacteria	Sphingomonadales	Sphingosinicellaceae
MAG	Bin_23_1	4450	1749	5755226	59.15	2.82	Actinomycetota	Rubrobacteria	Gaiellales	Gaiellaceae
MAG	Bin_23_2	2961	926	2378078	40.85	1.41	Pseudomonadota	Alphaproteobacteria	Sphingomonadales	Sphingosinicellaceae
MAG	Bin_24_1	1203	3092	3890366	49.30	0.00	Actinomycetota	Thermoleophilia	Solirubrobacterales	Capillimicrobiaceae
MAG	Bin_26_1	4975	346	1190296	84.21	1.32	Nitrososphaerota	unclassified Nitrososphaerota class	unclassified Nitrososphaerota order	unclassified Nitrososphaerota family
MAG	Bin_27_1	6375	995	4676799	0.00	0.00	Gemmatimonadota	Gemmatimonadetes	Gemmatimonadales	Gemmatimonadaceae
MAG	Bin_28_1	1739	1476	2571752	29.58	4.23	Myxococcota	Myxococcia	Myxococcales	Anaeromyxobacteraceae
MAG	Bin_29_1	1428	2702	4003186	61.97	4.23	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Rhodovibrionaceae
MAG	Bin_3_1	1585	3125	5021809	70.42	4.23	Pseudomonadota	Betaproteobacteria	Burkholderiales	Sphaerotilaceae
MAG	Bin_30_1	1300	2777	3762770	61.97	1.41	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Reyranellaceae
MAG	Bin_31_1	1731	2939	5066069	57.75	5.63	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Xanthobacteraceae
MAG	Bin_31_2	1318	1759	2418448	29.58	2.82	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Reyranellaceae
MAG	Bin_32_1	2290	3246	6968985	64.79	4.23	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Reyranellaceae
MAG	Bin_33_1	26581	477	5128680	87.32	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Hyphomicrobiaceae
MAG	Bin_33_2	1670	87	153373	0.00	0.00	Pseudomonadota	Betaproteobacteria	unclassified Betaproteobacteria order	unclassified Betaproteobacteria family
MAG	Bin_34_1	1724	419	711253	29.58	1.41	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Nitrobacteraceae
MAG	Bin_35_1	31637	416	2988539	95.77	2.82	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Hyphomicrobiaceae
MAG	Bin_36_1	3905	1873	5783662	66.20	2.82	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Hyphomicrobiaceae
MAG	Bin_37_1	1657	492	817757	39.44	2.82	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Phyllobacteriaceae
MAG	Bin_38	1864	794	1498231	35.21	0.00	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	Vicinamibacteraceae
MAG	Bin_39_1	2226	1071	2175637	71.83	4.23	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Xanthobacteraceae
MAG	Bin_4_1	1551	1251	1950009	63.38	5.63	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Phyllobacteriaceae
MAG	Bin_4_2	1226	283	362062	0.00	0.00	Actinomycetota	Actinomycetes	Kitasatosporales	Streptomycetaceae
MAG	Bin_40_1	1951	580	1131445	36.62	2.82	Actinomycetota	Acidimicrobiia	Acidimicrobiales	Ilumatobacteraceae
MAG	Bin_41	1573	438	700915	0.00	0.00	Nitrososphaerota	Nitrososphaeria	Nitrososphaerales	Nitrososphaeraceae
MAG	Bin_42_1	1348	939	1306728	36.62	1.41	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Reyranellaceae

ple	N50 (Kbp)	Number of contigs	Size (Kbp)	Completeness (%)	Contamination (%)	Phylum	Class	Order	Family	Genus	Species
	1992	1075	2079603	39.44	2.82	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Reyranellaceae	Reyranella	Reyranella soli
1	3061	1555	4090367	52.11	1.41	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Xanthobacteraceae	Pseudolabrys	Pseudolabrys taiwanensis
1	1581	1653	2673361	70.42	5.63	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Reyranellaceae	Reyranella	Reyranella soli
1	1945	2660	5055726	67.61	5.63	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Reyranellaceae	Reyranella	Reyranella soli
	1662	1174	1980832	61.97	1.41	Pseudomonadota	Betaproteobacteria	Burkholderiales	Sphaerotilaceae	Azohydromonas	Azohydromonas sediminis
	1503	1077	1649313	45.07	8.45	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Phyllobacteriaceae	Mesorhizobium	Mesorhizobium ephedrae
	1780	314	564528	40.85	5.63	Actinomycetota	Actinomycetes	Propionibacteriales	Nocardioidaceae	Nocardioides	Nocardioides halotolerans
	2946	1009	2582940	78.87	1.41	Actinomycetota	Rubrobacteria	Gaiellales	Gaiellaceae	Gaiella	Gaiella occulta
	1513	1059	1645470	39.44	2.82	Pseudomonadota	Betaproteobacteria	Burkholderiales	Sphaerotilaceae	Rubrivivax	Rubrivivax benzoatilyticus
6	25.8	17	133.7	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium soli
7	12.8	42	136.3	8.33	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae	Cutibacterium	Cutibacterium acnes
7	8.6	52	179.1	4.16	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae	Cutibacterium	Cutibacterium acnes
7	6.7	38	159	3.10	0.00	Bacteroidota	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	Mucilaginibacter	Mucilaginibacter auburnensis
7	49.3	9	76.5	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Nitrobacteraceae	Bradyrhizobium	Bradyrhizobium viridifuturi
4	4,8	88	222,6	13,79	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae	Mycobacterium	Mycobacterium heidelbergense
_bin	8,6	42	145,3	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Rhodopila	Rhodopila globiformis
	4.4	54	121.8	7.76	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae	Cutibacterium	Cutibacterium acnes
7	16.8	11	70	1.75	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae	Cutibacterium	Cutibacterium acnes
7	101.9	17	146.7	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae	Cutibacterium	Cutibacterium acnes
4	9,8	29	66,1	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae	Mycobacterium	Mycobacterium interjectum
7	5.2	40	110.3	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae	Cutibacterium	Cutibacterium acnes
6	7.1	19	58.6	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Lysobacter	Lysobacter oculi
_bin	13,5	139	373,5	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Rhodopila	Rhodopila globiformis
	22.1	31	136.4	6.90	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae	Cutibacterium	Cutibacterium acnes
4	5,8	25	75,6	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae	Chryseobacterium	Chryseobacterium indoltheticum
4_bin	102,1	13	115,9	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae	Mycobacterium	Mycobacterium lacus
4	2,7	41	79,1	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae	Methylobacterium	Methylobacterium tardum
4	146,4	40	220,8	6,9	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae	Cutibacterium	Cutibacterium acnes
_bin	36	31	162,8	5,33	0.00	Acidobacteriota	Terriglobia	Terriglobales	Acidobacteriaceae	Edaphobacter	Edaphobacter acidisoli
6	7.6	18	77.4	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	Candidatus Carsonella	Candidatus Carsonella ruddii
7	8.2	17	61.2	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae	Cutibacterium	Cutibacterium acnes
6	6.8	17	66.8	0.00	0.00	Actinomycetota	Actinomycetes	Jiangellales	Jiangellaceae	Phytoactinopolyspora	Phytoactinopolyspora alkaliphila
	104,2	50	399,8	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae	Methylobacterium	Methylobacterium radiotolerans
7	9.5	15	66.4	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae	Acinetobacter	Acinetobacter junii
7	3.9	36	77.7	0.86	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae	Cutibacterium	Cutibacterium acnes
7	38.5	8	70.9	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae	Epilithonimonas	Epilithonimonas vandammei
7	37.2	14	72	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Cellvibrionales	Halieaceae	Chromatocurvus	Chromatocurvus halotolerans
6	7.6	16	79.8	0.00	0.00	Cyanobacteriota	Cyanophyceae	Oscillatoriales	Microcoleaceae	Microcoleus	Microcoleus asticus
7	8.1	16	54.9	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae	Epilithonimonas	Epilithonimonas vandammei

Method	Sample	N50 (Kbp)	Number of contigs	Size (Kbp)	Completeness (%)	Contamination (%)	Phylum	Class	Order	Family
SAG	A_71_pH4	2,7	52	87,4	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	A_72_pH6	54.5	10	78.1	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Nitrobacteraceae
SAG	A_73_pH7	45.1	33	260.5	0.00	0.00	Acidobacteriota	Vicinamibacteria	Vicinamibacterales	Vicinamibacteraceae
SAG	A_75_pH6	3	26	57.7	2.08	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	A_77_pH4	5,3	3130	8807,7	57,65	5,11	Planctomycetota	Planctomycetia	Gemmatales	Gemmataceae
SAG	A_79_pH6	10.2	7	54	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Corynebacteriaceae
SAG	A_79_pH7	7.2	16	59.7	0.00	0.00	Bacillota	Clostridia	Eubacteriales	Desulfotomaculaceae
SAG	A_84_pH6	9.3	13	60.1	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Paracoccaceae
SAG	A_85_pH4	3,2	28	50,9	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	A_85_pH6	20.6	15	117.9	0.00	0.00	Bacteroidota	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae
SAG	A_86_pH6	5.1	26	83.7	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Nitrobacteraceae
SAG	A_89_pH6	15.2	16	76.6	0.00	0.00	Bacillota	Bacilli	Bacillales	Staphylococcaceae
SAG	A_89_pH7	5.1	39	102.2	0.00	0.00	Bacillota	Bacilli	Bacillales	Staphylococcaceae
SAG	A_90_pH6	19.9	27	120.5	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae
SAG	A_91_pH6_bin	4.4	40	82.1	0.00	0.00	Pseudomonadota	Betaproteobacteria	unclassified Betaproteobacteria order	unclassified Betaproteobacteria family
SAG	A_92_pH7	21.6	17	99.4	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Nitrobacteraceae
SAG	B_10_pH4_bin	5,4	168	369	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	B_10_pH6	11.6	31	152	0.00	0.00	Bacillota	Tissierellia	Tissierellales	Peptoniphilaceae
SAG	B_11_pH4	168,3	78	528,9	4,18	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	B_11_pH6	13	19	102.1	0.00	0.00	Bacillota	Tissierellia	Tissierellales	Peptoniphilaceae
SAG	B_11_pH7_bin	3.4	30	65.2	4.17	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	B_15_pH7	5.1	25	70.8	0.00	0.00	Bacteroidota	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae
SAG	B_18_pH6	14.5	7	54.9	4.17	0.00	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Roseobacteraceae
SAG	B_19_pH4	6	20	50,2	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	B_2_pH4	7,9	63	151	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	B_21_pH6	5	13	50	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	B_22_pH4	14,6	13	52,9	4,17	0.00	Bacillota	Bacilli	Lactobacillales	Streptococcaceae
SAG	B_25_pH4	53,6	73	497,4	6,67	0,33	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	B_28_pH6	3.3	24	67.1	0.00	0.00	Bacteroidota	Chitinophagia	Chitinophagales	Chitinophagaceae
SAG	B_29_pH7	26.8	11	52.8	4.17	0.00	Bacillota	Tissierellia	Tissierellales	Peptoniphilaceae
SAG	B_3_pH4	8,1	68	299,4	4,17	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	B_3_pH7_bin	5.5	17	62.8	4.17	0.00	Bacillota	Tissierellia	Tissierellales	Peptoniphilaceae
SAG	B_32_pH4	0,9	71	60,4	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	B_35_pH7	5.4	41	117.7	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	B_38_pH6	7	29	93.4	0.00	0.00	Bacillota	Tissierellia	Tissierellales	Peptoniphilaceae
SAG	B_40_pH4	2,2	92	149,6	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	B_42_pH7	3.2	32	64.9	1.85	0.34	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	B_43_pH7	4.9	18	55.3	1.85	0.34	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	B_5_pH4	17,8	50	295,5	0.00	0.00	Planctomycetota	Planctomycetia	Gemmatales	Gemmataceae

Method	Sample	N50 (Kbp)	Number of contigs	Size (Kbp)	Completeness (%)	Contamination (%)	Phylum	Class	Order	Family
SAG	B_50_pH6	6.3	47	105.4	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	B_51_pH6	6.3	54	167.6	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	B_52_pH6	3.1	31	71.9	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Paracoccaceae
SAG	B_58_pH7	6.2	30	94.4	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Corynebacteriaceae
SAG	B_59_pH7	5.9	33	95.7	0.00	0.00	Bacillota	Tissierellia	Tissierellales	Peptoniphilaceae
SAG	B_60_pH4	24,9	63	184	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	B_66_pH4	1,2	64	77,4	2,39	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	B_7_pH7	4.5	18	51.7	1.72	0.00	Bacteroidota	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae
SAG	B_74_pH4	6,5	123	226,7	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	B_75_pH4	8,1	145	300,7	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	B_75_pH7	5.7	33	102.6	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	B_81_pH6	4.5	41	118.7	2.66	0.00	Actinomycetota	Actinomycetes	Kitasatosporales	Streptomycetaceae
SAG	B_83_pH7	6.3	34	108.3	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	B_84_pH6	2.9	32	64.9	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Paracoccaceae
SAG	B_86_pH7	5	18	53.6	0.00	0.00	Bacteroidota	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae
SAG	B_89_pH4	3,7	157	281,1	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	B_90_pH6_bin	16.3	20	90.5	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	B_91_pH4	14,4	31	71,3	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	B_91_pH6	8	38	127.6	4.17	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	B_93_pH6	12	19	69	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Corynebacteriaceae
SAG	C_1_pH7	58.5	27	102.1	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Nitrobacteraceae
SAG	C_13_pH4	12,6	25	90,8	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	C_14_pH7	9.2	17	75	0.00	0.00	Bacteroidota	Cytophagia	Cytophagales	Hymenobacteraceae
SAG	C_15_pH7	7.6	16	89.2	4.17	0.00	Bacteroidota	Cytophagia	Cytophagales	Hymenobacteraceae
SAG	C_18_pH7	8	12	64.7	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	C_19_pH7	8.7	14	73.4	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	C_2_pH7	14.8	19	121.2	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	C_20_pH4	12	82	277	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	C_27_pH6	4.7	36	93.2	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Paracoccaceae
SAG	C_3_pH6_bin	26.9	8	98.7	0.00	0.00	Bacteroidota	Cytophagia	Cytophagales	Hymenobacteraceae
SAG	C_33_pH4	2,7	38	57	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	C_35_pH6	4.2	36	91.1	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	C_38_pH7	3.3	25	52.2	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Nitrobacteraceae
SAG	C_44_pH6	16.1	13	93.4	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	C_46_pH6	30.2	20	109	0.00	0.00	Bacillota	Bacilli	Bacillales	Bacillaceae
SAG	C_47_pH4	11,6	32	75,6	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	C_47_pH6	4.8	22	84.6	0.00	0.00	Bacteroidota	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae
SAG	C_47_pH7	5	31	85.9	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Reyranellaceae
SAG	C_5_pH4_bin	12,8	39	310,7	16,45	0.00	Pseudomonadota	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae

Method	Sample	N50 (Kbp)	Number of contigs	Size (Kbp)	Completeness (%)	Contamination (%)	Phylum	Class	Order	Family
SAG	C_50_pH4_bin	23,2	46	202	10,34	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	C_50_pH7_bin	7.3	35	89.6	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	C_51_pH4_bin	5	52	93,8	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	C_51_pH7_bin	5	46	111.4	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	C_53_pH4	34,4	29	204,7	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Pasteurellales	Pasteurellaceae
SAG	C_53_pH6	2.9	23	51.9	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Corynebacteriaceae
SAG	C_57_pH6	7.5	23	55	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Rhodobacterales	Paracoccaceae
SAG	C_62_pH4	215,2	23	264,2	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Nitrobacteraceae
SAG	C_65_pH7	10.9	23	113.3	1.72	0.00	Bacillota	Bacilli	Bacillales	Bacillaceae
SAG	C_66_pH6	35	6	56.2	0.00	0.00	Bacillota	Bacilli	Bacillales	Bacillaceae
SAG	C_69_pH4	5,6	70	144	0,86	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	C_71_pH6	4.2	68	181.9	1.72	0.00	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae
SAG	C_74_pH4	12,4	30	73,8	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	C_75_pH6	7.7	51	155.5	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	C_75_pH7	11.1	35	118.1	1.72	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	C_76_pH4	10,8	60	144,3	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	C_79_pH6	8.5	16	59.7	0.00	0.00	Bacillota	Bacilli	Bacillales	Bacillaceae
SAG	C_82_pH4	4,3	106	211,8	2,58	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	C_84_pH4	79,6	60	154,5	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	C_9_pH6	11.2	19	68.5	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	C_91_pH6	5.6	47	115.5	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	C_91_pH7	15.1	51	187	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	C_93_pH6	5	29	85.9	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Corynebacteriaceae
SAG	D_1_pH4	1,2	75	84,7	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	D_1_pH7	2.9	47	120.1	2.07	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Lawsonellaceae
SAG	D_12_pH7	9.7	7	53.4	2.63	0.00	Pseudomonadota	Betaproteobacteria	unclassified Betaproteobacteria order	unclassified Betaproteobacteria fam
SAG	D_14_pH7	4.6	27	60.3	0.00	0.00	Bacillota	Bacilli	Bacillales	Bacillaceae
SAG	D_19_pH6	4.3	21	63.5	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	D_19_pH7	5.9	28	91.9	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	D_2_pH7	4.6	55	118.2	1.25	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	D_20_pH6	21.3	17	92.5	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	D_26_pH4	5	41	76,6	4,17	0.00	Bacillota	Negativicutes	Veillonellales	Veillonellaceae
SAG	D_27_pH4	91,9	34	140,7	4,17	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	D_27_pH6	7.9	18	53.4	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	D_3_pH7	20.2	64	264.7	16.71	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	D_32_pH4_bin	11,4	29	106,6	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	D_35_pH6	3.7	38	74.3	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	D_38_pH4	8,3	57	209,9	0.00	0.00	Bacillota	Bacilli	Bacillales	Staphylococcaceae
SAG	D_39_pH7	6.3	19	66.3	0.00	0.00	Bacteroidota	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae

Method	Sample	N50 (Kbp)	Number of contigs	Size (Kbp)	Completeness (%)	Contamination (%)	Phylum	Class	Order	Family
SAG	D_40_pH7	42.6	4	67.2	0.00	0.00	Mycoplasmatota	Mollicutes	Mycoplasmatales	Mycoplasmataceae
SAG	D_41_pH7	6.3	21	75.3	7.83	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	D_42_pH4	11,7	33	238,1	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	D_42_pH7	5.3	18	50	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	D_47_pH4	14,9	8	73,4	0.00	0.00	Actinomycetota	Actinomycetes	Bifidobacteriales	Bifidobacteriaceae
SAG	D_48_pH6	5.9	18	55.9	2.14	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Corynebacteriaceae
SAG	D_49_pH4	4	40	100,7	0.00	0.00	Bacillota	Bacilli	Bacillales	Bacillaceae
SAG	D_5_pH4	11,6	29	61,5	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	D_51_pH4	3,4	51	116,8	3,45	0.00	Bacillota	Bacilli	Bacillales	Staphylococcaceae
SAG	D_54_pH6	14.2	15	92.7	4.17	0.00	Bacillota	Bacilli	Lactobacillales	Streptococcaceae
SAG	D_56_pH4	8,9	37	148,9	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Corynebacteriaceae
SAG	D_57_pH7	33.6	13	67	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	D_6_pH4	1,2	105	119,5	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	D_61_pH4	7,2	65	137,7	6,03	0.00	Pseudomonadota	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae
SAG	D_63_pH4	3,8	37	94,7	1,72	0.00	Pseudomonadota	Gammaproteobacteria	unclassified Gammaproteobacteria order	unclassified Gammaproteobacteria fa
SAG	D_64_pH7	10.9	18	79.9	0.00	0.00	Pseudomonadota	Betaproteobacteria	Burkholderiales	unclassified Burkholderiales family
SAG	D_65_pH4	21	14	124,3	0,34	0.00	Actinomycetota	Actinomycetes	Kitasatosporales	Streptomycetaceae
SAG	D_66_pH7	14.9	19	59.7	2.13	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	D_68_pH4	4,9	25	81,9	0.00	0.00	Bacillota	Bacilli	Bacillales	Staphylococcaceae
SAG	D_68_pH6	21.6	15	88.4	0.00	0.00	Bacteroidota	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae
SAG	D_70_pH6	5.8	19	68.1	0.00	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	D_74_pH4	5,7	313	952,1	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	D_74_pH6	5.6	38	121.6	1.72	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	D_75_pH4	9,5	628	2794,9	26,3	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	D_75_pH6	2.2	42	66.7	1.72	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	D_76_pH4	5,2	33	112,3	3,45	0.00	Bacillota	Bacilli	Bacillales	Staphylococcaceae
SAG	D_81_pH6	9.5	21	92.7	0.17	0.00	Actinomycetota	Actinomycetes	Micrococcales	Cellulomonadaceae
SAG	D_82_pH7	13	36	115.5	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	D_83_pH6	105	23	163	8.33	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	D_83_pH7	10.3	51	179.3	0.00	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	D_84_pH4	4,4	43	119,7	7,84	0,31	Bacillota	Bacilli	Bacillales	Staphylococcaceae
SAG	D_86_pH6_bin	4.3	76	153.5	8.07	0.00	Bacteroidota	Flavobacteriia	Flavobacteriales	Weeksellaceae
SAG	D_87_pH7	9.2	12	81.7	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae
SAG	D_89_pH4	3,8	120	262,3	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	D_91_pH4	5,2	36	106,4	0.00	0.00	Bacillota	Bacilli	Bacillales	Bacillaceae
SAG	D_95_pH4_bin	19,4	7	68,9	4,17	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	sc_13_S13_pH4	18,4	43	120,9	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	sc_18_S18_pH4	20,8	241	573,6	12,1	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	sc_19_S19_pH4	2	94	132	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae

Method	Sample	N50 (Kbp)	Number of contigs	Size (Kbp)	Completeness (%)	Contamination (%)	Phylum	Class	Order	Family
SAG	sc_2_S2_pH4	13,3	192	489,7	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	sc_20_S20_pH4	6,7	154	308,1	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	sc_23_S23_pH4	3,4	220	367,7	5,26	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	sc_28_S28_pH4	95,3	105	435,4	5,42	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	sc_29_S29_pH4	12,1	166	405,1	1,72	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	sc_3_S3_pH4	18,5	353	1395,3	10,9	0.00	Bacillota	Clostridia	Thermoanaerobacterales	Thermoanaerobacteraceae
SAG	sc_31_S31_pH4	7,7	165	542,9	20,69	0.00	Pseudomonadota	Betaproteobacteria	Nitrosomonadales	Nitrosomonadaceae
SAG	sc_38_S38_pH4	9,6	144	357,6	7,02	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	sc_4_S4_pH4	2,9	86	148,9	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Nitrobacteraceae
SAG	sc_40_S40_pH4	5,8	115	312?9	4,31	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	sc_47_S47_pH4	0,9	71	67,3	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	sc_48_S48_pH4	3,2	182	323,6	3,39	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Mycobacteriaceae
SAG	sc_49_S49_pH4	5,1	75	179,7	3,95	0,01	Bacillota	Bacilli	Bacillales	Bacillaceae
SAG	sc_50_S50_pH4	4,2	31	61,6	4,17	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	sc_53_S53_pH4	7,1	38	79,1	1,88	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	sc_57_S57_pH4	2,2	133	196,2	7,76	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
SAG	sc_58_S58_pH4	12,8	284	821,8	16,89	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	sc_64_S64_pH4	12,3	15	78,9	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
SAG	sc_69_S69_pH4	3,4	136	271,9	4,93	0,15	Pseudomonadota	Betaproteobacteria	Burkholderiales	Oxalobacteraceae
SAG	sc_7_S7_pH4_bin	31,6	29	61,1	0.00	0.00	Pseudomonadota	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae
SAG	sc_70_S70_pH4	7,2	34	154,2	0.00	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Corynebacteriaceae
SAG	sc_72_S72_pH4	5,1	42	127,7	0.00	0.00	Pseudomonadota	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae
SAG	sc_77_S77_pH4	15,8	159	1128,6	18,96	0.00	Actinomycetota	Actinomycetes	Mycobacteriales	Corynebacteriaceae
SAG	sc_8_S8_pH4	10,8	45	194,5	18,96	0.00	Actinomycetota	Actinomycetes	Propionibacteriales	Propionibacteriaceae
Mini_MAG	Bin_41	9493	401	2102018	73.24	1.41	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
Mini_MAG	Bin_3	21884	698	4765498	63.38	1.41	Pseudomonadota	Betaproteobacteria	Burkholderiales	Burkholderiaceae
Mini_MAG	Bin_37	10613	37	237589	46.48	1.41	Pseudomonadota	Gammaproteobacteria	Moraxellales	Moraxellaceae
Mini_MAG	Bin_1	18980	514	3297843	47.89	4.23	Pseudomonadota	Betaproteobacteria	Burkholderiales	Comamonadaceae
Mini_MAG	Bin_17	29775	344	3531448	47.89	2.82	Pseudomonadota	Betaproteobacteria	Burkholderiales	Burkholderiaceae
Mini_MAG	Bin_15_1	8179	545	2269907	49.30	0.00	Pseudomonadota	Betaproteobacteria	Burkholderiales	Burkholderiaceae
Mini_MAG	Bin_15_2	27745	440	3634900	70.42	2.82	Pseudomonadota	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae

General discussion and perspectives

The interest in single-cell omics application in microbial ecology has grown in the last few years and since 2010 has slowly developed to decipher ecological processes (Bowers, Kyrpides, et al., 2017). In the last decade, different methodologies emerged for sample preparation and data handling, mostly customized to suit the sample requirements (e.g. soil samples will not present the same extraction procedure that human intestinal microbiota and both datasets will contain different kinds of contamination sources). The use of single-cell genomics procedures applied to prokaryotes has however highlighted some similarities between the studies often cited as limitations of this approach: the required equipment can be very costly (Woyke et al., 2017), the cost for whole genome sequencing of a consequent number of cells - while still not being close to the amount of information processed by metagenomics - is still very high (Blainey, 2013; Kashima et al., 2020; Kaster & Sobol, 2020), the genomes recovered are partial (Alneberg et al., 2018; Bowers, Doud, et al., 2017; Landry et al., 2017; Zheng et al., 2022), the risk of contamination is high (López-Escardó et al., 2017), and the molecular reactions used for genome amplification such as MDA have biases and could introduce errors (Kaster & Sobol, 2020; Raghunathan et al., 2005; Sobol, 2023; Woyke et al., 2017). Are these limitations all responsible for the scarce use of single-cell omics in microbial ecology? Interestingly, metagenomics presents some of these limitations as well (i.e. amplification bias for amplicons approaches, contamination, partial genomes), and others not shared with single-cell omics (i.e. chimera genomes generation, no access to within-species diversity, systematic consideration of free environmental DNA). This does not prevent this method from being the current norm for studying uncultivated microorganisms. It is thus most likely that induced costs are the reason for low single-cell omics usage as well as handling time. In practice, library preparation of single cells demands a longer preparation time than metagenomics. With the developed protocol, starting with the Nycodenz cell extractions, the metagenomic procedure per sample was done in 3.5 days counting the quality controls while it required a minimum of 7.5 days for the single-cell protocol for 380 cells. The equipment for cell isolation is also not commonly adapted for other usage, while metagenomics only requires a thermal cycler.

In this last section, I will first discuss the main outcomes of this work, followed by its technical and biological limitations. In the third paragraph, I will propose improvement possibilities for single-cell omics in laboratory equipment, reaction volumes and database handling. The future single-cell genomics applications with recent studies as examples will then be discussed as well as the technical future of single-cell omics.

I- Main outcomes of the developed strategies

During my thesis, I aimed to improve the robustness of single-cell whole sequencing (scWGS) at the library preparation, and data handling steps to lower the cost to a minimum and broaden its possible applications. To test and choose each step of the protocol, multiple quality controls had first to be developed and will be used for future improvements of the workflow.

The application of the library preparation protocol for scWGS on environmental samples showed the efficiency of the protocol in recovering SAGs from unknown or under-described bacterial groups. The comparison with MAGs showed that SAGs were fit to be used for community composition description and were less prone to contain external gDNA from the soil sample. The two approaches seem to provide different information reflecting their genomic strategies: MAGs are the mirror of core genomes from a global sequence dataset whereas SAGs have a narrower but more direct approach since it is cell-centred. The utilization of metagenomics can lead to assemblies with better completeness than SAGs (within my work) but also to genomes that can be chimeric, with only a few marker genes and contain external DNA despite improvements in cleanup procedures (Lou et al., 2023; McArdle & Kaforou, 2020; Vollmers et al., 2022). Furthermore, there is no possible information provided from a population perspective with MAGs despite that many ecological processes occur within species or require the consideration of populational variants (Van Rossum et al., 2020).

We faced the lack of a universal strategy for single-cell data cleanup in the literature: data treatments are often customized resulting in difficulties in comparing the quality of the results between similar studies. The development of the automated decontamination pipeline SINCERE DATA, was shown to be more efficient than manual decontamination of the reads and to handle datasets with various contaminant types and origins. We have tested this

pipeline on two different single-cell genomics datasets with success and have demonstrated that it will greatly simplify and unify the procedure for future research on microbial single-cell data analyses.

II- Encountered limitations of the library preparation protocol

During the protocol execution and the data treatment, I encountered some limitations of the procedures that should be assessed for further improvements. The contamination estimated based on the diversity of sequences identified per sample was varying between the experiments. On pure strains (Chapter II), the Enterobacteriaceae and Bacillaceae families were dominant in the sequences. The Enterobacteriaceae were also present in sequences recovered from environmental samples. This family is suspected to originate from the MDA kit reagents and Bacillaceae from cross-contamination between wells containing gDNA and the wells containing single cells. The contamination induced by commercialized MDA kits has been identified and discussed in a few studies (Blainey & Quake, 2011; Woyke et al., 2010). If not deleted, the remaining contaminant sequences might alter the recovery and interpretation of unknown genomes. The elaboration of negative controls to subtract background noise with such an experiment is also very limited, as any trace of DNA will be amplified with the high-fidelity polymerase used for the MDA, be it from the kit, from the air, from the manipulator or from cross-contamination of the wells and therefore the negative replicates could present different taxonomic profiles.

I initially aimed at preparing one 380-well plate with single cells per soil sample (Chapter III), hoping we would recover at least 300 SAGs for each sample. In reality, few SAGs passed the quality filters (i.e. length of the assembly and identification of dominant taxa) as we obtained 210 SAGs out of the 1140 isolated cells. These numbers are however coherent with studies with similar sample preparation on single prokaryotic cell genomes (López-Escardó et al., 2017; Roux, Hawley, Beltran, et al., 2014), and the number of isolated cells is not systematically compared to the number of SAGs recovered (Martinez-Garcia et al., 2012; Swan et al., 2013), especially with recently developed microfluidics which presents many recovered SAGs out on an unknown high number of generated droplets (Kogawa et al., 2018; Nishikawa et al., 2022; Zheng et al., 2022).

The application of the developed protocol confirmed the difficulty of recovering single amplified genomes (SAGs) with high completeness, even from known and referenced bacterial strains (Zheng et al., 2022). The SAGs we recovered from pure cultures (Chapter II) and soil samples (Chapter III), all had a measured completeness that averaged at 2.6 % and did not exceed 37 %. Even though we argued in Chapter II that completeness alone is not a guarantee of assembly quality, some studies succeeded at improving the completeness rate (Berube et al., 2018b; Nishikawa et al., 2022). The major purpose of the single-cell approach is to decipher the populational scale of microbes; however, with only genome fragments, it is still early for proper exploitation of this potential. Therefore, there is room for improvements regarding the DNA preparation to increase genome recovery, but also to limit the contamination we identified as induced by our manipulations.

III- Improvement strategies

3.1- Laboratory equipment

The deep search for contamination in our samples was very complex and highlighted the extent to which single-cell library preparation was sensitive to either manipulation steps or free eDNA contained in the samples. Single-cell omics are exposed to many sources of contamination: external cellular liquid (Blainey, 2013), room air flows, or bacterial origin of some kit compounds (Woyke et al., 2011). To limit the risks of contamination by manipulation, single-cell library preparation should be executed in a dedicated laboratory space with air filtration (cleanroom), with dedicated tools and instruments in cleanrooms. From our experience with gDNA contamination origin, I highly recommend avoiding manipulating gDNA at the same time as single-cell samples which present low genetic material concentrations and will be underrepresented in the contaminated sample. The room dedicated to single-bacteria isolation and library preparation should not be used for other amplification procedures such as PCRs nor for applications to other samples than bacteria. The pipettes, and isolation device (for us, the celleONE interior) can also be systematically UV-treated. Just like any sensitive DNA or RNA sample preparation, the water, tubes, and plates must be highly clean and again dedicated to the single-cell sample preparation clean room.

3.2- Miniaturization

Improvements could be implemented in our protocol, which also started to be tested in the literature, and involved miniaturisation of DNA preparation. The lack of MDA coverage uniformity has been theorized as responsible for the low genome coverage of SAGs (Kaster & Sobol, 2020), as well as biased GC or chimera productions during the genome amplification (Lasken, 2007). Therefore, it seems that this step modification can engender significant improvements in scWGS. The replacement or modifications of the enzyme and amplification strategy (Berube et al., 2018b; Gonzalez-Pena et al., 2021; Stepanauskas et al., 2017) have been tested as well as lowering the reaction volumes with encouraging results (Hosokawa et al., 2017; Sobol, 2023). The work of Sobol (Sobol, 2023) shows the improvements in genome coverage and uniformity of amplified genomes in 1.25 µL reaction volume while still working in 384 well plates, and questions the efficiency of further miniaturization such as what is done in microfluidics (Nishikawa et al., 2022). This is consistent with other experiments run at Cellenion (data not shown) which showed an improvement of the MDA in 1uL reaction volume, but not below. This can explain the issues we observed in the MDA efficiency variability in the cellenCHIP, with a reaction volume of 150 nL. The miniaturisation would also decrease the amount of contaminant sequences in the samples, as well as costs which would enable more sample processing.

3.3- Quality control and databases

The amount of unreferenced data we obtained in our dataset testified to the limited annotated references in the databases (Z. Zhang et al., 2020). Completing these databases will decrease the biases they contain towards cultivated strains and possible misidentification of closely related uncultured bacteria to improve genomic data interpretation. Complete genome referencing is necessary for the quality assessment of recovered SAGs and MAGs, as the bioinformatic tools rely on either the percentage of genome coverage by the assembly (QUAST) or the detection of marker genes (CheckM). The outcome of these metrics is therefore highly dependent on the information stored on the found taxa. Today, only 105,953 genomes are referred to as complete on NCBI (Federhen, 2012), which represents at best 0.13 % of the total bacterial diversity, and 0.0001% at worst (Id et al., 2019; Locey & Lennon, 2016). To improve these numbers, the massive sequencing of single environmental bacterial cells is necessary to combine the similarities of populations but also their variabilities in gene content. The task can be eased by targeting specific taxa with 16srRNA gene (i.e. if we suppose an a

priori knowledge of the target in databses) PCRs following the genome amplification step on individual cells (Berube et al., 2018), or by selective isolation with fluorescent probes (Dam et al., 2020). This greatly helps the study of rare and underrepresented taxa that would otherwise be difficult to spot (Dam et al., 2020). With growing bacterial strain discoveries, the question of consistent nomenclature and data storage has been raised and some proposed a consortium to solve these growing issues (Bowers, Kyrpides, et al., 2017; Konstantinidis et al., 2017; Murray et al., 2020). The consistency in single-cell data generation (i.e. sampling, cell isolation, library preparation, SAG quality filtering) is essential for the efficient validation of newly discovered genomes with the least contamination and biases possible across different studies.

IV- The future of microbial ecology through the lens of single-cell genomics

An optimized single-cell sample preparation procedure as proposed above could generate many major advances in microbial ecology research. Theoretically, scWGS is a proxy for the evaluation of environmental DNA present in metagenomics data. Used at high scales and throughput, scWGS can become the alternative to metagenomics for a finer-scale evaluation of genome content and unbiased community composition and population structure. The more complete and purer the assembled genomes become, the more we will be able to identify and measure the diversity of unknown microbes and correct the genetic content of referenced genomes based on metagenomics only. The possibility to isolate cells based on cell features (i.e. live, dead/ in dormancy, respiratory activity, protein content…) offers infinite possibilities for fine-scale ecological hypothesis testing with access to the population scale of bacteria. Annotated genomes could be used for the evaluation of community structure and interactions and their evolutionary capacities with the modelling of metabolism networks to make a putative link between microbial diversity and richness and functions. This (r-)evolution in microbial ecology can be measured every year. Since the beginning of my thesis, I have noticed and read multiple single-cell-based studies participating in bacterial communities understanding improvement. Following is a short and non-exhaustive list of recent research on the topic.

The usage of single-cell omics has been done for diverse ecological purposes, with a focus on unveiling the unknown portion of bacteria, from genetic and functional standpoints. The exploration of bacterial populations has provided insights into preferable symbiont associations (Boscaro et al., 2022) or evolutionary gene polymorphisms in sub-populations of well-studied strains of *Salmonella* (Bawn et al., 2022). Large scale inventory of bacterial genomes in multiple ecosystems has been described, from the human body (Zheng et al., 2022), soil (Aoki et al., 2022; Nishikawa et al., 2022), and ocean (Anstett et al., 2023). These inventories will participate in the expansion of the microbial tree of life (Ahrendt et al., 2018). Among these studies, the majority used microfluidics for sample preparation with gel-based bacterial cell encapsulation (Aoki et al., 2022; Nishikawa et al., 2022; Zheng et al., 2022). Some still use the FACS technology for cell isolation (Anstett et al., 2023), and just like gel-based approaches are able to recover hundreds to thousands of SAGs. Most of these applications are still mainly descriptive of the bacterial diversity, but may help in the elaboration of testable ecological hypotheses (Prosser & Martiny, 2020; A. Tripathi et al., 2018).

V- The future of single-cell omics

Beyond genomic information of bacteria cells that provide a metabolic potential, single-cell RNA sequencing (scRNA seq) is necessary to evaluate the expressed functions of an individual in different environmental conditions. The phenotype of bacteria can vary very quickly depending on its direct neighbours (Scanlan & Buckling, 2012), environmental parameters (Beaumont et al., 2009), and host condition (Reese & Kearney, 2019) and therefore would be necessary to implement to understand the bacteria population and community functioning. This would allow us to answer the key question “Who does what, and how do they do it?” in a changing environment to understand the evolutionary dynamics of bacteria communities.

To magnify the extent of single-cell data interpretation, the combination of genomic, transcriptomics, proteomic, and epigenomic information from the same cell is an emerging topic but has never been applied to prokaryotes and has only been theorized for eukaryotic cell (Bock et al., 2016; Chappell et al., 2018; Kashima et al., 2020; Song et al., 2019). The possibility of linking each genome to its expressed genes and proteins would allow us to

answer many additional questions in ecology and health research, but we do not expect this application to be executed in the near future.

This work raised many questions, still unresolved, and are listed below.

VI- Outstanding questions

- How to deal with clustered microorganisms, treated as one object by the single-cell isolation instruments?
- How can we be certain that a lysis buffer will equally break the cell walls of all bacterial taxa?
- How can we be certain that the sequences in our samples are not contaminated?
- How to make the distinction within unassigned reads between what is unreferenced target DNA and artefactual reads?

Final thoughts

New technologies are emerging each year for single-cell omics applications, which is fundamental for this approach to evolve. The complexity in the elaboration of our protocol and its outputs raised many questions, technical and biological as discussed above, but also methodological. Mostly, I noticed that the race for technical improvements makes us put aside other approaches, sometimes less impressive and new than single-cell omics but with simple and efficient functioning. As our experimental possibilities grow, we should keep considering fundamental microbiology practices: time and local scales of sampling, hypothesis testing, cultivation, replicates, and negative controls (i.e. there are almost no negative controls in genomic sequencing). The oldest way of studying bacteria, culturing, can afford relevant information and has made big advances since Pasteur's first bacteria cultivation (Lewis & Ettema, 2019; Lewis et al., 2021). The race for technological improvement is worth nothing (scientifically) if not used for proper scientific testing. The focus must be set on finding the adequate technology to test hypotheses, not the other way around (Prosser, 2015). There will always be biases for any molecular strategy employed: PCRs, MDAs, single-cell omics, or metagenomics, but we have the tools and knowledge to combine these approaches to get the most out of what genomic information can provide (Alneberg et al., 2018; Hedlund et al., 2014; Mende et al., 2016), always with relevant and appropriate questions and hypothesis.

Bibliography

Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., List, M., & Neuhaus, K. (2021). Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. *MSphere*, *6*(1). https://doi.org/10.1128/MSPHERE.01202-20/SUPPL_FILE/MSPHERE.01202-20-ST005.PDF

Ahrendt, S. R., Quandt, C. A., Ciobanu, D., Clum, A., Salamov, A., Andreopoulos, B., Cheng, J.-F., Woyke, T., Pelin, A., Henrissat, B., Reynolds, N. K., Benny, G. L., Smith, M. E., James, T. Y., & Grigoriev, I. V. (2018). Leveraging single-cell genomics to expand the fungal tree of life. *Nature Microbiology*, *3*(12), 1417–1428. https://doi.org/10.1038/s41564-018-0261-0

Aigle, A., Gubry-Rangin, C., Thion, C., Estera-Molina, K. Y., Richmond, H., Pett-Ridge, J., Firestone, M. K., Nicol, G. W., & Prosser, J. I. (2020). Experimental testing of hypotheses for temperature- and pH-based niche specialization of ammonia oxidizing archaea and bacteria. *Environmental Microbiology*, *22*(9), 4032–4045. https://doi.org/10.1111/1462-2920.15192

Alneberg, J., Karlsson, C. M. G., Divne, A.-M., Bergin, C., Homa, F., Lindh, M. V., Hugerth, L. W., Ettema, T. J. G., Bertilsson, S., Andersson, A. F., & Pinhassi, J. (2018). Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome*, *6*(1), 173. https://doi.org/10.1186/s40168-018-0550-0

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389. https://doi.org/10.1093/NAR/25.17.3389

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data

Anstett, J., Plominsky, A. M., DeLong, E. F., Kiesser, A., Jürgens, K., Morgan-Lang, C., Stepanauskas, R., Stewart, F. J., Ulloa, O., Woyke, T., Malmstrom, R., & Hallam, S. J. (2023). A compendium of bacterial and archaeal single-cell amplified genomes from oxygen deficient marine waters. *Scientific Data*, *10*(1), 332. https://doi.org/10.1038/s41597-023-02222-y

Antoniewicz, M. R. (2020). A guide to deciphering microbial interactions and metabolic fluxes in microbiome communities. In *Current Opinion in Biotechnology* (Vol. 64, pp. 230–237). Elsevier Ltd. https://doi.org/10.1016/j.copbio.2020.07.001

Aoki, W., Kogawa, M., Matsuda, S., Matsubara, K., Hirata, S., Nishikawa, Y., Hosokawa, M., Takeyama, H., Matoh, T., & Ueda, M. (2022). Massively parallel single-cell genomics of microbiomes in rice paddies. *Frontiers in Microbiology*, *13*, 4400. https://doi.org/10.3389/FMICB.2022.1024640/BIBTEX

Archie, E. A., & Theis, K. R. (2011). Animal behaviour meets microbial ecology. *Animal Behaviour*, *82*(3), 425–436. https://doi.org/10.1016/j.anbehav.2011.05.029

Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J., & Polz, M. F. (2019). A Reverse Ecology

Approach Based on a Biological Definition of Microbial Populations. *Cell*, *178*(4), 820-834.e14. https://doi.org/10.1016/J.CELL.2019.06.033

Arikawa, K., Ide, K., Kogawa, M., Saeki, T., Yoda, T., Endoh, T., Matsuhashi, A., Takeyama, H., & Hosokawa, M. (2021). Recovery of strain-resolved genomes from human microbiome through an integration framework of single-cell genomics and metagenomics. *Microbiome 2021 9:1*, *9*(1), 1–16. https://doi.org/10.1186/S40168-021-01152-4

Baishya, J., Bisht, K., Rimbey, J. N., Yihunie, K. D., Islam, S., Mahmud, H. Al, Waller, J. E., & Wakeman, C. A. (2021). The impact of intraspecies and interspecies bacterial interactions on disease outcome. *Pathogens*, *10*(2), 1–11. https://doi.org/10.3390/pathogens10020096

Baker, B. J., & Dick, G. J. (2013). Omic approaches in microbial ecology: Charting the unknown. *Microbe*, *8*(9), 353–360. https://doi.org/10.1128/microbe.8.353.1

Banerjee, S., Schlaeppi, K., & van der Heijden, M. G. A. (2018). Keystone taxa as drivers of microbiome structure and functioning. In *Nature Reviews Microbiology* (Vol. 16, Issue 9, pp. 567–576). Nature Publishing Group. https://doi.org/10.1038/s41579-018-0024-1

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology, 19(5), 455-477.

Barnett, S. E., Youngblut, N. D., & Buckley, D. H. (2020). Soil characteristics and land-use drive bacterial community assembly patterns. *FEMS Microbiology Ecology*, *96*(1), 194. https://doi.org/10.1093/FEMSEC/FIZ194

Bartram, A. K., Jiang, X., Lynch, M. D. J., Masella, A. P., Nicol, G. W., Dushoff, J., & Neufeld, J. D. (2014). Exploring links between pH and bacterial community composition in soils from the Craibstone Experimental Farm. *FEMS Microbiology Ecology*, *87*(2), 403–415. https://doi.org/10.1111/1574-6941.12231

Bawn, M., Hernandez, J., Trampari, E., Thilliez, G., Quince, C., Webber, M. A., Kingsley, R. A., Hall, N., & Macaulay, I. C. (2022). Single-cell genomics reveals population structures from in vitro evolutionary studies of Salmonella. *Microbial Genomics*, *8*(9). https://doi.org/10.1099/mgen.0.000871

Beaumont, H. J. E., Gallie, J., Kost, C., Ferguson, G. C., & Rainey, P. B. (2009). Experimental evolution of bet hedging. *Nature*, *462*(7269), 90–93. https://doi.org/10.1038/nature08504

Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P., & Bapteste, E. (2018). Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biology and Evolution*, *10*(3), 707–715. https://doi.org/10.1093/GBE/EVY031

Berube, P. M., Biller, S. J., Hackl, T., Hogle, S. L., Satinsky, B. M., Becker, J. W., Braakman, R., Collins, S. B., Kelly, L., Berta-Thompson, J., Coe, A., Bergauer, K., Bouman, H. A., Browning, T. J., De Corte, D., Hassler, C., Hulata, Y., Jacquot, J. E., Maas, E. W., … Chisholm, S. W. (2018a). Data descriptor: Single cell genomes of Prochlorococcus,

Synechococcus, and sympatric microbes from diverse marine environments. *Scientific Data*, *5*(March), 1–11. https://doi.org/10.1038/sdata.2018.154

Berube, P. M., Biller, S. J., Hackl, T., Hogle, S. L., Satinsky, B. M., Becker, J. W., Braakman, R., Collins, S. B., Kelly, L., Berta-Thompson, J., Coe, A., Bergauer, K., Bouman, H. A., Browning, T. J., De Corte, D., Hassler, C., Hulata, Y., Jacquot, J. E., Maas, E. W., … Chisholm, S. W. (2018b). Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. *Scientific Data 2018 5:1*, *5*(1), 1–11. https://doi.org/10.1038/sdata.2018.154

Blainey, P. C. (2013). The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews*, *37*(3), 407–427. https://doi.org/10.1111/1574-6976.12015

Blainey, P. C., & Quake, S. R. (2011). Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Research*, *39*(4), e19–e19. https://doi.org/10.1093/NAR/GKQ1074

Bock, C., Farlik, M., & Sheffield, N. C. (2016). Multi-Omics of Single Cells: Strategies and Applications. *Trends in Biotechnology*, *34*(8), 605–608. https://doi.org/10.1016/J.TIBTECH.2016.04.004

Bokulich, N. A., Ziemski, M., Robeson, M. S., & Kaehler, B. D. (2020). Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. *Computational and Structural Biotechnology Journal*, *18*, 4048–4062. https://doi.org/10.1016/J.CSBJ.2020.11.049

Bonnet, M., Lagier, J. C., Raoult, D., & Khelaifia, S. (2020). Bacterial culture through selective and non-selective conditions: the evolution of culture media in clinical microbiology. *New Microbes and New Infections*, *34*, 100622. https://doi.org/10.1016/J.NMNI.2019.100622

Bordenstein, S. R., & Theis, K. R. (2015). Host biology in light of the microbiome: Ten principles of holobionts and hologenomes. *PLoS Biology*, *13*(8), 1–23. https://doi.org/10.1371/journal.pbio.1002226

Boscaro, V., Manassero, V., Keeling, P. J., & Vannini, C. (2022). Single-cell Microbiomics Unveils Distribution and Patterns of Microbial Symbioses in the Natural Environment. *Microbial Ecology*. https://doi.org/10.1007/s00248-021-01938-x

Bowers, R. M., Doud, D. F. R., & Woyke, T. (2017). Analysis of single-cell genome sequences of bacteria and archaea. *Emerging Topics in Life Sciences*, *1*(3), 249–255. https://doi.org/10.1042/ETLS20160028

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., … Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, *35*(8), 725–731. https://doi.org/10.1038/nbt.3893

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2018). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, *20*(4), 1125–1139.

https://doi.org/10.1093/bib/bbx120

Brown, C. T. (2015). Strain recovery from metagenomes. *Nature Biotechnology*, *33*(10), 1041–1043. https://doi.org/10.1038/nbt.3375

Bush, S. J., Foster, D., Eyre, D. W., Clark, E. L., de Maio, N., Shaw, L. P., Stoesser, N., Peto, T. E. A., Crook, D. W., & Walker, A. S. (2020). Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. *GigaScience*, *9*(2), 1–21. https://doi.org/10.1093/GIGASCIENCE/GIAA007

Busley, A. V., Kleinsorge, M., & Cyganek, L. (2023). Generation of a genetically-modified induced pluripotent stem cell line harboring an oncogenic gene variant KRAS p.G12V. *Stem Cell Research*, *69*, 103105. https://doi.org/10.1016/J.SCR.2023.103105

Carlström, C. I., Field, C. M., Bortfeld-Miller, M., Müller, B., Sunagawa, S., & Vorholt, J. A. (2019). Synthetic microbiota reveal priority effects and keystone strains in the Arabidopsis phyllosphere. *Nature Ecology and Evolution*, *3*(10). https://doi.org/10.1038/s41559-019-0994-z

Chappell, L., Russell, A. J. C., & Voet, T. (2018). Single-Cell (Multi)omics Technologies. *Annu. Rev. Genom. Hum. Genet*, *19*, 15–41. https://doi.org/10.1146/annurev-genom-091416

Chen, Z., Chen, L., & Zhang, W. (2017). Tools for Genomic and Transcriptomic Analysis of Microbes at Single-Cell Level. *Frontiers in Microbiology*, *8*, 1831. https://doi.org/10.3389/fmicb.2017.01831

Chijiiwa, R., Hosokawa, M., Kogawa, M., Nishikawa, Y., Ide, K., Sakanashi, C., Takahashi, K., & Takeyama, H. (2020). Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. *Microbiome*, *8*(1), 5. https://doi.org/10.1186/s40168-019-0779-2

Ciccarese, D., Zuidema, A., Merlo, V., & Johnson, D. R. (2020). Interaction-dependent effects of surface structure on microbial spatial self-organization. *Philosophical Transactions of the Royal Society B*, *375*(1798). https://doi.org/10.1098/RSTB.2019.0246

Coker, J., Zhalnina, K., Marotz, C., Thiruppathy, D., Tjuanta, M., D'Elia, G., Hailu, R., Mahosky, T., Rowan, M., Northen, T. R., & Zengler, K. (2022). A Reproducible and Tunable Synthetic Soil Microbial Community Provides New Insights into Microbial Ecology. *MSystems*, *7*(6). https://doi.org/10.1128/MSYSTEMS.00951-22/SUPPL_FILE/MSYSTEMS.00951-22-S0009.DOCX

Compant, S., Samad, A., Faist, H., & Sessitsch, A. (2019). A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application. *Journal of Advanced Research*, *19*, 29–37. https://doi.org/10.1016/j.jare.2019.03.004

Cordero, O. X., & Datta, M. S. (2016). Microbial interactions and community assembly at microscales. In *Current Opinion in Microbiology* (Vol. 31, pp. 227–234). Elsevier Ltd. https://doi.org/10.1016/j.mib.2016.03.015

Courtois, S., Sa Frostega, A. Ê., Go, P., Depret, G., Jeannin, P., & Simonet, P. (2001). Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environmental Microbiology*, *3*(7), 431–439. https://doi.org/10.1046/j.1462-2920.2001.00208.x

Cristinelli, S., & Ciuffi, A. (2018). The use of single-cell RNA-Seq to understand virus–host interactions. *Current Opinion in Virology*, *29*, 39–50. https://doi.org/10.1016/J.COVIRO.2018.03.001

Crits-Christoph, A., Olm, M. R., Diamond, S., Bouma-Gregson, K., & Banfield, J. F. (2020). Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. *ISME Journal*, *14*(7), 1834–1846. https://doi.org/10.1038/s41396-020-0655-x

D'Souza, G., Shitut, S., Preussger, D., Yousif, G., Waschina, S., & Kost, C. (2018). Ecology and evolution of metabolic cross-feeding interactions in bacteria. In *Natural Product Reports* (Vol. 35, Issue 5, pp. 455–488). Royal Society of Chemistry. https://doi.org/10.1039/c8np00009c

Dagan, T., Artzy-Randrup, Y., & Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(29), 10039–10044. https://doi.org/10.1073/pnas.0800679105

Dam, H. T., Vollmers, J., Sobol, M. S., Cabezas, A., & Kaster, A. K. (2020). Targeted Cell Sorting Combined With Single Cell Genomics Captures Low Abundant Microbial Dark Matter With Higher Sensitivity Than Metagenomics. *Frontiers in Microbiology*, *11*, 511433. https://doi.org/10.3389/FMICB.2020.01377/BIBTEX

Davis, K. M., & Isberg, R. R. (2016). Defining heterogeneity within bacterial populations via single cell approaches. *BioEssays*, *38*(8), 782–790. https://doi.org/10.1002/bies.201500121

de la Cruz, F., & Davies, J. (2000). Antibiotic resistance: the immediate response. *Trends Microbiology*, *8*(3), 128–133. https://pdf.sciencedirectassets.com/271202/1-s2.0-S0966842X00X00492/1-s2.0-S0966842X00017030/main.pdf?x-amz-security-token=AgoJb3JpZ2luX2VjEE4aCXVzLWVhc3QtMSJGMEQCICJF0G%2FFYxOjdscZPC4sklPTTn%2BDNqpelQfFOwhuQq2CAiBbjZK2C3mol48tATaItI%2F8l0w1VpAt1uo%2BoaR2

Defives, C., Guyard, S., Oularé, M. M., Mary, P., & Hornez, J. P. (1999). Total counts, culturable and viable, and non-culturable microflora of a French mineral water: a case study. *Journal of Applied Microbiology*, *86*(6), 1033–1038. https://doi.org/10.1046/J.1365-2672.1999.00794.X

Deng, Y., Ruan, Y., Ma, B., Timmons, M. B., Lu, H., Xu, X., Zhao, H., & Yin, X. (2019). Multi-omics analysis reveals niche and fitness differences in typical denitrification microbial aggregations. *Environment International*, *132*(June). https://doi.org/10.1016/j.envint.2019.105085

Dhungana, I., Kantar, M. B., & Nguyen, N. H. (2023). Root exudate composition from different plant species influences the growth of rhizosphere bacteria. *Rhizosphere*, *25*, 100645. https://doi.org/10.1016/J.RHISPH.2022.100645

Douglas, A. E. (2014). Symbiosis as a general principle in eukaryotic evolution. *Cold Spring Harbor Perspectives in Biology*, *6*(2). https://doi.org/10.1101/cshperspect.a016113

Douglas, A. E. (2020). The microbial exometabolome: Ecological resource and architect of

microbial communities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1798). https://doi.org/10.1098/rstb.2019.0250

Ellegaard, K. M., & Engel, P. (2016). Beyond 16S rRNA community profiling: Intra-species diversity in the gut microbiota. *Frontiers in Microbiology*, *7*(SEP), 1–16. https://doi.org/10.3389/fmicb.2016.01475

Engel, P., Stepanauskas, R., & Moran, N. A. (2014). Hidden Diversity in Honey Bee Gut Symbionts Detected by Single-Cell Genomics. *PLoS Genetics*, *10*(9). https://doi.org/10.1371/journal.pgen.1004596

Enomoto, S., Chari, A., Clayton, A. L., & Dale, C. (2017). Quorum Sensing Attenuates Virulence in Sodalis praecaptivus. *Cell Host and Microbe*, *21*(5), 629-636.e5. https://doi.org/10.1016/j.chom.2017.04.003

Eren, A. M., Esen, O. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: An advanced analysis and visualization platformfor 'omics data. *PeerJ*, *2015*(10), e1319. https://doi.org/10.7717/PEERJ.1319/SUPP-5

Escudeiro, P., Henry, C. S., & Dias, R. P. M. (2022). Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. *Current Research in Microbial Sciences*, *3*(July), 100159. https://doi.org/10.1016/j.crmicr.2022.100159

Estévez-Gómez, N., Prieto, T., Guillaumet-Adkins, A., Heyn, H., Prado-López, S., & Posada, D. (2018). Comparison of single-cell whole-genome amplification strategies. *BioRxiv*, 443754. https://doi.org/10.1101/443754

Estrela, S., Morris, J. J., & Kerr, B. (2016). Private benefits and metabolic conflicts shape the emergence of microbial interdependencies. *Environmental Microbiology*, *18*(5), 1415–1427. https://doi.org/10.1111/1462-2920.13028

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, *40*(D1), D136–D143. https://doi.org/10.1093/NAR/GKR1178

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Will FitzHugh, Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L., Glodek, A., … Venter, J. C. (1995). Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. *Science*, *269*(5223), 496–512. https://doi.org/10.1126/SCIENCE.7542800

Fu, Y., Li, C., Lu, S., Zhou, W., Tang, F., Xie, X. S., & Huang, Y. (2015). Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(38), 11923–11928. https://doi.org/10.1073/PNAS.1513988112/SUPPL_FILE/PNAS.1513988112.SAPP.PDF

Fung, T. C., Olson, C. A., & Hsiao, E. Y. (2017). Interactions between the microbiota, immune and nervous systems in health and disease. *Nature Neuroscience 2016 20:2*, *20*(2), 145–155. https://doi.org/10.1038/nn.4476

Fykse, E. M., Olsen, J. S., & Skogan, G. (2003). Application of sonication to release DNA from Bacillus cereus for quantitative detection by real-time PCR. *Journal of Microbiological Methods*, *55*(1), 1–10. https://doi.org/10.1016/S0167-7012(03)00091-5

García-García, N., Tamames, J., Linz, A. M., Pedrós-Alió, C., & Puente-Sánchez, F. (2019). Intra-species diversity ensures the maintenance of functional microbial communities under changing environmental conditions. In *bioRxiv* (p. 530022). bioRxiv. https://doi.org/10.1101/530022

Garcia, S. L., Stevens, S. L. R., Crary, B., Martinez-Garcia, M., Stepanauskas, R., Woyke, T., Tringe, S. G., Andersson, S. G. E., Bertilsson, S., Malmstrom, R. R., & McMahon, K. D. (2018). Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. *ISME Journal*, *12*(3), 742–755. https://doi.org/10.1038/s41396-017-0001-0

Garner, R. E., Kraemer, S. A., Onana, V. E., Fradette, M., Varin, M. P., Huot, Y., & Walsh, D. A. (2023). A genome catalogue of lake bacterial diversity and its drivers at continental scale. *Nature Microbiology 2023*, 1–15. https://doi.org/10.1038/S41564-023-01435-6

Gawad, C., Koh, W., & Quake, S. R. (2016a). Single-cell genome sequencing: Current state of the science. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2015.16

Gawad, C., Koh, W., & Quake, S. R. (2016b). Single-cell genome sequencing: Current state of the science. *Nature Reviews Genetics*, *17*(3), 175–188. https://doi.org/10.1038/nrg.2015.16

Gest, H. (2004). The discovery of microorganisms by Robert Hooke and Antoni van Leeuwenhoek, Fellows of The Royal Society. *Notes and Records of the Royal Society of London*, *58*(2), 187–201. https://doi.org/10.1098/RSNR.2004.0055

Ghylin, T. W., Garcia, S. L., Moya, F., Oyserman, B. O., Schwientek, P., Forest, K. T., Mutschler, J., Dwulit-Smith, J., Chan, L.-K., Martinez-Garcia, M., Sczyrba, A., Stepanauskas, R., Grossart, H.-P., Woyke, T., Warnecke, F., Malmstrom, R., Bertilsson, S., & McMahon, K. D. (2014). Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage. *The ISME Journal*, *8*(12), 2503–2516. https://doi.org/10.1038/ismej.2014.135

Gilbert, S. F., Sapp, J., & Tauber, A. I. (2012). A Symbiotic View of Life: We Have Never Been Individuals. *Https://Doi.Org/10.1086/668166*, *87*(4), 325–341. https://doi.org/10.1086/668166

Gonzalez-Pena, V., Natarajan, S., Xia, Y., Klein, D., Carter, R., Pang, Y., Shaner, B., Annu, K., Putnam, D., Chen, W., Connelly, J., Pruett-Miller, S., Chen, X., Easton, J., & Gawad, C. (2021). Accurate genomic variant detection in single cells with primary template-directed amplification. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(24), 1–12. https://doi.org/10.1073/pnas.2024176118

Gopal, M., & Gupta, A. (2016). Gopal, M., & Gupta, A. (2016). Microbiome selection could spur next-generation plant breeding strategies. Frontiers in Microbiology, 7(DEC), 1–10. https://doi.org/10.3389/fmicb.2016.01971Microbiome selection could spur next-generation plant breeding strat. *Frontiers in Microbiology*, *7*(DEC), 1–10. https://doi.org/10.3389/fmicb.2016.01971

Gorter, F. A., Manhart, M., & Ackermann, M. (2020). Understanding the evolution of interspecies interactions in microbial communities. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 375, Issue 1798). Royal Society Publishing.

https://doi.org/10.1098/rstb.2019.0256

Gubry-Rangin, C., Kratsch, C., Williams, T. A., McHardy, A. C., Embley, T. M., Prosser, J. I., & Macqueen, D. J. (2015). Coupling of diversification and pH adaptation during the evolution of terrestrial Thaumarchaeota. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(30), 9370–9375. https://doi.org/10.1073/pnas.1419329112

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, *29*(8), 1072–1075. https://doi.org/10.1093/BIOINFORMATICS/BTT086

Guy, L., & Ettema, T. J. G. (2011). The archaeal "TACK" superphylum and the origin of eukaryotes. *Trends in Microbiology*, *19*(12), 580–587. https://doi.org/10.1016/j.tim.2011.09.002

Haeckel, E. (1866). Generelle Morphologie der Organismen. *Generelle Morphologie Der Organismen*. https://doi.org/10.1515/9783110848281/HTML

Hall, E., Kim, S., Appadoo, V., Zare, R., Hall, E. W., Kim, S., Appadoo, V., & Zare, R. N. (2013). Lysis of a Single Cyanobacterium for Whole Genome Amplification. *Micromachines*, *4*(3), 321–332. https://doi.org/10.3390/mi4030321

Hammer, T. J., Janzen, D. H., Hallwachs, W., Jaffe, S. P., & Fierer, N. (2017). Caterpillars lack a resident gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(36), 9641–9646. https://doi.org/10.1073/pnas.1707186114

Hammer, T. J., Sanders, J. G., & Fierer, N. (2019). Not all animals need a microbiome. *FEMS Microbiology Letters*, *366*(10), 1–11. https://doi.org/10.1093/femsle/fnz117

Hassani, M. A., Durán, P., & Hacquard, S. (2018). Microbial interactions within the plant holobiont. In *Microbiome* (Vol. 6, Issue 1, p. 58). NLM (Medline). https://doi.org/10.1186/s40168-018-0445-0

He, J., Du, S., Tan, X., Arefin, A., & Han, C. S. (2016). Improved lysis of single bacterial cells by a modified alkaline-thermal shock procedure. *BioTechniques*, *60*(3), 129–135. https://doi.org/10.2144/000114389

Hedlund, B. P., Dodsworth, J. A., Murugapiran, S. K., Rinke, C., & Woyke, T. (2014). Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter." In *Extremophiles* (Vol. 18, Issue 5, pp. 865–875). Springer-Verlag Tokyo. https://doi.org/10.1007/s00792-014-0664-7

Henry, L. P., Bruijning, M., Forsberg, S. K. G., & Ayroles, J. F. (2021). The microbiome extends host evolutionary potential. *Nature Communications*, *12*(1), 1–13. https://doi.org/10.1038/s41467-021-25315-x

Hoang, K. L., Gerardo, N. M., & Morran, L. T. (2021). Association with a novel protective microbe facilitates host adaptation to a stressful environment. *Evolution Letters*, *5*(2), 118–129. https://doi.org/10.1002/evl3.223

Hosokawa, M., Endoh, T., Kamata, K., Arikawa, K., Nishikawa, Y., Kogawa, M., Saeki, T., Yoda,

T., & Takeyama, H. (2022). Strain-level profiling of viable microbial community by selective single-cell genome sequencing. *Scientific Reports*, *12*(1), 4443. https://doi.org/10.1038/s41598-022-08401-y

Hosokawa, M., Nishikawa, Y., Kogawa, M., & Takeyama, H. (2017). Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics. *Scientific Reports 2017 7:1*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-05436-4

Houwenhuyse, S., Stoks, R., Mukherjee, S., & Decaestecker, E. (2021). Locally adapted gut microbiomes mediate host stress tolerance. *ISME Journal*, *15*(8), 2401–2414. https://doi.org/10.1038/s41396-021-00940-y

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5), 1–6. https://doi.org/10.1038/nmicrobiol.2016.48

Id, S. L., Id, F. M., & Doebeli, M. (2019). *A census-based estimate of Earth ' s bacterial and archaeal diversity*. 1–30.

Isobe, K., Bouskill, N. J., Brodie, E. L., Sudderth, E. A., & Martiny, J. B. H. (2020). Phylogenetic conservation of soil bacterial responses to simulated global changes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1798). https://doi.org/10.1098/rstb.2019.0242

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology 2016 17:1*, *17*(1), 1–11. https://doi.org/10.1186/S13059-016-1103-0

Jiao, J. Y., Liu, L., Hua, Z. S., Fang, B. Z., Zhou, E. M., Salam, N., Hedlund, B. P., & Li, W. J. (2021). Microbial dark matter coming to light: Challenges and opportunities. *National Science Review*, *8*(3), 1–5. https://doi.org/10.1093/nsr/nwaa280

Jiao, S., & Lu, Y. (2020). Soil pH and temperature regulate assembly processes of abundant and rare bacterial communities in agricultural ecosystems. *Environmental Microbiology*, *22*(3), 1052–1065. https://doi.org/10.1111/1462-2920.14815

Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., Küsel, K., Rillig, M. C., Rivett, D. W., Salles, J. F., Van Der Heijden, M. G. A., Youssef, N. H., Zhang, X., Wei, Z., & Hol, G. W. H. (2017). Where less may be more: How the rare biosphere pulls ecosystems strings. In *ISME Journal* (Vol. 11, Issue 4, pp. 853–862). Nature Publishing Group. https://doi.org/10.1038/ismej.2016.174

Kamada, N., Seo, S. U., Chen, G. Y., & Núñez, G. (2013). Role of the gut microbiota in immunity and inflammatory disease. *Nature Reviews Immunology 2013 13:5*, *13*(5), 321–335. https://doi.org/10.1038/nri3430

Kang, Y., McMillan, I., Norris, M. H., & Hoang, T. T. (2015). Single prokaryotic cell isolation and total transcript amplification protocol for transcriptomic analysis. *Nature Protocols*, *10*(7), 974–984. https://doi.org/10.1038/nprot.2015.058

Kans, J. (2023). Entrez direct: e-utilities on the UNIX command line. In Entrez programming

utilities help [Internet]. National Center for Biotechnology Information (US).

Kashima, Y., Sakamoto, Y., Kaneko, K., Seki, M., Suzuki, Y., & Suzuki, A. (2020). Single-cell sequencing techniques from individual to multiomics analyses. *Experimental & Molecular Medicine*, *52*, 1419–1427. https://doi.org/10.1038/s12276-020-00499-2

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. R., Stocker, R., Follows, M. J., Stepanauskas, R., & Chisholm, S. W. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science*, *344*(6182), 416–420. https://doi.org/10.1126/science.1248575

Kaster, A. K., & Sobol, M. S. (2020). Microbial single-cell omics: the crux of the matter. In *Applied Microbiology and Biotechnology* (Vol. 104, Issue 19, pp. 8209–8220). Springer. https://doi.org/10.1007/s00253-020-10844-0

Katz, L. S., Griswold, T., Morrison, S. S., Caravas, J. A., Zhang, S., Bakker, H. C. den, Deng, X., & Carleton, H. A. (2019). Mashtree: a rapid comparison of whole genome sequence files. *Journal of Open Source Software*, *4*(44), 1762. https://doi.org/10.21105/JOSS.01762

Kelly, S., Wickstead, B., & Gull, K. (2011). Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1708), 1009–1018. https://doi.org/10.1098/rspb.2010.1427

Kemp, J. S., Paterson, E., Gammack, S. M., Cresser, M. S., & Kiliham, K. (1992). Biology and Fertility of Soils Leaching of genetically modified Pseudomonas fluorescens through organic soils: Influence of temperature, soil pH, and roots. In *Biol Fertil Soils* (Vol. 13).

Knack, J. J., Wilcox, L. W., Delaux, P. M., AnÉ, J. M., Piotrowski, M. J., Cook, M. E., Graham, J. M., & Graham, L. E. (2015). Microbiomes of streptophyte algae and bryophytes suggest that a functional suite of microbiota fostered plant colonization of land. *International Journal of Plant Sciences*, *176*(5), 405–420. https://doi.org/10.1086/681161

Knudsen, G. R. (2010). Bacteriology of Soils and Plants. *Topley & Wilson's Microbiology and Microbial Infections*. https://doi.org/10.1002/9780470688618.TAW0008

Kogawa, M., Hosokawa, M., Nishikawa, Y., Mori, K., & Takeyama, H. (2018). Obtaining high-quality draft genomes from uncultured microbes by cleaning and co-assembly of single-cell amplified genomes. *Scientific Reports*, *8*(1), 2059. https://doi.org/10.1038/s41598-018-20384-3

Konstantinidis, K. T., Rosselló-Móra, R., & Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. In *ISME Journal* (Vol. 11, Issue 11, pp. 2399–2406). Nature Publishing Group. https://doi.org/10.1038/ismej.2017.113

Labonté, J. M., Swan, B. K., Poulos, B., Luo, H., Koren, S., Hallam, S. J., Sullivan, M. B., Woyke, T., Eric Wommack, K., & Stepanauskas, R. (2015). Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME Journal*, *9*(11), 2386–2399. https://doi.org/10.1038/ismej.2015.48

Ladau, J., & Eloe-Fadrosh, E. A. (2019). Spatial, Temporal, and Phylogenetic Scales of Microbial Ecology. In *Trends in Microbiology* (Vol. 27, Issue 8, pp. 662–669). Elsevier Ltd.

https://doi.org/10.1016/j.tim.2019.03.003

Lammel, Daniel R., Barth, G., Ovaskainen, O., Cruz, L. M., Zanatta, J. A., Ryo, M., de Souza, E. M., & Pedrosa, F. O. (2018). Direct and indirect effects of a pH gradient bring insights into the mechanisms driving prokaryotic community structures. *Microbiome*, *6*(1). https://doi.org/10.1186/s40168-018-0482-8

Lammel, Daniel Renato, Nüsslein, K., Tsai, S. M., & Cerri, C. C. (2015). Land use, soil and litter chemistry drive bacterial community structures in samples of the rainforest and Cerrado (Brazilian Savannah) biomes in Southern Amazonia. *European Journal of Soil Biology*, *66*, 32–39. https://doi.org/10.1016/J.EJSOBI.2014.11.001

Lan, F., Demaree, B., Ahmed, N., & Abate, A. R. (2017). Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nature Biotechnology*, *35*(7), 640–646. https://doi.org/10.1038/nbt.3880

Landry, Z., Swa, B. K., Herndl, G. J., Stepanauskas, R., & Giovannoni, S. J. (2017). SAR202 genomes from the dark ocean predict pathways for the oxidation of recalcitrant dissolved organic matter. *MBio*, *8*(2). https://doi.org/10.1128/MBIO.00413-17/SUPPL_FILE/MBO002173270S1.DOC

Langsrud, S., & Sundheim, G. (1996). Flow cytometry for rapid assessment of viability after exposure to a quaternary ammonium compound. *Journal of Applied Bacteriology*, *81*(4), 411–418. https://doi.org/10.1111/J.1365-2672.1996.TB03527.X

Lasken, R. S. (2007). Single-cell genomic sequencing using Multiple Displacement Amplification. *Current Opinion in Microbiology*, *10*(5), 510–516. https://doi.org/10.1016/J.MIB.2007.08.005

Lawson, C. E., Harcombe, W. R., Hatzenpichler, R., Lindemann, S. R., Löffler, F. E., O'Malley, M. A., García Martín, H., Pfleger, B. F., Raskin, L., Venturelli, O. S., Weissbrodt, D. G., Noguera, D. R., & McMahon, K. D. (2019). Common principles and best practices for engineering microbiomes. *Nature Reviews Microbiology*, *17*(12), 725–741. https://doi.org/10.1038/s41579-019-0255-9

Lebaron, P., Catala, P., & Parthuisot, N. (1998). Effectiveness of SYTOX green stain for bacterial viability assessment. *Applied and Environmental Microbiology*, *64*(7), 2697–2700. https://doi.org/10.1128/AEM.64.7.2697-2700.1998/ASSET/E4F8ACF0-F024-415E-944F-2241390067B6/ASSETS/GRAPHIC/AM0780074003.JPEG

Leducq, J. B., Sneddon, D., Santos, M., Condrain-Morel, D., Bourret, G., Martinez-Gomez, N. C., Lee, J. A., Foster, J. A., Stolyar, S., Shapiro, B. J., Kembel, S. W., Sullivan, J. M., & Marx, C. J. (2022). Comprehensive Phylogenomics of Methylobacterium Reveals Four Evolutionary Distinct Groups and Underappreciated Phyllosphere Diversity. *Genome Biology and Evolution*, *14*(8). https://doi.org/10.1093/GBE/EVAC123

Lee, S., Sieradzki, E. T., Nicol, G. W., & Hazard, C. (2022). Propagation of viral genomes by replicating ammonia-oxidising archaea during soil nitrification. *The ISME Journal 2022 17:2*, *17*(2), 309–314. https://doi.org/10.1038/s41396-022-01341-5

Leimbach, A., Hacker, J., & Dobrindt, U. (2013). E. coli as an all-rounder: The thin line between commensalism and pathogenicity. *Current Topics in Microbiology and*

Immunology, *358*, 3–32. https://doi.org/10.1007/82_2012_303/COVER

Lewis, W. H., & Ettema, T. J. G. (2019). Culturing the uncultured. *Nature Biotechnology*, *37*(11), 1278–1279. https://doi.org/10.1038/s41587-019-0300-2

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., & Ettema, T. J. G. (2021). Innovations to culturing the uncultured microbial majority. In *Nature Reviews Microbiology* (Vol. 19, Issue 4, pp. 225–240). Nature Research. https://doi.org/10.1038/s41579-020-00458-8

Li, X., Garbeva, P., Liu, X., klein Gunnewiek, P. J. A., Clocchiatti, A., Hundscheid, M. P. J., Wang, X., & de Boer, W. (2020). Volatile-mediated antagonism of soil bacterial communities against fungi. *Environmental Microbiology*, *22*(3), 1025–1035. https://doi.org/10.1111/1462-2920.14808

Li, Y., Chapman, S. J., Nicol, G. W., & Yao, H. (2018). Nitrification and nitrifiers in acidic soils. *Soil Biology and Biochemistry*, *116*, 290–301. https://doi.org/10.1016/J.SOILBIO.2017.10.023

Liu, Y., Jeraldo, P., Jang, J. S., Eckloff, B., Jen, J., & Walther-Antonio, M. (2019). Bacterial Single Cell Whole Transcriptome Amplification in Microfluidic Platform Shows Putative Gene Expression Heterogeneity. *Analytical Chemistry*, *91*(13), 8036–8044. https://doi.org/10.1021/acs.analchem.8b04773

Liu, Y., Schulze-Makuch, D., de Vera, J.-P., Cockell, C., Leya, T., Baqué, M., Walther-Antonio, M., Liu, Y., Schulze-Makuch, D., De Vera, J.-P., Cockell, C., Leya, T., Baqué, M., & Walther-Antonio, M. (2018). The Development of an Effective Bacterial Single-Cell Lysis Method Suitable for Whole Genome Amplification in Microfluidic Platforms. *Micromachines*, *9*(8), 367. https://doi.org/10.3390/mi9080367

Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J., & Crosby, L. (2018). Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *MSystems*, *3*(5). https://doi.org/10.1128/msystems.00055-18

Locey, K. J., & Lennon, J. T. (2016). *Scaling laws predict global microbial diversity*. *2016*, 30–35. https://doi.org/10.1073/pnas.1521291113

López-Escardó, D., Grau-Bové, X., Guillaumet-Adkins, A., Gut, M., Sieracki, M. E., & Ruiz-Trillo, I. (2017). Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate Monosiga brevicollis. *Scientific Reports*, *7*(1), 11025. https://doi.org/10.1038/s41598-017-11466-9

Lou, Y. C., Hoff, J., Olm, M. R., West-Roberts, J., Diamond, S., Firek, B. A., Morowitz, M. J., & Banfield, J. F. (2023). Using strain-resolved analysis to identify contamination in metagenomics data. *Microbiome*, *11*(1), 1–14. https://doi.org/10.1186/s40168-023-01477-2

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279. https://doi.org/10.1016/J.GPB.2016.05.004

Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E. E., Brochado, A. R., Fernandez, K. C., Dose, H., Mori, H., Patil, K. R., Bork, P., & Typas, A. (2018). Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature 2018 555:7698*,

555(7698), 623–628. https://doi.org/10.1038/nature25979

Malik, A. A., Puissant, J., Buckeridge, K. M., Goodall, T., Jehmlich, N., Chowdhury, S., Gweon, H. S., Peyton, J. M., Mason, K. E., van Agtmaal, M., Blaud, A., Clark, I. M., Whitaker, J., Pywell, R. F., Ostle, N., Gleixner, G., & Griffiths, R. I. (2018). Land use driven change in soil pH affects microbial carbon cycling processes. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-05980-1

Mangot, J. F., Logares, R., Sánchez, P., Latorre, F., Seeleuthner, Y., Mondy, S., Sieracki, M. E., Jaillon, O., Wincker, P., Vargas, C. De, & Massana, R. (2017). Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Scientific Reports 2017 7:1*, *7*(1), 1–12. https://doi.org/10.1038/srep41498

Marcy, Y., Ouverney, C., Bik, E. M., Lösekann, T., Ivanova, N., Martin, H. G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D. A., & Quake, S. R. (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(29), 11889–11894. https://doi.org/10.1073/pnas.0704662104

Marotz, C. A., Sanders, J. G., Zuniga, C., Zaramela, L. S., Knight, R., & Zengler, K. (2018). Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*, *6*(1), 42. https://doi.org/10.1186/s40168-018-0426-3

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/10.14806/EJ.17.1.200

Martinez-Garcia, M., Swan, B. K., Poulton, N. J., Gomez, M. L., Masland, D., Sieracki, M. E., & Stepanauskas, R. (2012). High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *The ISME Journal*, *6*(1), 113–123. https://doi.org/10.1038/ismej.2011.84

Martinez-Medina, M., Denizot, J., Dreux, N., Robin, F., Billard, E., Bonnet, R., Darfeuille-Michaud, A., & Barnich, N. (2014). Western diet induces dysbiosis with increased E coli in CEABAC10 mice, alters host barrier function favouring AIEC colonisation. *Gut*, *63*(1), 116–124. https://doi.org/10.1136/GUTJNL-2012-304119

Martinez-Rabert, E., Sloan, W. T., & Gonzalez-Cabaleiro, R. (2023). Multiscale models driving hypothesis and theory-based research in microbial ecology. *Interface Focus*, *13*(4). https://doi.org/10.1098/RSFS.2023.0008

Mas, A., Jamshidi, S., Lagadeuc, Y., Eveillard, D., & Vandenkoornhuyse, P. (2016). Beyond the Black Queen Hypothesis. *ISME Journal*, *10*(9), 2085–2091. https://doi.org/10.1038/ismej.2016.22

Mataigne, V., Vannier, N., Vandenkoornhuyse, P., & Hacquard, S. (2022). Multi-genome metabolic modeling predicts functional inter-dependencies in the Arabidopsis root microbiome. *Microbiome 2022 10:1*, *10*(1), 1–20. https://doi.org/10.1186/S40168-022-01383-Z

Mauger, S., Monard, C., Thion, C., & Vandenkoornhuyse, P. (2022). Contribution of single-cell omics to microbial ecology. In *Trends in Ecology and Evolution* (Vol. 37, Issue 1, pp. 67–78). Elsevier Ltd. https://doi.org/10.1016/j.tree.2021.09.002

McArdle, A. J., & Kaforou, M. (2020). Sensitivity of shotgun metagenomics to host DNA: abundance estimates depend on bioinformatic tools and contamination is the main issue. *Access Microbiology*, *2*(4). https://doi.org/10.1099/ACMI.0.000104

Mee, M. T., Collins, J. J., Church, G. M., & Wang, H. H. (2014). Syntrophic exchange in synthetic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(20), 2149–2156. https://doi.org/10.1073/pnas.1405641111

Meisner, A., Wepner, B., Kostic, T., van Overbeek, L. S., Bunthof, C. J., de Souza, R. S. C., Olivares, M., Sanz, Y., Lange, L., Fischer, D., Sessitsch, A., & Smidt, H. (2022). Calling for a systems approach in microbiome research and innovation. *Current Opinion in Biotechnology*, *73*, 171–178. https://doi.org/10.1016/j.copbio.2021.08.003

Mende, D. R., Aylward, F. O., Eppley, J. M., Nielsen, T. N., & DeLong, E. F. (2016). Improved Environmental Genomes via Integration of Metagenomic and Single-Cell Assemblies. *Frontiers in Microbiology*, *7*(FEB), 143. https://doi.org/10.3389/fmicb.2016.00143

Mendes, R., Kruijt, M., De Bruijn, I., Dekkers, E., Van Der Voort, M., Schneider, J. H. M., Piceno, Y. M., DeSantis, T. Z., Andersen, G. L., Bakker, P. A. H. M., & Raaijmakers, J. M. (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science*, *332*(6033), 1097–1100. https://doi.org/10.1126/science.1203980

Moran, N. A., & Sloan, D. B. (2015). The Hologenome Concept: Helpful or Hollow? *PLOS Biology*, *13*(12), e1002311. https://doi.org/10.1371/JOURNAL.PBIO.1002311

Morris, J. J., Lenski, R. E., & Zinser, E. R. (2012). The Black Queen Hypothesis : Evolution of Dependencies through Adaptative Gene Loss. *Mbio*, *3*(2), 1–7. https://doi.org/10.1128/mBio.00036-12.Copyright

Murray, A. E., Freudenstein, J., Gribaldo, S., Hatzenpichler, R., Hugenholtz, P., Kämpfer, P., Konstantinidis, K. T., Lane, C. E., Papke, R. T., Parks, D. H., Rossello-Mora, R., Stott, M. B., Sutcliffe, I. C., Thrash, J. C., Venter, S. N., Whitman, W. B., Acinas, S. G., Amann, R. I., Anantharaman, K., … Reysenbach, A. L. (2020). Roadmap for naming uncultivated Archaea and Bacteria. *Nature Microbiology*, *5*(8), 987–994. https://doi.org/10.1038/s41564-020-0733-x

Myers, C. R., & Nealson, K. H. (1988). Bacterial manganese reduction and growth with manganese oxide as the sole electron acceptor. *Science Science*, *240*(4857), 1319–1321. https://doi.org/10.1126/science.240.4857.1319

Nealson, K. H., & Hastings, J. W. (1979). Bacterial bioluminescence: Its control and ecological significance. *Microbiological Reviews*, *43*(4), 496–518. https://doi.org/10.1128/mmbr.43.4.496-518.1979

Nealson, K. H., Platt, T., & Hastings, J. W. (1970). Cellular control of the synthesis and activity of the bacterial luminescent system. *Journal of Bacteriology*, *104*(1), 313–322. https://doi.org/10.1128/jb.104.1.313-322.1970

Nealson, Kenneth H., & Hastings, J. W. (2006). Quorum sensing on a global scale: Massive numbers of bioluminescent bacteria make milky seas. *Applied and Environmental Microbiology*, *72*(4), 2295–2297. https://doi.org/10.1128/AEM.72.4.2295-2297.2006

Neina, D. (2019). The Role of Soil pH in Plant Nutrition and Soil Remediation. *Applied and Environmental Soil Science*, *2019*. https://doi.org/10.1155/2019/5794869

Neuenschwander, S. M., Ghai, R., Pernthaler, J., & Salcher, M. M. (2017). Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. *The ISME Journal 2018 12:1*, *12*(1), 185–198. https://doi.org/10.1038/ismej.2017.156

Niccum, B. A., Kastman, E. K., Kfoury, N., Robbat, A., & Wolfe, B. E. (2020). Strain-Level Diversity Impacts Cheese Rind Microbiome Assembly and Function. *MSystems*, *5*(3). https://doi.org/10.1128/msystems.00149-20

Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., & Pettersson, S. (2012). Host-Gut Microbiota Metabolic Interactions. *Science*, *336*(6086), 1262–1267. https://doi.org/10.1126/SCIENCE.1223813

Nishikawa, Y., Hosokawa, M., Maruyama, T., Yamagishi, K., Mori, T., & Takeyama, H. (2015). Monodisperse Picoliter Droplets for Low-Bias and Contamination-Free Reactions in Single-Cell Whole Genome Amplification. *PLOS ONE*, *10*(9), e0138733. https://doi.org/10.1371/JOURNAL.PONE.0138733

Nishikawa, Y., Kogawa, M., Hosokawa, M., Wagatsuma, R., Mineta, K., Takahashi, K., Ide, K., Yura, K., Behzad, H., Gojobori, T., & Takeyama, H. (2022). Validation of the application of gel beads-based single-cell genome sequencing platform to soil and seawater. *ISME Communications 2022 2:1*, *2*(1), 1–11. https://doi.org/10.1038/s43705-022-00179-4

Nunan, N., Schmidt, H., & Raynaud, X. (2020). The ecology of heterogeneity: Soil bacterial communities and C dynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1798). https://doi.org/10.1098/rstb.2019.0249

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(Database issue), D733. https://doi.org/10.1093/NAR/GKV1189

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., & Pop, M. (2018). Metagenomic assembly through the lens of validation: Recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in Bioinformatics*, *20*(4), 1140–1150. https://doi.org/10.1093/bib/bbx098

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/S13059-016-0997-X/FIGURES/5

Ouyang, Y., Chen, D., Fu, Y., Shi, W., Provin, T., Han, A., van Shaik, E., Samuel, J. E., de Figueiredo, P., Zhou, A., & Zhou, J. (2021). Direct cell extraction from fresh and stored soil samples: Impact on microbial viability and community compositions. *Soil Biology and Biochemistry*, *155*, 108178. https://doi.org/10.1016/J.SOILBIO.2021.108178

Pace, N. R., Sapp, J., & Goldenfeld, N. (2012). Phylogeny and beyond: Scientific, historical,

and conceptual significance of the first tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(4), 1011–1018. https://doi.org/10.1073/PNAS.1109716109/ASSET/BDFAC99D-F377-47A6-B343-7941F368EAB9/ASSETS/GRAPHIC/PNAS.1109716109FIG02.JPEG

Pachiadaki, M. G., Brown, J. M., Brown, J., Bezuidt, O., Berube, P. M., Biller, S. J., Poulton, N. J., Burkart, M. D., La Clair, J. J., Chisholm, S. W., & Stepanauskas, R. (2019a). Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell*, *179*(7), 1623-1635.e11. https://doi.org/10.1016/j.cell.2019.11.017

Pachiadaki, M. G., Brown, J. M., Brown, J., Bezuidt, O., Berube, P. M., Biller, S. J., Poulton, N. J., Burkart, M. D., La Clair, J. J., Chisholm, S. W., & Stepanauskas, R. (2019b). Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell*, *179*(7), 1623-1635.e11. https://doi.org/10.1016/j.cell.2019.11.017

Paneth, N., Vinten-Johansen, P., Brody, H., & Rip, M. (1998). A rivalry of foulness: Official and unofficial investigations of the London cholera epidemic of 1854. *American Journal of Public Health*, *88*(10), 1545–1553. https://doi.org/10.2105/AJPH.88.10.1545

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P. A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology 2018 36:10*, *36*(10), 996–1004. https://doi.org/10.1038/NBT.4229

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Pascual-García, A., Bonhoeffer, S., & Bell, T. (2020). Metabolically cohesive microbial consortia and ecosystem functioning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1798). https://doi.org/10.1098/rstb.2019.0245

Pfeiffer, R. (1892). Vorläufige Mittheilungen über die Erreger der Influenza. *DMW - Deutsche Medizinische Wochenschrift*, *18*(02), 28–28. https://doi.org/10.1055/S-0029-1198870

Pierce, J. V., & Bernstein, H. D. (2016). Genomic Diversity of Enterotoxigenic Strains of Bacteroides fragilis. *PLOS ONE*, *11*(6), e0158171. https://doi.org/10.1371/JOURNAL.PONE.0158171

Polz, M. F., Alm, E. J., & Hanage, W. P. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics*, *29*(3), 170–175. https://doi.org/10.1016/j.tig.2012.12.006

Prosser, J. I. (2015). *Dispersing misconceptions and identifying opportunities for the use of "omics" in soil microbial ecology*. https://doi.org/10.1038/nrmicro3468

Prosser, J. I. (2020). Putting science back into microbial ecology: a question of approach. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1798), 20190240. https://doi.org/10.1098/rstb.2019.0240

Prosser, J. I. (2022). How and why in microbial ecology: An appeal for scientific aims, questions, hypotheses and theories. *Environmental Microbiology*, *24*(11), 4973–4980.

https://doi.org/10.1111/1462-2920.16221

Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., Green, J. L., Green, L. E., Killham, K., Lennon, J. J., Osborn, A. M., Solan, M., van der Gast, C. J., & Young, J. P. W. (2007). The role of ecological theory in microbial ecology. *Nature Reviews Microbiology*, *5*(5), 384–392. https://doi.org/10.1038/nrmicro1643

Prosser, J. I., Hink, L., Gubry-Rangin, C., & Nicol, G. W. (2020). Nitrous oxide production by ammonia oxidizers: Physiological diversity, niche differentiation and potential mitigation strategies. *Global Change Biology*, *26*(1), 103–118. https://doi.org/10.1111/GCB.14877

Prosser, J. I., & Martiny, J. B. H. (2020). Conceptual challenges in microbial community ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1798), 2–4. https://doi.org/10.1098/rstb.2019.0241

Prosser, J. I., & Raaijmakers, J. M. (2020). *Correspondence Towards meaningful scales in ecosystem microbiome research*. *October*. https://doi.org/10.1111/1462-2920.15276

Raddadi, N., Giacomucci, L., Marasco, R., Daffonchio, D., Cherif, A., & Fava, F. (2018). Bacterial polyextremotolerant bioemulsifiers from arid soils improve water retention capacity and humidity uptake in sandy soil. *Microbial Cell Factories*, *17*(1), 1–12. https://doi.org/10.1186/s12934-018-0934-7

Raghunathan, A., Ferguson, H. R., Bornarth, C. J., Song, W., Driscoll, M., & Lasken, R. S. (2005). Genomic DNA amplification from a single bacterium. *Applied and Environmental Microbiology*, *71*(6), 3342–3347. https://doi.org/10.1128/AEM.71.6.3342-3347.2005

Rainey, P. B., & Quistad, S. D. (2020). Toward a dynamical understanding of microbial communities. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *375*(1798), 20190248. https://doi.org/10.1098/rstb.2019.0248

Ramoneda, J., Stallard-olivera, E., Hoffert, M., Winfrey, C. C., Stadler, M., Fierer, N., Biology, E., Biologiques, S., & Centre-ville, S. (2023). *Building a genome-based understanding of bacterial pH preferences*. *8998*(12). https://doi.org/10.1126/sciadv.adf8998

Reese, A. T., & Kearney, S. M. (2019). Incorporating functional trade-offs into studies of the gut microbiota. In *Current Opinion in Microbiology* (Vol. 50, pp. 20–27). Elsevier Ltd. https://doi.org/10.1016/j.mib.2019.09.003

Rezaeinejad, S., & Ivanov, V. (2011). Heterogeneity of Escherichia coli population by respiratory activity and membrane potential of cells during growth and long-term starvation. *Microbiological Research*, *166*(2), 129–135. https://doi.org/10.1016/J.MICRES.2010.01.007

Rigottier-Gois, L. (2013). Dysbiosis in inflammatory bowel diseases: The oxygen hypothesis. *ISME Journal*, *7*(7), 1256–1261. https://doi.org/10.1038/ismej.2013.80

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W. T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437. https://doi.org/10.1038/nature12352

Ritpitakphong, U., Falquet, L., Vimoltust, A., Berger, A., Métraux, J. P., & L'Haridon, F. (2016). The microbiome of the leaf surface of Arabidopsis protects against a fungal pathogen. *New Phytologist*, *210*(3), 1033–1043. https://doi.org/10.1111/nph.13808

Ross, B. N., & Whiteley, M. (2020). Ignoring social distancing: advances in understanding multi-species bacterial interactions. *Faculty Reviews*, *9*(23). https://doi.org/10.12703/r/9-23

Rousk, J., Bååth, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., Knight, R., & Fierer, N. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME Journal*, *4*(10), 1340–1351. https://doi.org/10.1038/ismej.2010.58

Rousk, J., & Bengtson, P. (2014). Microbial regulation of global biogeochemical cycles. *Frontiers in Microbiology*, *5*(MAR), 103. https://doi.org/10.3389/fmicb.2014.00103

Roux, S., Hawley, A. K., Beltran, M. T., Scofeld, M., Schwientek, P., Stepanauskas, R., Woyke, T., Hallam, S. J., & Sullivan, M. B. (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *ELife*, *2014*(3). https://doi.org/10.7554/eLife.03125.001

Roux, S., Hawley, A. K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R., Woyke, T., Hallam, S. J., & Sullivan, M. B. (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *ELife*, *3*. https://doi.org/10.7554/elife.03125

Rütting, T., Schleusner, P., Hink, L., & Prosser, J. I. (2021). The contribution of ammonia-oxidizing archaea and bacteria to gross nitrification under different substrate availability. *Soil Biology and Biochemistry*, *160*, 108353. https://doi.org/10.1016/J.SOILBIO.2021.108353

Sanchez, A., & Gore, J. (2013). Feedback between Population and Evolutionary Dynamics Determines the Fate of Social Microbial Populations. *PLoS Biology*, *11*(4), e1001547. https://doi.org/10.1371/journal.pbio.1001547

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sangwan, N., Xia, F., & Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. In *Microbiome* (Vol. 4, Issue 1, p. 8). BioMed Central Ltd. https://doi.org/10.1186/s40168-016-0154-5

Savage, D. C. (2003). MICROBIAL ECOLOGY OF THE GASTROINTESTINAL TRACT. *Https://Doi.Org/10.1146/Annurev.Mi.31.100177.000543*, *31*, 107–133. https://doi.org/10.1146/ANNUREV.MI.31.100177.000543

Scanlan, P. D., & Buckling, A. (2012). Co-evolution with lytic phage selects for the mucoid phenotype of Pseudomonas fluorescens SBW25. *ISME Journal*, *6*(6), 1148–1158. https://doi.org/10.1038/ismej.2011.174

Schaule, G., Flemming, H. C., & Ridgway, H. F. (1993). Use of 5-cyano-2,3-ditolyl tetrazolium chloride for quantifying planktonic and sessile respiring bacteria in drinking water. *Applied and Environmental Microbiology*, *59*(11), 3850–3857.

https://doi.org/10.1128/AEM.59.11.3850-3857.1993

Schlatter, D. C., Kahl, K., Carlson, B., Huggins, D. R., & Paulitz, T. (2020). Soil acidification modifies soil depth-microbiome relationships in a no-till wheat cropping system. *Soil Biology and Biochemistry*, *149*. https://doi.org/10.1016/j.soilbio.2020.107939

Schönknecht, G., Chen, W.-H., Ternes, C. M., Barbier, G. G., Shrestha, R. P., Stanke, M., Bräutigam, A., Baker, B. J., Banfield, J. F., Garavito, R. M., Carr, K., Wilkerson, C., Rensing, S. A., Gagneul, D., Dickenson, N. E., Oesterhelt, C., Lercher, M. J., & Weber, A. P. M. (2013). Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science (New York, N.Y.)*, *339*(6124), 1207–1210. https://doi.org/10.1126/science.1231707

Sender, R., Fuchs, S., & Milo, R. (2016). Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell*, *164*(3), 337–340. https://doi.org/10.1016/J.CELL.2016.01.013

Shade, A., Jones, S. E., Gregory Caporaso, J., Handelsman, J., Knight, R., Fierer, N., & Gilbert, J. A. (2014). Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *MBio*, *5*(4), 1371–1385. https://doi.org/10.1128/mBio.01371-14

Shaiber, A., & Eren, A. M. (2019). Composite metagenome-assembled genomes reduce the quality of public genome repositories. *MBio*, *10*(3). https://doi.org/10.1128/MBIO.00725-19/ASSET/7F6BB782-23B4-49A7-9CE7-CC66BE9AD72C/ASSETS/GRAPHIC/MBIO.00725-19-F0001.JPEG

Sharon, G., Garg, N., Debelius, J., Knight, R., Dorrestein, P. C., & Mazmanian, S. K. (2014). Specialized metabolites from the microbiome in health and disease. *Cell Metabolism*, *20*(5), 719–730. https://doi.org/10.1016/j.cmet.2014.10.016

Shen, W., & Ren, H. (2021). TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics*, *48*(9), 844–850. https://doi.org/10.1016/J.JGG.2021.03.006

Smith, S. E., Huang, W., Tiamani, K., Unterer, M., Khan Mirzaei, M., & Deng, L. (2022). Emerging technologies in the study of the virome. *Current Opinion in Virology*, *54*, 101231. https://doi.org/10.1016/J.COVIRO.2022.101231

Sobol, M. S. (2023). *Back to Basics : A Simplified Improvement to Multiple Displacement Amplification for Microbial Single- Cell Genomics*.

Song, Y., Xu, X., Wang, W., Tian, T., Zhu, Z., & Yang, C. (2019). Single cell transcriptomics: moving towards multi-omics. *The Analyst*, *144*(10), 3172–3189. https://doi.org/10.1039/C8AN01852A

Steen, A. D., Crits-Christoph, A., Carini, P., DeAngelis, K. M., Fierer, N., Lloyd, K. G., & Cameron Thrash, J. (2019). High proportions of bacteria and archaea across most biomes remain uncultured. *ISME Journal*, *13*(12), 3126–3130. https://doi.org/10.1038/s41396-019-0484-y

Stepanauskas, R., Fergusson, E. A., Brown, J., Poulton, N. J., Tupper, B., Labonté, J. M., Becraft, E. D., Brown, J. M., Pachiadaki, M. G., Povilaitis, T., Thompson, B. P., Mascena, C. J., Bellows, W. K., & Lubys, A. (2017). Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nature*

Communications, *8*(1), 1–10. https://doi.org/10.1038/s41467-017-00128-z

Stubbendieck, R. M., Vargas-Bautista, C., & Straight, P. D. (2016). Bacterial communities: Interactions to scale. *Frontiers in Microbiology*, *7*(AUG), 1–19. https://doi.org/10.3389/fmicb.2016.01234

Swan, B. K., Martinez-Garcia, M., Preston, C. M., Sczyrba, A., Woyke, T., Lamy, D., Reinthaler, T., Poulton, N. J., Masland, E. D. P., Gomez, M. L., Sieracki, M. E., DeLong, E. F., Herndl, G. J., & Stepanauskas, R. (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science (New York, N.Y.)*, *333*(6047), 1296–1300. https://doi.org/10.1126/science.1203690

Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., González, J. M., Luo, H., Wright, J. J., Landry, Z. C., Hanson, N. W., Thompson, B. P., Poulton, N. J., Schwientek, P., Acinas, S. G., Giovannoni, S. J., Moran, M. A., Hallam, S. J., Cavicchioli, R., Woyke, T., & Stepanauskas, R. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences*, *110*(28), 11463–11468. https://doi.org/10.1073/PNAS.1304246110

Tahon, G., Geesink, P., & Ettema, T. J. G. (2021). Expanding Archaeal Diversity and Phylogeny: Past, Present, and Future. *Annual Review of Microbiology*, *75*, 359–381. https://doi.org/10.1146/annurev-micro-040921-050212

Tang, Q., Huang, J., Zhang, S., Qin, H., Dong, Y., Wang, C., Li, D., & Zhou, R. (2022). Keystone microbes affect the evolution and ecological coexistence of the community via species/strain specificity. *Journal of Applied Microbiology*, *132*(2), 1227–1238. https://doi.org/10.1111/jam.15255

Tatsumi, C., Taniguchi, T., Du, S., Yamanaka, N., & Tateno, R. (2020). Soil nitrogen cycling is determined by the competition between mycorrhiza and ammonia-oxidizing prokaryotes. *Ecology*, *101*(3), e02963. https://doi.org/10.1002/ECY.2963

Theis, K. R., Dheilly, N. M., Klassen, J. L., Brucker, R. M., Baines, J. F., Bosch, T. C. G., Cryan, J. F., Gilbert, S. F., Goodnight, C. J., Lloyd, E. A., Sapp, J., Vandenkoornhuyse, P., Zilber-Rosenberg, I., Rosenberg, E., & Bordenstein, S. R. (2016). Getting the Hologenome Concept Right: an Eco-Evolutionary Framework for Hosts and Their Microbiomes. *MSystems*, *1*(2). https://doi.org/10.1128/msystems.00028-16

Tripathi, A., Marotz, C., Gonzalez, A., Vázquez-Baeza, Y., Song, S. J., Bouslimani, A., McDonald, D., Zhu, Q., Sanders, J. G., Smarr, L., Dorrestein, P. C., & Knight, R. (2018). Are microbiome studies ready for hypothesis-driven research? In *Current Opinion in Microbiology* (Vol. 44, pp. 61–69). Elsevier Ltd. https://doi.org/10.1016/j.mib.2018.07.002

Tripathi, B. M., Stegen, J. C., Kim, M., Dong, K., Adams, J. M., & Lee, Y. K. (2018). Soil pH mediates the balance between stochastic and deterministic assembly of bacteria. *ISME Journal*, *12*(4), 1072–1083. https://doi.org/10.1038/s41396-018-0082-4

Triplett, E. W., & Sadowsky, M. J. (2003). GENETICS OF COMPETITION FOR NODULATION OF LEGUMES. *Https://Doi.Org/10.1146/Annurev.Mi.46.100192.002151*, *46*(1), 399–422. https://doi.org/10.1146/ANNUREV.MI.46.100192.002151

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature 2003 428:6978*, *428*(6978), 37–43. https://doi.org/10.1038/nature02340

Van Der Heijden, M. G. A., Bruin, S. De, Luckerhoff, L., Van Logtestijn, R. S. P., & Schlaeppi, K. (2016). A widespread plant-fungal-bacterial symbiosis promotes plant biodiversity, plant nutrition and seedling recruitment. *ISME Journal*, *10*(2), 389–399. https://doi.org/10.1038/ismej.2015.120

Van Rossum, T., Ferretti, P., Maistrenko, O. M., & Bork, P. (2020). Diversity within species: interpreting strains in microbiomes. In *Nature Reviews Microbiology* (Vol. 18, Issue 9, pp. 491–506). Nature Research. https://doi.org/10.1038/s41579-020-0368-1

Vandenkoornhuyse, P., Dufresne, A., Quaiser, A., Gouesbet, G., Binet, F., Francez, A. J., Mahé, S., Bormans, M., Lagadeuc, Y., & Couée, I. (2010). Integration of molecular functions at the ecosystemic level: Breakthroughs and future goals of environmental genomics and post-genomics. *Ecology Letters*, *13*(6), 776–791. https://doi.org/10.1111/j.1461-0248.2010.01464.x

Vandenkoornhuyse, P., Quaiser, A., Duhamel, M., Le Van, A., & Dufresne, A. (2015). The importance of the microbiome of the plant holobiont. In *New Phytologist* (Vol. 206, Issue 4, pp. 1196–1206). Blackwell Publishing Ltd. https://doi.org/10.1111/nph.13312

Vannier, N., Agler, M., & Hacquard, S. (2019). Microbiota-mediated disease resistance in plants. *PLOS Pathogens*, *15*(6), e1007740. https://doi.org/10.1371/JOURNAL.PPAT.1007740

Vollmers, J., Wiegand, S., Lenk, F., & Kaster, A. K. (2022). How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner. *Nucleic Acids Research*, *50*(13), e76–e76. https://doi.org/10.1093/NAR/GKAC294

Wan, W., Tan, J., Wang, Y., Qin, Y., He, H., Wu, H., Zuo, W., & He, D. (2020). Responses of the rhizosphere bacterial community in acidic crop soil to pH: Changes in diversity, composition, interaction, and function. *Science of the Total Environment*, *700*. https://doi.org/10.1016/j.scitotenv.2019.134418

Wang, B., Yeun, L. H., Xue, J. Y., Liu, Y., Ané, J. M., & Qiu, Y. L. (2010). Presence of three mycorrhizal genes in the common ancestor of land plants suggests a key role of mycorrhizas in the colonization of land by plants. *New Phytologist*, *186*(2), 514–525. https://doi.org/10.1111/j.1469-8137.2009.03137.x

Wang, J., Mei, X., Wei, Z., Raza, W., & Shen, Q. (2021). Effect of bacterial intra-species community interactions on the production and activity of volatile organic compounds. *Soil Ecology Letters*, *3*(1), 32–41. https://doi.org/10.1007/s42832-020-0054-2

Wang, P., Wang, X., Nie, J., Wang, Y., Zang, H., Peixoto, L., Yang, Y., & Zeng, Z. (2022). Manure Application Increases Soil Bacterial and Fungal Network Complexity and Alters Keystone Taxa. *Journal of Soil Science and Plant Nutrition*, *22*(1), 607–618. https://doi.org/10.1007/S42729-021-00673-Z/METRICS

Wee, W. Y., Dutta, A., & Choo, S. W. (2017). Comparative genome analyses of mycobacteria give better insights into their evolution. *PLOS ONE*, *12*(3), e0172831. https://doi.org/10.1371/JOURNAL.PONE.0172831

Weitz, J. S., & Wilhelm, S. W. (2012). Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biology Reports*, *4*(1), 17. https://doi.org/10.3410/B4-17

Williams, T. A., Foster, P. G., Cox, C. J., & Embley, T. M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, *504*(7479), 231–236. https://doi.org/10.1038/nature12779

Williams, T. J., Allen, M. A., Panwar, P., & Cavicchioli, R. (2022). Into the darkness: the ecologies of novel 'microbial dark matter' phyla in an Antarctic lake. *Environmental Microbiology*, *24*(5), 2576–2603. https://doi.org/10.1111/1462-2920.16026

Wilpiszeski, R. L., Gionfriddo, C. M., Wymore, A. M., Moon, J. W., Lowe, K. A., Podar, M., Rafie, S., Fields, M. W., Hazen, T. C., Ge, X., Poole, F., Adams, M. W. W., Chakraborty, R., Fan, Y., van Nostrand, J. D., Zhou, J., Arkin, A. P., & Elias, D. A. (2020). In-field bioreactors demonstrate dynamic shifts in microbial communities in response to geochemical perturbations. *PLOS ONE*, *15*(9), e0232437. https://doi.org/10.1371/JOURNAL.PONE.0232437

Wilson, K. H., & Blitchington, R. B. (1996). Human colonic biota studied by ribosomal DNA sequence analysis. *Applied and Environmental Microbiology*, *62*(7), 2273–2278. https://doi.org/10.1128/aem.62.7.2273-2278.1996

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090. https://doi.org/10.1073/PNAS.74.11.5088

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 1–13. https://doi.org/10.1186/S13059-019-1891-0/FIGURES/2

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A Primer on Metagenomics. *PLOS Computational Biology*, *6*(2), e1000667. https://doi.org/10.1371/JOURNAL.PCBI.1000667

Woyke, T., Doud, D. F. R., & Schulz, F. (2017). The trajectory of microbial single-cell sequencing. In *Nature Methods* (Vol. 14, Issue 11, pp. 1045–1054). Nature Publishing Group. https://doi.org/10.1038/nmeth.4469

Woyke, T., Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., Malmstrom, R., Stepanauskas, R., & Cheng, J. F. (2011). Decontamination of MDA reagents for single cell whole genome amplification. *PLoS ONE*, *6*(10). https://doi.org/10.1371/JOURNAL.PONE.0026161

Woyke, T., Tighe, D., Mavromatis, K., Clum, A., Copeland, A., Schackwitz, W., Lapidus, A., Wu, D., Mccutcheon, J. P., Mcdonald, B. R., Moran, N. A., Bristow, J., & Cheng, J. F. (2010). One Bacterial Cell, One Complete Genome. *PLOS ONE*, *5*(4), e10314. https://doi.org/10.1371/JOURNAL.PONE.0010314

Wu, C., Yan, B., Wei, F., Wang, H., Gao, L., Ma, H., Liu, Q., Liu, Y., Liu, G., & Wang, G. (2023).

Long-term application of nitrogen and phosphorus fertilizers changes the process of community construction by affecting keystone species of crop rhizosphere microorganisms. *Science of The Total Environment*, 165239. https://doi.org/10.1016/J.SCITOTENV.2023.165239

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M., Tindall, B. J., Hooper, S. D., Pati, A., Lykidis, A., Spring, S., Anderson, I. J., Dhaeseleer, P., Zemla, A., Singer, M., Lapidus, A., … Eisen, J. A. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, *462*(7276), 1056–1060. https://doi.org/10.1038/nature08656

Xu, Q., Ling, N., Quaiser, A., Guo, J., Ruan, J., Guo, S., Shen, Q., & Vandenkoornhuyse, P. (2021). Rare Bacterial Assembly in Soils Is Mainly Driven by Deterministic Processes. *Microbial Ecology*. https://doi.org/10.1007/s00248-021-01741-8

Xu, R., Sun, X., Häggblom, M. M., Dong, Y., Zhang, M., Yang, Z., Xiao, E., Xiao, T., Gao, P., Li, B., & Sun, W. (2022). Metabolic potentials of members of the class Acidobacteriia in metal-contaminated soils revealed by metagenomic analysis. *Environmental Microbiology*, *24*(2), 803–818. https://doi.org/10.1111/1462-2920.15612

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., Whitman, W. B., Euzéby, J., Amann, R., & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *NATURE REVIEWS | MICROBIOLOGY*, *12*, 635. https://doi.org/10.1038/nrmicro3330

Yavitt, J. B., Roco, C. A., Debenport, S. J., Barnett, S. E., & Shapleigh, J. P. (2021). Community Organization and Metagenomics of Bacterial Assemblages Across Local Scale pH Gradients in Northern Forest Soils. *Microbial Ecology*, *81*(3), 758–769. https://doi.org/10.1007/s00248-020-01613-7

Yin, Y., Jiang, Y., Lam, K. W. G., Berletch, J. B., Disteche, C. M., Noble, W. S., Steemers, F. J., Camerini-Otero, R. D., Adey, A. C., & Shendure, J. (2019). High-Throughput Single-Cell Sequencing with Linear Amplification. *Molecular Cell*, *76*(4), 676-690.e10. https://doi.org/10.1016/j.molcel.2019.08.002

Yu, F. B., Blainey, P. C., Schulz, F., Woyke, T., Horowitz, M. A., & Quake, S. R. (2017). Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *ELife*, *6*. https://doi.org/10.7554/eLife.26580

Yuval, B. (2017). Symbiosis: Gut Bacteria Manipulate Host Behaviour. *Current Biology*, *27*(15), R746–R747. https://doi.org/10.1016/j.cub.2017.06.050

Zamkovaya, T., Foster, J. S., de Crécy-Lagard, V., & Conesa, A. (2021). A network approach to elucidate and prioritize microbial dark matter in microbial communities. *ISME Journal*, *15*(1), 228–244. https://doi.org/10.1038/s41396-020-00777-x

Zhalnina, K., Dias, R., de Quadros, P. D., Davis-Richardson, A., Camargo, F. A. O., Clark, I. M., McGrath, S. P., Hirsch, P. R., & Triplett, E. W. (2015). Soil pH Determines Microbial Diversity and Composition in the Park Grass Experiment. *Microbial Ecology*, *69*(2), 395–406. https://doi.org/10.1007/s00248-014-0530-2

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., & Wang, J. (2019). Comparative

Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell*, *73*(1), 130-142.e5. https://doi.org/10.1016/j.molcel.2018.10.020

Zhang, Z., Wang, J., Wang, J., Wang, J., & Li, Y. (2020). *Estimate of the sequenced proportion of the global prokaryotic genome*. 1–9.

Zheng, W., Zhao, S., Yin, Y., Zhang, H., Needham, D. M., Evans, E. D., Dai, C. L., Lu, P. J., Alm, E. J., & Weitz, D. A. (2022). High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome. *Science*, *376*(6597). https://doi.org/10.1126/science.abm1483

COLLEGE ECOLOGIE
DOCTORAL GEOSCIENCES
BRETAGNE AGRONOMIE ALIMENTATION

Université de Rennes

Titre : Développement d'une approche innovante de génomique sur cellules uniques et application aux communautés bactériennes du sol

Mots clés : Ecologie microbienne, génomique, cellule unique

Résumé : Le séquençage du génome entier d'une seule cellule (scWGS) pour étudier les bactéries s'est développé ces dernières années. Les microbiologistes sont particulièrement intéressés par cette approche pour accéder à l'échelle de la population de l'organisation des bactéries et évaluer le potentiel d'évolution et d'interaction de ces structures bactériennes. Le scWGS devrait également accélérer la découverte de bactéries inconnues et fournir un contenu génomique plus précis que la métagénomique traditionnellement utilisée, qui n'est pas adaptée à la description des communautés bactériennes à des échelles organisationnelles fines. Dans la pratique, les scWGS sont peu utilisés en raison de leurs limites en termes de coût, d'équipement nécessaire, de temps de manipulation long et de génomes récupérés partiels et contaminés. J'ai développé une préparation de librairies génomique unicellulaire afin de proposer une approche facile à utiliser avec un coût limité et discute ses possibles améliorations futures. Pour compenser le manque de procédures de décontamination universelles pour de tels jeux de données, un pipeline de décontamination automatisé a été développé et permet l'unification du traitement des données unicellulaires. Je démontre, sur des souches bactériennes pures et environnementales, la nécessité d'une procédure de décontamination systématique et souligne les avantages du scWGS par rapport à la métagénomique. Enfin, je discute des perspectives techniques et écologiques que cette approche a à offrir à la microbiologie.

Title: Elaboration of an innovative single-cell genomics approach and application to soil bacterial communities

Keywords: Microbial ecology, genomics, single-cell

Abstract: The development of single-cell whole genome sequencing (scWGS) to study bacteria has grown in recent years. Microbiologists are particularly interested in this approach to access the population scale of bacteria organisation and evaluate the potential of these bacterial structures to evolve and interact. scWGS is also expected to accelerate the discovery of unknown bacteria and to provide more accurate genome content than the traditionally used metagenomics which is not suited for bacterial community description at fine organisational scales. In practice, scWGS are timidly used for their limitations regarding their cost, equipment requirements, long handling time, and partial and contaminated recovered genomes. Here, I developed an improved single-cell genomic library preparation to propose an easy-to-use approach with limited cost and discuss its possible future improvements. To compensate for the lack of universal decontamination procedures for such datasets, an automated decontamination pipeline was developed and allows the unification of single-cell data handling. I demonstrate, on pure and environmental bacterial strains, the necessity for systematic decontamination procedure and highlight the advantages of scWGS compared to metagenomics. Finally, I discuss the technical and ecological perspectives that single-cell omics have to offer to microbiology.