“ GeOMICS ”, nouveaux concepts de bioinformatique pour un nouvel outil de diagnostic environnemental basé sur l’alliance de la géochimie et des omiques

Virginie Jouffret

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Bioinformatique

École doctorale : Sciences Chimiques et Biologiques pour la Santé (CBS2 ED n°168)

Unité de recherche : Laboratoire d'Innovations technologiques pour la Détection et le Diagnostic (Li2D)

« GeOMICS », nouveaux concepts de bioinformatique pour un nouvel outil de diagnostic environnemental basé sur l'alliance de la géochimie et des omiques

Présentée par Virginie Jouffret
Le 30 novembre 2022

Sous la direction de Jean Armengaud
et Sophie Ayrault et encadrée par Olivier Pible

Devant le jury composé de

Mme Guillermina HERNANDEZ-RAQUET, Directrice de Recherche, INRAE Toulouse	Présidente du jury
Mme Fabienne BATTAGLIA-BRUNET, Directrice de Recherche, BRGM Orléans	Rapporteur
M. Jean ARMENGAUD, Directeur de Recherche, CEA Marcoule	Directeur de thèse
Mme Sophie AYRAULT, Directrice de Recherche, CEA Saclay	Co-Directrice de thèse
M. Olivier PIBLE, Ingénieur de Recherche, CEA Marcoule	Invité
M. Christophe HIRTZ, Professeur, CHU de Montpellier	Invité

UNIVERSITÉ DE MONTPELLIER

« The more we learn the more we realize how little we know. »

Richard Buckminster Fuller

Sommaire_Toc129294603

Remerciements

Je voudrais tout d'abord adresser mes remerciements aux personnes qui ont accepté de rapporter ma thèse. Merci Mme Fabienne Battaglia, directrice de recherche et experte au BRGM, d'apporter son expertise en microbiologie des sols pollués notamment par les éléments traces métalliques dans un contexte minier, de remédiation des sols, de pollution des sédiments et des traitements des eaux. Merci Mme Guillermina Hernandez-Raquet, directrice de recherche INRAE, d'apporter son expertise sur la biodégradation de composés organiques par des communautés microbiennes qui utilise notamment des approches méta-omiques et notamment métaprotéomique. Ce fut un plaisir de vous rencontrer au congrès international de métaprotéomique qui s'est tenu il y a un an au Luxembourg. Merci Mr Christophe Hirtz, professeur des universités, de nous honorer de votre présence en tant que spécialiste en protéomique et métaprotéomique clinique. J'ai eu l'occasion de participer au pipeline d'analyse qui a servi à identifier les microorganismes associés aux colorations dentaires.

Je tiens également à remercier mes directeurs de thèse Mme Sophie Ayrault et Mr Jean Armengaud pour m'avoir accompagné le long de ces quatre années de thèse.

Merci Jean pour tes multiples relectures ainsi que ton soutien dans la rédaction du manuscrit. Merci de m'avoir accueilli pendant ces six années au sein du laboratoire et de m'avoir donné l'occasion de mener à bien ce projet de thèse.

Merci Sophie pour tes conseils et ta bonne humeur, merci pour nos échanges sur la géochimie, et notamment pour toutes ces petites histoires autour de Notre-Dame de Paris.

Je remercie également Olivier, mon encadrant de thèse, qui m'a fait découvrir la protéomique d'un point de vue bioinformatique, la détection de micro-organismes dans un contexte NRBC dans un premier temps puis métaprotéomique des sols dans un second. Merci de m'avoir fait confiance au long de toutes ces années dans le développement du pipeline bioinformatique. Merci pour tes conseils et tes retours enrichissants sur l'analyse des données ainsi que ton aide pour l'installation des outils sous Windows.

Je souhaite aussi remercier l'équipe d'Olivier et Sophie pour leur accueil très chaleureux au LSCE durant cette journée passée à Saclay. Cette journée m'a permis de rencontrer l'équipe et de découvrir les laboratoires de géochimie. Une pensée va à Matthieu Roy-Barman et sa famille qui nous a quitté en mars dernier. Il avait participé à la journée de carottage en 2018.

Je souhaite remercier l'équipe de protéomique qui m'a accueillie il y a six ans alors que je sortais tout juste de master. Merci Jean-Charles alias JC pour m'avoir expliqué le fonctionnement du spectromètre de masse, pour tous ces moments autour d'un café à discuter de la vie et de tout pleins de choses. Merci d'avoir testé les outils et d'avoir trouvé tous les bugs improbables tels que l'ajout d'accents dans les bases de données en francisant tous les mots anglais (je me souviendrais de « ble » transformé en blé dans les fichiers FASTA). Merci Guylaine qui a dû réaliser toutes les manips au labo pour générer les données du projet de thèse et qui m'a appris à faire une manip (une seule c'est bien) ! Merci Gérard, maintenant à la retraite, pour les petits projets d'analyses et toutes tes petites blagues. Merci Lucia et Clément. Merci Béatrice pour ces moments de cultures "jeuns", de littérature, de discussions autour de l'actualité.

Merci aux anciens du bureau, Charlotte Mappa, Karim Hayoun, Duarte Gouveia, Cédric Pisani et Karen Culotta pour votre bonne humeur, vos conseils, pour tous ces moments de vie au quotidien. Merci

Charlotte de m'avoir fait découvrir des hits musicaux incontournables français. Merci Karim pour ta gentillesse et tous ces moments également !

Karen, ma partenaire bioinfo durant 5 ans avec qui j'ai partagé le bureau, pour m'avoir fait aimer la salle de sport à y aller presque 15h par semaine (0h aujourd'hui). Pour tous les moments de SAV, les débats bioinfos de débuggage, de développement, les pauses clopes avec les filles à parler de la vie !

Bonne chance à Pauline et Madison qui sont sur la route du doctorat, à Mélodie et aux nouveaux de l'équipe de protéomique.

Merci à toutes les personnes du laboratoire, Laurent, Joëlle, Alicia, Alexia, Anastasia, Fabrice, Fabienne, Gauthier, Yannick, Pascale, Stéphanie, Virginie et Anne ; aux anciens permanents Nicole et Charles à la retraites et aux anciens collègues qui ont suivi leurs routes, Niza, Charlotte Foissard, Basile, Emilie & Rémy, Esther, Florent, Hélène, Marie-Anne, Martine, Noémie.

Merci à Lola Reynaud et Valérie Chazel, pour leur patience et leur enthousiasme à réaliser une vidéo portant sur mon projet de recherche présenté au GYSS et rediffusée sur la chaîne Youtube CEA.

Merci Céline et Laëtitia pour tous ces moments avec Christine et Karen. Merci Céline pour m'avoir fait découvrir une autre image de la Camargue, celle de l'intérieur. Merci également pour toutes les anecdotes de la vie courante qui mettaient du baume au cœur dès le lundi matin. Je fais régulièrement "des Céline".

Un grand merci à Christine, pour ton éternel soutien dans les moments difficiles et les plus joyeux, pour tes conseils, tes relectures, ta gentillesse, je ne saurais comment te remercier, merci ! J'aurais certainement tout laissé tomber si tu n'avais pas été là mais j'ai enfin atteint la fin du tunnel.

Merci à mes collègues de bureau du CBI, Vincent, et Marion de la plateforme bioinformatique BigA et Sébastien, le pharmacien bioinfo ainsi qu'à l'ensemble de l'équipe Trouche avec Didier, Estelle, Lisa, Fabrice, Mahdia, Julien et Luana. Un merci particulier à Marion Aguirrebengoa et Didier Trouche qui m'ont accordé leur confiance dans ce nouveau projet au CNRS qui est à l'opposé de ce que j'ai fait jusqu'à maintenant.

Merci à mes amis, à ma petite famille et ma belle-famille pour leur soutien ; à leur éternelle question "C'est quand que tu soutiens" ou alors "C'est pas déjà fini ?", "C'est sur quoi déjà ton truc ? Ça sert à quoi ? Non mais dans la vraie vie".

Merci à mon père et à ma mère qui n'ont jamais cessé de croire en moi, qui m'ont accompagné tout au long de mes études, en me répétant « Je n'y comprends pas grand-chose mais si tu travailles tu y arriveras ». A mon chat Méli qui a su égayer mes journées et aux chatons Tweety et Sly pour les animer. Merci Floriane pour tes relectures et tes conseils sur Docker ; merci pour ces 25 ans d'amitié à toujours me supporter et à partager la passion des chats (Petite pensée à Pollen) ! Merci à mes amies de longues dates, ma maman de fac Cécile ainsi que Christophe, Marie-Anne (alias Kaly), Jérome, Genjo. Merci aux amis de Montpellier Marion, Jules, Matthieu et Quentin.

Un grand merci à mon chéri, Yannick Cogne, avec qui j'ai partagé cette expérience au CEA en tant que collègue bioinfo et qui m'accompagne au jour le jour et pour qui j'ai le plus grand des respects ; merci mon cœur de me supporter au quotidien et d'être là à mes côtés, de m'encourager à viser le sommet et à me dépasser jour pour jour.

Merci à tous, vous, lecteur expérimenté ou plus jeune, je vous souhaite une bonne lecture !

Liste des abréviations et des acronymes

ADN	Acide désoxyribonucléique
AP	Alkylphénol
ARN	Acide ribonucléique
ARNr	ARN ribosomique
CDS	Cadre ouvert de lecture
DDA	Acquisition dépendante des données (data dependant analysis)
DIA	Acquisition indépendante des données (data independant analysis)
ESI	Source d'ionisation par électonébuliseur
ETM	Eléments traces métalliques
GO	Ontologie des gènes (Gene Ontology)
HAP	Hydrocarbures aromatiques polycycliques
KEGG	Encyclopédie de Kyoto sur les gènes et les génomes (Kyoto Encyclopedia of Genes and Genomes)
LC	Chromatographie liquide (liquid chromatography)
MAGs	Génomes assemblés à partir de données de séquençage métagénomique (metagenome assembled genomes)
MRM	Surveillance multiréactionnelle (multi reaction monitoring)
MS	Spectrométrie de masse
MS/MS	Spectrométrie de masse en tandem
NCBI	National Center for Biotechnology Information
PBDE	Polybromodiphényléthers
PCB	Polychlorobiphényls (ou biphényls polychlorés)
POP	Polluant organique persistent
PRM	Surveillance parallèle des réactions (parallel reaction monitoring)
PSM	Peptide associé à un spectre (Peptide-Spectrum Match)
p-value	Probabilité d'obtenir le résultat par hasard
RT	Temps de rétention
SDS	Dodécylsulfate de sodium
TOF	Temps de vol (time-of-flight)
WGS	Séquençage de génome complet (whole genome sequencing)
XIC	Chromatogramme d'ions extraits (extracted ion chromatogram)

Liste des figures

Liste des tables

Préambule

Les sols représentent un habitat complexe, constitué pendant des temps très longs par l'action conjuguée de l'érosion des roches et la dégradation de la matière organique apportée par la faune et la flore. Par la complexité de leurs compositions, ils constituent une matrice difficile pour les analyses omiques. Les communautés microbiennes qui y vivent se sont adaptées à leurs écosystèmes qui ont des caractéristiques physico-chimiques différentes selon la nature du sol, sa localisation et sa profondeur. Les perturbations d'origine naturelle telles que les aléas climatiques et celles liées aux activités humaines avec l'agriculture, l'urbanisation et l'industrialisation participent à la grande diversité des paramètres de ces écosystèmes. La grande diversité et leurs potentiels d'adaptation permettent aux communautés bactériennes de se maintenir et de se développer dans des conditions extrêmement variées que ce soit en termes de température, de salinité, d'humidité, d'acidité ou de disponibilité en oxygène et en nutriments, mais aussi la présence de contaminants qui résultent des activités anthropiques comme les hydrocarbures aromatiques polycycliques (HAP), les éléments traces métalliques (ETM), les radionucléides, les pesticides et les antibiotiques. La composition de ces sols est également impactée par ce qui s'y trouve en surface, par exemple les plantes. Celles-ci structurent la physique et la chimie du sol grâce aux apports de matière organique et à la croissance des racines. Les micro-organismes présents dans ces sols interviennent dans les cycles biogéochimiques d'éléments majeurs tels que le carbone, l'azote, le soufre et le phosphore. Les bactéries interviennent également dans la transformation et le transfert de polluants dans l'environnement. Leurs implications fortes dans les fonctions du sol leur donnent un rôle central dans la fertilité et plus généralement, l'état de santé de ce sol.

Le sol étudié durant cette thèse provient d'une plaine d'inondation située dans la partie aval du bassin de la Seine. Ce sol s'est formé à partir des sédiments déposés lors de crues successives. Le site draine 96 % du bassin versant de la Seine. Les contaminants présents ont une forte affinité pour les alluvions et s'y fixent. Ces alluvions qui ont évolué en sol sont également appelés sédiments hérités (« legacy sediments »). Ils sont les témoins des activités anthropiques et sont des archives sédimentaires des contaminations présentes et passées. Ces sols présentent une sédimentation régulière qui peut être mise à profit après datation pour reconstruire la dynamique temporelle de la contamination de la Seine, à l'aide de mesures des contaminants le long d'une colonne de sédiments. Des travaux ont notamment porté sur les éléments traces métalliques (Ayrault et al., 2010 ; Le Cloarec et al., 2011) et plus particulièrement le plomb (Ayrault et al., 2012) ainsi que sur les antibiotiques (Tamtam et al., 2011) et d'autres contaminations organiques (Lorgeoux et al., 2016).

Dans un contexte d'une seule santé, le concept « One Health », la compréhension de la santé de notre environnement est importante. Ce concept repose sur le principe que la santé de l'Homme repose sur celle de l'animal et de l'environnement. Le diagnostic de l'état de santé d'un sol nécessite de connaître les différents états d'un sol et de paramétrer des seuils définissant des états différents. Mieux diagnostiquer cet état de santé permet aussi d'envisager des solutions pour y remédier et mesurer les actions de rétablissement. De manière générale, les indicateurs de la santé des sols peuvent être classés comme physiques, chimiques ou biologiques. Dans les études d'analyse de la santé des sols, les indicateurs couramment utilisés sont la quantification de la matière organique, le pH, le phosphore et le potassium assimilables par les plantes (Lehmann et al., 2020). Des indicateurs tels que la respiration du sol, de la biomasse microbienne ou de la minéralisation de l'azote sont également recommandés dans un tiers des études analysées par Lehman (Lehmann et al., 2020). La santé du sol pourrait-elle être déterminée à partir des communautés microbiennes des sols présentes, de leurs structures et de

leurs fonctions ? Singh a proposé pour évaluer les risques en écotoxicologie d'utiliser les micro-organismes comme outil (Singh et al., 2016). Les écosystèmes microbiens composés de bactéries, de protistes, de champignons et d'archées peuvent être analysés grâce à l'essor des méthodes d'analyses métataxonomiques telles que les analyses de séquences du gène codant l'ARNr 16S et 18S (Fierer et al., 2006) mais aussi celles des approches de métagénomique et de métaprotéomique. L'analyse globale de ces communautés microbiennes des sols pourrait fournir une vision holistique de la santé d'un sol. Les méthodes à grande échelle telles que la métagénomique (Dubey et al., 2019) et la métaprotéomique (Bastida et al., 2008) peuvent contribuer au développement de nouveaux bioindicateurs de la santé des sols.

Les approches de métaprotéomique basées sur des résultats massifs de spectrométrie de masse ultra-rapide permettent l'identification des organismes présents dans l'échantillon. Dans le cas du sol, les données obtenues sur l'identification et la quantification des protéines ainsi que de leurs analyses fonctionnelles pourraient être corrélées à la présence de contaminants d'origines anthropiques. La métaprotéomique environnementale est un domaine en pleine croissance. Des protocoles d'extraction et l'optimisation de l'obtention des données de métaprotéomique sont régulièrement le sujet d'études (Chourey et al., 2010 ; Bastida et al., 2014 ; Quinn et al., 2022) afin de maximiser la quantité de signal interprétable. En effet, le sol est une matrice complexe contenant des interférents tels que la matière humique. Une deuxième limitation se situe sur la méthode utilisée pour interpréter les spectres. En effet, ces analyses reposent sur l'utilisation d'une base de données de séquences qui doit contenir l'ensemble des organismes présents dans l'échantillon afin de permettre leur identification et leur quantification au niveau taxonomique et fonctionnel. Cependant, la base de données idéale car complète n'existe pas et la construction de la base de données utilisée impacte directement le résultat de toute étude métaprotéomique. Différentes méthodologies peuvent être utilisées notamment par l'utilisation de données métagénomiques pour reconstruire par assemblage des séquences protéiques attendues ou de données métataxonomiques par l'identification des organismes à un certain rang taxonomique et la récupération des protéomes des organismes identifiés (Tanca et al., 2016). Différentes méthodologies d'analyse peuvent être utilisées avec notamment des stratégies d'analyses en plusieurs étapes (Jagtap et al., 2013). Ces études permettant d'optimiser les résultats obtenus en métaprotéomique sont développées pour étudier des échantillons de microbiotes humains ou de matières fécales (Tanca et al., 2016). Cependant, quelles méthodologies devons-nous utiliser pour analyser les données de spectrométrie de masse dans le cas d'échantillons de sol ?

La reconstruction de l'histoire géochimique de ce sol couplée à l'analyse stratifiée des communautés microbiennes d'une carotte de sol permet-elle de comprendre les impacts anthropogéniques sur les communautés microbiennes et d'étudier l'état de santé de ce sol ? Dans un premier temps, la géochimie appliquée à l'analyse de sédiments et le site d'étude du bassin de la Seine seront présentés. Ensuite, l'introduction bibliographique s'axera sur les micro-organismes présents dans les sols puis sur les approches métagénomiques et métaprotéomiques, avant de conclure sur le contexte et les objectifs de la thèse.

1 L'état de l'art

1.1 La géochimie appliquée à l'analyse des sédiments

Les sédiments fluviaux ou marins peuvent être analysés d'un point de vue dynamique sédimentaire, en termes de transport et de dépôt mais les sédiments sont également, d'un point de vue historique de la contamination des fleuves, une archive des perturbations d'origine naturelle ou anthropique. La granulométrie des sédiments est un bon indicateur des caractéristiques hydrodynamiques d'un milieu aquatique (Figure 1). L'origine des sédiments transportés dans les rivières peut être déterminée par des méthodes nommées « fingerprinting » pour étudier leurs dynamiques de transport (Walling, 2013). La composition en éléments chimiques des sédiments et plus précisément la teneur en matière organique et de divers contaminants peut être interprétée pour reconstruire l'histoire et retracer les activités anthropiques du bassin versant à l'aide de techniques appartenant à la géochimie.

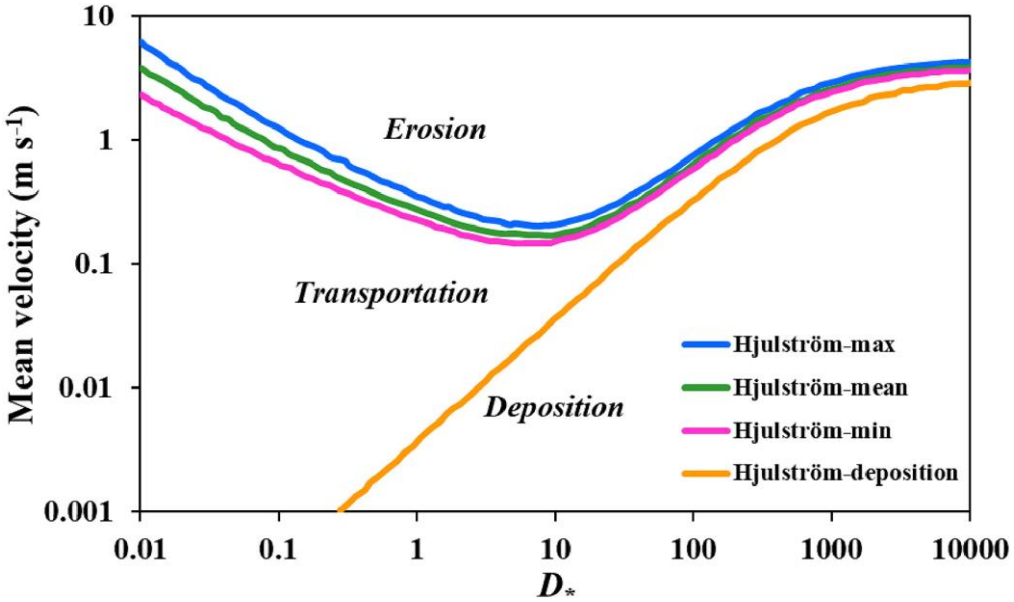


Figure 1: Diagramme de Hjulström où D est la taille du grain sans dimension (Yang et al., 2019).*

1.1.1 Qu'est-ce que la géochimie ?

La géochimie est une discipline relativement récente dont le terme a été utilisé pour la première fois en 1838 par Mr Schönbein, un chimiste suisse. Elle étudie l'origine, l'évolution et la répartition des éléments chimiques et de leurs isotopes dans les différents environnements terrestres et cosmiques. Elle analyse la composition et la dynamique à court et long terme des processus géochimiques tels que la cristallisation des carbonates marins ou les dépôts de sédiments de manière quantitative. La géochimie est une discipline vaste subdivisée en de nombreux domaines.

La géochimie organique étudie la transformation de la matière biologique et son origine, la genèse des hydrocarbures ou des composés organiques. La géochimie inorganique, appelée également géochimie minérale, porte sur la composition en éléments et leurs cycles et transferts dans les environnements terrestres et extra-terrestres. La géochimie isotopique porte, quant à elle, sur la mesure et l'interprétation des compositions isotopiques des éléments chimiques. L'analyse en profondeur de la composition géochimique d'un environnement, notamment l'étude des transferts des contaminants, nécessite d'étudier la chronologie, la dynamique et d'effectuer le traçage des éléments chimiques en utilisant des traceurs tels que les isotopes radioactifs et leurs descendants radiogéniques.

1.1.1.1 La géochimie minérale

L'étude des éléments chimiques et de leur abondance dans les minéraux et roches est au centre du domaine de la géochimie minérale. Les éléments chimiques sont ainsi classés selon :

- Leur abondance : éléments majeurs, mineurs et traces ;
- La position dans le tableau périodique : les gaz rares, les lanthanides ou terres rares, les métaux de transition et les éléments du groupe platine ;
- Leur comportement lors de la fusion partielle selon leur affinité avec la phase liquide ou solide ;
- Leur charge ionique et leur rayon ionique qui va indiquer s'ils seront mobiles ou immobiles lors de l'altération des roches ;
- L'électronégativité avec la classification de Goldschmidt, décrite ci-dessous ;
- La température de condensation des éléments avec les éléments réfractaires et volatiles.

La classification de Goldschmidt, publiée en 1937 (Goldschmidt, 1937), classe les éléments en quatre grandes classes selon leur affinité dominante (Figure 2: les sidérophiles pour le fer, les chalcophiles pour le soufre, les lithophiles pour les silicates, les atmophiles pour les phases fluides. Des modifications ont ensuite été proposées notamment avec l'inclusion de l'arsenic, du mercure et du plomb en tant qu'éléments biophiles (Hollabaugh, 2007).

Iron, siderophile.	Sulphide, chalcophile.	Silicate, lithophile.	Gases, atmophile.	Organisms, biophile.
Fe, Ni, Co	((O)), S, Se, Te	O, (S), (P), (H)	H, N, C, (O)	C, H, O, N, P
P, (As), C	Fe, Cr, (Ni), (Co)	Si, Ti, Zr, Hf, Th	Cl, Br, I	S, Cl, I (B)
Ru, Rh, Pd	Cu, Zn, Cd, Pb	(Sn)		(Ca, Mg, K, Na)
Os, Ir, Pt, Au *	Sn, Ge, Mo	F, Cl, Br, I	He, Ne, Ar	(V, Mn, Fe, Cu)
Ge,* Sn *	As, Sb, Bi	B, Al, (Ga), Sc, Y	Kr, X	
Mo, (W)	Ag, (Au), Hg	La, Ce, Pr, Nd, Sm		
(Nb), Ta	Pd, Ru, (Pt)	Eu, Gd, Tb, Dy		
(Se), (Te)	Ga, In, Tl	Ho, Er, Tu, Yb, Cp		
	(Cr)	Li, Na, K, Rb, Cs		
		Be, Mg, Ca, Sr, Ba		
		(Fe), V, Cr, Mn		
		((Ni)), ((Co)), Nb, Ta		
		W, U, ((C))		

Figure 2: Classification de Goldschmidt (Goldschmidt, 1937).

L'origine de chaque élément est ainsi étudiée que ce soit dans un contexte de présence naturelle ou influencée par les activités humaines. De nombreuses études sont réalisées, par exemple sur les éléments traces métalliques (ETM) tels que l'arsenic (As) (Shrivastava et al., 2015), le cadmium (Cd) (Wang et al., 2015), le chrome (Cr) (Dhal et al., 2013), le cuivre (Cu) (Ballabio et al., 2018), le manganèse (Mn) (Pavilonis et al., 2014), le plomb (Pb) (Ayrault et al., 2012), le zinc (Zn) (Shaheen et al., 2014) et le mercure (Hg) (Xu et al., 2015).

1.1.1.2 La géochimie organique

En parallèle des contaminants tels que les éléments traces métalliques, d'autres types de contaminants sont également problématiques dans les milieux. En effet, la rémanence des contaminants organiques limite l'amélioration de la qualité des milieux tels que le bassin de la Seine avec les hydrocarbures aromatiques polycycliques (HAP) (Gateuille et al., 2014) ou le bassin du Rhône avec les polychlorobiphényles (PCB) (Mourier et al., 2014). L'émergence de nouveaux contaminants comme les micropolluants démontrent la nécessité d'agir et de comprendre précisément le fonctionnement et l'impact de ces molécules sur le vivant. Parmi les micropolluants, on retrouve des composés pharmaceutiques tel que les antibiotiques et autres médicaments, les phtalates, les nitrosamines, les polybromodiphényléthers (PBDE), les PCB et les HAP. Ces derniers, les HAP, sont des contaminants d'intérêt majeur mis en évidence par le développement des méthodes analytiques telles que la

chromatographie en phase gazeuse (Keith, 2014). Ainsi, 16 HAP ont été définis comme polluants prioritaires par l'EPA (Environmental Protection Agency, USA) comme composés représentatifs parmi les 65 composés organiques identifiés comme polluant les eaux potables de la Nouvelle-Orléans (Keith, 2014). La directive cadre européenne sur l'eau retient 8 HAP parmi ces 16 HAP : le naphtalène, l'anthracène, le fluoranthène, le benzo(b)fluoranthène, le benzo(k)fluoranthène, le benzo(ghi)perylène et l'indeno(1,2,3-cd)pyrène (Gateuille et al., 2020).

1.1.2 Cycles biogéochimiques des éléments traces

Seulement six éléments chimiques constituent 99 % des silicates de notre planète : l'oxygène, le magnésium, la silice, le fer, l'aluminium et le calcium. Contrairement à ces éléments majoritaires, les éléments-traces sont présents naturellement en faibles concentrations dans l'environnement et ne sont pas à l'origine de la disparition ou apparition de phases minérales c'est-à-dire que leur comportement sera dicté par l'évolution du système (Géochimie, Dunod, 2009). Cependant, selon les milieux, un élément considéré comme élément trace dans la plupart des environnements peut être un élément majeur dans d'autres comme le potassium dans les granites (Carron et Lagache, 1971). Les éléments majeurs ont dans les roches une teneur supérieure à 1 %, les éléments mineurs une teneur entre 0,1 et 1 % et les éléments traces, une teneur inférieure à 0,1 %.

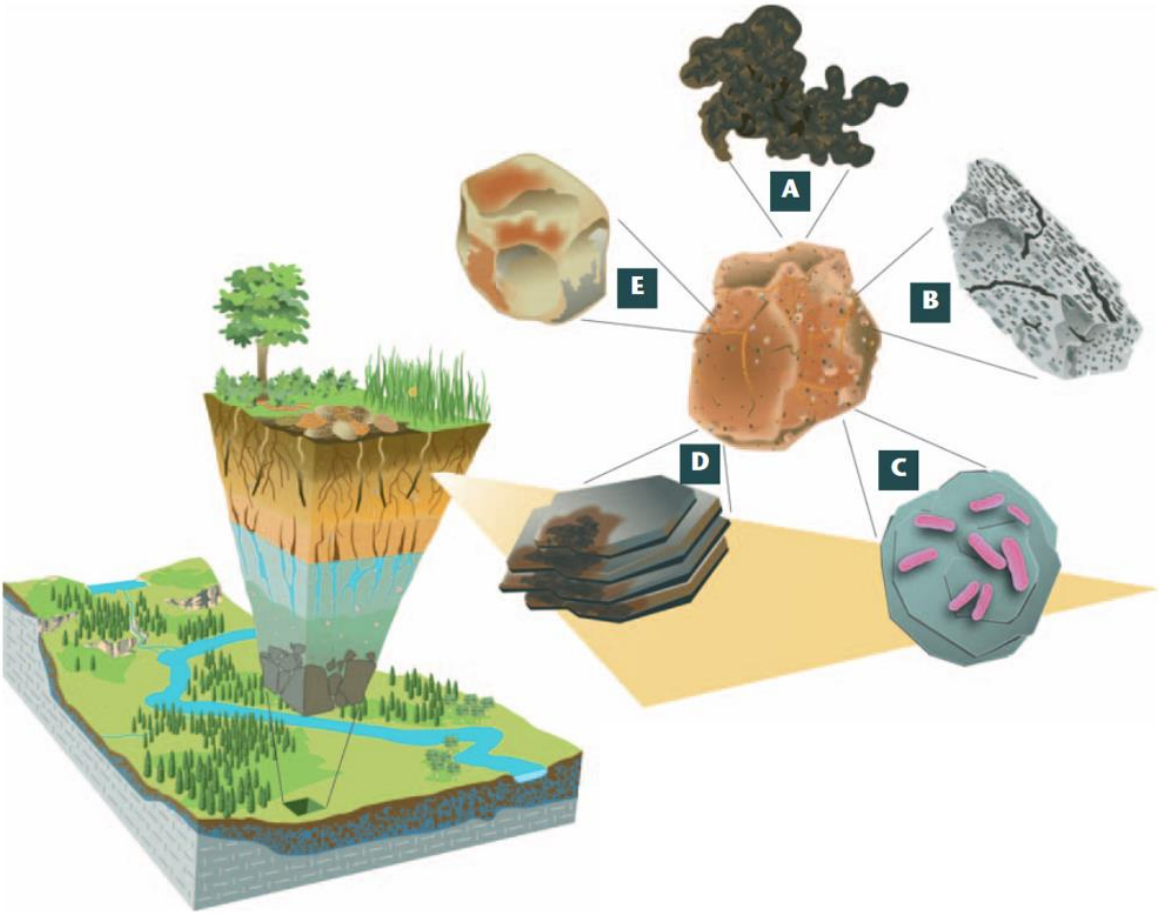


Figure 3: A un climat et une position géographique donnés, la composition géochimique du sol dépend des conditions biogéochimiques et comprend de la matière organique naturelle (A), des minéraux silicatés (B), des complexes minéraux-micro-organismes (C), des argiles (D) ainsi que des carbonates et oxydes (E), issu de Chorover et al., 2007.

1.1.2.1 Les cycles naturels

Les cycles biogéochimiques sont essentiels à l'existence de la vie, car ils transforment l'énergie et la matière en formes utilisables pour soutenir le fonctionnement des écosystèmes tel que les cycles de l'eau, du carbone et bien d'autres. Les éléments en trace sont naturellement présents dans les roches. Leurs concentrations dans celles-ci sont variables en fonction de la nature de ces roches. Le transfert des éléments en trace depuis les roches vers les sols, l'eau et la biosphère est gouverné par les cycles biogéochimiques. Ces transferts dépendent de nombreux facteurs biotiques, par exemple, la présence de communautés microbiennes pouvant promouvoir ou au contraire ralentir ces transferts, et de facteurs abiotiques, par exemple le pH et l'humidité. Ainsi, leurs concentrations « naturelles » dans les sols, appelées également « bruit de fond géochimique » varient d'un site à l'autre. Il est donc nécessaire de définir le « bruit de fond géochimique local » pour constater de façon fiable une éventuelle pollution en éléments traces.

La composition géochimique du sol est fortement variable d'un site à l'autre, même pour un même climat. En effet cette composition dépend des conditions biogéochimiques de formation de sol qui sera composé de parts variables de matière organique naturelle (A), de minéraux silicatés (B), de complexes minéraux-micro-organismes (C), d'argiles (D) ainsi que de carbonates et oxydes (E) (Figure 3).

1.1.2.2 Origine de la contamination

Avant de discuter de l'origine des contaminations et de leurs diversités, il est nécessaire de définir les termes « pollution » et « contamination » qui sont souvent utilisés dans le même contexte. Parmi les définitions du terme pollution, celle de Holdgate est l'introduction par l'homme dans l'environnement de substances ou d'énergie susceptibles de présenter des risques pour la santé humaine, de nuire aux ressources vivantes et aux systèmes écologiques, de porter atteinte à la structure ou à l'agrément de l'environnement ou de gêner son utilisation légitime (Holdgate et al., 1979 ; Thi Thu Dung et al., 2013). Cette définition pointe du doigt les activités anthropiques comme à l'origine de la pollution. L'évaluation de la pollution implique donc de différencier les concentrations naturelles et anthropiques des éléments dans les milieux (Thi Thu Dung et al., 2013). Chapman propose une définition claire de la contamination : « La contamination est simplement la présence d'une substance là où elle ne devrait pas être ou à des concentrations supérieures au niveau de fond. La pollution est une contamination qui entraîne ou peut entraîner des effets biologiques néfastes. » (Chapman et al., 2007). Afin d'évaluer la contamination d'un milieu par un élément, il est donc nécessaire de mesurer sa concentration et de la comparer à la concentration naturelle de l'élément si celui-ci est présent naturellement dans ce milieu, ce qui est le cas des éléments traces métalliques.

Les activités anthropiques peuvent fournir des quantités excessives de métaux et de métalloïdes à l'environnement (Resongles et al., 2014), en particulier, les activités industrielles, domestiques et agricoles (Le pape et al., 2012; Rosolen et al., 2015). Dans les bassins versants, les retombés atmosphériques humides (i.e., la pluie) et sèches (i.e., les poussières) associées aux activités anthropiques sont l'une des principales voies de contamination des sols. Les éléments traces sont transportés et déposés sur les sols où ils peuvent s'accumuler et être stockés sur de longues périodes mais ils peuvent se retrouver dans l'environnement fluvial par des processus tels que l'érosion ou le lessivage (ou lixiviation) des sols (Le Gall et al., 2018). Depuis la révolution industrielle, les sources d'émissions anthropiques sont l'exploitation minière et la métallurgie mais une multitude de sources majeures différentes y contribuent, telles que le rejet de boues d'épuration, l'utilisation d'engrais et de pesticides et le rejet d'eaux usées (Viers et al., 2008). Ces activités contribuent à une contamination directe des sols et des eaux, et une contamination indirecte des eaux par l'érosion des sols (Le Gall et al., 2018, Gateuille et al., 2014).

Notre société industrialisée emploie de nombreux métaux ce qui influence la concentration de ces éléments dans les sols. On peut citer quelques exemples : l'utilisation de cuivre pour l'électricité, le plomb pour les canalisations d'eau et les toits, ou encore le cadmium et le nickel pour les batteries. Ces usages peuvent conduire à la contamination de milieux tels que les sols, les sédiments et l'eau. Mais certains de ces éléments sont aussi essentiels aux êtres vivants comme l'indique la classification de Frieden (Frieden, 1974; Bhattacharya et al., 2016).

1.1.3 Apport de la géochimie dans l'analyse des contaminants dans les sédiments

La plupart des contaminants ont une forte affinité pour les particules et notamment les sédiments. Une fois introduits dans la rivière, ils se fixent sur les particules présentes dans la colonne d'eau et se déposent avec les particules. Ainsi les sédiments sont des archives des contaminations (Vauclin, 2021).

L'analyse des contaminants inorganiques et organiques déjà caractérisés tels que le plomb et certains micropolluants (HAP, PCB, etc.) en différents points d'un bassin versant permet de déterminer les trajectoires des polluants et de les caractériser à l'échelle d'un bassin versant (Figure 4). Les contaminants organiques issus des produits pharmaceutiques, des microplastiques ou des marqueurs fécaux retrouvés dans les sédiments permettent de suivre les usages et l'évolution des réseaux d'assainissement. Au niveau industriel, des molécules telles que les chloroalcanes, les composés perfluorés et les phtalates permettent le suivi des usages et des règlementations (Ayrault et al., 2020).

Par exemple, le plomb (Pb) est un contaminant métallique dans de nombreux systèmes fluviaux à travers le monde et reconstituer l'histoire de la contamination par le Pb et quantifier sa dynamique de transport dans les systèmes fluviaux est important du fait de sa toxicité (Ayrault et al., 2012). L'analyse de la composition isotopique des sédiments a permis d'élucider son origine pour déterminer la nature et la contribution des différentes sources de Pb dans le bassin de la Seine (Ayrault et al., 2012). Une partie est issue de son accumulation dans Paris depuis le XIXème siècle dans les peintures et les éléments de toitures par exemple. L'autre partie, dont le pic est autour de 1986, est lié la présence d'additifs contenant du plomb dans l'essence, interdits en France depuis le 2 janvier 2000 (Ayrault et al., 2010).

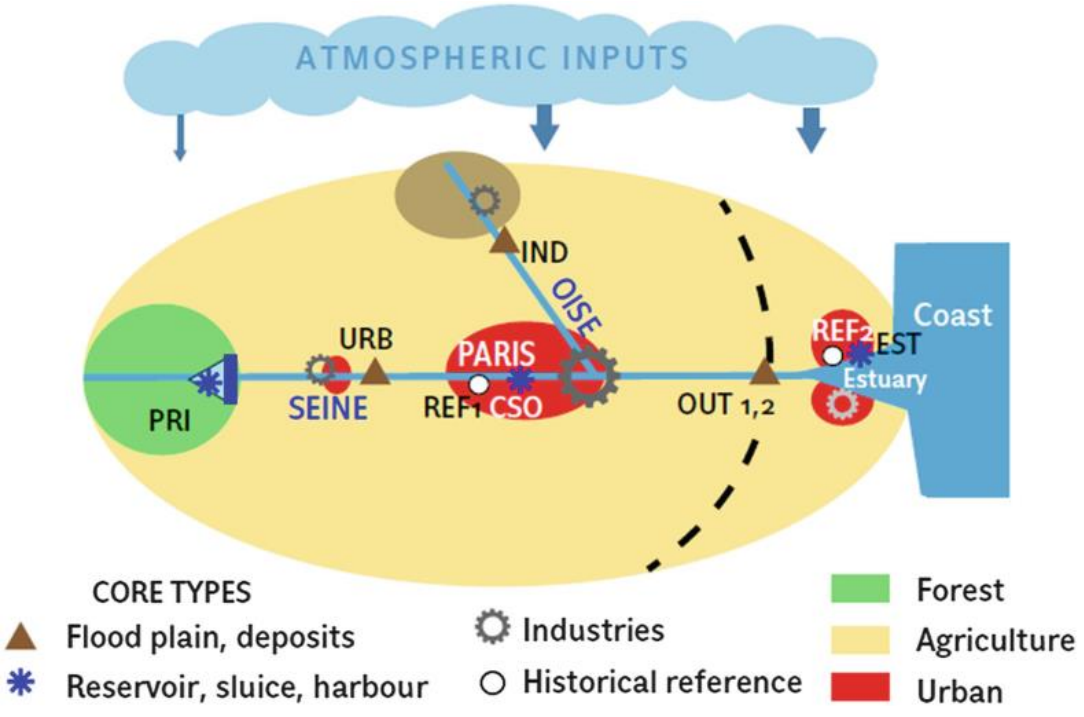


Figure 4: Sites de prélèvements de carotte de sédiments du bassin de la Seine (Ayrault et al., 2020).

1.2 Le bassin de la Seine et ses sédiments

1.2.1 Contexte socio-économique et territorial

1.2.1.1 Histoire du bassin de la Seine : un bassin au cœur des hommes

La Seine occupe une place centrale dans l'histoire de ce territoire et ce depuis la Préhistoire. Au Moyen-Age, les hommes vivent au fond des vallées. Des étangs sont aménagés pour nourrir les villageois et de nombreux moulins sont construits, modifiant la morphologie du fleuve. Ces aménagements du bassin ne cessent de s'accroître et s'accentuent au XIXème et XXème siècle. Les ouvrages mis en place dans le bassin de la Seine répondent aux besoins croissants d'une population en pleine expansion. Au IXème siècle, les activités dans le bassin sont très contrastées spatialement et se poursuivent sur plusieurs siècles. A l'est, la forêt est présente sur toute la partie supérieure du bassin dont le plateau de Langres et le Morvan (Source : PIREN-Seine, https://www.piren-seine.fr/sites/default/files/piren_documents/fascicules/Collection_AESN_PIREN-Seine_10_-_L_eau_dans_les_campagnes_avant_l_ere_industrielle.pdf). Ces forêts sont utilisées pour l'élevage et fournissent le bois pour la construction et le chauffage, le bois étant transporté vers les villes par flottage. Le centre est quant à lui occupé par des plaines céréalières, des vignes sur les coteaux et, dans les zones humides, on trouve les cultures de l'osier et du lin. Paris prend de l'importance au cours du Moyen-Age jusqu'à atteindre 220 000 habitants au XIVème siècle. Durant la période du Moyen-Age, les évolutions économiques et sociétales ont fortement impacté les plaines d'inondation (i.e. les surfaces jouxtant le fleuve et inondées lors des crues) et les cours d'eau (Lewin, 2010). Une deuxième phase de développement importante s'intensifie dès la révolution industrielle au XVIIIème siècle (Lewin, 2013).

Un canal est même construit pour relier la Loire et la Seine : le Canal de Briare utilisable dès 1642. Le transport intensif de bois avait, déjà à l'époque, des répercussions sur l'environnement avec des effets d'eutrophisation des milieux, c'est-à-dire, une forte croissance des végétaux dans le milieu aquatique conduisant à l'épuisement de l'oxygène présent au détriment de la faune piscicole. Aujourd'hui encore le bassin est menacé par ce phénomène naturel malgré des progrès dont le traitement des eaux usées pour contrôler les rejets de nutriments tels que le phosphore et l'azote (Source : PIREN-Seine, https://www.piren-seine.fr/sites/default/files/piren_documents/fascicules/Collection_AESN_PIREN-Seine_06_-_eutrophisation.pdf).

Figure 5: La Seine en aval du pont Neuf à Paris avec, à gauche, le Louvre et, à droite, le collège des Quatre-Nations, Peinture de RAGUENET Jean-Baptiste-Nicolas (1715 - 1793) réalisée en 1754, conservé au musée du Louvre (Paris). Source : Histoire-image.org)

Petite anecdote : *La Seine a été aussi un atout stratégique dans l'histoire de France. Lors du blocus de Paris en 1649 par les troupes fidèles au roi pour mettre fin à la fronde, la crue exceptionnelle de la Seine, le froid ainsi que la prise de la ville de Charenton, ont permis au roi Louis XIV de récupérer la ville de Paris avec la signature de la paix de Rueil (Hubac, 2018). Le peintre Le Conte Sauveur a illustré ce blocus avec différents tableaux dont celui intitulé « Les Actions du Grand Condé, Blocus de Paris 1649 » (Figure 5).*

Figure 6: Les Actions du Grand Condé, Blocus de Paris 1649. LE CONTE Sauveur (1659 - 1694). Peints en 1687.

L'étude de l'histoire du bassin à travers les cartes historiques permettent d'identifier les modifications du paysage qui ont eu lieu dans le bassin de la Seine au cours de ces derniers siècles (Figure 7)(Lestel et al., 2020). L'étude à plus court terme de ces cartes, et notamment des sédiments, permet d'identifier des zones qui ont été soumises à de très fortes pressions anthropiques avec par exemple la présence d'une industrie mais aussi d'identifier des zones non perturbées où l'homme n'a pas apporté de modifications. Ces zones sont des archives sédimentaires des contaminations passées (Ayrault et al., 2020).

1.2.1.2 L'empreinte environnementale de l'agglomération parisienne

Avec près de 17 millions d'habitants sur tout le bassin, 12 millions habitent en Ile-de-France dont 10,6 millions dans l'agglomération parisienne. Ces chiffres indiquent une densité de population inégale, avec en moyenne 225 habitants au km² et plus de 21 000 habitants au km² à Paris (Source : PIREN-Seine, https://www.piren-seine.fr/sites/default/files/piren_documents/fiches_4_pages/Fiche%20PS-Hydrologie_NUMERIQUE.pdf). Près de 90 % des communes du bassin de la Seine comptent moins de 2 000 habitants. Le bassin versant accueille 30 % de la population française et représente 40 % de l'industrie nationale et 25 % de l'agriculture nationale. Les rejets associés à cette forte activité anthropique sont reçus par le bassin marqué par une forte activité industrielle : les industries de transformation (pétrochimie, chimie de spécialités, papeteries) et les industries manufacturières (aéronautique, industrie mécanique) (eau-seine-normandie.fr). En comparaison avec le XIXème siècle (Figure 7), la croissance de l'agglomération parisienne est visible (Figure 8).

Figure 7: Carte centrée sur la région de Paris avec la zone urbaine en rose datant de 1820 à 1866 (carte au 1/40 000ème de l'Etat-Major). Issu de : https://www.apur.org/dataviz/evolution_nature/index.html

Figure 8: Photo aérienne de la région parisienne prise en 2015 avec les zones urbaines en gris par l'atelier parisien d'urbanisme (Apur). Issu de : https://www.apur.org/dataviz/evolution_nature/index.html

1.2.1.3 L'axe fluvial de la Seine

Prenant sa source à Source-Seine en Côte-d'Or, le bassin versant de la Seine a une superficie de 76441 km² soit 14 % du territoire français métropolitain (Figure 9). Le bassin versant de la Seine est composé de 47 affluents. La zone en amont est constituée de têtes de bassin en montagnes et de cours d'eau présentant des eaux claires et rapides. Le débit diminue dans la vallée avec l'apparition de méandres et un dénivelé plus faible. Dans la zone aval, une vallée large avec une pente très douce accueille une rivière large aux multiples méandres. Enfin, à l'estuaire est une zone de mélange des eaux douces et des eaux marines (Figure 10).

La Seine a un débit moyen de 310 m^3/s à la station de Paris Austerlitz. Avec les variations saisonnières, le débit varie de 100 m3/s en été jusqu'à 600 m3/s en hiver et peut atteindre des niveaux beaucoup plus hauts pendant les épisodes de crues, comme par exemple, 1750 m3/s lors des crues de juin 2016. Au commencement de l'estuaire, au barrage de Poses, le débit est de 485 m3/s en moyenne annuelle. En quelques chiffres, la Seine c'est 1400 km de voies navigables, 800 km de rivières et 600 km de canaux (Source : PIREN-Seine, https://www.piren-seine.fr/sites/default/files/piren_documents/fiches_4_pages/Fiche%20PS-Hydrologie_NUMERIQUE.pdf).

En 2019, le trafic fluvial représentait 23,7 millions de tonnes transportées dans le bassin de la Seine parmi les 56,3 millions de tonnes transportées au niveau national. Au niveau du tourisme fluvial, près de 8 millions de passagers ont voyagé en bateau sur la Seine le temps d'une promenade en île de France (Source : Voie navigables de France, https://www.vnf.fr/vnf/brochure-et-lettress/chiffres-cles-2019-de-voies-navigables-de-france-bassin-de-la-seine/). En comparaison, le Louvre représente 10,2 millions de visiteurs.

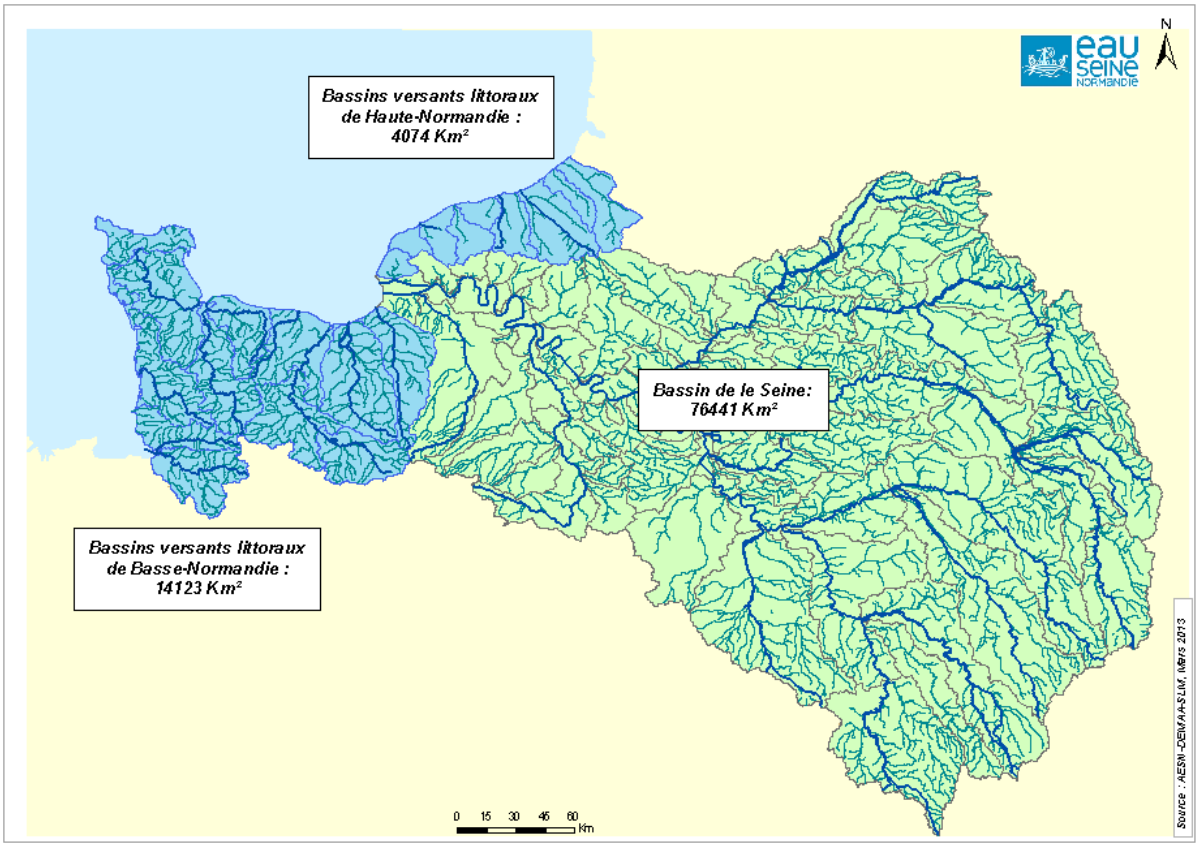


Figure 9; Bassin versant de la Seine et des littoraux de Haute-Normandie et Basse-Normandie. Issu de http://www.eau-seine-normandie.fr.

Figure 10: Schéma d'un bassin versant simplifié avec le cours d'eau principal (1), un affluent (2), un exutoire (3) avec une zone d'estuaire subissant les marées et ce bassin est délimité par la ligne de partage des eaux (4). Issu de l'Agence régionale pour l'environnement PACA – RRGMA – Azoé – 2016. (source https://www.syndicat-huveaune.fr/le-bassin-versant-de-lhuveaune/quest-ce-quun-bassin-versant/)

1.2.2 Le bassin de la Seine sous l'œil des scientifiques : le projet PIREN-SEINE

Le Projet PIREN-Seine voit le jour en 1989 né des projets Projets Interdisciplinaires de Recherche sur l'Environnement (PIREN) lancés par le CNRS. Plusieurs projets PIREN-Grands Fleuves sont créés (PIREN-Rhin, PIREN-Rhône, …), le PIREN-Seine est le seul encore en cours. Basé sur l'interdisciplinarité, ce projet a différents objectifs dont la compréhension dans son ensemble du continuum fluvial, du bassin versant et une meilleure gestion qualitative et quantitative de la ressource en eau en ayant une vue d'ensemble des milieux du bassin de la Seine tant sur les aspects physiques, géologiques, chimiques, biologiques, climatiques que historiques (Source : PIREN-Seine, https://www.piren-seine.fr/sites/default/files/piren_documents/fiches_4_pages/Fiche_6p_PIREN-Seine.pdf). Après 30 ans, le projet PIREN-Seine est dans sa 8ème phase divisé en cinq axes :

- Axe 1 : Trajectoires du bassin, de ses tissus urbains et agricoles et ses territoires ;
- Axe 2 : Fonctionnement du bassin soumis à des extrêmes hydroclimatiques ;
- Axe 3 : Construction de la qualité des milieux aquatiques conciliant risques hydrologiques et biodiversité ;
- Axe 4 : Ambition et enjeux pour la Métropoles en 2024 et après … ;
- Axe 5 : Dynamique des contaminants : de la compréhension des processus au métabolisme territorial ;
- Axe 6 : Transfert de connaissance et mise à disposition des données.

1.2.3 Les sédiments de plaine d'inondation

Les sédiments sont des archives des contaminations passées et présentes. Nées de l'érosion des sols lors d'évènements pluvieux, ces particules de sols sont transportées par le ruissellement jusqu'au cours d'eau le plus proche (Hudson, 1993) ou déposées selon des paramètres tels que le type de sol et le dénivelé. C'est ce que l'on appelle l'érosion hydrique (Lepage, 2015). Lorsque les particules se retrouvent dans les cours d'eau, celles-ci sont soit en suspension dans la colonne d'eau, soit transportées par charriage (Figure 11). L'érosion, le transport ou le dépôt d'une particule dans un cours d'eau dépendent de deux paramètres : la vitesse du courant et la dimension des particules, détaillé par le diagramme de Hjulström (Figure 1). La granulométrie et la densité de ces particules sont des paramètres dhydrodynamiques d'un milieu aquatique.

L'analyse précise de ces sédiments et notamment de leurs teneurs en matière organique et/ou la présence de macro-restes sont un indicateur de l'évolution de l'hydrosystème au cours du temps (Vauclin, 2021). Mais les sédiments sont également un indicateur des tendances temporelles de contaminations et de l'évolution de l'activité humaine et plus généralement des pressions anthropiques.

Les sédiments déposés au sein d'un bassin versant qui ont subi des perturbations liées aux activités humaines prennent le nom de sédiments hérités (ou « legacy sediments ») (James, 2013). Les sédiments et plus généralement les particules transportées sont affectés par la présence de polluants qui sont issus des retombées atmosphériques et des rejets directs lors de leur transport : ils enregistrent ces contaminations. Les sédiments qui en découlent peuvent être stockés dans de multiples types de milieu dans le bassin versant. Ils constituent donc une archive sédimentaire qui permet de retracer les trajectoires des contaminants. Lorsque les sédiments sont déposés en couches successives, l'analyse de ces sédiments, s'il n'y a pas eu de perturbation des dépôts, permet de retracer les perturbations passées, de quelques décennies à quelques siècles, (Gardes, 2021 ; Marcus et al., 1993 ; Gellis et al., 2009 ; Hupp et al., 2009 ; James, 2013).

Dans un fleuve comme la Seine, il n'y a pas de sédiment au fond de la rivière du fait de la faible érosivité du bassin principalement calcaire qui produit peu de particules ainsi que de la forte navigation qui

brasse le sédiment en permanence. Dans ces conditions, les plaines d'inondations, une zone adjacente au cours d'eau régulièrement submergée lors de crues et où les sédiments sont alors déposés lorsque le débit de la rivière diminue, sont des sites privilégiés pour collecter des archives sédimentaires (Figure 12). Les sédiments s'accumulent crue après crue et se transforment en sol lorsque la plaine redevient hors d'eau ce qui représente plus de 90 % du temps. Ce mode de dépôt crée des sols stratifiés qui sont des archives de la contamination. Idéalement, lors d'études de la contamination sur plusieurs décennies, la sédimentation doit être régulière dans le temps et non-perturbée ni par les activités humaines ni lors des crues. L'érosion doit être faible ou modérée afin de ne pas emporter les restes de sédiments issus de crues précédentes. Si ces critères sont remplis, alors on peut procéder aux analyses des carottes de sédiments qui peuvent atteindre plusieurs mètres. La première série d'analyses est réalisée par spectrométrie gamma pour obtenir le « modèle d'âge » de la carotte (Section 1.2.4.3.

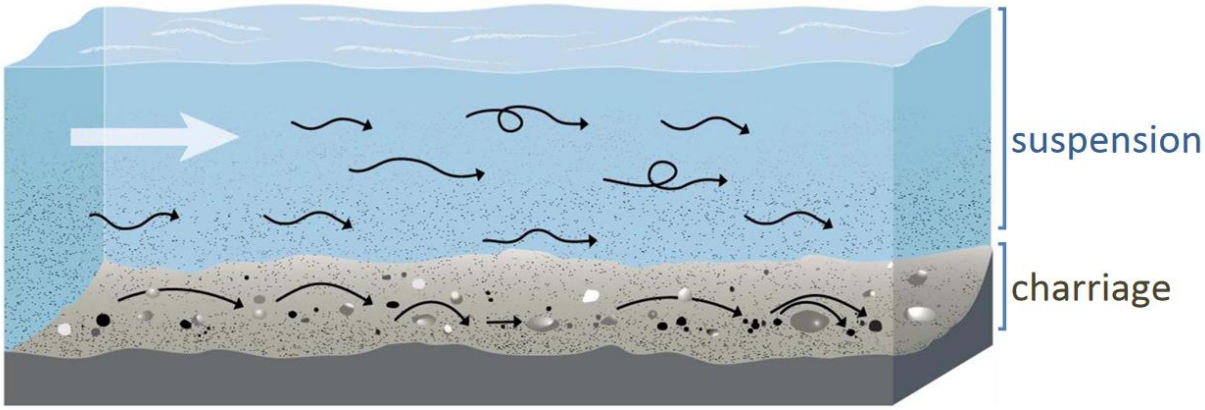


Figure 11: Transport de particules de sol dans les cours d'eau par charriage ou en suspensions (Des rivières et des hommes, 2015, https://lms.fun-mooc.fr/courses/grenobleinp/19001S02/session02/info)

Figure 12: Crue de 2016, apport de sédiments dans des zones inondables en dehors du lit mineur de la Seine. (Source : PIREN-Seine, https://www.piren-seine.fr/sites/default/files/piren_documents/fascicules/Fascicule_qualite_crue_PIREN-Seine.pdf).

De nombreuses études ont réalisé l'analyse des contaminants présents le long des carottes de sédiments et de sols en Seine dont notamment des études portant sur la trajectoire historique des éléments métalliques et les polluants organiques (Ayrault et al., 2010 ; Ayrault et al., 2020 ; Lorgeoux et al., 2016 ; Le Cloarec et al., 2011). D'autres fleuves, en France et à l'étranger, sont ainsi étudiés et comparés (Schäfer et al., 2022 ; Dendievel et al., 2020 ; Borrego et al., 2002).

1.2.4 Le prélèvement et l'analyse des sédiments

1.2.4.1 La recherche d'un site

Le prélèvement de sédiments de plaine d'inondation nécessite de choisir des sites de prélèvement adéquats avec une sédimentation continue et régulière. Le site de prélèvement ne doit pas être perturbé ni cultivé ni pâturé et être utilisable à long terme tout en étant facile d'accès. Cela nécessite un travail cartographique approfondi en utilisant les cartes historiques et actuelles afin d'identifier dans un premiers temps les zones inondables avec une bonne représentativité du bassin versant, ensuite le site utilisable et enfin la position (x, y) du carottage le jour du prélèvement. Lorsqu'un site est candidat, plusieurs carottages courts peuvent ainsi être extraits afin de vérifier la régularité de sédimentation en fonction de la profondeur. Un simple labour des premiers centimètres gomme les différences entre les couches en termes de contamination et rendent le site inexploitable pour de futures études (Figure 13).

Les sites sélectionnés peuvent être à des endroits stratégiques, en amont et en aval d'une zone sous forte pression anthropique ou disséminés sur différents affluents permettant de connaître les apports de chaque affluent dans la contamination mesurée en aval (Figure 4).

Figure 13: Carotte de sédiment prélevé à Bouafles présentant une zone labourée de 16 cm (rectangle jaune), photo de Sophie Ayrault.

1.2.4.2 Méthode de prélèvement

L'extraction de la carotte de sol ou de sédiment peut se faire avec un carottier « sol » à percussion en inox (Figure 13) mais il existe d'autres types tels que le carottier UWITEC (pour carotter sous eau) ou le carottier « russe » pour les milieux très meubles (comme les tourbes) par exemple. Le carottier à percussions en inox a plusieurs avantages, il perturbe très peu les séquences (compaction ou extension des couches), et n'ajoute pas de contaminations plastiques contrairement à des carottiers utilisant des tubes en polymère, et les enregistrements sont visibles immédiatement.

1.2.4.3 La datation au césium 137

Le césium ^{137}Cs est un radionucléide artificiel utilisé pour le traçage des particules et la datation. Il a été émis dans l'environnement lors des essais nucléaires des années 1950-1960 et par les accidents nucléaires, notamment Tchernobyl en 1986. Emis dans l'atmosphère, il retombe avec les pluies et se fixe aux particules de sol. Du fait de sa période de demi-vie de 30 ans, le ^{137}Cs est utilisé dans les études environnementales afin d'étudier l'érosion des sols et parce qu'il permet de dater les carottes de sédiments en fonction de la profondeur. A partir des pics observés de ^{137}Cs et de son profil (Figure 14), les différentes couches de la carotte de sédiments peuvent être datées. Le ^{210}Pb est également un traceur radioactif naturel et permet d'estimer le taux de sédimentation de la carotte de sédiment. Il a également une forte affinité pour les particules mais son utilisation dans les milieux n'est sûre que pour les archives maintenues constamment sous eau.

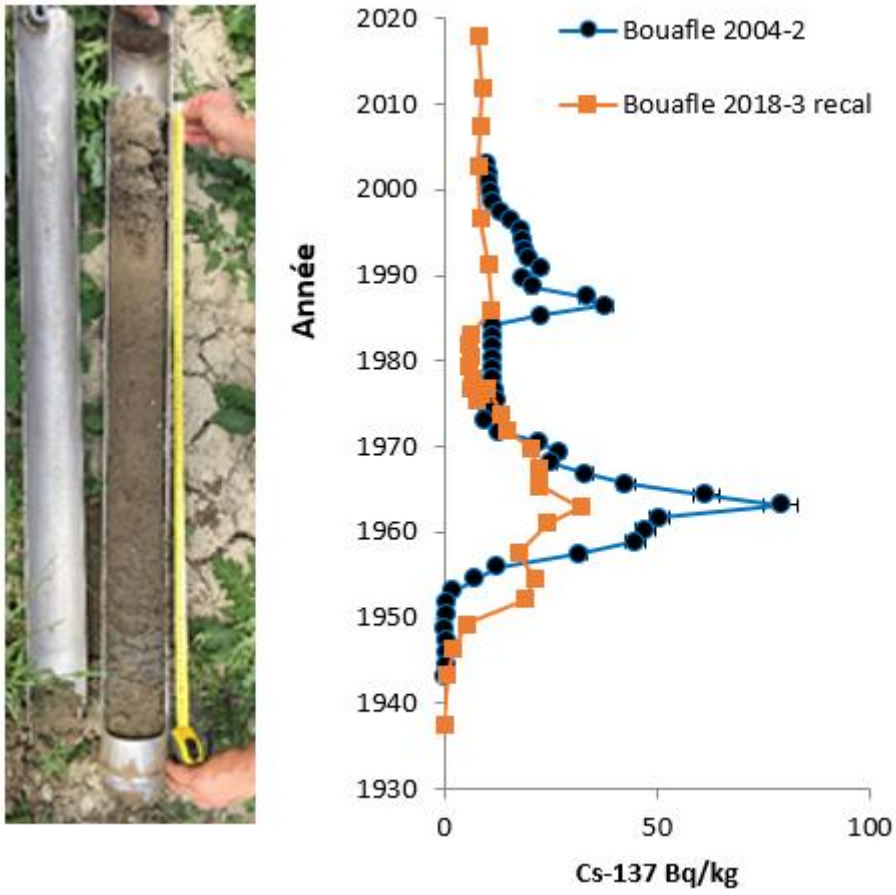


Figure 14: Profil de datation des carottes de sédiments Bouafles 2004-2 (carotte de référence prélevée en 2004) et Bouafle 2018-3 (carotte utilisée pour les travaux de cette thèse) à partir des taux mesurés de césium 137.

1.2.4.4 Les méthodes de mesures

Les mesures d'activités des radionucléides tels que le ^{137}Cs et le ^{210}Pb sont réalisées avec un spectromètre gamma (Figure 15) dont le principe et d'identifier et quantifier les éléments en mesurant l'énergie des rayons gamma que les échantillons émettent. Les échantillons sont séchés préalablement à la mesure. Le spectromètre gamma permet la mesure des teneurs en potassium et thorium qui sont des indices de la texture du dépôt et de la quantité d'argiles. Les éléments traces métalliques (ETM) ont été mesurés à l'aide de la spectrométrie de masse (ICP-MS) qui permet d'estimer les concentrations élémentaires en éléments après mise en solution totale par un mélange d'acides concentrés à chaud (Le Gall, 2016). Les composés organiques ont été analysés après extraction assistée par micro-ondes associés à un mélange de solvants (chlorure de méthylène/méthanol), d'élimination

des interférents et purification qui permettent d'analyser les HAPs, PCB et AP (Lorgeoux et al., 2016). Les HAP, PCB et PBDE sont quantifiés par un chromatographe en phase gazeuse couplé à un spectromètre de masse GC-MS et les AP par un chromatographe liquide couplé à un spectromètre de masse en tandem LC-MS/MS (Lorgeoux et al., 2016).

Figure 15: Spectromètre Gamma du LSCE de l'équipe d'Olivier Evrard.

A partir des mesures de différents contaminants, des indicateurs tels que le facteur d'enrichissement (FE) peut être calculé. Le thorium Th est un indicateur de la quantité d'argiles présents dans les sédiments. Il est utilisé pour normaliser la quantité de contaminants présents puisque ceux-ci ont une forte affinité pour les argiles. En effet, si la quantité d'argile est forte, la concentration en élément trace pourra être forte sans que cette concentration puisse être attribuée à une contamination, ce qui induit un biais dans la mesure. La concentration mesurée (M) dans l'échantillon est comparée à la concentration naturelle (bruit de fond géochimique local, M naturel). Le facteur d'enrichissement permet de comparer la contamination d'un site à l'autre, ou d'une période à l'autre pour un même site.

$$Facteur\ d'enrichissement\ FE = \frac{M}{Th}.\frac{Th\ naturel}{M\ naturel}$$

1.2.4.5 Exemple de la trajectoire temporelle des contaminants

La trajectoire temporelle des ETM dans une carotte de sol telle que celle collectée à Bouafles est analysée après une datation précise de la carotte (Figure 16) ainsi que des polluants organiques (Figure 17: Concentrations des polluants organiques mesurées à Bouafles, figure issus de Lorgeoux et al., 2016).

Les facteurs d'enrichissement des ETM (Figure 16) indiquent une abondance non naturelle des différents éléments avec un facteur supérieur à 1. Le cadmium par exemple présente un pic de contamination dans les années 1960-1970 avec un facteur d'enrichissement supérieur à 100 fois la concentration naturelle et décroit ensuite à 10 (Figure 16). La trajectoire temporelle des polluants organiques suit une tendance différente selon leurs catégories (Figure 17). Les tendances temporelles des composés sont en accord avec les changements d'utilisation et de réglementation (Lorgeoux et al., 2016). Ainsi, on note que les pics élevés de concentration des PCBs, utilisés depuis les années 30, se situent dans les années 1960 et 1975. Les PCBs sont interdits d'utilisation en France depuis 1987 et il est interdit de détenir des appareils dont le fluide contient des PCB depuis le 1er janvier 2017. Ces interdictions consécutives à la prise de conscience de leur toxicité a permis une réduction de leur concentration dans les sols depuis ces années.

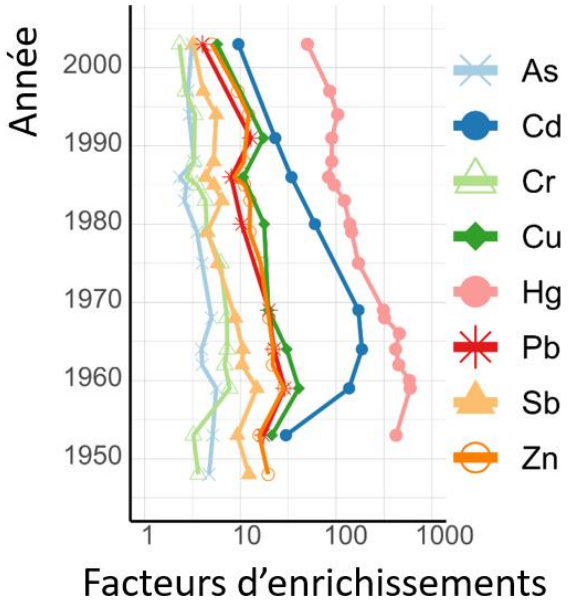


Figure 16: Facteurs d'enrichissement d'éléments traces métalliques obtenus sur le site de Bouafles, le site de carottage de la thèse, figure réalisée à partir des données de Le Cloarec et al., 2011.

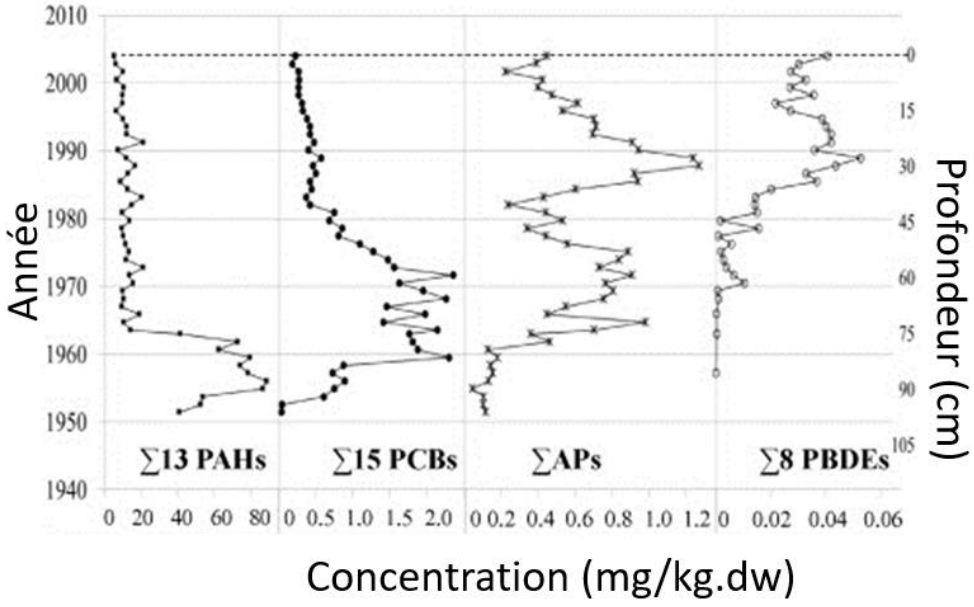


Figure 17: Concentrations des polluants organiques mesurées à Bouafles, figure issus de Lorgeoux et al., 2016

Cette approche par les archives sédimentaires peut être utilisée pour comparer plusieurs fleuves. Dendievel et al (2020) a ainsi comparé la dynamique des PCBs sur les fleuves français les plus importants (Seine, Rhône, Loire et Garonne) pour la période 1945-2018. Deux tendances ont ainsi été mises en évidence. La première est une augmentation rapide de 1945 à 1975 atteignant une concentration jusqu'à 4 $mg.kg^{-1}$ de sol sec suivi d'une forte diminution dans les années 1980 pour la Seine et la Loire. La seconde concerne le Rhône et la Garonne avec une augmentation modérée suivi d'une diminution après les années 1990. Le Rhône est le fleuve ayant présenté une charge en PCBs jusqu'à 25 % supérieure à la Seine et la Loire entre 1977 et 1987. Les flux français de PCB ont contribué à la contamination des mers et sont inférieurs de deux ordres de grandeurs à ceux trouvés dans les rivières américaines ou asiatiques (Dendievel et al., 2020).

La même approche a été appliquée aux ETM. La Seine présente une pointe de contamination par les métaux en Seine vers 1940-1960, suivie d'une forte baisse des concentrations qui est la conséquence de la désindustrialisation du bassin et de la mise en place des premières réglementations sur les rejets directs en rivière (Le Cloarec et al., 2011). C'est une dynamique temporelle qu'on retrouve sur les fleuves français (Dendievel et al., 2022). De même, on retrouve pour ces fleuves la dynamique amont-aval, avec un amont peu anthropisé et donc peu contaminé, et un aval plus contaminé par la présence de zones urbaines et d'industries. Si on compare la Seine à d'autres rivières françaises en utilisant les facteurs d'enrichissement, la Seine est plus contaminée que le Rhône mais moins que la Meuse (Dendievel et al., 2022) et de même en termes de concentration (Ayrault et al., 2020).

En conclusion de cette partie sur le bassin versant de la Seine, les archives sédimentaires sont des éléments essentiels pour l'analyse des trajectoires des contaminants et notamment dans ce bassin versant fortement anthropisé. L'étude d'un bassin nécessite la recherche de sites permettant de dater et retracer les usages des contaminants en utilisant également l'histoire pédologique de ce bassin.

1.3 Les micro-organismes présents dans les sols

Le sol est un écosystème complexe qui abrite une grande diversité de micro-organismes. Dans le chapitre précédent, nous avons démontré l'importance des sédiments en tant qu'archives sédimentaires. Ici, la complexité biologique de ces sédiments ayant évolués pour constituer un sol va être abordée ainsi que les paramètres impliqués.

L'association française pour l'étude des sols (AFES) édite un référentiel pédologique (le dernier est de 2008) qui constitue une base de référence pour nommer et désigner les sols. La définition des sols déterminée par l''AFES est la suivante, je cite : « *Le sol est un volume qui s'étend depuis la surface de la Terre jusqu'à une profondeur marquée par l'apparition d'une roche dure ou meuble, peu altérée, ou peu marquée par la pédogenèse. L'épaisseur du sol peut varier de quelques centimètres à quelques dizaines de mètres, ou plus. Il constitue, localement, une partie de la couverture pédologique qui s'étend à l'ensemble de la surface de la Terre. Il comporte le plus souvent plusieurs horizons correspondant à une organisation des constituants organiques et/ou minéraux (la terre). Cette organisation est le résultat de la pédogenèse et de l'altération du matériau parental. Il est le lieu d'une intense activité biologique (racines, faune et micro-organismes ».*

1.3.1 La diversité des sols

1.3.1.1 Les paramètres influençant la diversité

Les micro-organismes présents dans les sols dépendent du type de sols pédologiques mais également des paramètres physico-chimiques et du climat. A grande échelle, les régions tropicales et arctiques ont des modèles biogéographiques distincts où les régions tropicales, chaudes et humides, ont généralement une activité métabolique et une richesse en espèces plus élevées (Crowther et al., 2019). La distribution de la composition des communautés bactériennes semble être davantage liée à des facteurs locaux tels que le type de sol et la couverture du sol qu'à des facteurs plus globaux tels que les caractéristiques climatiques et les caractéristiques géomorphologiques (Dequiedt et al., 2009 ; Hermans et al., 2017). Les facteurs environnementaux influençant le plus la distribution des organismes dans le sol, sont ordonnés ci-après selon leur influence décroissante sur la distribution des phyla : le pH , la gestion des terres, la texture du sol, les nutriments du sol et le climat (Karimi et al., 2018). Prenons en exemple une augmentation du pH, Karimi et al. montrent l'influence positive de ce paramètre sur la diversité taxonomique de neuf phyla, une influence négative pour huit phyla et sans impact pour 3 phyla (Karimi et al., 2018). L' étude à grande échelle via un échantillonnage intensif et systématique constituerait une contribution prometteuse pour comprendre le déterminisme et la biodiversité du sol puisque ne serait-ce qu'à l'échelle de la France, les modèles biogéographiques sont très diversifiés (Karimi et al., 2018).

Le laboratoire du professeur Thomas Crowther a édité une carte indiquant la répartition du pH des sols dans le monde. L'échelle de pH s'étend de pH 4 dans les zones de latitude correspondantes à celles de la Russie, à 9 dans les zones dont la latitude est plutôt au niveau de celle de l'Afrique du nord et centrale (Figure 18A). Concernant la distribution de la biomasse bactérienne, elle s'échelonne de 25 à 350 ng PLFA (acides gras phospholipidiques)/ g de sol. La localisation géographique de la région la plus riche est plutôt en bord des océans (Figure 18B). La péninsule de la Norvège et Suède et une partie du Royaume Uni se détachent particulièrement du reste du monde du fait de sa forte concentration en biomasse bactérienne du sol (Figure 18C). L'ensemble des données enregistrées pour constituer les différentes cartes est une mine d'or pour l'exploration de l'influence des différents paramètres environnementaux sur la modulation des caractéristiques des sols et de la végétation.

L'influence du climat est aussi très importante sur la diversité des micro-organismes. En effet, dans les zones tempérées, la diversité génétique des bactéries est plus élevée mais pas celle des champignons suggérant un impact plus grand des variables environnementales que la capacité de dispersion pour déterminer les distributions mondiales (Bahram et al., 2018). Au niveau de la diversité et la géographie des champignons, la diversité est plus élevée dans les écosystèmes tropicaux avec une endémicité plus élevée alors que l'étendue géographique des organismes augmente vers les pôles (Tedersoo et al., 2014). La diversité des champignons dépend de conditions tels que le pH, le calcium et le phosphore mais est également liée à l'abondance des plantes hôtes plutôt qu'à la diversité de ces hôtes (Tedersoo et al., 2014). Par exemple, la teneur en carbone du sol ainsi que la couverture végétale déterminent l'uniformité et la diversité en protéines fongiques du sol alors que la richesse en protéines est plus corrélée à la température annuelle moyenne et au pH. De même, la richesse en protéines des champignons varie selon les zones étudiées : elle est par exemple plus élevée dans les zones de forêts que dans les zones d'arbustes (Parente Fernandes et al., 2022).

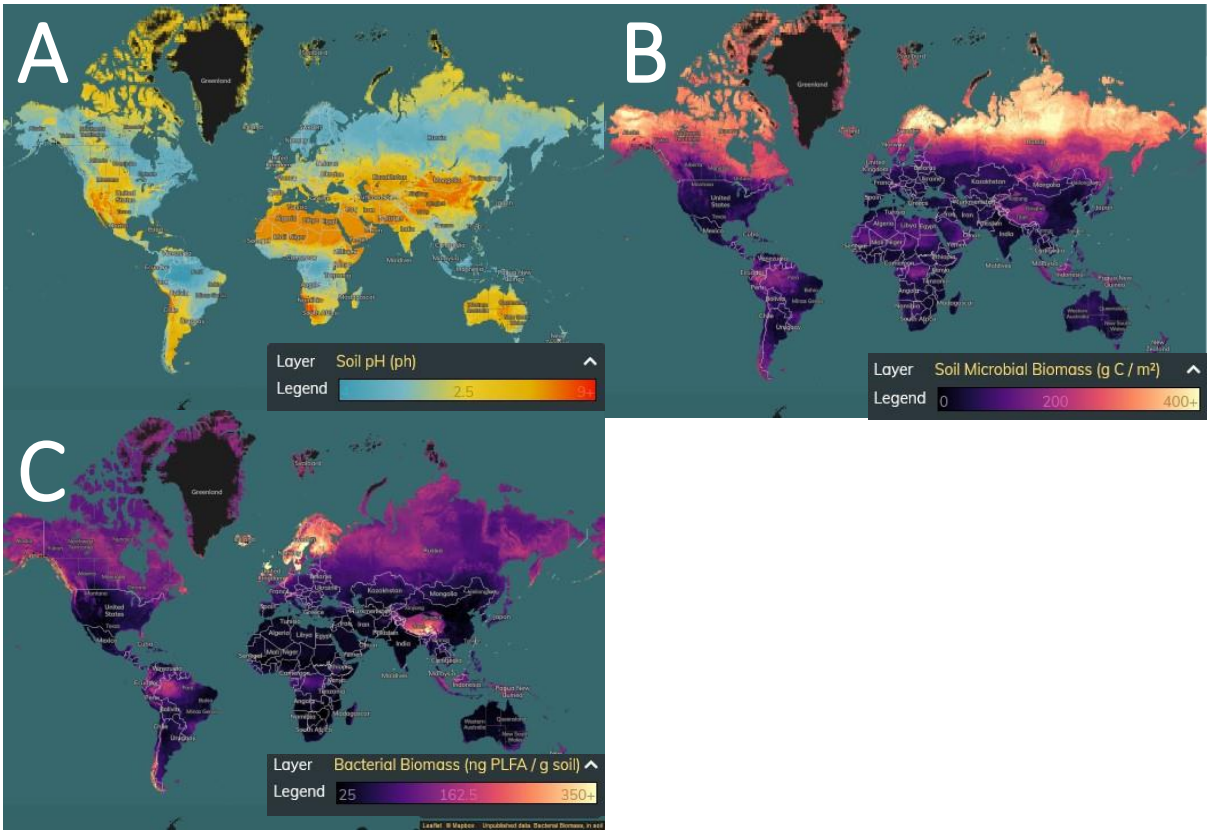


Figure 18 : Carte de la distribution du pH dans les sols (A), de la biomasse microbienne (g C / m²) (B) et de la biomasse bactérienne (ng PLFA / g sol). Cartes issues du site du Laboratoire de Crowther (https://crowtherlab.com/maps/).

1.3.1.2 Les micro-organismes les plus abondants

En 1934, Baas Becking a écrit « Tout est partout, mais l'environnement sélectionne » (Baas Becking, 1934). Pour comprendre la biogéographie des communautés microbiennes du sol dans leur ensemble, Karimi et ses collaborateurs ont étudié plus de 2173 échantillons de sol couvrant la France entière via une analyse de la séquence de l'ARNr 16S. Cette analyse de grande ampleur a mis en évidence 32 phyla bactériens et 3 phyla d'archées, dont 20 phyla sont cosmopolites et abondants (Figure 19)(Karimi et al., 2018). Ce sont les organismes les plus cosmopolites qui sont les plus abondants (Nemergut et al.,

2011) et l'abondance de chaque phyla présente une distribution hétérogène et spatialement structurée (Karimi et al., 2018).

Les actinobactéries, parmi les plus répandues, auraient une abondance principalement déterminée par la latitude puis par le pH mais qui n'est pas affectée par les conditions climatiques selon une étude portant sur les avant-pays des glaciers (Zhang et al., 2016). D'autres phyla comme les Proteobacteria, Acidobacteria, Planctomycetes, Bacteroidetes, et Firmicutes présentent une sensibilité au pH, au rapport C/N et à la teneur en phosphore (Hermans et al., 2017).

Au contraire, certains micro-organismes sont endémiques à des régions (Cho et al., 2000) ou des types d'habitat tel que les sources chaudes (Whitaker et al., 2003), c'est-à-dire, des organismes dont la présence est limitée à un lieu, une région ou un type d'habitat particulier (Hanson et al., 2012).

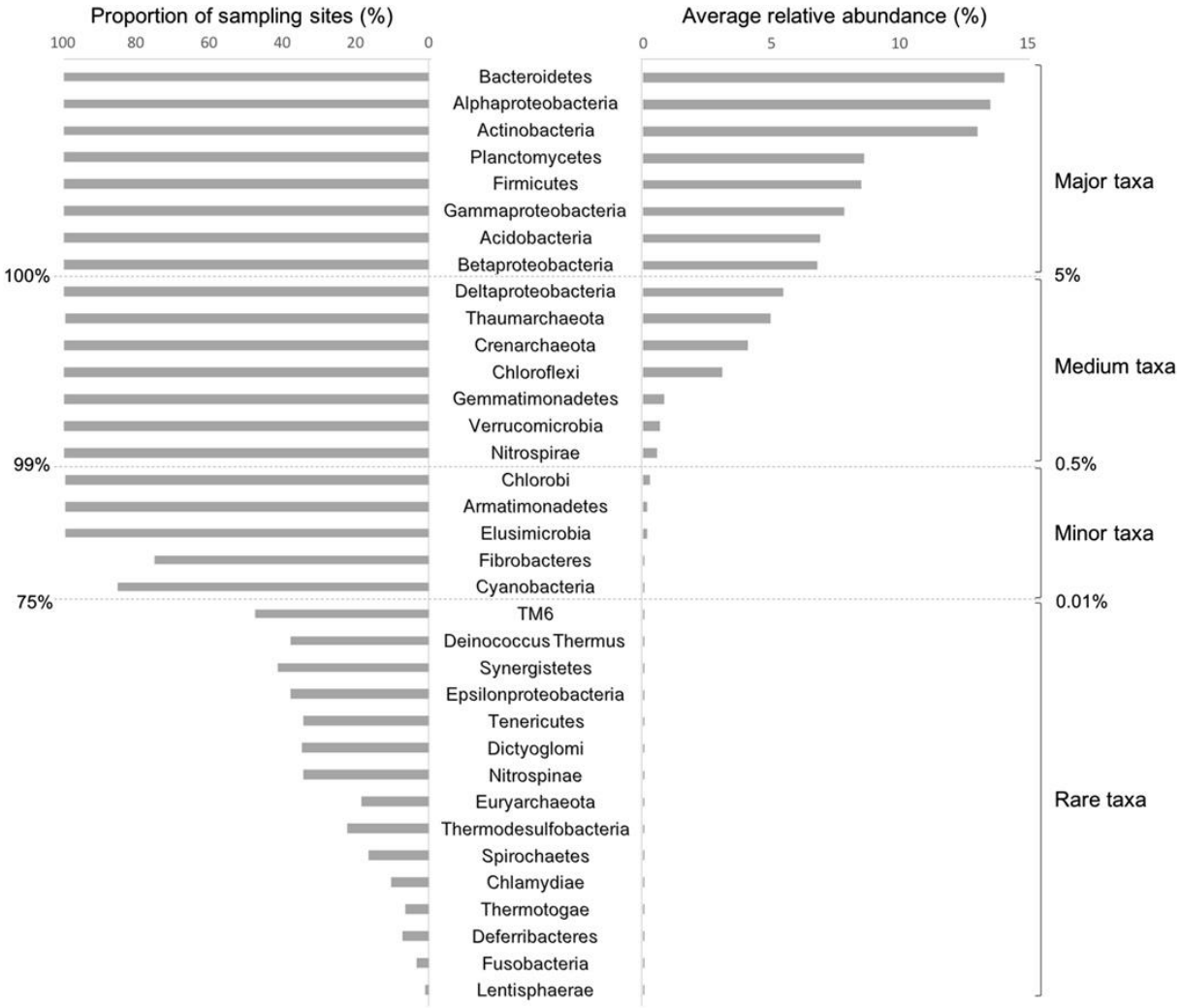


Figure 19: Proportion et abondances des phyla bactériens et archéens dans les sols français. A gauche, la proportion de sites d'échantillonnage où les phyla étaient présents et à droite, l'abondance relative des phyla (Karimi et al., 2018).

1.3.1.3 Etudier la diversité à l'échelle locale

La diversité à l'échelle locale est contrairement à l'échelle globale, dépendante de l'historique du site constitué par les différents évènements climatiques et environnementaux. Dans un gramme de sol, des milliards de bactéries, des dizaines de milliers de protistes, des kilomètres d'hyphes fongiques sont

présents (Geisen et al., 2019). A l'inverse des études visant à élucider les modèles de distribution des organismes, la plupart des études sont focalisées, soit sur le rôle d'un organisme dans son écosystème ou l'étude de la complexité d'un milieu contaminé ou sur les processus et les cycles biogéochimiques.

Bastida et al. ont étudié l'impact de la déforestation à long terme sur une communauté bactérienne sous un climat semi-aride et ont conclu à une perte de biomasse bactérienne et d'activité enzymatique et une augmentation de diversité notamment la présence de protéines de cyanobactéries (Bastida et al., 2015). Dans les sols restaurés par l'application d'amendements organiques, Bastida a également mis en évidence le lien entre les processus écosystémiques, les fonctionnalités cellulaires et les modes de vie des communautés microbiennes (oligotrophie et copiotrophie) (Bastida et al., 2015).

Ces études impliquent l'analyse en profondeur de la diversité taxonomique, des fonctions potentielles et de la diversité fonctionnelle selon les méthodes utilisées. C'est ainsi que sont déchiffrés les mécanismes mis en jeu par les micro-organismes pour répondre à un changement ou simplement perdurer dans un milieu.

1.3.2 La dynamique du sol

Les capacités d'adaptation des micro-organismes aux changements liés à des contaminants ou des modifications physico-chimiques dépendent du potentiel fonctionnel des communautés. A travers des études sur la présence de contaminants dans les sols et l'étude du continuum du sol dans le temps ou en profondeur, je vais illustrer cette dynamique.

1.3.2.1 L'influence des contaminants sur les micro-organismes

Certains éléments traces métalliques tels que le Cu, Zn, Ni, Cr, Co, Mo, Fe, et Mn sont des micro-nutriments essentiels pour les organismes et sont également nocifs à doses élevées. Lorsque les concentrations sont élevées, les métaux induisent des effets physiologiques, biochimiques et génotoxiques (Jaishankar et al., 2014). Cela peut se traduire par l'inhibition de fonctions métaboliques, la dénaturation de protéines ou des modulations du matériel génétique en inhibant par exemple le processus de transcription (Jacob et al., 2018). Par ailleurs, l'aluminium, qui est un métal non essentiel toxique, provoque l'inhibition d'enzymes telles que l'hexokinase et la phosphoxydase chez les micro-organismes (Barabasz et al., 2002).

Les communautés fonctionnelles bactériennes impliquées dans les processus d'absorption des métaux possèdent des groupes fonctionnels tels que les groupes hydroxyle, carboxyle, sulfonate, amide et phosphonate (Jacob et al., 2018). Ces communautés sont intéressantes et étudiées pour la remédiation d'environnements contaminés. La bioremédiation consiste à adsorber, réduire ou éliminer les contaminants de l'environnement par le biais des ressources biologiques tels que les micro-organismes ou les plantes.

Pour répondre à la présence de contaminants ou d'autres changements, les micro-organismes mettent en place des mécanismes de défense comme la sécrétion d'enzymes telles que les oxydoréductases ou les oxygénases qui influencent les taux de biorémédiation (Jacob et al., 2018).

1.3.2.2 L'évolution du sol au cours du temps

Le sol n'est pas inerte mais est en perpétuelle évolution. Par exemple, le temps de génération d'un Bacillus est de l'ordre de seulement 30 minutes. Les phénomènes de sélection, de compétition vont contraindre les micro-organismes à maintenir un certain dynamisme et répondre aux changements. Schneider et ses collaborateurs ont étudié les principaux acteurs microbiens dans la décomposition de la litière en termes d'abondance et d'activité au cours des saisons qui est stimulée par la teneur élevée en azote et phosphore (Schneider et al., 2012). Les Ascomycota sont les principaux décomposeurs fongiques dans les premiers stades de décomposition, ensuite ce sont les Basidiomycota qui

deviennent les champignons majoritaires (Schneider et al., 2012). Les auteurs émettent l'hypothèse selon laquelle la partie active de la communauté microbienne a changé au fil du temps. En agriculture, l'amendement de sol est courant pour prévenir notamment les carences en éléments chimiques essentiels aux plantes comme le phosphore qui joue un rôle important dans la productivité (Bastida et al., 2019). Ces amendements entraînent des modifications au niveau des communautés et notamment pour les Verrumicrobia où leur abondance, visualisée par leurs protéines, diminue. Cette diminution avait déjà été observée dans des conditions d'augmentation d'azote, de phosphore et de potassium disponible dans le sol (Huang et al., 2011).

1.3.2.3 Profil de profondeur d'un sol

La plupart des études menées sur le sol sont des analyses réalisées uniquement sur les premiers centimètres de 0 à 10 cm de profondeur. Quelques études se sont concentrées sur l'étude à différentes profondeurs. Zhang a analysé la composition en communautés bactériennes à différentes profondeurs pour tester le paillage au lieu d'une fertilisation artificielle (Zhang et al., 2019). Il a montré une évolution distincte des communautés en surface et en profondeur. Xiao a travaillé sur deux carottes de sol contaminé par l'arsenic et l'antimoine issues de deux types de sol. Les communautés microbiennes du sol étaient plus similaires à l'intérieur des carottes qu'entre les deux types de sols carottés (Xiao et al., 2017). Ainsi via l'utilisation de réseaux de co-occurrence, Xiao identifie des organismes fortement liés à la fraction contaminé (Xiao et al., 2017). La composition des communautés microbiennes du sol est différente selon la profondeur. Les phénomènes de diffusion verticale peuvent également impacter la diversité des communautés le long de la carotte.

En conclusion, la géodistribution des micro-organismes dans les sols dépendent de nombreux paramètres, tels que le pH, qui impacteront chacun à leur manière les différentes communautés microbiennes présentes. Ces sols sont encore à ce jour peu connus et les modèles régissant les interactions et la diversité fonctionnelle et taxonomique sont méconnus. Cela est dû aux limites des méthodologies actuelles utilisées pour les étudier. Les principales techniques sont le métabarcoding, la métagénomique et la métaprotéomique qui seront décrites dans les prochains paragraphes. Cependant, ce savoir pourrait par exemple servir d'indicateur biologique de l'état de santés des sols en se basant sur les abondances relatives des taxons (Hermans et al., 2017).

1.4 La métagénomique, outil d'exploration de la diversité microbienne

Bien que les micro-organismes et macro-organismes soient au centre du fonctionnement des écosystèmes, les organismes présents dans les sols restent mal connus que ce soient au niveau des souches et espèces identifiées, des interactions entre les organismes ou de l'impact des types de sols et des paramètres du milieu sur ces organismes. Les projets à grande échelle participent à diminuer cette méconnaissance des écosystèmes. Des projets tel que Tara Ocean (Louca et al., 2016 ; Sunagawa et al., 2015) pour les océans participent à la classification de milliers d'organismes dans des classes fonctionnelles et dénouent les facteurs qui façonnent les communautés bactériennes et archéales en analysant les profils taxonomiques et fonctionnels. Au niveau des sols, il y a par exemple le projet « Microbiome Stress Project » qui s'attèle à l'analyse des facteurs de stress environnementaux et de leurs effets sur les communautés microbiennes (Rocca et al., 2019).

Dans un premier temps, nous décrirons les approches moléculaires de métabarcoding et de métagénomique puis nous nous axerons sur les techniques de séquençage en métagénomique, les techniques d'assemblages et enfin les outils utilisés.

1.4.1 Les différentes approches moléculaires

Les progrès des technologies moléculaires de séquençage à haut débit permettent de révéler les structures des communautés dans les biomes du sol. Parmi les approches moléculaires, les méthodes ciblées tel que le métabarcoding ou les méthodes shotgun tel que la métagénomique permettent de décrire la biodiversité d'un sol (Figure 20). Le metabarcoding est la description de la diversité microbienne par l'utilisation d'une amplification d'un gène précis ou portion de gène tel que celui codant pour l'ARNr 16S. La métagénomique permet la description des génomes de nombreux organismes en terme de diversité mais également le potentiel de ces micro-organismes. Ces deux approches couramment utilisées sont clairement à différencier. La plupart des espèces microbiennes n'ont pas été obtenues en culture pure et ne sont étudiées que par des techniques indépendantes de la culture. Ces micro-organismes non encore cultivés, isolés, et caractérisés du point du vue taxonomique par les microbiologistes sont appelés la matière noire microbienne (Microbial Dark Matter, MDM) (Jiao et al., 2020). Certains micro-organismes sont identifiables uniquement sur la base de la séquence de l'ARN ribosomal 16S. Les avancées technologiques en culturomique, génomique, métagénomique et dans le domaine émergent du séquençage de cellule unique ainsi qu'une réduction des coûts de séquençage et d'analyses permettent d'accroître le nombre d'organismes actuellement séquencé (Lasken et McLean, 2014).

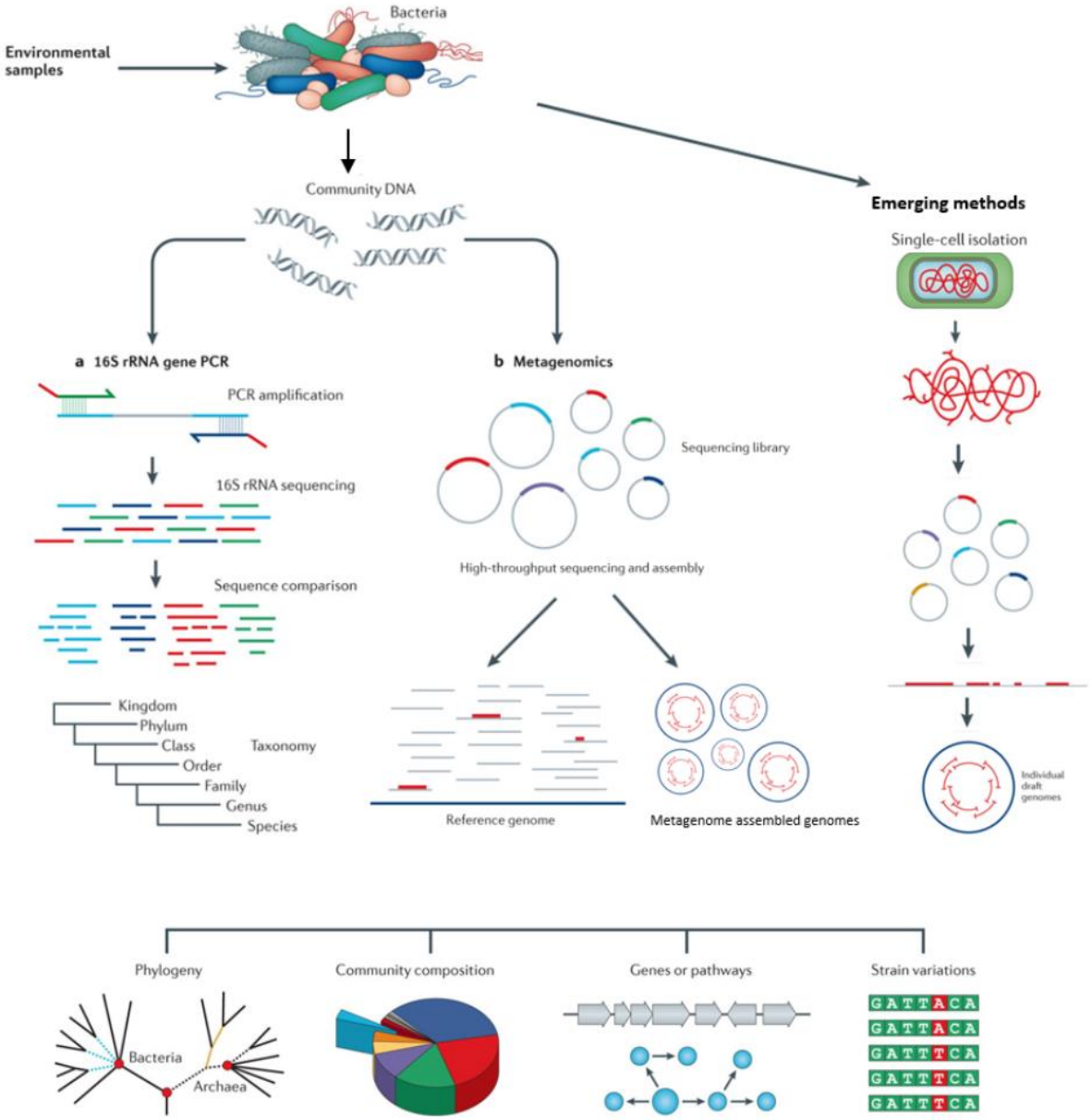


Figure 20: Analyses d'échantillons environnementaux tel que le sol en utilisant les approches moléculaires, adapté de Lasken et McLean, 2014.

1.4.1.1 Les méthodologies ciblées : l'analyse de l'ARNr 16S

Le ribosome est un complexe ribonucléoprotéique en charge de la production des protéines (traduction de l'information des ARNm). Il est composé d'une grande (50S) et une petite sous-unité (30S) chez les bactéries. Cette dernière est composée d'ARN ribosomique 16S, codé par un gène de longueur 1500 nucléotides environ, et d'une vingtaine de protéines. La séquence du gène codant l'ARNr 16S est couramment utilisée pour l'identification des espèces bactériennes et réaliser des études taxonomiques, car sa contrainte évolutive est très forte de par le complexe qu'elle constitue avec les protéines ribosomiques associées et elle est représentative de l'évolution de l'ensemble des bactéries. La séquence est composée de neuf régions hypervariables (V1-V9) et de segments conservés chez la plupart des bactéries permettant une amplification par PCR des séquences cibles à l'aide d'amorces universelles (Chakravorty et al., 2007). Les régions hypervariables ont des degrés de diversité de séquence et l'usage d'une unique région ne permet pas de distinguer toutes les bactéries.

Par exemple, les régions V3-V4 et V4-V5 semblent produire des résultats de profils bactériens plus reproductibles (Teng et al., 2018).

L'analyse de la séquence du gène de l'ARNr 16S par PCR à l'avantage d'être peu coûteuse et c'est une analyse à haute résolution et à haut débit. Cependant, l'efficacité de l'amplification dépend de l'environnement de la séquence de l'amorce et donc peut fluctuer en fonction des microorganismes, ce qui fausse la structure observée de la communauté (Knight et al., 2018). L'identification taxonomique est basée sur une recherche de similarité de séquence par rapport à une base de données listant l'ensemble des séquences d'ARNr16S connus à ce jour. Des outils tels que mothur (Schloss et al., 2020) ou FROGS (Escudié et al., 2018) peuvent ensuite être utilisés pour analyser les séquences de l'ARNr 16S d'un échantillon environnemental et ainsi identifier les groupes d'organismes, identifiés sous le terme d'unités taxonomiques (OTU, operational taxonomic unit), qui sont créé en fonction d'un degré de similarité de séquences au sein de l'échantillon ou sous le terme de phylotypes où les séquences sont comparés et assignés à la base de données. Les OTUs sont utilisés comme proxy des espèces avec un seuil de clusterisation généralement utilisé de 97 % d'identité et de 95 % pour les genres (Schloss et Handelsman, 2005). Proposé en 1994 par Stackebrandt (Stackebrandt et al., 1994), ce seuil est aussi discuté aux vues des avancées technologiques (Edgar, 2018 ; Johnson et al., 2019). L'outil mothur (Schloss et al., 2009), un des plus cités pour l'analyse des séquences du gène de l'ARNr 16S, permet l'analyse complète de plusieurs échantillons simultanément séquencés.

1.4.1.2 Les méthodologies shotgun : la métagénomique

Les nouvelles technologies de séquençage permettent aujourd'hui d'analyser à haut débit des échantillons d'origines environnementales via le séquençage du matériel génétique extrait des communautés microbiennes. Ne se limitant pas à une région cible du génome, elles permettent l'analyse taxonomique ainsi que l'étude du potentiel fonctionnel des communautés et de leurs structures. Le premier métagénome shotgun de la mer des Sargasses a été publié par Craig Venter et ses collaborateurs (Venter et al., 2004), pionniers en métagénomique.

La métagénomique utilise les technologies de séquençage d'acides nucléiques tels que ceux utilisé en génomique ou transcriptomique à la différence que la complexité et la diversité de l'échantillon nécessite l'utilisation d'outils bioinformatiques spécifiques pour l'analyse des données produites. L'étude d'un échantillon se fait dans un premier temps par l'extraction de l'ADN de l'échantillon par des traitements physiques, chimiques ou enzymatiques. Puis, l'ensemble du contenu génomique des communautés environnementales est séquencé.

Les avantages de la métagénomique sont sa haute résolution et la possibilité d'analyser simultanément tous les organismes identifiables par le biais de leurs gènes, ce qui permet d'obtenir une image du potentiel fonctionnel de l'échantillon, résolu taxonomiquement ou non. Les désavantages restent le coût élevé d'une analyse et la quantité de données à traiter certaines n'étant que peu informative ou ne serait-ce qu'attribuable (Delmont et al., 2010 ; Geisen et al., 2019).

Le cas particulier du sol :

Particulièrement, le sol est un habitat complexe qui de par ses spécificités induisent des problèmes et des limites lors de l'utilisation de la métagénomique (Semenov, 2021):

- L'hétérogénéité de la couverture du sol induit des échantillons de sol non représentatifs et difficilement reproductibles,
- L'hétérogénéité des communautés microbiennes au sein des agrégats induit des échantillons de sol non représentatifs et nécessite d'augmenter le volume des échantillons et d'étudier des agrégats de différentes tailles,

- La présence d'argiles qui induisent une extraction incomplète de l'ADN du sol nécessite une meilleure homogénéisation des échantillons de sol,
- Les substances humiques inhibent l'amplification. Cet effet est réduit par des étapes supplémentaires de purification et de dilution de l'ADN,
- La biodiversité étant riche, la profondeur de séquençage doit être élevée,
- La présence de reliques d'ADN anciens complexifie l'identification des organismes actifs, cet effet pouvant être réduit en éliminant l'ADN extracellulaire.

1.4.2 Les technologies de séquençages d'acides nucléiques

Le séquençage de l'ADN connait aujourd'hui sa troisième génération de technologie de séquençage. Pendant près de trois décennies, Sanger, une technologie de première génération a dominé le marché du séquençage mais son coût élevé et un temps d'analyse long a laissé place à la seconde génération de séquençage dès 2005. Cette dernière est généralement nommée NGS (Next Generation Sequencing) ou HTS (High-Throughput Sequencing) et a réduit le coût du séquençage en générant des millions de lectures courtes. Les principaux acteurs de cette seconde génération sont les technologies Roche / 454, Illumina / Solexa, ABI / SOLiD ainsi que la technologie Ion Torrent. La troisième génération axée sur le séquençage de lectures plus longues ne nécessite pas d'étape d'amplification par PCR. L'augmentation de la longueur des lectures rend possible l'assemblage des régions répétées qui étaient complexes pour les séquenceurs de deuxième génération (Morisse, 2019).

1.4.2.1 La technologie Illumina

La technologie Illumina est la technologie de séquençage qui s'est imposée dans le domaine du séquençage de l'ADN. C'est une technologie peu coûteuse présentant un taux d'erreur de 0.1 à 1%, en moyenne 0.76% (Quail et al., 2012). Ces erreurs de séquençage ne sont généralement pas des insertions qui provoqueraient des décalages de cadres de lectures lors de la traduction mais plutôt principalement des substitutions. La nature de ces erreurs et leurs taux faibles ajoutés au faible coût ont permis à Illumina de s'imposer comme la technologie de séquençage de deuxième génération.

Le système HiSeq 4000, par exemple, permet de générer entre 1300 et 1500 Gb en 1 à 3.5 jours avec des lectures de 150 bases en une analyse dont le score de qualité Phred est supérieure à une valeur de 30 dans plus de 75% des bases, c'est-à-dire 99.9 % de chances que la base ait été correctement assignée (données constructeur, https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/hiseq-3000-4000-specification-sheet-770-2014-057.pdf).

Cependant, les lectures courtes sont plus difficiles à assembler en un contig ou génome ou jeu de génomes tel un puzzle composé de petites pièces, notamment dans certaines régions du génome fortement répétées.

1.4.2.2 Les technologies long-reads

Deux technologies concurrentes sont disponibles pour de longues lectures de séquences d'acides nucléiques : PacBIO (Pacific Biosciences) et Oxford Nanopore. Considérés comme la troisième génération de séquençage, elles permettent de générer des fragments de séquences de plusieurs dizaines de kb. Plus longues, de telles séquences permettent de couvrir les régions répétées de l'ADN mais ont un taux d'erreurs compris entre 10 et 30% avec des erreurs de types insertions et délétions plus fréquentes. Idéalement, le séquençage long-read et le séquençage court-red se complètent et peuvent permettre d'obtenir des génomes entiers de qualité.

1.4.3 Les méthodologies d'assemblage

1.4.3.1 Qu'est-ce qu'un assemblage ?

L'assemblage est une étape majeure, elle permet de reconstruire la séquence initiale d'ADN du ou des organismes séquencés. La tâche consiste à décomposer les lectures de 100 à 600 nucléotides en k-mers, c'est-à-dire, en fragments homogènes de taille k qui sont ensuite utilisés dans un graphe, le graphe de De Bruijn, en tant que nœuds. L'objectif de l'assembleur est de rechercher le chemin passant par toutes les arêtes du graphe. Afin d'évaluer la qualité d'un assemblage, différentes constantes sont calculées avec notamment le N50 qui correspond à la longueur du fragment qui lorsque l'on met bout à bout l'ensemble des fragments assemblés triés par ordre décroissants de longueur sépare en deux groupes ayant le même nombre de base. Le deuxième indicateur est la longueur du plus grand fragment assemblé.

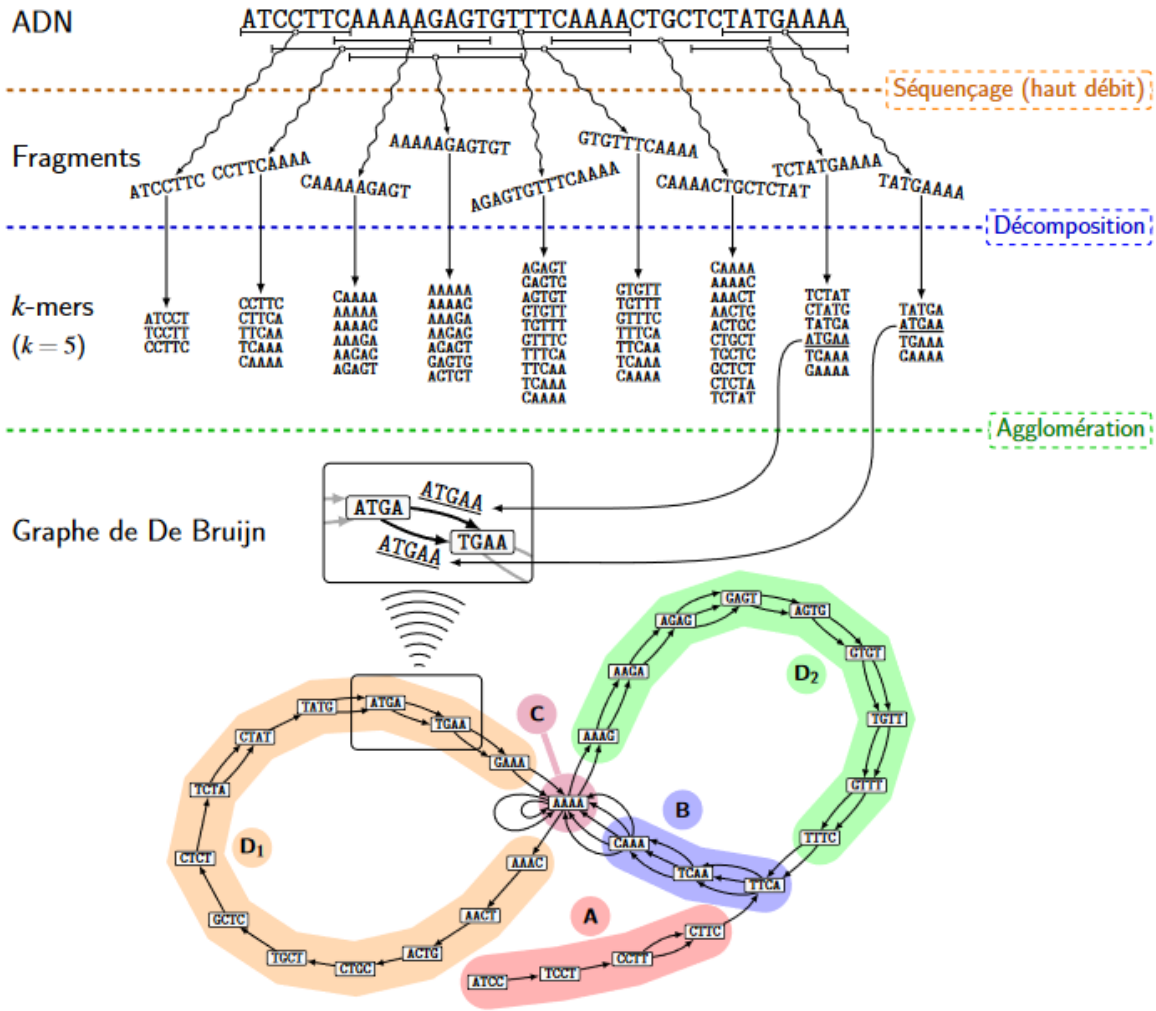


Figure 21: Les différentes étapes constituant l'assemblage ainsi que de la création du graphe de De Bruijn (source Ponty, 2014).

Par défaut, l'assemblage de métagénomes est plus difficile puisque de nombreux organismes taxonomiquement proches peuvent être séquencés simultanément induisant la présence de gènes orthologues. Dans le cas d'un unique génome, l'algorithme d'assemblage utilise la couverture de séquence pour identifier les copies répétées ou distinguer les « vraies » séquences des erreurs de séquençage en supposant que la couverture de séquence le long du génome sera approximativement

uniforme (Quince et al., 2017). Dans le cas de métagénomes, la couverture dépend aussi de l'abondance de chaque génome dans la communauté microbienne étudiée.

1.4.3.2 Assemblage sans génome de référence

Contrairement à l'étude génomique d'un organisme séquencé où le programme qui assemble les séquences peut utiliser le génome de l'organisme cible pour venir cartographier les lectures sur le génome et ainsi assembler les données en séquences continues, l'algorithme *de novo* utilise les graphes de De Bruijn seulement pour assembler les données (*Figure 21*).

MEGAHIT (Li et al., 2015), par exemple, est un algorithme d'assemblage *de novo* spécialiste des données métagénomiques ayant pour vocation de limiter les coûts en temps et en mémoire de cette étape. Contrairement à d'autres algorithmes tel que SOAPdenovo2 (Luo et al., 2012) ou IDBA-UD (Peng et al., 2012) qui peuvent nécessiter jusqu'à au moins 4 TB de mémoire requise pour assembler des données métagénomiques de sol, MEGAHIT utilise seulement 768 Go avec un seul nœud (Li et al., 2015).

1.4.3.3 Assemblage des organismes présents dans le métagénome, les MAGs

Les génomes des microorganismes présents dans le métagénome peuvent être reconstruits à partir de l'assemblage des données métagénomiques, appelé génomes assemblés à partir de métagénomes (metagenome-assembled genomes ou MAGs). Après un contrôle de qualité des données brutes, les données métagénomiques sont assemblés puis filtrés pour ne garder que les fragments assemblés les plus longs, contigs ou scaffolds. La complétude et la contamination sont ensuite évaluées en utilisant par exemple CheckM (Parks et al., 2015) qui utilise des gènes ubiquitaires. Les séquences issues d'une même population microbienne peuvent être fusionnés. CheckM, par exemple, utilise le taux de GC moyen, la couverture et la classification identique. Par exemple, Parks et al. ont décrit l'assemblage de 7903 MAGs à partir de 1 550 métagénomes publics dont 93 % ont une couverture moyenne supérieure à 10x et un degré de contamination inférieur à 5 % et de complétude supérieure à 70 % (Parks et al., 2017).

Cette méthode est émergente. Avec 39 articles en 2018, près de 183 articles scientifiques ont été publiés avec les mots « MAGs » et « metagenome » sur PubMed en 2021 avec par exemple les études suivantes (Parks et al., 2020 ; Xue et al., 2020 ; Tully et al., 2018). Les outils et les techniques utilisés pour construire les MAGs à partir des technologies short-reads et long-reads sont indiqués par Yang (Yang et al., 2021).

1.4.3.4 Assemblage sans graphe de De Bruijn

Afin d'augmenter l'efficacité des algorithmes d'assemblage, certains algorithmes s'affranchissent de l'utilisation de graphes ce qui computationnellement permet à PLASS (Steinegger et al., 2019) de s'exécuter en temps linéaire. L'algorithme de PLASS (Steinegger et al., 2019) traduit les cadres de lecture ouverte en séquences protéiques et réalise l'assemblage en se basant sur le calcul de chevauchement d'alignements complets et non seulement de k-mers comme avec les graphes de De Bruijn.

1.4.3.5 Les défis et les limites de la méthode d'assemblage

Sur quels critères juger la qualité d'un assemblage ? Des thèses se sont axées sur cette problématique ainsi que sur l'efficacité de corrections de données de séquençage (Morisse, 2019). Comme mentionné précédemment, les principales limites de la méthode d'assemblage en métagénomique sont les coûts en temps et en mémoire.

Parmi les défis de l'assemblage de données métagénomiques, il y a l'assemblage de novo de génomes étroitement apparentés, c'est-à-dire, présentant une identité nucléotidique moyenne supérieure ou

égale à 95 % (Sczyrba et al., 2017). Ces génomes d'organismes proches ajoutent de nombreuses branches dans le graphe d'assemblage où les séquences peuvent différer d'un seul nucléotide, d'un gène ou d'un opéron entier (Quince et al., 2017).

Dans le cadre d'échantillons complexes, un optimum sur la profondeur de séquençage doit être trouvé. En effet, une faible profondeur de séquençage ne permet pas de couvrir les génomes des microorganismes de faible abondance mais dans le cas contraire, le temps de calcul ou la mémoire lors de l'étape d'assemblage de données trop denses sont des facteurs très limitants (Quince et al., 2017).

Des études et des outils permettent d'évaluer la qualité des génomes assemblés (Mikheenko et al., 2016) et des méthodologies permettant de valider l'assemblages de métagénomes (Olson et al., 2019 ; Vollmers et al., 2017).

Récemment les méthodes d'assemblages hybrides mêlant les technologies de séquençage de lectures courtes et de lectures longues sont appliquées aux données métagénomique de sol (Xu et al., 2022). Dans cette étude, les trois types d'assemblages sont testées à partir des technologies de séquençage Illumina et PacBio. La méthode d'assemblage hybride présente les avantages des deux assemblages individuels, c'est-à-dire, la sensibilité de séquençage et l'intégrité des gènes. L'assemblage des lectures courtes fournit un plus grand nombre de contigs et de séquences de gènes alors que l'assemblage des lectures longues fournit des contigs plus longs et des séquences de gènes relativement intacts (Xu et al., 2022).

1.4.4 Les outils de métagénomique

1.4.4.1 Le challenge CAMI

Le domaine de la métagénomique a connu une croissance forte en termes de quantité de données et de méthodes. Les sites hébergeant des données sur les microbiomes tel que IMG/M (Chen et al., 2019), SRA (Leinonen et al., 2011 ; Katz et al., 2021) ou encore MG-RAST (Keegan et al., 2016) grossissent rapidement et disposent de centaines de milliers de jeux de données. IMG/M permet de trouver un génome à partir de la taxonomie ou d'un écosystème particulier ou de trouver un gène ou encore de comparer des génomes. Le SRA stocke les données brutes de séquençage et les informations d'alignement. Enfin MG-RAST stocke également les données brutes mais comporte un pipeline d'analyse taxonomique et fonctionnelle pour les analyser.

L'initiative CAMI (Critical Assessment of Metagenome Interpretation) a évalué la performance d'algorithmes d'assemblage, de regroupements des séquences et d'annotations taxonomiques en établissant des standards de comparaison (Sczyrba et al., 2017) ainsi que la définition de normes pour l'évaluation des performances d'un outil ainsi que l'étalonnage de ses performances. Dernièrement, CAMI a publié un pipeline et son tutoriel (Meyer et al., 2021) pour aider les développeurs de logiciels à comparer de façon homogène les nouveaux outils et leurs différentes versions.

1.4.4.2 Les outils de recherche de taxonomies

Différents outils permettent l'annotation taxonomique des génomes ou des gènes identifiés à partir d'échantillons environnementaux. Quel que soit la méthode, la limite principale de ces outils est la présence des organismes séquencés dans une base de données de référence qui vont permettre leurs identifications. Dans le cas échant, l'annotation sera réalisée dans des rangs taxonomiques plus élevés.

- Kaiju

Kaiju (Menzel et al., 2016) est basé sur la comparaison de séquences à une base de données de référence de protéines microbiennes pour l'annotation taxonomique. Il est capable d'analyser des millions de lectures par minute et être lancé sur un ordinateur de bureau avec un service en ligne.

- KRAKEN 2

KRAKEN (Wood et al., 2019) est une approche basée sur l'utilisation de k-mers et de calcul d'ancêtre commun le plus proche (lowest common ancestor, LCA) pour une classification taxonomique. Contrairement à des méthodes d'alignements, cet algorithme est basé sur la recherche de séquences spécifiques aux différents rangs taxonomiques sans notion de couverture ou de seuils d'identité.

- MG-RAST

MG-RAST (Keegan et al., 2016) est une ressource publique pour l'annotation et l'analyse de données métagénomiques mais aussi il s'agit d'un répertoire pour des dizaines de milliers de jeux de données accessibles publiquement. Le serveur permet à l'utilisateur de télécharger sur la plateforme les données brutes issues du séquençage et de les annoter taxonomiquement et fonctionnellement.

1.4.4.3 Les outils de recherche d'annotations fonctionnelles

Le séquençage du génome et son assemblage est souvent suivi par l'annotation systématique de la fonction des protéines basée sur l'hypothèse que des séquences similaires auront des fonctions similaires (Devos et Valencia, 2001). La recherche d'annotation permet d'annoter précisément le génome d'un organisme tel qu'un MAGS et ainsi d'en évaluer le potentiel fonctionnel. La prédiction de gène peut être effectué avec par exemple FragGeneScan (Rho et al., 2010) ou Prodigual (Hyatt et al., 2010).

En revanche, l'analyse fonctionnelle des échantillons via de l'enrichissement des gènes (en anglais gene-set enrichment analysis ou GSEA) permet d'analyser le potentiel fonctionnel de la communauté microbienne selon les paramètres mesurés et notamment de réaliser une analyse différentielle. L'utilisation d'algorithmes basés sur la similarité de séquences peuvent entraîner par dérives successives des erreurs d'annotations potentielles.

- DAVID

DAVID (Huang et al., 2009) est un outil d'annotation fonctionnelle centré sur les gènes. Ils sont ainsi classés en groupes fonctionnels où les termes annotés sont enrichis permettant d'explorer les annotations de chaque gène et d'explorer toutes les annotations pour un ensemble de gènes.

- PANTHER

PANTHER (Mi et al., 2013 ; Mi et al., 2015) est un outil d'annotation similaire à DAVID. L'outil combine l'annotation de la fonction des gènes, l'ontologie et les voies métaboliques.

- GhostKOALA

GhostKOALA (Kanehisa et al., 2016) sont des outils d'annotations des métagénomes permettant l'annotation fonctionnelle des séquences en termes d'annotation KEGG associée au métabolisme.

1.5 La métaprotéomique

1.5.1 Introduction sur les notions de protéomique et de métaprotéomique

La protéomique, la métaprotéomique et les techniques omiques offrent une vue intégrative des organismes d'un environnement afin de mieux le comprendre en termes de structure, de diversité, d'interactions, etc. Ces environnements contiennent une multitude d'organismes dont le métagénome n'est pas intégralement résolu. L'expression de leurs gènes aboutit à la production de protéines. Ces dernières sont les molécules effectrices des voies moléculaires activées pour le maintien de l'homéostasie cellulaire. La séquence des protéines est déterminée par la séquence du gène codant correspondant. Puisque la séquence des protéines est étroitement liée à celle de son gène d'origine, les méthodes d'analyses des protéines sont tout à fait adaptées pour distinguer les organismes entre eux et déterminer leurs caractères phylogénétiques spécifiques directement liés à leurs caractéristiques phénotypiques. Cette approche, dite de protéotypage, a été discutée dans une récente revue (Grenga et al., 2019).

1.5.1.1 La définition de protéome

Le protéome est constitué de l'ensemble des protéines synthétisées par un génome d'un organisme (Wilkins et al., 1996). L'étude des protéomes permet donc de déterminer les acteurs clés des différentes fonctions biologiques d'un organisme. Par exemple, lorsqu'un organisme est soumis à un stress externe, les protéines sont activées ou produites pour répondre à ce stress et atteindre l'homéostasie cellulaire pour maintenir l'organisme vivant. De ce fait, chaque protéine devient la signature du type de stress induit. C'est ainsi que les organismes ont pu s'adapter à leur environnement. Dans les environnements marins, *Ruegeria pomeroyi* est capable de s'adapter à l'influence anthropique ou aux faibles concentrations de nutriments grâce à l'expression de son génome dont près de 50% des protéines traduites est consacré à l'adaptation aux variations physiologiques cellulaires générales (Christie-Oleza et al., 2012). D'autres bactéries ont développé des résistances aux métaux grâce à l'expression de gènes codant pour des protéines capables de les capter. C'est le cas de la superfamille des protéines du type cupredoxines qui assure la résistance des cyanobactéries au cuivre (Dupont et al., 2011). De la même façon, les organismes eucaryotes s'adaptent à leur environnement en modulant l'expression de leurs gènes. La taille du génome varie, du plus petit chez les virus au plus grand chez les espèces eucaryotes supérieures. Cependant, la taille du génome n'est pas directement proportionnelle à la complexité du protéome. En effet, la chenille possède environ 20 000 gènes et les différentes protéines produites à partir de ces gènes sont dénombrées à pratiquement un million (Paczesny et al., 2014) traduisant une plus grande complexité au niveau du protéome, nécessaire pour une régulation fine du métabolisme de l'organisme.

1.5.1.2 La complexité du protéome

Le protéome est complexe et dynamique. Les protéines qui le composent ont des spécificités physico-chimiques qui leur confèrent des fonctions spécifiques. Certaines assurent l'homéostasie des espèces oxydatives à la fois importante pour la régulation des complexes mais aussi toxique lorsque le niveau de tolérance est dépassé. Les protéines en charge de la régulation des espèces oxydatives comme la famille des superoxydes dismutases sont dotée de sites actifs spécifiques, communs entre tous les organismes procaryotes et eucaryotes. D'autres assurent des fonctions liées à la structuration et l'intégrité cellulaires, la signalisation intra/inter-cellulaire et catalysent les réactions qui interviennent dans le métabolisme (Timp and Timp, 2020). La structure de la protéine détermine sa fonction et une modification peut entraîner son dysfonctionnement. Les niveaux d'expression de gènes et la durée de vie de la protéine, produit du gène, sont aussi important pour la régulation des processus biologiques. Par exemple, la chenille et le papillon possèdent le même génome mais la transformation de l'un à l'autre est liée à des différences d'expression au niveau du transcriptome et du protéome à l'origine

des modifications phénotypiques (*Figure 22*). L'analyse de la dynamique du protéome permet de mieux comprendre les mécanismes moléculaires impliqués dans les fonctions cellulaires et le phénotype des organismes étudiés. Deux stratégies peuvent être utilisées pour étudier le protéome et les mécanismes moléculaires : l'analyse de protéines spécifiques et isolées par des méthodes de biochimie et biophysique ou l'analyse à grande échelle des protéomes par des méthodes basées sur la spectrométrie de masse (Aebersold and Mann, 2016).

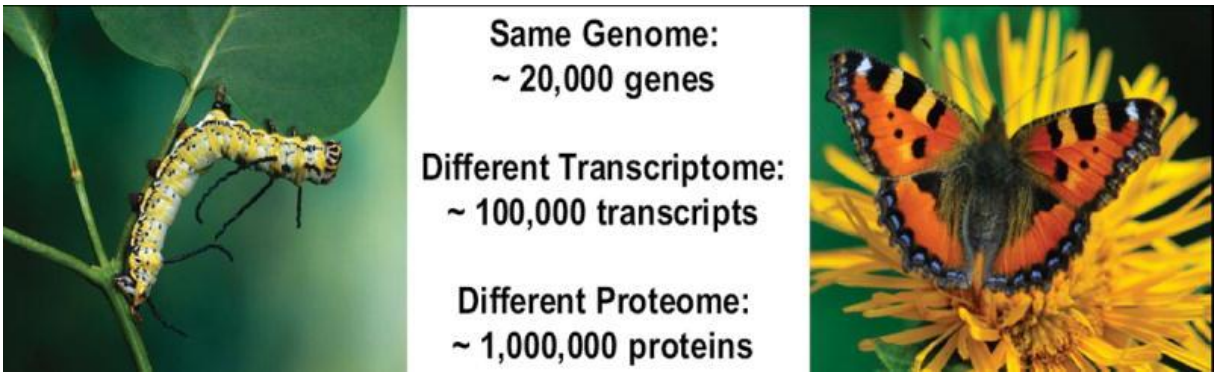


Figure 22: La complexité des omiques représentée par la transformation d'une chenille en papillon (source : Paczesny et al., 2014)

1.5.1.3 La protéomique

La protéomique vise à identifier et à quantifier les protéines d'un protéome, y compris l'abondance directement liée au niveau d'expression du gène, la localisation cellulaire, les interactions, les modifications post-traductionnelles (PTM) et le renouvellement en fonction du temps, de l'espace et du type de cellule (Zhang et al., 2013). Des projets de grande ampleur sont lancés tel que le Earth BioGenome Project (EBP) lancé en 2018 qui vise à séquencer toutes les espèces eucaryotes connues en dix ans (Lewin et al., 2022). La protéomique pourra donc être appliquée à l'ensemble de ces espèces.

Les domaines d'applications de la protéomique sont variés : elle peut être utilisée en recherche clinique et pharmaceutique, dans des études écotoloxicologiques, environnementales et animales. Lors d'étude à grande échelle, elle peut être utilisée pour une analyse intégrative et combinée à d'autres types de données telles que des données génomiques, transcriptomiques et métabolomiques ; l'ensemble sont des analyses dites multi-omiques.

1.5.1.4 La métaprotéomique

Dès lors que plusieurs organismes sont présents dans un échantillon, la caractérisation à grande échelle des protéines de l'échantillon à un moment donné est appelée la métaprotéomique (Herbst et al., 2015). L'ensemble des protéines de cet écosystème s'appelle le métaprotéome. Elle permet l'étude du fonctionnement des communautés microbiennes au sein d'un écosystème environnemental ou d'un hôte (Seifert & Muth, 2019). Les microbiotes environnementaux sont complexes et sont constitués d'une multitude d'espèces en majeure partie non cultivables intervenant notamment dans les processus métaboliques et dans les cycles biogéochimiques. La métaprotéomique fournit également des informations taxonomiques et fonctionnelles sur les différentes espèces et populations constituant le microbiote. Des échantillons biologiques d'origines diverses ont été analysé par métaprotéomique (Wilmes et al., 2015) provenant entre autres de systèmes industriels tels que les eaux usées de stations d'épuration (Wilmes et al., 2008), de digesteurs anaérobies (Hanreich et al., 2012), d'environnements océaniques (Saito et al., 2019), de sols (Tartaglia et al., 2020) mais aussi d'échantillons cliniques (Schaubeck et al., 2016) ou de fluides corporels (Feig et al., 2013).

1.5.2 La préparation d'échantillons

1.5.2.1 L'approche classique en protéomique

Dans le cadre d'analyse de protéomique classique, la préparation d'échantillons est une étape critique du bon déroulement de l'analyse et permet notamment l'extraction des protéines. Elle se décompose en plusieurs étapes : la lyse des cellules du micro-organisme avec une méthode mécanique ou chimique, l'extraction des protéines, leur purification et la protéolyse avec la trypsine des protéines en peptides (Hayoun et al., 2019). Les peptides obtenus sont injectés dans la chromatographie liquide (LC) puis dans le spectromètre de masse dans le cas d'une analyse de spectrométrie de masse LC-MS/MS. Chacune des étapes de protéomique classique peut être améliorée individuellement.

1.5.2.2 La spécificité de la préparation d'échantillons de sols

Le sol est une matrice complexe et hétérogène constitué de composés organiques tels que les glucides complexes, les lipides, les composés phénoliques (par exemple la lignine) et les substances humiques ainsi que les composés inorganiques tels que les limons et les minéraux argileux. La coextraction des substances humiques ou des minéraux argileux ne complique pas seulement l'extraction protéique mais interfère aussi avec la séparation des peptides (Bastida et al., 2009), l'identification des protéines (Arenalla et al., 2014) et la quantification des protéines en raison de modifications physico-chimiques des protéines (Keiblinger et al., 2016). Ces limitations peuvent être liées à l'adsorption, la liaison, l'ancrage ou l'incorporation des protéines sur ou dans des particules solides telles que l'argile, les minéraux argileux et les complexes organo-minéraux de la matière organique (Keiblinger et al., 2016) ce qui réduit l'efficacité d'extraction.

De nombreuses méthodes d'extraction de protéines ont été développées (Chourey et al., 2010 ; Keiblinger et al., 2012) soit par extraction indirecte où les micro-organismes sont enrichis avant extraction, soit par séparation par centrifugation par gradient de densité, soit par l'extraction directe avec une lyse dans la matrice du sol (Keiblinger et al., 2016) (*Figure 23*). La dernière méthode permet de ne pas se focaliser uniquement sur les organismes cultivables et de ne pas biaiser l'extraction (Bastida et al., 2012). L'extraction directe se décompose en plusieurs sous-étapes. La lyse cellulaire peut s'effectuer avec une méthode physique telle que le broyage par billes (en anglais « bead-beating »), une méthode chimique telle que des tampons de lyse contenant des détergents ioniques, comme le dodécylsulfate de sodium (SDS), ou des détergents non-ioniques ou des méthodes enzymatiques qui peuvent être utilisées seules ou en complément des méthodes physiques ou chimiques (Keiblinger et al., 2016) (*Figure 23*). La purification des protéines peut également être effectuée par filtration sur gel, ultrafiltration ou précipitation. Des kits commerciaux d'extraction de protéines dans les échantillons de sols sont une alternative comme le kit NoviPure Soil Protein Extraction des laboratoires Mo-Bio. Les protéines sont ensuite protéolysées en peptides.

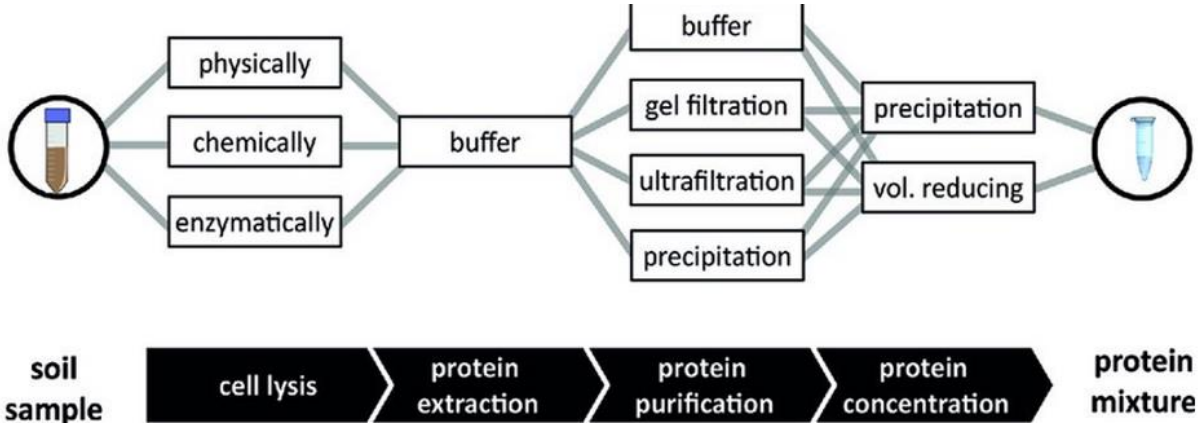


Figure 23: Extraction directe des protéines dans un échantillon de sol (Keiblinger et al., 2016).

1.5.3 Les approches de spectrométrie de masse

L'utilisation de la spectrométrie de masse pour l'analyse d'un échantillon est une technique instrumentale basée sur la séparation, l'identification et la quantification des peptides constituants l'échantillon en fonction de leurs charges et de leurs masses. Plusieurs méthodes d'analyses peuvent être utilisées en protéomique. La protéomique non ciblée permet d'analyser de manière globale l'ensemble des protéines d'un échantillon. La protéomique ciblée permet de quantifier et suivre quelques protéines d'intérêts. Dans les approches non ciblées, deux méthodes d'analyses complémentaires peuvent être utilisées : la méthode top-down par laquelle on analyse les protéines entières et la méthode bottom-up qui consiste à identifier les protéines par l'analyse des peptides suite à une digestion enzymatique. Cette dernière méthode peut utiliser indépendamment deux protocoles d'acquisition, l'acquisition dépendante des données (data-dependent acquisition ou DDA) et l'acquisition indépendante des données (data-independent acquisition ou DIA).

1.5.3.1 Quantification ciblée

Les méthodes de quantification ciblée qualifiées aussi de protéomique ciblée permettent de quantifier des ensembles prédéterminés de protéines dans de multiples échantillons de manière cohérente, reproductible et précise (Picotti et Aebersold, 2012). La protéomique ciblée est basée sur l'analyse des composés de mêmes masses/charges que les peptides protéotypiques dérivés des protéines cibles qui seront fragmentées et analysées pour être quantifiées (Figure 24). Contrairement aux méthodes non ciblées, ces méthodes sont plus spécifiques et sensibles car elles utilisent un filtrage de masse en deux étapes des ions précurseurs et des ions fragments avec une haute résolution (Nakayasu et al., 2021). C'est une technique permettant par exemple de vérifier et valider des biomarqueurs candidats préalablement sélectionnés par protéomique shotgun. Les peptides biomarqueurs candidats sont mesurés en même temps que leurs homologues synthétiques marqués avec des isotopes. Les différentes techniques de quantifications ciblées sont le suivi de réactions multiples (selected reaction monitoring ou SRM), aussi connu sous le nom de MRM (multiple reaction monitoring), et le suivi de réactions parallèles (parallel reaction monitoring ou PRM).

1.5.3.2 La méthode top-down

L'approche protéomique descendante ou top-down est l'analyse de la masse des protéines entières par un spectromètre de masse. Un gène peut être à l'origine de différentes protéines suite à différentes formes d'épissages ou de modifications post-traductionnelles. Le terme « protéoforme » est utilisé pour décrire les différentes formes moléculaires selon lesquelles les produits protéiques d'un seul gène peuvent être trouvés (Smith et al., 2013). La méthode top-down donne une vue d'ensemble du protéome et permet d'identifier de nouvelles protéoformes, de caractériser en profondeur les séquences et de quantifier les modifications post-traductionnelles (Chen et al., 2018). Cependant cette méthode est extrêmement complexe sur des échantillons simples et est inenvisageable sur des échantillons complexes tel que les sols. Les massifs isotopiques observés dans les spectres MS/MS d'une seule protéine sont complexes à analyser.

1.5.3.3 La méthode bottom-up

L'approche protéomique ascendante ou bottom-up est une approche shotgun. Elle permet l'analyse de l'intégralité d'un protéome par le biais de peptides générés par protéolyse des polypeptides, des fragments de protéines. La caractérisation par spectrométrie de masse des peptides, molécules plus petites, est facilitée et nécessite peu de matériel car la méthodologie est sensible. Le principe repose sur la mesure par un spectromètre de masse des masses/charges de peptides issus de la protéolyse des protéines, puis du séquençage de ces peptides après fragmentation. La protéase la plus utilisée est la trypsine qui clive les protéines en peptides après les acides aminés basiques, arginine (R) et lysine

(K), chargés positivement. Les peptides sont séparés selon leur hydrophobicité par chromatographie liquide, ionisés par la source ionisante et analysés par le spectromètre de masse. Après la mesure du ratio masse/charge des peptides ionisés (MS1), les ions les plus abondants dans le cas de l'acquisition dépendante des données et l'ensemble des ions dans le cas de l'acquisition indépendante des données sont sélectionnés et fragmentés dans la chambre de collision et le ratio masse/charge des fragments est mesuré (MS2). Les données masse/charge des ions dits "parents", MS1 et des ions dits "fils", MS2 sont ainsi obtenus sous la forme de spectres MS et de spectres MS/MS. Ces données associées au temps de rétention sont interprétées par une suite de logiciels spécifiques afin d'assigner une séquence peptidique à un spectre (PSM pour Peptide-Spectrum Match).

1.5.3.4 La différence entre DDA et DIA de la méthode bottom-up

Dans le cadre d'une analyse bottom-up, les méthodologies d'acquisition principales utilisées aujourd'hui sont l'acquisition dépendante des données (data-dependent acquisition ou DDA) et l'acquisition indépendante des données (data-independent acquisition ou DIA). La principale différence est la sélection d'un nombre d'ions dits « parents » les plus abondants et leurs analyses successives dans le cas de la DDA, la fragmentation de l'ensemble des ions et leurs analyses ininterrompues dans le cas de la DIA (Figure 24). Cette différence entraîne l'utilisation de pipelines d'interprétation des spectres différents (Zhang et al., 2020). Dans le cadre de la DIA, des outils tels que Spectronaut (Bruderer et al., 2015), Skyline (MacLean et al., 2010), DIA-NN (Demichev et al., 2020), Diatools (Aakko et al., 2020) peuvent être utilisés. Dans le cas de la DDA, des moteurs de recherche tels que Mascot (Perkins et al., 1999), Andromeda (Cox et al., 2011), X!Tandem (Fenyö et Beavis, 2003), Sequest (Eng et al., 1994), OMSSA (open mass spectrometry search algorithm) (Geer et al., 2004) et MS-GF+ (Kim et Pevzner, 2014) peuvent être utilisés.

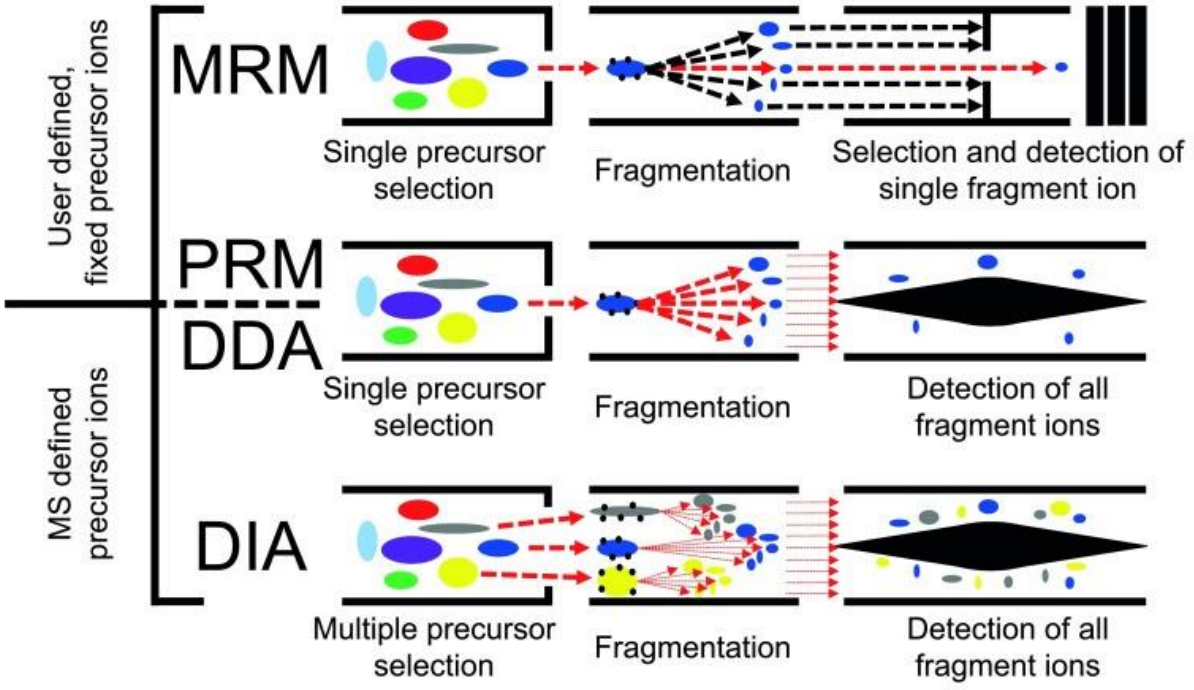


Figure 24: Représentation schématique des différentes méthodes d'acquisition utilisées en protéomique : la MRM, PRM, DDA et DIA. Les peptides sont isolés, fragmentés et analysés par le spectromètre de masse en spectre MS/MS en MRM, PRM et DDA et non isolés en DIA (Hu et al., 2016).

1.5.3.5 La spectrométrie de masse en métaprotéomique appliquée aux échantillons de sols

Pour l'analyse d'échantillons de sols, l'approche bottom-up est la seule utilisée à l'heure actuelle avec une acquisition DDA. Cette méthode shotgun permet d'analyser les métaprotéomes sans a priori, basée sur l'analyse de groupes fonctionnels détectés dans l'échantillon (Van Den Bossche et al., 2021) et notamment des fonctions de chaque organisme ainsi que leurs interactions dans l'écosystème. Non dépendante des bases de données de spectres MS/MS, cette méthode permet d'identifier les protéines les plus abondantes dans les échantillons et permet d'analyser taxonomiquement et fonctionnellement l'échantillon. Van der Bossche, à travers le projet CAMPI, a analysé l'effet des étapes de préparation d'échantillons et de l'analyses bioinformatiques de différents laboratoires partenaires du projet (Van Den Bossche et al., 2021). Les principales différences ont été mises en évidence au niveau peptides et sont liées principalement à la méthode expérimentale. Ces différences disparaissent au niveau des groupes de protéines et des profils fonctionnels similaires sont obtenus pour tous les types d'analyses, permettant ainsi de montrer la robustesse de la recherche actuelle et servant de support pour les futurs développements (Van Den Bossche et al., 2021).

1.5.3.5.1 Co-élution et co-fragmentation

La métaprotéomique accentue des problèmes qui se posent en protéomique, la co-élution et la co-fragmentation de peptides. La séparation de dizaines de milliers de peptides selon leur hydrophobicité dans la colonne de la chromatographie liquide conduit à l'analyse simultanée de peptides ayant des propriétés physico-chimiques très proches. La co-élution de peptides complexifie l'analyse en spectrométrie de masse où à un instant *t*, le nombre d'espèces présentes dans le spectromètre de masse est élevé (Lahrichi et al., 2013). De plus, la sélection des espèces les plus abondantes se fait sur une fenêtre de masse/charge réduite. De ce fait, plusieurs espèces peuvent être fragmentées en même temps et donner un spectre MS/MS chimérique. Ce phénomène de co-fragmentation réduit la qualité du spectre pour la suite de l'analyse bioinformatique puisque les fragments analysés proviennent de différents peptides ayant des séquences proches.

1.5.3.5.2 Gamme dynamique

La fréquence d'acquisition des spectromètres de masse en mode DDA impose la sélection des peptides les plus abondants et induit une sélection aléatoire des ions précurseurs de faible abondance (Van Den Bossche et al., 2021). De ce fait, la fiabilité et la sensibilité de détection en spectrométrie de masse ne permet pas de couvrir la gamme dynamique des concentrations physiologiques des protéines d'un seul organisme En effet, celle-ci est de l'ordre de 10^4-10^6 pour des isolats microbiens (Hettich et al., 2012) alors que la gamme dynamique de détection par spectrométrie de masse est de l'ordre de 10^3 (Documentation Thermo Fisher, https://assets.thermofisher.com/TFS-Assets/CMD/brochures/BR-64052-LC-MS-Q-Exactive-HF-Orbitrap-BR64052-EN.pdf) pour le Q-Exactive HF jusqu'à 10^6 pour le plus récent spectromètre de masse Q Exactive Focus MS de chez Thermo Fisher (Documentation Thermo Fisher, https://assets.thermofisher.com/TFS-Assets/CMD/brochures/BR-64278-LC-MS-Q-Exactive-Focus-Orbitrap-BR64278-EN.pdf). Un manque de reproductibilité dans la mesure des ions de basse concentration est constaté, empêchant l'identification fiable des protéines qui sont à l'origine du peptide. Pour élargir l'identification des protéines à des niveaux de concentration bas, la solution est de décomposer l'échantillon en plusieurs fractions. Cette méthode permet d'accéder à des protéines de niveaux de concentration plus faibles et importantes pour différencier l'activation des mécanismes moléculaires finement régulés par ces dernières ou pour discriminer des espèces dont la spécificité est portée par ces protéines mais nécessite des quantités d'échantillons plus importantes.

1.5.3.5.3 La DIA n'est-elle pas envisageable ?

La DIA est une solution alternative à la simplification des matrices complexes par le fractionnement préalable à l'analyse des échantillons. En effet, elle permet une meilleure couverture des protéines de

faible abondance (Searle et al., 2018) puisqu'il n'y a pas de présélection des ions précurseurs qui simplifient la lecture des ions fragmentés. En revanche, les méthodes d'interprétation des spectres doivent être améliorées pour distinguer chaque spectre spécifique dans le mélange complexe d'ions fragmentés. Une fois les outils d'interprétation optimisés, la DIA pourrait être une technologie d'avenir dans le cadre d'analyses d'échantillons de métaprotéomique de sol. Elle assurera l'identification des microorganismes en faible quantité permettant ainsi de distinguer les communautés de microorganismes à partir de prélèvements réalisés à partir d'un sol de même nature, mais à différents points de localisation rapprochés. Cependant, l'identification des microorganismes de sols dépend aussi du niveau de connaissances acquises dans le domaine de la microbiologie environnementale. Plusieurs méthodologies d'exploration des données sont développées telles que Diatools (Aakko et al., 2020) et DDIA (Data Dependent-Independent Acquisition) (Guan et al. 2020).

1.5.3.6 L'analyse d'un échantillon de sol en spectrométrie de masse Q-Exactive HF en mode DDA

Il existe différentes technologies de spectromètres de masse et de nombreux appareils. Prenons en exemple le spectromètre de masse Q-Exactive HF (Scheltema et al., 2014) couplé à un système de chromatographie liquide nano Ultimate 3000 (nano LC) en utilisant le mode DDA, configuration à la base de toutes les analyses réalisées dans le cadre de la thèse.

L'analyse d'un échantillon peptidique par un spectromètre de masse couplée à une chromatographie en phase liquide (HPLC) se déroule en plusieurs étapes. La première étape est la séparation des peptides de l'échantillon selon leurs propriétés physico-chimiques le long d'un gradient d'eau/acétonitrile/TFA (acide trifluoroacétique) dans la pré-colonne et la colonne de la chromatographie en phase liquide (LC). Ce gradient sépare les peptides selon leur hydrophobicité. Le temps que le peptide mette à traverser la colonne est le temps de rétention. Les peptides les plus hydrophiles se présentent les premiers à la source d'ionisation en un flux continu jusqu'à la fin de l'analyse de l'échantillon.

La source d'ionisation (ESI) est un électronébuliseur (en anglais, electrospray) qui ionise les molécules ensuite guidés par des capteurs électromagnétiques. La source assure le transfert des peptides de la phase aqueuse à la phase gazeuse. Selon le principe, un champ haute tension est appliqué à une gouttelette issue de l'échantillon où les ions, en suspension dans le liquide, vont migrer à la surface de la gouttelette et s'accumuler. A cet endroit, le champ électrique est plus élevé et la force électrostatique tire la surface de la gouttelette où les ions (=les peptides ionisés) sont récupérés.

Les ions récupérés en continu à la source d'ionisation sont ensuite pré-filtrés par une succession de guide d'ions, la source S-lens constituée d'une variété de lentilles électrostatiques, puis d'une série d'optiques d'ions (le Flatapole) avant d'arriver dans le quadrupôle où les ions sont sélectionnés dans la gamme de balayage. Les ions neutres non chargés sont filtrés et uniquement les ions d'intérêts sont conservés pour la suite de l'analyse selon leur intensité et leur état de charge. Dans le quadrupôle, les ions sont sélectionnés sur la base de leur ratio masse sur charge (m/z) en fonction de la stabilité de leurs trajectoires dans les champs électriques oscillants appliqués aux quatre cylindres constituant le quadrupôle. La gamme de masse des ions analysés est définie par les caractéristiques de l'appareil, configuré dans notre cas entre 350 et 1800 m/z à une résolution de 60000.

Les ions traversent le quadrupôle en un flux continu, chaque seconde (un cycle), les ions sont collectés en paquets dans la C-Trap, une trappe ionique. Les ions sont envoyés dans l'analyseur de masse Orbitrap pour détection. Dans celui-ci, le mouvement orbital des ions induit une fréquence de rotation reliée au rapport masse sur charge suivant la transformation de Fourier. Les ions ayant des ratios m/z différents ont des fréquences d'oscillations différentes. Les ions analysés fournissent un spectres MS, ou MS1, où les ions des espèces les plus abondantes sont identifiés. Dans le flux continu d'ions, les ions

de chaque espèce sont successivement isolés sur une fenêtre de masse/charge jusqu'à sélectionner un maximum de vingt ions les plus abondants (méthode TOP 20) en un cycle d'une seconde. Les ions de chaque espèce sont ainsi stabilisés dans la C-Trap et mis en présence d'un gaz de collision, l'azote (N_2). Ils sont ensuite fragmentés dans la cellule de collision pour former les ions dit "fils". Les fragments chargés sont analysés par l'Orbitrap à une résolution de 15000 fournissant les spectres MS/MS, ou MS2. Chaque seconde, jusqu'à vingt ions sont analysés. Lorsqu'un ion est sélectionné, celui-ci ne peut être de nouveau sélectionné pendant dix secondes.

Le Q-Exactive HF est capable d'analyser en parallèle les ions précurseurs (= sans fragmentation) et les ions fragments, c'est-à-dire les événements MS (MS1), les événements MS/MS (MS2). Lors de l'analyse des ions dans l'Orbitrap, la génération suivante, les ions N+1 sont accumulés dans la C-Trap avant d'être détectés et analysés.

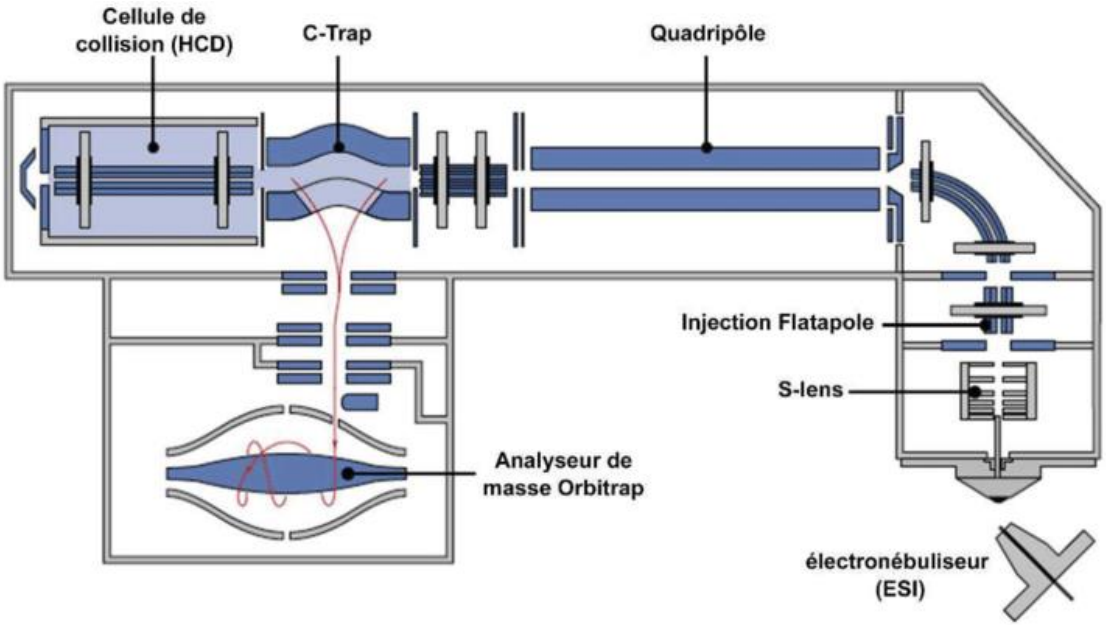


Figure 25: Représentation schématique du spectromètre de masse ESI Q-Exactive HF (Thermo Fisher Scientific).

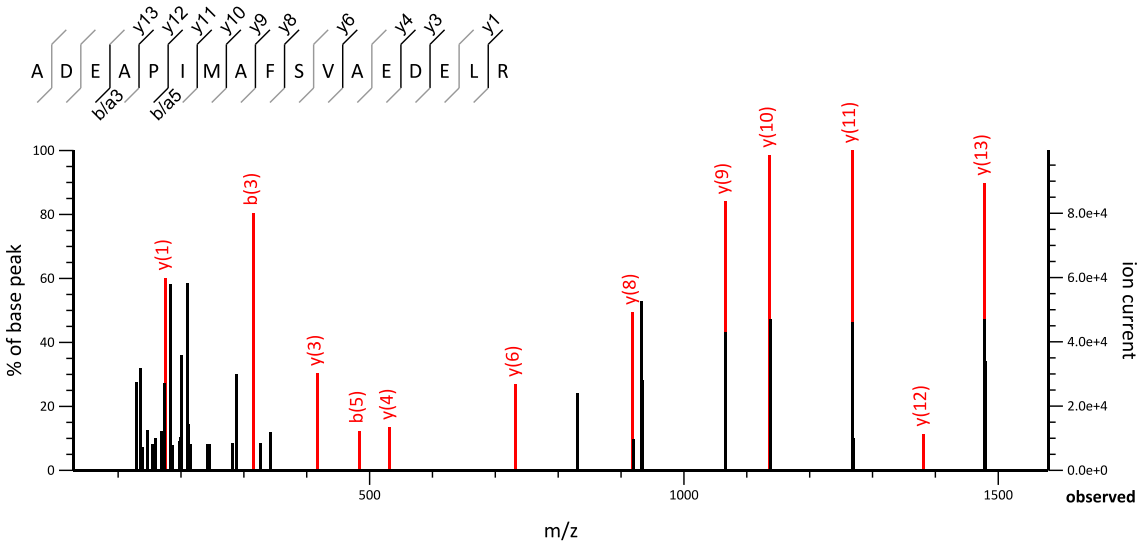


Figure 26: Exemple d'un spectre MS/MS : en abscisse, la masse/charge et en ordonnée à droite, l'intensité du courant ionique mesuré des fragments et en ordonnée à gauche la valeur en pourcentage. En noir, ce sont les

valeurs mesurées expérimentalement et en rouge, les valeurs obtenus in silico lors de l'interprétation des spectres.

1.5.4 Les méthodes d'interprétation des données issues du spectromètre de masse

1.5.4.1 L'interprétation des spectres MS/MS en protéomique classique

1.5.4.1.1 Les données brutes : les spectres MS/MS

Un spectre MS/MS correspond à la mesure d'intensité et à la masse/charge (m/z) des ions fils issus de la fragmentation d'un peptide (Figure 26). Les types d'ions fragmentés observés dans un spectre MS/MS dépendent de nombreux facteurs, notamment la séquence primaire, la quantité d'énergie interne, l'état de charge (+1, +2, +3), etc. Les fragments ne seront détectés que s'ils portent au moins une charge. Si cette charge est retenue sur le fragment N-terminal, l'ion est classé comme a, b ou c. Si la charge est retenue sur le C-terminal, le type d'ion est soit x, y ou z. Le spectre MS/MS est ensuite interprété par les outils adéquats qui analysent l'adéquation entre la fragmentation *in silico* du peptide et la fragmentation expérimentale (Figure 26).

1.5.4.1.2 L'attribution d'une séquence peptidique à un spectre MS/MS

L'interprétation des spectres pour l'identification peptidique en protéomique shotgun se fait principalement via l'utilisation d'un moteur de recherche utilisant une base de données de séquences de protéines connues représentative de l'échantillon analysé ou non.

Attribution *de novo* des peptides :

Des méthodes de séquençage *de novo* des peptides peuvent être appliquées aux données MS/MS. Elles n'utilisent pas de bases de données de référence. Elles utilisent les ratios masses/charges de l'ion parent et des ions fils pour reconstruire la séquence du peptide. Des outils tels que Novor (Ma, 2015), PEAKS (Tran et al., 2019), DirecTag (Tabb et al., 2008), PepNovo (Frank et Pevzner, 2005) et Antilope (Andreotti et al., 2011) peuvent être utilisés. Mais la précision de prédiction des séquences peptidiques exacte moyenne est d'environ 35 % (Muth et Renard, 2018).

Attribution des peptides avec un moteur de recherche et une base de données :

Le moteur de recherche génère *in silico*, à partir de la base de données de protéines, les peptides potentiellement présents dans l'échantillon en clivant les protéines de la base de données avec l'enzyme utilisée *in vivo*. Chaque peptide est ensuite fragmenté *in silico* afin d'obtenir les ratios masses/charges des fragments. Ainsi, pour l'interprétation des résultats, après sélection des possibilités issues de la mesure de la masse du peptide parent, les spectres MS/MS obtenus en spectrométrie de masse sont comparés aux données de digestion et fragmentation *in silico* en tenant compte des paramètres d'erreurs de mesure sélectionnés. Cette méthode permet d'intégrer aussi les modifications post-traductionnelles. La quantification de ces peptides peut se faire sur la base du nombre de spectres (spectral count) ou évaluée en termes d'intensité (XIC). L'identification et la quantification des peptides permettent ainsi de remonter aux abondances relatives de chacune des protéines pour être extrapolées ensuite aux organismes les contenant. La méthode de quantification XIC est dite plus précise que la méthode « spectral count » car elle utilise l'aire sous la courbe d'intensité des pics des ions analysés à un instant *t* pour quantifier les peptides. Avec le XIC, on se rapproche d'une méthode de quantification alors que le spectral count est une quantification relative. A contrario, en « spectral count », chaque spectre a une valeur de « un » quelle que soit l'intensité du peptide (ion parent) et les spectres associés à un même peptide sont comptés. La méthode XIC est cependant fortement dépendante de la qualité de la base de données, celle-ci peut être que partielle ou contenir des séquences erronées, et de la densité des signaux des peptides élués. Cette méthode est notamment implémentée dans le logiciel Maxquant (Tyanova et al., 2016).

Parmi la liste des logiciels qui peuvent être utilisé pour interpréter les données de spectrométrie de masse, prenons en exemple le logiciel Mascot (Perkins et al., 1999) afin d'expliciter le fonctionnement et les paramètres pour l'identification et la validation des peptides puis des protéines.

1.5.4.1.3 Le principe de fonctionnement du moteur de recherche Mascot

Le logiciel Mascot est un logiciel sous licence commerciale payante. Il offre la possibilité de paramétrer la requête sur le fichier MGF en fonction des caractéristiques de l'appareil telles que la précision de masse mais aussi l'enzyme utilisée, le seuil d'identité ou le seuil d'homologie, la p-value ou le taux de faux positifs (False Discovery Rate, ou FDR). Cependant, les calculs des scores et des seuils de validation des spectres ne sont pas accessibles aux utilisateurs. Seule une librairie disponible dans plusieurs langages informatiques, Mascot Parser, peut être utilisée pour visualiser les fichiers de résultats de Mascot ainsi que l'interface de Mascot. Le principal avantage lié à l'utilisation de la librairie est que l'on peut automatiser le traitement des résultats Mascot ainsi que leurs visualisations. Le moteur de recherche fonctionne en quatre étapes.

Première étape : la base de données et sa digestion *in silico*

La première étape est la configuration de la base de données à utiliser par le moteur de recherche qui va ensuite la digérer *in silico*. Selon la capacité du serveur hébergeant Mascot, de nombreuses bases de données peuvent être paramétrées simultanément et être disponibles pour les différentes requêtes d'un laboratoire. Les plus courantes sont les bases de données :

- ☑ NCBInr: une base de données de protéines non redondantes issue de NCBI, contenant des millions de protéines d'organismes séquencés issus des trois règnes du vivant (195 731 717 protéines au 3 janvier 2020),
- ☑ UniprotKB-TrEMBL: une base de données automatiquement annotée mais non vérifiée (219 174 961 protéines au 17 août 2021),
- ☑ UniprotKB-SwissProt: une base de données annotée et vérifiée manuellement. Les informations sont extraites de la littérature (565 254 protéines au 17 août 2021),
- ☑ Les bases de données mono-organismes d'organismes séquencés contenant l'ensemble des protéines extraites à partir des bases de données NCBI ou Uniprot,
- ☑ Les bases de données de séquences issues du séquençage de l'échantillon d'intérêt (génomique, métagénomique ou transcriptomique) de l'organisme cible ou de l'écosystème d'intérêt.

Chaque base de données configurée est ensuite digérée *in silico* par l'enzyme utilisée pour cliver les protéines lors de la préparation de l'échantillon, généralement, la trypsine. Les masses des peptides générées *in silico* sont ensuite comparées aux ions parents des spectres MS obtenus par le spectromètre de masse donnant un score nommé Qmatch. Il correspond au nombre de peptides candidats pour une masse/charge donnée. Par exemple, pour 20 candidats possibles, le Qmatch est de 20.

Deuxième étape : l'attribution d'une séquence peptidique à un spectre MS/MS

Dans un deuxième temps, la correspondance du spectre et de chaque peptide candidat est évalué *in silico* par un score nommé ion score. Il est basé sur la correspondance entre les ions fragments MS/MS pré-calculé *in silico* pour une séquence peptidique candidate et les fragments MS/MS mesurés expérimentalement (Figure 26). Ce score est indépendant du paramètre de tolérance d'erreur sur la masse/charge théorique et expérimentale. Pour un spectre donné, les différents peptides candidats

sont à ion score non dépendant des paramètres tels que la taille de la base de données, le nombre de candidats ou les paramètres d'analyse Mascot.

A partir de l'ion score, il est possible de calculer une e-value (Expectation Value) qui indique le niveau de probabilité que l'attribution peptidique soit liée au hasard. Ce score est équivalent au score de e-value d'une recherche Blast. Le lien entre l'ion score et la e-value est la suivante :

$$S = -10Log10(P)$$

Où P est la probabilité que la correspondance observée entre les données expérimentales et la séquence de la base de données soit un évènement aléatoire. Dans le cas d'un nombre de peptides candidats constants, la e-value des séquences peptidiques des spectres MS/MS sera inversement proportionnel à la tolérance de masse. Au cours d'une recherche, si 1500 peptides sont dans la fenêtre de tolérance de masse autour de la masse du peptide précurseur (= parent) et que le seuil de significativité est choisi à p <= 0.05, c'est-à-dire une chance sur 20 d'être un faux positif, le seuil de score correspondant devrait être -10*Log*(1/(20 x 1500)) = 45. Le ou les peptides candidats ayant les scores les plus élevés et supérieurs à l'un des seuils de validation (le seuil d'identité et le seuil d'homologie) dépendant du spectre MS/MS en cours sont considérés validés.

Troisième étape : les seuils de validation des peptides

Premièrement, le seuil d'identité Mascot (MIT, Mascot Identity Threshold) est dépendant du nombre de peptides candidats n et de la e-value précédemment calculée. Il s'énonce comme suit :

$$MIT = -10Log10(\frac{20S}{n})$$

Le MIT est représentatif du niveau de significativité choisi et de la taille de l'espace de recherche impacté directement par les paramètres de la requête choisis. Dans le cas d'un spectre MS/MS d'une qualité médiocre, le rapport signal/bruit est faible et aucun peptide peut ne pas dépasser le seuil MIT qui est un score dit absolu malgré un score de correspondance peptide candidat et spectre élevé. Ce qui a conduit Mascot à mettre en place un second seuil de validation basé sur l'homologie.

Deuxièmement, le seuil d'homologie Mascot (MHT, Mascot Homologie Threshold) est basé sur la caractérisation de la distribution des scores des différents peptides candidats lorsque leur nombre est élevé. Il est plus faible que le MIT. La formule de calcul n'est pas communiquée par Mascot.

Quatrième étape : l'attribution des peptides identifiés aux protéines

A partir de l'ensemble des peptides candidats attribués aux spectres dont les scores sont supérieurs aux seuils de validation, l'intérêt de la protéomique est de connaître les protéines présentes dans un échantillon. Cependant, une séquence peptidique est contenue généralement dans plusieurs protéines telles que les protéines isoformes ou partageant des fonctions communes. Le défi de l'inférence protéique consiste à sélectionner l'ensemble des protéines expliquant le jeu de peptides identifiés. Il peut être résolu par le principe de parcimonie qui sélectionne le sous-jeu minimal de protéines expliquant l'ensemble des spectres identifiés en utilisant un score au niveau protéine afin de les classer et les sélectionner. Des groupes de protéines sont ainsi déduits de la sélection de protéines représentantes de l'échantillon : les protéines « same-set » où un ensemble de protéines partage le même ensemble de peptide et les protéines « sub-set » qui partagent un sous-jeu de peptides. Pour chaque groupe, c'est la protéine ayant le score MudPit le plus élevé et la première dans l'ordre alphabétique dans le cas de scores identiques qui sera nommé comme représentante du groupe, celle-ci possédant le plus grand nombre de peptides discriminants. L'ensemble du groupe est ainsi compté comme une seule occurrence validée. La principale limitation de cette méthode est l'identification et la quantification des différentes isoformes dans de multiples échantillons.

Le score Mudpit de chaque protéine est calculé en additionnant les ions scores maxima de chaque peptide attribué appartenant à la protéine, plus précisément, c'est la quantité au-dessus du seuil qui est prise en compte, du seuil MHT s'il est disponible ou du MIT. La moyenne des seuils utilisés est ensuite ajoutée au score et une correction sur le score est appliquée afin de réduire la contribution des correspondances aléatoires de faibles scores.

1.5.4.1.4 La gestion des faux positifs

Basé sur cette méthodologie, Mascot permet l'identification et la quantification par spectral count des protéines identifiées dans une analyse à une p-value donnée. La p-value choisie par l'utilisateur représente le risque de se tromper en affirmant que le test est significatif. Cependant, lorsque de nombreuses comparaisons et de nombreux tests sont effectués à une certaine p-value, le risque d'obtenir de fausses attributions significatives augmentent. Par conséquent, il est nécessaire de prendre en compte le nombre de tests effectués et d'estimer le taux de fausse découverte dit FDR (False Discovery Rate) (Burger, 2017 ; Muth et al., 2015). La méthode la plus courante pour corriger le score calculé est la correction de Benjamini et Hochberg (Benjamini et Hochberg, 1994).

Dans le cadre de la protéomique, le contrôle de la FDR s'effectue au niveau spectre en utilisant le Qmatch dans le calcul du score, ce qui impacte les seuils de validation MIT et MHT. La gestion de la FDR peut être effectuée plus généralement au niveau spectre, peptide et protéine en utilisant une base de données leurre par lecture inverse de la base de données cible utilisée par la requête ou en utilisant des outils tels que DecoyPyrat (Wright et Choudhary, 2016). Le nombre d'attribution à une p-value donnée sur la base leurre permet une estimation du taux de faux positifs. Cette méthode est couramment utilisée en métaprotéomique mais implique des biais d'identification connus en fonction de la taille de l'espace de recherche (Muth et al., 2015). Le seuil de FDR appliqué pour des analyses en métaprotéomique est généralement de 1% (Pérez-Cobas et al., 2013), 5% (Ng et al., 2010) ou 10% (Sollanek et al., 2017).

Une autre méthode proposée par Mascot est l'utilisation de Percolator (Käll et al., 2007), un algorithme basé sur l'apprentissage automatique semi-supervisé pour améliorer la discrimination entre les identifications correctes et incorrectes du spectre. Les correspondances obtenues lors de la recherche dans une base de données de leurres fournissent les exemples négatifs pour le classificateur et les sous-ensembles des correspondances à score élevé de la base de données cible fournit les exemples positifs.

1.5.4.2 L'interprétation des spectres en métaprotéomique des sols et autres échantillons complexes

1.5.4.2.1 Le moteur de recherche

Lors de l'analyse de données de métaprotéomique des sols, le choix de la stratégie à employer n'est pas défini. Il n'existe pas à ce jour de comparaison fiable des stratégies d'analyses pour maximiser les résultats de spectrométrie de masse obtenus en métaprotéomique des sols. Dans le contexte général de la métaprotéomique, Seifert et Muth ont mis en évidence le manque d'outils et de workflows dédiés à l'identification, l'annotation et la quantification de protéines pour déterminer avec précision la fonction de la communauté microbienne et comparer de manière robuste des échantillons hétérogènes dans le temps et l'espace (Seifert & Muth, 2019). Les outils d'interprétation de spectres comme Mascot et X!Tandem n'ont pas été conçu pour faire face à la complexité des ensembles de données protéomiques (Saito et al., 2019). Dans certaines études, la combinaison des résultats obtenus avec plusieurs moteurs de recherches fournit des résultats qui semblent adaptés à l'analyse des systèmes complexes (Searle et al., 2015). En effet, l'utilisation de plusieurs moteurs de recherche dotés de leurs propres paramètres d'identification augmente le nombre de spectres interprétés et donc de protéines identifiées. Cependant, la plupart des études en métaprotéomique n'utilisent qu'un

seul moteur de recherche (Muth et al., 2015) ce qui permet de réduire le temps d'analyse et la complexité associé à la gestion des résultats issus de plusieurs moteurs de recherche.

Après avoir choisi un moteur de recherche, les paramètres de requête doivent être choisis rigoureusement afin d'augmenter le nombre de protéines identifiées et améliorer la couverture de leur séquence par le nombre de peptides identifiés. (Révész et al., 2021).

1.5.4.2.2 La base de données utilisée en métaprotéomique

Les spectres MS/MS acquis en métaprotéomique sont interprétés par comparaison avec une base de données répertoriant les séquences de toutes les protéines potentiellement présentes dans l'échantillon. Pour créer une telle base de données, la stratégie la plus appropriée est de réaliser un séquençage métagénomique ou métatranscriptomique sur le même échantillon. Ces bases de données peuvent ensuite être traduites pour en déduire les séquences protéiques théoriques. Il est également possible de compiler les séquences protéiques des organismes identifiés dans des échantillons similaires pour compléter les informations obtenues par métagénomique (Zampieri et al., 2016 ; Hultman et al., 2015). Une méthode alternative consiste à assembler une base de données à partir des organismes identifiés lors d'un séquençage d'ARNr 16S (Xiao et al., 2018) ou étant potentiellement présents dans l'habitat échantillonné (Tanca et al., 2013). Des bases de données généralistes telles que NCBInr ou UniProtKB/Swiss-Prot peuvent également être utilisées (Heyer et al., 2016).

1.5.4.2.3 La taille de l'espace de recherche

La biodiversité élevée d'un sol entraîne un espace de recherche dans les bases de données utilisées qui est naturellement beaucoup plus grand que celui requis pour la protéomique d'un seul organisme. La base de données utilisée est souvent de grande taille et la capacité à distinguer les correspondances correctes des correspondances incorrectes est fortement altérée (Rechenberger et al., 2019). En augmentant l'espace de recherche, le potentiel de faux positifs est fortement augmenté. Généralement, en métaprotéomique des sols, la base de données utilisée contient des millions de séquences telle que Uniref100 (Bastida et al., 2014), NCBI ou un métagénome.

Plusieurs stratégies ont été proposées pour contrer les effets négatifs d'une base de données trop volumineuse sur la sensibilité et la précision de l'appariement peptide-spectre. Notamment, une stratégie consiste en la réduction de la base de données à l'aide d'une recherche dans la base de données en deux étapes où les protéines identifiées à la première requête constituent la base de données de la seconde requête (Jagtap et al., 2013).

1.5.5 L'analyse de données protéomiques et métaprotéomiques : les outils bioinformatiques

L'analyse de données métaprotéomiques repose sur l'analyse des peptides identifiés et des protéines associées. Ensuite, l'ensemble des protéines et peptides identifiés constitue la base pour l'annotation taxonomique qui classe les protéines et/ou peptides en fonction de leur spécificité de séquence pour un organisme donné et pour l'annotation fonctionnelle des peptides et/ou protéines identifiés en attribuant un ensemble de fonctions et de processus biologiques au mélange. L'ensemble de ces données permet ensuite de réaliser l'analyse des structures des communautés microbiennes et de leurs activités fonctionnelles.

L'analyse de données métaprotéomiques nécessite des outils bioinformatiques performants permettant de gérer de grands volumes de données et une complexité plus élevée que les données de protéomique. De nombreux outils ne sont pas uniquement spécifiques à la métaprotéomique mais peuvent être utilisés plus généralement aux données de protéomique. En 2019, il existait déjà plus de

1000 logiciels dédiés à la protéomique dont plus de 750 sont référencés dans la base de données bio.tools (https://bio.tools/t?domain=proteomics) (Tsiamis et al., 2019).

1.5.5.1 L'annotation fonctionnelle et taxonomique

Les méthodes d'annotation peuvent être regroupées en deux catégories, les méthodes basées sur la spécificité des séquences tel que Unipept, et eggNOG ou celles sur leurs similarités de séquence en utilisant des méthodes d'alignements telles que MEGAN, Prophana ou GhostKOALA.

La première approche utilise la composition en acide aminés des peptides pour les attribuer à un rang taxonomique et une fonction donnée selon sa spécificité, grâce à un algorithme très performant d'indexation des séquences. Cette méthode est directement dépendante du contenu de la base de données initiale pour l'analyse de la proximité phylogénétique des organismes analysés et de ceux présents dans la base de données.

La deuxième approche basée sur la similarité de séquence est couramment utilisée pour l'annotation fonctionnelle des protéines. Elle est basée sur l'alignement des protéines identifiés sur une base d'annotation tel que les bases de données Uniprot ou Uniref. Les protéines similaires sont sélectionnées à partir de seuils basés sur la couverture et l'identité évaluées plus précisément par la e-value. On considère qu'au-delà de 30 % d'identité de séquences, deux protéines possèdent la même structure tridimensionnelle. Lorsque ce pourcentage est supérieur à 50%, il est admis que les protéines partagent la même fonction et sont donc homologues, mais ceci n'est pas forcément le cas, certaines mutations du site actif par exemple peuvent changer la fonction d'une protéine.

Sajulga a mis en évidence par une analyse BLAST contre la base de données non redondante de NCBI, que la sensibilité et la spécificité de l'annotation fonctionnelle variaient entre les outils d'alignement. Par exemple, eggNOG-mapper sélectionne un plus grand nombre de termes GO pour un ensemble de protéines donné, tandis qu'Unipept génère des termes GO plus précis (Sajulga et al., 2020).

Les méthodes basées sur la métaprotéomique pour évaluer la structure des communautés microbiennes en utilisant l'abondance des protéines comme mesure des contributions à la biomasse des populations individuelles sont moins sujettes à certains des biais trouvés dans les méthodes basées sur le séquençage (Kleiner et al., 2017)

1.5.5.2 Exemple d'outils bioinformatiques dédiés à la (méta)protéomique

1.5.5.2.1 SearchGUI et PeptideShaker

SearchGUI (Vaudel et al., 2011) est une interface open-source permettant de configurer et exécuter des requêtes sur différents moteurs de recherche en protéomique (X ! Tandem, MyriMatch, MS Amanda, MS-GF+, OMSSA, Comet, Tide, Andromeda, MetaMorpheus, Novor et DirecTag). Pour visualiser et analyser les résultats de la recherche obtenus par SearchGUI et d'autres outils, PeptideShaker (Vaudel et al., 2015) permet l'interprétation des résultats d'identification protéomique provenant de multiples moteurs de recherche et de novo. PeptideShaker regroupe les résultats en un seul ensemble d'identification, annote les spectres, calcule un score de consensus, cartographie les séquences et effectue une inférence protéique, évalue la localisation des modifications post-traductionnelles, effectue une validation statistique, un contrôle de qualité et annote les résultats en utilisant de multiples sources d'information comme l'annotation de Gene Ontology, UniProt et Ensembl, et les structures protéiques. Ces outils sont mis à disposition par le groupe CompOmics.

De mon point de vue, ces outils sont très prometteurs pour l'épanouissement d'un domaine tel que la métaprotéomique puisque ces outils sont libres de droits, fonctionnent sur Windows, Mac et Linux et ne nécessite pas d'un serveur pour fonctionner. Cependant, l'utilisation de grandes bases de données et d'échantillons complexes rendent difficiles son utilisation au quotidien.

1.5.5.2.2 OmicsPlayground

OmicsPlayground (Akhmedov et al., 2020) est une plateforme pour l'analyse, l'exploration et la visualisation des données omiques dédiée au prétraitement des données, aux tests statistiques et l'interprétation des données. Développé avec Rshiny, l'outil inclut des méthodes telles que le clustering, les analyses d'enrichissement, d'expression différentielles et le profilage cellulaire. L'un des inconvénients de cet outil est la présence d'une interface pour générer l'analyse des données qui impose une certaine plasticité et un modèle de données d'entrée et de sorties spécifiques.

1.5.5.2.3 mixOmics

mixOmics (Rohart et al., 2017) est une librairie R avec un large éventail de méthodes multivariées pour l'exploration et l'intégration des données biologiques, avec un accent particulier sur la sélection des variables. Elle inclut des méthodes dédiées à l'analyse d'un type de données omiques ou à l'intégration de données multi-omiques. Axée sur les méthodes multivariées, il propose des fonctions permettant d'avoir une vision globale des résultats ainsi que des tutoriels approfondis afin de s'approprier l'outil et obtenir le type de résultats souhaités et adaptés à son analyse.

1.5.5.2.4 RforProteomics

RforProteomics (Gatto et Christoforou, 2014) est une libraire R permettant d'analyser les données de protéomiques et de visualiser les données de spectrométrie de masse. Cet outil permet de charger les données brutes de spectrométrie de masse et d'analyser les spectres MS1 et MS2. Il permet également de charger des données issues de répertoire public tel que ProteomeXchange.

1.5.5.3 Exemples d'outils bioinformatiques développés pour la métaprotéomique

1.5.5.3.1 MetaproteomeAnalyzer (MPA)

Le logiciel MetaProteomeAnalyzer (Muth et al., 2015) est un outil open-source pour l'analyse et l'interprétation des données métaprotéomiques qui comprend plusieurs moteurs de recherche. Développé spécifiquement pour la métaprotéomique, MPA présente l'avantage d'inclure de nombreuses fonctions axées sur l'analyse taxonomique et fonctionnelle des protéines découvertes.

1.5.5.3.2 Unipept

L'application web Unipept (https://unipept.ugent.be) (Gurdeep Singh et al., 2019) ainsi que la version récente pour ordinateur permet l'analyse fonctionnelle et taxonomique d'échantillons métaprotéomiques complexes (Mesuere et al., 2012, Gurdeep Singh et al., 2019). Cette application comprend quatre fonctionnalités distinctes : l'analyse des peptides tryptiques, l'analyse métaprotéomique, la recherche de peptides uniques et le regroupement de peptides. Elle est basée sur l'analyse d'une liste de peptides tryptiques fournie par l'utilisateur dans une base de données optimisée et indexée originaires des séquences d'UniProt.

1.5.6 Les applications de la métaprotéomique des sols

Dans les différentes études qui portent sur l'analyse des écosystèmes terrestres, la métaprotéomique est un outil primordial. Il permet l'exploration des organismes vivants des sols (Abiraami et al., 2019), des sédiments (Glass et al., 2014), des océans (Saito et al., 2019 ; Wang et al., 2014) mais aussi à l'issue du traitement des eaux usées par boues actives dans les stations d'épurations (Wilmes et al., 2008) ou la composition des microorganismes des eaux souterraines (Benndorf et al., 2007). La métaprotéomique peut être un outil d'avenir pour l'identification des protéines d'environnements extrêmes (Maseh et al., 2021) comme les milieux salins (Hanson et al., 2014), arides (Bastida et al., 2015), chauds (Hensley et al., 2014), ou froids (Bell et al., 2013).

Le sol est un système biologique complexe présentant une grande variance spatiale et temporelle (Crowther et al., 2019). Les microorganismes du sol sont essentiels aux cycles géochimiques et nutritifs

et notamment dans les processus de minéralisation et de décomposition, aux interactions avec la flore et la faune et dans l'élimination des polluants. L'utilisation de la métaprotéomique sur ce type d'échantillon permet d'analyser les microorganismes à l'échelle de la communauté et de comprendre le fonctionnement de l'écosystème du sol. Cependant, pour réaliser une analyse fiable, un nombre suffisant de réplicas techniques et biologiques sont nécessaires (Abiraami et al., 2019 ; Elias et al., 2005). Dans le cadre de l'analyse du sol où la truffe se développe autour de l'arbre hôte, quatre réplicas ont été utilisé pour élucider les voies métaboliques dans cette zone, appelée le brûlé, qui ont mis en évidence une forte activité métabolique notamment des processus liés à des réponses à de multiples types de stress malgré une diversité microbienne et végétale réduite (Zampieri et al., 2016). Plusieurs types d'échantillons de sols ont déjà été exploré révélant les premières connaissances moléculaires des communautés du sol et de leur interaction.

1.5.6.1 L'analyse métaprotéomique de profils de profondeur

L'analyse stratifiée d'un ensemble d'échantillon dans un même continuum tel que la profondeur, la proximité spatiale ou le temps constitue un profil. L'étude de ce profil peut être orientée sur une analyse physico-chimique mais également sur la taxonomie des organismes présents (le profil taxonomique) ou les groupes fonctionnels exprimés par les gènes des communautés microbiennes (le profil fonctionnel). L'analyse d'un profil peut ainsi revêtir différentes natures.

Le profilage des communautés microbiennes permet de révéler des différences métaboliques ou taxonomiques le long d'un gradient, mais l'essentiel de ce type d'études n'a pas été conduit sur des échantillons de sols. Par exemple, Colatriano a mis en évidence des différences de compositions des communautés bactériennes et leurs fonctions le long d'une colonne d'eau d'un estuaire (Colatriano et al., 2015). Plus saumâtre, les eaux de surface sont habitées par des populations hétérotrophes impliquées dans le traitement de la matière organique alors que dans les eaux profondes, ce sont des activités liées à la chimiosynthèse et à la nitrification (Colatriano et al., 2015). Cependant, dans ce contexte, seulement trois profondeurs ont été étudié ne permettant pas réellement d'établir un profil de profondeur. Des études plus complexes ont été menées telles que celle de Bergauer portant sur les communautés microbiennes actives dans les processus de production de la matière organique. Quatorze niveaux de profondeurs, de 100 à 4050 mètres, ont été explorés avec les approches de métaprotéomique (Bergauer et al., 2018). Les échantillons ont été regroupés selon des profondeurs. Des études de profils ont été mené à l'aide d'autres approches génétiques, telles que l'analyse de l'ARNr 16S. Par exemple, Rocalkvam a étudié l'influence des taux d'infiltration de méthane sur la stratification, l'abondance et la diversité des méthanotrophes anaérobies dans des carottes de sables de 3 mètres (Roaclkvam et al., 2012).

L'analyse de profils de communautés microbiennes nécessite d'utiliser des méthodes d'analyse à grande échelle telles que les méthodes d'analyses statistiques multi variées supervisées (par exemple, la PLS-DA) et non-supervisée (par exemple, l'ACP) permettant d'extraire les données pertinentes des analyses. Les méthodes de corrélation de profils tels que coseq (Rau et al., 2020) sont également préconisés pour mettre en évidence le lien entre les différents paramètres biologiques et physico-chimiques mesurés.

1.5.6.2 Les types d'applications

1.5.6.2.1 La dégradation de la biomasse

La décomposition de la matière organique est le processus central du cycle de carbone. Son analyse permet d'étudier la structure et la fonction des communautés microbiennes impliquées dans la décomposition de la litière et de l'activité de certains décomposeurs (Schneider et al., 2012). Schneider et ses collaborateurs montrent que la teneur en nutriments de la litière et la stœchiométrie du carbone, de l'azote et du phosphore affectent la structure et l'activité de la communauté des

décomposeurs (Schneider et al., 2012). Le type et l'origine des protéines de la matière organique dissoute ont été analysées dans les sols de forêts (Schulze et al., 2005). La métaprotéomique a mis en exergue l'activité des champignons dans la production d'enzymes hydrolytiques extracellulaires non détectée dans les communautés bactériennes (Schneider et al., 2012).

1.5.6.2.2 Les interactions sol-plante et l'étude de la rhizosphère

L'étude des interactions sol-plantes au niveau de la rhizosphère présente un réel intérêt économique dans le domaine de l'agriculture. Les analyses métaprotéomiques de la rhizosphère des repousses de la canne à sucre (Lin et al., 2013), de la rhizosphère de cultures de maïs (Renu et al., 2019) ou de la rhizoremédiation pour l'élimination des HAP (Kotoky et al., 2018) ont apporté des informations précieuses sur les interactions sol-plantes mises en place . Lin (Lin et al., 2013), cherche à élucider les mécanismes du déclin du rendement des canne à sucre. Il a mis en évidence les activités enzymatiques pour la transformation du carbone, de l'azote et du phosphore plus faibles, assurées par la peroxydase, l'uréase et la phosphomonoestérase. Ces trois enzymes mises en évidence illustrent l'efficacité de ce processus biologique et peuvent donc être considérées comme des protéines signatures pour évaluer les performances de dégradation spécifique de la matrice constituée par le type de sol

1.5.6.2.3 La bioremédiation et restauration des sols

La bioremédiation de sites pollués lié à l'intensification des pratiques agricoles, de l'industrialisation en grande partie liée à l'activité humaine a été étudié par métaprotéomique afin d'analyser la dégradation des différents types de contaminants (Daffonchio et al., 2013 ; Kohle et al., 2018). L'objectif est d'analyser les réponses à long terme des communautés microbiennes aux effets de l'amendement organique par exemple (Bastida et al., 2015).

1.6 Contexte et objectifs de thèse

Ce projet de thèse s'effectue en collaboration entre deux laboratoires, le Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D) et le Laboratoire des Sciences du Climat et de l'Environnement (LSCE), tous les deux affiliés à la direction de la Recherche Fondamentale du Commissariat de l'énergie atomique et des énergies alternatives (CEA/DRF). La thèse a été financée par un programme de thèse phare Amont-Aval soutenu par le Haut-Commissaire du CEA. A l'initiative de ce sujet, le projet Geomics, est un projet transdisciplinaire financé par la DRF dans le cadre de l'appel à projets DRF Impulsion. L'objectif du projet est de développer un outil innovant de diagnostic de la santé des sols s'appuyant sur l'étude des micro-organismes des sols en lien avec la pression anthropique.

1.6.1 La dynamique de la contamination des sédiments de la Seine

Le LSCE, et particulièrement l'équipe Géochimie des impacts (GEDI), a pour mission l'étude de l'influence des activités anthropiques sur les transferts de matière sur les surfaces continentales et dans l'océan en utilisant des outils de géochimie. L'équipe a notamment travaillé sur la dynamique de la contamination des sédiments de la Seine comme archives du passé. L'objectif est d'étudier comment les changements technologiques et réglementaires ont modifié le type de contaminants et le niveau de contamination de ce fleuve fortement anthropisé.

Les archives sédimentaires, si elles sont correctement datées, permettent de suivre les émissions de contaminants à long terme et celles dont nous disposons actuellement pour la Seine permettent de remonter jusqu'à 1910 . Différents types de contaminants ont ainsi pu être étudiés par l'équipe et les collaborateurs du programme PIREN-Seine (cf paragraphe 1.2.2). Des résultats importants ont été obtenus sur un large éventail de contaminants dont les éléments traces métalliques, les radionucléides, des produits pharmaceutiques et des polluants organiques persistants. Cela représente environ 200 contaminants particulaires qui, pour la plupart sont réglementés ou interdits. A travers le projet PIREN-Seine, de nombreuses études ont porté, par exemple, sur les bilans de flux de métaux, les tendances historiques, l'impact des rejets urbains par temps de pluie.

Parmi les projets en cours, il y a notamment les études réalisées sur le plomb. Pour reconstruire les trajectoires d'un élément tel que le plomb dans le Bassin de la Seine, la stratégie est de confronter des archives sédimentaires situées en aval d'une source de pollution ou représentant l'ensemble du bassin au bruit de fond géochimique local établi sur des échantillons de sédiments datés de 5000 ans. En effet, dans le bassin de la Seine, nul site ne peut être considéré comme non contaminé. Les carottes sont datées et les mesures des teneurs en métaux ou autres contaminants sont réalisées. Pour évaluer la contribution de l'essence plombée dans la contamination au plomb, la signature urbaine parisienne du plomb a pu être définie en se basant sur les signatures des incinérateurs de déchets de la ville de Paris ainsi que la signature de l'essence plombée utilisée en France. Les données isotopiques du plomb des carottes ont ainsi pu être analysées montrant que le maximum d'impact de l'essence plombée est atteint en 1986, alors que l'âge d'or de l'essence plombée sont les années 70. Des analyses plus poussées ont mis en exergue la forte capacité du plomb de l'essence à contaminer la rivière, supérieure au plomb urbain. En 2003, la teneur en plomb des sédiments de la Seine était à 70 % d'origine urbaine et seulement à 30 % d'origine naturelle (Ayrault et al., 2012, Ayrault et al., 2020).

Plus largement, les travaux sur les archives sédimentaires de la Seine ont démontré un niveau de contamination très élevé dans le cours inférieur de la Seine notamment une contamination par les métaux et des polluants organiques, résultant des émissions de la zone très urbanisée autour de Paris. Ces études ont révélé l'historique de la contamination à long terme qui n'avait jamais été évaluée avant 2006.

L'étude des contaminants ont mis en évidence des histoires et des trajectoires différentes selon le contaminant considéré, tel que :

- ➔ L'absence de contamination pour des éléments dont les concentrations sont du même ordre que les valeurs naturelles comme le thallium (Ayrault et al., 2010) ;
- ➔ Des contaminants anciens tel que l'argent (Ayrault et al., 2010) ;
- ➔ Des évolutions en accord avec l'utilisation de ces contaminants avec tout de même un écart tel que le plomb (Ayrault et al., 2012) dont la contamination décroît également (Le Gall et al., 2018) ;
- ➔ Des fluctuations liées à l'usage pour certains contaminants tel que les POP, PCB et HAP (Lorgeoux et al., 2016) et également des agents antimicrobiens (Tamtam et al., 2011) ;
- ➔ Des effets d'accumulation à long terme suite aux retombées atmosphériques tels que les HAP qui ne cessent de croître (Froger et al., 2019).

L'utilisation des archives sédimentaires est une méthode efficace permettant de suivre les contaminations présentes et passées à l'échelle d'un bassin. Les prélèvements de sédiments en rivière donnent une image ponctuelle dans le temps.

1.6.2 L'étude des microbiotes par l'équipe ProGénoMix - Li2D

Le Li2D a pour mission le développement de méthodologies et de technologies pour la détection d'agents pathogènes ou toxiques présents dans l'environnement. L'équipe du Dr Jean Armengaud est spécialisée en protéomique et notamment dans l'étude de microbiotes cliniques et environnementaux. La plateforme de spectrométrie de masse de cette équipe, nommée ProGénoMix, fait partie du réseau IBiSA et est dédiée aux analyses protéogénomiques d'organismes non modèles, la caractérisation de microbiotes ainsi que les analyses métaprotéomiques. L'équipe est également très active dans la communauté de métaprotéomique et est notamment un membre fondateur de la « Metaproteomics Initiative » (https://metaproteomics.org/) regroupant les acteurs clés en métaprotéomique de par le monde et qui a comme objectif d'améliorer l'expertise dans ce domaine et populariser son application.

L'équipe a au fil des années mis au point des méthodes d'analyses notamment celle portant sur l'identification des organismes dans un échantillon simple ou complexe, la phylopeptidomique, permettant in fine d'estimer leurs biomasses (Pible et al., 2020 ; Hayoun et al., 2020). La phylopeptidomique considère à la fois les peptides discriminants et partagés entre les organismes. L'approche propose un modèle mathématique qui produit une signature phylopeptidomique spécifique pour un organisme décrivant la proportion de peptides partagés avec les autres organismes et leur phylogénie respective. Sans *a priori*, cette méthode à l'avantage d'être rapide dans la caractérisation des données (méta)protéomiques.

En parallèle, les stratégies d'analyses de données métaprotéomiques de microbiotes d'organismes sentinelles de l'environnement tel que le gammare *Gammarus fossarum*, un amphipode d'eau douce utilisé comme bio-indicateur de son milieu, a mis en évidence l'importance de stratégie intégrative de données omiques (Gouveia et al., 2020). Notamment, une stratégie de requêtes sur bases de données en multi-étapes basée sur l'utilisation de la base de données généraliste NCBI non-redondante et du transcriptome de l'organisme hôte a été élaborée. Cette stratégie en multi-étapes a permis de caractériser la composition microbienne du microbiote en extrayant les informations de l'hôte et de l'alimentation.

La caractérisation taxonomique et fonctionnelle d'un écosystème issu d'un hôte (Gouveia et al., 2020), de patients (Hardouin et al., 2021) ou d'un environnement révèlent des enjeux bien différents à partir d'une même technique qu'est la métaprotéomique.

1.6.3 Le projet Geomics

Le projet Geomics réunit deux équipes qui n'étaient pas naturellement destinées à travailler ensemble, mettant à profit leurs expertises pour l'analyse métaprotéomique d'archives sédimentaires en alliant les résultats aux données géochimiques. Ce projet interdisciplinaire a comme objectif d'étudier les communautés microbiennes sous pressions anthropiques d'un point de vue temporel en reliant les changements de pression anthropique mesurés sur un site dont l'historique est bien connu grâce à des archives sédimentaires et la structuration des communautés prélevées sur ces archives. D'un point de vue biologique, ces communautés microbiennes ont évolué dans un milieu soumis à de nombreuses contaminations et ont pu développer des mécanismes de résistances, d'adaptation et ont été sélectionné par ce milieu. L'alliance de la géochimie et de la multi-omique pourrait permettre d'établir un outil de diagnostic innovant de l'état de santé d'un sol.

En amont du projet lui-même, des échantillons de laisses de la crue exceptionnelle de la Seine en juin 2016 ont permis de caractériser le microbiote au niveau taxonomique et fonctionnelle. Ces résultats préliminaires ont mis en évidence la faisabilité du projet Geomics. Celui-ci se décompose en différentes phases. Dans un premier temps, des tests de protocoles d'extractions de protéines du sol ont permis de sélectionner un protocole permettant de maximiser la quantité de protéines extraites des sols. Ces résultats préliminaires ont mis en évidence des limitations au niveau du taux de spectres interprétables à partir de bases généralistes qui s'est avéré être extrêmement faible.

1.6.4 Les objectifs de la thèse

Dans le cadre du projet Geomics initié par le CEA dans le cadre du programme DRF-Impulsion, ma thèse avait pour mission de contribuer à une analyse stratifiée des communautés microbiennes soumises aux pressions anthropiques afin de caractériser leurs dynamiques en alliant les données géochimiques et omiques. Pour atteindre cette mission, la stratégie d'interprétation des spectres MS/MS doit être améliorée pour que l'identification taxonomique et fonctionnelle des communautés microbiennes de sol par métaprotéomique soit la meilleure possible. Contrairement à des échantillons de microbiotes humains ou de mammifères pour lesquels plusieurs revues comparant les méthodologies et proposant des bases de données adaptées sont disponibles, la nature des échantillons de sols et la méconnaissance des communautés qui y sont hébergées impactent fortement l'identification en métaprotéomique et peu de données méthodologiques sont accessibles pour la communauté scientifique. Ainsi, le premier objectif de ma thèse a été de (1) concevoir et tester différentes stratégies d'interprétation des données métaprotéomiques en travaillant la nature des bases de données utilisées et (2) tester l'apport de cascades de recherches successives. Cette étude systématique a fait l'objet d'une publication dans « Microbiome » le journal de référence du domaine, les résultats obtenus permettant une avancée majeure dans le domaine de la métaprotéomique des sols, puisque les résultats d'interprétation y sont multipliés par un facteur 4 !

Pour contribuer à l'analyse de la dynamique des communautés microbiennes sous pressions anthropiques, l'analyse intégrative des résultats doit établir le lien entre la diversité microbienne, la diversité fonctionnelle et la contamination présente sachant que la formation de ce sol est le résultat de dépôts successifs. Le deuxième objectif de ma thèse a été d'interpréter les données acquises le long d'une carotte de sol prélevée sur le site de Bouafles qui en terme géochimique permet de retracer l'histoire des contaminations du bassin parisien. Les interprétations sur un jeu de données très important regroupant 105 analyses nanoLC-MS/MS ont permis d'identifier la structure des communautés microbiennes avec une granulométrie assez unique dans le domaine de la

métaprotéomique des sols puisque l'analyse verticale de sédiments est la première à être conduite. Cette étude a permis de mettre en évidence les fluctuations verticales en termes de structure des communautés, ainsi que de décrire les fonctionnalités de ces communautés. Les liens possibles entre ces communautés, leurs fonctions métaboliques, et la présence de contaminants y sont également abordés.

Chapitre 2 : Matériels et Méthodes

Le projet de thèse s'axe sur un site bien documenté, le site de Bouafles, situé en aval de Paris le long de la Seine, dans le département de l'Eure. La zone de prélèvement est une plaine d'inondation régulièrement inondée permettant un suivi des concentrations des contaminations et une datation précise des carottes prélevées sur le site. Ce site répond aux critères de sédimentation laminée, continue et régulière, indispensable pour ce type d'archives historiques.

Les 23 et 24 mai 2018, une carotte de sol de 1 mètre a été prélevée avec un carottier « sol » à percussion en inox (Figure 13, Figure 27-b). Le site de prélèvement de la carotte a été labouré en 2016, suite à la mise en culture qui limite désormais l'utilisation de ce site pour de futures analyses sur les 20 dernières décennies. La partie supérieure de la carotte (~16 cm) correspondant à la profondeur de labour est attendue homogène au niveau des concentrations géochimiques mais pas forcément au niveau des communautés microbiennes.

La carotte a été analysée en métaprotéomique, métagénomique, métabarcoding à l'aide du gène 16S rRNA et géochimique de manière stratifiée en délimitant 35 couches successives selon l'axe de la profondeur. Chacune de ces couches a été analysée trois fois par métaprotéomique (réplica technique), résultant en 105 métaprotéomes cumulant 5.5 millions de spectres MS/MS. Des données partielles ont également été acquises grâce aux méthodes de séquences de l'ARNr 16S (séquenceur MiSeq Illumina) et métagénomique (séquenceur HiSeq 4000 Illumina) en regroupant les 35 couches en 5 couches de profondeur variables (Tableau 1). En parallèle, au LSCE, la carotte a été datée au Césium^{137} par spectrométrie gamma et 26 éléments géochimiques par spectrométrie de masse (ICP-MS) ont été mesurés (Tableau 2) . Chaque mesure est associée à une mesure d'incertitude correspondant à l'erreur de mesure associée à l'expérimentation. Les 4 dernières couches de la carotte n'ont pu être datées, la concentration en ^{137}Cs étant trop faible, et en conséquence les mesures des éléments chimiques n'ont pas été faites. Seulement 31 couches sont associées à des données géochimiques.

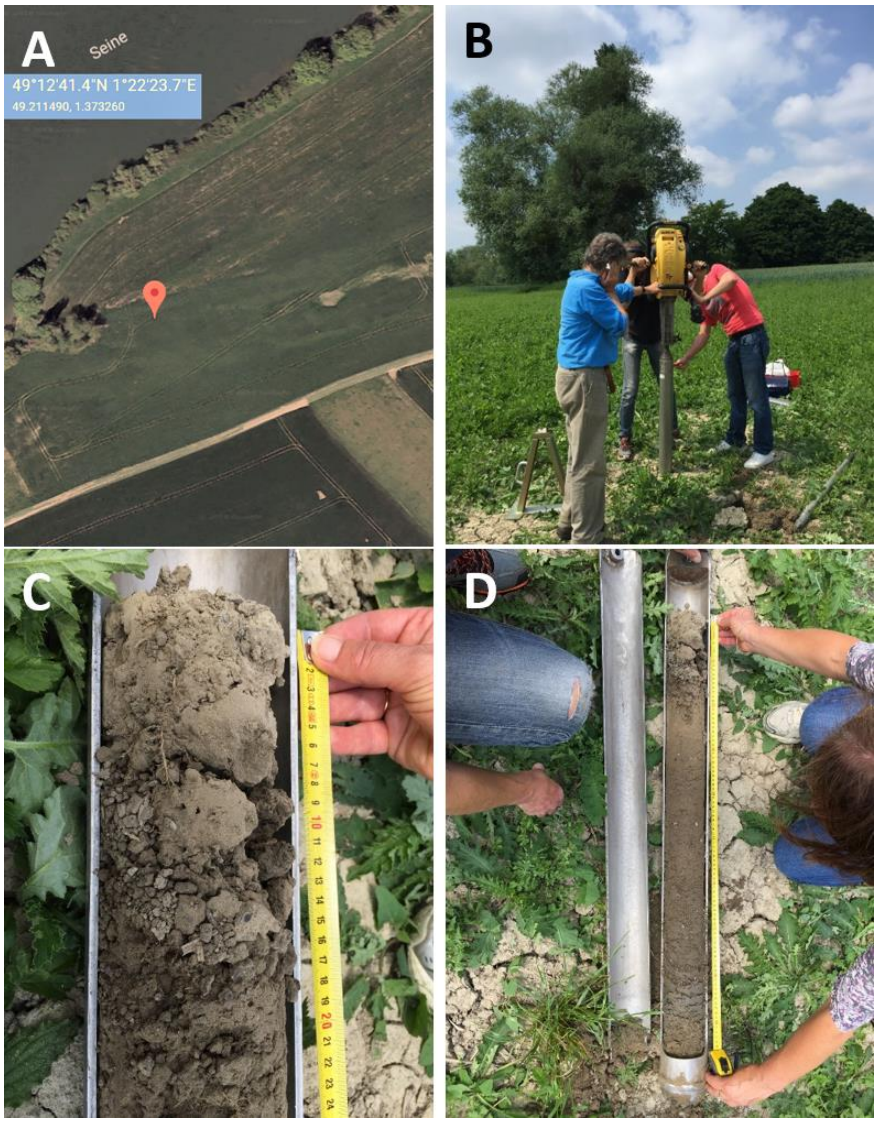


Figure 27: La carotte de sédiment a été prélevée à Bouafles aux coordonnées (A) à l'aide d'un carottier « sol » à percussion (B). Une zone de labour en surface de la carotte a été constatée et mesurée (C) ainsi que la carotte entière a été mesuré (D).

Tableau 1 : Profondeur des couches de la carotte de sédiments utilisées pour le séquençage métagénomique et de l'ADNr 16S.

	Profondeur (cm)	Métagénomique	ADNr 16S
B18.3.G1	0 à 6	41 891 615	39 328
B18.3.G2	6 à 15	39 407 940	43 598
B18.3.G3	15 à 308	43 399 203	40 048
B18.3.G4	30 à 56	40 418 529	33 659
B18.3.G5	56 à 97	41 730 223	35 827
TOTAL		**206 847 510**	**192 460**

Tableau 2 : Eléments géochimiques mesurés et leurs incertitudes calculées en pourcentage.

Eléments	**Na**	**Mg**	**Al**	**P**	**K**	**Ca**	**Ti**	**V**	**Cr**	**Mn**	**Fe**
Numéro atomique	23	25	27	31	39	43	48	51	52	55	57
Incertitude (%)	10	10	10	30	15	5	20	5	5	5	5
Eléments	**Co**	**Ni**	**Cu**	**Zn**	**As**	**Rb**	**Sr**	**Mo**	**Ag**	**Cd**	**Sb**
Numéro atomique	59	60	65	66	75	85	88	98	109	111	132
Incertitude (%)	5	2	2	2	20	5	10	20	10	5	5
Eléments	**Cs**	**Ba**	**Tl**	**Pb**							
Numéro atomique	133	138	205	207							
Incertitude (%)	5	5	5	5							

Chapitre 3 : Améliorer l'interprétation des données métaprotéomiques du sol

Contexte :

Les sols sont une matrice complexe pour les analyses métaprotéomiques, notamment les micro-organismes présents sont très diversifiés et sont largement sous-représentés dans les bases de données publiques rendant leur caractérisation difficile pour ne pas dire impossible. Il est essentiel de construire une base de données de séquences de protéines la plus adéquate possible afin de maximiser la découverte de protéines d'intérêt dans les échantillons cibles. Cette base de données doit être exhaustive en incluant les organismes d'intérêt qui sont présents dans l'échantillon tout en ayant une taille réduite pour limiter les biais d'analyses et augmenter la sensibilité. Généralement, il est conseillé en métaprotéomique d'utiliser des métagénomes réalisés sur les mêmes échantillons afin de constituer la base de données nécessaire aux analyses protéiques. Une autre approche alternative, mais qui est pour l'heure peu exploitée, consiste à utiliser les bases généralistes de séquences de protéines telles que NCBInr ou Uniprot. Cette alternative permet de réduire le cout des analyses métaprotéomiques mais aussi le temps dédié à l'expérimentation et l'interprétation, puisqu'il ne serait pas nécessaire de réaliser de métagénomes profonds sur les mêmes échantillons.

L'étude présentée dans le chapitre 2 a pour objectif de maximiser le taux de spectres interprétables dans le cas d'échantillon de sol à travers la comparaison de différentes stratégies d'interprétation des données métaprotéomiques. Cette étude a été conçue en préalable à l'application plus systématique sur une cohorte d'échantillons et nous a permis d'appréhender les limites des deux approches : métaprotéomique guidée par métagénomique d'une part, et métaprotéomique à partir de bases généralistes d'autre part. Pour cette étude, différentes stratégies de construction de bases de données et de requêtes en cascade ont été testées et comparées. Les résultats des requêtes utilisant soit des bases de données métagénomiques spécifiques aux échantillons, soit des bases de données publiques, ont conduit à des résultats similaires, et un faible taux d'identification attendu dans le cas de métaprotéomique de sols dû à la complexité et diversité microbienne des échantillons. Cependant, une recherche en cascade en deux étapes a permis d'obtenir de meilleurs résultats quelques soit le type de bases de données utilisés. Enfin, la stratégie de combiner les bases de données généralistes, spécifiques au sol et les métagénomes dédiés a permis de maximiser les résultats d'interprétation des spectres en termes d'annotation fonctionnelle des peptides. Cette étude, publiée dans le journal de référence dans le domaine, « Microbiome », a été très favorablement appréciée par les reviewers et l'éditeur en charge de son évaluation qui souligne qu'elle permet à la communauté scientifique de mesurer l'intérêt d'investir dans des stratégies complexes d'interprétation permettant de maximiser le résultat.

Jouffret et al. Microbiome (2021) 9:195
https://doi.org/10.1186/s40168-021-01139-1

Microbiome

METHODOLOGY Open Access

Increasing the power of interpretation for soil metaproteomics data

Virginie Jouffret[1,2,3], Guylaine Miotello[1], Karen Culotta[1], Sophie Ayrault[2], Olivier Pible[1] and Jean Armengaud[1*]

Abstract

Background: Soil and sediment microorganisms are highly phylogenetically diverse but are currently largely underrepresented in public molecular databases. Their functional characterization by means of metaproteomics is usually performed using metagenomic sequences acquired for the same sample. However, such hugely diverse metagenomic datasets are difficult to assemble; in parallel, theoretical proteomes from isolates available in generic databases are of high quality. Both these factors advocate for the use of theoretical proteomes in metaproteomics interpretation pipelines. Here, we examined a number of database construction strategies with a view to increasing the outputs of metaproteomics studies performed on soil samples.

Results: The number of peptide-spectrum matches was found to be of comparable magnitude when using public or sample-specific metagenomics-derived databases. However, numbers were significantly increased when a combination of both types of information was used in a two-step cascaded search. Our data also indicate that the functional annotation of the metaproteomics dataset can be maximized by using a combination of both types of databases.

Conclusions: A two-step strategy combining sample-specific metagenome database and public databases such as the non-redundant NCBI database and a massive soil gene catalog allows maximizing the metaproteomic interpretation both in terms of ratio of assigned spectra and retrieval of function-derived information.

Keywords: Bioinformatics, Cascaded search, Database, Interpretation, Metagenomics, Metaproteomics, Microbiome, Soil, Tandem mass spectrometry

Background

Soil hosts complex microbial ecosystems which are crucial for numerous ecosystem services, including plant growth and animal life [71]. These ecosystems can be affected by anthropogenic pressure and climate change [29]; therefore, it is important to understand their structure and how they function [6]. Due to the broad diversity of components they include and their dynamic relationships, soil microbial ecosystems are complex by nature [17]. Indeed, soils are open systems exposed to highly variable environmental parameters such as temperature, hygrometry, gas, metal, and chemical contaminants, which can influence microbial populations and their functions. Thanks to improved meta-omics technologies, the number of in-depth molecular studies of soil environments is increasing [58]. Since the pioneering metagenomics works almost two decades ago, molecular phenotyping approaches such as metatranscriptomics, metaproteomics, and meta-metabolomics have emerged and been used to attempt to understand how these systems function at various levels. Specifically, metaproteomics allows the identification and quantification of proteins, which are the workhorses of the cells, and can be used to monitor more integrated levels, such as pathways and general functions [56, 70]. Humic acids and potential contaminants may interfere with protein

* Correspondence: jean.armengaud@cea.fr
[1]Université Paris-Saclay, CEA, INRAE, Département Médicaments et Technologies pour la Santé (DMTS), SPI, F-30200 Bagnols-sur-Cèze, France
Full list of author information is available at the end of the article

extraction, thus metaproteomics methods must be specifically developed to suit each soil type [32, 63]. Despite these difficulties, several pioneering studies have been performed on soils extracted from forests [40, 74], arid environments [7], agricultural areas [39, 50], permafrost [27], and from mining drainage [53]. Sediments — deposited material arising from weathering, erosion, and transport processes — also contain complex microbial ecosystems [19, 64].

Metaproteomics involves protein extraction and trypsin proteolysis, detection of the resulting peptides by tandem mass spectrometry, interpretation of MS/MS spectra to assign peptide identities, and higher-level interpretation in terms of taxonomy and function [34, 49]. MS/MS spectra acquired in metaproteomics studies are interpreted by comparison to a database listing the sequences of all the proteins potentially present in the sample. To create such a database, the most appropriate strategy is to perform metagenomics or metatranscriptomics on the same sample. These databases can then be translated (in six- or three-frames) to derive the theoretical protein sequences. Alternatively, protein sequences from the organisms identified in similar samples can be compiled for complementing metagenomics information [27, 75]. Another alternative is to assemble a specific database based on the organisms identified after 16S rRNA amplicon sequencing and taxonomical assignment [73] or potentially present in the habitat where the sample was obtained [9]. The choice made between read- or contig-based databases may influence the identification rate. For animal metaproteomics, a contig-based database has been shown to be the most productive strategy [62]. However, if the necessary metagenomics information is not available, generalist databases such as NCBInr or UniProtKB/Swiss-Prot can also be used [24]. Despite these multiple options, the large diversity and dynamic range of taxa contained in some samples, such as soils and sediments, represents a true challenge for metaproteomics interpretation and limits protein identification [58, 70]. Indeed, this diversity results in a search space for metaproteomics databases that is naturally much larger than that required for single-organism proteomics. To counteract the negative effects of an inflated database size on sensitivity and accuracy of peptide-to-spectrum matching (PSM), several strategies have been proposed. These include database reduction using a two-step search [28], where matches derived from the first search — performed without false discovery rate (FDR) threshold — are used for a second search round, during which a stringent threshold is applied. This type of cascaded search was successfully implemented to define the metaproteome of the gut microbiota from a sentinel, non-sequenced animal [20], and lichen-associated bacterial communities [11]. These databases are protein-centric, i.e., focused on the main proteins across all clades, and can thus successfully highlight the main functions at play within the most abundant microbial organisms from the ecosystem sampled.

Soil/sediment metaproteomics is currently challenging because a large proportion of organisms in soil samples have yet to be taxonomically characterized [47] and only a small fraction of reference genome sequences are available in public data repositories. Furthermore, the community structure of such samples may vary dramatically over time and space. Although numerous large-scale metagenomics studies have been performed on soil samples [5, 51], the contribution of specific soil gene catalogs to improving metaproteomics interpretation has not yet been estimated. In this study, we recorded metaproteomics data from a soil core consisting of the annual sediment deposit in a floodplain which provides long-term records of particle-bound pollutants (metals, radionuclides, pharmaceuticals, and numerous persistent organic pollutants) released by the Seine River (France), including effluents from the Parisian megacity [1]. We tested several strategies when interpreting the metaproteomics data acquired for this soil sample. These strategies included sample-specific metagenomics data, a topsoil gene catalog constructed from a large diversity of sites [5], and genome sequences from reference microorganisms. We found that a significant increase in the numbers of MS/MS spectra interpreted and functionally annotated was obtained when a combination of all types of information was used in an appropriate cascaded search strategy. The results substantially improved our understanding of the soil microbiota.

Results

Benchmarking databases created from sample-specific metagenomics data

Different databases built from metagenomics data acquired on a sediment sample were evaluated for metaproteomics based on the number of PSMs as main criterion. For this, a soil core was collected from the Seine River floodplain at the Bouafles site (France) located downstream of Paris. The 1-m core was cut up into 3-cm slices. A shotgun metagenome sequencing dataset comprising ~ 87 million Illumina paired-end reads was acquired for the slices corresponding to 17–28-cm depth in the soil core after extracting DNA from a pool of the five corresponding slices. Figure 1 shows the five options used to construct the sequence databases: (i) reads were assembled with MEGAHIT and the resulting contigs were translated in the six possible reading frames (MGF-6RF), (ii) selected based on coding gene sequences predicted by FragGeneScan tool (MGH-FGS), (iii) reads were assembled directly at the protein level using PLASS assembler (PLASS), (iv) coding sequences were selected

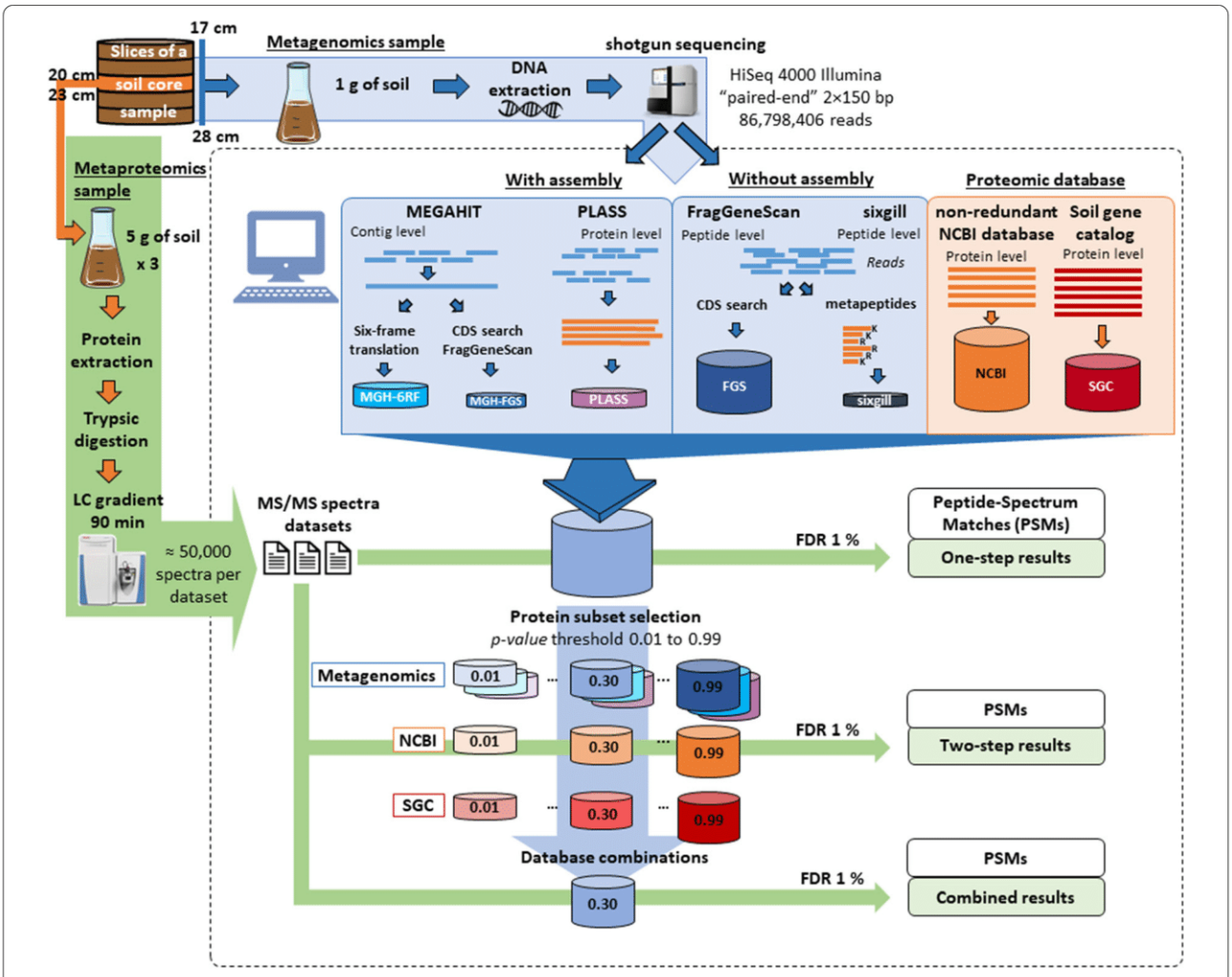


Fig. 1 Workflows assessed for metaproteomics analysis of environmental samples. The left (green) and top (blue) arrows show the experimental steps tested to obtain MS/MS spectra and metagenomics reads. The dashed-line panel presents the bioinformatics analysis, with, at the top, database construction either from metagenomics data with or without assembly step, or from generic databases. The bottom part shows the process used to refine and enhance the attribution rate through multiple search rounds. A first analysis on full databases was used to select protein-based subsets at various *p*-values. New searches on these subsets then identified the databases performing best

from reads by FragGeneScan without assembly step (FGS), or (v) selected only tryptic peptides capable of undergoing tandem mass spectrometry, as intended by sixgill (sixgill). Table 1 indicates the number of sequences and size of the resulting databases. First, reads were directly assembled using the MEGAHIT tool, which has been benchmarked as one of the best assemblers [69], resulting in 972,629 contigs with 60.8% GC content, 939 N50, and 101,728 L50. The largest contig length was 48,284. A systematic six-reading-frame translation was used to produce the MGH-6RF database, which comprises almost 22 million possible protein sequences and a billion amino acid residues. To decrease the size of the database and remove erroneous polypeptide sequences, the FragGeneScan tool was then used to select predicted protein-coding sequences (CDS). The resulting MGH-FGS database is much more focused, retaining only 17% of the information contained in MGH-6RF. A third database was created by assembling reads at the protein level using the PLASS assembler. This strategy bypasses silent single nucleotide sequencing errors and compresses the possible single nucleotide polymorphisms that could occur across closely phylogenetically related strains present in the sample. It should be noted, however, that this tool may lead to chimeric assemblies between similar protein sequences. Application of the PLASS assembler resulted in a database containing 16 million proteins with a mean length of 112 amino acids, which is a significant increase in size (+ 80%) compared to the proteins listed in MGH-6RF. To avoid possible bias due to assembly of metagenome reads either at the nucleotide sequence level or at the

Table 1 Sample-specific metagenomic databases and generic databases

Databases	Tools used/database origin	Computational time[a] (hours)	Size of the database (in residues)	Number of protein entries
MGH-6RF	MEGAHIT + six-frames translation	13	1,028,880,437	21,883,653
MGH-FGS	MEGAHIT + FragGeneScan	13	168,662,946	1,269,322
PLASS	PLASS	6	1,784,677,737	16,004,028
FGS	FragGeneScan	43[b]	2,939,955,188	72,130,656
sixgill	Sixgill	5.5	82,314,892	2,577,349
NCBI	Non-redundant NCBI	-	41,817,980,956	108,307,546
SGC	Soil gene catalog	-	21,962,323,955	159,657,012

[a]Computer used: 10 CPU, 240 Gb RAM memory
[b]Single thread

amino acid sequence level, small truncated polypeptide sequences can be directly predicted from the short reads. The FragGeneScan tool performs this type of prediction and was used to produce the FGS database, which is three times larger than MGH-6RF. Finally, the sixgill algorithm was applied to directly produce a list of putative tryptic peptides amenable to tandem mass spectrometry which are well represented in at least two reads. The resulting sixgill database was rather small, representing only 8% of the size of MGH-6RF.

Proteins were extracted from three equal aliquots of the same section of the soil sample (slice 20–23 cm of the sampled soil core). The peptides derived from the biological triplicates after trypsin proteolysis were analyzed by nanoLC-MS/MS, producing 59,501, 59,917, and 59,141 MS/MS spectra. These three datasets were subsequently used separately to estimate search variability across the different databases even if the biological samples taken for metagenomics and metaproteomics do not match perfectly. Figure 1 shows the two strategies used to interpret MS/MS signals with the five databases. First, databases were queried at the same 1% FDR in a one-step search strategy. Because decoy database searches are problematic for large database [15, 26] with increased occurrence of reversed peptide sequences corresponding to true peptide sequences and variability depending on how the decoy is constructed, we used a decoy-free FDR evaluation for this. As recommended by Jagtap et al. [28], a two-step database search strategy was also conducted. The first search round selected protein sequences at low stringency, whereas the second search performed with this sub-database validated the most relevant hits. In this case, several *p*-value thresholds (0.01, 0.05, 0.10, 0.20, 0.30, 0.50, 0.70, 0.80, 0.90, 0.99) were tested for the first-round search to estimate the impact of this parameter on the final results. The second search was performed at decoy-free 1% FDR. Figure 2 presents the results obtained following application of the two strategies, in terms of PSM attribution rate. The *X*-axis represents the size of the databases used in the final step of the cascaded search. Notably, for all conditions tested, the result variability estimated on the three experimental metaproteomic datasets was quite low, at less than 0.5% in most cases. The one-step search method allowed between 2.2 and 6.5% of MS/MS spectra to be assigned, with the maximum reached using the PLASS database. The sixgill database search performed poorly (only 2.2% MS/MS spectra assigned) even though it was the smallest, and theoretically the best-adapted to the proteomic data format. The two-step database search method significantly increased the proportion of MS/MS spectra assigned, with 3-fold higher values recorded for most conditions. Although this increase was expected, the results reveal that the improvement ratio depends strongly on the stringency of protein selection during the first identification round. Here, optimal *p*-values could clearly be identified for each database: 0.10 for sixgill, 0.20 for MGH-6RF and PLASS, and 0.30 for MGH-FGS and FGS. Using the two-step search method, higher numbers of confident PSMs were assigned, reaching at best 7.3% for sixgill, 13.0% for MGH-FGS, 14.1% for MGH-6RF, 17.3% for PLASS, and 20.5% for FGS. As with the one-step search, the two-step search strategy performed better with the FGS and PLASS databases, but a clear advantage was noted for the FGS database. Unexpectedly, among the sequencing-read-assembly strategies, a better attribution rate was obtained for PLASS compared to MEGAHIT. This result highlights the power and reliability of a strategy based on assembly of peptide sequences rather than nucleic acid sequences and demonstrates the added value of retaining variants that are discarded by the MEGAHIT algorithm. These results also show that predicting coding sequences after assembly (MGH-FGS) does not provide significant advantages over six-frame translation (MGH-6RF) in the two-step search method, as these databases allowed 13.0% and 14.1% MS/MS assignment, respectively. This result directly contrasted with that of the one-step search strategy, where 6.1% and 4.5% of MS/MS spectra were assigned, respectively. In conclusion, the highest attribution rate (20.5%) and coverage of the microbial metaproteome was obtained with the FGS database

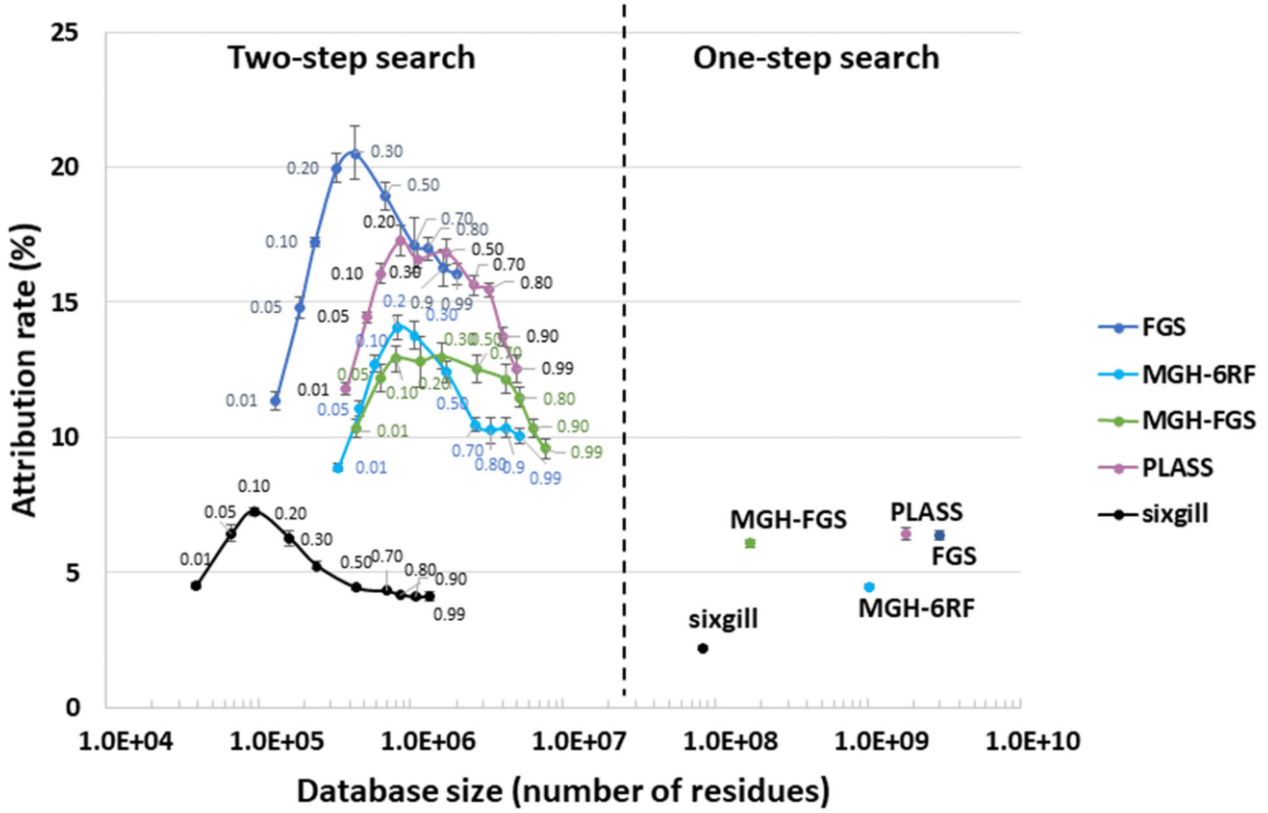


Fig. 2 Attribution rates for MS/MS spectra, obtained with sample-specific metagenomics databases. Databases were generated using the following workflows: FragGeneScan only (FGS), MEGAHIT and FragGeneScan (MGH-FGS), MEGAHIT and six-frame translation (MGH-6RF), sixgill, and PLASS. The results obtained when using the complete databases are indicated on the right (one-step search), whereas the results obtained on refined databases containing proteins selected at distinct *p*-values during the first-round search are shown on the left. Searches were performed at decoy-free FDR 1% and results are ordered based on the database size, using a logarithmic X-axis. The small numbers indicate the *p*-value threshold used for each search. Three biological replicates were used to calculate the mean and standard deviation values.

when queries were performed at *p*-value 0.30 in the first search round, using a large database with only short sequences (mean length, 40.7 amino acids).

Assessing the potential of generic databases

As shown in Fig. 1, two generic databases were also used to interpret the three MS/MS datasets: the giant NCBInr database, totaling 41.8 billion residues and comprising the protein sequences from 27,137 species; and the Soil Gene Catalog (SGC) database compiling an extensive catalog of genes established by metagenomics of numerous topsoil samples [5]. This SGC database is twofold smaller than the NCBInr database. Figure 3 shows the results of the one-step and two-step search methods. The proportion of MS/MS spectra assigned at decoy-free FDR 1% with these two databases was low when used directly: 0.9% for NCBInr and 2.8% for SGC. Such a result was expected with the two generic databases, as the first one is not specifically representative of microorganisms likely to be present in soil samples, but also due to the huge size of the two generic databases, which hinders correct FDR estimates. Using the two-step search method, the ratio of assigned spectra increased significantly (Fig. 3). At the optimal *p*-value thresholds, SGC performed better than NCBInr, with 19.5% versus 15.5% of MS/MS spectra assigned to peptide sequences. Because the two databases could be complementary in terms of environmental sequence coverage, we also assessed the effect of merging SGC and NCBI sub-databases at various *p*-values (SGC+NCBI). As indicated in Fig. 3, the proportion of MS/MS spectra assigned was increased to 23.8% when using this combined database, for which the optimal *p*-value threshold was 0.30. Interestingly, this assignment ratio was higher than that obtained with the classical approach, consisting in nucleic acid sequence assembly and protein sequence prediction (MGH-FGS). These results suggest that, in the future, generic databases — which are continuously expanding to include new environmental metagenomics projects and NCBInr updates — could perform as well as sample-specific metagenomics databases, even when treating difficult environmental samples. This prospect

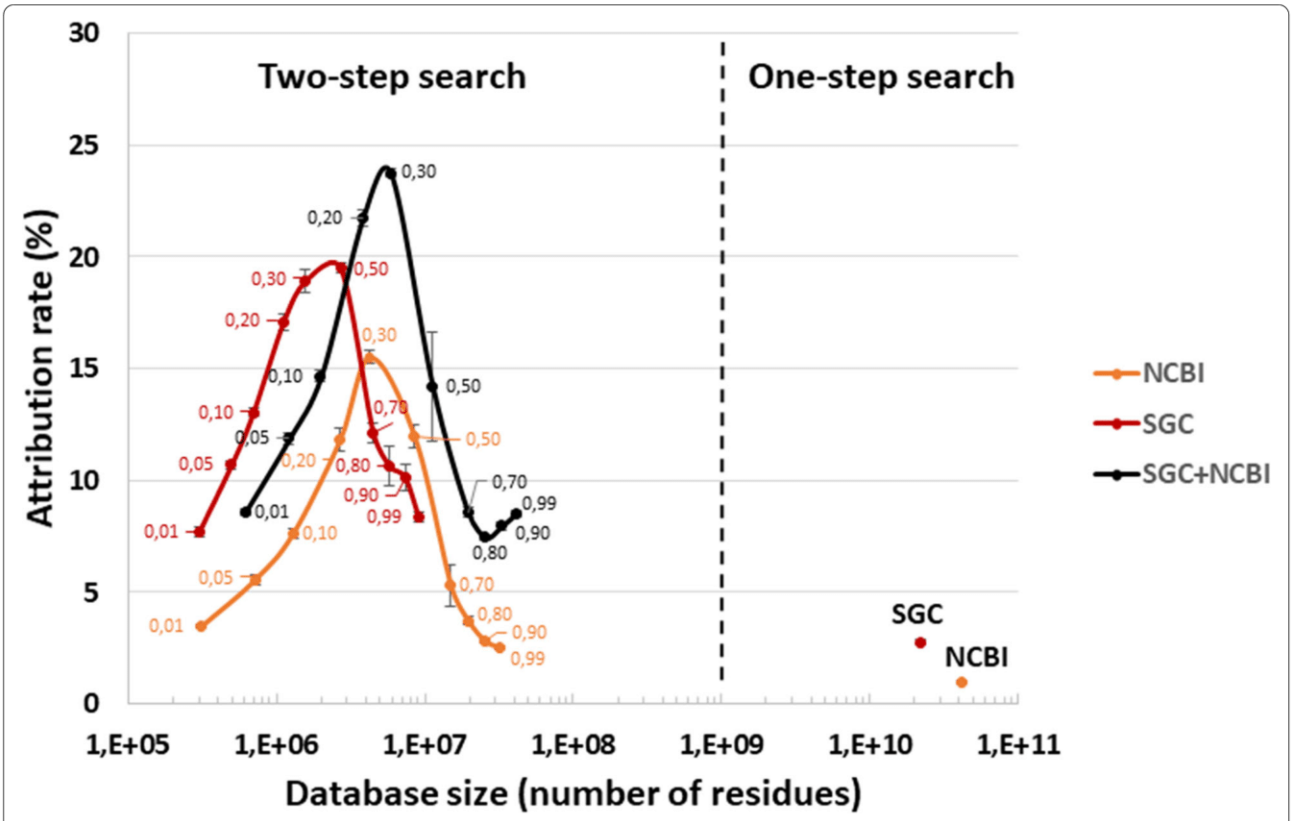


Fig. 3 Attribution rates for MS/MS spectra when using generic databases. The results obtained with SGC and NCBI are shown as a function of the database size using a logarithmic *X*-axis. The results obtained with the whole databases are presented on the right (one-step search), whereas the results obtained on refined databases containing proteins selected at distinct *p*-values during the first-round search are indicated on the left (two-step search). The SGC+NCBI databases were obtained by merging the reduced NCBI and SGC protein databases at each *p*-value cutoff. Searches were performed at decoy-free FDR 1%. The small numbers indicate the *p*-value threshold used for each search. Three biological replicates were used to calculate the mean and standard deviation values.

would decrease the per-sample cost of metaproteomics analyses.

Combining sample-specific metagenomics data and generic databases

We next went on to test the effect of a combination of sample-specific metagenomics databases and generic databases on the attribution rate for MS/MS spectra. The best-performing reduced databases from the two-step search strategy were selected: FGS_0.30, PLASS_0.30, MGH-6RF_0.30, NCBI_0.30, and SGC_0.30 (at the most common optimal *p*-value of 0.30). Table 2 reports the number of sequences and residues contained in these reduced databases. In addition, we created two new databases comprising only the peptides detected in the two-step search performed with the generalist databases, resulting in the NCBIp_0.30 and SGCp_0.30 databases. Table 2 shows the 16 combinations of databases tested in this new round of MS/MS interpretation, their sizes, and the assignment rate obtained at decoy-free FDR 1%. Combining the reduced FGS_0.30 and NCBI_0.30 databases for a single search resulted in an average of 24.9% of MS/MS spectra assigned for the three metaproteomic datasets. This proportion represents a significant increase compared to the optimal FGS_0.30 database (20.5%). Reduced FGS_0.30 and NCBIp_0.30 also performed well, with 24.8% spectra assigned, but a greater variability was noted. Use of the reduced SGC_0.30 and FGS_0.30 databases also resulted in a higher number of PSMs (25.9%) compared to FGS_0.30 alone. Concatenation of the FGS_0.30, SGC_0.30, and NCBI_0.30 sub-databases slightly improve results (26.2% MS/MS assignment). The same trend was observed with combinations of PLASS_0.30 and general sub-databases. Indeed, PLASS_0.30+SGC_0.30 (24.8%) performed better than PLASS_0.30+NCBI_0.30 (22.4%) and PLASS_0.30+SGC_0.30+NCBI_0.30 (24.6%). MGH-FGS_0.30+SGC_0.30 (24.4%) performed less than MGH-FGS_0.30+SGC_0.30+NCBI_0.30 (24.8%). The alternative MGH-6RF database performed slightly less with 23.4% combined with SGC_0.30 and 23.6% with SGC_0.30+NCBI_0.30. Decreasing the size of the merged database by selecting only the peptide sequences detected in a two-round search did not systematically increase the assignment

Table 2 Combining sample-specific metagenomic databases and generic databases

Combined databases	Size of the database (in residues)		Number of sequence entries		Attribution rate (%)	
	Mean	sd (%)	Mean	sd (%)	Mean	sd (%)
SGC_0.30 +NCBI_0.30	5,849,545	1.9	15,340	2.1	23.76	0.2
SGCp_0.30 +NCBIp_0.30	220,207	8.8	18,644	8.8	23.73	6.9
FGS_0.30 +NCBI_0.30	4,670,124	2.3	17,101	2.4	24.98	0.6
FGS_0.30 +SGC_0.30	1,990,818	1.7	18,073	2.5	25.94	0.4
FGS_0.30 +SGC_0.30 +NCBI_0.30	6,275,438	1.9	25,298	2.3	26.21	0.9
FGS_0.30 +NCBIp_0.30	525,294	2.6	18,878	2.5	24.84	1.1
FGS_0.30 +SGCp_0.30 +NCBIp_0.30	646,100	1.9	28,603	4.8	27.24	5.8
PLASS_0.30 +NCBI_0.30	5,351,198	2.1	15,541	2.0	22.42	1.2
PLASS_0.30 +SGC_0.30	2,673,152	1.8	16,512	2.2	24.83	0.6
PLASS_0.30 +SGC_0.30 +NCBI_0.30	6,957,772	1.8	23,738	2.1	24.62	0.5
PLASS_0.30 +NCBIp_0.30	1,206,368	2.2	17,318	2.2	22.05	0.7
PLASS_0.30 +SGCp_0.30 +NCBIp_0.30	1,327,067	1.2	27,042	5.2	25.42	4.7
MGH-FGS_0.30 +SGC_0.30	3,158,743	1.8	15,862	2.1	24.40	0.4
MGH-FGS_0.30 +SGC _0.30 +NCBI_0.30	7,443,363	1.7	23,088	2.0	24.83	0.8
MGH-6RF_0.30 + SGC _0.30	2,629,015	2.0	15,664	1.9	23.43	0.8
MGH-6RF_0.30 + SGC _0.30 +NCBI_0.30	6,913,635	1.9	22,890	1.9	23.58	0.4

Mean and standard deviation (sd) were calculated based on the three biological replicates

rate, as shown when comparing PLASS_0.30+NCBIp_0.30 (22.1%) and PLASS_0.30+NCBI_0.30 (22.4%). In conclusion, our results demonstrated that, for experimental soil metaproteomics data, the assembly of metagenomics reads at either the nucleic acid (MGH) or the polypeptide level (PLASS) could be detrimental to the MS/MS spectrum assignment rate compared to direct use of reads (FGS). Here, the highest MS/MS spectral assignment rate was obtained when a sample-specific metagenomics database was combined with generalist databases in a two-round search strategy.

Functional annotation with the optimal combined databases

As the aim of metaproteomics is to analyze the function of the proteins identified, we next assessed the levels of functional annotation obtained with the four databases performing best in terms of attribution rates, when used alone or in combination. Figure 4 shows the functional annotation obtained following three processes. First, peptides identified at FDR 1% using the FGS_0.30, PLASS_0.30, SGC_0.30, and NCBI_0.30 databases, and combined databases were annotated by applying the Unipept tool which is based on the lowest common ancestor approach. This peptide-based functional annotator returns molecular function (MF) and biological process (BP) Gene Ontology (GO) terms, and enzyme commission (EC) numbers. In parallel, identified proteins were annotated using GO slim level and KEGG Orthology (KO) terms by the Diamond BLASTP and GhostKOALA tools, respectively. Notably, here, Unipept annotation produced less annotated MS/MS spectra than Uniref50 BLASTP searches, suggesting that protein level functional annotation is more powerful than peptide level. In terms of databases, PLASS_0.30 and FGS_0.30 performed well, as judged using the Uniref50-based GO BP annotation, with 13.2% and 13.1% of MS/MS spectra functionally annotated, respectively (Fig. 4). Interestingly, PLASS_0.30 performed better at the functional level than at the attribution rate level. SGC_0.30 database performed better than NCBI_0.30 database with 15.3% and 11.3% of MS/MS spectra functionally annotated, respectively. For the four databases, between 64 and 81% of PSMs were functionally annotated. SGC_0.30+NCBI_0.30 performed better than individual metagenomics databases, with 19.7% of spectra annotated. The combination of metagenomics and generic databases was very efficient to improve the functional attribution rate compared to the standalone databases: the combinations of each standalone MGH-FGS_0.30, PLASS_0.30, and FGS_0.30 databases with SGC_0.30 and NCBI_0.30 databases allowed 20.6%, 20.3%, and 20.0% of spectra to be functionally annotated, respectively. Regarding the GhostKOALA-based KO annotation (Fig. 4), the same trend was observed, with maximized functional annotation obtained when using databases combining metagenomics and generic information. Thus, MGH-FGS_0.30+SGC_0.30+NCBI_0.30 and PLASS_0.30+SGC_0.30+NCBI_0.30 provided 13.4% and 13.1% of functionally annotated spectra, respectively, with a slightly higher contribution from soil gene catalog database. With the reduced

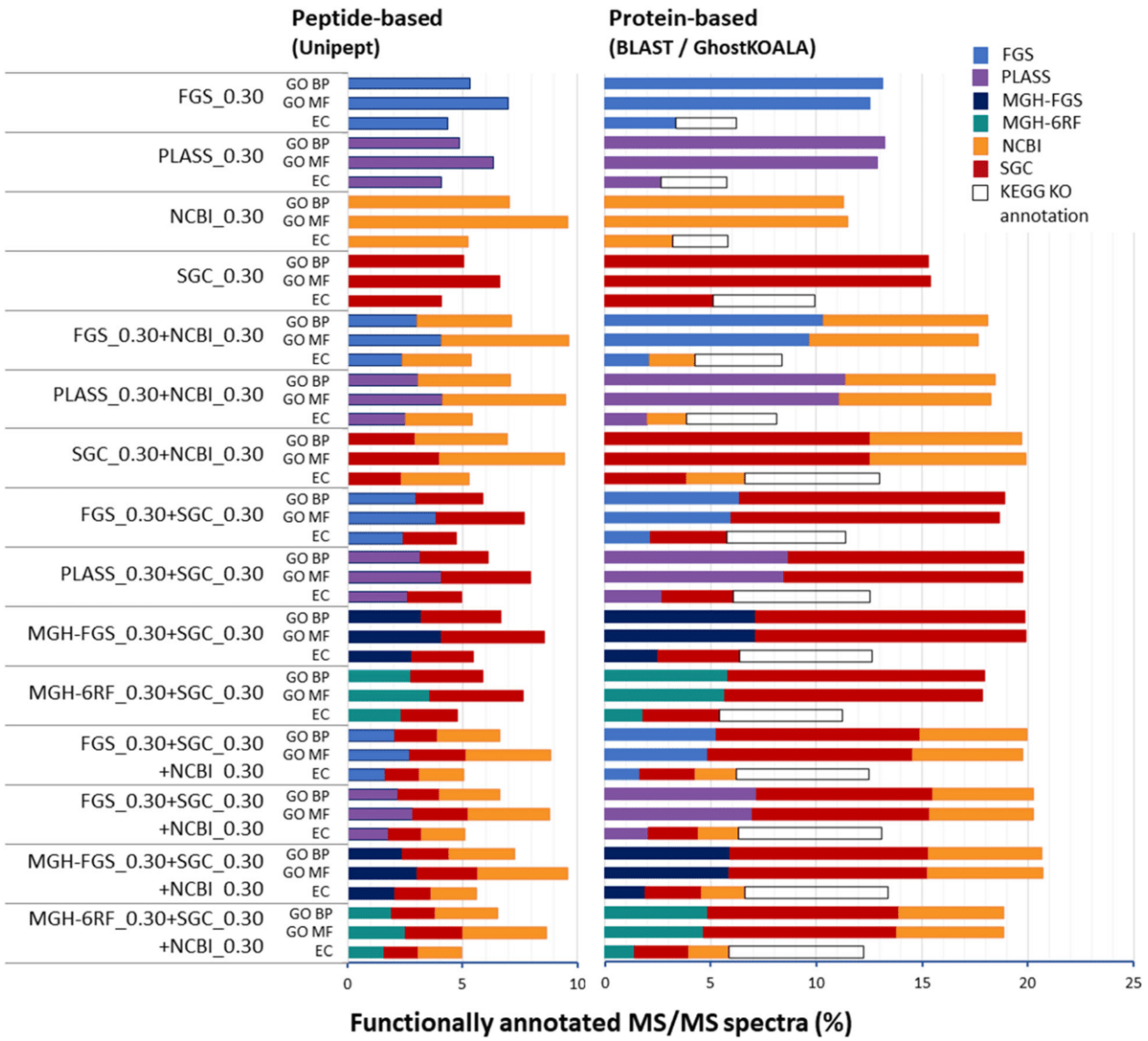


Fig. 4 Functional annotation of peptides and proteins identified at FDR 1%. The databases used to identify proteins are indicated on the left. For each database, the percentage of PSMs annotated are indicated in terms of GO_BP, GO_MF terms, and EC numbers obtained either by Unipept or using Diamond BLASTp queries on Uniref50. In the latter case, the GO OWL tools were used to retrieve GO slim annotations; EC numbers and KO entries were retrieved using the GhostKOALA web-service. The grey squared areas on the EC lines represent the KEGG KO annotation level, from which EC numbers were extracted

SGC_0.30 and NCBI_0.30 databases, 9.9% and 5.8% of spectra were functionally annotated, respectively. When combined as SGC_0.30+NCBI_0.30, 13.0% of spectra were functionally annotated, representing 55% of PSMs for the database, with a higher contribution of SGC. In conclusion, the best result in terms of functional annotation was obtained when a sample-specific metagenomics database was combined with a generalist database and soil-specific database in a two-round search strategy. Remarkably, combined generic databases allowed 55% of PSMs to be functionally assigned when the different procedures tested were merged.

Consistency of functional and taxonomic annotations

Metagenomics and metaproteomics results were compared in terms of functional annotation based on KO. For this comparison, metagenomics reads obtained for the soil sample were analyzed using the MG-RAST pipeline [44] to produce KO annotations. These results were compared to those obtained following GhostKOALA functional annotation for the metaproteomics dataset, which grouped together the three biological replicates, and was interpreted using the FGS_0.30+SGC_0.30+NCBI_0.30 database. Figure 5A shows three main functional groups: “metabolism”, “genetic information

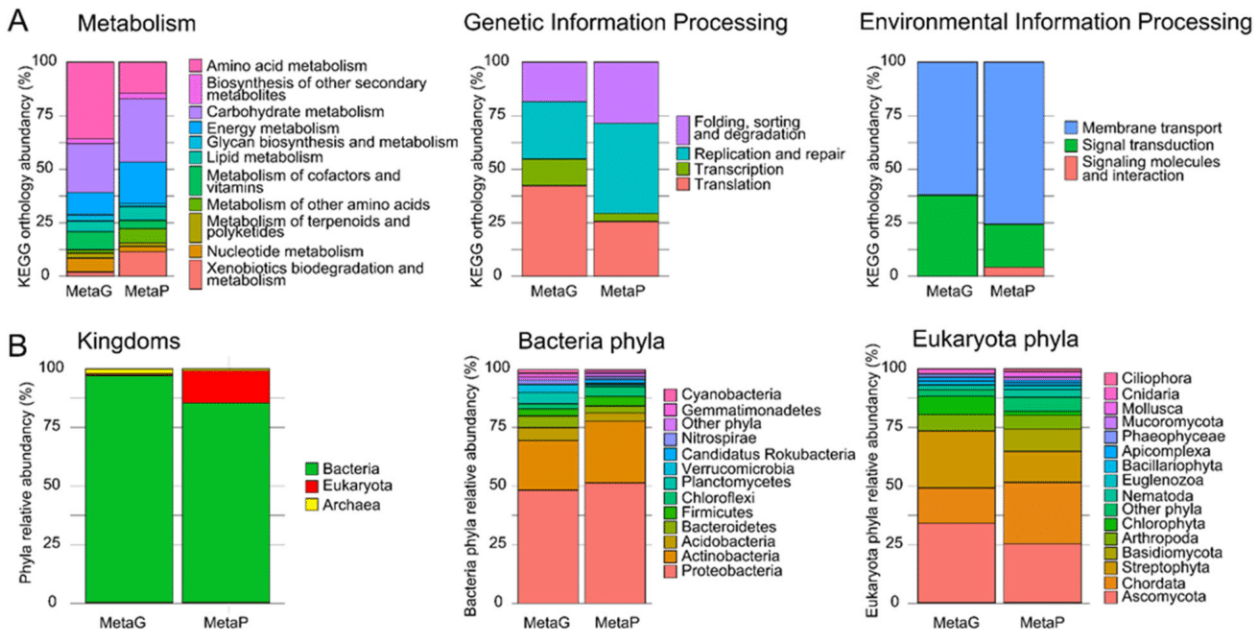


Fig. 5 KO and taxonomical annotation of soil metagenomics data and metaproteomics data. Metagenomics data were annotated using the MG-RAST pipeline, with KO and taxonomic analysis. Peptides and sequences attributed at FDR 1% of the combined reduced sample-specific and generic database FGS_0.30+SGC_0.30+NCBI_0.30 were analyzed using Unipept and GhostKOALA to retrieve taxonomic and functional annotations, respectively. The PSMs for the three biological replicates were cumulated. **A** Functional annotation of three KEGG categories: metabolism (left), genetic information processing (center), and environmental information processing (right). A total of 16% of reads and 72% of PSMs were annotated in terms of function. **B** Taxonomic annotation of metagenomics data (MetaG) and metaproteomics data (MetaP) at Kingdom (left), Bacterial phyla (center), and Eukaryotic phyla (right) levels. Phyla represented less than 1% of the total were merged under "Other phyla". A total of 17% of reads and 21% of PSMs were unequivocally annotated in terms of taxonomy

processing", and "environmental information processing". Remarkably, the different activities within each of these groups were relatively consistent when assessed by the two methodologies, even though they do not rely on the same molecules or measurements. In this soil sample, "amino acid metabolism", and "carbohydrate metabolism" and "energy metabolism" were the most abundant functional categories according to metagenomics and metaproteomics data. Interestingly, metaproteomics allows us to descend deeper into the functional category "signaling molecules and interaction pathway", as proteins classified as "signaling molecules and interaction" are identified. In contrast, this category is under-represented by the metagenomics analysis (Fig. 5A).

Metagenomics and metaproteomics datasets were also interpreted at the taxonomic level. As shown in Fig. 5B, the phyla identified and their ratios were consistent. At the domain level, both methodologies indicated a vast predominance of bacteria in the sample. Within this superkingdom, Proteobacteria and Actinobacteria are the most abundant phyla, but a large diversity of phyla were represented. Remarkably, both methodologies can highlight the presence of a candidatus phylum, namely *Candidatus Rokubacteria*, which was previously reported to predominate in Amazonian rainforest soil [35]. Some discrepancies were noted for the estimated Eukaryote ratio. Clearly, metagenomics underestimated the presence of eukaryotic cells compared to metaproteomics. However, this underestimation is expected as the volume of these cells is much higher than the volume of bacteria, whereas their nucleic acid molecule content is similar, leading to a higher ratio when protein biomass is measured compared to nucleic acid estimation. The eukaryotic phyla identified and their respective quantities are consistent between the two technologies, although the huge diversity present in the sample could have been a source of bias.

Discussion

The aim of this study was to determine the database construction and search approach that would maximize the information extracted by metaproteomics analysis of soil sampled from a floodplain along the Seine River, downstream of Paris (France). Through the list of proteins it provides, and their abundances, metaproteomics brings a new dimension to the study of microbiota by delivering the list of organisms present in a sample and their respective biomasses [52], and by providing information on how the microbial community functions

[12, 34]. Most metaproteomics interpretation pipelines up until now have been evaluated using human microbiome samples such as saliva [28] or feces [48, 62, 67] or laboratory-assembled microbial mixtures [61, 67]. As shown previously with these samples, the choice of the workflow in metaproteomics is critical as it controls the peptide identification. An average of 21% of attribution rate at FDR 1% was obtained with human fecal samples using a combination of the search algorithms X!Tandem and OMSSA against a customized protein sequence database containing 6 millions of proteins from different sources such as metagenomes, bacterial, and human genomes [49]. In comparison, less than 3% of spectra were assigned in our study with the SGC database which comprises 159 million of sequences. The sample preparation and mass spectrometry acquisition parameters are also critical as they may impact the attribution rate. Based on the same bioinformatics workflows, identification rates varying in the range 12 to 35 % were obtained at FDR 1% on a fecal sample using the same reference database (van der Bossche, Kunath et al. 2021). Because of its inherent characteristics, soil is a difficult matrix to work with for metagenomics and metaproteomics [58]. The extent of the diversity of microorganisms in soils is considered a significant bottleneck for the interpretation of omics data in general.

To construct the most appropriate database for use when interpreting metaproteomics data, it is generally recommended to use metagenomics data acquired for the same sample. From the sampled soil, ~ 87 million Illumina paired-end reads were recorded, corresponding to 13 Gbp of sequenced nucleotides. Whether this sequencing effort comprehensively represents the microbial community found in the sample is a key question. In some soil studies, the cumulated efforts made to analyze a large collection of samples is considerably greater. For example, a total of 730 Gbp of sequenced nucleotides were obtained for the analysis of soil communities in phosphorus-deficient and phosphorus-rich tropical soils [74]. Similarly, a total of 250 Gbp of sequenced nucleotides were recorded when reconstructing the microbial metabolic network in a host geological nuclear waste repository [4]. For the present study, we assumed that the depth of metaproteomics achieved with a standard analysis (here, a 90-min nanoLC-MS/MS run) would be relatively limited, and that the metagenomics information obtained should be sufficient to effectively represent the most abundant microorganisms. If the objective was to analyze the whole 1-m soil core, the metagenomics efforts would have to be multiplied, along with the monetary costs of nucleic acid sequencing, to produce a database representative of the whole core. Sequencing depth directly influences the outcome of any attempt to assemble metagenomics data, but more importantly, the use of short-read next-generation sequencing combined with long-read technology should also be taken into account in such projects [23]. Once again, the corresponding costs will be the main factor driving the implementation of these combined sequencing technologies for soil analysis, but we are confident that a combined approach could boost metagenomics-based metaproteomics.

The metagenomics reads obtained in this study were treated either with MEGAHIT, FragGeneScan, or sixgill. Our results using the five constructed databases confirmed that a strategy with two query rounds, as recommended for unusually large databases [28], performs better than direct assignment, whatever the database used. However, at least in our set-up, an optimum should be considered to select the entries used to create the sub-database. Indeed, a proteomics approach was recently used to assess the quality of transcriptomics data and their assembly [14], and metaproteomics data could be used in a similar way to assess the quality of metagenomics data assemblies.

With the datasets considered here, the best PSM attribution rate was obtained for the FragGeneScan CDSs predicted directly from trimmed reads (20.5%). As expected, this attribution rate was lower than those obtained with similar instruments when studying single organisms for which a well-annotated genome is available. For example, a rate of 61% PSM assignment was reported for the *Microbacterium oleivorans* A9 strain [18]. However, our rate it is quite similar to that reported for an animal proteogenomics study (21% [66];. The complexity of soil samples in terms of strains means that many possible peptide co-elutions and thus chimeric MS/MS spectra can be produced. We therefore expect that the rate of assignment would be further improved using higher-performance acquisition instruments.

The high quality of theoretical proteomes from isolates available in generic databases such as NCBInr and their large numbers advocate for use of these resources in metaproteomics interpretation pipelines. Indeed, the use of selected annotated genomes has previously been explored [13, 27, 50, 75], as has the use of the Uniref100 database [55] or the NCBInr database [30, 68, 72]. Here, two generic databases were assessed for their usefulness in interpreting the soil metaproteomics data: NCBInr and the SGC soil gene catalog. The two databases were complementary in terms of environmental sequence coverage, and the spectrum attribution rate of the combined database was 23.8%, which is higher than with a search against a sample-specific metagenomics database, but without the cost. Therefore, this strategy could be advantageous whenever numerous samples of diverse origins are to be analyzed.

Previous studies indicated that merging protein sequence databases from several samples might improve the peptide identification rate [59, 62]. Here, we combined metagenomics data analyzed with FragGeneScan, SGC, and generic database such as NCBI in a two-step search strategy. This approach produced the best assignment rate, with 26.2% of MS/MS spectra assigned. We therefore recommend this approach for use with other experimental metagenomics and metaproteomics datasets. Another previous study indicated that combining Uniprot with sample-specific metagenomics data could improve the number of peptides identified for samples from a biogas plant [25]. We found that the dedicated SGC database performed better than the generalist NCBInr database in the present study. Combining metagenomics sequencing data with data from a generic database could be performed while applying taxonomical constraints, as proposed previously [73]. However, this strategy is highly dependent on the presence of the identified organisms in the generic database and will consequently be sample-specific. Defining the optimal strategy in metaproteomics may depend on the research question to tackle as the objective may be either a focus on a few microorganisms with interesting metabolism, or the overall picture. In the first case, the design of a dedicated database emphasizing the genomes or metagenome-assembled genomes (MAGs) of interest may be well worth the effort required. In this vein, using the most abundant proteins identified by metaproteomics as guides to derive the taxonomic composition of the microbial community and expanding the search database with the genomes from the identified abundant species appears a promising two-stage strategy [57]. However, missing the identification of accessory proteins not present in the database could impact the understanding of the functionality of the microbial system. In the latter case, sequencing data allows MAGs binning, but a more globalized approach is often applied, either imposed by insufficient sequencing depth or preferred for speed, cost, sample, or resource availability. Taxonomical and functional assignation is then often performed at family or phylum levels using peptides, proteins, reads, genes, contigs, or scaffolds taxonomical and functional mapping. In that case, the assessment of metaproteomic databases can be performed using the PSM attribution yield.

Two significant criteria to consider when assessing the power of metaproteomics is how many of the peptides/proteins identified have taxonomy- and function-derived annotations. In metaproteomics, the taxonomical annotation is commonly performed with taxon-specific peptides using the lowest common ancestor approach, such as with the Unipept tool [21]. However, functional annotation works best at the protein level for metaproteomics, as shown here. The length of the sequences used to find a GO or KO has an impact on the percentage of PSMs functionally annotated. As shown here, peptide level functional annotation is improved using a sequence-based search for functional homologs at protein level, which both allows to annotate peptides missing in large protein databases (e.g., NCBI, Uniprot) and to enlarge the pool of proteins functionally associated with a given peptide, and thus the probability to gather GO or KEGG annotated proteins. Here, we found the optimal strategy in terms of both MS/MS attribution ratio and functional annotation ratio to be a combination of FGS, SGC, and NCBInr databases with 26.2% and 20.0% respectively. Combining SGC and NCBI databases results in a MS/MS attribution ratio of 23.8% and a functional annotation ratio of 19.7%. Therefore, this later strategy represents an interesting alternative for soil samples in the absence of sample-specific metagenomics sequencing data.

Conclusions

In conclusion, combining sample-specific metagenomics data and generic databases in a two-step database search performed best for the soil sample analyzed in the present study, both in terms of ratio of assigned spectra and retrieval of function-derived information. Amalgaming a massive soil gene catalog and the generalist NCBInr database resulted in almost the same outcome. This result opens up broad prospects for the application of metaproteomics to soil samples, which includes a highly challenging matrix, as well as broad microbial diversity, and extensive complexity.

Materials and methods

Soil material

A soil core was sampled on May 23 2018 from a floodplain at Bouafles near the Seine River (France). The site has already been well characterized in terms of sedimentation and chemicals [2, 3, 37, 41]. The section of the core between 17 and 28 cm depth was sliced into five layers. Two grams of each layer were pooled and homogenized for DNA extraction. The mid-layer (20–23 cm depth) was used for protein extraction.

DNA extraction from soil and sequencing

Soil DNA was extracted and sequenced by GenoScreen (Lille, France) from 1 g of lyophilized sample using an optimized protocol [65]. Briefly, soil was mixed with 100 mM Tris-HCl (pH 8), 100 mM EDTA (pH 8), 100 mM NaCl, 2% (w/v) polyvinylpyrrolidone (40 g/mol), and 2% (w/v) sodium dodecyl sulfate and subjected to bead-beating. DNA was precipitated with isopropanol, washed with 70% ethanol, and further purified using the MP Biomedials GeneClean Turbo kit (Fisher scientific).

DNA libraries were constructed with the Nextera XT DNA Library Preparation kit (Illumina) and sequenced on a HiSeq 4000 Illumina run in 2 × 150 bp. Raw reads have been deposited in the Sequence Read Archive under dataset identifier SRX8818139, as part of Bioproject PRJNA648365. Reads were analyzed using the phylogenetic MG-RAST pipeline [44].

Metagenomics analysis

Paired-reads were processed using the MEGAHIT workflow into the ASaiM Galaxy framework [8]. They were quality controlled and trimmed using FastQC and Trim Galore v0.4.3.1 with a Phred quality score cutoff of 20. MEGAHIT v1.1.2 [38] was used to assemble trimmed paired-reads into contigs with default parameters with a minimum kmer size of 21, maximum kmer size of 141, k-step of 12, and merge complex bubbles with length up to 20,098. The estimation of the assembly quality statistics was done with MetaQUAST [45] and the identification of potential assembly error signature with VALET. The percentage of unmapped reads were determined with Bowtie2 [36] and combined with MultiQC [16]. The MGH-6RF database was obtained by six-frame translation of the assembly, retaining only tryptic peptide sequences composed of at least five residues. PLASS [59, 60] was used with default parameters. Sixgill v0.2.4 [42] was used with the following parameters: minlength 10, minqualscore 30, minorflength 40, minlongesttryppeplen 7, and minreadcount 2. The paired-reads were processed with WHORMSS (Genoscreen) workflow consisting in demultiplexing and removing indexes in reads. The reads were trimmed and a Phred quality score cutoff of 30 was applied. The reads with a length lower than 75 bases were removed. Paired-reads were reassembled and low complexity sequences were removed as well as various contaminants including *Homo sapiens* sequences. FragGeneScan v1.3 [54] was applied with Illumina sequencing reads with about 0.01% error rate model to construct the FGS database.

Soil gene catalog and NCBInr databases

The soil gene catalog [5] was downloaded from http://vm-lux.embl.de/~hildebra/Soil_gene_cat/ (accessed on 22 March 2021). NCBnr was downloaded from https://www.ncbi.nlm.nih.gov/ on 3 January 2018).

Protein extraction and proteolysis

The proteins from 5 g of soil were extracted using the NoviPure Soil Protein Extraction Kit (Mo-Bio) as recommended by the supplier. After centrifugation, proteins from the 10-ml supernatant were precipitated by adding 2.5 ml trichloroacetic acid (50% w/v). Proteins were collected by centrifugation for 10 min at 6000×*g*. The resulting pellet was resuspended in 40 µL LDS 1X (Invitrogen) containing 5% beta-mercaptoethanol, sonicated for 5 min in an ultra sound bath and then heated to 99 °C for 5 min. Soluble proteins (25 µL per well) were subjected to SDS-PAGE gel electrophoresis on NuPAGE 4–12% Bis-Tris gel (Invitrogen) for 5 min at 200 V in MES/SDS 1X running buffer (Invitrogen). Proteins were stained for 15 min with Coomassie SimplyBlue SafeStain (Thermo Fisher Scientific), and then in-gel proteolyzed with trypsin gold (Promega) for 1 h at 50 °C, as recommended [22].

NanoLC-MS/MS and interpretation

Peptides were analyzed on a Q-Exactive HF mass spectrometer (Thermo) coupled to an Ultimate 3000 nano LC system (Thermo), as described previously [33]. Tryptic peptides (8 µl) were desalted on a reverse-phase PepMap 100 C18 µ-precolumn (5 µm, 100 Å, 300 µm i.d. × 5 mm, Thermo) before separating peptides on a nanoscale PepMap 100 C18 nanoLC column (3 µm, 100 Å, 75 µm i.d. × 50 cm, Thermo) at a flow rate of 0.2 µL min^{-1} using a 90-min gradient of mobile phase A (0.1% HCOOH/100% H_2O) and phase B (0.1% HCOOH/80% CH3CN). The gradient used was developed from 4 to 25% B in 70 min and then from 25 to 40% B in 20 min. The mass spectrometer was operated in Top20 data-dependent acquisition mode. Full MS scans were acquired from 350 to 1800 *m/z* at a resolution of 60,000 and the 20 most abundant precursor ions were sequentially selected for fragmentation with a dynamic exclusion time of 10 s. The resolution for the fragment scans was 15,000. Only ions with 2 or 3 positive charges were selected for fragmentation. MS/MS spectra were interpreted using Mascot Daemon software (version 2.6.1; Matrix Science) indicating 5-ppm tolerance for the parent ion and 0.02-Da tolerance for secondary fragments, 2+ and 3+ as possible peptide charges, a maximum of two missed cleavages, carbamidomethylation of cysteine as fixed modification, oxidation of methionine as variable modification, and trypsin as proteolytic enzyme. The FDR threshold was set at 0.01 using a decoy-free FDR method based on a mixture-model of four beta distributions which has been shown well adapted for handling large proteogenomics and metaproteomics datasets and databases [52]. The two-step database search strategy was initiated using several Mascot *p*-value thresholds (0.01, 0.05, 0.10, 0.20, 0.30, 0.50, 0.70, 0.80, 0.90, 0.99) for the first search round to select the protein sequences. The most time-consuming search (13 h) was noted for the first step NCBI database interrogation.

Functional and taxonomic annotation and gene ontology

Functional annotation of identified proteins was based on sequence similarity searches carried out with Diamond BLASTP (v0.8.22.84) [10] against the Uniref50

[46] database (release August 24, 2018). The following parameters were applied: top five hits, *e*-value threshold 10, and percentage identity above 50%. The GOSlim terms (release January 30, 2017) associated with the Uniprot accession number were retrieved for each protein. KEGG annotation was performed using the GhostKOALA [31] web server. Peptides identified at FDR 1% were functionally annotated using the Unipept [43] desktop application version 1.2.1, activating the "equate I and L" and "advanced missing cleavage handling" options. Unipept peptide taxonomical information was used to calculate kingdom and phylum abundances.

Abbreviations

NCBInr: National Center for Biotechnology Information non-redundant; PSMs: Peptide-to-spectrum matches; FDR: False discovery rate; MAGs: Metagenome-assembled genomes; CDS: Protein-coding sequence; MF: Molecular function; BP: Biological process; GO: Gene Ontology; EC: Enzyme commission; KO: KEGG Orthology

Acknowledgements

Not applicable.

Authors' contributions

OP, SA, and JA conceived the project. VJ, OP, and JA designed the overall experimental approach and wrote the manuscript. VJ and OP analyzed the data. VJ created the databases. VJ and KC contributed post-processing of database searches under the supervision of OP. GM performed the protein extraction and tandem mass spectrometry measurements under the supervision of OP and JA. The authors read and approved the final manuscript.

Funding

This work was funded in part by the DRF impulsion program from CEA (Geomics project).

Availability of data and materials

The mass spectrometry proteomics data have been submitted to the ProteomeXchange Consortium via the PRIDE partner repository under dataset identifier PXD026798 and project DOI https://doi.org/10.6019/PXD026798.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Université Paris-Saclay, CEA, INRAE, Département Médicaments et Technologies pour la Santé (DMTS), SPI, F-30200 Bagnols-sur-Cèze, France. [2]Laboratoire des Sciences et de l'Environnement (LSCE-IPSL), UMR 8212 (CEA/CNRS/UVSQ), CEA Saclay, Université Paris-Saclay, Orme des Merisiers, F-91191 Gif-sur-Yvette, France. [3]Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Université de Montpellier, F-30207 Bagnols-sur-Cèze, France.

Received: 5 July 2021 Accepted: 29 July 2021
Published online: 29 September 2021

References

1. Ayrault, S., M. Meybeck, J.-M. Mouchel, J. Gaspéri, L. Lestel, C. Lorgeoux and D. Boust (2019). Sedimentary archives reveal the concealed history of micropollutant contamination in the Seine River basin. Berlin, Heidelberg, Springer Berlin Heidelberg: 1-32.
2. Ayrault S, Priadi CR, Evrard O, Lefevre I, Bonte P. Silver and thallium historical trends in the Seine River basin. J Environ Monit. 2010;12(11):2177–85. https://doi.org/10.1039/c0em00153h.
3. Ayrault S, Roy-Barman M, Le Cloarec MF, Priadi CR, Bonte P, Gopel C. Lead contamination of the Seine River, France: geochemical implications of a historical perspective. Chemosphere. 2012;87(8):902–10. https://doi.org/10.1016/j.chemosphere.2012.01.043.
4. Bagnoud A, Chourey K, Hettich RL, de Bruijn I, Andersson AF, Leupin OX, et al. Reconstructing a hydrogen-driven microbial metabolic network in Opalinus Clay rock. Nat Commun. 2016;7(1):12770. https://doi.org/10.1038/ncomms12770.
5. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. Nature. 2018;560(7717):233–7. https://doi.org/10.1038/s41586-018-0386-6.
6. Bastida F, Jehmlich N, Martínez-Navarro J, Bayona V, García C, Moreno JL. The effects of struvite and sewage sludge on plant yield and the microbial community of a semiarid Mediterranean soil. Geoderma. 2019;337:1051–7. https://doi.org/10.1016/j.geoderma.2018.10.046.
7. Bastida F, Torres IF, Moreno JL, Baldrian P, Ondono S, Ruiz-Navarro A, et al. The active microbial diversity drives ecosystem multifunctionality and is physiologically related to carbon availability in Mediterranean semi-arid soils. Mol Ecol. 2016;25(18):4660–73. https://doi.org/10.1111/mec.13783.
8. Batut B, Gravouil K, Defois C, Hiltemann S, Brugere JF, Peyretaillade E, et al. ASaiM: a Galaxy-based framework to analyze microbiota data. Gigascience. 2018;7(6). https://doi.org/10.1093/gigascience/giy057.
9. Becher D, Bernhardt J, Fuchs S, Riedel K. Metaproteomics to unravel major microbial players in leaf litter and soil environments: challenges and perspectives. Proteomics. 2013;13(18-19):2895–909. https://doi.org/10.1002/pmic.201300095.
10. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60. https://doi.org/10.1038/nmeth.3176.
11. Cernava T, Erlacher A, Aschenbrenner IA, Krug L, Lassek C, Riedel K, et al. Deciphering functional diversification within the lichen microbiota by metaomics. Microbiome. 2017;5(1):82. https://doi.org/10.1186/s40168-017-0303-5.
12. Cheng K, Ning Z, Zhang X, Li L, Liao B, Mayne J, et al. MetaLab: an automated pipeline for metaproteomic data analysis. Microbiome. 2017;5(1):157. https://doi.org/10.1186/s40168-017-0375-2.
13. Chourey K, Nissen S, Vishnivetskaya T, Shah M, Pfiffner S, Hettich RL, et al. Environmental proteomics reveals early microbial community responses to biostimulation at a uranium- and nitrate-contaminated site. Proteomics. 2013;13(18-19):2921–30. https://doi.org/10.1002/pmic.201300155.
14. Cogne Y, Gouveia D, Chaumot A, Degli-Esposti D, Geffard O, Pible O, et al. Proteogenomics-guided evaluation of RNA-Seq assembly and protein database construction for emergent model organisms. Proteomics. 2020;20(10):e1900261. https://doi.org/10.1002/pmic.201900261.
15. Coute Y, Bruley C, Burger T. Beyond target-decoy competition: stable validation of peptide and protein identifications in mass spectrometry-based discovery proteomics. Anal Chem. 2020;92(22):14898–906. https://doi.org/10.1021/acs.analchem.0c00328.
16. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8. https://doi.org/10.1093/bioinformatics/btw354.
17. Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat Rev Microbiol. 2017;15(10):579–90. https://doi.org/10.1038/nrmicro.2017.87.
18. Gallois N, Alpha-Bazin B, Ortet P, Barakat M, Piette L, Long J, et al. Proteogenomic insights into uranium tolerance of a Chernobyl's microbacterium bacterial isolate. J Proteomics. 2018;177:148–57. https://doi.org/10.1016/j.jprot.2017.11.021.
19. Glass JB, Yu H, Steele JA, Dawson KS, Sun S, Chourey K, et al. Geochemical, metagenomic and metaproteomic insights into trace metal utilization by methane-oxidizing microbial consortia in sulphidic marine sediments. Environ Microbiol. 2014;16(6):1592–611. https://doi.org/10.1111/1462-2920.12314.
20. Gouveia D, Pible O, Culotta K, Jouffret V, Geffard O, Chaumot A, et al. Combining proteogenomics and metaproteomics for deep taxonomic and functional characterization of microbiomes from a non-sequenced host. NPJ Biofilms Microbiomes. 2020;6(1):23. https://doi.org/10.1038/s41522-020-0133-2.

21. Gurdeep Singh R, Tanca A, Palomba A, Van der Jeugt F, Verschaffelt P, Uzzau S, et al. Unipept 4.0: functional analysis of metaproteome data. J Proteome Res. 2019;18(2):606–15. https://doi.org/10.1021/acs.jproteome.8b00716.
22. Hartmann EM, Allain F, Gaillard JC, Pible O, Armengaud J. Taking the shortcut for high-throughput shotgun proteomic analysis of bacteria. Methods Mol Biol. 2014;1197:275–85. https://doi.org/10.1007/978-1-4939-1261-2_16.
23. Henson J, Tischler G, Ning Z. Next-generation sequencing and large genome assemblies. Pharmacogenomics. 2012;13(8):901–15. https://doi.org/10.2217/pgs.12.72.
24. Heyer R, Benndorf D, Kohrs F, De Vrieze J, Boon N, Hoffmann M, et al. Proteotyping of biogas plant microbiomes separates biogas plants according to process temperature and reactor type. Biotechnol Biofuels. 2016;9(1):155. https://doi.org/10.1186/s13068-016-0572-4.
25. Heyer R, Schallert K, Zoun R, Becher B, Saake G, Benndorf D. Challenges and perspectives of metaproteomic data analysis. J Biotechnol. 2017;261:24–36. https://doi.org/10.1016/j.jbiotec.2017.06.1201.
26. Hubler SL, Kumar P, Mehta S, Easterly C, Johnson JE, Jagtap PD, et al. Challenges in peptide-spectrum matching: a robust and reproducible statistical framework for removing low-accuracy, high-scoring hits. J Proteome Res. 2020;19(1):161–73. https://doi.org/10.1021/acs.jproteome.9b00478.
27. Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, Blazewicz SJ, et al. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature. 2015;521(7551):208–12. https://doi.org/10.1038/nature14238.
28. Jagtap P, Goslinga J, Kooren JA, McGowan T, Wroblewski MS, Seymour SL, et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics. 2013;13(8):1352–7. https://doi.org/10.1002/pmic.201200352.
29. Jansson JK, Hofmockel KS. Soil microbiomes and climate change. Nat Rev Microbiol. 2020;18(1):35–46. https://doi.org/10.1038/s41579-019-0265-7.
30. Johnson-Rollings AS, Wright H, Masciandaro G, Macci C, Doni S, Calvo-Bado LA, et al. Exploring the functional soil-microbe interface and exoenzymes through soil metaexoproteomics. ISME J. 2014;8(10):2148–50. https://doi.org/10.1038/ismej.2014.130.
31. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol. 2016;428(4):726–31. https://doi.org/10.1016/j.jmb.2015.11.006.
32. Keiblinger KM, Wilhartitz IC, Schneider T, Roschitzki B, Schmid E, Eberl L, et al. Soil metaproteomics - comparative evaluation of protein extraction protocols. Soil Biol Biochem. 2012;54(15-10):14–24. https://doi.org/10.1016/j.soilbio.2012.05.014.
33. Klein G, Mathe C, Biola-Clier M, Devineau S, Drouineau E, Hatem E, et al. RNA-binding proteins are a major target of silica nanoparticles in cell extracts. Nanotoxicology. 2016;10(10):1555–64. https://doi.org/10.1080/17435390.2016.1244299.
34. Kleiner, M. (2019). "Metaproteomics: much more than measuring gene expression in microbial communities." mSystems 4(3).
35. Kroeger ME, Delmont TO, Eren AM, Meyer KM, Guo J, Khan K, et al. New biological insights into how deforestation in Amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. Front Microbiol. 2018;9:1635. https://doi.org/10.3389/fmicb.2018.01635.
36. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
37. Le Cloarec MF, Bonte PH, Lestel L, Lefèvre I, Ayrault S. Sedimentary record of metal contamination in the Seine River during the last century. Physics and Chemistry of the Earth, Parts A/B/C. 2011;36(12):515–29. https://doi.org/10.1016/j.pce.2009.02.003.
38. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102:3–11. https://doi.org/10.1016/j.ymeth.2016.02.020.
39. Lin W, Wu L, Lin S, Zhang A, Zhou M, Lin R, et al. Metaproteomic analysis of ratoon sugarcane rhizospheric soil. BMC Microbiol. 2013;13(1):135. https://doi.org/10.1186/1471-2180-13-135.
40. Liu D, Keiblinger KM, Leitner S, Wegner U, Zimmermann M, Fuchs S, et al. Response of microbial communities and their metabolic functions to drying(-)rewetting stress in a temperate forest soil. Microorganisms. 2019;7(5). https://doi.org/10.3390/microorganisms7050129.
41. Lorgeoux C, Moilleron R, Gasperi J, Ayrault S, Bonte P, Lefevre I, et al. Temporal trends of persistent organic pollutants in dated sediment cores: chemical fingerprinting of the anthropogenic impacts in the Seine River basin, Paris. Sci Total Environ. 2016;541:1355–63. https://doi.org/10.1016/j.scitotenv.2015.09.147.
42. May DH, Timmins-Schiffman E, Mikan MP, Harvey HR, Borenstein E, Nunn BL, et al. An alignment-free "metapeptide" strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. J Proteome Res. 2016;15(8):2697–705. https://doi.org/10.1021/acs.jproteome.6b00239.
43. Mesuere B, Debyser G, Aerts M, Devreese B, Vandamme P, Dawyndt P. The Unipept metaproteomics analysis pipeline. Proteomics. 2015;15(8):1437–42. https://doi.org/10.1002/pmic.201400361.
44. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008;9(1):386. https://doi.org/10.1186/1471-2105-9-386.
45. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics. 2016;32(7):1088–90. https://doi.org/10.1093/bioinformatics/btv697.
46. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017;45(D1):D170–6. https://doi.org/10.1093/nar/gkw1081.
47. Murray, A. E., J. Freudenstein, S. Gribaldo, R. Hatzenpichler, P. Hugenholtz, P. Kampfer, K. T. Konstantinidis, C. E. Lane, R. T. Papke, D. H. Parks, R. Rossello-Mora, M. B. Stott, I. C. Sutcliffe, J. C. Thrash, S. N. Venter, W. B. Whitman, S. G. Acinas, R. I. Amann, K. Anantharaman, J. Armengaud, B. J. Baker, R. A. Barco, H. B. Bode, E. S. Boyd, C. L. Brady, P. Carini, P. S. G. Chain, D. R. Colman, K. M. DeAngelis, M. A. de Los Rios, P. Estrada-de Los Santos, C. A. Dunlap, J. A. Eisen, D. Emerson, T. J. G. Ettema, D. Eveillard, P. R. Girguis, U. Hentschel, J. T. Hollibaugh, L. A. Hug, W. P. Inskeep, E. P. Ivanova, H. P. Klenk, W. J. Li, K. G. Lloyd, F. E. Loffler, T. P. Makhalanyane, D. P. Moser, T. Nunoura, M. Palmer, V. Parro, C. Pedros-Alio, A. J. Probst, T. H. M. Smits, A. D. Steen, E. T. Steenkamp, A. Spang, F. J. Stewart, J. M. Tiedje, P. Vandamme, M. Wagner, F. P. Wang, P. Yarza, B. P. Hedlund and A. L. Reysenbach (2020). "Roadmap for naming uncultivated Archaea and bacteria." Nat Microbiol 5(8): 987-994, DOI: https://doi.org/10.1038/s41564-020-0733-x.
48. Muth T, Kolmeder CA, Salojärvi J, keskitalo S, Varjosalo M, Verdam FJ, Rensen SS, Reichl U, de Vos WM, Rapp E, Martens L. "Navigating through metaproteomics data: a logbook of database searching." Proteomics. 2015;15(20):3439–53.
49. Muth T, Renard BY, Martens L. Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. Expert Rev Proteomics. 2016;13(8):757–69. https://doi.org/10.1080/14789450.2016.1209418.
50. Orellana LH, Hatt JK, Iyer R, Chourey K, Hettich RL, Spain JC, et al. Comparing DNA, RNA and protein levels for measuring microbial dynamics in soil microcosms amended with nitrogen fertilizer. Sci Rep. 2019;9(1):17630. https://doi.org/10.1038/s41598-019-53679-0.
51. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2(11):1533–42. https://doi.org/10.1038/s41564-017-0012-7.
52. Pible O, Allain F, Jouffret V, Culotta K, Miotello G, Armengaud J. Estimating relative biomasses of organisms in microbiota using "phylopeptidomics". Microbiome. 2020;8(1):30. https://doi.org/10.1186/s40168-020-00797-x.
53. Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake RC 2nd, et al. Community proteomics of a natural microbial biofilm. Science. 2005;308(5730):1915–20. https://doi.org/10.1126/science.1109070.
54. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38(20):e191. https://doi.org/10.1093/nar/gkq747.
55. Schneider T, Keiblinger KM, Schmid E, Sterflinger-Gleixner K, Ellersdorfer G, Roschitzki B, et al. Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. ISME J. 2012;6(9):1749–62. https://doi.org/10.1038/ismej.2012.11.
56. Seifert J, Muth T. Editorial for special issue: metaproteomics. Proteomes. 2019;7(1). https://doi.org/10.3390/proteomes7010009.
57. Stamboulian M, Li S, Ye Y. Using high-abundance proteins as guides for fast and effective peptide/protein identification from human gut metaproteomic data. Microbiome. 2021;9(1):80. https://doi.org/10.1186/s40168-021-01035-8.
58. Starke R, Jehmlich N, Bastida F. Using proteins to study how microbes contribute to soil ecosystem services: the current state and future

perspectives of soil metaproteomics. J Proteomics. 2019;198:50–8. https://doi.org/10.1016/j.jprot.2018.11.011.
59. Starr AE, Deeke SA, Li L, Zhang X, Daoud R, Ryan J, et al. Proteomic and metaproteomic approaches to understand host-microbe interactions. Anal Chem. 2018;90(1):86–109. https://doi.org/10.1021/acs.analchem.7b04340.
60. Steinegger M, Mirdita M, Soding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods. 2019;16(7):603–6. https://doi.org/10.1038/s41592-019-0437-4.
61. Tanca A, Palomba A, Deligios M, Cubeddu T, Fraumene C, Biosa G, et al. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. PLoS One. 2013;8(12):e82981. https://doi.org/10.1371/journal.pone.0082981.
62. Tanca A, Palomba A, Fraumene C, Pagnozzi D, Manghina V, Deligios M, et al. The impact of sequence database choice on metaproteomic results in gut microbiota studies. Microbiome. 2016;4(1):51. https://doi.org/10.1186/s40168-016-0196-8.
63. Tartaglia M, Bastida F, Sciarrillo R, Guarino C. Soil metaproteomics for the study of the relationships between microorganisms and plants: a review of extraction protocols and ecological insights. Int J Mol Sci. 2020;21(22). https://doi.org/10.3390/ijms21228455.
64. Taubert M, Grob C, Crombie A, Howat AM, Burns OJ, Weber M, et al. Communal metabolism by methylococcaceae and methylophilaceae is driving rapid aerobic methane oxidation in sediments of a shallow seep near Elba, Italy. Environ Microbiol. 2019;21(10):3780–95. https://doi.org/10.1111/1462-2920.14728.
65. Terrat S, Christen R, Dequiedt S, Lelievre M, Nowak V, Regnier T, et al. Molecular biomass and MetaTaxogenomic assessment of soil microbial communities as influenced by soil DNA extraction procedure. Microb Biotechnol. 2012;5(1):135–41. https://doi.org/10.1111/j.1751-7915.2011.00307.x.
66. Trapp J, Almunia C, Gaillard JC, Pible O, Chaumot A, Geffard O, et al. Proteogenomic insights into the core-proteome of female reproductive tissues from crustacean amphipods. J Proteomics. 2016;135:51–61. https://doi.org/10.1016/j.jprot.2015.06.017.
67. Van Den Bossche T, Kunath B, Schallert K, Schäpe S Abraham P, Armengaud J, Arntzen M, Bassignanin A, Benndorf D, Fuchs S, et al. "Critical Assessment of Metaproteome Investigation (CAMPI): A Multi-Lab Comparison of Established Workflows." BioRxiV. 2021. https://doi.org/10.1101/2021.03.05.433915.
68. Wang HB, Zhang ZX, Li H, He HB, Fang CX, Zhang AJ, et al. Characterization of metaproteomics in crop rhizospheric soil. J Proteome Res. 2011;10(3):932–40. https://doi.org/10.1021/pr100981r.
69. Wang Z, Wang Y, Fuhrman JA, Sun F, Zhu S. Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences. Brief Bioinform. 2020;21(3):777–90. https://doi.org/10.1093/bib/bbz025.
70. Wilmes P, Heintz-Buschart A, Bond PL. A decade of metaproteomics: where we stand and what the future holds. Proteomics. 2015;15(20):3409–17. https://doi.org/10.1002/pmic.201500183.
71. Wilpiszeski RL, Aufrecht JA, Retterer ST, Sullivan MB, Graham DE, Pierce EM, et al. Soil aggregate microbial communities: towards understanding microbiome interactions at biologically relevant scales. Appl Environ Microbiol. 2019;85(14). https://doi.org/10.1128/AEM.00324-19.
72. Wu L, Wang H, Zhang Z, Lin R, Zhang Z, Lin W. Comparative metaproteomic analysis on consecutively *Rehmannia glutinosa*-monocultured rhizosphere soil. PLoS One. 2011;6(5):e20611. https://doi.org/10.1371/journal.pone.0020611.
73. Xiao J, Tanca A, Jia B, Yang R, Wang B, Zhang Y, et al. Metagenomic taxonomy-guided database-searching strategy for improving metaproteomic analysis. J Proteome Res. 2018;17(4):1596–605. https://doi.org/10.1021/acs.jproteome.7b00894.
74. Yao Q, Li Z, Song Y, Wright SJ, Guo X, Tringe SG, et al. Community proteogenomics reveals the systemic impact of phosphorus availability on microbial functions in tropical soil. Nat Ecol Evol. 2018;2(3):499–509. https://doi.org/10.1038/s41559-017-0463-5.
75. Zampieri E, Chiapello M, Daghino S, Bonfante P, Mello A. Soil metaproteomics reveals an inter-kingdom stress response to the presence of black truffles. Sci Rep. 2016;6(1):25773. https://doi.org/10.1038/srep25773.

Publisher's Note

Chapitre 4 : Structure des communautés microbiennes d'une archive sédimentaire évaluée par métaprotéomique

Contexte :

Après un premier axe de travail porté sur l'amélioration de l'interprétation des données de métaprotéomique focalisé sur une seule couche de la colonne de sédiment, nous nous sommes intéressés à l'analyse de la colonne entière. Cette colonne a été divisée en 35 couches qui ont été datées et dont les concentrations en contaminants ont été mesurées systématiquement. Ces mesures de contaminants et notamment des éléments traces métalliques permettent de retracer l'histoire des contaminants au sein de cette carotte alimentée par les sédiments de crues successives de la Seine. Ces contaminants ont-ils un impact sur les communautés microbiennes présentes dans cette carotte de sédiments ? Pour répondre à cette question, une approche multi-omiques a été choisie nécessitant des moyens importants et du temps, mais résultant en un jeu de données très riche en information. Ce projet n'a pu être conduit que grâce aux efforts des deux laboratoires impliqués.

L'analyse de la communauté microbienne a nécessité des ajustements de la stratégie d'interprétation de métaprotéomique afin de pouvoir tirer pleinement profit des données métagénomiques et séquençage des amplicons 16S rRNA. L'analyse s'est volontairement focalisée à un niveau taxonomique élevé « intégratif », le rang phylum, car la profondeur de l'analyse métaprotéomique de chaque couche ne peut être très exhaustive pour les raisons de couts. A des niveaux inférieurs en termes de rang taxonomique, aucune espèce ou genre n'est suffisamment abondant de par lui-même pour pouvoir être qualifié statistiquement comme différentiellement présent entre les couches. Ces analyses ont produit de grandes matrices éparses que ce soit au niveaux taxonomiques ou fonctionnelles. L'objectif a été de décrire des tendances entre la présence de micro-organismes et la présence d'éléments traces métalliques. Ce focus particulier sur les interactions entre les microorganismes et les métaux a été contraint par les mesures géochimiques précises qui n'ont été possible que sur les métaux, les autres composés d'intérêt dans ces couches (antibiotiques par exemple) n'ayant pas été dosé mais éventuellement dérivables de données acquises antérieurement sur des carottes similaires et datées elles aussi. Les hypothèses qui en découlent pourront être par la suite testées en laboratoire.

Vertical structure of microbial communities of a sediment archive of past contaminations assessed by metaproteomics

Virginie Jouffret[1,2,3], Olivier Pible[1], Guylaine Miotello[1],

Sophie Ayrault[2], Jean Armengaud[1#]

[1]Université Paris-Saclay, CEA, INRAE, Département Médicaments et Technologies pour la Santé (DMTS), SPI, F-30200 Bagnols-sur-Cèze, France.

[2]Laboratoire des Sciences et de l'Environnement (LSCE-IPSL), UMR 8212 (CEA/CNRS/UVSQ), Université Paris-Saclay, CEA Saclay, Orme des Merisiers, F-91191 Gif-sur-Yvette, France.

[3]Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Université de Montpellier, F-30207 Bagnols-sur-Cèze, France.

[#]Correspondence: jean.armengaud@cea.fr; Tel.: +33 4 66 79 62 77

Running title: Sediment archive metaproteomics

Word count: 8162 *Character count: 47608*

Keywords: metaproteomics, metallomics, geochemistry, archive, sediments, soil

ABSTRACT

Floodplain soils, formed from sediments deposited during each flood, drain the alluvium of the watershed with which contaminants have a strong affinity. These soils are the witnesses of anthropic activities and archives of past contaminations. The Seine River collects and carries numerous contaminants of anthropogenic origin from a heavily industrialized basin. Here, we analyzed a Seine sediment core sampled downstream of the Paris conurbation which was subdivided into 35 dated layers. We measured for each of these layers the concentrations of 26 trace metals and established the vertical structure of the microbial communities with metataxonomics, metagenomics and metaproteomics data. We conceived a new pipeline for comparing the metaproteomics results of all the sediment layers on the very same basis revealing the presence of 52 classes, 85 orders, and 115 families of organisms. These taxonomical and functional results obtained through the peptides identified by tandem mass spectrometry were correlated with the geochemical elements deposited over the years. Alphaproteobacteria and Actinobacteria are the dominant phyla in the core. By clustered phyla abundances and geochemistry element concentrations, the Oomycetes, Candidatus_Lokiarchaeota, and Candidatus Eisenbacteria phyla are correlated with peak concentrations of Cd, Ag and Cr. The strategies and computer tools for data interpretation on a well-characterized core at the geochemical level open the way to an application in rapid diagnosis of microbial communities in soils.

INTRODUCTION

The soil microbiota is considered extremely diverse, replete with bacteria, archaea, yeasts, fungi, protists, and viruses, and performing vital functions such as conducting the turnover of organic matter and favoring plant growth and fertility. It has received a great deal of attention in recent years as soil microbiota could be a game-changer in restoring degraded land, improving the quality and sustainable yields of agriculture, but also regulating the efflux of carbon dioxide to the atmosphere. It is a key parameter for understanding and modeling climate change and predicting the effects of global warming. But due to its high complexity, results in space and time are still scarce and far from sufficiently comprehensive. The soil microbiome depends on biotic and abiotic factors, such as soil structure, moisture, pH, aeration, temperature, as well as vegetation type (Fierer and Jackson, 2006; Zhou et al., 2016; Bru et al., 2011). Unifying patterns of the biomass, diversity, and composition of some soil microbial players across the globe have been recently discovered (Crowther et al, 2019). A global atlas of the dominant bacteria found in many soils on all continents indicated that only a limited set of bacterial phylotypes accounted for nearly half of the soil microbiome. These phyla include Alphaproteobacteria, Betaproteobacteria, Actinobacteria, Acidobacter and Planctomycetes (Delgado-Baquerizo et al., 2018).

Over the past decade, many studies have described how soil microbial communities respond to environmental changes. For example, its dynamics have been probed over time and found significantly influenced by season (Lohmann et al., 2020). The impact of drought on soil microbial biomass has also been documented (Bastida et al., 2017), as well as the change in microbial decomposition of soil organic carbon in response to fresh carbon inputs, the so –called soil priming (Bastida et al., 2019). The relationships between soil microbial biodiversity and microbial biomass have been shown to be determined by soil carbon content (Bastida et al, 2021). A high concentration of heavy-metals strongly influences the soil microbiome (Chen et al., 2018).

While 16S rRNA gene amplicon sequencing or shotgun metagenomics are the most widely used methodologies to assess the structure of soil microbial communities, they are of limited relevance for explaining the functioning of these complex biological systems and delineating how their components interact. Metaproteomics has been shown to uniquely describe this functioning by identifying and quantifying the proteins that are the real workhorses of the system (Van den Bossche et al. 2021). However, its application to soil remained till recently a challenge (Abiraami et al., 2019; Starke et al., 2019), but new computer and database developments have greatly enhanced the value of interpretation (Jouffret et al., 2021). Several examples are worth citing to illustrate the interest of the approach. Recently, the proteins secreted by the most active microbial taxa in the rhizosphere of the oilseed rape have been identified (Lidbury et al., 2022), leading to a better understanding of the interactions between free-living microorganisms and plants which are inherently elusive and difficult to be evaluated while being crucial. The main microbial players of leaf-litter decomposition, which is a central process of the carbon cycle, have been studied by metaproteomics (Shneider et al., 2012), as well as the link between active microbial diversity and carbon availability in semi-arid soils (Bastida et al., 2016). The effects of toxic substances such as toluene on soil microbial communities have also been documented with this methodology (Williams et al., 2010).

The Paris conurbation, which has more than 10 million inhabitants, is among the most densely populated cities in the world. The Seine collects and carries numerous contaminants of anthropogenic origin, the entire Seine basin being very fertile and heavily industrialized. Downstream from Paris, the floodplains adjoining the river are regularly submerged during river floods. As a result, sediments are deposited there when the river flow decreases. These floodplains are prime sites for collecting sedimentary records, analyzing pollutants deposited over the years, and understanding the temporal

trajectory of contaminants. Bouafles is a well-documented site in terms of historical and geochemical context (Le Cloarec et al., 2011; Lorgeoux et al., 2016; Tamtam et al., 2011). In the present study, the vertical structure of the microbial communities of a sediment archive of past contaminations of Paris conurbation sampled in Bouafles in May 2018 was established by metaproteomics. We conceived a new pipeline for interpreting the microbiome results of tens of sediment layers and compare on the very same basis the layers. A specific focus on the taxa and functions correlated to the geochemical elements deposited over the years has been proposed.

RESULTS

Workflow to generate multi-omics and geochemical data along the strata of a sediment archive

A soil core was sampled from a floodplain, downstream from Paris (France) in the Seine basin, well characterized as a sediment archive of past contamination. The soil core was divided from top to the bottom into 35 layers of 3 cm depth. These samples were dated using Cs-137 radioactivity. The site was ploughed in 2016, homogenizing the first 21 cm of the core and affecting the last sediment deposits since 1987. The oldest sediment layer was dated from 1941. The high regularity of the sedimentation rate is shown in **Supplementary Figure 1** demonstrating the good linear correlation between the depth of the layer and the dating of the year obtained by Cs137 radioactivity. These measurements confirmed previous results reported for the same site which highlighted regular sedimentation of about 1.5 cm per year (e.g., Ayrault et al., 2012).

Figure 1 shows the workflow for generating multi-omics data on the microbiota present in the different layers of the core. Proteins from each of the 35 layers were extracted in triplicate. The 93 metaproteome samples were proteolyzed with trypsin and the resulting peptidomes were analyzed by tandem mass spectrometry. A dataset comprising 5,539,483 MS/MS spectra was generated. In parallel, the total DNA was extracted from five groups of layers representing the whole core but with a greater effort for the upper layers compared to the most buried layers because of their respective expected microbial diversity. A total of 365 million single Illumina reads were obtained when results from the five samples were merged. These data were exploited to generate five specific metagenomic database for the metaproteomics interpretation of each of the five layer groups. As shown in **Figure 1**, the interpretation of MS/MS data was carried out with a strategy to increase the sensitivity as much as possible and allow direct comparison of the different layers. Mass spectrometry results from triplicate analyses of each layer were merged. The MS/MS data of each layer was interpreted against the metagenome-derived database specific of the core soil group on the one hand, and against a generalist database including representatives of the taxa encompassed in NCBInr on the other. In the latter, the proteins of the identified genera of each layer were compiled in a layer-specific subdatabase and a second search identified the proteins of each layer. Then, the sequence of metagenomics and NCBI protein subsets were combined for each layer. This combined database was queried to obtain the protein subset at 1% FDR for each one of the 35 layers. The third stage consists of combining all the proteins identified in the 35 layers into a single database. This last common database was then interrogated at 1% FDR to obtain directly comparable metaproteomics results between layers. Thus, the cascade search comprised a total of 5 searches for each layer, so a consequent interpretation stage with a total of 155 searches and the use of 217 different databases.

A total of 617,878 MS/MS spectra could be assigned to peptide sequences, representing an assignment ratio of 12.4%. This ratio is relatively correct for soil metaproteomics dataset. The number of peptide-to-spectrum matches per peptide sequence is low. This is expected for such a highly diverse microbial community considering the mass spectrometry parameters chosen to increase as much as possible the peptidome coverage. The 428,708 identified peptide sequences were interpreted to obtain the structure of the microbial communities present in each layer, as well as an in-depth functional view of each metaproteome layer. Furthermore, 16S rRNA gene amplicon sequencing was performed on all five sample groups to confirm the metaproteomic-derived taxonomical results.

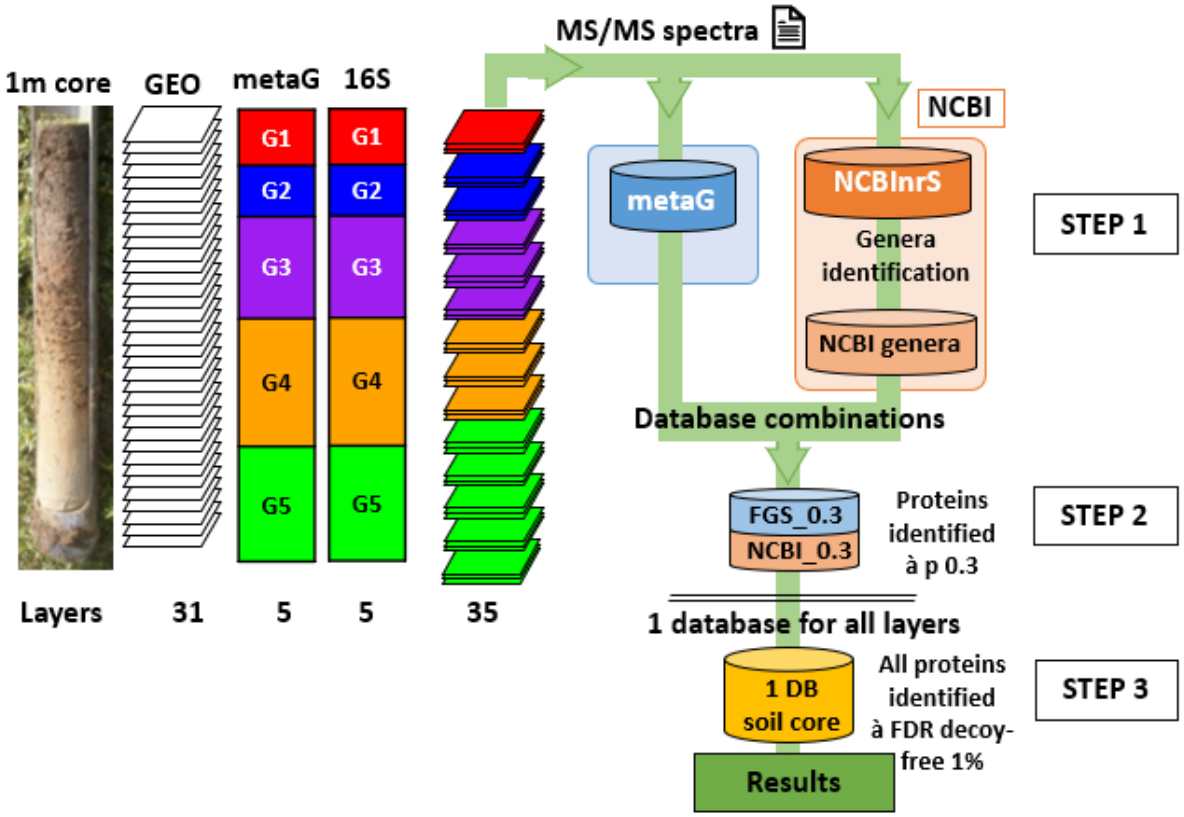


Figure 1: Workflow proposed for the analysis of the soil core from the Bouafles site, downstream of the Seine basin river. Metagenomics data were obtained by shotgun sequencing of total DNA extracted from five groups of soil layers. For each soil layer, MS/MS spectra and geochemical concentrations of elements were obtained. The multi-step process to refine and enhance the attribution rate of each of the 35 metaproteomes is schematized.

In parallel, the concentrations of 26 geochemical elements were measured in each of the layers, including major, minor and trace elements, some of them being essential elements: titanium (Ti), mercury (Hg), vanadium (V), chromium (Cr), manganese (Mn), cobalt (Co), nickel (Ni), copper (Cu), arsenic (As), rubidium (Rb), strontium (Sr), molybdenum (Mo), silver (Ag), cadmium (Cd), antimony (Sb), cesium (Cs), barium (Ba), thallium (Tl), and lead (Pb). In addition, the concentrations of additional elements such as phosphorus (P), magnesium (Mg), aluminum (Al), calcium (Ca), zinc (Zn), iron (Fe) and potassium (K) were evaluated.

Contamination by trace metals in the sediment core

We were particularly interested in monitoring metal contaminants over time by analyzing all the 31 layers. **Supplementary Figure 2** presents the corresponding results plotted along the core dating for each of the 26 geochemical elements. **Figure 2** shows the enrichment factors of nine metal contaminants established along the soil core to describe past and more recent floodplain contaminations. Enrichment factors (EF) in Co and Ni were found close to 1, which means that these trace metal concentrations are similar to the geochemical background (Le Cloarec et al, 2011). Copper, zinc and lead exhibit the same contamination trends with high contamination levels from 1950 to 1970, followed by a gradual decrease in recent years. This trend is directly explained by the use of these metals during the industry apogee of this basin, and the deindustrialization of the Seine basin in the late 1960s (Ayrault et al, 2020). Peak EFs were 14.8 for lead and 15.8 for zinc in 1953, and 18.5 for copper in 1956, indicating a significant contamination. Their concentrations are therefore close to or greater than 15 times the geochemical background. EFs for Cd have been observed in the range between 14.2 and 86.9 (showing extremely high contamination), with a maximum reached in the 1960s. The most recent layers have levels closely related to those of the early 1940s, in agreement with the flow of release of Cd into the environment due to plating industries, an activity which has greatly decreased since the end of the 1960s (Ayrault et al, 2020).

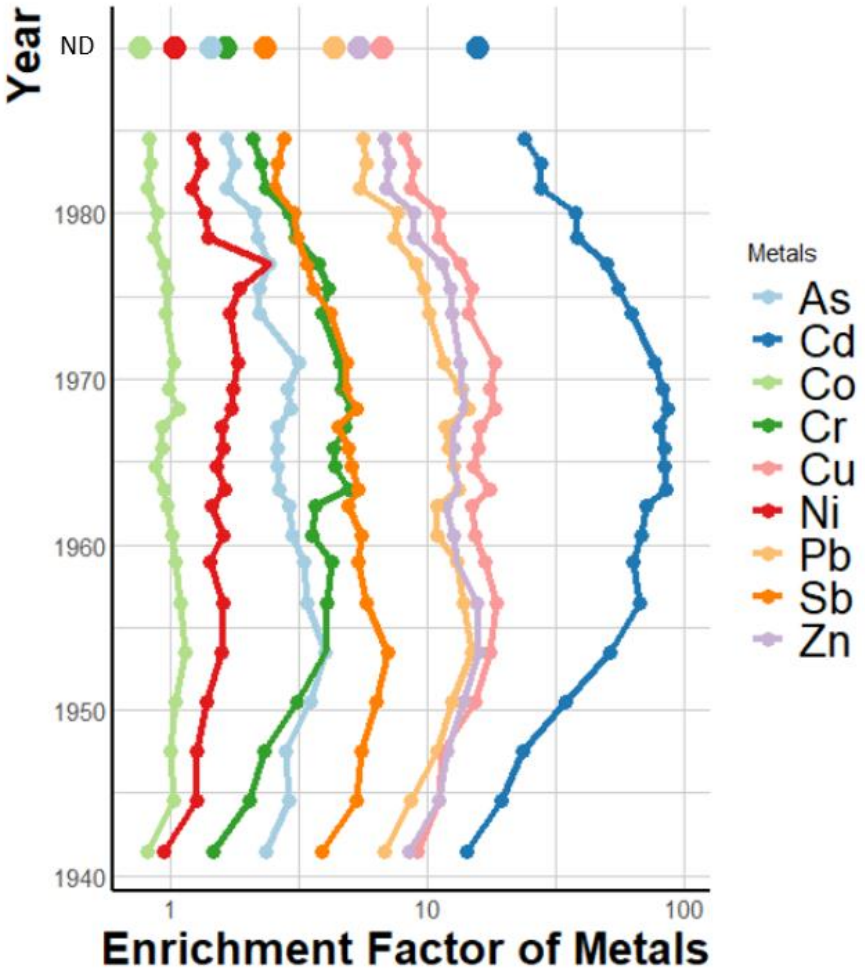
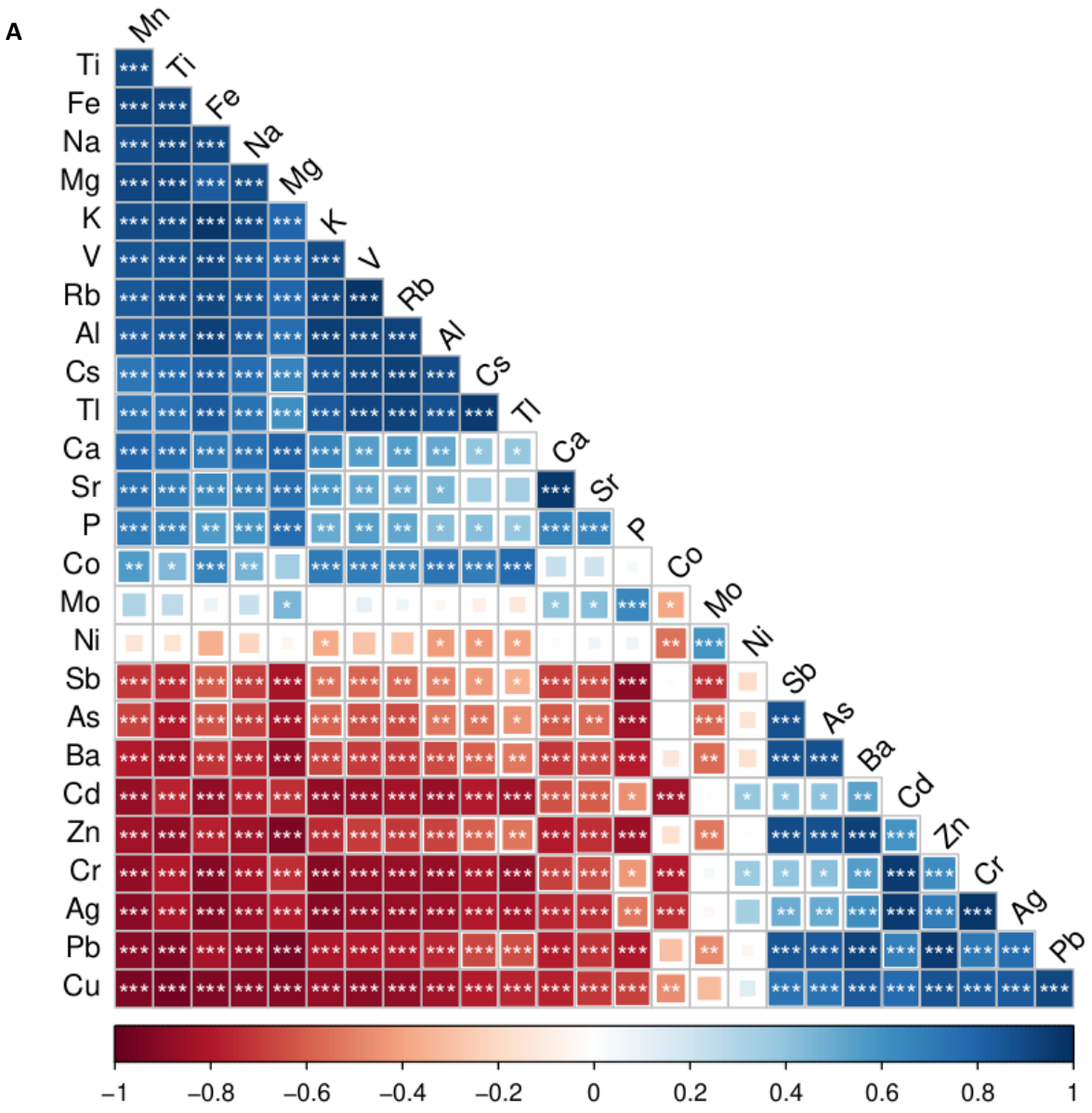


Figure 2. Enrichment factors (EF) of metal contaminants in the soil core. From 1987 until now, mean of elements were calculated to evaluate the EF corresponding to the ploughing zone (ND time point indicated at the top).

Geochemical element profiles in the soil core define clusters of soil layers



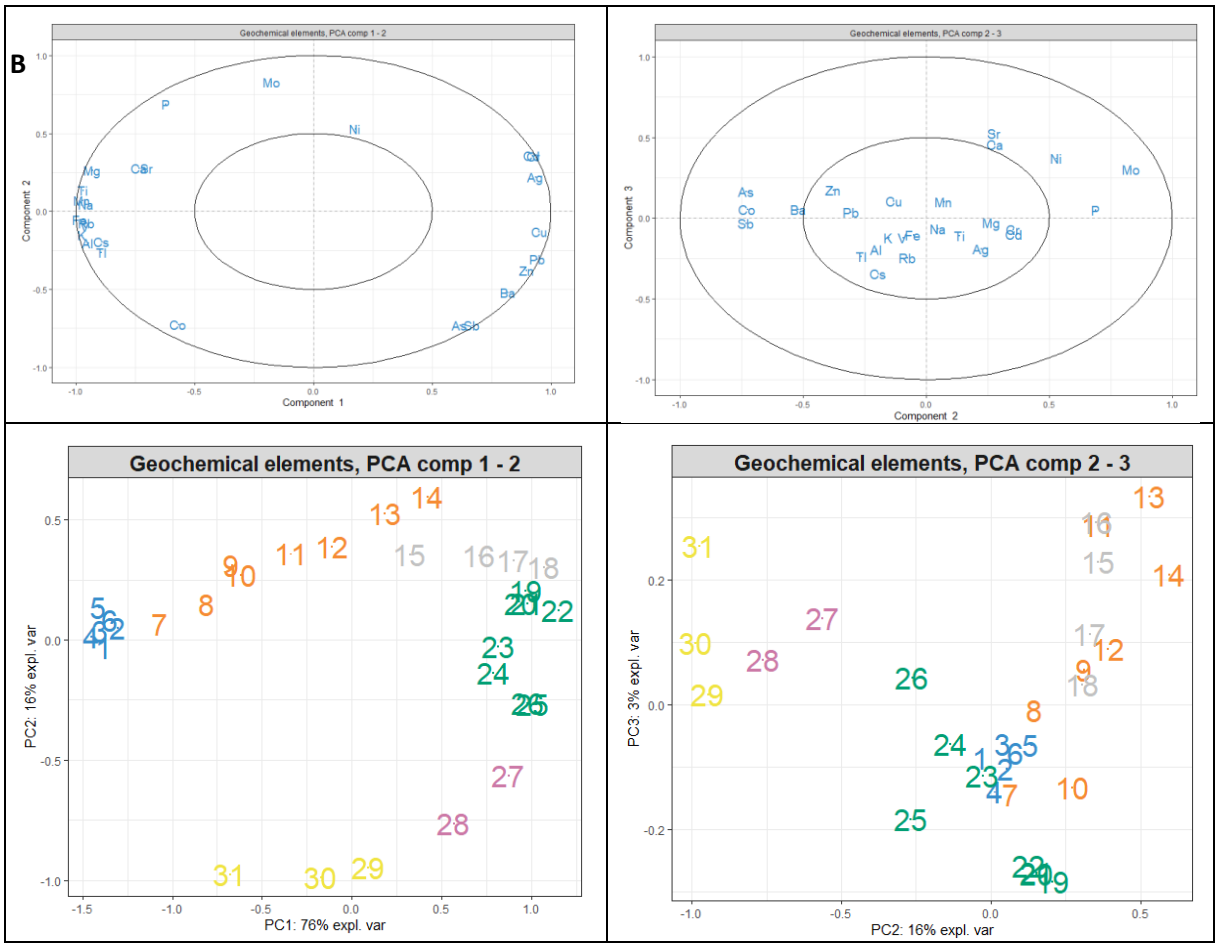


Figure 3. (A) Spearman correlation plot of geochemical elements concentrations with positive correlation in blue gradient and negative in red gradient and stars are used for statistical correlation test (p-value < 0.05 (*), <0.01 (**), < 0.001 (***)). **(B)** Principal component analysis of geochemistry elements with CLR normalized metal concentrations by displaying PCA components 1-2 and 2-3 in upper panel of geochemistry elements and those on sediment layer in bottom panel.

Figure 3 (Panel A) shows the correlation between the profiles of geochemical element concentrations along the core. For this, the element concentrations were normalized by the centered log ratio (CLR) method (Grunsky et al., 2014). Spearman correlation plot shows elements with more or less strong positive and negative correlations, delineating four major distinctive clusters of layers with correlated metal element abundancies (Figure 3A). These results can be explained by either similar level metal contamination over several years or metal species diffusion during or after flood deposits in neighboring layers. A principal component analysis (PCA) was performed on CLR normalized element concentrations (Figure 3B). The first principal component (PC1) accounted for 76% of total variance, compared to 16% and 3% for the second and third components, respectively. According to axes 1, 2 and 3 of the PCA, the most superficial layers corresponding to the ploughing depth (1-6), the recent layers (7-14), the intermediate layers (15-26), and the oldest (27-31) were distinguished. Additionally, the two oldest layer groups can be further subdivided: 15-18 and 19-26 on the one hand, and 27-28 and 29-31 on the other, resulting in six layer-clusters. A proposed interpretation is that the major elements/contaminations are represented on axis 1, and the old/more recent contaminations on axis 2. Thallium is not a contaminant in the Seine where it is present in natural quantities (Ayrault et al., 2010) and therefore its position in the PCA near Al is expected. Calcium and strontium (r=0.97) on the

one hand, and potassium and rubidium (r=0.92) on the other hand, are couples of chemically closely related elements and their correlation is seen in the correlation plot and by PCA. The third component can be explained by the Ca and Al ratio. As Ca increases, Al decreases. There is more clay at the bottom of the core and more calcium at the top (shown by axis 3, although at 3%). Last, we made a Spearman correlation plot for each of these layer groups. The recent layers (8-14) show the most variability between the elements with high positive and negative correlations, more than the ploughing zone (group 1) and the intermediate layers. The oldest layers (group 5 + 6) show different patterns.

Identification of main phyla along the depth profile

As multi-omics data were obtained on five predefined groups of layers (**Figure 1**), the results of 16S rRNA gene amplicon metabarcoding, metagenomics, and metaproteomics could be directly compared as illustrated in **Figure 4 (Panel A)**. While metagenomics and metaproteomics have a much wider scope for typing without any a priori the organisms present in the sample, the 16S rRNA gene amplicon metabarcoding is generally limited to Bacteria. Due to this limitation, only this super kingdom was considered for a fair comparison. As expected, the three methodologies are concordant for the identification of the main phyla. In terms of biomass ratio, a strong homogeneity was observed for the metaproteomics and metagenomics results which reproduce the same variations along the five groups of layers that have been defined prior metagenomics sequencing while the two approaches measure very different molecules. This high level of similarity confirms the validity of these two approaches when quantitative measures are needed. Meanwhile, the results of 16S rRNA gene amplicon metabarcoding are slightly discordant, possibly due to amplification with universal primers which is known to not give homogeneous results for all bacterial phyla (Yeh et al., 2021). As expected for soils (Delgado-Baquerizo et al., 2018), Alphaproteobacteria and Actinobacteria are dominant and quantified by metaproteomics to represent 28%, and 21% of the total signal, respectively, followed by Betaproteobacteria, Acidobacteria, Chloroflexi and Firmicutes with abundances between 5 and 7% each.

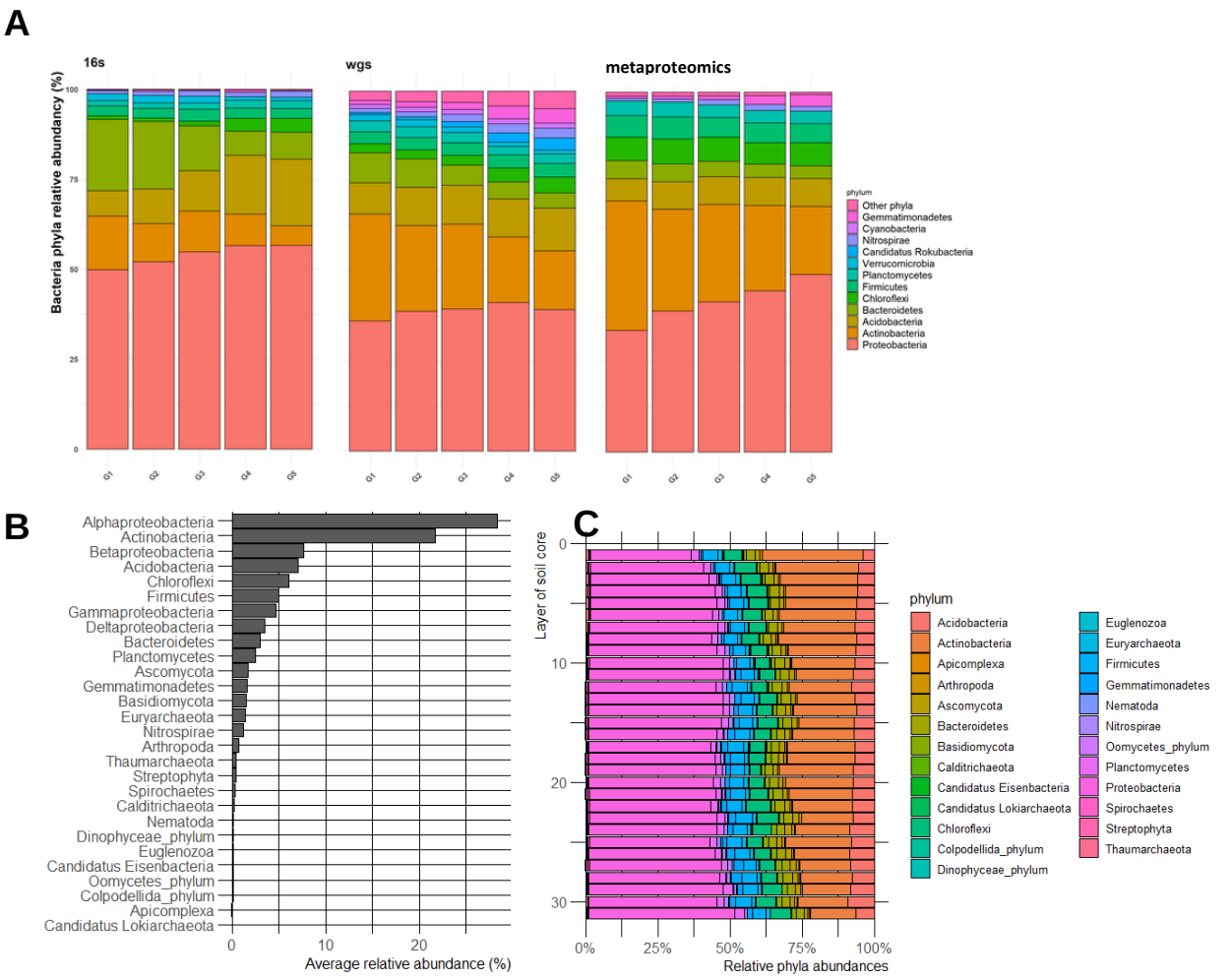


Figure 4. Taxonomical structure of microbial communities assessed by metaproteomics. Panel A. Comparison of abundances of phyla obtained with complementary omics on the five groups of layers: Metabarcoding with 16S rRNA gene amplicon sequencing, metaproteomics and metagenomics (left to right graphs). Panel B. Ranking of phyla by their average abundances. Panel C. Profile of dominant phyla along the core soil layers.

While metaproteomics results per layer-groups **in Figure 4 (Panel A)** were averaged, individual metaproteomics results for each of the layers were also obtained, using the taxonomical information associated with the peptides identified by tandem mass spectrometry. Results allow the identification of microorganisms present at the various possible taxonomic ranks in all superkingdoms, and the estimation of their biomass ratio. When considering the dataset in its whole, 25 different phyla were identified belonging to Eukaryotes (10), Archaea (3), and Bacteria (12) superkingdoms. Interestingly, two candidatus phyla were detected: candidatus Eisenbacteria and candidatus Lokiarchaeota by means of 10 and 7 taxon-specific peptides in layers, respectively. The first phylum was discovered by metagenomics from aquifer sediments and groundwater in 2016 (Anantharaman et al 2016) and named in honor of Jonathan A. Eisen from the University of California. The second phylum was also discovered by metagenomics (Orsi et al. 2020) and an isolate was recently obtained (Imachi et al., 2020). It is worth noting that these two poorly characterized branches of the Tree of Life represent 0.11% and 0.01%, respectively, of the total biomass of the soil core. Their presence is higher in layer 27 (0.25%) and layer 22 (0.07%), respectively. Despite their low abundance, the pipeline used to interpret the large dataset is performing well. Overall, the biomass percentage of Eukaryota, Archaea,

and Bacteria super-kingdoms are 5, 2, and 93%. **Figure 4** (Panel B) shows the relative abundances of identified phyla ranked from the most to the least abundant.

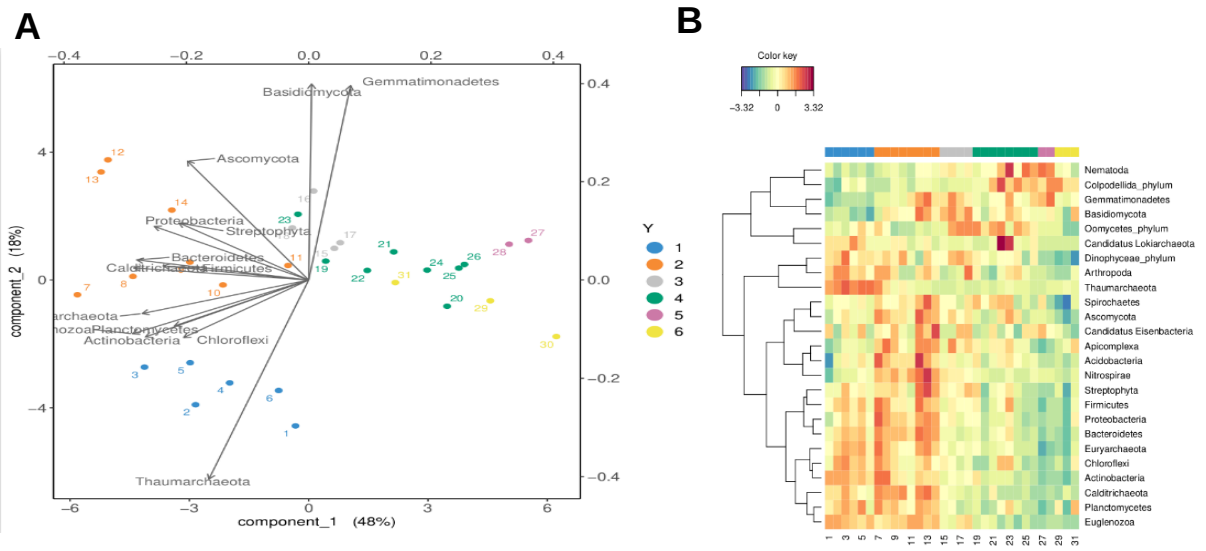


Figure 5. Correlation between geochemical and metaproteomics datasets. **Panel A**: sPLSDA using phyla abundancies to separate geochemically-clustered layers. **Panel B**: Corresponding Clustered image map.

A more in-depth statistical analysis of the metaproteomics data using sPLSDA and the six layer-clusters defined on the basis of their biogeochemical elements makes it possible to highlight the phyla able to separate the six layer-clusters identified previously using metal profiles (**Figure 5, Panel A**). **Figure 5 (Panel B)** shows the corresponding color coded heatmap termed Clustered image map in the mixOmics package. For example, the phyla *Dinophyceae*, *Arthropoda*, and *Thaumarchaeota* are highly correlated and are found in greatest abundance near the surface. The first cited phylum corresponds to the dinoflagellates which are unicellular aquatic eukaryotes responsible for blooms and capable of producing phycotoxins. The presence of arthropods is explained by the release of proteins by these animals near the surface. The last phylum includes chemolithoautotrophic archaea known to be found in non-extreme and oxic environments and found associated with protists and animals (Borrel et al., 2020). Thus the presence of representatives of these three branches of the Tree of Life near the surface of the soil core is in some way expected. Another cluster associated the phyla *Colpodellida* and *Nematoda* in the deep layers. The first mentioned are the alveolate eukaryotes which group together small predatory species which consume small protists by sucking up the cellular content of the prey. Nematodes are very mobile roundworms that can harbor a large number of microorganisms and release them with their excrement. The grouping of *Gemmatimonadetes* and *Basiodomycota*, both abundant in the middle layers, is a third example. The former is a group of gram-negative aerobic bacteria while the latter includes filamentous fungi composed of hyphae.

A Wilcoxon statistical test was performed in order to identify differently abundant phyla between layer-clusters 1 and 4, cluster 1 encompassing surface layers and cluster 4 layers where geochemical conditions are drastically different, including concentrations of metals such as Cd, Cu, Cs concentrations are at their highest (Figure S2). As shown in **Supplementary Figure 3**, *Gemmatimonadetes*, *Nematoda* and *Colpodellida* were found significantly more abundant in depth and *Actinobacteria, Chloroflexi, Euryarchaeta*, *Thaumarchaeta, Calditrichaeota* and *Euglenozoa* were more abundant near the surface (BH adjusted-pvalue 0.05, abundancy difference above 20%).

Metaproteomics analysis at lower taxonomical ranks indicated the presence in the core of 52 class of organisms, 85 orders, and 115 families. However, we focused our attention on the phylum taxonomical

rank to see the overall effects of the depth of the sediments. Interestingly, the main components of the layers have not changed overall. The abundance of Proteobacteria and Acidobacteria were found increased slightly with depth, while the abundance of Actinobacteria decreases from 35% to 16% as shown in **Figure 3 (Panel B)**.

Microbial community structure and the presence of contaminants

A hierarchical clustering of the different profiles along the soil core was proposed to highlight the possible correlation between phyla and specific geochemical elements. As shown in **Figure 6 (Panel A)**, the first items listed at the top of the figure are variables found in abundance in layers 17 to 25: Cd and Ag, two toxic metals, are found to be correlated with the presence of *Ascomycota*, *Streptophyta*, and *Candidatus Eisenbacteria*. While clearly the first two phyla are already known to be resistant to these toxic metals and can accumulate them, nothing is known about the *Candidatus Eisenbacteria* phylum. It would be interesting to isolate strains of this phylum from the material found in deep layers and to evaluate on these isolates their capacity to resist to these two specific toxicants.

Canonical Correlation Analysis (CCA) was also used to analyze more specifically the direct correlation between metaproteomics phyla and geochemical datasets. **Figure 6 (Panel B)** shows several results relating directly metals to phyla. For example the Oomycetes phylum is shown to be more correlated with slices were Cd, Ag and Cr are at peak concentration as visible in Figure S4, possibly related to a better tolerance for these metals than other phyla. *Nematoda* appear correlated with deeper slices where Co, Sb or As are at maximal concentration as it can be verified on **Figure 4 (Panel A)** and Supplementary **Figure S5**, as well as *Colpodellida*, which are again Eukaryota, possibly because of a better tolerance to heavy metals or chemicals. **Figure S5A** highlights (i) a decreasing abundancy of Eukaryota : *Euglenozoa* and *Apicomplexa* with depth coherently with the correlation with surface layers and metals Mn, Ti in **Figure 4C**, and (ii) an abundancy maximized in intermediate depths for Eukaryota: *Colpodellida*, *Nematoda, Oomycetes* and Archaea: *Candidatus_Lokiarchaeota*, in coherence with **Figure 6 (Panel B)** with a correlation of these phyla with As, Sb, Co, Ba, Pb, Cu, Cr, Ag, Cd metals with Colpodellida and Nematoda possibly more tolerant to the former and Oomycetes to the latter metals of this series. **Figure 6 (Panel B)** also indicates *Candidatus Eisenbacteria* as a possibly Ag, Cr and Cd tolerant Bacteria. Interestingly, *Candidatus Eisenbacteria* was found in a Hg contaminated soil and presents putative heavy metal resistance genes (Kim et al., 2021).

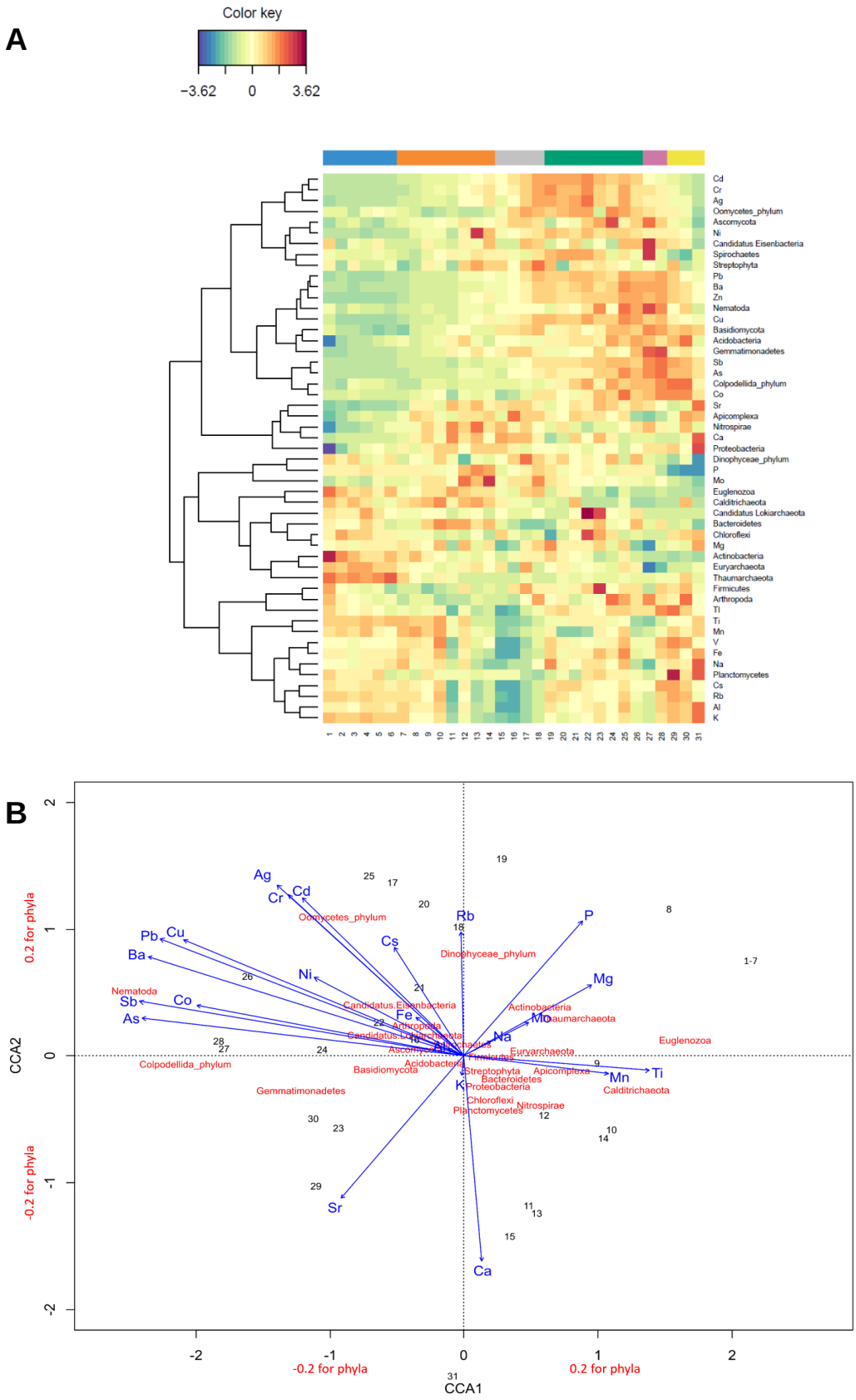


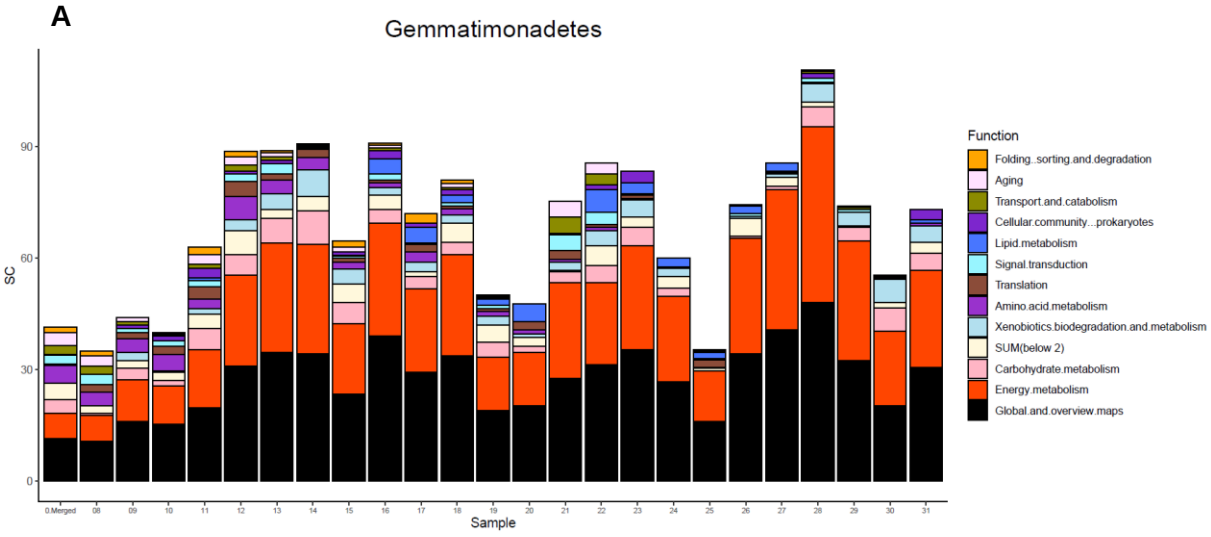
Figure 6. (A) Clustered image maps of phyla and geochemical elements profile along the layers using Euclidean distances and complete clustering method. **(B)** Correlation between geochemical and metaproteomics datasets using Canonical-correlation analysis (CCA) of the phyla, geochemical variables and samples.

Soil microbial functions uniquely highlighted by metaproteomics

Taxonomical and functional mapping on peptides as can be done in metaproteomics allows either global taxonomical or functional assessment, but also taxonomically-resolved functional analysis, at different taxonomical and functional globalisation levels. Depending on the signal level, analysis can involve broad categorisation such as superkingdom-resolved KEGG pathway-level 2 as seen in **Figure S4A**, more resolved categorisation such as family-resolved KEGG pathway map analysis as seen in **Figure S4D**. The analysis of the whole dataset using GHOSKOALA resulted in 1882 KO assignations which were also mapped to KEGG pathway map or pathway level 2 entries. **Figure S4** details several results concerning one of the major metabolic pathways in microbial systems, the carbon metabolism pathway, illustrating the power of interpretation of metaproteomics even for difficult samples such as soil.

For the assessment of functions associated with geochemical profiles, we focused our attention on two of the bacterial phyla previously shown to be significantly different between layer-clusters 1 and 4, namely Gemmatimonadetes more abundant in depth and Actinobacteria more abundant near the surface. **Figure 7** displays stacked barplots of the two phyla functionally informed at KEGG pathway level 2 and KEGG pathway map resolutions. Of note, relative stacked-barplots were deemed efficient to monitor altered gene expression of an organism, since a major cell response to local contaminants would result in a corresponding shift in the functional profiles. None of the phyla allowed us to pinpoint functional KEGG terms which could be associated with heavy-metal tolerance, including ko terms (data not shown) specific of P-type ATPase, RND transporters or HME family (Nies, 1999 ; Nikaido, 2018).

This suggests that the chronical exposure of the monitored microorganisms to metals released by anthropogenic activities was not done at a sufficient concentration to alter specifically and significantly the main functional expression profiles of the monitored phyla, whatever their gene repertoires.



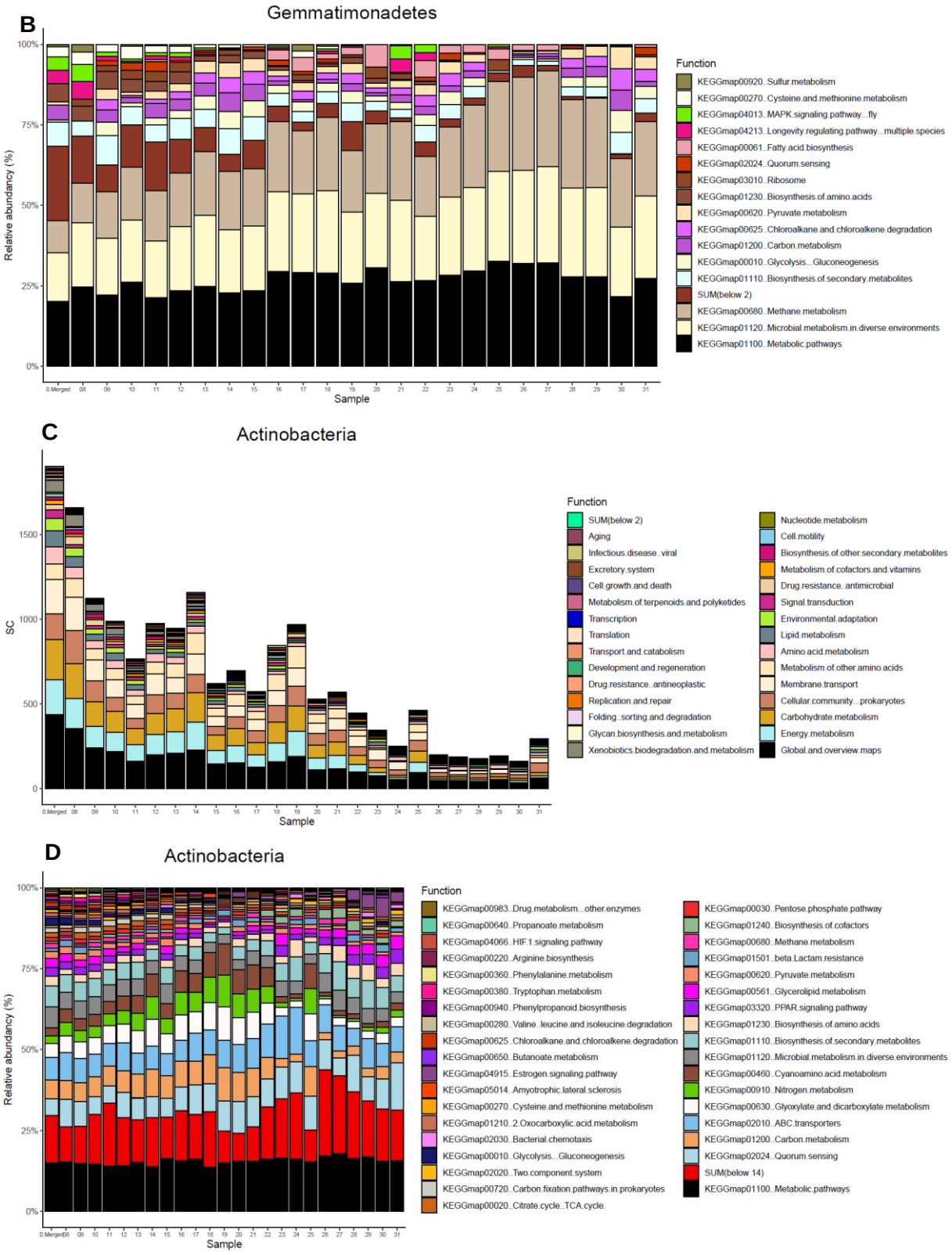


Figure 7: **Functional profiles of differencially abundant phyla** between surface (Actinobacteria more abundant) and cluster-layer with maximal heavy-metals concentrations (Gemmatimonadetes more abundant). Panel A: absolute stacked barplot of Gemmatimonadetes at KEGG pathway level 2 resolution, Panel B: relative stacked barplot of Gemmatimonadetes at KEGG pathway ma resolution, Panel C: absolute stacked barplot of Actinobacteria at KEGG pathway level 2 resolution, Panel D: relative stacked barplot of Actinobacteria at KEGG pathway ma resolution. Relative barplots can

indicate shifts in bacteria metabolism in relation with the geochemical environment. No profile could be associated with both heavy-metals profiles and KEGG pathway maps linked with metal detoxification.

DISCUSSION

Microbial communities are an indicator of soil ecosystem health. It had been proposed to manipulate them to improve the recovery of degraded ecosystems or to boost agriculture performances. In this context, increasing the understanding of the functioning of these microbial communities is a prerequisite before any further terraforming. Focused on the proteins that are the real workhorses of the cells, metaproteomics is able to give valuable insights into the functioning of the microbial communities. Furthermore, metaproteomics delivers the whole panorama of organisms present in the sample with quite precise estimation of their respective biomasses (Pible et al., 2020). Here, the metaproteomics dataset acquired on the soil core comprises 4,967,934 million of high-quality MS/MS spectra, resulting in one of the most extended metaproteomics dataset on the same sampled soil. Its interpretation identified 25 different phyla, 52 classes, 85 orders, and 115 families of organisms. Fungi are important players of soil ecosystems, especially fungal hyphae contribute to soil networks actively structuring their environment. Interestingly, metaproteomics gives comparative results for any phyla, thus including fungi and unicellular eukaryotes. Among the identified eukaryotic phyla, Ascomycota and Basidiomycota are among the significant contributors of the soil protein biomass. Other unicellular eukaryotes could be detected but in very low abundance: *Oomycota* which form a distinct phylogenetic lineage of fungus-like eukaryotic microorganisms, *Euglenozoa* which are unicellular microorganisms with flagellum, and *Colpodellida* which includes small predatory species. These identifications prove the depth of analysis obtained with the dataset and the proposed pipeline for its interpretation.

Most soil activity usually occurs in the upper layers, where organic matter and mineral content are higher, and along plant roots that secreted exudates rich in sugars for shaping their surrounding microbial communities (Lidbury et al., 2022). Such a result was observed in the present study with a strong metabolic activity in the surface layers marked with prominent carbohydrate catabolic activities. Because of the importance of topsoil, most of previous studies on soils are focused on this specific compartment, while results on the vertical distribution of microorganisms in soils is rather scarce. Previous studies have shown microbial abundance and richness are decreasing from topsoil to subsoil. For example, it has been shown that the bacterial diversity dropped by 20 to 40% in the deepest layers compared to topsoil in a forested montane watershed in Colorado (Eilers et al., 2012), with mainly a decline in the relative abundance of *Bacteroidetes* with depth and peak in the relative abundance of *Verrucomicrobia* between 10 and 50 cm. In our dataset, *Bacteroidetes* (Y% in the top layer) is not a major phylum contrary to what has been previously observed for other soils across France (Karimi et al. 2018) or elsewhere (Eilers et al., 2012). However, the same trend is observed in the analyzed soil core regarding microbial diversity along depth. Noteworthy, the grain size of the vertical analyses is generally not sufficient to obtain an accurate profile of the variations. While 20 (Li et al., 2020) or 5 (Wu et al., 2019) cm layers have been reported, our strategy was to decrease this grain size to the maximum (3 cm layers) resulting in a rich and redundant dataset where trends in microbial diversity and functionalities can be better assessed. The specific site of Bouafles that was sampled is a historical archive of anthropogenic contaminations over almost a century located ideally downstream of Paris. Paris area being highly impacted by long-term anthropogenic activities, no other comparable site could be identified in the course of this study to have a control soil core equivalent to this site in terms of sediments and microbial communities. For this reason, we focused our attention

on this single soil core to draw possible correlations that could then be further experienced with specific experimental set-up.

The metaproteomics pipeline used in the present study is based on the synergy of generalist database such as NCBInr and specific metagenome-derived database more adapted to each sample (Jouffret et al., 2021). To gain in sensitivity when searching giant metaproteomics databases, the cascaded search approach proposed previously (Jagtap et al., 2013, Bassignani et al., 2021) is performing well, but is resulting in results that cannot be equally compared between samples. Delineating 35 layers and applying the cascaded search to each layer tend to amplify the possible bias in such comparative studies with a dramatic increase of missing values. Here, we have introduced a new round of search merging all the databases created per layer in a single, unique database to interrogate each MS/MS dataset with strictly equal FDR chances. This strategy allows a direct comparison of the results between layers and avoid the frequent missing values conundrum previously highlighted in metaproteomics (Plancade et al., 2022).

With the profile of geochemical elements in hands, we could define 5 groups of layers in the soil core with probably similar physical and chemical properties. This delineation allows to observe whether diversity is different amongst these conditions. A multivariate ordination technique, CCA, has highlighted which variables between metaproteomics-established phyla and geochemical elements are correlated. Interestingly, oomycetes are strongly correlated with the presence of Ag, Cd, and Cr. Even if these microorganisms have been counterselected in these layers due to their resistance to the toxicity of these metals should be further tested after their isolation and systematic phenotypic characterization. *Calditrichaeota*, which comprises Hg methylating microbial groups (Lin et al., 2021), is strongly correlated to the presence of Mn and Ti. Whether specific enzymatic arsenal of enzymes is associated to the tolerance to these metals would be interesting to establish on isolates of this hitherto poorly characterized phylum. Thus, this analysis points at the importance of incorporating culturomics and phenotypic assays of isolates to such microbiome characterization.

MATERIALS AND METHODS

Soil material

The soil core sampled on May 23rd 2018 from a floodplain at Bouafles near the Seine River (France) as described in Jouffret et al., 2021 was sliced in 35 equals samples every 3 cm. Two grams of each layer were pooled and homogenized for DNA extraction. For each layer, extraction of proteins was done in triplicate from five grams of soil that were homogenized, i.e. fifteen grams in total.

Geochemistry data

Elemental concentration analyses were performed in the LSCE Laboratory (Gif-Sur-Yvette, France). The samples dried at 40°C for 24 h where ground and sieved to 2 mm. One hundred milligrams were totally digested using a four acids protocols described in (Froger et al., 2018). The metal concentrations were determined using Thermo ScientificTM iCAPTM TQ ICP-MS. Certified reference materials (NIST1640a, SL-1) were used to evaluate the quality of the determined concentrations (< 5 % for all elements except Ni, Cu and Zn with <2% ; Na, Mg, Al, Sr, Ag with <10% ; K with < 15 % ; Ti, As, Mo with < 20% and P with 30% of uncertainties).

Enrichment factor (EF) of a metal (M) compared to its natural value (M*natural*) is defined by the formula (M/Al)/(M/Al)*natural ratio* where metal concentrations were normalized to aluminum (Al) concentrations in the samples and in the geochemical background (Al*natural*).

Metagenomics analysis

DNA was extracted from five samples of soil (groups of layers 1-2 (G1), 3-5 (G2), 6-10 (G3), 11-21 (G4), and 22-35 (G5) from 1 g of lyophilized sample as starting material. The five DNA samples were sequenced on a HiSeq 4000 Illumina run in 2x150 bp by GenoScreen (Lille, France). Reads were analyzed using the phylogenetic MG-RAST pipeline (Meyer, Paarmann et al. 2008). For creating the metagenomics database, FragGeneScan v1.3 (Rho, Tang et al. 2010) was applied on reads with 0.01% error rate model.

Protein extraction, protein proteolysis, and mass spectrometry

Proteins were extracted using the NoviPure Soil Protein Extraction Kit (Mo-Bio). After centrifugation, proteins were precipitated and resuspended in 40 µL LDS 1X (Invitrogen) containing 5% beta-mercaptoethanol as previously described (Jouffret et al., 2021). Protein samples were subjected to SDS-PAGE gel electrophoresis on NuPAGE 4–12% Bis-Tris gel (Invitrogen) for 5 min. The total proteins from the 105 samples were treated and in-gel proteolyzed with trypsin gold (Promega) as previously described (Rubiano-Labrador et al., 2014). Tryptic peptides were analyzed on a Q-Exactive HF mass spectrometer (Thermo) coupled to an Ultimate 3000 nano LC system (Thermo). Briefly, peptides (8 µl) were injected and desalted on a reverse-phase PepMap 100 C18 µ-precolumn (5 µm, 100 Å, 300 µm i.d. × 5 mm, Thermo). Then, they were resolved along their hydrophobicity on a nanoscale PepMap 100 C18 nanoLC column (3 µm, 100 Å, 75 µm i.d. × 50 cm, Thermo). The flow rate was 0.2 µL per min. A biphasic 90-min gradient of mobile phase A (0.1% HCOOH/100% H2O) and phase B (0.1% HCOOH/80% CH3CN) was developed from 4 to 25% B in 70 min and then from 25 to 40% B in 20 min. The mass spectrometer was operated in Top20 data-dependent acquisition mode with the same parameters as previously described (Jouffret et al., 2021). MS/MS spectra were interpreted using the 2.6.1 version of Mascot Daemon software (Matrix Science) with the following parameters: tolerance of 5 ppm and 0.02 Da for the parent ion and the secondary fragments, respectively, only 2+ and 3+ peptide charges, two missed cleavages at maximum, carbamidomethylation of cysteine (+57.0215) as fixed modification, oxidation of methionine (+15.9949) as variable modification, and trypsin as proteolytic enzyme. The three nanoLC-MS/MS datasets of the same layer were merged. Three rounds of interpretation were carried out. The databases used in the first round were: i) a selection of coding gene sequences predicted by FragGeneScan on the metagenome of the group of layers, and ii) the NCBInrS database (Grenga et al., 2022) followed by a specific NCBI derived database of all representatives from the identified genera in each layer. Both queries were performed selecting peptide-to-spectrum matches (PSMs) at a Mascot p-value threshold of 0.3. The second round was performed systematically for each layer against a database of the identified proteins in round 1. All the protein sequences selected at p-value 0.3 in each layer were grouped in a unique final database for round 3. This database representative for the whole soil core was queried at FDR decoy-free 1% for the 35 metaproteomes. The metaproteomics data have been deposited to the PRIDE repository.

Metaproteomics interpretation

The identified peptide sequences were mapped to taxa using in-house SQLite databases built from NCBI data as previously described (Gouveia et al., 2020). Taxonomies were identified on the basis of the raw PSMs for each taxon, the numbers of matching peptide sequences and taxon-specific peptides. KEGG Orthology (KO) functions were assigned to the peptides identified by tandem mass spectrometry in each layer with GhostKOALA (Kanehisa et al., 2016) and their abundances were directly assessed by the associated spectral counts as previously described (Jouffret et al., 2021). Multivariate statistical analysis was performed using R packages mixOmics for PCA and sPLS-DA, vegan for canonical-correlation analysis (CCA), tidyverse for transform and visualize data.

REFERENCES

Abiraami, T. V., Singh, S., & Nain, L. (2020). Soil metaproteomics as a tool for monitoring functional microbial communities: promises and challenges. *Re/Views in Environmental Science and Bio/Technology*, *19*(1), 73–102. https://doi.org/10.1007/s11157-019-09519-8

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., Thomas, B. C., Singh, A., Wilkins, M. J., Karaoz, U., Brodie, E. L., Williams, K. H., Hubbard, S. S., & Banfield, J. F. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, *7*(1), 13219. https://doi.org/10.1038/ncomms13219

Ayrault, S., Meybeck, M., Mouchel, J.-M., Gaspéri, J., Lestel, L., Lorgeoux, C., & Boust, D. (2020). Sedimentary archives reveal the concealed history of micropollutant contamination in the Seine River basin. In *The Handbook of Environmental Chemistry* (pp. 269–300). Springer International Publishing.

Ayrault, S., Priadi, C. R., Evrard, O., Lefèvre, I., & Bonté, P. (2010). Silver and thallium historical trends in the Seine River basin. *Journal of Environmental Monitoring: JEM*, *12*(11), 2177–2185. https://doi.org/10.1039/c0em00153h

Ayrault, S., Roy-Barman, M., Le Cloarec, M.-F., Priadi, C. R., Bonté, P., & Göpel, C. (2012). Lead contamination of the Seine River, France: geochemical implications of a historical perspective. *Chemosphere*, *87*(8), 902–910. https://doi.org/10.1016/j.chemosphere.2012.01.043

Bassignani, A., Plancade, S., Berland, M., Blein-Nicolas, M., Guillot, A., Chevret, D., Moritz, C., Huet, S., Rizkalla, S., Clément, K., Doré, J., Langella, O., & Juste, C. (2021). Benefits of iterative searches of large databases to interpret large human gut metaproteomic data sets. *Journal of Proteome Research*, *20*(3), 1522–1534. https://doi.org/10.1021/acs.jproteome.0c00669

Bastida, F., Eldridge, D. J., García, C., Kenny Png, G., Bardgett, R. D., & Delgado-Baquerizo, M. (2021). Soil microbial diversity-biomass relationships are driven by soil carbon content across global biomes. *The ISME Journal*, *15*(7), 2081–2091. https://doi.org/10.1038/s41396-021-00906-0

Bastida, F., García, C., Fierer, N., Eldridge, D. J., Bowker, M. A., Abades, S., Alfaro, F. D., Asefaw Berhe, A., Cutler, N. A., Gallardo, A., García-Velázquez, L., Hart, S. C., Hayes, P. E., Hernández, T., Hseu, Z.-Y., Jehmlich, N., Kirchmair, M., Lambers, H., Neuhauser, S., … Delgado-Baquerizo, M. (2019). Global ecological predictors of the soil priming effect. *Nature Communications*, *10*(1), 3481. https://doi.org/10.1038/s41467-019-11472-7

Bastida, F., Torres, I. F., Andrés-Abellán, M., Baldrian, P., López-Mondéjar, R., Větrovský, T., Richnow, H. H., Starke, R., Ondoño, S., García, C., López-Serrano, F. R., & Jehmlich, N. (2017). Differential sensitivity of total and active soil microbial communities to drought and forest management. *Global Change Biology*, *23*(10), 4185–4203. https://doi.org/10.1111/gcb.13790

Bastida, F., Torres, I. F., Moreno, J. L., Baldrian, P., Ondoño, S., Ruiz-Navarro, A., Hernández, T., Richnow, H. H., Starke, R., García, C., & Jehmlich, N. (2016). The active microbial diversity drives ecosystem multifunctionality and is physiologically related to carbon availability in Mediterranean semi-arid soils. *Molecular Ecology*, *25*(18), 4660–4673. https://doi.org/10.1111/mec.13783

Bru, D., Ramette, A., Saby, N. P. A., Dequiedt, S., Ranjard, L., Jolivet, C., Arrouays, D., & Philippot, L. (2011). Determinants of the distribution of nitrogen-cycling microbial communities at the landscape scale. *The ISME Journal*, *5*(3), 532–542. https://doi.org/10.1038/ismej.2010.130

Chen, Y., Jiang, Y., Huang, H., Mou, L., Ru, J., Zhao, J., & Xiao, S. (2018). Long-term and high-concentration heavy-metal contamination strongly influences the microbiome and functional genes in Yellow River sediments. *The Science of the Total Environment*, *637–638*, 1400–1412. https://doi.org/10.1016/j.scitotenv.2018.05.109

Crowther, T. W., van den Hoogen, J., Wan, J., Mayes, M. A., Keiser, A. D., Mo, L., Averill, C., & Maynard, D. S. (2019). The global soil community and its influence on biogeochemistry. *Science (New York, N.Y.)*, *365*(6455), eaav0550. https://doi.org/10.1126/science.aav0550

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., Maestre, F. T., Singh, B. K., & Fierer, N. (2018). A global atlas of the dominant bacteria found in soil. *Science (New York, N.Y.)*, *359*(6373), 320–325. https://doi.org/10.1126/science.aap9516

Eilers, K. G., Debenport, S., Anderson, S., & Fierer, N. (2012). Digging deeper to find unique microbial communities: The strong effect of depth on the structure of bacterial and archaeal communities in soil. *Soil Biology & Biochemistry*, *50*, 58–65. https://doi.org/10.1016/j.soilbio.2012.03.011

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(3), 626–631. https://doi.org/10.1073/pnas.0507535103

Gottschling, M., Czech, L., Mahé, F., Adl, S., & Dunthorn, M. (2021). The windblown: Possible explanations for dinophyte DNA in forest soils. *The Journal of Eukaryotic Microbiology*, *68*(1), e12833. https://doi.org/10.1111/jeu.12833

Gouveia, D., Pible, O., Culotta, K., Jouffret, V., Geffard, O., Chaumot, A., Degli-Esposti, D., & Armengaud, J. (2020). Combining proteogenomics and metaproteomics for deep taxonomic and functional characterization of microbiomes from a non-sequenced host. *Npj Biofilms and Microbiomes*, *6*(1), 23. https://doi.org/10.1038/s41522-020-0133-2

Grenga, L., Pible, O., Miotello, G., Culotta, K., Ruat, S., Roncato, M.-A., Gas, F., Bellanger, L., Claret, P.-G., Dunyach-Remy, C., Laureillard, D., Sotto, A., Lavigne, J.-P., & Armengaud, J. (2022). Taxonomical and functional changes in COVID-19 faecal microbiome could be related to SARS-CoV-2 faecal load. *Environmental Microbiology*, *24*(9), 4299–4316. https://doi.org/10.1111/1462-2920.16028

Grunsky, E. C., Mueller, U. A., & Corrigan, D. (2014). A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: Applications for predictive geological mapping. *Journal of Geochemical Exploration*, *141*, 15–41. https://doi.org/10.1016/j.gexplo.2013.07.013

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., Matsui, Y., Miyazaki, M., Murata, K., Saito, Y., Sakai, S., Song, C., Tasumi, E., Yamanaka, Y., Yamaguchi, T., … Takai, K. (2020). Isolation of an archaeon at the prokaryote-

eukaryote interface. *Nature*, *577*(7791), 519–525. https://doi.org/10.1038/s41586-019-1916-6

Jagtap, P., Goslinga, J., Kooren, J. A., McGowan, T., Wroblewski, M. S., Seymour, S. L., & Griffin, T. J. (2013). A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics*, *13*(8), 1352–1357. https://doi.org/10.1002/pmic.201200352

Jouffret, V., Miotello, G., Culotta, K., Ayrault, S., Pible, O., & Armengaud, J. (2021). Increasing the power of interpretation for soil metaproteomics data. *Microbiome*, *9*(1), 195. https://doi.org/10.1186/s40168-021-01139-1

Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of Molecular Biology*, *428*(4), 726–731. https://doi.org/10.1016/j.jmb.2015.11.006

Karimi, B., Terrat, S., Dequiedt, S., Saby, N. P. A., Horrigue, W., Lelièvre, M., Nowak, V., Jolivet, C., Arrouays, D., Wincker, P., Cruaud, C., Bispo, A., Maron, P.-A., Bouré, N. C. P., & Ranjard, L. (2018). Biogeography of soil bacteria and archaea across France. *Science Advances*, *4*(7), eaat1808. https://doi.org/10.1126/sciadv.aat1808

Kim, M., Wilpiszeski, R. L., Wells, M., Wymore, A. M., Gionfriddo, C. M., Brooks, S. C., Podar, M., & Elias, D. A. (2021). Metagenome-assembled genome sequences of novel prokaryotic species from the mercury-contaminated East Fork Poplar Creek, Oak Ridge, Tennessee, USA. *Microbiology Resource Announcements*, *10*(17). https://doi.org/10.1128/MRA.00153-21

Kostygov, A. Y., Karnkowska, A., Votýpka, J., Tashyreva, D., Maciszewski, K., Yurchenko, V., & Lukeš, J. (2021). Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. *Open Biology*, *11*(3), 200407. https://doi.org/10.1098/rsob.200407

Le Cloarec, M.-F., Bonte, P. H., Lestel, L., Lefèvre, I., & Ayrault, S. (2011). Sedimentary record of metal contamination in the Seine River during the last century. *Physics and Chemistry of the Earth (2002)*, *36*(12), 515–529. https://doi.org/10.1016/j.pce.2009.02.003

Li, S., Zhao, B., Jin, M., Hu, L., Zhong, H., & He, Z. (2020). A comprehensive survey on the horizontal and vertical distribution of heavy metals and microorganisms in soils of a Pb/Zn smelter. *Journal of Hazardous Materials*, *400*(123255), 123255. https://doi.org/10.1016/j.jhazmat.2020.123255

Lidbury, I. D. E. A., Raguideau, S., Borsetto, C., Murphy, A. R. J., Bottrill, A., Liu, S., Stark, R., Fraser, T., Goodall, A., Jones, A., Bending, G. D., Tibbet, M., Hammond, J. P., Quince, C., Scanlan, D. J., Pandhal, J., & Wellington, E. M. H. (2022). Stimulation of distinct rhizosphere bacteria drives phosphorus and nitrogen mineralization in Oilseed rape under field conditions. *MSystems*, *7*(4), e0002522. https://doi.org/10.1128/msystems.00025-22

Lin, H., Ascher, D. B., Myung, Y., Lamborg, C. H., Hallam, S. J., Gionfriddo, C. M., Holt, K. E., & Moreau, J. W. (2021). Mercury methylation by metabolically versatile and cosmopolitan marine bacteria. *The ISME Journal*, *15*(6), 1810–1825. https://doi.org/10.1038/s41396-020-00889-4

Lohmann, P., Benk, S., Gleixner, G., Potthast, K., Michalzik, B., Jehmlich, N., & Bergen, M. von. (2020). Seasonal patterns of dominant microbes involved in central nutrient cycles in the subsurface. *Microorganisms*, *8*(11), 1694. https://doi.org/10.3390/microorganisms8111694

Lorgeoux, C., Moilleron, R., Gasperi, J., Ayrault, S., Bonté, P., Lefèvre, I., & Tassin, B. (2016). Temporal trends of persistent organic pollutants in dated sediment cores: Chemical fingerprinting of the anthropogenic impacts in the Seine River basin, Paris. *The Science of the Total Environment*, *541*, 1355–1363. https://doi.org/10.1016/j.scitotenv.2015.09.147

Nies, D. H. (1999). Microbial heavy-metal resistance. *Applied Microbiology and Biotechnology*, *51*(6), 730–750. https://doi.org/10.1007/s002530051457

Nikaido, H. (2018). RND transporters in the living world. *Research in Microbiology*, *169*(7–8), 363–371. https://doi.org/10.1016/j.resmic.2018.03.001

Orsi, W. D., Vuillemin, A., Rodriguez, P., Coskun, Ö. K., Gomez-Saez, G. V., Lavik, G., Mohrholz, V., & Ferdelman, T. G. (2020). Metabolic activity analyses demonstrate that Lokiarchaeon exhibits homoacetogenesis in sulfidic marine sediments. *Nature Microbiology*, *5*(2), 248–255. https://doi.org/10.1038/s41564-019-0630-3

Plancade, S., Berland, M., Blein-Nicolas, M., Langella, O., Bassignani, A., & Juste, C. (2022). A combined test for feature selection on sparse metaproteomics data-an alternative to missing value imputation. *PeerJ*, *10*(e13525), e13525. https://doi.org/10.7717/peerj.13525

Rubiano-Labrador, C., Bland, C., Miotello, G., Guérin, P., Pible, O., Baena, S., & Armengaud, J. (2014). Proteogenomic insights into salt tolerance by a halotolerant alpha-proteobacterium isolated from an Andean saline spring. *Journal of Proteomics*, *97*, 36–47. https://doi.org/10.1016/j.jprot.2013.05.020

Schneider, T., Keiblinger, K. M., Schmid, E., Sterflinger-Gleixner, K., Ellersdorfer, G., Roschitzki, B., Richter, A., Eberl, L., Zechmeister-Boltenstern, S., & Riedel, K. (2012). Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. *The ISME Journal*, *6*(9), 1749–1762. https://doi.org/10.1038/ismej.2012.11

Starke, R., Jehmlich, N., & Bastida, F. (2019). Using proteins to study how microbes contribute to soil ecosystem services: The current state and future perspectives of soil metaproteomics. *Journal of Proteomics*, *198*, 50–58. https://doi.org/10.1016/j.jprot.2018.11.011

Tamtam, F., Le Bot, B., Dinh, T., Mompelat, S., Eurin, J., Chevreuil, M., Bonté, P., Mouchel, J.-M., & Ayrault, S. (2011). A 50-year record of quinolone and sulphonamide antimicrobial agents in Seine River sediments. *Journal of Soils and Sediments*, *11*(5), 852–859. https://doi.org/10.1007/s11368-011-0364-1

Van Den Bossche, T., Arntzen, M. Ø., Becher, D., Benndorf, D., Eijsink, V. G. H., Henry, C., Jagtap, P. D., Jehmlich, N., Juste, C., Kunath, B. J., Mesuere, B., Muth, T., Pope, P. B., Seifert, J., Tanca, A., Uzzau, S., Wilmes, P., Hettich, R. L., & Armengaud, J. (2021). The Metaproteomics Initiative: a coordinated approach for propelling the functional characterization of microbiomes. *Microbiome*, *9*(1), 243. https://doi.org/10.1186/s40168-021-01176-w

Williams, M. A., Taylor, E. B., & Mula, H. P. (2010). Metaproteomic characterization of a soil microbial community following carbon amendment. *Soil Biology & Biochemistry*, *42*(7), 1148–1156. https://doi.org/10.1016/j.soilbio.2010.03.021

Wu, Q., Du, Y., Huang, Z., Gu, J., Leung, J. Y. S., Mai, B., Xiao, T., Liu, W., & Fu, J. (2019). Vertical profile of soil/sediment pollution and microbial community change by e-waste recycling operation.

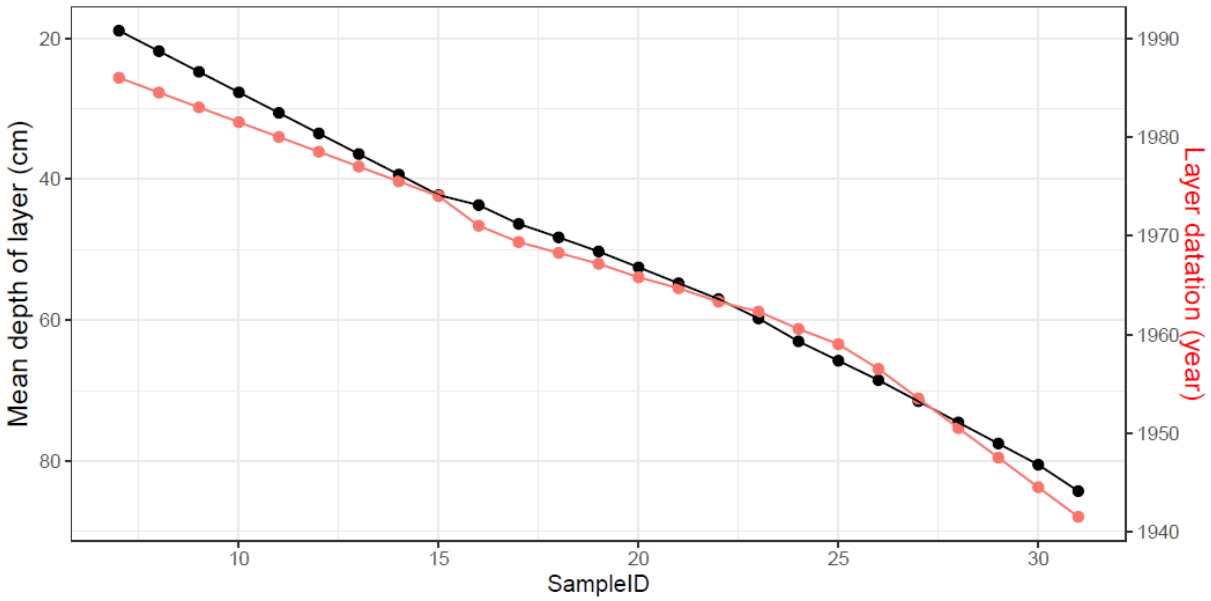
The Science of the Total Environment, *669*, 1001–1010. https://doi.org/10.1016/j.scitotenv.2019.03.178

Wu, Y., Zeng, J., Zhu, Q., Zhang, Z., & Lin, X. (2017). pH is the primary determinant of the bacterial community structure in agricultural soils impacted by polycyclic aromatic hydrocarbon pollution. *Scientific Reports*, *7*, 40093. https://doi.org/10.1038/srep40093

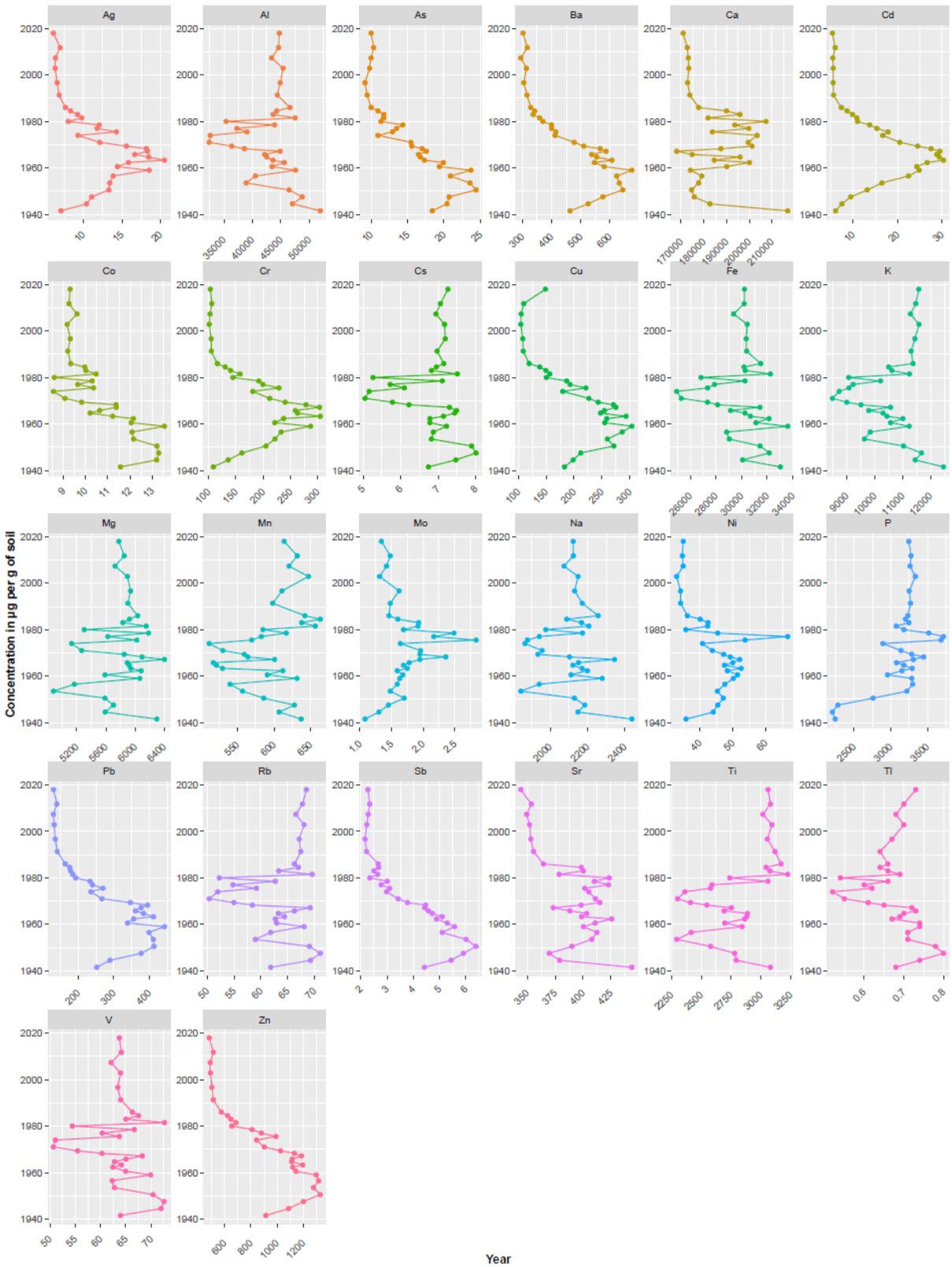
Yeh, Y.-C., McNichol, J., Needham, D. M., Fichot, E. B., Berdjeb, L., & Fuhrman, J. A. (2021). Comprehensive single-PCR 16S and 18S rRNA community analysis validated with mock communities, and estimation of sequencing bias against 18S. *Environmental Microbiology*, *23*(6), 3240–3250. https://doi.org/10.1111/1462-2920.15553

Zhou, J., Deng, Y., Shen, L., Wen, C., Yan, Q., Ning, D., Qin, Y., Xue, K., Wu, L., He, Z., Voordeckers, J. W., Van Nostrand, J. D., Buzzard, V., Michaletz, S. T., Enquist, B. J., Weiser, M. D., Kaspari, M., Waide, R., Yang, Y., & Brown, J. H. (2016). Temperature mediates continental-scale diversity of microbes in forest soils. *Nature Communications*, *7*, 12083. https://doi.org/10.1038/ncomms12083

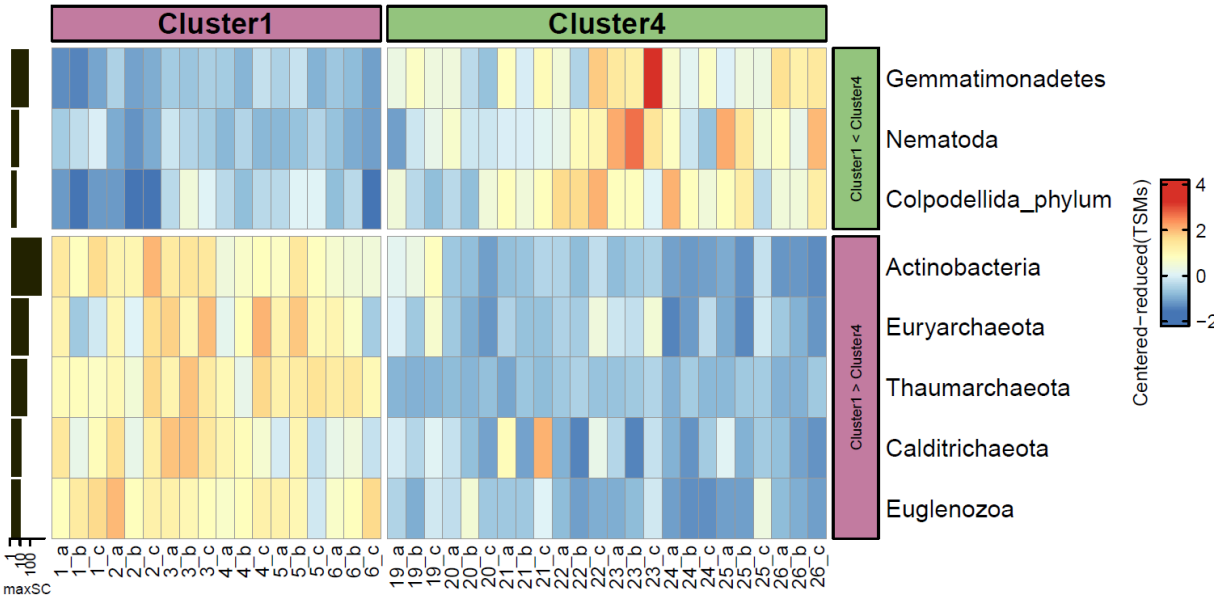
SUPPLEMENTARY FIGURES



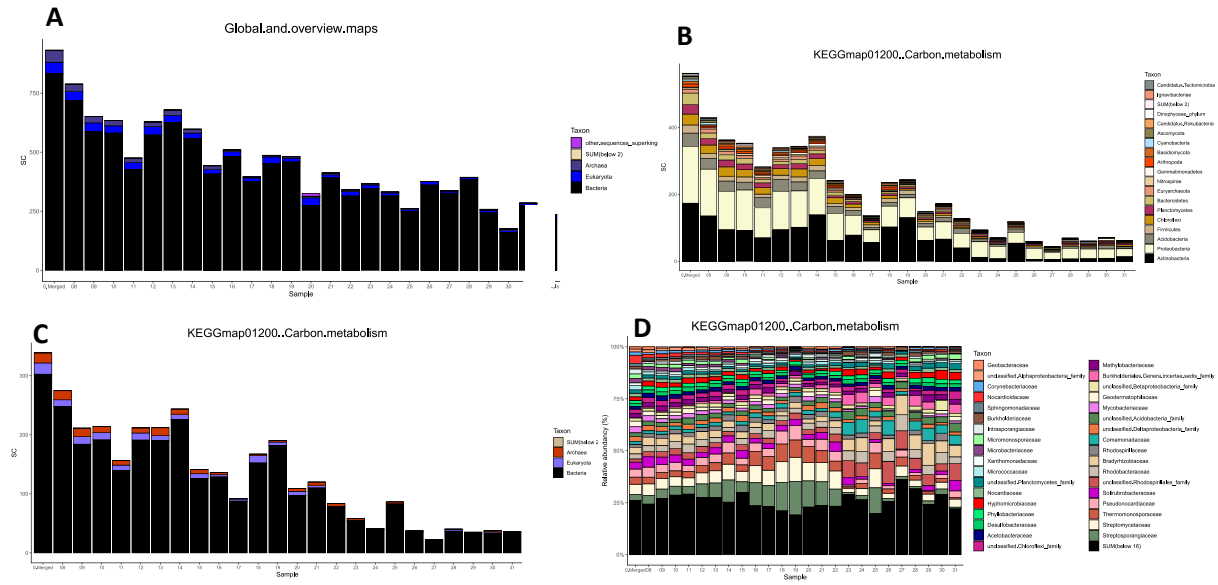
Supplementary Figure 1: Analysis of the regular deposit pattern of the layers of the sediment core, dating in years and depth in centimeters from surface.



Supplementary Figure 2: Geochemical element measurements of the soil core in µg per g of soil.



Supplementary Figure 3. Heatmap of differentially abundant phyla (Wilcoxon test, adjusted p-value 0.05, abundancy difference above 20%) between layer-clusters 1 and 4.



Supplementary Figure 4. **Carbon metabolism** pathway. **Panel A**: KEGG pathway level 2 « Global and overview maps » corresponding to the « Carbon metabolism » map01200 at superkingdom level. **Panel B**: KEGG pathway map01200 at superkingdom level indicating a greater decrease of this metabolism with depth compared to « Global and overview maps ». **Panel C**: KEGG pathway map01200 at phylum level, detailing phyla contributing to the KEGG pathway map. **Panel D**: KEGG pathway map01200 at family level, detailing relative contributions of families to the KEGG pathway map. For example, the Nocardioidaceae family contributes more near the surface whereas the Streptosporangiaceae contribution is maximized in intermediate layers. Stacked barplots are absolute for Panels A, B, C and relative for Panel D.

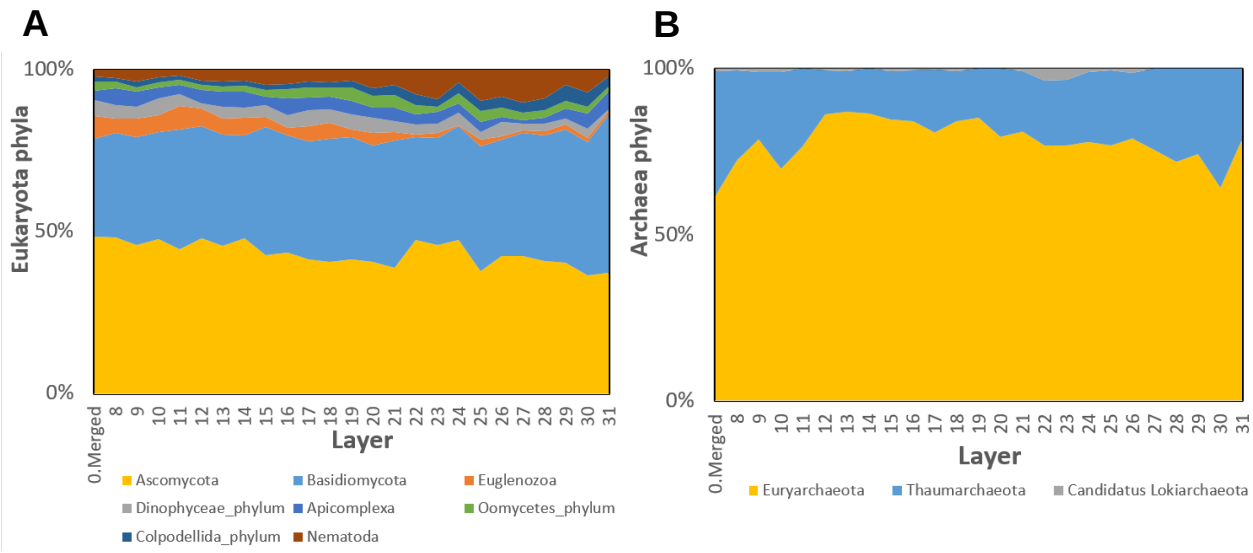


Figure S5. 100% stacked area charts of Eukaryota (A) and Archaea (B) classified phyla.

SUPPLEMENTARY TABLES

Supp. table I. Geochemistry element concentrations.

SampleID	Sample	Mean depth (cm)	Year	Na	Mg	Al	P	K	Ca	Ti	V	Cr	Mn	Fe	Co
1	B18-3_1	1.42	undated	2122.67	5777	44744.38	3245	11577.73	170859.5	3060.69	63.72	103.35	613.78	30222.62	9.3
2	B18-3_2	4.58	undated	2123.46	5850.09	44617.05	3276.13	11471.33	172685.8	3080.8	64.12	106.08	631.88	30203.17	9.25
3	B18-3_3	7	undated	2073.57	5724.55	43365.81	3262.76	11280.18	173138.1	3017.37	62.08	102.79	620.61	29376.67	9.61
4	B18-3_4	9.92	undated	2149	5894.07	45388.56	3335.05	11588.32	173386.3	3092.81	63.96	101.3	647.28	30439.24	9.17
5	B18-3_5	13.08	undated	2130.36	5937.93	44888.56	3251.24	11439.85	172834.3	3054.52	63.41	104.76	610.89	30354.5	9.31
6	B18-3_6	16	undated	2171.3	5900.83	44378.1	3269.2	11306.12	173830.9	3118.72	63.99	105.08	598.19	30389.12	9.2
7	B18-3_7	18.92	1986	2256.47	6034.9	46598.49	3233.19	11374.14	177673.1	3169.34	66.31	116.15	642.3	31491.48	9.33
8	B18-3_8	21.83	1984.5	2087.83	5921.85	44275.99	3200.1	10498.74	189965.3	3043.5	67.51	130.51	663.37	30187.19	9.96
9	B18-3_9	24.75	1983	2167.34	5830.18	43632.33	3246.66	10614.34	195888.3	3076.39	65.01	140.39	637.97	30279.9	10.01
10	B18-3_10	27.67	1981.5	2206.81	6151.35	47507.91	3074.25	11246.27	181869.7	3225.41	72.64	157.67	656.54	32245.86	10.47
11	B18-3_11	30.58	1980	1975.66	5302.45	35363.03	3177.86	9068.38	207360.5	2739.38	54.43	144.56	584.89	26833.3	8.59
12	B18-3_12	33.5	1978.5	2172.13	6183.52	43901.32	3517.27	10209.24	193604.7	3059.84	66.67	192.04	616.79	30269.49	10.29
13	B18-3_13	36.42	1977	1940.94	5626.05	37265.87	3717.99	9226.31	199786.1	2590.18	60.34	199.54	582.72	27888.89	9.64
14	B18-3_14	39.33	1975.5	1877.33	6021.8	39043.62	3685.45	9076.01	183766.2	2576.87	63.72	228.84	569.51	27350.49	10.36
15	B18-3_15	42.25	1974	1863.98	5129.74	32590.3	2892.62	8725.7	203441.2	2362.17	51.09	180.92	511.38	24932.56	8.55
16	B18-3_16	43.67	1971	1955.96	5266.78	32415.38	3139.68	8485.8	199521.3	2300.35	50.69	211.71	530.03	25293.79	9.07
17	B18-3_17	46.33	1969.35	1933.46	5849.87	36364.57	3284.23	8998.38	201201	2409.75	55.48	240.16	559.42	27336.96	9.82
18	B18-3_18	48.25	1968.25	2102.77	6095.06	38591.48	3446.59	9506.56	187479	2547.41	60.31	278.22	564.14	28124.52	11.37
19	B18-3_19	50.25	1967.15	2344.02	6401.85	44855.54	3307.61	10567.75	168013.8	2753.35	68.26	303.23	600.88	31433.1	11.38
20	B18-3_20	52.5	1965.775	2151.06	5891.16	42158.99	3077.93	9787.31	174839.2	2690.95	65.03	258.77	516.88	29137.97	10.63
21	B18-3_21	54.75	1964.675	2121.13	5918.17	42530.19	3178.58	10298.76	195929.3	2887.84	62.75	263.02	521.22	30265.87	10.2
22	B18-3_22	57	1963.3	2170.28	5940.44	43608.92	3291.21	10438.67	184559	2881.49	64.12	304.44	529.54	30685.76	11.22
23	B18-3_23	59.75	1962.3	2198.58	6082.14	45585.16	3151.84	11013.52	200060	2863.89	62.44	237.5	612.18	32144.74	12.15
24	B18-3_24	63	1960.55	2110.89	5586.69	43448.56	2955.34	10567.09	190105	2697.45	64.98	221.2	590.79	31120.47	12.05
25	B18-3_25	65.75	1959	2278.08	6063.6	47566.97	3285.94	11236.18	174092.9	2842.91	69.9	286.75	631.55	33595.3	13.55
26	B18-3_26	68.5	1956.5	1940.29	5166.77	40514.04	3296.97	9840.49	179082.7	2414.77	62.31	232.71	539.8	28837.92	12.09
27	B18-3_27	71.5	1953.5	1841.91	4880.28	38890.68	3219.59	9632.61	177708.8	2293.14	62.75	221.56	556.94	29037.86	12.16
28	B18-3_28	74.5	1950.5	2129.1	5586.52	46445.14	2758.81	11031.9	174574.5	2578.28	70.36	205.08	585.96	31440.56	13.21
29	B18-3_29	77.5	1947.5	2184.98	5703.41	48734.58	2280.43	11676.77	175760.4	2780.2	72.55	161.03	627.95	32158.83	13.29
30	B18-3_30	80.5	1944.5	2147.92	5588.53	47040.06	2205.9	11450.43	182633.7	2793.51	71.97	136.04	607.14	30071.88	13.2
31	B18-3_31	84.25	1941.5	2437.19	6297.27	51916.74	2243.77	12465	216898.8	3080.37	63.94	108.89	637.3	33026.08	11.57

SampleID	Ni	Cu	Zn	As	Rb	Sr	Mo	Ag	Cd	Sb	Cs	Ba	Tl	Pb
1	34.96	148.47	482.16	9.98	68.49	344.19	1.34	6.11	5.35	2.27	7.26	302.03	0.73	130.47
2	34.71	109.43	513.28	10.33	67.78	353.59	1.48	6.98	5.96	2.34	7.06	316.9	0.7	139.77
3	34.95	104.69	489.51	9.98	66.53	349.39	1.42	6.38	5.44	2.28	6.94	294.64	0.68	129.99
4	32.96	104.19	491.68	9.77	68.06	352.01	1.31	6.33	5.53	2.23	7.17	313.62	0.7	132.03
5	34.14	107.74	502.72	9.15	67.16	353.15	1.62	6.59	5.51	2.16	7.19	304.83	0.67	135.09
6	34.18	108.5	512.67	9.44	67.47	355.63	1.48	6.85	5.59	2.22	6.97	315.92	0.64	140.92
7	36.19	118.99	574.46	9.98	66.27	364.41	1.46	7.67	7.34	2.66	7.15	328.51	0.66	163.27
8	40.1	138.17	622.71	10.93	67.02	398.41	1.6	8.35	8.67	2.68	6.95	342.21	0.64	175.88
9	42.49	149.25	648.05	11.73	63.33	400.24	1.93	9.25	9.86	2.5	6.82	336.7	0.66	178.32
10	42.34	156.29	686.41	11.72	69.59	378.54	1.92	9.79	10.84	2.64	7.51	358.86	0.69	184.34
11	35.73	150.01	654.74	11.33	52.23	423.42	1.69	8.02	10.98	2.34	5.26	370.22	0.54	192.77
12	45.43	186.52	809.2	14.34	62.69	410	2.49	12.07	13.72	3.01	7.11	399.65	0.66	233.29
13	66.65	192.71	879.83	13.49	54.77	422.68	2.17	11.77	15.33	2.77	5.72	399.67	0.6	239.71
14	53.73	221.26	991.76	12.93	59.18	401.45	2.84	14.3	17.89	3.1	6.1	414.83	0.62	269.25
15	40.79	179.5	843.48	10.93	51.94	405.1	1.64	9.3	16.66	2.98	5.16	411.72	0.52	235.58
16	43.78	226.53	902.58	15.46	50.35	410.99	1.96	12.13	20.49	3.42	5.05	477.31	0.55	266.43
17	47.22	243.24	1027.43	15.57	54.98	415.05	1.95	15.59	24.5	3.78	5.78	510.23	0.61	346.66
18	49.24	270.65	1131.98	17	58.42	398.24	2.36	18.18	27.52	4.47	6.22	566.46	0.65	395.03
19	52.09	275.14	1186.93	17.57	69.21	373.21	1.95	18.33	29.51	4.43	7.3	586.6	0.72	376.63
20	50.01	254.33	1115.2	16.49	66.27	388.18	1.78	16.67	28.83	4.57	7.49	537.17	0.73	360.9
21	47.47	247.54	1112.73	16.74	63.26	403.3	1.69	18.48	29.38	4.72	7.44	554.33	0.7	383.57
22	52.55	293.46	1196.05	17.3	64.42	398.58	1.75	20.53	30.32	5.1	7.14	606.61	0.69	411.47
23	48.33	258.78	1119.93	19.91	62.7	425.71	1.59	15.86	26.64	4.88	6.78	546.47	0.67	355.86
24	51.43	255.02	1143.21	19.36	62.95	410.8	1.68	14.42	24.29	5.29	6.78	580.26	0.74	338.75
25	50.17	303.9	1298.49	23.7	68.08	400.3	1.63	18.55	24.87	5.57	7.22	674.97	0.74	442.49
26	47.59	286.44	1317.91	20.86	61.82	412.43	1.59	13.88	22.45	5.1	6.87	622.98	0.71	398.92
27	45.32	259.83	1278.44	23.57	58.96	407.67	1.48	13.42	16.47	6.01	6.82	631.67	0.71	411.28
28	47.13	271.58	1330.16	24.4	69.07	389.92	1.7	13.31	13.17	6.39	7.89	644.13	0.78	413
29	45.43	211.83	1201.39	20.7	71.1	369.72	1.45	11.08	9.46	5.91	8.01	575.46	0.8	376.85
30	44.07	198.26	1088.78	20.4	69.28	379.03	1.3	10.41	7.48	5.44	7.47	525.01	0.74	289.09
31	35.79	182.67	914.31	18.38	61.85	443.73	1.08	7.08	6.06	4.42	6.74	463.17	0.68	251.84

Discussion

2.1 La métaprotéomique et la métagénomique main dans la main

De nos jours, la métaprotéomique est couramment utilisée pour analyser les microbiotes car elle permet de caractériser fonctionnellement ainsi que taxonomiquement les organismes présents, d'identifier des mécanismes de réponses métaboliques ou des changements de communautés microbiennes dans un environnement donné. Pour les échantillons plus complexes tels que les sols, le développement de stratégies d'interprétation de spectres est essentiel afin de maximiser les résultats que l'on peut extraire. Jusqu'à très récemment, un assemblage métagénomique est utilisé pour l'analyse métaprotéomique, car une base de données recensant l'ensemble des séquences de protéines potentiellement présentes dans l'échantillon peut être déduite des données métagénomiques acquises sur le même échantillon. La qualité de cette base de données ainsi que sa taille sont essentielles à la qualité de l'interprétation en métaprotéomique. En effet, la méthodologie utilisée pour construire la base de données, la complétude, l'exactitude et la représentativité des séquences ainsi que la taille de la base de données sont des choix critiques et vont impacter directement l'interprétation des données de spectrométrie de masse. Un organisme ou des protéines absentes de la base de données utilisée ne pourront être identifiés. Toutefois, en métaprotéomique se pose la question de l'espace de recherche à considérer qui influe fortement sur le résultat de l'interprétation. Notamment, la taille des bases de données de métaprotéomique n'a aucune commune mesure avec la taille des bases de données de protéomique où seul un organisme est à analyser. Dans le cas d'échantillons complexes, l'utilisation de requêtes en cascades utilisant différentes origines de séquences protéiques permet d'obtenir une base de données spécifiques à l'échantillon, beaucoup plus adaptée à la réalité et maximisant son interprétation. Un métagénome avec une profondeur de séquençage élevée permet une meilleure couverture en lectures des génomes d'organismes assemblés jusqu'à produire des MAGS lorsque les données le permettent. De nouvelles technologies telles que le séquençage HiFi de Pacbio avec une précision supérieure à 99% permet d'obtenir des lectures longues de 13.5kb de longueur moyenne (Wenger et al., 2019), tout en combinant des avantages des séquençages de deuxième et troisième génération. Par exemple, en utilisant la technologie HiFi, les génomes assemblés sont moins fractionnés qu'en utilisant la technologie short reads (Castinel et al., 2022) jouant directement sur la complétude des protéomes théoriques présents dans les bases de données. L'identification de MAGS à partir de cette technologie peut fournir un grand nombre de génomes dans des échantillons de fèces (Bickhart et al., 2022) ou la référence ZymoBIOMICS TruMatrix de fecès (Portik et al., 2022). De même, les assemblages hybrides permettrait éventuellement une meilleure représentativité des protéomes des organismes ou le développement du séquençage dit sur cellule unique (« single cell ») d'échantillons complexes. Ainsi les efforts de séquençage et les nouvelles technologies pour le séquençage profond vont accroitre les informations disponibles dans les bases de données, permettant d'augmenter le niveau d'interprétation des données métaprotéomiques. Toutefois, le problème des tailles des bases de données sera exacerbé et les méthodes d'interprétation devront s'adapter pour en tirer le meilleur parti. Les travaux réalisés lors de cette thèse ont permis d'ouvrir une nouvelle voie, montrant que l'alliance entre bases de données dédiées et bases de données généralistes permet une interprétation améliorée des résultats de spectrométrie de masse pour la métaprotéomique.

A l'instar de la base de données qui est un point critique des analyses métaprotéomiques de sol, on peut utiliser la métaprotéomique pour juger de la qualité de cette base de données en utilisant différents marqueurs tel que le taux de spectres interprétés, le taux de spectres annotés taxonomiquement et/ou fonctionnellement. L'assemblage métagénomique est une question ardue associée à un algorithme souvent basé sur un graphe de De Bruijn demandant un serveur de calcul puissant. De nombreux outils permettent de réaliser l'assemblage mais les paramètres permettant de juger de la qualité de l'assemblage sont les mêmes que ceux utilisés en génomique et transcriptomique comme le N50 ou le pourcentage de lectures alignés sur le métagénome assemblé avec tout de même la fraction du génome assemblé. Lorsque les organismes ne sont pas séquencés, il est plus difficile d'évaluer la complétude de l'assemblage métagénomique. Des critères fonctionnels ou de complétude de protéomes pourrait être mis en place pour juger de la qualité d'un métagénome tels que ça a été précédemment fait en protéogénomique en calculant le ratio de séquences peptidiques consensuelles entre les différents benchmarks d'assemblages (Cogne et al., 2020) par exemple. La technologie de séquençage HiFi étant récente, des collaborations avec des plateformes disposant de cette technologie pourrait être envisagées afin d'optimiser la méthodologie d'assemblage dans le cas d'échantillons complexes. La même démarche peut s'appliquer pour les annotations des génomes assemblés, les logiciels d'annotation des gènes étant pour l'heure entrainés sur des règles classiques de biologie moléculaire issues des travaux sur *Escherichia coli*. L'étude de microorganismes plus atypiques a démontré l'existence parfois de règles alternatives, comme par exemple les démarrages de traduction qui ne sont pas systématiquement faites par le codon ATG (Baudet et al., 2010). Les données de métaprotéomiques pourraient donc servir à mieux identifier les séquences codantes (CDS) dans les métagénomes. L'identification taxonomique et fonctionnelle en métaprotéomique en serait améliorée et permettrait de développer de nouvelles stratégies d'évaluation de la qualité de métagénomes et de métaprotéomes. Celles-ci pourraient par exemple permettre d'évaluer sur des critères du vivant de nouveaux algorithmes d'assemblages que ceux-ci soient au niveau nucléotidique (metaSPADES, Megahit) ou protéique (PLASS). La métaprotéomique des sols est également un enjeu technique complexe, du laboratoire jusqu'aux serveurs.

Une comparaison des données de métaprotéomiques au niveau multi-laboratoires a été menée par la communauté CAMPI (Van Den Bossche et al., 2021). Les différences observées au niveau des peptides disparaissent en partie au niveau des groupes de protéines et s'atténuent au niveau des profils fonctionnels obtenus sur des échantillons de fèces et un mix connu. En effet, 3.4% des peptides sont communs entre 5 préparations d'échantillons et leurs analyses alors que 34.6% des groupes protéines sont communs au 5 (Van Den Bossche et al., 2021). Cette disparité au niveau de l'analyse a été mesurée par ce consortium mais il serait bien d'intégrer des échantillons plus complexes tels que les échantillons de sols et d'océans où le traitement informatique jouent un grand rôle dans l'obtention des résultats. Quel est l'impact d'autres outils d'interprétation ? Dans le cadre de ma thèse, le logiciel Mascot a été principalement utilisé malgré des tests utilisant X !Tandem pour l'annotation taxonomique d'échantillons de sol ou plus simple tel que des échantillons mono-organismes. L'utilisation d'autres pipeline nécessite un travail important au niveau informatique puisque tous les outils ne sont pas développés pour gérer des bases de données avec des millions de séquences et sous des systèmes d'exploitation tel que Windows. Un échantillon bien caractérisé pourrait être utilisé avec différents pipelines d'analyses pour comparer ces outils (benchmark) et améliorer les futures annotations.

Les échantillons de sol constituent une matrice complexe riche en organismes et en molécules interférentes. Contrairement aux échantillons d'eau où les micro-organismes doivent être concentrés, la densité microbienne dans le sol est élevée et seulement quelques mg sont nécessaires pour une analyse exhaustive de métaprotéomique. Pour le sol, le nombre d'organismes est très élevé mais aussi

la diversité des microorganismes présents est impressionnante, induisant une diversité incroyable de molécules. Ceci entraine un problème de co-élution des ions correspondant aux peptides produits à partir de ce type d'échantillons, et un besoin d'optimisation des paramétrages de chromatographie et de spectrométrie de masse.

Pour analyser la diversité des échantillons de sols, il existe différentes méthodologies. La première est d'analyser sans *a priori* les organismes ; la seconde est de réduire la diversité afin de se focaliser sur certaines caractéristiques de la communauté. Celle utilisée pendant la thèse est de séquencer et de digérer l'ensemble du signal sans étape de culture et qui permet d'avoir une image de la diversité globale de l'échantillon. Cette démarche est indispensable pour avoir une vision non biaisée de l'ensemble de la communauté microbienne. Toutefois, nous avons mis en avant par cette approche des microrganismes spécifiques qui mériteraient d'être mieux caractérisés. Dans certains cas, ces microorganismes n'ont pas été encore isolés et ne sont connus que par la séquence de leur génome et l'identification de quelque unes de leurs protéines. Dans ce travail de thèse, il a été mis en évidence l'intérêt d'adjoindre aux analyses de métaprotéomique des données de culturomique qui a pour objectif d'avoir un maximum d'isolats microbiens. Cette méthode est expérimentalement très lourde puisque des centaines de micro-organismes sont mis en culture, et une fois que les colonies sont isolées, elles sont successivement analysées par 16S rRNA metabarcode ou protéotypage de type MALDI-TOF ou protéotypage par spectrométrie de masse en tandem tel que développé dans l'équipe de recherche du Li2D qui s'apparente plus à de la protéomique classique. En effet, cette dernière méthodologie permet de caractériser les organismes analysés et les identifiés au niveau taxonomique. Une voie intermédiaire est d'enrichir certains microorganismes par culture sélective pour analyser des consortia microbiens simplifiés. Pour cela, les micro-organismes pourraient être mis en culture dans des milieux de culture sélectifs de certains clades afin de mieux les caractériser, notamment afin de mieux les prendre en compte du point de vue fonctionnel et taxonomique. L'étape de culture et cette sélection entraîne la perte d'une grande partie du signal au profit d'une analyse plus fine des éléments sélectionnés. Ce type de dé-complexification permet d'amplifier certaines activités de la communauté microbiennes telles que les micro-organismes évoluant dans des milieux anaérobies, ou sur certains types de sucres ou évoluant dans des milieux avec une certaine concentration de contaminants/antibiotiques, ou pour notre intérêt sur les métaux, des incubations sélectives avec de tels produits. Par exemple, Vanoukian discutent des possibilités de cultures des bactéries « incultivable » (Vartoukian et al., 2010).

Concernant le problème de co-élution des ions lors de la chromatographie en amont du couplage avec le spectromètre de masse en tandem à haute résolution, ce problème pourrait être atténué en utilisant la mobilité ionique. Cette technologie permet de fractionner les peptides selon une séparation ionique des différentes formes chargées sous l'influence d'un champ électrique. Plusieurs versions existent selon l'instrumentation : la spectrométrie de mobilité ionique à tube de dérive (DTIMS) et la spectrométrie de mobilité ionique à forme d'onde asymétrique à haut champ (FAIMS). La mobilité ionique permet en principe de séparer les peptides co-élués limitant ensuite la co-fragmentation des peptides. Ainsi l'utilisation de la mobilité ionique peut améliorer considérablement la précision et l'exhaustivité des analyses protéomiques multiplexées (Pfammatter et al., 2016). Lors de la fragmentation de chaque ion du top n sélectionné par le spectromètre de masse, plusieurs ions vont être isolés au lieu d'un et vont produire un spectre MS/MS correspondant à plusieurs ions. Ces spectres ne résultant pas de la fragmentation d'un seul peptide auront des scores faibles et leur interprétation en peptides en sera compliquée. Le développement de protocoles basés sur la mobilité ionique proposée entre autres par le spectromètre TimsTOF de Bruker permet d'intégrer une quatrième dimension en plus du temps de rétention, la masse sur charge et l'intensité des ions. La mobilité ionique dépend de la charge de l'ion et de sa friction dans le gaz tel un "effet parachute" où lorsque le

parachute est déployé, la masse ne change pas mais permet de ralentir la chute, la conformation de l'ion va impacter sa vitesse. Les peptides isomères seront ainsi différenciés et ne seront plus co-élués. A ma connaissance, ces méthodes n'ont pas été encore déployées dans le cadre de la métaprotéomique mais mériteraient d'être optimisées pour améliorer les résultats.

La présence de molécules interférentes tels que les composés humiques va impacter la séparation des ions dans la pré-colonne et la colonne résultant en un encrassage prématuré de ces éléments vitaux en amont du spectromètre de masse. La purification des peptides et la suppression de ces composés nécessite le développement de protocoles. Des efforts d'optimisation de la préparation des échantillons de sols pour la métaprotéomique sont donc nécessaires mais doivent être testés sur de nombreux échantillons, chaque sol ayant ses propres caractéristiques. Au niveau bioinformatique, la reproductibilité entre réplica peut être délicate avec un biais d'identification pour les protéines les plus abondantes et les plus conservées de l'échantillon comme le métabolisme central majoritaire et les espèces les plus abondantes. L'identification au niveau des protéines produisent des matrices éparses qui nécessitent des post-traitements vers des niveaux d'identifications plus élevés telles que les niveaux d'annotations fonctionnelles ou des rangs taxonomiques élevés. L'analyse d'une multitude d'échantillons d'un continuum d'un site par métaprotéomique a nécessité l'utilisation d'une base de données commune tel que développé dans la deuxième partie de la thèse afin d'éviter des incohérences d'identifications liées au nombre élevé de données manquantes, basiquement des zéros dans les matrices d'identification. Ces données manquantes peuvent être imputées mais seulement si un nombre de réplica biologiques ou techniques suffisant est enregistré permettant d'imputer avec confiance les données.

2.2 Les perspectives de l'étude

Des données métagénomiques et métaprotéomiques ont été acquises sur des données de talus et de laisses de crues prélevés lors de la crue exceptionnelle de la Seine en juin 2016. Ces données pourraient être explorées au niveau taxonomique et fonctionnel et comparées entre eux et avec la carotte de sédiment. Ce sont des échantillons du même bassin et de composition pédologique proche. Les contaminations mesurées dans ces échantillons sont différentes dans le sens où les échantillons de laisses sont issus de l'érosion actuelle des sols, mais aussi de l'effet de purge des petites rivières urbaines très contaminées, et de leur dépôt dans les jours suivant la crue (Le Gall et al 2018). Cet échantillon frais permet de retracer les pollutions actuelles telle que l'antimoine, un contaminant réémergent (Ayrault et al., 2013), et les contaminations plus anciennes liées aux sols érodés par l'eau. La comparaison et une étude approfondie pour positionner ces échantillons par rapport aux couches de la carotte pourrait être intéressant. Nous pourrions également analyser des sédiments prélevés au niveau de l'estuaire, endroit privilégié de dépôt important de sédiments contaminés. Malheureusement, le temps imparti pour la réalisation de cette thèse n'a pas permis de nous plonger dans l'analyse de ces échantillons d'intérêt.

Pour une analyse plus fine des relations contaminants et communautés microbiennes, il serait intéressant d'intégrer la méthodologie utilisée en écotoxicologie permettant d'étudier la résilience des organismes en faisant varier des paramètres fixes (Oliver et al., 2015). Par exemple, le suivi de la communauté microbienne des sédiments avec ajout croissant d'un ou plusieurs contaminants présents « naturellement » et le suivi de l'évolution de certaines fonctions. Ces résultats pourraient permettre de modéliser la résilience de la communauté microbienne actuelle face aux concentrations de contaminants qui pourrait croître avec les activités humaines. De telles expériences peuvent être désormais imaginées sur la base des résultats obtenus lors de l'étude de la carotte de Bouafles.

A l'échelle globale du bassin, la Seine est un bassin calcaire où les concentrations en éléments traces métalliques (ETM) sont faibles. La Seine peut apparaître comme un fleuve peu contaminé lorsqu'on compare les concentrations avec d'autres fleuves européens (Spree en Allemagne, la Zenne en Belgique et le Lambro en Italie ; Ayrault et al, 2020). Mais si on utilise les facteurs d'enrichissement, qui permettent de tenir compte de la concentration naturelle des ETM, la Seine se révèle être l'un des fleuves les plus contaminés d'Europe. Les ETM sont des éléments naturels, ils sont présents dans les sédiments à des concentrations variées suivant la géologie du bassin. Ainsi, comparer les concentrations des ETM peut conduire à des erreurs d'interprétation. Il serait intéressant de comparer de multiples carottes de sols au niveau métagénomique et métaprotéomique avec systématiquement un échantillon situé en amont et en aval du bassin et de nature pédologique proche permettant de limiter les effets batch et pouvoir estimer la diversité taxonomique et fonctionnelle des microbiotes à une échelle entre des bassins qui relèvent d'histoire de contaminations différentes. Ces données pourraient également participer à l'amélioration des bases de données de séquences des organismes spécifiques à ces environnements. En incluant des sols ayant des caractéristiques proches comme le taux d'argiles dans d'autres zones sédimentaires de grands fleuves européens, l'analyse géochimique montre que l'histoire de la population peut être retracée à travers l'analyse des sédiments (Ayrault et al., 2020). En fixant le paramètre de type de sol, on pourrait comparer les communautés microbiennes. Celles-ci sont peu fluctuantes à l'échelle d'une même carotte mais pourraient être plus variées entre différents sites d'un même bassin et inter-bassins.

En termes de développements bioinformatiques, les capacités grandissantes de calcul des ordinateurs permettent d'envisager des projets de grande ampleur nécessitant des téraoctets de mémoire vive et de stockage. Plusieurs bases de données publiques permettent de stocker et rendre accessible des données de spectrométrie de masse d'échantillons de sols issus du monde entier, d'écosystèmes complexes, mais aussi d'organismes isolés. Il serait envisageable de développer des algorithmes dit d'apprentissage (« machine learning ») basés sur ces données massives qui ne dépendent pas de bases de données de spectres MS/MS ou de séquences protéiques tel que fait par Feng (Feng et al., 2021). L'algorithme pourrait apprendre sur des échantillons simples et complexes à identifier des spectres MS/MS par interprétation *de novo* à partir de ces connaissances sur le milieu. Cela permettrait de s'affranchir des problèmes de taille de bases de données et d'estimation du taux de fausse découverte (FDR) tout en intégrant des notions de co-élution possible des peptides. Une autre possibilité gourmande en capacité de calcul est de développer les bases de données de spectres MS/MS pour des données de métaprotéomiques basées sur la génération théorique des spectres de bases telles que à partir des catalogues de gènes de sols par exemple. A ma connaissance, un outil basé sur la DIA est développé par le laboratoire de Laura Elo (Pietilä et al., 2022) qui utilise soit les spectres MS/MS des analyses DDA pour l'identification en DIA des mêmes échantillons et récemment intégrant une voie utilisant uniquement des données DIA pour des microbiotes. Il pourrait y avoir un développement approfondi axé environnement utilisant les données publiques comme source d'informations pour l'algorithme. Une analyse plus globale des données de masse environnementale générées permettrait éventuellement d'identifier des caractéristiques propres, par exemple, certains types de modifications post-traductionnelles plus abondantes à certains seuils de temps de rétention ou de caractéristiques environnementales de l'échantillons. Enfin, une des voies à explorer pour faciliter l'interprétation des données obtenues en métaprotéomique concerne l'intégration des données taxonomiques et fonctionnelles dans le contexte biologique. Pour cela, de nouveaux concepts d'interprétation et de nouvelles bases de données sur les microbiotes seraient à développer.

Conclusion et perspectives

En conclusion, ce projet de thèse a permis d'améliorer considérablement l'interprétation des données métaprotéomiques de sols et d'explorer d'un point de vue métaprotéomique, métagénomique et géochimique (métalomique) une colonne de sol issue de dépôts réguliers de sédiments lors de crues dans un fleuve très anthropisé : la Seine. Le premier axe d'amélioration a été d'augmenter le taux d'interprétation des spectres MS/MS acquis par spectrométrie à haute résolution en mettant en place une méthodologie de requêtes en cascades et de combinaison de base de données généraliste, spécifique au type d'échantillons et spécifique à l'échantillon. Cette méthodologie a été saluée par la communauté scientifique et a été publiée par le journal de référence dans le domaine. L'analyse de la carotte de sédiment est particulièrement riche en nouvelles connaissances et a permis de faire des corrélations entre certains microorganismes et certains métaux. Il est important désormais de définir si ces corrélations sont fortuites ou liées à un aspect physiologique de ces microorganismes. Des analyses de microbiologie pourraient être menées afin de sélectionner et analyser par métaprotéomique des groupes d'organismes résistants à des concentrations croissantes de contaminants. L'intégration de données géochimiques concernant d'autres contaminants que les ETM précédemment obtenues sur le site de Bouafles pourrait également être explorée. Par exemple, les données sur les molécules chlorées et sur les antibiotiques qui ont été mesurées sur des carottes prélevées sur le même site de Bouafles n'ont pu être explorées pendant le laps de temps de la thèse. Les échelles de temps, les données manquantes rendent cette tâche difficile. L'analyse d'une carotte supplémentaire ayant des caractéristiques pédologiques similaires permettrait de compléter ces données géochimiques mais le nombre de réplicas de carotte nécessaires pour ce type de comparaison et validations statistiques se pose. Au vu de l'effort d'analyses réalisées sur une carotte, le répéter sur plusieurs carottes de façon identique est un projet dantesque à l'heure actuelle, mais pourrait être envisagé désormais à moyen terme. La recherche en cours d'un nouveau site remplaçant le site de Bouafles, qui a été labouré rendant impossible la datation des dépôts les plus récents, est complexe et permettrait d'avoir une vision sur les contaminants plus récents ou le devenir récent des polluants historiques qui n'ont pas pu être différenciés dans l'étude réalisée.

Le travail effectué au cours de cette thèse a donc abouti à améliorer le pipeline d'interprétation de données métaprotéomique de sols et a permis de montrer la faisabilité de la métaprotéomique et ses nombreux atouts quand elle est alliée à des analyses de géochimie.

Références bibliographiques

Aakko, J., Pietilä, S., Suomi, T., Mahmoudian, M., Toivonen, R., Kouvonen, P., Rokka, A., Hänninen, A., & Elo, L. L. (2020). Data-independent acquisition mass spectrometry in metaproteomics of gut Microbiota-implementation and computational analysis. *Journal of Proteome Research*, *19*(1), 432–436. https://doi.org/10.1021/acs.jproteome.9b00606

Abiraami, T. V., Singh, S., & Nain, L. (2020). Soil metaproteomics as a tool for monitoring functional microbial communities: promises and challenges. *Re/Views in Environmental Science and Bio/Technology*, *19*(1), 73–102. https://doi.org/10.1007/s11157-019-09519-8

Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, *537*(7620), 347–355. https://doi.org/10.1038/nature19949

Akhmedov, M., Martinelli, A., Geiger, R., & Kwee, I. (2020). Omics Playground: a comprehensive self-service platform for visualization, analytics and exploration of Big Omics Data. *NAR Genomics and Bioinformatics*, *2*(1), lqz019. https://doi.org/10.1093/nargab/lqz019

Andreotti, S., Klau, G. W., & Reinert, K. (2012). Antilope--a Lagrangian relaxation approach to the de novo peptide sequencing problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(2), 385–394. https://doi.org/10.1109/TCBB.2011.59

Arenella, M., Giagnoni, L., Masciandaro, G., Ceccanti, B., Nannipieri, P., & Renella, G. (2014). Interactions between proteins and humic substances affect protein identification by mass spectrometry. *Biology and Fertility of Soils*, *50*(3), 447–454. https://doi.org/10.1007/s00374-013-0860-0

Ayrault, S., Le Pape, P., Evrard, O., Priadi, C. R., Quantin, C., Bonté, P., & Roy-Barman, M. (2014). Remanence of lead pollution in an urban river system: a multi-scale temporal and spatial study in the Seine River basin, France. *Environmental Science and Pollution Research International*, *21*(6), 4134–4148. https://doi.org/10.1007/s11356-013-2240-6

Ayrault, Sophie, Meybeck, M., Mouchel, J.-M., Gaspéri, J., Lestel, L., Lorgeoux, C., & Boust, D. (2020). Sedimentary archives reveal the concealed history of micropollutant contamination in the Seine River basin. In *The Handbook of Environmental Chemistry* (pp. 269–300). Springer International Publishing.

Ayrault, Sophie, Priadi, C. R., Evrard, O., Lefèvre, I., & Bonté, P. (2010). Silver and thallium historical trends in the Seine River basin. *Journal of Environmental Monitoring: JEM*, *12*(11), 2177–2185. https://doi.org/10.1039/c0em00153h

Ayrault, Sophie, Priadi, C. R., Pape, P. L., & Bonté, P. (2013). Occurrence, sources and pathways of antimony and silver in an urban catchment. In *Urban Environment* (pp. 425–435). Springer Netherlands.

Ayrault, Sophie, Roy-Barman, M., Le Cloarec, M.-F., Priadi, C. R., Bonté, P., & Göpel, C. (2012). Lead contamination of the Seine River, France: geochemical implications of a historical perspective. *Chemosphere*, *87*(8), 902–910. https://doi.org/10.1016/j.chemosphere.2012.01.043

Baas, L. G. M. (1934). *Geobiologie of Inleiding Tot de Milieukunde (W)*. W.P. Van Stockum & Zoon.

Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., Bengtsson-Palme, J., Anslan, S., Coelho, L. P., Harend, H., Huerta-Cepas, J., Medema, M. H., Maltz, M. R., Mundra, S., Olsson, P. A., Pent, M., Põlme, S., Sunagawa, S., Ryberg, M., … Bork, P. (2018). Structure and function of the global topsoil microbiome. *Nature*, *560*(7717), 233–237. https://doi.org/10.1038/s41586-018-0386-6

Ballabio, C., Panagos, P., Lugato, E., Huang, J.-H., Orgiazzi, A., Jones, A., Fernández-Ugalde, O., Borrelli, P., & Montanarella, L. (2018). Copper distribution in European topsoils: An assessment based on LUCAS soil survey. *The Science of the Total Environment*, *636*, 282–298. https://doi.org/10.1016/j.scitotenv.2018.04.268

Barabasz, W., Albińska, D., Jaśkowska, M., & Lipiec, J. (01 2002). Ecotoxicology of Aluminium. *Polish Journal of Environmental Studies*, *11*.

Bastida, F., Algora, C., Hernández, T., & García, C. (2012). Feasibility of a cell separation-proteomic based method for soils with different edaphic properties and microbial biomass. *Soil Biology & Biochemistry*, *45*, 136–138. https://doi.org/10.1016/j.soilbio.2011.10.017

Bastida, F., Hernández, T., & García, C. (2014). Metaproteomics of soils from semiarid environment: functional and phylogenetic information obtained with different protein extraction methods. *Journal of Proteomics*, *101*, 31–42. https://doi.org/10.1016/j.jprot.2014.02.006

Bastida, F., Jehmlich, N., Martínez-Navarro, J., Bayona, V., García, C., & Moreno, J. L. (2019). The effects of struvite and sewage sludge on plant yield and the microbial community of a semiarid Mediterranean soil. *Geoderma*, *337*, 1051–1057. https://doi.org/10.1016/j.geoderma.2018.10.046

Bastida, F., Moreno, J. L., Nicolás, C., Hernández, T., & García, C. (2009). Soil metaproteomics: a review of an emerging

environmental science. Significance, methodology and perspectives. *European Journal of Soil Science*, *60*(6), 845–859. https://doi.org/10.1111/j.1365-2389.2009.01184.x

Bastida, F., Selevsek, N., Torres, I. F., Hernández, T., & García, C. (2015). Soil restoration with organic amendments: linking cellular functionality and ecosystem processes. *Scientific Reports*, *5*(1), 15550. https://doi.org/10.1038/srep15550

Bastida, F., Zsolnay, A., Hernández, T., & García, C. (2008). Past, present and future of soil quality indices: A biological perspective. *Geoderma*, *147*(3–4), 159–171. https://doi.org/10.1016/j.geoderma.2008.08.007

Bastida, Felipe, García, C., von Bergen, M., Moreno, J. L., Richnow, H. H., & Jehmlich, N. (2015). Deforestation fosters bacterial diversity and the cyanobacterial community responsible for carbon fixation processes under semiarid climate: a metaproteomics study. *Applied Soil Ecology: A Section of Agriculture, Ecosystems & Environment*, *93*, 65–67. https://doi.org/10.1016/j.apsoil.2015.04.006

Baudet, M., Ortet, P., Gaillard, J.-C., Fernandez, B., Guérin, P., Enjalbal, C., Subra, G., de Groot, A., Barakat, M., Dedieu, A., & Armengaud, J. (2010). Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of non-canonical translation initiation codons. *Molecular & Cellular Proteomics: MCP*, *9*(2), 415–426. https://doi.org/10.1074/mcp.M900359-MCP200

Bell, T. H., Yergeau, E., Maynard, C., Juck, D., Whyte, L. G., & Greer, C. W. (2013). Predictable bacterial composition and hydrocarbon degradation in Arctic soils following diesel and nutrient disturbance. *The ISME Journal*, *7*(6), 1200–1210. https://doi.org/10.1038/ismej.2013.1

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benndorf, D., Balcke, G. U., Harms, H., & von Bergen, M. (2007). Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *The ISME Journal*, *1*(3), 224–234. https://doi.org/10.1038/ismej.2007.39

Bergauer, K., Fernandez-Guerra, A., Garcia, J. A. L., Sprenger, R. R., Stepanauskas, R., Pachiadaki, M. G., Jensen, O. N., & Herndl, G. J. (2018). Organic matter processing by microbial communities throughout the Atlantic water column as revealed by metaproteomics. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(3), E400–E408. https://doi.org/10.1073/pnas.1708779115

Bhattacharya, P. T., Misra, S. R., & Hussain, M. (2016). Nutritional aspects of essential trace elements in oral health and disease: An extensive review. *Scientifica*, *2016*, 5464373. https://doi.org/10.1155/2016/5464373

Bickhart, D. M., Kolmogorov, M., Tseng, E., Portik, D. M., Korobeynikov, A., Tolstoganov, I., Uritskiy, G., Liachko, I., Sullivan, S. T., Shin, S. B., Zorea, A., Andreu, V. P., Panke-Buisse, K., Medema, M. H., Mizrahi, I., Pevzner, P. A., & Smith, T. P. L. (2022). Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nature Biotechnology*, *40*(5), 711–719. https://doi.org/10.1038/s41587-021-01130-z

Borrego J. Morales M. de la Torre, J. (2002). Geochemical characteristics of heavy metal pollution in surface sediments of the Tinto and Odiel river estuary (southwestern Spain). *Environmental Geology*, *41*(7), 785–796. https://doi.org/10.1007/s00254-001-0445-3

Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinović, S. M., Cheng, L.-Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., Vitek, O., Rinner, O., & Reiter, L. (2015). Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Molecular & Cellular Proteomics: MCP*, *14*(5), 1400–1410. https://doi.org/10.1074/mcp.M114.044305

Burger, T. (2018). Gentle introduction to the statistical foundations of false discovery rate in quantitative proteomics. *Journal of Proteome Research*, *17*(1), 12–22. https://doi.org/10.1021/acs.jproteome.7b00170

Carron, J.-P., & Lagache, M. (1971). La distribution des éléments alcalins Li, Na, K, Rb, dans les minéraux essentiels des granites et granodiorites du sud de la Corse. *Bulletin de la Société française de Minéralogie et de Cristallographie*, *94*(1), 70–80. https://doi.org/10.3406/bulmi.1971.6553

Castinel, A., Mainguy, J., Bouchez, O., Combes, S., Iampietro, C., Gaspin, C., Milan, D., Donnadieu, C., Pascal, G., & Hoede, C. (2021). *Benefits of PacBio HiFi long reads for metagenomic whole genome analysis*. https://hal.archives-ouvertes.fr/hal-03538657

Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*, *69*(2), 330–339. https://doi.org/10.1016/j.mimet.2007.02.005

Chapman, P. M. (2007). Determining when contamination is pollution - weight of evidence determinations for sediments and effluents. *Environment International*, *33*(4), 492–501. https://doi.org/10.1016/j.envint.2006.09.001

Chen, B., Brown, K. A., Lin, Z., & Ge, Y. (2018). Top-down proteomics: Ready for prime time? *Analytical Chemistry*, *90*(1), 110–127. https://doi.org/10.1021/acs.analchem.7b04747

Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J. R., Seshadri, R., Smirnova, T., Kirton, E., Jungbluth, S. P., Woyke, T., Eloe-Fadrosh, E. A., Ivanova, N. N., & Kyrpides, N. C. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research*, *47*(D1), D666–D677. https://doi.org/10.1093/nar/gky901

Chiffres clés 2019 de Voies navigables de France Bassin de la Seine. (2020, May 29). VNF; Voies Navigables de France. https://www.vnf.fr/vnf/brochure-et-lettress/chiffres-cles-2019-de-voies-navigables-de-france-bassin-de-la-seine/

Cho, J.-C., & Tiedje, J. M. (2000). Biogeography and degree of endemicity of Fluorescent Pseudomonas strains in soil. *Applied and Environmental Microbiology*, *66*(12), 5448–5456. https://doi.org/10.1128/aem.66.12.5448-5456.2000

Chorover, J., Kretzschmar, R., Garcia-Pichel, F., & Sparks, D. L. (2007). Soil biogeochemical processes within the critical zone. *Elements (Quebec, Quebec)*, *3*(5), 321–326. https://doi.org/10.2113/gselements.3.5.321

Chourey, K., Jansson, J., VerBerkmoes, N., Shah, M., Chavarria, K. L., Tom, L. M., Brodie, E. L., & Hettich, R. L. (2010). Direct cellular lysis/protein extraction protocol for soil metaproteomics. *Journal of Proteome Research*, *9*(12), 6615–6622. https://doi.org/10.1021/pr100787q

Christie-Oleza, J. A., Fernandez, B., Nogales, B., Bosch, R., & Armengaud, J. (2012). Proteomic insights into the lifestyle of an environmentally relevant marine bacterium. *The ISME Journal*, *6*(1), 124–135. https://doi.org/10.1038/ismej.2011.86

Cogne, Y., Gouveia, D., Chaumot, A., Degli-Esposti, D., Geffard, O., Pible, O., Almunia, C., & Armengaud, J. (2020). Proteogenomics-guided evaluation of RNA-seq assembly and protein database construction for emergent model organisms. *Proteomics*, *20*(10), e1900261. https://doi.org/10.1002/pmic.201900261

Colatriano, D., Ramachandran, A., Yergeau, E., Maranger, R., Gélinas, Y., & Walsh, D. A. (2015). Metaproteomics of aquatic microbial communities in a deep and stratified estuary. *Proteomics*, *15*(20), 3566–3579. https://doi.org/10.1002/pmic.201500079

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, *10*(4), 1794–1805. https://doi.org/10.1021/pr101065j

Crowther, T. W., van den Hoogen, J., Wan, J., Mayes, M. A., Keiser, A. D., Mo, L., Averill, C., & Maynard, D. S. (2019). The global soil community and its influence on biogeochemistry. *Science (New York, N.Y.)*, *365*(6455), eaav0550. https://doi.org/10.1126/science.aav0550

Daffonchio, D., Ferrer, M., Mapelli, F., Cherif, A., Lafraya, A., Malkawi, H. I., Yakimov, M. M., Abdel-Fattah, Y. R., Blaghen, M., Golyshin, P. N., Kalogerakis, N., Boon, N., Magagnini, M., & Fava, F. (2013). Bioremediation of Southern Mediterranean oil polluted sites comes of age. *New Biotechnology*, *30*(6), 743–748. https://doi.org/10.1016/j.nbt.2013.05.006

Delmont, T. O., Robe, P., Cecillon, S., Clark, I. M., Constancias, F., Simonet, P., Hirsch, P. R., & Vogel, T. M. (2011). Accessing the soil metagenome for studies of microbial diversity. *Applied and Environmental Microbiology*, *77*(4), 1315–1324. https://doi.org/10.1128/AEM.01526-10

Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., & Ralser, M. (2020). DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, *17*(1), 41–44. https://doi.org/10.1038/s41592-019-0638-x

Dendievel, A.-M., Grosbois, C., Ayrault, S., Evrard, O., Coynel, A., Debret, M., Gardes, T., Euzen, C., Schmitt, L., Chabaux, F., Winiarski, T., Van Der Perk, M., & Mourier, B. (2022). Key factors influencing metal concentrations in sediments along Western European Rivers: A long-term monitoring study (1945-2020). *The Science of the Total Environment*, *805*(149778), 149778. https://doi.org/10.1016/j.scitotenv.2021.149778

Dendievel, A.-M., Mourier, B., Coynel, A., Evrard, O., Labadie, P., Ayrault, S., Debret, M., Koltalo, F., Copard, Y., Faivre, Q., Gardes, T., Vauclin, S., Budzinski, H., Grosbois, C., Winiarski, T., & Desmet, M. (2020). Spatio-temporal assessment of the polychlorinated biphenyl (PCB) sediment contamination in four major French river corridors (1945–2018). *Earth System Science Data*, *12*(2), 1153–1170. https://doi.org/10.5194/essd-12-1153-2020

Dendievel, A.-M., Mourier, B., Dabrin, A., Delile, H., Coynel, A., Gosset, A., Liber, Y., Berger, J.-F., & Bedell, J.-P. (2020). Metal pollution trajectories and mixture risk assessed by combining dated cores and subsurface sediments along a major European river (Rhône River, France). *Environment International*, *144*(106032), 106032. https://doi.org/10.1016/j.envint.2020.106032

Dequiedt, S., Thioulouse, J., Jolivet, C., Saby, N. P. A., Lelievre, M., Maron, P.-A., Martin, M. P., Prévost-Bouré, N. C., Toutain, B., Arrouays, D., Lemanceau, P., & Ranjard, L. (2009). Biogeographical patterns of soil bacterial communities. *Environmental Microbiology Reports*, *1*(4), 251–255. https://doi.org/10.1111/j.1758-2229.2009.00040.x

Devos, D., & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in Genetics: TIG*, *17*(8), 429–431.

https://doi.org/10.1016/s0168-9525(01)02348-4
Dhal, B., Thatoi, H. N., Das, N. N., & Pandey, B. D. (2013). Chemical and microbial remediation of hexavalent chromium from contaminated soil and mining/metallurgical solid waste: a review. *Journal of Hazardous Materials*, *250–251*, 272–291. https://doi.org/10.1016/j.jhazmat.2013.01.048
Dubey, A., Malla, M. A., Khan, F., Chowdhary, K., Yadav, S., Kumar, A., Sharma, S., Khare, P. K., & Khan, M. L. (2019). Soil microbiome: a key player for conservation of soil health under changing climate. *Biodiversity and Conservation*, *28*(8–9), 2405–2429. https://doi.org/10.1007/s10531-019-01760-5
Dung, T. T. T., Cappuyns, V., Swennen, R., & Phung, N. K. (2013). From geochemical background determination to pollution assessment of heavy metals in sediments and soils. *Re/Views in Environmental Science and Bio/Technology*, *12*(4), 335–353. https://doi.org/10.1007/s11157-013-9315-1
Dupont, C. L., Grass, G., & Rensing, C. (2011). Copper toxicity and the origin of bacterial resistance--new insights and applications. *Metallomics: Integrated Biometal Science*, *3*(11), 1109–1118. https://doi.org/10.1039/c1mt00107h
Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, *34*(14), 2371–2375. https://doi.org/10.1093/bioinformatics/bty113
Elias, J. E., Haas, W., Faherty, B. K., & Gygi, S. P. (2005). Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature Methods*, *2*(9), 667–675. https://doi.org/10.1038/nmeth785
Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, *5*(11), 976–989. https://doi.org/10.1016/1044-0305(94)80016-2
Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., Maman, S., Hernandez-Raquet, G., Combes, S., & Pascal, G. (2018). FROGS: Find, rapidly, OTUs with Galaxy solution. *Bioinformatics (Oxford, England)*, *34*(8), 1287–1294. https://doi.org/10.1093/bioinformatics/btx791
Feig, M. A., Hammer, E., Völker, U., & Jehmlich, N. (2013). In-depth proteomic analysis of the human cerumen-a potential novel diagnostically relevant biofluid. *Journal of Proteomics*, *83*, 119–129. https://doi.org/10.1016/j.jprot.2013.03.004
Feng, S., Sterzenbach, R., & Guo, X. (2021). Deep learning for peptide identification from metaproteomics datasets. *Journal of Proteomics*, *247*(104316), 104316. https://doi.org/10.1016/j.jprot.2021.104316
Fenyö, D., & Beavis, R. C. (2003). A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry*, *75*(4), 768–774. https://doi.org/10.1021/ac0258709
Fernandes, M. L. P., Bastida, F., Jehmlich, N., Martinović, T., Větrovský, T., Baldrian, P., Delgado-Baquerizo, M., & Starke, R. (2022). Functional soil mycobiome across ecosystems. *Journal of Proteomics*, *252*(104428), 104428. https://doi.org/10.1016/j.jprot.2021.104428
Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(3), 626–631. https://doi.org/10.1073/pnas.0507535103
Frank, A., & Pevzner, P. (2005). PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, *77*(4), 964–973. https://doi.org/10.1021/ac048788h
Frieden, E. (1974). The evolution of metals as essential elements (with special reference to iron and copper). *Advances in Experimental Medicine and Biology*, *48*(0), 1–29. https://doi.org/10.1007/978-1-4684-0943-7_1
Froger, C., Quantin, C., Gasperi, J., Caupos, E., Monvoisin, G., Evrard, O., & Ayrault, S. (2019). Impact of urban pressure on the spatial and temporal dynamics of PAH fluxes in an urban tributary of the Seine River (France). *Chemosphere*, *219*, 1002–1013. https://doi.org/10.1016/j.chemosphere.2018.12.088
Gardes, T. (2020). *Reconstruction temporelle des contaminations métalliques et organiques particulaires dans le bassin versant de l'Eure et devenir des sédiments suite à l'arasement d'un barrage.* [Normandie Université]. https://tel.archives-ouvertes.fr/tel-03122674
Gateuille, D., Evrard, O., Lefevre, I., Moreau-Guigon, E., Alliot, F., Chevreuil, M., & Mouchel, J.-M. (2014). Mass balance and decontamination times of Polycyclic Aromatic Hydrocarbons in rural nested catchments of an early industrialized region (Seine River basin, France). *The Science of the Total Environment*, *470–471*, 608–617. https://doi.org/10.1016/j.scitotenv.2013.10.009
Gateuille, D., Evrard, O., Moreau-Guigon, E., Chevreuil, M., & Mouchel, J.-M. (2014). Long term impact of PAH contamination in soils on the water quality in rivers. *EGU General Assembly 2014*, *16*, EGU2014-4051.
Gatto, L., & Christoforou, A. (2014). Using R and Bioconductor for proteomics data analysis. *Biochimica et Biophysica Acta*, *1844*(1 Pt A), 42–51. https://doi.org/10.1016/j.bbapap.2013.04.032
Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., & Bryant, S. H. (2004). Open mass spectrometry search algorithm. *Journal of Proteome Research*, *3*(5), 958–964. https://doi.org/10.1021/pr0499491

Geisen, S., Briones, M. J. I., Gan, H., Behan-Pelletier, V. M., Friman, V.-P., de Groot, G. A., Hannula, S. E., Lindo, Z., Philippot, L., Tiunov, A. V., & Wall, D. H. (2019). A methodological framework to embrace soil biodiversity. *Soil Biology & Biochemistry*, *136*(107536), 107536. https://doi.org/10.1016/j.soilbio.2019.107536

Gellis, A. C., Hupp, C. R., Pavich, M. J., Landwehr, J. M., Banks, W. S. L., Hubbard, B. E., Langland, M. J., Ritchie, J. C., & Reuter, J. M. (2009). Sources, transport, and storage of sediment at selected sites in the Chesapeake bay watershed. In *Scientific Investigations Report*. US Geological Survey. https://doi.org/10.3133/sir20085186

Glass, J. B., Yu, H., Steele, J. A., Dawson, K. S., Sun, S., Chourey, K., Pan, C., Hettich, R. L., & Orphan, V. J. (2014). Geochemical, metagenomic and metaproteomic insights into trace metal utilization by methane-oxidizing microbial consortia in sulphidic marine sediments: Metal micronutrients for anaerobic oxidation of methane. *Environmental Microbiology*, *16*(6), 1592–1611. https://doi.org/10.1111/1462-2920.12314

Goldschmidt, V. M. (1937). The principles of distribution of chemical elements in minerals and rocks. The seventh Hugo Müller Lecture, delivered before the Chemical Society on March 17th, 1937. *Journal of the Chemical Society*, *0*(0), 655–673. https://doi.org/10.1039/jr9370000655

Gouveia, D., Pible, O., Culotta, K., Jouffret, V., Geffard, O., Chaumot, A., Degli-Esposti, D., & Armengaud, J. (2020). Combining proteogenomics and metaproteomics for deep taxonomic and functional characterization of microbiomes from a non-sequenced host. *Npj Biofilms and Microbiomes*, *6*(1), 23. https://doi.org/10.1038/s41522-020-0133-2

Grenga, L., Pible, O., & Armengaud, J. (2019). Pathogen proteotyping: A rapidly developing application of mass spectrometry to address clinical concerns. *Clinical Mass Spectrometry (Del Mar, Calif.)*, *14 Pt A*, 9–17. https://doi.org/10.1016/j.clinms.2019.04.004

Guan, S., Taylor, P. P., Han, Z., Moran, M. F., & Ma, B. (2020). Data dependent-independent acquisition (DDIA) proteomics. *Journal of Proteome Research*, *19*(8), 3230–3237. https://doi.org/10.1021/acs.jproteome.0c00186

Gurdeep Singh, R., Tanca, A., Palomba, A., Van der Jeugt, F., Verschaffelt, P., Uzzau, S., Martens, L., Dawyndt, P., & Mesuere, B. (2019). Unipept 4.0: Functional analysis of metaproteome data. *Journal of Proteome Research*, *18*(2), 606–615. https://doi.org/10.1021/acs.jproteome.8b00716

Hanreich, A., Heyer, R., Benndorf, D., Rapp, E., Pioch, M., Reichl, U., & Klocke, M. (2012). Metaproteome analysis to determine the metabolically active part of a thermophilic microbial community producing biogas from agricultural biomass. *Canadian Journal of Microbiology*, *58*(7), 917–922. https://doi.org/10.1139/w2012-058

Hanson, B. T., Hewson, I., & Madsen, E. L. (2014). Metaproteomic survey of six aquatic habitats: discovering the identities of microbial populations active in biogeochemical cycling. *Microbial Ecology*, *67*(3), 520–539. https://doi.org/10.1007/s00248-013-0346-5

Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C., & Martiny, J. B. H. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews. Microbiology*, *10*(7), 497–506. https://doi.org/10.1038/nrmicro2795

Hardouin, P., Chiron, R., Marchandin, H., Armengaud, J., & Grenga, L. (2021). Metaproteomics to decipher CF host-Microbiota interactions: Overview, challenges and future perspectives. *Genes*, *12*(6), 892. https://doi.org/10.3390/genes12060892

Hayoun, K., Gouveia, D., Grenga, L., Pible, O., Armengaud, J., & Alpha-Bazin, B. (2019). Evaluation of sample preparation methods for fast proteotyping of microorganisms by tandem mass spectrometry. *Frontiers in Microbiology*, *10*, 1985. https://doi.org/10.3389/fmicb.2019.01985

Hayoun, K., Pible, O., Petit, P., Allain, F., Jouffret, V., Culotta, K., Rivasseau, C., Armengaud, J., & Alpha-Bazin, B. (2020). Proteotyping environmental microorganisms by phylopeptidomics: Case study screening water from a radioactive material storage pool. *Microorganisms*, *8*(10), 1525. https://doi.org/10.3390/microorganisms8101525

Hensley, S. A., Jung, J.-H., Park, C.-S., & Holden, J. F. (2014). Thermococcus paralvinellae sp. nov. and Thermococcus cleftensis sp. nov. of hyperthermophilic heterotrophs from deep-sea hydrothermal vents. *International Journal of Systematic and Evolutionary Microbiology*, *64*(Pt 11), 3655–3659. https://doi.org/10.1099/ijs.0.066100-0

Herbst, F.-A., Lünsmann, V., Kjeldal, H., Jehmlich, N., Tholey, A., von Bergen, M., Nielsen, J. L., Hettich, R. L., Seifert, J., & Nielsen, P. H. (2016). Enhancing metaproteomics--The value of models and defined environmental microbial systems. *Proteomics*, *16*(5), 783–798. https://doi.org/10.1002/pmic.201500305

Hermans, S. M., Buckley, H. L., Case, B. S., Curran-Cournane, F., Taylor, M., & Lear, G. (2017). Bacteria as emerging indicators of soil condition. *Applied and Environmental Microbiology*, *83*(1). https://doi.org/10.1128/AEM.02826-16

Hettich, R. L., Sharma, R., Chourey, K., & Giannone, R. J. (2012). Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. *Current Opinion in Microbiology*, *15*(3), 373–380. https://doi.org/10.1016/j.mib.2012.04.008

Heyer, R., Benndorf, D., Kohrs, F., De Vrieze, J., Boon, N., Hoffmann, M., Rapp, E., Schlüter, A., Sczyrba, A., & Reichl, U.

(2016). Proteotyping of biogas plant microbiomes separates biogas plants according to process temperature and reactor type. *Biotechnology for Biofuels*, *9*(1), 155. https://doi.org/10.1186/s13068-016-0572-4

Hollabaugh, C. L. (2007). Chapter 2 Modification of Goldschmidt's geochemical classification of the elements to include arsenic, mercury, and lead as biophile elements. In *Concepts and Applications in Environmental Geochemistry* (pp. 9–31). Elsevier.

Hu, A., Noble, W. S., & Wolf-Yadlin, A. (2016). Technical advances in proteomics: new developments in data-independent acquisition. *F1000Research*, *5*, 419. https://doi.org/10.12688/f1000research.7042.1

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Huang, W., Bai, Z., Hoefel, D., Hu, Q., Lv, X., Zhuang, G., Xu, S., Qi, H., & Zhang, H. (2012). Effects of cotton straw amendment on soil fertility and microbial communities. *Frontiers of Environmental Science & Engineering*, *6*(3), 336–349. https://doi.org/10.1007/s11783-011-0337-z

Hultman, J., Waldrop, M. P., Mackelprang, R., David, M. M., McFarland, J., Blazewicz, S. J., Harden, J., Turetsky, M. R., McGuire, A. D., Shah, M. B., VerBerkmoes, N. C., Lee, L. H., Mavrommatis, K., & Jansson, J. K. (2015). Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature*, *521*(7551), 208–212. https://doi.org/10.1038/nature14238

Hupp, C. R., Pierce, A. R., & Noe, G. B. (2009). Floodplain geomorphic processes and environmental impacts of human alteration along Coastal Plain rivers, USA. *Wetlands (Wilmington, N.C.)*, *29*(2), 413–429. https://doi.org/10.1672/08-169.1

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119

Jacob, J. M., Karthik, C., Saratale, R. G., Kumar, S. S., Prabakar, D., Kadirvelu, K., & Pugazhendhi, A. (2018). Biological approaches to tackle heavy metal pollution: A survey of literature. *Journal of Environmental Management*, *217*, 56–70. https://doi.org/10.1016/j.jenvman.2018.03.077

Jagtap, P., Goslinga, J., Kooren, J. A., McGowan, T., Wroblewski, M. S., Seymour, S. L., & Griffin, T. J. (2013). A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics*, *13*(8), 1352–1357. https://doi.org/10.1002/pmic.201200352

Jaishankar, M., Tseten, T., Anbalagan, N., Mathew, B. B., & Beeregowda, K. N. (2014). Toxicity, mechanism and health effects of some heavy metals. *Interdisciplinary Toxicology*, *7*(2), 60–72. https://doi.org/10.2478/intox-2014-0009

Jambon, A., & Thomas, A. (01 2009). *Géochimie : Géodynamique et Cycles*.

James, L. A. (2013). Legacy sediment: Definitions and processes of episodically produced anthropogenic sediment. *Anthropocene*, *2*, 16–26. https://doi.org/10.1016/j.ancene.2013.04.001

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), nwaa280. https://doi.org/10.1093/nsr/nwaa280

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 5029. https://doi.org/10.1038/s41467-019-13036-1

Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, *4*(11), 923–925. https://doi.org/10.1038/nmeth1113

Karimi, B., Terrat, S., Dequiedt, S., Saby, N. P. A., Horrigue, W., Lelièvre, M., Nowak, V., Jolivet, C., Arrouays, D., Wincker, P., Cruaud, C., Bispo, A., Maron, P.-A., Bouré, N. C. P., & Ranjard, L. (2018). Biogeography of soil bacteria and archaea across France. *Science Advances*, *4*(7), eaat1808. https://doi.org/10.1126/sciadv.aat1808

Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Brister, J. R., & O'Sullivan, C. (2022). The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Research*, *50*(D1), D387–D390. https://doi.org/10.1093/nar/gkab1053

Keegan, K. P., Glass, E. M., & Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods in Molecular Biology (Clifton, N.J.)*, *1399*, 207–233. https://doi.org/10.1007/978-1-4939-3369-3_13

Keiblinger, K. M., Fuchs, S., Zechmeister-Boltenstern, S., & Riedel, K. (2016). Soil and leaf litter metaproteomics-a brief guideline from sampling to understanding. *FEMS Microbiology Ecology*, *92*(11). https://doi.org/10.1093/femsec/fiw180

Keiblinger, K. M., Wilhartitz, I. C., Schneider, T., Roschitzki, B., Schmid, E., Eberl, L., Riedel, K., & Zechmeister-Boltenstern, S. (2012). Soil metaproteomics - Comparative evaluation of protein extraction protocols. *Soil Biology & Biochemistry*, *54*(15–10), 14–24. https://doi.org/10.1016/j.soilbio.2012.05.014

Keith, L. H. (2015). The source of U.s. epa's sixteen PAH priority pollutants. *Polycyclic Aromatic Compounds*, *35*(2–4), 147–160. https://doi.org/10.1080/10406638.2014.892886

Kim, S., & Pevzner, P. A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, *5*(1), 5277. https://doi.org/10.1038/ncomms6277

Kleiner, M., Thorson, E., Sharp, C. E., Dong, X., Liu, D., Li, C., & Strous, M. (2017). Assessing species biomass contributions in microbial communities via metaproteomics. *Nature Communications*, *8*(1), 1558. https://doi.org/10.1038/s41467-017-01544-x

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews. Microbiology*, *16*(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Kolhe, N., Zinjarde, S., & Acharya, C. (2018). Responses exhibited by various microbial groups relevant to uranium exposure. *Biotechnology Advances*, *36*(7), 1828–1846. https://doi.org/10.1016/j.biotechadv.2018.07.002

Kotoky, R., Rajkumari, J., & Pandey, P. (2018). The rhizosphere microbiome: Significance in rhizoremediation of polyaromatic hydrocarbon contaminated soil. *Journal of Environmental Management*, *217*, 858–870. https://doi.org/10.1016/j.jenvman.2018.04.022

Lahrichi, S. L., Affolter, M., Zolezzi, I. S., & Panchaud, A. (2013). Food peptidomics: large scale analysis of small bioactive peptides--a pilot study. *Journal of Proteomics*, *88*, 83–91. https://doi.org/10.1016/j.jprot.2013.02.018

Lasken, R. S., & McLean, J. S. (2014). Recent advances in genomic DNA sequencing of microbial species from single cells. *Nature Reviews. Genetics*, *15*(9), 577–584. https://doi.org/10.1038/nrg3785

Le bassin de la Seine. (n.d.). Agence de l'Eau Seine-Normandie. Retrieved September 14, 2022, from http://www.eau-seine-normandie.fr/agence-de-leau/le-bassin-de-la-seine

Le blocus de Paris en 1649. (n.d.). Histoire-image.org. Retrieved September 14, 2022, from http://histoire-image.org/fr/etudes/blocus-paris-1649

Le Cloarec, M.-F., Bonte, P. H., Lestel, L., Lefèvre, I., & Ayrault, S. (2011). Sedimentary record of metal contamination in the Seine River during the last century. *Physics and Chemistry of the Earth (2002)*, *36*(12), 515–529. https://doi.org/10.1016/j.pce.2009.02.003

Le Gall, M. (2016). *Traçage des sources de sédiments à l'amont des hydrosystèmes agricoles : apport de la géochimie élémentaire, du rapport ⁸⁷Sr/⁸⁶Sr et des radionucléides* [Université Paris Saclay (COmUE)]. https://tel.archives-ouvertes.fr/tel-01412184

Le Gall, M., Ayrault, S., Evrard, O., Laceby, J. P., Gateuille, D., Lefèvre, I., Mouchel, J.-M., & Meybeck, M. (2018). Investigating the metal contamination of sediment transported by the 2016 Seine River flood (Paris, France). *Environmental Pollution (Barking, Essex: 1987)*, *240*, 125–139. https://doi.org/10.1016/j.envpol.2018.04.082

Le Pape, P., Ayrault, S., & Quantin, C. (2012). Trace element behavior and partition versus urbanization gradient in an urban river (Orge River, France). *Journal of Hydrology*, *472–473*, 99–110. https://doi.org/10.1016/j.jhydrol.2012.09.042

Lehmann, J., Bossio, D. A., Kögel-Knabner, I., & Rillig, M. C. (2020). The concept and future prospects of soil health. *Nature Reviews Earth & Environment*, *1*(10), 544–553. https://doi.org/10.1038/s43017-020-0080-8

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database issue), D19-21. https://doi.org/10.1093/nar/gkq1019

Lepage, H. (2015). *Traçage de la dispersion des sédiments contaminés dans les bassins versants côtiers de Fukushima*. Université Paris Sud - Paris XI.

Lestel, L., Eschbach, D., Meybeck, M., & Gob, F. (2020). The evolution of the seine basin water bodies through historical maps. In *The Handbook of Environmental Chemistry* (pp. 29–57). Springer International Publishing.

Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Bálint, M., Barker, K. B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M. L., Cai, J., Caperello, N. D., Carlson, K., Castilla-Rubio, J. C., Chaw, S.-M., Chen, L., Childers, A. K., Coddington, J. A., … Zhang, G. (2022). The Earth BioGenome Project 2020: Starting the clock. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), e2115635118. https://doi.org/10.1073/pnas.2115635118

Lewin, J. (2010). Medieval environmental impacts and feedbacks: The lowland floodplains of England and Wales. *Geoarchaeology*, *25*(3), 267–311. https://doi.org/10.1002/gea.20308

Lewin, J. (2013). Enlightenment and the GM floodplain: Enlightenment and the gm floodplain. *Earth Surface Processes and Landforms*, *38*(1), 17–29. https://doi.org/10.1002/esp.3230

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Lin, W., Wu, L., Lin, S., Zhang, A., Zhou, M., Lin, R., Wang, H., Chen, J., Zhang, Z., & Lin, R. (2013). Metaproteomic analysis of ratoon sugarcane rhizospheric soil. *BMC Microbiology*, *13*, 135. https://doi.org/10.1186/1471-2180-13-135

Lorgeoux, C., Moilleron, R., Gasperi, J., Ayrault, S., Bonté, P., Lefèvre, I., & Tassin, B. (2016). Temporal trends of persistent organic pollutants in dated sediment cores: Chemical fingerprinting of the anthropogenic impacts in the Seine River basin, Paris. *The Science of the Total Environment*, *541*, 1355–1363. https://doi.org/10.1016/j.scitotenv.2015.09.147

Louca, S., Parfrey, L. W., & Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science (New York, N.Y.)*, *353*(6305), 1272–1277. https://doi.org/10.1126/science.aaf4507

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217X-1-18

Ma, B. (2015). Novor: real-time peptide de novo sequencing software. *Journal of the American Society for Mass Spectrometry*, *26*(11), 1885–1894. https://doi.org/10.1007/s13361-015-1204-0

MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., & MacCoss, M. J. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics (Oxford, England)*, *26*(7), 966–968. https://doi.org/10.1093/bioinformatics/btq054

Marcus, W. A., Nielsen, C. C., & Cornwell, J. C. (1993). Sediment budget-based estimates of trace metal inputs to a Chesapeake estuary. *Environmental Geology*, *22*(1), 1–9. https://doi.org/10.1007/bf00775277

Mascot database search: MS/MS Results Interpretation. (n.d.). Matrixscience.com. Retrieved September 14, 2022, from http://www.matrixscience.com/help/interpretation_help.html

Maseh, K., Ehsan, N., Mukhtar, S., Mehnaz, S., & Malik, K. A. (2021). Metaproteomics: an emerging tool for the identification of proteins from extreme environments. *Environmental Sustainability*, *4*(1), 39–50. https://doi.org/10.1007/s42398-020-00158-2

McDonald, A. (1981). A perspective of environmental pollution. M. W. Holdgate, Cambridge University Press, Price: £15.00. *Earth Surface Processes and Landforms*, *6*(3–4), 398–398. https://doi.org/10.1002/esp.3290060319

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, *7*(1), 11257. https://doi.org/10.1038/ncomms11257

Mesuere, B., Devreese, B., Debyser, G., Aerts, M., Vandamme, P., & Dawyndt, P. (2012). Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. *Journal of Proteome Research*, *11*(12), 5773–5780. https://doi.org/10.1021/pr300576s

Meyer, F., Lesker, T.-R., Koslicki, D., Fritz, A., Gurevich, A., Darling, A. E., Sczyrba, A., Bremges, A., & McHardy, A. C. (2021). Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. *Nature Protocols*, *16*(4), 1785–1801. https://doi.org/10.1038/s41596-020-00480-3

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, *8*(8), 1551–1566. https://doi.org/10.1038/nprot.2013.092

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2016). PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Research*, *44*(D1), D336-42. https://doi.org/10.1093/nar/gkv1194

Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics (Oxford, England)*, *32*(7), 1088–1090. https://doi.org/10.1093/bioinformatics/btv697

Morisse, P. (2019). *Correction de données de séquençage de troisième génération* [Normandie Université]. https://tel.archives-ouvertes.fr/tel-02320413

Mourier, B., Desmet, M., Van Metre, P. C., Mahler, B. J., Perrodin, Y., Roux, G., Bedell, J.-P., Lefèvre, I., & Babut, M. (2014). Historical records, sources, and spatial trends of PCBs along the Rhône River (France). *The Science of the Total Environment*, *476–477*, 568–576. https://doi.org/10.1016/j.scitotenv.2014.01.026

Muth, T., Behne, A., Heyer, R., Kohrs, F., Benndorf, D., Hoffmann, M., Lehtevä, M., Reichl, U., Martens, L., & Rapp, E. (2015). The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *Journal of Proteome Research*, *14*(3), 1557–1565. https://doi.org/10.1021/pr501246w

Muth, T., Kolmeder, C. A., Salojärvi, J., Keskitalo, S., Varjosalo, M., Verdam, F. J., Rensen, S. S., Reichl, U., de Vos, W. M., Rapp, E., & Martens, L. (2015). Navigating through metaproteomics data: a logbook of database searching. *Proteomics*, *15*(20), 3439–3453. https://doi.org/10.1002/pmic.201400560

Muth, T., & Renard, B. Y. (2018). Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Briefings in Bioinformatics*, *19*(5), 954–970. https://doi.org/10.1093/bib/bbx033

Nakayasu, E. S., Gritsenko, M., Piehowski, P. D., Gao, Y., Orton, D. J., Schepmoes, A. A., Fillmore, T. L., Frohnert, B. I., Rewers, M., Krischer, J. P., Ansong, C., Suchy-Dicey, A. M., Evans-Molina, C., Qian, W.-J., Webb-Robertson, B.-J. M., & Metz, T. O. (2021). Tutorial: best practices and considerations for mass-spectrometry-based protein

biomarker discovery and validation. *Nature Protocols*, *16*(8), 3737–3760. https://doi.org/10.1038/s41596-021-00566-6

Nemergut, D. R., Costello, E. K., Hamady, M., Lozupone, C., Jiang, L., Schmidt, S. K., Fierer, N., Townsend, A. R., Cleveland, C. C., Stanish, L., & Knight, R. (2011). Global patterns in the biogeography of bacterial taxa. *Environmental Microbiology*, *13*(1), 135–144. https://doi.org/10.1111/j.1462-2920.2010.02315.x

Ng, C., DeMaere, M. Z., Williams, T. J., Lauro, F. M., Raftery, M., Gibson, J. A. E., Andrews-Pfannkoch, C., Lewis, M., Hoffman, J. M., Thomas, T., & Cavicchioli, R. (2010). Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. *The ISME Journal*, *4*(8), 1002–1019. https://doi.org/10.1038/ismej.2010.28

Oliver, T. H., Heard, M. S., Isaac, N. J. B., Roy, D. B., Procter, D., Eigenbrod, F., Freckleton, R., Hector, A., Orme, C. D. L., Petchey, O. L., Proença, V., Raffaelli, D., Suttle, K. B., Mace, G. M., Martín-López, B., Woodcock, B. A., & Bullock, J. M. (2015). Biodiversity and resilience of ecosystem functions. *Trends in Ecology & Evolution*, *30*(11), 673–684. https://doi.org/10.1016/j.tree.2015.08.009

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., & Pop, M. (2019). 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

Paczesny, S., Raiker, N., Brooks, S., & Mumaw, C. (2013). Graft-versus-host disease biomarkers: omics and personalized medicine. *International Journal of Hematology*, *98*(3), 275–292. https://doi.org/10.1007/s12185-013-1406-9

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., & Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, *38*(9), 1079–1086. https://doi.org/10.1038/s41587-020-0501-8

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

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, *2*(11), 1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Pavilonis, B. T., Lioy, P. J., Guazzetti, S., Bostick, B. C., Donna, F., Peli, M., Zimmerman, N. J., Bertrand, P., Lucas, E., Smith, D. R., Georgopoulos, P. G., Mi, Z., Royce, S. G., & Lucchini, R. G. (2015). Manganese concentrations in soil and settled dust in an area with historic ferroalloy production. *Journal of Exposure Science & Environmental Epidemiology*, *25*(4), 443–450. https://doi.org/10.1038/jes.2014.70

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics (Oxford, England)*, *28*(11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Pérez-Cobas, A. E., Gosalbes, M. J., Friedrichs, A., Knecht, H., Artacho, A., Eismann, K., Otto, W., Rojo, D., Bargiela, R., von Bergen, M., Neulinger, S. C., Däumer, C., Heinsen, F.-A., Latorre, A., Barbas, C., Seifert, J., dos Santos, V. M., Ott, S. J., Ferrer, M., & Moya, A. (2013). Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut*, *62*(11), 1591–1601. https://doi.org/10.1136/gutjnl-2012-303184

Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, *20*(18), 3551–3567. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

Pfammatter, S., Bonneil, E., & Thibault, P. (2016). Improvement of quantitative measurements in multiplex proteomics using high-field asymmetric waveform spectrometry. *Journal of Proteome Research*, *15*(12), 4653–4665. https://doi.org/10.1021/acs.jproteome.6b00745

Pible, O., Allain, F., Jouffret, V., Culotta, K., Miotello, G., & Armengaud, J. (2020). Estimating relative biomasses of organisms in microbiota using “phylopeptidomics.” *Microbiome*, *8*(1), 30. https://doi.org/10.1186/s40168-020-00797-x

Picotti, P., & Aebersold, R. (2012). Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature Methods*, *9*(6), 555–566. https://doi.org/10.1038/nmeth.2015

Pietilä, S., Suomi, T., & Elo, L. L. (2022). Introducing untargeted data-independent acquisition for metaproteomics of complex microbial samples. *ISME Communications*, *2*(1). https://doi.org/10.1038/s43705-022-00137-0

Ponty, Y. (2014). Bio-algorithmique des ARN : petite promenade aux interfaces. In E. Sopena (Ed.), *1024 - Bulletin de la société informatique de France* (Vol. 4, pp. 23–53). SIF.

Portik, D., Ashby, M., Zhang, S., Locken, K., Tang, S., Farthing, B., Weinsten, M., Carlin, M., Cano, R., & Wilkinson, J. (n.d.). *A new standard: high MAG recovery and precision species profiling of a pooled human gut microbiome reference using PacBio HiFi sequencing*. Pacb.com. Retrieved September 14, 2022, from https://www.pacb.com/wp-content/uploads/ASM-Zymo-Wilkinson.pdf

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*, 341. https://doi.org/10.1186/1471-2164-13-341

Qu'est ce qu'un bassin versant ? - Syndicat Mixte du Bassin Versant de l'Huveaune. (2017, September 27). Syndicat Mixte du Bassin Versant de l'Huveaune. https://www.syndicat-huveaune.fr/le-bassin-versant-de-lhuveaune/quest-ce-quun-bassin-versant/

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833–844. https://doi.org/10.1038/nbt.3935

Quinn, G. A., Abdelhameed, A., Banat, I. M., Berrar, D., Doerr, S. H., Dudley, E., Francis, L. W., Gazze, S. A., Hallin, I., Matthews, G. P., Swain, M. T., Whalley, W. R., & van Keulen, G. (2022). Complementary protein extraction methods increase the identification of the Park Grass Experiment metaproteome. *Applied Soil Ecology: A Section of Agriculture, Ecosystems & Environment*, *173*(104388), 104388. https://doi.org/10.1016/j.apsoil.2022.104388

Renu, Gupta, S. K., Rai, A. K., Sarim, K. M., Sharma, A., Budhlakoti, N., Arora, D., Verma, D. K., & Singh, D. P. (2019). Metaproteomic data of maize rhizosphere for deciphering functional diversity. *Data in Brief*, *27*(104574), 104574. https://doi.org/10.1016/j.dib.2019.104574

Resongles, E., Casiot, C., Freydier, R., Dezileau, L., Viers, J., & Elbaz-Poulichet, F. (2014). Persisting impact of historical mining activity to metal (Pb, Zn, Cd, Tl, Hg) and metalloid (As, Sb) enrichment in sediments of the Gardon River, Southern France. *The Science of the Total Environment*, *481*, 509–521. https://doi.org/10.1016/j.scitotenv.2014.02.078

Révész, Á., Milley, M. G., Nagy, K., Szabó, D., Kalló, G., Csősz, É., Vékey, K., & Drahos, L. (2021). Tailoring to search engines: Bottom-up proteomics with collision energies optimized for identification confidence. *Journal of Proteome Research*, *20*(1), 474–484. https://doi.org/10.1021/acs.jproteome.0c00518

Rho, M., Tang, H., & Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research*, *38*(20), e191. https://doi.org/10.1093/nar/gkq747

Roalkvam, I., Dahle, H., Chen, Y., Jørgensen, S. L., Haflidason, H., & Steen, I. H. (2012). Fine-scale community structure analysis of ANME in Nyegga sediments with high and low methane flux. *Frontiers in Microbiology*, *3*, 216. https://doi.org/10.3389/fmicb.2012.00216

Rocca, J. D., Simonin, M., Blaszczak, J. R., Ernakovich, J. G., Gibbons, S. M., Midani, F. S., & Washburne, A. D. (2018). The Microbiome Stress Project: Toward a global meta-analysis of environmental stressors and their effects on microbial communities. *Frontiers in Microbiology*, *9*, 3272. https://doi.org/10.3389/fmicb.2018.03272

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K.-A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, *13*(11), e1005752. https://doi.org/10.1371/journal.pcbi.1005752

Rosolen, V., De-Campos, A. B., Govone, J. S., & Rocha, C. (2015). Contamination of wetland soils and floodplain sediments from agricultural activities in the Cerrado Biome (State of Minas Gerais, Brazil). *Catena*, *128*, 203–210. https://doi.org/10.1016/j.catena.2015.02.007

Saito, M. A., Bertrand, E. M., Duffy, M. E., Gaylord, D. A., Held, N. A., Hervey, W. J., 4th, Hettich, R. L., Jagtap, P. D., Janech, M. G., Kinkade, D. B., Leary, D. H., McIlvin, M. R., Moore, E. K., Morris, R. M., Neely, B. A., Nunn, B. L., Saunders, J. K., Shepherd, A. I., Symmonds, N. I., & Walsh, D. A. (2019). Progress and challenges in ocean metaproteomics and proposed best practices for data sharing. *Journal of Proteome Research*, *18*(4), 1461–1476. https://doi.org/10.1021/acs.jproteome.8b00761

Sajulga, R., Easterly, C., Riffle, M., Mesuere, B., Muth, T., Mehta, S., Kumar, P., Johnson, J., Gruening, B. A., Schiebenhoefer, H., Kolmeder, C. A., Fuchs, S., Nunn, B. L., Rudney, J., Griffin, T. J., & Jagtap, P. D. (2020). Survey of metaproteomics software tools for functional microbiome analysis. *PloS One*, *15*(11), e0241503. https://doi.org/10.1371/journal.pone.0241503

Schäfer, J., Coynel, A., & Blanc, G. (2022). Impact of metallurgy tailings in a major European fluvial-estuarine system: Trajectories and resilience over seven decades. *The Science of the Total Environment*, *805*(150195), 150195. https://doi.org/10.1016/j.scitotenv.2021.150195

Schaubeck, M., Clavel, T., Calasan, J., Lagkouvardos, I., Haange, S. B., Jehmlich, N., Basic, M., Dupont, A., Hornef, M., von Bergen, M., Bleich, A., & Haller, D. (2016). Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence. *Gut*, *65*(2), 225–237. https://doi.org/10.1136/gutjnl-2015-309333

Scheltema, R. A., Hauschild, J.-P., Lange, O., Hornburg, D., Denisov, E., Damoc, E., Kuehn, A., Makarov, A., & Mann, M. (2014). The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Molecular & Cellular Proteomics: MCP*, *13*(12), 3698–3708. https://doi.org/10.1074/mcp.M114.043489

Schloss, P. D., & Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology*, *71*(3), 1501–1506.

https://doi.org/10.1128/AEM.71.3.1501-1506.2005

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schneider, T., Keiblinger, K. M., Schmid, E., Sterflinger-Gleixner, K., Ellersdorfer, G., Roschitzki, B., Richter, A., Eberl, L., Zechmeister-Boltenstern, S., & Riedel, K. (2012). Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. *The ISME Journal*, *6*(9), 1749–1762. https://doi.org/10.1038/ismej.2012.11

Schulze, W. X., Gleixner, G., Kaiser, K., Guggenberger, G., Mann, M., & Schulze, E.-D. (2005). A proteomic fingerprint of dissolved organic carbon and of soil particles. *Oecologia*, *142*(3), 335–343. https://doi.org/10.1007/s00442-004-1698-9

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., … McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nature Methods*, *14*(11), 1063–1071. https://doi.org/10.1038/nmeth.4458

Searle, B. C., Pino, L. K., Egertson, J. D., Ting, Y. S., Lawrence, R. T., MacLean, B. X., Villén, J., & MacCoss, M. J. (2018). Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications*, *9*(1), 5128. https://doi.org/10.1038/s41467-018-07454-w

Searle, B. C., Turner, M., & Nesvizhskii, A. I. (2008). Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *Journal of Proteome Research*, *7*(1), 245–253. https://doi.org/10.1021/pr070540w

Seifert, J., & Muth, T. (2019). Editorial for special issue: Metaproteomics. *Proteomes*, *7*(1), 9. https://doi.org/10.3390/proteomes7010009

Semenov, M. V. (2021). Metabarcoding and metagenomics in soil ecology research: Achievements, challenges, and prospects. *Biology Bulletin Reviews*, *11*(1), 40–53. https://doi.org/10.1134/s2079086421010084

Shaheen, S. M., & Rinklebe, J. (2014). Geochemical fractions of chromium, copper, and zinc and their vertical distribution in floodplain soil profiles along the Central Elbe River, Germany. *Geoderma*, *228–229*, 142–159. https://doi.org/10.1016/j.geoderma.2013.10.012

Shrivastava, A., Ghosh, D., Dash, A., & Bose, S. (2015). Arsenic contamination in soil and sediment in India: Sources, effects, and remediation. *Current Pollution Reports*, *1*(1), 35–46. https://doi.org/10.1007/s40726-015-0004-2

Singh, B., & Singh, K. (2016). Microbial degradation of herbicides. *Critical Reviews in Microbiology*, *42*(2), 245–261. https://doi.org/10.3109/1040841X.2014.929564

Smith, L. M., Kelleher, N. L., & Consortium for Top Down Proteomics. (2013). Proteoform: a single term describing protein complexity. *Nature Methods*, *10*(3), 186–187. https://doi.org/10.1038/nmeth.2369

Sollanek, K. J., Burniston, J. G., Kavazis, A. N., Morton, A. B., Wiggs, M. P., Ahn, B., Smuder, A. J., & Powers, S. K. (2017). Global proteome changes in the rat diaphragm induced by endurance exercise training. *PloS One*, *12*(1), e0171007. https://doi.org/10.1371/journal.pone.0171007

Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology*, *44*(4), 846–849. https://doi.org/10.1099/00207713-44-4-846

Steinegger, M., Mirdita, M., & Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods*, *16*(7), 603–606. https://doi.org/10.1038/s41592-019-0437-4

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., … Bork, P. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science (New York, N.Y.)*, *348*(6237), 1261359. https://doi.org/10.1126/science.1261359

Tabb, D. L., Ma, Z.-Q., Martin, D. B., Ham, A.-J. L., & Chambers, M. C. (2008). DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *Journal of Proteome Research*, *7*(9), 3838–3846. https://doi.org/10.1021/pr800154p

Tamtam, F., Le Bot, B., Dinh, T., Mompelat, S., Eurin, J., Chevreuil, M., Bonté, P., Mouchel, J.-M., & Ayrault, S. (2011). A 50-year record of quinolone and sulphonamide antimicrobial agents in Seine River sediments. *Journal of Soils and Sediments*, *11*(5), 852–859. https://doi.org/10.1007/s11368-011-0364-1

Tanca, A., Palomba, A., Deligios, M., Cubeddu, T., Fraumene, C., Biosa, G., Pagnozzi, D., Addis, M. F., & Uzzau, S. (2013). Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PloS One*, *8*(12), e82981. https://doi.org/10.1371/journal.pone.0082981

Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., Muth, T., Rapp, E., Martens, L., Addis, M. F., & Uzzau, S. (2016). The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome*, *4*(1), 51. https://doi.org/10.1186/s40168-016-0196-8

Tartaglia, M., Bastida, F., Sciarrillo, R., & Guarino, C. (2020). Soil metaproteomics for the study of the relationships between microorganisms and plants: A review of extraction protocols and ecological insights. *International Journal of Molecular Sciences*, *21*(22), 8455. https://doi.org/10.3390/ijms21228455

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N. S., Wijesundera, R., Villarreal Ruiz, L., Vasco-Palacios, A. M., Thu, P. Q., Suija, A., Smith, M. E., Sharp, C., Saluveer, E., Saitta, A., Rosas, M., Riit, T., Ratkowsky, D., Pritsch, K., Põldmaa, K., … Abarenkov, K. (2014). Fungal biogeography. Global diversity and geography of soil fungi. *Science (New York, N.Y.)*, *346*(6213), 1256688. https://doi.org/10.1126/science.1256688

Teng, F., Darveekaran Nair, S. S., Zhu, P., Li, S., Huang, S., Li, X., Xu, J., & Yang, F. (2018). Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. *Scientific Reports*, *8*(1), 16321. https://doi.org/10.1038/s41598-018-34294-x

Timp, W., & Timp, G. (2020). Beyond mass spectrometry, the next step in proteomics. *Science Advances*, *6*(2), eaax8978. https://doi.org/10.1126/sciadv.aax8978

Tran, N. H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A., & Li, M. (2019). Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature Methods*, *16*(1), 63–66. https://doi.org/10.1038/s41592-018-0260-3

Tsiamis, V., Ienasescu, H.-I., Gabrielaitis, D., Palmblad, M., Schwämmle, V., & Ison, J. (2019). One thousand and one software for proteomics: Tales of the toolmakers of science. *Journal of Proteome Research*, *18*(10), 3580–3585. https://doi.org/10.1021/acs.jproteome.9b00219

Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, *5*, 170203. https://doi.org/10.1038/sdata.2017.203

Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, *11*(12), 2301–2319. https://doi.org/10.1038/nprot.2016.136

Van Den Bossche, T., Kunath, B. J., Schallert, K., Schäpe, S. S., Abraham, P. E., Armengaud, J., Arntzen, M. Ø., Bassignani, A., Benndorf, D., Fuchs, S., Giannone, R. J., Griffin, T. J., Hagen, L. H., Halder, R., Henry, C., Hettich, R. L., Heyer, R., Jagtap, P., Jehmlich, N., … Muth, T. (2021). Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nature Communications*, *12*(1), 7305. https://doi.org/10.1038/s41467-021-27542-8

Vartoukian, S. R., Palmer, R. M., & Wade, W. G. (2010). Strategies for culture of "unculturable" bacteria: Culturing the unculturable. *FEMS Microbiology Letters*, *309*(1), 1–7. https://doi.org/10.1111/j.1574-6968.2010.02000.x

Vauclin, S. (2020). *Influence des aménagements et des contaminations sur les héritages sédimentaires des fleuves anthropisés : le cas du Rhône* [Université de Lyon]. https://tel.archives-ouvertes.fr/tel-03448844

Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A., & Martens, L. (2011). SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics*, *11*(5), 996–999. https://doi.org/10.1002/pmic.201000595

Vaudel, M., Burkhart, J. M., Zahedi, R. P., Oveland, E., Berven, F. S., Sickmann, A., Martens, L., & Barsnes, H. (2015). PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology*, *33*(1), 22–24. https://doi.org/10.1038/nbt.3109

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., … Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)*, *304*(5667), 66–74. https://doi.org/10.1126/science.1093857

Viers, J., Dupré, B., & Gaillardet, J. (2009). Chemical composition of suspended sediments in World Rivers: New insights from a new database. *The Science of the Total Environment*, *407*(2), 853–868. https://doi.org/10.1016/j.scitotenv.2008.09.053

Vollmers, J., Wiegand, S., & Kaster, A.-K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PloS One*, *12*(1), e0169662. https://doi.org/10.1371/journal.pone.0169662

Vue de la Seine au XVIIIe siècle. (n.d.). Histoire-image.org. Retrieved September 14, 2022, from https://histoire-image.org/fr/etudes/vue-seine-xviiie-siecle

Walling, D. E. (2013). The evolution of sediment source fingerprinting investigations in fluvial systems. *Journal of Soils and Sediments*, *13*(10), 1658–1675. https://doi.org/10.1007/s11368-013-0767-2

Wang, D.-Z., Xie, Z.-X., & Zhang, S.-F. (2014). Marine metaproteomics: current status and future directions. *Journal of Proteomics*, *97*, 27–35. https://doi.org/10.1016/j.jprot.2013.08.024

Wang, L., Cui, X., Cheng, H., Chen, F., Wang, J., Zhao, X., Lin, C., & Pu, X. (2015). A review of soil cadmium contamination

in China including a health risk assessment. *Environmental Science and Pollution Research International*, *22*(21), 16441–16452. https://doi.org/10.1007/s11356-015-5273-1

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Whitaker, R. J., Grogan, D. W., & Taylor, J. W. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science (New York, N.Y.)*, *301*(5635), 976–978. https://doi.org/10.1126/science.1086909

Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J. C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L., & Hochstrasser, D. F. (1996). From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/Technology*, *14*(1), 61–65. https://doi.org/10.1038/nbt0196-61

Wilmes, P., Heintz-Buschart, A., & Bond, P. L. (2015). A decade of metaproteomics: where we stand and what the future holds: PROTEOMICS. *Proteomics*, *15*(20), 3409–3417. https://doi.org/10.1002/pmic.201500183

Wilmes, P., Wexler, M., & Bond, P. L. (2008). Metaproteomics provides functional insight into activated sludge wastewater treatment. *PloS One*, *3*(3), e1778. https://doi.org/10.1371/journal.pone.0001778

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wright, J. C., & Choudhary, J. S. (2016). DecoyPyrat: Fast non-redundant hybrid decoy sequence generation for large scale proteomics. *Journal of Proteomics & Bioinformatics*, *9*(6), 176–180. https://doi.org/10.4172/jpb.1000404

Xiao, E., Krumins, V., Xiao, T., Dong, Y., Tang, S., Ning, Z., Huang, Z., & Sun, W. (2017). Depth-resolved microbial community analyses in two contrasting soil cores contaminated by antimony and arsenic. *Environmental Pollution (Barking, Essex: 1987)*, *221*, 244–255. https://doi.org/10.1016/j.envpol.2016.11.071

Xiao, J., Tanca, A., Jia, B., Yang, R., Wang, B., Zhang, Y., & Li, J. (2018). Metagenomic taxonomy-guided database-searching strategy for improving metaproteomic analysis. *Journal of Proteome Research*, *17*(4), 1596–1605. https://doi.org/10.1021/acs.jproteome.7b00894

Xu, G., Zhang, L., Liu, X., Guan, F., Xu, Y., Yue, H., Huang, J.-Q., Chen, J., Wu, N., & Tian, J. (2022). Combined assembly of long and short sequencing reads improve the efficiency of exploring the soil metagenome. *BMC Genomics*, *23*(1), 37. https://doi.org/10.1186/s12864-021-08260-3

Xu, J., Bravo, A. G., Lagerkvist, A., Bertilsson, S., Sjöblom, R., & Kumpiene, J. (2015). Sources and remediation techniques for mercury contaminated soil. *Environment International*, *74*, 42–53. https://doi.org/10.1016/j.envint.2014.09.007

Xue, Y., Jonassen, I., Øvreås, L., & Taş, N. (2020). Metagenome-assembled genome distribution and key functionality highlight importance of aerobic metabolism in Svalbard permafrost. *FEMS Microbiology Ecology*, *96*(5). https://doi.org/10.1093/femsec/fiaa057

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., & Zhang, L. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal*, *19*, 6301–6314. https://doi.org/10.1016/j.csbj.2021.11.028

Yang, Y., Gao, S., Wang, Y. P., Jia, J., Xiong, J., & Zhou, L. (2019). Revisiting the problem of sediment motion threshold. *Continental Shelf Research*, *187*(103960), 103960. https://doi.org/10.1016/j.csr.2019.103960

Zampieri, E., Chiapello, M., Daghino, S., Bonfante, P., & Mello, A. (2016). Soil metaproteomics reveals an inter-kingdom stress response to the presence of black truffles. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep25773

Zhang, B., Wu, X., Zhang, G., Li, S., Zhang, W., Chen, X., Sun, L., Zhang, B., Liu, G., & Chen, T. (2016). The diversity and biogeography of the communities of Actinobacteria in the forelands of glaciers at a continental scale. *Environmental Research Letters*, *11*(5), 054012. https://doi.org/10.1088/1748-9326/11/5/054012

Zhang, F., Ge, W., Ruan, G., Cai, X., & Guo, T. (2020). Data-independent acquisition mass spectrometry-based proteomics and software tools: A glimpse in 2020. *Proteomics*, *20*(17–18), e1900276. https://doi.org/10.1002/pmic.201900276

Zhang, Tayyab, Abubakar, Yang, Pang, Islam, Lin, Li, Luo, Fan, Fallah, & Zhang. (2019). Bacteria with different assemblages in the soil profile drive the diverse nutrient cycles in the sugarcane straw retention ecosystem. *Diversity*, *11*(10), 194. https://doi.org/10.3390/d11100194

Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C., & Yates, J. R., 3rd. (2013). Protein analysis by shotgun/bottom-up proteomics. *Chemical Reviews*, *113*(4), 2343–2394. https://doi.org/10.1021/cr3003533

Annexes

2.3 Annexe de la publication « Increasing the power of interpretation for soil metaproteomics data »

Les données de spectrométrie de masse ont été déposés sur PRIDE avec l'identifiant PXD026798 (https://doi.org/10.6019/PXD026798).

Listes des publications et des communications

2.4 Publications

Certains de mes travaux antérieurs au Li2D n'étaient pas directement liés à mes travaux de thèse, mais ont conduit à des publications dont je suis co-auteur. Ces publications parues récemment sont citées ci-après :

- Gouveia, D., Pible, O., Culotta, K., **Jouffret, V**., Geffard, O., Chaumot, A., Degli-Esposti, D., & Armengaud, J. (2020). Combining proteogenomics and metaproteomics for deep taxonomic and functional characterization of microbiomes from a non-sequenced host. *Npj Biofilms and Microbiomes*, *6*(1), 23. https://doi.org/10.1038/s41522-020-0133-2

- Hayoun, K., Pible, O., Petit, P., Allain, F., **Jouffret, V.**, Culotta, K., Rivasseau, C., Armengaud, J., & Alpha-Bazin, B. (2020). Proteotyping environmental microorganisms by phylopeptidomics: Case study screening water from a radioactive material storage pool. *Microorganisms*, *8*(10), 1525. https://doi.org/10.3390/microorganisms8101525

- Hirtz, C., Mannaa, A. M., Moulis, E., Pible, O., O'Flynn, R., Armengaud, J., **Jouffret**, **V.**, Lemaistre, C., Dominici, G., Martinez, A. Y., Dunyach-Remy, C., Tiers, L., Lavigne, J.-P., Tramini, P., Goldsmith, M.-C., Lehmann, S., Deville de Périère, D., & Vialaret, J. (2022). Deciphering black extrinsic tooth stain composition in children using metaproteomics. *ACS Omega*, *7*(10), 8258–8267. https://doi.org/10.1021/acsomega.1c04770

- Pible, O., Allain, F., **Jouffret, V.**, Culotta, K., Miotello, G., & Armengaud, J. (2020). Estimating relative biomasses of organisms in microbiota using "phylopeptidomics." *Microbiome*, *8*(1), 30. https://doi.org/10.1186/s40168-020-00797-x

2.5 Communications orales

- **Challenges and solutions for exploring taxonomical and functional changes of microbiota along a soil core.** Virginie Jouffret
  Congrès international de métaprotéomique. 27 – 29 septembre 2021, Luxembourg, Luxembourg.

- **Probing microbial life of soil: a new EcoHealth diagnosis.** Virginie Jouffret, réalisé par Lola Reynaud et Valérie Chazel de la communication de Marcoule.
  Congrès Global Young Scientists Summit (GYSS). 14 – 17 janvier 2020, Singapour.
  La vidéo a été traduite et publié sur la chaine YouTube CEA Sciences avec comme titre : « Sonder la vie microbienne du sol, un nouveau diagnostic d'éco-santé ».

- **La métaprotéomique et la Métagénomique : une alliance pour Explorer la santé des sols.** Virginie Jouffret
  Groupe de Travail CEA. 9 décembre 2020, en visioconférence.

2.6 Présentations posters

- **Deep exploration of the Microbial community in Sediments from the Seine River basin.**
  Virginie Jouffret, Guylaine Miotello, Sophie Ayrault, Jean Armengaud, Olivier Pible.
  Congrès Ecotoxicomic. 6 – 9 octobre 2020, Montpellier, France. (format vidéo)

- **GEOMICS – l'alliance de la GEOchimie et des approches OMICS.**
  Virginie Jouffret, Olivier Pible, Guylaine Miotello, Jean Armengaud, Sophie Ayrault, Matthieu Roy-Barman, Louise Bordier, Irène Lefèvre.
  DRF Impulsion. CEA

- **Evaluation of tools and workflows for database construction for environmental metaproteomics.**
  Virginie Jouffret, Olivier Pible, Yannick Cogne, Sophie Ayrault, Jean Armengaud.
  Congrès GCC. 1 – 6 juillet 2019, Freiburg, Allemagne.

- **Automatized functional annotation of virulence and resistance factors to complement phylopeptidomics results.**
  Virginie Jouffret
  Conférence CBRNE. 29 mai – 1er juin 2017, Lyon, France.

2.7 Liste des formations

- **Machine Learning.** Université de Stanford (Formation en ligne Coursera). *25 heures.*
- **Gérez votre temps efficacement**. Formation en ligne OpenClassroom. *12 heures*.
- **One-day workshop on Ethics in Biological Research.** Formation ADUM. *7 heures*.
- **Apprenez à coder avec Javascript.** Formation en ligne OpenClassroom. *20 heures*.
- **Environnement et santé : un homme sain dans un environnement sain.** Université Paris Cité (Formation en ligne FUN MOOC). *10 heures*.
- **Unlock your english.** Institut Mines-Télécom (Formation en ligne FUN MOOC). *10 heures*.
- **L'anglais pour tous – Spice up Your English.** Université libre de Bruxelles (Formation en ligne FUN MOOC). *25 heures*.

Résumé

Les sols représentent un habitat complexe et une matrice difficile pour leurs analyses omiques. Les communautés microbiennes qui y vivent se sont adaptées à leurs écosystèmes dont les caractéristiques physico-chimiques diffèrent selon la nature du sol, sa localisation et sa profondeur. Les sols de plaine d'inondation, formés à partir des sédiments déposés lors de chaque crue, drainent les alluvions du bassin versant avec qui les contaminants ont une forte affinité . Ces sols sont les témoins des activités anthropiques formant des archives des contaminations passées. Le choix d'un site de prélèvement où la sédimentation est régulière permet la datation des couches successives de ces archives. Les travaux de thèse ont pour objectif d'étudier les communautés microbiennes sous pressions anthropiques d'un point de vue temporel en reliant les changements de pression anthropique mesurés sur un site dont l'historique est bien connu et la structuration des communautés prélevées sur ces archives. Les approches de métaprotéomique basées sur des résultats massifs de spectrométrie de masse ultra-rapide permettent l'identification des organismes présents dans l'échantillon et d'apporter des données fonctionnelles uniques. Un effort particulier a été porté sur l'amélioration du pipeline d'interprétation des données basé sur des requêtes en cascade aboutissant à un taux d'identification bien supérieur au standard actuel. La combinaison de bases de données publiques et de données métagénomiques plus spécifiques de l'échantillon est une stratégie gagnante. Dans un second temps, une archive sédimentaire de la Seine, subdivisée en 35 couches datées, a été analysée de façon exhaustive par métallomique, métagénomique, métataxonomique, et métaprotéomique. Les micro-organismes identifiés et quantifiés par métaprotéomique ont conduit à des corrélations aux concentrations en éléments traces métalliques. Les stratégies et outils informatiques d'interprétation des données sur une carotte bien caractérisée au niveau géochimique ouvrent la voie à une application en diagnostic rapide des communautés microbiennes dans les sols.

Mots-clés : bioinformatique, métaprotéomique, géochimie, contaminants, sédiments, spectrométrie de masse en tandem, communautés, élément trace métallique, identification, micro-organismes

Abstract

Soil represents a complex habitat and a difficult matrix for omic analysis. The microbial communities that live in them are adapted to their ecosystems whose physico-chemical characteristics differ according to the nature of the soil, its location and its depth. Floodplain soils, formed from sediments deposited during each flood, drain the alluvium of the watershed with which contaminants have a strong affinity. These soils are the witnesses of anthropic activities and are archives of past contaminations. The choice of a sampling site where sedimentation is regular allows the dating of successive layers of these archives. The objective of this thesis is to study the microbial communities under anthropic pressure from a temporal point of view by linking the changes in anthropic pressure measured on a site whose history is well known and the structuring of the communities sampled on these archives. Metaproteomics approaches based on massive ultrafast mass spectrometry results allow the identification of organisms present in the sample and provide unique functional data. A particular effort has been made to improve the data interpretation pipeline based on cascading queries leading to an identification rate well above the current standard. The combination of public databases and more sample-specific metagenomic data is a winning strategy. In a second step, a Seine River sediment core, subdivided into 35 dated layers, was exhaustively analyzed by metallomics, metagenomics, metataxonomics, and metaproteomics. Microorganisms identified and quantified by metaproteomics led to correlations with trace metal elements concentrations. The strategies and computer tools for data interpretation on a well-characterized core at the geochemical level open the way to an application in rapid diagnosis of microbial communities in soils.

Keywords : bioinformatics, metaproteomics, geochemistry, contaminants, sediments, tandem mass spectrometry, communities, trace metals, identification, microorganisms