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

**En Sciences de l'Évolution et de la Biodiversité**

**École doctorale GAIA**

**Unité de recherche Centre d'Ecologie Fonctionnelle et Evolutive**

## Une approche éco-évolutive de la propagation de la résistance antibiotique : l'exemple de la résistance aux aminoglycosides


**Présentée par Léa PRADIER**
**Le 12 avril 2022**

**Sous la direction de Stéphanie BEDHOMME**


**Devant le jury composé de**

| | |
|---|---|
| Eric BAPTESTE, Directeur de Recherche, CNRS Paris | Rapporteur |
| Stéphanie BEDHOMME, Chargée de Recherche, CNRS Montpellier | Directrice |
| Aude BERNHEIM, Chargée de Recherche, INSERM Paris | Examinatrice |
| Sylvain GANDON, Directeur de Recherche, CNRS Montpellier | Président |
| Marisa HAENNI, Chargée de projet, ANSES Lyon | Examinatrice |
| Eduardo ROCHA, Directeur de Recherche, CNRS Paris | Rapporteur |

UNIVERSITÉ DE MONTPELLIER

# Remerciements



Dans un premier temps, je remercie les divers encadrants que j'ai pu avoir au cours de mes stages : Oliver Kaltz, James Gurney, Sylvain Gandon, Hélène Chabas. Ces derniers ont su susciter chez moi le goût de la recherche et stimuler une curiosité intellectuelle que seule une expérience académique peut éveiller. De plus, ces introductions bienveillantes au monde de la recherche ont grandement contribué à mon envie de me plonger plus intensément en biologie évolutive. Cette source de motivation a également participé à mon parcours académique et chacune de ces expériences a, à sa manière, servi de brique à la construction de l'édifice que représente cette thèse.

Un grand merci à Stéphanie pour m'avoir permis de me lancer dans cette aventure. Merci pour tes conseils avisés et tes corrections qui m'ont permis de raffiner mes analyses et d'augmenter la clarté de mes arguments tout au long de cette thèse. Merci à toi pour ton ouverture d'esprit. J'ai l'impression que nous avons pu apprendre l'une de l'autre, et j'espère pouvoir à nouveau avoir ce genre d'interaction.

J'ai eu la chance de collaborer avec des personnes qui ont su m'aiguiller avec bienveillance. Je remercie d'ailleurs chaleureusement Anna-Sophie Fiston-Lavier pour nos nombreuses conversations tant personnelles que professionnelles, et ses précieux conseils qui ont grandement aidé à l'amélioration de mes différents papiers. Je remercie également les membres de mon comité de suivi. Tout particulièrement Laurent Duret pour sa vision synthétique et ses questions pertinentes ; Guillaume Martin pour son enthousiasme et pour les nombreux conseils statistiques que lui et Sébastien Lion ont pu me donner.

Un grand merci aux membres de mon jury Aude Bernheim, Sylvain Gandon, Marisa Haenni, et en particulier aux rapporteurs Éric Bapteste et Eduardo Rocha, de me faire l'honneur d'examiner mon travail. J'espère que, malgré la charge de travail que ma thèse va représenter, vous prendrez plaisir à la lire.

Sur un plan plus personnel, je remercie ma famille et mes amis du CEFE et d'ailleurs qui m'ont permis (même à distance) de m'évader un peu pendant cette thèse.

Cécile et Lorelei, qui ont expérimenté en même temps que moi le marathon de la thèse et qui ont permis une entraide, rendant le parcours bien plus appréciable.

Christelle, pour son amour des chats, pour le soutien scientifique qu'elle m'a apporté.

Enrique, pour nos errances verbales, pour [insérer ici une blague incluant le mot patate].

Mon binôme Louise, pour nos nombreux projets partagés et finis dans le rush entre deux tasses de thé froides.

Camille, Lucie, Pierre$_1$, Pierre$_2$, Margaux, Maurine, et tous les autres... merci à vous d'être dans ma vie.

Enfin, cette thèse n'aurait pas été possible sans Tazzio Tissot.

Un long discours ne suffirait pas à exprimer ma reconnaissance. Au brouhaha, je préfère le silence.

# Table des matières


**I Introduction** **1**

**1 Emergence et conséquences de la résistance aux antibiotiques** **1**

1.1 Définition fonctionnelle : les antibiotiques et les résistances . . . . . 1

1.1.1 Que ciblent les antibiotiques ? . . . . . 1

a Inhibition de la synthèse de la paroi bactérienne . . . . . 1

b Action sur la membrane plasmique . . . . . 1

c Inhibition de la synthèse des acides nucléiques . . . . . 1

d Inhibition de la synthèse protéique . . . . . 2

e Inhibition du métabolisme des folates . . . . . 2

1.1.2 Comment les bactéries y échappent-elles ? . . . . . 2

1.2 Une histoire humaine de la résistance aux antibiotiques . . . . . 3

1.2.1 Utilisation des antibiotiques et émergence des résistances . . . . . 3

1.2.2 État des lieux actuel sur l'utilisation des antibiotiques . . . . . 5

1.2.3 Conséquences de la résistance pour la santé humaine . . . . . 6

**2 Evolution de la résistance aux antibiotiques** **6**

2.1 D'où proviennent les résistances aux antibiotiques ? . . . . . 6

a Des origines naturelles de la résistance . . . . . 7

b Des origines anciennes de la résistance . . . . . 7

2.2 La résistance sous l'effet des forces évolutives . . . . . 8

2.2.1 Le couple mutation-sélection . . . . . 8

a La mutation . . . . . 8

b La sélection . . . . . 9

c L'équilibre mutation-sélection . . . . . 10

2.2.2 Le couple dérive-migration . . . . . 10

a La dérive . . . . . 10

b La migration . . . . . 11

2.3 Recombinaison et transfert horizontal de gènes . . . . . 11

2.3.1 Mécanismes . . . . . 11

2.3.2 Succès du transfert horizontal et évolution post-transfert . . . . . 12

a Facteurs impactant le succès du transfert horizontal . . . . . 12

b Evolution post-transfert horizontal . . . . . 14

**3 Structuration, modulation, et dynamique de la résistance aux antibiotiques 14**
3.1 Régulation génétique et génomique du phénotype résistant . . . . . 14
3.1.1 Les contextes génomiques de la résistance . . . . . 15
a Les éléments génétiques mobiles . . . . . 15
b Interactions entre chromosome et éléments génétiques mobiles . . . 16
3.1.2 L'encodage des gènes de résistance . . . . . 16
a Origines du biais d'usage de codons . . . . . 17
b Impacts sur la résistance aux antibiotiques . . . . . 18
3.2 Problématiques écologiques de la résistance aux antibiotiques . . . . . 18
3.2.1 Rôle et impacts de la résistance dans les communautés bactériennes . . . . . 19
a Propagation des résistances dans les communautés . . . . . 19
b Les conséquences écologiques de la résistance . . . . . 19
3.2.2 Échanges entre écosystèmes naturels et écosystèmes anthropisés . . . . . 19
a Antibiotiques et pollutions . . . . . 20
b Des bactéries résistantes jusqu'aux réservoirs de résistance . . . . . 20
3.3 Les sociétés humaines face aux résistances antibiotiques . . . . . 21
3.3.1 Historique de la réponse humaine aux résistances antibiotiques . . . . . 21
a De 1945 à 1963 : de l'optimisme face à un problème mineur . . . . . 22
b De 1963 à 1981 : une préoccupation croissante . . . . . 23
c De 1981 à 1992 : une prise de conscience mondiale . . . . . 23
d De 1992 à 2013 : une accélération des initiatives . . . . . 23
e Depuis 2013 : un problème aussi important que le changement climatique . . . . . 24
3.3.2 De la santé individuelle aux approches One health et Global health . . . . . 25

**II Problématique générale de la thèse 27**

**III Caractérisation des éléments génétiques mobiles dans grands jeux de données génomiques 29**

**IV Le rôle de la mobilité dans la propagation des gènes de résistance aux antibiotiques 48**

**V Structuration phylogénétique, écologique, et génomique du transfert horizontal des gènes de résistance 89**

**VI Discussion** **121**

**4 Structuration environnementale de la résistance** **121**
4.1 Les antibiotiques anthropogéniques parmi d'autres forces . . . 121
4.2 La connectivité écologique et géographique . . . 123
4.2.1 Mondialisation et connectivité géographique . . . 123
4.2.2 Flux naturels et anthropiques entre écosystèmes . . . 124
4.2.3 La modulation écologique du transfert horizontal de gènes . . . 125
4.2.4 Les bactéries généralistes, des navettes pour la résistance ? . . . 126
4.3 Origines et circulation des résistances antibiotiques . . . 127
4.3.1 Les environnements naturels sont-ils des réservoirs de résistance ? . . . 127
4.3.2 Dans quelle(s) direction(s) les résistances se propagent-elles ? . . . 128

**5 Structuration génomique et phylogénétique de la résistance** **129**
5.1 Une résistance ubiquitaire chez les bactéries . . . 129
5.1.1 Les AME sont présents dans toute la phylogénie des Eubactéries . . . 129
5.1.2 Une distribution différente en fonction des familles de gènes . . . 130
5.1.3 Le transfert horizontal des résistances se fait-il à large échelle ? . . . 131
5.2 Le rôle structurant des éléments génétiques mobiles . . . 132
5.2.1 Diversification fonctionnelle et émergence de la multirésistance . . . 133
5.2.2 Des facilitateurs du transfert horizontal . . . 133
5.3 Les préférences d'usage de codons : une barrière au transfert horizontal des résistances ? 134
5.3.1 Une barrière entre sélection diversifiante et sélection stabilisante . . . 135
5.3.2 Une barrière qui peut être contournée . . . 136

**6 Vers de nouvelles approches de gestion de la résistance antibiotique** **136**
6.1 Dépasser une conception adaptationniste de la résistance . . . 137
6.2 Intégrer la gestion écologique dans la santé publique . . . 137
6.2.1 Vers une surveillance accrue des écosystèmes naturels . . . 138
6.2.2 Vers une réduction de la connectivité écologique . . . 139
6.2.3 Vers la décontamination des écosystèmes ? . . . 139
6.3 Un problème global nécessitant une coordination globale . . . 140

**7 Limitations méthodologiques** **141**
7.1 Les limites de la détection bioinformatique . . . 141
7.2 Des jeux de données biaisés et incomplets . . . 142

**8 Perspectives** **143**

8.1 L'apport des données métagénomiques à l'étude de la propagation des résistances antibiotiques . . . . . 143
8.2 Le rôle du système CRISPR-Cas dans le transfert horizontal de gènes . . . . . 144
8.3 Modéliser la propagation d'un gène de résistance dans une métacommunauté . . . . . 145

**VII Bibliographie 147**

**VIII Liste des abréviations 168**

**IX Annexes 169**

**9 Annexe 1 : Informations supplémentaires de la section IV 169**

**10 Annexe 2 : Informations supplémentaires de la section V 212**

**11 Annexe 3 : Sources de sélection sur le biais d'usage de codons 239**

# Section I : Introduction

## 1 Emergence et conséquences de la résistance aux antibiotiques

### 1.1 Définition fonctionnelle : les antibiotiques et les résistances

#### 1.1.1 Que ciblent les antibiotiques ?

Un antibiotique est un composé qui tue les bactéries ou bloque leur croissance. La plupart de ces molécules sont naturelles et synthétisées par une diversité d'organismes (plantes, champignons, procaryotes), mais il existe aujourd'hui sur le marché un grand nombre de molécules de synthèses. Le principe d'action de ces molécules consiste à bloquer une ou plusieurs étapes du cycle cellulaire de la bactérie, essentielle à sa survie et/ou sa multiplication. Il existe une grande variété de mécanisme d'action pour les antibiotiques (figure 1).

##### a Inhibition de la synthèse de la paroi bactérienne

Les bactéries Gram-positives sont protégées des stress environnementaux par une membrane plasmique, mais aussi grâce à une paroi de peptidoglycanes. Cette dernière doit croître quand la bactérie se divise, et nécessite donc une machinerie de synthèse des composants et un système de transport vers la surface cellulaire. Certains antibiotiques, dont les représentants principaux sont les β-lactamines (pénicilline, amoxicilline, etc.), bloquent la production et/ou l'acheminement de ces composants. Une membrane ainsi incapable de se renouveler se fragilise, ce qui amène in fine à la lyse cellulaire.

##### b Action sur la membrane plasmique

L'existence d'une membrane plasmique intacte est nécessaire à la survie bactérienne. Son rôle est double. D'une part, elle permet de séquestrer métabolites et ions nécessaires à l'intérieur du cytoplasme. D'autre part, elle permet de maintenir un gradient de protons entre l'intérieur et l'extérieur de la cellule, généré par la chaîne respiratoire et le cycle de Krebs, et qui permet le stockage de l'énergie cellulaire. Un certain nombre de molécules antibiotiques agissent sur la membrane des cellules, soit en agissant comme des détergents qui désorganisent les lipides (par exemple la polymyxine), soit en formant un pore dans la membrane qui va permettre la fuite des composés cellulaires (par exemple la gramicidine).

##### c Inhibition de la synthèse des acides nucléiques

La synthèse des acides nucléiques (ADN et ARN) est absolument vitale pour les cellules. Sans elle, la division cellulaire et la fabrication des protéines est impossible. C'est pourquoi un certain nombre d'antibiotiques peuvent bloquer de manière directe ou indirecte les voies de synthèse des acides nucléiques.

Les quinolones peuvent, par exemple, bloquer l'action des gyrases, et ainsi provoquer une perte du surenroulement de l'ADN essentielle à la réplication de ce dernier. D'autres, tels que la mitomycine, peuvent bloquer la réplication en créant des pontages entre bases voisines, créant une déformation de l'ADN qui ne peut plus s'ouvrir pour se répliquer. Enfin, il existe des inhibiteurs spécifiques de l'ARN polymérase bactérienne qui bloquent la transcription des gènes et la synthèse des ARN messagers. Parmi ces antibiotiques, on trouve en particulier la rifampicine qui est aujourd'hui utilisée en association avec d'autres antibiotiques pour le traitement de la tuberculose.

#### d Inhibition de la synthèse protéique

Le processus de synthèse protéique est essentiel au bon fonctionnement cellulaire. Ce dernier consiste en la traduction de l'ARN messager en protéine par le ribosome. Le processus de traduction étant sensiblement différent entre procaryotes et eucaryotes, environ la moitié des antibiotiques thérapeutiques utilisent ces différences pour ne cibler que les cellules bactériennes. Les modes d'actions peuvent cependant varier. De nombreuses classes d'antibiotiques ciblent ce processus cellulaire (phénicolés, puromycine, acide fusidique), mais les classes les plus importantes (aminoglycosides, macrolides, cyclines) interagissent avec l'ARN ribosomique : par exemple, les aminoglycosides se fixent sur la petite sous-unité des ribosomes au niveau du site du décodage des codons, empêchant ainsi la traduction de l'ARNm et conduisant à des erreurs de lecture.

#### e Inhibition du métabolisme des folates

Les folates ou vitamine B9 interviennent dans la régulation de voies particulièrement importantes à la production de métabolites essentiels pour la cellule comme les lipides, les acides aminés, ou les nucléotides. En effet, ses dérivés (dihydrofolate et tétrahydrofolates) interviennent dans le transfert de groupements (en particulier dans le processus de méthylation), impactant in fine la formation de bases azotées (comme la thymine) ou la formation de certains acides aminés (méthionine, glycine, sérine). Différentes étapes de la voie de biosynthèse des folates peuvent être ciblées, mais par exemple, les sulfamidés peuvent constituer des inhibiteurs compétitifs dans la synthèse des folates.

### 1.1.2 Comment les bactéries y échappent-elles ?

Face à la pression que représentent les antibiotiques, une grande variété de mécanismes de résistance aux antibiotiques existent. On oppose deux grands types de résistance : (i) la résistance intrinsèque et (ii) la résistance acquise. Dans le premier cas, la résistance est une exaptation : un processus d'adaptation d'un protéine existante, selon lequel une ou plusieurs fonctions s'ajoutent à sa fonction d'origine, ou la remplacent. Par exemple, la membrane externe présente chez les bactéries Gram-négatives leur confère également une résistance à de nombreux antibiotiques. Au contraire, les résistances acquises constituent une adaptation : une modification héréditaire permettant une meilleure adéquation de l'or-

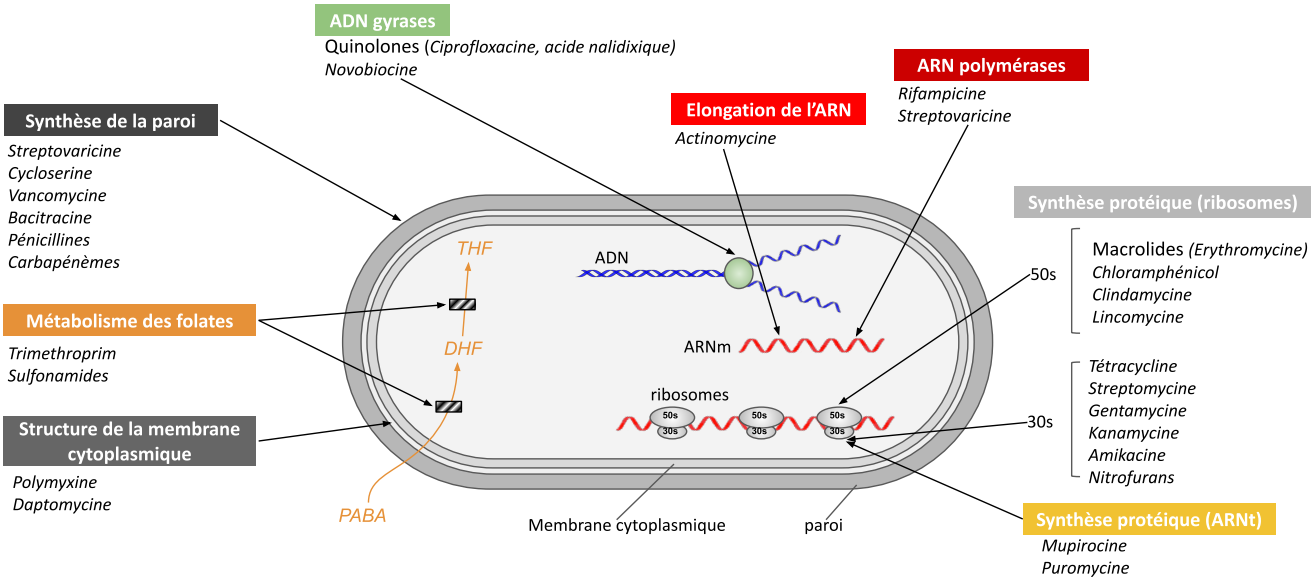


**FIGURE 1 –** Cibles des antibiotiques.

ganisme à son environnement et évoluée par un processus de sélection naturelle. Par exemple, certaines mutations de l'ARN ribosomique 30S chez *Mycobacterium tuberculosis* empêchent la fixation de la streptomycine (DOUGLASS et STEYN 1993).

Pour autant, les résistances antibiotiques empêchent l'antibiotique d'entraver les processus physiologiques bactériens (BLAIR et al. 2015, figure 2) : soit en développant des voies enzymatiques alternatives (voir l'exemple des résistances aux pénicillines, (LIM et STRYNADKA 2002), soit en empêchant le contact entre l'antibiotique et sa cible. Dans ce dernier cas, la cible peut être exprimée en excès par rapport à l'antibiotique, ou simplement être modifiée par mutation (DOUGLASS et STEYN 1993) ou de manière secondaire par une enzyme (par exemple via une méthylation de l'ARNr 23s, voir LONG et al. 2006). De plus, il est également possible de diminuer la perméabilité membranaire pour prévenir l'entrée de l'antibiotique (par exemple, via une régulation à la baisse des porines voir TAMBER et HANCOCK 2003), de l'expulser grâce à des pompes d'efflux (FLOYD et al. 2010), voire encore de l'altérer chimiquement de sorte à lui enlever toute efficacité (voir l'exemple des $\beta$-lactamases, LIVERMORE 2008).

## 1.2 Une histoire humaine de la résistance aux antibiotiques

### 1.2.1 Utilisation des antibiotiques et émergence des résistances

L'utilisation anthropique des antibiotiques commence avec Paul Ehrlich. Ce dernier propose l'idée de composés chimiques qui pourraient détruire spécifiquement un parasite sans affecter son hôte. En 1904, lui et ses collaborateurs se lancent dans une expérience à grande échelle sur la syphilis. Leurs projet consiste à produire des dérivés organo-arsenicaux de l'atoxyl (un dérivé de l'arsenic utilisé au 19ème siècle pour lutter contre *Trypanosoma gambiense*) et de les tester ensuite sur des lapins infec-

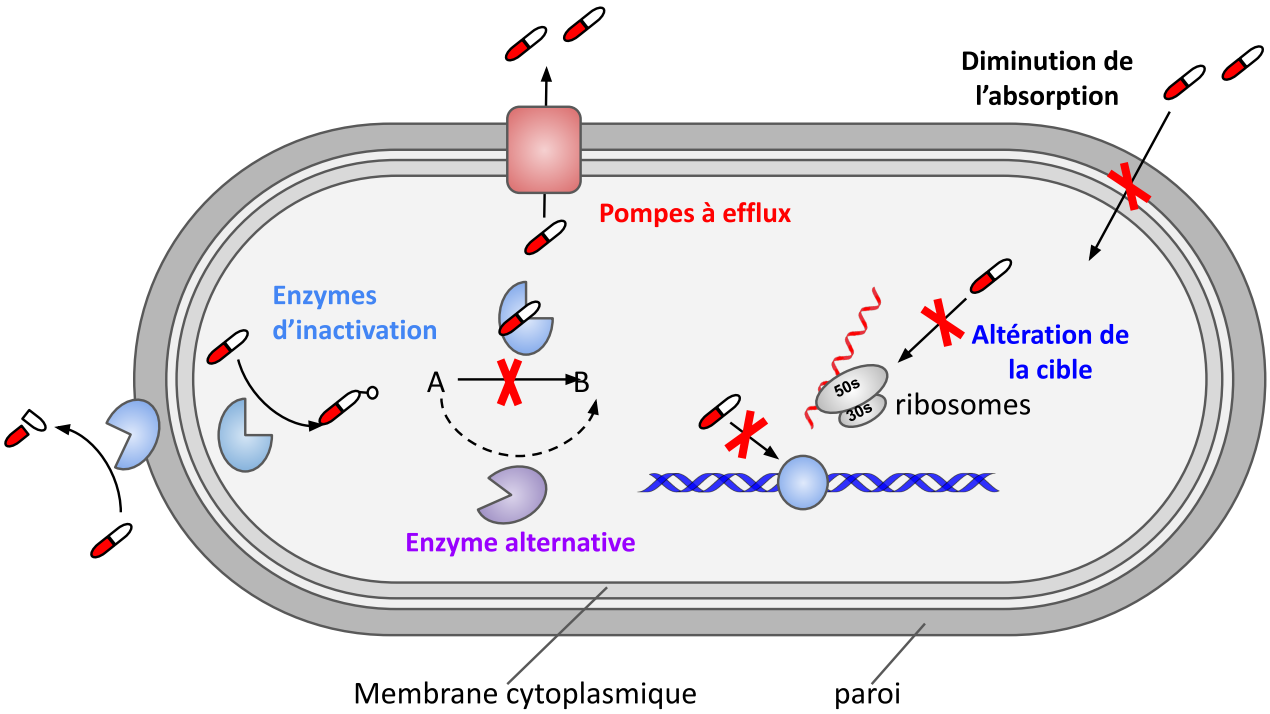


**FIGURE 2 –** Modalités de la résistance antibiotique.

tés. En 1910, ils trouvent une molécule qui semble soigner les lapins mais donne aussi des résultats prometteurs sur l'Homme, l'arsphénamine (EHRLICH et HATA 1910). Malgré la procédure contraignante d'injection, il reste le médicament le plus utilisé jusqu'à son remplacement par la pénicilline dans les années 1940. La stratégie de criblage biochimique employée par Ehrlich a ensuite été beaucoup utilisée par l'industrie pharmaceutique et a notamment contribué à la découverte des premières molécules antimicrobiennes en 1935 : les sulfonamides. Le coût de production de ces molécules étant relativement peu cher, leur utilisation a été massive. Il en a résulté un des cas de résistance aux médicaments les plus largement diffusés : la résistance aux sulfamidés, qui est presque universellement liée aux intégrons de classe 1 (MACHADO et al. 2005). Pourtant, c'est avec la découverte de la pénicilline par Alexander Fleming en 1929, et la publication du protocole de purification de Florey et Chain en 1945, que la production et la distribution de masse d'antibiotiques a débuté. Bien que des résistances aient déjà été répertoriées dès la fin des années 1940, les années 1950 à 1970 constituent l'âge d'or de la découverte de nouvelles classes d'antibiotiques. Cependant, à partir des années 1980, le taux de découvertes a décliné. L'approche dominante pour le développement de nouveaux médicaments, notamment pour lutter contre la résistance émergente et ré-émergente des agents pathogènes aux antibiotiques, a été la modification des antibiotiques existants. Cela signait ainsi l'entrée dans une course aux armements contre la résistance (CHOPRA 2002 ; PODOLSKY 2018).

L'utilisation massive des antibiotiques par l'Homme est assez récente : seulement depuis les années 1940. Le problème de la résistance a pu être contourné pendant quelques décennies par la découverte

de nouvelles molécules, puis par la modification des chaînes carbonées (Lewis 2013). Cependant, à l'heure actuelle, la grande majorité des possibilités de modifications ont été explorées et cette source de nouveaux antibiotiques se tarit. En effet, les structures moléculaires de ces composés sont quasiment identiques. En outre, de plus en plus de souches bactériennes, dites *superbugs*, résistent à différentes classes d'antibiotiques en possédant dans leurs génomes des combinaisons de gène de résistance (Martínez 2007).

Pour autant, certaines approches plus récentes tentent de contourner le problème en utilisant soit du criblage de molécules extraites de plantes et de champignons (McCutcheon et al. 1992), soit des approches de biologie synthétique visant l'identification de nouvelles voies métaboliques jusque là peu touchées par les antibiotiques connus (par exemple la synthèse des lipides, voir Fernández-Oliva et al. 2019). Cependant, même en exploitant de nouvelles molécules et en en faisant une utilisation raisonnable, rien ne garantit à long terme que la multirésistance sera évitée. Il devient donc urgent de comprendre les causes de l'évolution de la résistance aux antibiotiques, aussi bien les mécanismes d'émergence que de diffusion, pour pouvoir mettre en place des stratégies de limitation.

### 1.2.2 État des lieux actuel sur l'utilisation des antibiotiques

Aujourd'hui, les antibiotiques sont encore utilisés dans le traitement de maladies infectieuses liées à des pathogènes bactériens. Leur utilisation est assez régulée dans les pays européens (Mueller et Östergren 2016), mais reste plus difficilement traçable dans les pays en voie de développement (Mitchell et al. 2020). Malgré des différences de collecte et de traitement des données sur ce sujet, deux grandes tendances semblent être opposées (Klein et al. 2018) : une diminution des consommation et des ventes d'antibiotiques via des programmes de réduction dans plusieurs pays industrialisés ; et au contraire une tendance à la hausse dans les pays à revenus faibles ou intermédiaires. Ces tendances peuvent en partie être expliquées par le fait que la consommation d'antibiotiques est un problème multidimensionnel ne pouvant se réduire aux problèmes médicaux.

Les instituts de santé publique, comme l'Organisation mondiale de la santé (WHO, OMS) ou le Centre européen de prévention et de contrôle des maladies (ECDC), se sont emparés de la question. Les directives générales de l'OMS en 2015 s'appuient sur trois axes : optimiser l'usage des antibiotiques chez l'humain et l'animal et éduquer les populations au bon usage des antibiotiques ; réduire l'incidence des infections en améliorant l'hygiène et les mesures prophylactiques ; augmenter l'investissement dans les vaccins, les traitements alternatifs, et les outils diagnostics (WHO 2015).

Cependant, au niveau national, les politiques des pays industrialisés tendent majoritairement à restreindre la prescription des antibiotiques (Mueller et Östergren 2016 ; Sabuncu et al. 2009). Ces politiques sanitaires semblent pourtant ne pas être les seules à permettre une diminution de l'utilisation : en effet, en 2020, au plus fort de la crise du COVID, les confinements et l'utilisation de mesures d'hygiène ont permis aux prescriptions d'antibiotiques de diminuer en Europe (Rosier 2021). De plus, l'utilisation

médicale anthropique n'est aujourd'hui pas la plus problématique. En effet, deux-tiers des antibiotiques sont utilisés en élevage et pisciculture (Done et al. 2015). Il est connu que les antibiotiques peuvent dans ces milieux être utilisés soit dans un but thérapeutique soit comme facteurs de croissance. Une part non négligeable de leurs utilisation est donc liée à des considérations économiques et non médicale. Il reste cependant compliqué de les interdire complètement ou de contrôler fermement leurs distributions, même si certains pays ont déjà légiférés contre leurs utilisations comme facteurs de croissance.

### 1.2.3 Conséquences de la résistance pour la santé humaine

La résistance antibiotique est un problème de santé publique persistant, en augmentation (Witzany et al. 2020). Pour la seule année 2019, il a été estimé 1.3 millions de décès attribuables aux bactéries résistantes (Murray et al. 2022) et l'*AMR review* en prévoit 10 millions par an d'ici 2050, dépassant largement les décès dus au cancer (O'Neill 2014, mais voir Kraker et al. 2016). Pourtant, les antimicrobiens qui tuent ou inhibent les maladies infectieuses sont des outils cliniques essentiels et la résistance aux antibiotiques rend déjà les opérations chirurgicales de routine et les visites à l'hôpital de plus en plus risquées. L'épidémie est particulièrement problématique dans les établissements de soins de longue durée, où plus de 25 % des infections nosocomiales sont causées par des bactéries résistantes aux antibiotiques (O'Neill 2014). Si ces observations se maintiennent, plusieurs prédictions ont été formulées :

- des difficultés pour les opérations, les chirurgies, avec une surmortalitée liée aux septicémies et/ou aux surinfections ;
- une surmortalité pour les patients immunodéprimés qui sont déjà plus susceptibles d'être exposés à des environnements cliniques (Conlan et al. 2014) ;
- une potentielle modification des microbiomes humains pouvant favoriser l'acquisition de différentes maladies chez l'homme (Livanos et al. 2018) et d'autres hôtes (par exemple, le porc ou les abeilles, Raymann et al. 2017).

# 2 Evolution de la résistance aux antibiotiques

Derrière la résistance aux antibiotiques se cachent un certain nombre de caractères phénotypiques héritables, dont la fréquence varie d'une population bactérienne à une autre. La biologie de l'évolution peut étudier ces caractères et ainsi aider à retracer les origines de ces caractères, et les phénomènes qui favorisent leur maintien et leur rapide propagation au sein des communautés bactériennes du monde entier.

## 2.1 D'où proviennent les résistances aux antibiotiques ?

La résistance est souvent considérée comme un mécanisme récent, dont l'émergence est corrélée à l'utilisation des antibiotiques par l'homme. En effet, de nombreux exemples d'évolution rapide sont

présents dans la littérature (Perron et al. 2005). De plus, il est possible de ré-évoluer très rapidement des résistances en laboratoire par sélection (Lukačišinová et al. 2020) . Pourtant, même si la résistance est un mécanisme pouvant émerger rapidement et facilement, on n'a été en mesure de la détecter qu'à partir du moment où les antibiotiques ont commencé à être utilisés pour traiter des infections. La résistance a donc pu émerger beaucoup plus tôt, et il ne faut pas confondre son émergence chez les bactéries en général de la résistance et sa propagation aux bactéries infectant l'homme. Il est actuellement consensuel que l'apparition de la résistance, en plus d'avoir eu lieu de multiples fois, est un phénomène ancien et naturel.

### a Des origines naturelles de la résistance

Les antibiotiques sont des molécules produites par certaines bactéries comme molécules de signalisation ou pour réguler les populations. On pense donc que l'origine des gènes de résistance s'est probablement faite chez des bactéries sauvages non pathogènes pour se protéger des antimicrobiens autoproduits ou produits par d'autres membres de la communauté bactérienne (Benveniste et Davies 1973 ; Cox et Wright 2013). Cependant, même si l'émergence de la résistance est multiple et ancienne, il reste certain que la découverte puis l'utilisation massive en agriculture et en médecine des antibiotiques a considérablement augmenté les prévalences de ces phénotypes. On note par exemple une augmentation allant jusqu'à 1500 % des résistances dans les sols depuis les années 1940 (Knapp et al. 2010).

### b Des origines anciennes de la résistance

Bien que souvent étudiée en milieu anthropisé et sur des périodes assez récentes, la résistance aux antibiotiques (et par conséquent la constitution de résistomes) est un processus bien antérieur à l'ère d'utilisation anthropique des antibiotiques (Aminov et Mackie 2007 ; Kobayashi et al. 2007). Actuellement, on sait que les résistances antibiotiques remontent à au moins 30 000 ans, par la datation d'ADN ancien. Par exemple, une étude du permafrost nord-canadien (D'Costa et al. 2011) a mis en évidence la présence de gènes encodant des résistances aux β-lactamines, à la tétracycline, et aux antibiotiques glycopeptides, dans des métagénomes, co-occurrents avec de l'ADN d'animaux propres au Pléistocène. Des observations similaires ont pu être faites dans une étude de métagénomique effectuée sur le permafrost sibérien, même si les résistances retrouvées n'ont été datées que de 5000 ans (Petrova et al. 2009).

Le portage de résistances par des bactéries du microbiote ou de pathogènes humains est également attesté bien avant la découverte et l'utilisation massive des antibiotiques. Par exemple, des gènes homologues à ceux encodant des résistances aux β-lactamines, à la fosfomycine, au chloramphénicol, aux aminoglycosides, aux macrolides, aux sulfamidés, aux quinolones, à la tétracycline et à la vancomycine, ont été retrouvés chez une momie péruvienne précolombienne (Santiago-Rodriguez et al. 2015). De même, les premières occurrences de la plupart des résistances étudiées sont associées à des infections

chez des humains ou des animaux (Ebmeyer et al. 2021). L'évolution de la résistance est donc bien antérieure à l'âge d'or de l'utilisation anthropique des antibiotiques (Martinez 2009).

## 2.2 La résistance sous l'effet des forces évolutives

La découverte de la résistance aux antibiotiques s'est faite à partir du moment où les antibiotiques ont commencé à être utilisés par l'Homme. De plus, on a constaté une augmentation de la prévalence des résistances avec l'augmentation de la consommation d'antibiotiques. La résistance aux antibiotiques a donc essentiellement été considérée sous l'angle de la sélection par les antibiotiques. Cependant, d'autres forces évolutives participent à moduler l'émergence et la propagation des résistances.

### 2.2.1 Le couple mutation-sélection

La sélection et la mutation sont les forces évolutives dont le rôle dans l'émergence, le maintien et la propagation de la résistance, a été le plus décrit et discuté.

#### a La mutation

L'allèle d'un gène conférant une résistance à un ou plusieurs antibiotiques peut émerger par une simple mutation ponctuelle : substitution, insertion, ou délétion (voir par exemple (Wang, Wilson et al. 2001 chez *Helicobacter pylori*). Ces événements peuvent arriver dans une très grande diversité de gènes, dont plus particulièrement des gènes du métabolisme énergétique (Lopatkin et al. 2021). De plus, certains gènes confèrent à leurs porteurs une susceptibilité à certains antibiotiques. L'insertion d'un transposon dans ces gènes, perturbant ainsi leur expression, peut donc également conférer une résistance à ces antibiotiques (Poros-Gluchowska et al. 2003 chez *Listeria monocytogenes*). Toutefois, les taux de mutations peuvent varier chez les bactéries, et il serait donc possible pour une lignée de moduler ainsi le taux d'apparition de nouveaux variants génétiques conférant potentiellement des résistances aux antibiotiques. En effet, le système MMR (mismatch repair system) est crucial pour éviter les mutations et maintenir la fidélité réplicative. Dans les populations bactériennes sauvages, la majorité des isolats ont de faibles taux de mutation, ce qui suggère que la stabilité génétique à long terme est favorisée (Sniegowski et al. 2000). L'inactivation d'un ou plusieurs gènes du MMR diminue la capacité à réparer les dommages à l'ADN et les erreurs de réplication, et les bactéries deviennent ainsi hypermutatrices. De tels phénotypes mutateurs ne sont pas sélectionnables directement mais, comme un taux de mutation accru peut faciliter une adaptation rapide à de nouveaux environnements, ils peuvent être co-sélectionnés par association avec des mutations favorables, telles que la résistance aux antibiotiques (Chao et Cox 1983 ; Taddei et al. 1997). Il a donc été suggéré que l'hypermutabilité pourrait avoir un rôle prépondérant dans l'émergence de la résistance (voire de la multirésistance) dans certaines espèces bactériennes (Blázquez et Gómez-Gómez 2014), notamment chez *Mycobacterium tuberculosis*. Cependant, si des taux de mutation élevés peuvent s'accompagner de nombreux avantages compéti-

tifs à court terme, ces avantages s'estompent une fois que la population est installée, voire deviennent maladaptatifs lors de changements d'environnements (Giraud et al. 2001).

### b La sélection

La mutation permet l'émergence du phénotype résistant, mais encore faut-il que ce dernier persiste. En effet, un gène de résistance peut conférer un coût à la bactérie qui le porte : par exemple, l'ARNt muté de *M. tuberculosis* lui confère certes une résistance aux aminoglycosides, mais réduit également sa vitesse de traduction (Douglass et Steyn 1993). Et on peut donc prédire que les gènes de résistances s'éteindraient en l'absence d'une sélection positive. Aussi, plusieurs mécanismes peuvent permettre le maintien de génotypes résistants, parmi lesquels le plus étudié est la sélection. Dans ce cadre, l'exposition à des antibiotiques cytotoxiques ou cytostatiques agit comme un filtre, permettant à celles portant des résistances de se perpétuer plus que les autres phénotypes. Cependant, s'il a été considéré pendant longtemps que l'inhibition bactérienne (et donc la mesure de concentrations inhibitrices minimales, MICs) était nécessaire à la sélection, on sait désormais que de simples résidus d'antibiotiques, à des concentrations 100 fois plus faibles que la MIC, suffisent à sélectionner des résistances (d'où l'émergence du concept de concentration sélective minimale, Gullberg et al. 2011). Du reste, la sélection s'observe toujours à l'échelle de la population, et plusieurs scénarios s'écartent du simple modèle émergence-sélection.

Dans un premier scénario, la sélection peut favoriser le maintien de phénotypes généralistes, par exemple lorsque les bactéries sont exposées à plusieurs antibiotiques. Dans ce cas, les bactéries qui présentent le spectre de résistance le plus large sont les phénotypes majoritaires de la population (par exemple chez les souches multirésistantes de *Staphylococcus aureus*, Mwangi et al. 2007). Sous ce scénario, une sélection directionnelle pour une diversification du spectre de résistance serait observée.

Un autre scénario est celui d'une sélection fréquence-dépendante. En effet, si les concentrations en antibiotiques fluctuent dans l'environnement, il peut être bénéfique de conserver un phénotype sensible et un phénotype résistant dans la population : les uns étant avantagés en l'absence d'antibiotique, les autres quand la concentration réaugmente. Ce type de sélection peut également se reposer sur des stratégies collectives où la résistance se comporte comme un bien public : par exemple, par la mise en place de biofilms protecteurs (Hall et Mah 2017) ou encore par la dégradation par les phénotypes résistants des antibiotiques présents dans le milieu (Dugatkin et al. 2005).

Un dernier cas beaucoup moins considéré est celui d'une plasticité phénotypique adaptative, lorsqu'un génotype est capable de produire différents phénotypes. Cette capacité peut permettre la conservation de la viabilité malgré les changements environnementaux. Par exemple, chez les bactéries, il existe beaucoup de switches métaboliques permettant de passer d'actifs à inactifs (Brückner et Titgemeyer 2002). Ainsi, l'entrée dans un état quiescent lors de l'exposition aux antibiotiques permet à

certaines souches bactériennes d'y survivre, ce qui fait de ces *persisters* un problème tout particulier en milieu clinique (Bartell et al. 2020). La survie de ces lignées s'accompagne d'une plus grande chance d'accumuler des mutations sur un temps long, et donc in fine d'acquérir une résistance à ces antibiotiques (Levin-Reisman et al. 2017). De même, la résistance antibiotique peut évoluer sous un pression de sélection autre que la pression de sélection antibiotique : ainsi, chez *Escherichia coli*, la tolérance à une grande gamme de conditions de pH confère une résistance constitutive aux antibiotiques (Mueller, Egan et al. 2019).

#### c L'équilibre mutation-sélection

Bien qu'il soit intéressant d'étudier la part de chaque force dans l'évolution du phénotype résistant, mutation et sélection n'agissent jamais de manière indépendante et il reste crucial pour notre compréhension des systèmes biologiques d'étudier les relations et équilibres entre ces dernières. Des approches de modélisation se sont cependant penchées sur la question et il semble que l'équilibre possible entre la mutation et la sélection dépend majoritairement d'un trade-off entre la diversification et l'unification (Novak et al. 2013). D'une part, la mutation permet de générer de la variance, renforçant ainsi l'évolvabilité de la population, puisque la sélection se nourrit de la variabilité phénotypique. D'autre part, la mutation peut aussi faire disparaître des allèles avantageux, et ainsi défaire l'action de la sélection. Cette théorie permet de décrire deux tendances évolutives opposées. D'un côté, l'hypermutation et la recombinaison peuvent être interprétées comme une stratégie de diversification extrême. De l'autre, la sélection stabilisante tend à renforcer les structures phénotypiques déjà sélectionnées par le passé, et peut donc être perçue comme une conservation extrême, freinant la diversification. Il est possible qu'un équilibre évolutif se mette en place entre ces deux tendances dans le cas de la résistance aux antibiotiques (Reding-Roman et al. 2017).

### 2.2.2 Le couple dérive-migration

Même si la sélection et la mutation sont les forces évolutives les plus étudiées dans le cycle de la résistance antibiotique, elles ne sont pas les seules à permettre l'émergence, le maintien, et surtout la propagation de la résistance. La dérive et la migration ont été beaucoup moins étudiées pour deux raisons. D'une part, on considère généralement que les bactéries ont des tailles efficaces très importantes, ce qui minimise l'impact de la dérive par rapport à celui de la sélection sur l'évolution des caractères. D'autre part, la question de la résistance a longtemps été posée dans le secteur médical uniquement, ce qui a pu conduire à négliger l'importance de la migration dans l'apport et le maintien des variants résistants (Finley et al. 2013).

#### a La dérive

Cette force a longtemps été ignorée dans l'évolution bactérienne. En effet, la notion de taille efficace est très liée à l'impact que peut avoir la dérive sur une population. La notion de taille efficace — c'est-

à-dire le nombre d'individus reproducteurs dans une population théorique (qui montrerait la même dispersion des fréquences alléliques sous dérive qu'une population observée, Wright 1931) — a été définie dans les modèles classiques sur des populations eucaryotes à reproduction sexuée. Pourtant, les effets de la dérive sur les procaryotes résistants sont observables. Des études d'évolution expérimentales ont suggéré que des variants résistants aux antibiotiques peuvent évoluer même en l'absence d'antibiotiques, sous l'effet de l'adaptation génétique des bactéries à diverses conditions de croissance (Knöppel et al. 2017). Par ailleurs, des variants résistants présents en faible fréquence dans une population peuvent ensuite se fixer dans la population suite à un goulot d'étranglement (Mahrt et al. 2021), ou par auto-stop suite à un balayage sélectif sur un locus lié génétiquement (co-sélection de la résistance aux antibiotiques avec la résistance aux métaux lourds, Baker-Austin et al. 2006). De même, chez les bactéries pathogènes, lors du cycle d'infection-transmission-dissémination intra- et inter-hôte, de petites populations peuvent se fragmenter en fonction des ressources disponibles, des tissus d'implantation, ou encore par la formation de biofilms (voir par exemple Jorth et al. 2015) sur des patients atteints de mucoviscidose). Cette organisation en métapopulations autorise potentiellement une exploration plus large du paysage adaptatif – à la fois du fait de l'exposition à une gamme plus large de conditions environnementales et du fait de la dérive – et ainsi de maximiser les chances d'envahissement de la population la mieux adaptée à l'hôte par migrations entre populations (Weinreich et Chao 2005).

#### b La migration

Comme expliqué précédemment, la dérive est un terreau fertile à la migration permettant l'exploration et l'établissement via des flux migratoires réguliers de populations dans de nouveaux environnements. En effet, il semble que même au sein des hôpitaux, la connectivité entre hôpitaux peut être un facteur prédictif majeur de la distribution des bactéries résistantes (Shapiro et al. 2020). De plus, des effets fondateurs ont déjà été démontrés chez les bactéries dans le cas de la résistance aux antibiotiques. Par exemple, bien qu'il existe partout dans le monde une très grande diversité de bêta-lactamases, la bêta-lactamase TEM-1 est restée pendant des années une des sources de résistance prédominantes chez les *Enterobacteriaceae*. Ceci peut s'expliquer par l'émergence précoce de cette enzyme peu après l'introduction des bêta-lactamines, et par conséquent son fort succès évolutif à cette époque (Medeiros 1997).

## 2.3 Recombinaison et transfert horizontal de gènes

### 2.3.1 Mécanismes

Le transfert horizontal de gènes (HGT, horizontal gene transfer) se définit comme la transmission de matériel génétique d'un organisme à un autre, ou d'une lignée à une autre, sans relation d'ancêtre à descendant (Dagan et al. 2008 ; Ochman et al. 2000 ; Sorek et al. 2007). En une seule génération,

une lignée peut ainsi acquérir un gène paralogue (lorsque ce gène et un gène du génome récepteur partagent un ancêtre commun) ou xénologue (lorsque ce gène ne possède pas d'ancêtre commun avec un gène du génome récepteur). Trois mécanismes de HGT ont été identifiés (Furuya et Lowy 2006, voir figure 3) :

— La transformation correspond à la simple récupération par une bactérie réceptrice, d'un morceau d'ADN libre (Griffith 1928).
— La transduction met en jeu un virus vecteur, qui permet le transfert génétique d'une cellule à l'autre.
— La conjugaison correspond au transfert de matériel génétique entre cellules procaryotes par le biais de pili sexuels. Ce mode de transfert nécessite un contact entre cellules donneuses et réceptrices.

Le HGT implique essentiellement une grande variabilité de séquences autoréplicantes (éléments génétiques mobiles, *mobile genetic elements*, ci-après MGE), telles que les bactériophages, les plasmides, ou les éléments intégratifs et conjugatifs (ICE, *integrative and conjugative elements*).

Le HGT est un processus relativement commun au cours de l'évolution, et une fraction significative des gènes procaryotes y sont sujets (Ochman et al. 2000). Tantôt associé à la mutation ou à une forme de para-sexualité, le HGT peut même s'avérer être un puissant moteur de l'évolution des organismes, permettant autant l'acquisition de nouvelles protéines (Marijuán et al. 2010) que de grandes transitions écologiques (Schaack et al. 2010).

### 2.3.2 Succès du transfert horizontal et évolution post-transfert

On sait que le HGT est un mécanisme particulièrement important pour l'acquisition de résistance, il est particulièrement pervasif dans les communautés bactériennes et entre communautés (Koonin, Lopez-Garcia et al. 2016). Cependant, il est plus rarement discuté du succès de ces événements de transferts horizontaux et des mécanismes qui pourraient faciliter ou empêcher un gène transféré de s'implanter durablement chez le receveur.

#### a Facteurs impactant le succès du transfert horizontal

L'entrée de matériel génétique dans l'hôte et la compatibilité du nouveau gène avec l'hôte constitue une barrière très importante à l'occurrence du HGT. En effet, dans un premier temps, les vecteurs de transfert horizontal (plasmides, phages, etc.) ont généralement un spectre d'hôte restreint à quelques espèces tout au plus. Même lorsqu'un gène n'est pas associé à un vecteur, il existe une forte variabilité inter- et intra-spécifique de la compétence naturelle (capacité à intégrer de l'ADN libre) des bactéries (Johnsborg et al. 2007), ce qui induit une forte variabilité de la propension des génomes/espèces à recevoir du matériel génétique par transformation. De plus, la compétence naturelle nécessite une proximité écologique et physique (même si cette dernière peut varier selon le mécanisme). Dans un deuxième temps, une barrière à l'installation peut s'effectuer rapidement via des mécanismes de vigi-

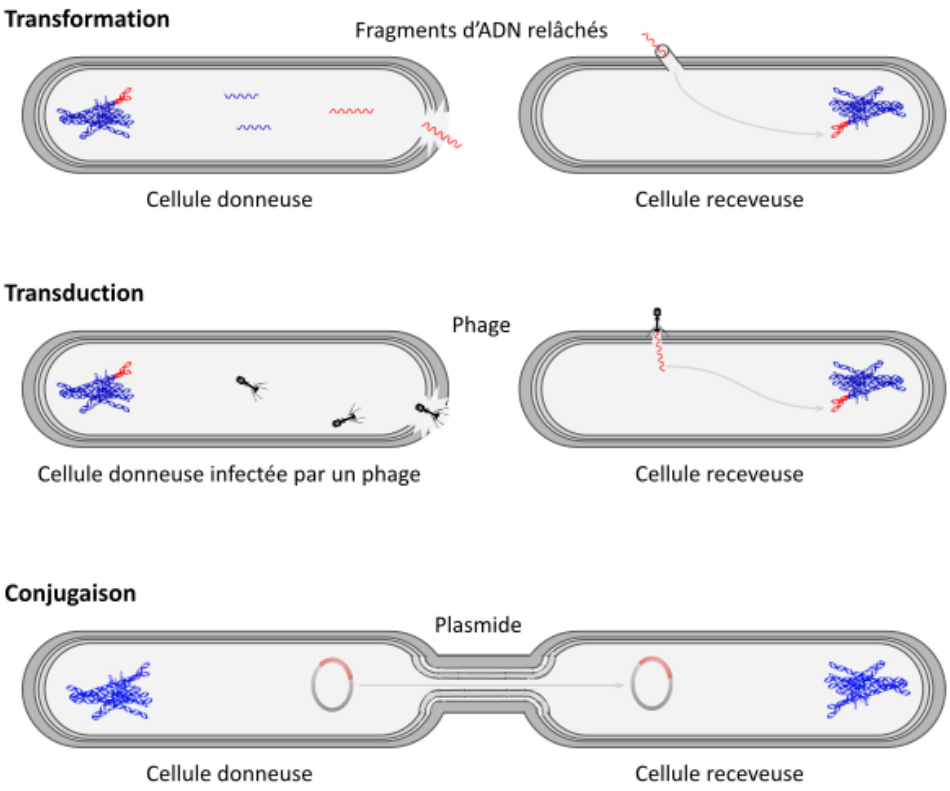


**FIGURE 3 –** Mécanismes de transfert horizontal de gènes.

lance immunitaire visant la destruction de l'ADN étranger (système *CRISPR-Cas*, enzymes de restriction, etc.).

Par ailleurs, même lorsque le matériel génétique exogène a été intégré, plusieurs coûts peuvent être associés au HGT. Avant tout, le HGT se produit fréquemment grâce à des vecteurs, et la réplication et l'expression de ces derniers sont coûteuses pour la cellule en termes de ressources énergétiques (ATP) et matérielles (ribosomes, bases azotées). Mais d'autres coûts peuvent s'ajouter : l'intégration chromosomique du MGE qui peut *(1)* interférer avec le réseau de régulation des gènes et avec les voies métabolique du receveur (un exemple par transduction, CHEN, GOLDING et al. 2005), *(2)* entraîner des effets cytotoxiques des protéines exprimées (SOREK et al. 2007), et *(3)* entrainer la séquestration des ressources métaboliques et des machines de transcription et de traduction (exemple chez *E. coli*, BONOMO et GILL 2005). De plus, le gène transféré n'a pas co-évolué avec les autres gènes du génome : il peut donc présenter un niveau d'expression trop faible/élevé (PARK et ZHANG 2012) ou encore constituer une rupture de coadaptation avec la machinerie de traduction (BALTRUS 2013, voir également partie 3.1.2). Pour se maintenir dans les populations, les gènes transférés doivent donc conférer au receveur un bénéfice suffisamment grand (ou suffisamment limiter leurs coûts) à leur hôte pour éviter d'être perdus par sélection purificatrice ou par dérive génétique (BALTRUS 2013). De nombreux gènes acquis par HGT comportent ainsi des fonctions accessoires car l'avantage que celles-ci confèrent peuvent compenser le

coût d'acquisition (Hyun et al. 2022). C'est notamment ce qui explique le succès des gènes de résistance aux antibiotiques, présents dans un très grand nombre de lignées bactériennes, par rapport à d'autres gènes conférant d'autres fonctions accessoires.

#### b Evolution post-transfert horizontal

Face aux différents coûts de transfert et de maintien, une réponse évolutive est observée de la part des structures transférées. La forme la plus décrite d'évolution post-HGT est le phénomène d'évolution compensatoire des MGE. En l'absence d'une pression de sélection sur les gènes accessoires portés par les MGE, il est possible que la lignée receveuse s'éteigne. Cependant, l'association entre un génome et un MGE peut être stabilisée par évolution compensatoire. La plupart des mutations répertoriées concernent le chromosome (Andersson et Hughes 2010), incluant par exemple le système d'expression (hélicases, ARN polymérase, Loftie-Eaton et al. 2017) ou le système de régulation de l'expression (Harrison et al. 2015). Les mutations évoluées après l'association à un MGE donné peuvent même augmenter la permissivité à de nouveaux MGE (Loftie-Eaton et al. 2017). Cependant, ces mutations compensatoires concernant souvent des gènes essentiels, elles sont coûteuses et leur maintien à long terme dépend lui-même de la présence de résistances portées par des MGE (MacLean et Vogwill 2015).

D'autres formes d'évolution peuvent compenser les coûts associés aux gènes transférés. Par exemple, l'inactivation du système de régulation de l'homéostasie protéique autorise l'expression de protéines d'origine exogène portant des fonctions avantageuses (Bershtein et al. 2015). De manière générale, de très larges modifications du protéome peuvent évoluer suite à un événement de HGT, et ce de manière dépendante du contexte environnemental (Bedhomme, Amorós-Moya et al. 2019).

# 3 Structuration, modulation, et dynamique de la résistance aux antibiotiques

L'évolution et la propagation de la résistance sont des phénomènes complexes, qui se jouent à plusieurs échelles. Ces derniers sont aussi bien régulés par des facteurs génétiques et génomiques, que par des facteurs environnementaux et sociaux. L'étude et la compréhension de la résistance doivent donc intégrer différents éléments pour essayer de retracer les trajectoires et de mettre en place des stratégies efficaces.

## 3.1 Régulation génétique et génomique du phénotype résistant

Le contexte génomique est l'une des premières choses prises en compte dans l'étude de la propagation des résistances aux antibiotiques. C'est même par l'intermédiaire des plasmides qu'on a découvert que la résistance pouvait se propager horizontalement. Cependant, d'autres types de MGE et surtout

d'autres facteurs génomiques, comme l'épistasie et le fond génétique, sont moins pris en compte mais tout aussi importants.

### 3.1.1 Les contextes génomiques de la résistance

#### a Les éléments génétiques mobiles

Les MGE ont été décrits pour la première fois en 1940 après des observations dans le génome du maïs (McClintock 1950, 1951). Ces derniers sont identifiés comme des fragments d'ADN codant une grande variété de facteurs de virulence et de résistance, ainsi qu'un ensemble d'enzymes permettant leurs transferts et intégration chez un nouvel hôte (Frost et al. 2005). En effet, la réplication des MGE est dépendante d'un hôte. Ils démontrent une mobilité intracellulaire et intercellulaire, et ceux d'une cellule particulière forment un mobilome (Siefert 2009).

Il existe une grande variété de MGE, certains ayant une capacité de mobilité intergénomique, et d'autres une capacité de mobilité intragénomique. Parmi les premiers, les bactériophages sont des virus bactériens susceptibles de capter des gènes de leur hôte au cours de leur réplication, tandis que les plasmides et les ICE peuvent être échangés entre bactéries par conjugaison. Parmi les derniers, les intégrons sont capables de capter de nombreux gènes au sein de cassettes grâce à leur intégrase (Mazel 2006), et les éléments transposables sont capables de se transposer et éventuellement de se dupliquer entre plusieurs locus.

Du fait de leur capacité de déplacement, de réplication, et des gènes qu'ils encodent, les MGE peuvent donc participer à une évolution rapide des bactéries. En effet, leurs rôles dans l'évolution des génomes chez les procaryotes, notamment sur les fonctions métaboliques est déjà bien établie (Frost et al. 2005). Ils peuvent notamment servir de vecteur de réarrangement génomique par recombinaison homologue et donc favoriser la néofonctionnalisation (Ding et al. 2009 ; Johnson et Nolan 2009).

Dans le cadre de l'étude de la résistance antibiotique, le rôle des MGE est souvent focalisé sur les plasmides. En effet, ils sont historiquement liés à la découverte des mécanismes de HGT (Watanabe 1963), et sont depuis très étudiés par le milieu hospitalier (León-Sampedro et al. 2021). Par exemple, la résistance à médiation plasmidique aux β-lactamines fournit une excellente illustration de la façon dont le HGT exacerbe les défis de la résistance antibiotique dans les hôpitaux (Paterson et Bonomo 2005). De plus, les gènes de résistance aux β-lactamines sont généralement localisés sur des plasmides et se diffusent donc par conjugaison inter- et intraspécifique chez les *Enterobacteriaceae*, *Pseudomonas* et *Acinetobacter* (Conlan et al. 2014 ; Gorrie et al. 2018 ; Mathers et al. 2011 ; Tofteland et al. 2013 ; Weingarten et al. 2018). À l'échelle mondiale, on peut même retracer la propagation du plasmide pCT porteur de la résistance *blaCTX-M-14* aux animaux et aux humains sur plusieurs continents (Cottell et al. 2011).

#### b Interactions entre chromosome et éléments génétiques mobiles

Même s'ils semblent jouer un rôle important dans le maintien et la propagation de la résistance aux antibiotiques, les plasmides sont souvent étudiés pour leurs caractéristiques intrinsèques (réplication, maintenance et mobilisation, Smillie et al. 2011 ; Solar et al. 1998). Avec une augmentation de la disponibilité des séquences génomiques plasmidiques, de plus en plus d'études ont cherché à étudier les relations entre plasmides et leurs contributions dans l'évolution bactérienne (Redondo-Salvo et al. 2020). Ces études ont également mis en lumière le fait que les plasmides ne sont pas isolés dans la cellule hôte et que de nombreux réarrangements avec les chromosomes sont possibles (Fondi et Fani 2010). Par exemple, une reconstruction phylogénétique des gènes partagés montrent que le plasmidome et le pangénome chromosomique de *Bacillus cereus* présentent des compositions géniques très similaires, ce qui témoigne d'échanges de gènes massifs entre les plasmides et les chromosomes (Zheng et al. 2015).

De plus, les plasmides peuvent également être présents en plusieurs copies (Rodríguez-Beltrán et al. 2020), et au sein d'une même cellule peuvent co-exister une diversité de plasmides (Bedhomme, Perez Pantoja et al. 2017). Du fait des échanges entre plasmides et chromosome, et du fait de la persistance conjointe de plusieurs plasmides, il peut apparaître un phénomène de redondance fonctionnelle, où plusieurs gènes identiques ou encodant la même fonction co-existent au sein d'un même génome. Cette redondance peut éventuellement être éliminée, lorsque les copies supplémentaires du gène en question ne confèrent pas d'avantage supplémentaire. Cependant, les gènes partagés entre plasmidome et pan-génome présentent généralement des séquences de promoteur et de terminaison différentes (Zheng et al. 2015). Cela suggère que l'expression d'un même gène est régulée par différents éléments au niveau transcriptionnel, en fonction de sa localisation dans le génome. Ainsi, conserver plusieurs copies d'un même gène pourrait être avantageux si cela permet à la cellule de contrôler plus finement les conditions d'expression de ce gène. De plus, ce mécanisme de conservation peut également permettre un élargissement du spectre de résistance passant par la diversification d'une des copies. La cellule peut ainsi maintenir deux allèles présentant des spectres différents (Rodriguez-Beltran et al. 2018).

### 3.1.2 L'encodage des gènes de résistance

Il existe une grande diversité de gènes conférant une résistance à des antibiotiques, par une toute aussi grande diversité de mécanismes (voir partie 1.1.2). Ces gènes ont tous en commun le fait d'être encodés par le même code génétique, et pour la plupart (à l'exception des ARN ribosomiques), de devoir être exprimés sous la forme de protéines. Toutefois, plusieurs mécanismes liés à l'encodage-même des gènes de résistance peuvent interférer avec leur expression. Cette section se concentre sur un des mécanismes les plus étudiés de modulation de l'expression par l'encodage : le biais d'usage de codons.

### a Origines du biais d'usage de codons

La redondance du code génétique autorise plusieurs codons à coder pour le même acide aminé. Les différents codons correspondant au même acide aminé sont définis comme des codons synonymes. Du fait de cette redondance, les substitutions synonymes dans les séquences codantes n'entraînent pas de changement de séquences d'acide aminés et ont été interprétées comme n'ayant aucune conséquence fonctionnelle, et donc considérées comme évolutivement neutres. Cependant, les codons synonymes sont rarement présents à des fréquences similaires dans les génomes. L'utilisation des codons synonymes ne résulte donc pas d'un processus aléatoire. Il existe une préférence pour certains triplets, appelée biais d'usage de codon (CUB, *codon usage bias*). Les CUB ont été observés dans une grande diversité d'espèces eucaryotes et procaryotes. Ils varient fortement, en préférence et/ou en intensité, d'une espèce à l'autre : ils peuvent même être extrêmement forts chez certaines espèces, qui évitent presque entièrement certains codons. Par exemple, *Pseudomonas aeruginosa* n'utilise que très peu le triplet TTA pour la leucine, alors qu'il s'agit du codon majoritaire chez *Bacillus cereus*. De même, en plus de variations interspécifiques, les CUB peuvent varier d'un gène à l'autre, au sein d'un même génome

L'origine évolutive de ce CUB est aujourd'hui bien identifiée. La préférence résulte en partie d'un biais de composition des séquences. En effet, chez les organismes eucaryotes et procaryotes, le CUB est en grande partie expliqué par le contenu nucléotidique GC, aussi bien aux niveaux des séquences codantes qu'au niveau des régions intergéniques. La variation de CUB, entre séquences et entre génomes, provient donc initialement de biais dans les mécanismes d'évolution moléculaire. D'une part, la mutation est universellement biaisée en faveur des nucléotides AT chez les Eubactéries (Hershberg et Petrov 2010). D'autre part, lors des évènements de recombinaison homologue, la conversion génique peut également introduire des biais en faveur de certains nucléotides. La conversion génique biaisée vers GC semble expliquer la majorité des biais de codon chez les Eucaryotes (Galtier et al. 2001), et pourrait également avoir un rôle chez les Eubactéries (Lassalle et al. 2015). Si ces deux mécanismes expliquent l'origine de la variation de CUB observée entre différentes espèces, ils n'expliquent pas à eux seuls l'intensité du CUB chez certaines espèces ou certains gènes. En outre, une co-adaptation entre le CUB et la machinerie de traduction de la cellule se met en place : la prévalence d'un codon dans un génome est positivement corrélée au nombre de copies pour l'ARNt qui présente l'anti-codon correspondant (Rocha 2004). De même, le CUB semble être plus élevé pour les gènes fréquemment exprimés (Stoletzki et Eyre-Walker 2007) et chez les espèces avec des taux de croissance maximaux élevés (Rocha 2004 ; Sharp et al. 2005). Chez les espèces à fort effectif efficace, comme c'est le cas pour de nombreuses bactéries, le CUB est donc façonné par la sélection traductionnelle qui optimise l'efficacité et la fidélité de la traduction. En effet, il a été montré que l'utilisation de codons rares conduisait à une traduction lente, donc à la production de protéine en petite quantité et de protéine erronées, tronquées et mal repliées. Plus récemment, il a été établi que la sélection non-traductionnelle participait également à façonner le CUB (Callens, Scornavacca et al. 2021).

### b Impacts sur la résistance aux antibiotiques

Le CUB pourrait avoir une influence sur le succès du HGT. En effet, le succès d'un gène acquis par HGT peut dépendre du contexte génomique, et plus précisément des différences entre le génome donneur et le génome receveur. En effet, dans un génome donné, un gène peut acquérir un CUB assez fort au cours de son histoire évolutive, notamment s'il code pour une protéine fonctionnellement avantageuse. Suite au transfert de ce gène dans un nouveau génome, il peut se produire une incompatibilité entre son CUB et celui du génome receveur (Amorós-Moya et al. 2010). Cette incompatibilité induit l'expression d'une faible quantité de protéines fonctionnelles, et l'accumulation de protéines non fonctionnelles qui doivent être éliminées. L'incompatibilité entre CUB peut donc conférer un désavantage sélectif à un gène, suite à un événement de HGT.

Il a été montré par des approches comparatives (Callens, Scornavacca et al. 2021) que les gènes tranférés dans un génome ont une probabilité plus faible de s'y maintenir au cours du temps, quand la différence de CUB entre les gènes et leur génome receveur augmente. Cependant, de nombreux transferts entre lignées phylogénétiquement éloignées ont été observés (Dunning Hotopp 2011), et il est donc possible qu'un gène au CUB maladapté soit maintenu dans le génome récepteur. Il existe donc des processus évolutifs capables de réduire le coût d'incompatibilité entre le CUB du transféron (matériel génétique transféré) et le CUB de l'espèce réceptrice. D'une part, on peut s'attendre à un processus d'amélioration des gènes transférés : les transférons peuvent accumuler des mutations synonymes qui rapprochent le CUB du gène à celui de son génome récepteur. Cependant, ce phénomène ne s'observe que sur des temps phylogénétiquement longs (Lawrence et Ochman 1997), et pourrait ne participer qu'au maintien d'une faible proportion des transférons lorsque le coût d'incompatibilité est fort. D'autre part, des mutations compensatoires peuvent émerger au sein du chromosome procaryote. Ces mutations peuvent augmenter la valeur sélective de l'individu, et compenser le coût de l'incompatibilité. Cela peut passer par une production accrue de protéines fonctionnelles : soit en exprimant davantage le transféron, soit en modifiant la machinerie de traduction. Cette dernière issue peut avoir lieu beaucoup plus rapidement et a été observée en moins de 350 générations chez *Escherichia coli* (Bedhomme, Amorós-Moya et al. 2019).

## 3.2 Problématiques écologiques de la résistance aux antibiotiques

L'étude de la résistance antibiotique s'est le plus souvent faite dans des environnements cliniques, en particulier sous l'angle des infections à bactéries résistantes ou sous l'angle de la sélection par exposition aux antibiotiques. Cependant, on sait désormais que les résistances antibiotiques existent aussi dans les environnements naturels, qu'elles y ont même émergé avant l'utilisation anthropiques des antibiotiques, et qu'il existe des échanges entre environnements naturels et environnements anthropisés. Un corpus grandissant d'études pointe le rôle de facteurs environnementaux dans la dynamique d'apparition, de maintien et de propagation de la résistance, mais aussi les conséquences potentielles ou

observées de la résistance dans les milieux naturels.

### 3.2.1 Rôle et impacts de la résistance dans les communautés bactériennes

#### a Propagation des résistances dans les communautés

Si on sait que les résistances antibiotiques ont une origine naturelle, leur impact dans les communautés bactériennes naturelles est encore peu connu. En effet, beaucoup d'études se focalisent sur un petit nombre d'espèces pathogènes, souvent en environnement clinique. Pourtant il semble que l'acquisition de résistances antibiotiques soit plus difficile dans les communautés naturelles (Klümper et al. 2019). En effet, des communautés variées pourraient impliquer une plus grande compétition qui tendrait à rendre le coût de la résistance difficilement supportable. Les interactions au sein des communautés jouent donc un rôle-clé dans l'acquisition et le maintien des gènes de résistance. La coopération avec la mise en place de biofilms et la compétition pour les ressources semblent particulièrement importantes à prendre en compte pour une compréhension globale du système de la résistance. En effet, ces dernières peuvent façonner le cycle de la résistance en entraînant par exemple des trade-offs évolutifs. Un bon exemple est le compromis entre résistance antibiotique et résistance aux phages (Gurney et al. 2020).

#### b Les conséquences écologiques de la résistance

La résistance antibiotique n'est pas seulement un problème pour le traitement des infections humaines, animales, et végétales. En effet, des antibiotiques, des biocides, des métaux lourds et d'autres substances d'origine anthropique sont relarguées dans les environnements naturels, que ce soit en grandes quantités ou en traces. Lorsqu'elles sont présentes dans l'environnement, ces substances modifient le paysage adaptatif et favorisent l'émergence ou l'augmentation en fréquence de populations résistantes dans les communautés microbiennes. Ces bactéries résistantes pourraient remplacer des populations sensibles : par exemple, les cyanobactéries, principaux producteurs primaires des écosystèmes aquatiques, sont sensibles à la plupart des antibiotiques les plus largement utilisés (Grinten et al. 2010 ; Guo, Boxall et al. 2015). De plus, les changements dans les communautés pourraient également perturber les chaînes trophiques en affectant les préférences alimentaires des décomposeurs (Bundschuh et al. 2009). Ainsi, les déséquilibres produits par les antibiotiques parmi les producteurs primaires et les décomposeurs pourraient perturber les écosystèmes naturels, avec de graves conséquences pour l'environnement dans son ensemble (Martínez 2008). Outre les potentielles conséquences pour les services écosystémiques, l'homéostasie des communautés bactériennes est également à préserver dans le cadre de la lutte contre l'antibiorésistance, puisqu'elle favorise la biodégradation des antibiotiques (Yu et al. 2018) et limite la propagation des gènes de résistance (Chen, An et al. 2019).

### 3.2.2 Échanges entre écosystèmes naturels et écosystèmes anthropisés

### a Antibiotiques et pollutions

Les antibiotiques sont utilisés par l'homme dans les hôpitaux, dans les soins communautaires, mais également dans les élevages et en aquaculture. Ces différentes sources de consommations entraînent une production de polluants organiques persistants (POP) dans différents environnements difficiles à éliminer (Lei et al. 2019). En effet, les antibiotiques consommés dans ces différents contextes ne sont pas absorbés ou dégradés dans leur intégralité (jusqu'à 80 % sont relargués dans les urines, Jjemba 2006), et les résidus d'antibiotiques sont ensuite évacués avec les déchets, par les eaux usées notamment (Gothwal et Shashidhar 2015). De même, le relargage dans l'environnement peut être volontaire, les eaux usées étant utilisées comme engrais (Bondarczuk et al. 2016). Par ailleurs, les déchets issus de la fabrication de médicaments contiennent régulièrement de très fortes concentrations de résidus d'antibiotiques (Kotwani et al. 2021), qui se retrouvent très régulièrement dans les eaux effluentes et avoisinantes des usines concernées (par exemple Tong et al. 2020). Des résidus d'antibiotiques tendent donc à s'accumuler dans différents environnements, et en premier lieu dans les usines de traitement des eaux (Kümmerer 2009). Même lorsqu'ils sont acheminés jusqu'aux usines de traitement des eaux, les résidus d'antibiotiques peuvent ensuite se retrouver dans le circuit d'eau ou dans l'environnement, soit parce qu'ils ne sont pas toujours correctement dégradés, soit à cause de relargages accidentels (Park, Kim et al. 2007). De ce fait, on peut retrouver des résidus d'antibiotiques dans tout le cycle de l'eau : de l'eau courante et des plans d'eau naturels (Valcárcel et al. 2011) jusqu'aux aquifères (López-Serna et al. 2013) et aux océans (Na et al. 2013). On retrouve également une forte pollution antibiotique des sols et sédiments, du fait de l'épandage d'eaux usées, mais également du fumier (Chen, An et al. 2019 ; Yang et al. 2020 ; Zhang, Zheng et al. 2021a). Enfin, les antibiotiques issus de la pollution peuvent également être bio-accumulés, notamment dans les tissus des plantes et des animaux aquatiques (Eggen et al. 2011 ; Na et al. 2013).

Cette pollution antibiotique est susceptible de contribuer à la sélection locale de gènes de résistance (Martinez 2009 ; Tello et al. 2012), et donc indirectement à la création de réservoirs de résistance, mais également de bouleverser les communautés microbiennes de l'environnement (Cycoń et al. 2019).

### b Des bactéries résistantes jusqu'aux réservoirs de résistance

Les bactéries peuvent se propager entre écosystèmes, et il y a donc toujours un flux continu d'informations génétiques entre les différents compartiments écologiques. Une fois que les gènes potentiels de résistance aux antibiotiques pénètrent, même en petit nombre ou à faible fréquence, dans le microbiote commensal/pathogène humain/animal, la sélection d'antibiotiques conduit immédiatement à l'amplification et à la dissémination de ces gènes. En effet, il existe des preuves sur l'origine environnementale de certains gènes de résistance cliniquement pertinents (Wright 2010), et on sait que certains gènes de résistance cliniquement pertinents se retrouvent à l'identique dans des bactéries du sol (Forsberg et al. 2012). De même, des études récentes suggèrent que la source la plus probable d'émergence des résistances est le résistome environnemental aux antibiotiques (Martínez 2008 ; Wright 2010) avec un rôle

fonctionnel de signalisation/régulation au sein des populations (Aminov 2009 ; Davies 2006 ; Fajardo et Martínez 2008). De même, on a pu observer des résistances dans les bactéries des microbiotes de nombreux animaux sauvages, tels que des gibiers sauvages (Guerrero-Ramos et al. 2016 ; Mateus-Vargas et al. 2017).

Même au sein des environnements anthropisés, de nombreuses activités peuvent conduire au transfert de résistances. En particulier, l'agriculture et l'élevage constituent un carrefour pour la transmission de résistances. Le fumier et les eaux usées des exploitations agricoles, parfois épandus sur les cultures en tant qu'engrais, peuvent contenir des résidus d'antibiotiques, des gènes de résistance aux antibiotiques, et des bactéries résistantes aux antibiotiques (Chen, An et al. 2019 ; Yang et al. 2020 ; Zhang, Liu et al. 2021)). Ainsi, les sols amendés au fumier contiennent plus de gènes de résistance que les sols non amendés ou les sols fertilisés à partir d'engrais chimiques commerciaux (He, Yuan et al. 2020). Cela pourrait être dû à la fois à l'introduction de bactéries résistantes aux antibiotiques, et à la pression de sélection exercée par les résidus d'antibiotiques. De plus, cet effet est durable : les sols amendés contiennent encore davantage de gènes de résistance que les sols non amendés 5 ans après l'épandage (Larouche et al. 2020). Cette contamination des sols peut ensuite se propager. D'une part, les légumes cultivés sur des sols amendés au fumier sont régulièrement contaminés par des gènes de résistance (Gao et al. 2020). D'autre part, il existe une corrélation entre la densité de bétail et la présence de bactéries résistantes dans les rivières attenantes (Teng et al. 2021). Du reste, le stockage externe du fumier produit des bioaérosols qui facilitent le transfert des résistances dans l'environnement par voie aérienne (Zhang, Zheng et al. 2021b). In fine, l'isolement de micro-organismes résistants aux antibiotiques à partir de divers types de viande a été abondamment rapporté dans la littérature (voir par exemple Abass et al. 2020 ; Geser et al. 2012 ; Lee, Cho et al. 2018 ; Mashak et al. 2020 ; Puangseree et al. 2021). De fait, lors de l'abattage, la surface de la carcasse et les découpes de détail peuvent être contaminées par des organismes provenant de la peau et du contenu intestinal de l'animal, mais aussi des mains des travailleurs et de l'environnement de l'abattage (Driessen et al. 2020). Il n'est donc pas surprenant que si les animaux élevés pour la consommation de viande abritent des micro-organismes résistants aux antibiotiques, ceux-ci se retrouvent dans la viande à consommer (Massot et al. 2021).

## 3.3 Les sociétés humaines face aux résistances antibiotiques

### 3.3.1 Historique de la réponse humaine aux résistances antibiotiques

La réponse apportée face au problème de la résistance antibiotique n'a pas toujours été la même et notre situation actuelle est en partie le produit des mesures prises ou non prises. Il est donc crucial de regarder comment notre réponse actuelle s'inscrit dans l'histoire des politiques de réponse précédentes. Au cours des 80 dernières années, la préoccupation face à la résistance aux antibiotiques a dépendu de plusieurs facteurs tels que : la distribution des bactéries résistantes, la capacité et les efforts de détection, les pratiques de l'agro-industrie, la capacité de l'industrie pharmaceutique a conserver

une longueur d'avance sur la résistance, et la coordination des efforts financiers et de développement des infrastructures pour attirer l'attention sur la résistance microbienne et y faire face. Une observation rétrospective englobant l'Europe et les Etats-Unis a permis de mettre en évidence 5 périodes mondiales occidentales et sont synthétisées dans la figure 4.

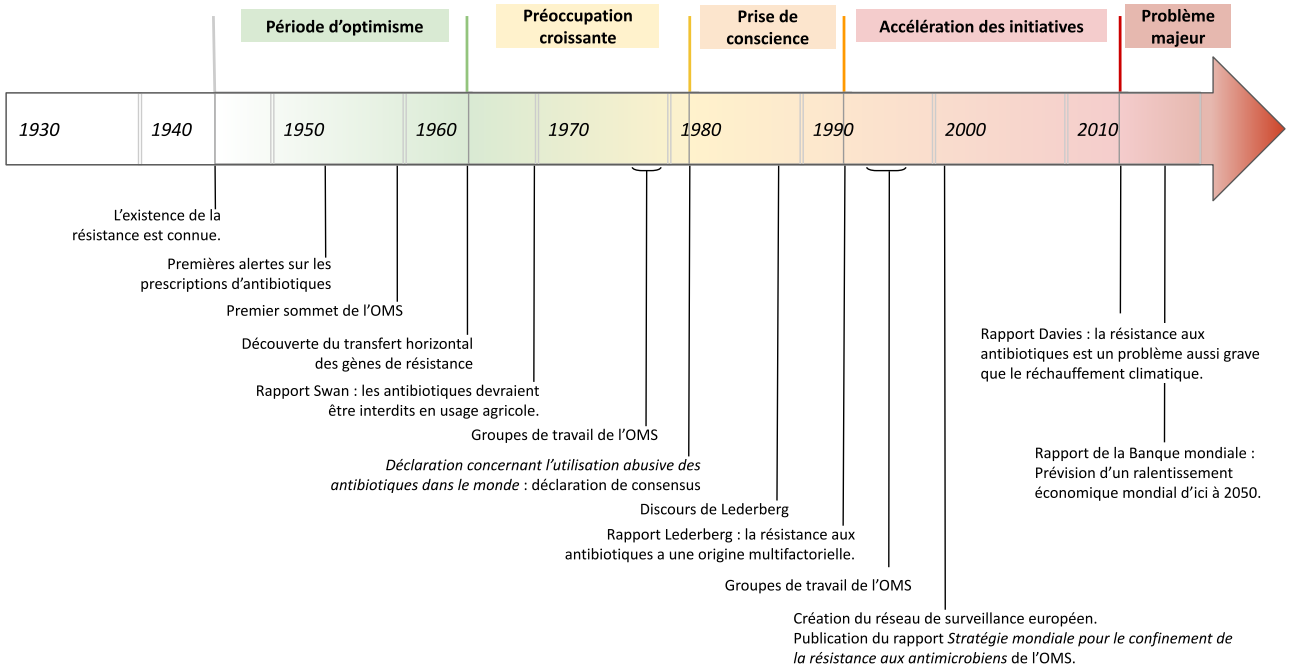


**FIGURE 4 –** Historique de la réponse humaine aux résistances antibiotiques.

### a De 1945 à 1963 : de l'optimisme face à un problème mineur

Le potentiel de résistance aux antibiotiques cliniquement pertinents était déjà reconnu dans les années 1940 (FLEMING 1945), et l'existence de résistance aux sulfamidés pendant la Seconde Guerre mondiale avait placé les cliniciens en alerte (LESCH 2007). Dès les années 1950, certains spécialistes des maladies infectieuses et médecins généralistes prédisaient que l'utilisation des antibiotiques deviendrait de plus en plus difficile (BATTEN 1955), et la recherche pointait déjà le rôle de la prescription des antibiotiques à l'aveugle (HUSSAR et HOLLEY 1954). Cependant, à cette époque, le problème reste ignoré en dehors du monde académique et peu de moyens sont investis pour coordonner la résistance au-delà de l'échelle locale. Notamment, un sommet d'experts a été réuni par l'OMS en 1959, mais aucune coordination internationale en matière de surveillance ou d'utilisation n'est ressortie (PODOLSKY 2018). De plus, le développement de nouveaux antibiotiques à large spectre (tétracycline, chloramphénicol, etc.) dans les années 1950 entretenait un climat optimiste quant à la capacité de l'industrie pharmaceutique à gagner la course aux armements engagée avec les microbes via le développement de nouvelles molécules (PODOLSKY 2015).

### b De 1963 à 1981 : une préoccupation croissante

L'optimisme commence pourtant à décliner dans les années 1960, avec la découverte du transfert horizontal intra- et inter-espèce de la résistance par le biais des MGE (Watanabe 1963). En 1969, le rapport Swan préconise l'interdiction de l'utilisation des antibiotiques thérapeutiquement pertinents comme facteurs de croissance dans le milieu agricole (Bud 2007). Cependant, l'impact de ce rapport a été limité (Thoms 2012) : dans les pays qui ont adopté le texte, des stratégies pour compenser les pertes économiques n'ont pas été mises en place et ont conduit à des comportements abusifs comme la surprescription thérapeutique vétérinaire. De plus, le manque de données sur l'impact des antibiotiques en milieu agricole a permis aux groupes d'influence de l'agro-industrie de faire valoir l'impact économique du retrait des antibiotiques pour empêcher les interdictions jusque dans les années 1970 (McKenna 2007). Au contraire, dans les années 1970, un plus grand effort est portée à restreindre la prescription d'antibiotiques en milieu médical, qualifiées "d'irrationnelle" (Gibbs et al. 1973 ; Roberts et Visconti 1972 ; Simmons et Stolley 1974).

### c De 1981 à 1992 : une prise de conscience mondiale

La préoccupation n'ayant pas percolé au-delà des pays industrialisés, de nouvelles infections résistantes à médiation plasmidique ont continué à être décrites au cours des années 1970. L'OMS a commencé à s'emparer du problème en organisant des groupes de travail en 1976 puis en 1978. Cependant, ce n'est qu'en 1981 que la Déclaration concernant l'utilisation abusive d'antibiotiques dans le monde, impliquant des chercheurs de 27 pays, a établi un consensus sur le « problème de santé publique mondial » posé par la résistance aux antibiotiques. Cette déclaration rassemblait un ensemble de préoccupations et de recommandations exigeant une coordination et une diversification de la part de la communauté mondiale. Cette déclaration a servi de point d'ancrage à la constitution de l'Alliance pour l'utilisation prudente des antibiotiques (APUA), groupe international dédié à l'éducation et la réglementation de la vente et de l'utilisation des antibiotiques.

### d De 1992 à 2013 : une accélération des initiatives

Un prix Nobel va aider à la diffusion publique du problème de la résistance (Podolsky 2015) : Joshua Lederberg a reçu en 1958 la distinction pour ses études sur les échanges génétiques bactériens. S'exprimant en 1988, Lederberg a évoqué la nécessité pragmatique et morale de faire face aux maladies infectieuses comme une préoccupation mondiale partagée comparable aux maladies émergentes (Lederberg 1988). Un rapport dont il sera le commendataire quelques années plus tard posera l'origine multifactorielle de la résistance aux antibiotiques (Lederberg et Oaks 1992). Ce rapport a déclenché une série de rapports politiques gouvernementaux et d'experts, mais aussi une présence durable dans des revues à fort impact. Des pathogènes résistants deviennent emblématiques comme les infections à staphylocoques résistants à la méthicilline (SARM), ou la pneumonie à pneumocoques, ainsi que cer-

tains ignorées par les hôpitaux et cliniques occidentales comme le paludisme résistant et la tuberculose (Garrett 1994). L'expertise en matière de lutte contre la résistance aux antibiotiques s'est progressivement déplacé vers l'Europe où, grâce à l'expérience acquise durant l'épidémie d'encéphalite spongiforme bovine, un comité d'experts a conduit à la création d'un système européen de surveillance des résistances (EARSS, puis EARS-Net) en 2001 (Bud 2007). Notamment, les pays nordiques sont devenus des modèles mondiaux en ce qui concerne la gestion des antibiotiques : le consortium de la Finlande, la Suède et du Danemark ont organisé des campagnes nationales d'éducation au milieu des années 1990 pour préserver l'utilité clinique de ces médicaments (Sørensen et Monnet 2000). Enfin, l'OMS s'est pleinement engagée sur la question des résistances à partir des années 1990 avec des groupes de travail et des réunions tout au long de la décennie (World Health Organization 1994, 1997, 1998). En 2001, il a publié sa Stratégie mondiale pour le confinement de la résistance aux antimicrobiens en commençant à attirer l'attention sur les implications économiques et nationales plus larges d'une résistance généralisée (World Health Organization 2001). La nécessité d'une approche mondiale partagée de la résistance aux antibiotiques est ainsi devenue l'un des principes centraux des déclarations de l'OMS. La résistance aux antibiotiques est maintenant présentée publiquement comme un problème mondial qui nécessite des interventions multisectorielles et collaboratives.

### e Depuis 2013 : un problème aussi important que le changement climatique

En 2013, deux rapports ont pointé l'urgence de la situation. D'une part, le rapport Davies met sur un pied d'égalité le changement climatique et une résistance antimicrobienne généralisée (Davies 2013), et prédit, si la résistance n'est pas contrôlée, des taux de mortalité liés aux infections comparable à ceux de l'ère victorienne (Shallcross et al. 2015). D'autre part, le rapport des CDC américains attirent l'attention sur des agents pathogènes nouvellement prioritaires, tels que les entérobactéries résistantes aux carbapénèmes ou multirésistants (Centers for Disease Control and Prevention 2013). L'attention est particulièrement portée sur l'augmentation mondiale de la consommation d'antibiotiques (35% entre 2000 et 2010), dont 76% est imputable au Brésil, la Russie, l'Inde, la Chine, et l'Afrique du Sud, même si on conçoit la nécessité de préserver ou d'améliorer l'accès aux antibiotiques dans les pays en développement (Årdal et al. 2016). Dans ce contexte, l'OMS a publié son Rapport mondial sur la surveillance de la résistance aux antimicrobiens et son Plan d'action mondial sur la résistance aux antimicrobiens (WHO 2015 ; World Health Organization 2014) et, en 2017, un rapport de la Banque Mondiale affirme que, sans contrôle, la résistance pourrait entraîner des ralentissements massifs de l'économie mondiale (World Bank 2017). Finalement, l'ONU a créé un groupe de coordination inter-institutions sur la résistance aux antimicrobiens (Årdal et al. 2016), pour plaider en faveur d'une action contre la résistance antibiotique au plus haut niveau politique et pour attirer l'attention et le financement sur ceux dont la mission déclarée est de prévenir une ère post-antibiotique. Parmi les différentes actions recommandées : augmenter la surveillance et le dénombrement réel de l'utilisation et de la résistance aux antimicrobiens, réduire l'incidence de la résistance aux antimicrobiens, et augmenter l'offre de nou-

veaux antimicrobiens. Cependant, à l'heure actuelle, la surveillance dépend majoritairement de groupes philanthropiques, et moins d'un tiers des états membres de l'OMS participent au système mondial de surveillance des antimicrobiens.

### 3.3.2 De la santé individuelle aux approches One health et Global health

A travers une rétrospective historique, on peut donc constater que les sociétés ont tenté de contourner le problème de la résistance aux antibiotiques par plusieurs angles d'attaques, tous ayant en commun une approche de santé individuelle. Selon cette approche, il faut chercher à caractériser les résistances, et à les traiter/contourner les unes après les autres, ainsi qu'à éviter la propagation notamment au sein des hôpitaux. Cependant, comme expliqué dans les sections précédentes, de nombreuses résistances émergent et/ou se propagent hors des hôpitaux et des élevages (c'est-à-dire des lieux où les antibiotiques sont majoritairement consommés), notamment dans les écosystèmes naturels.

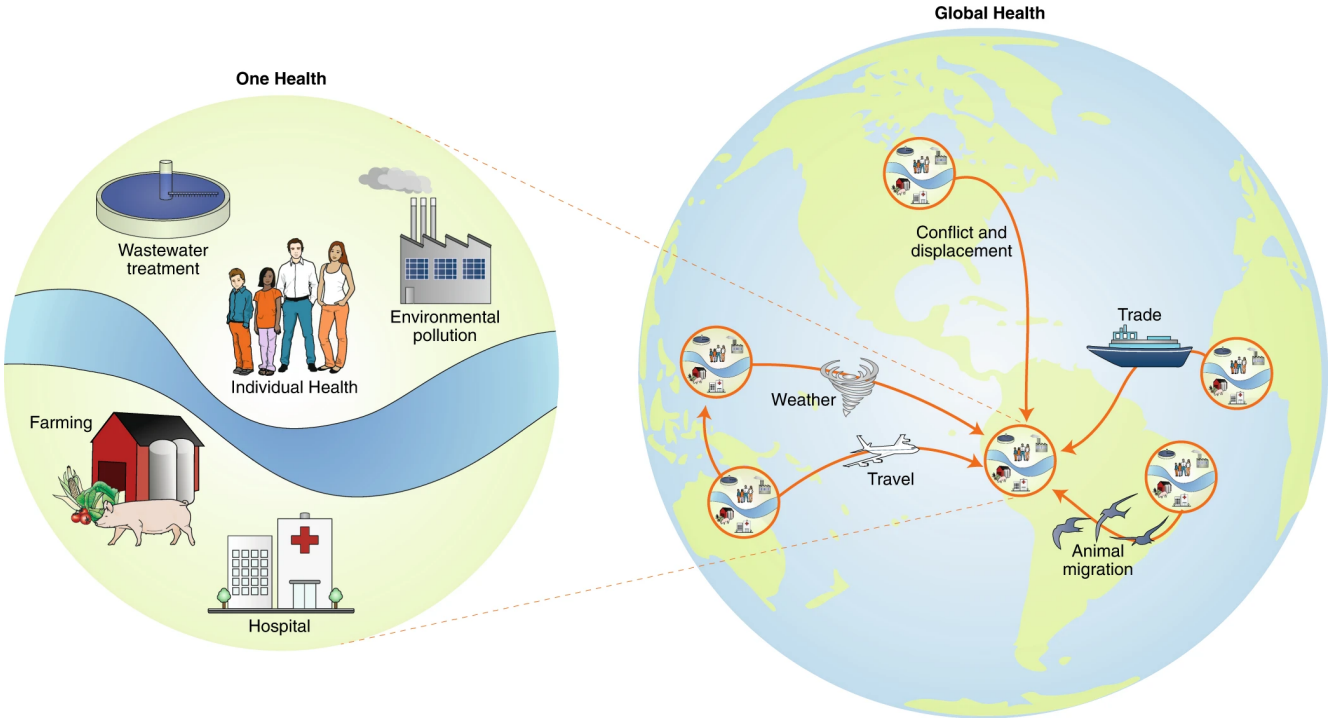


**FIGURE 5 –** Approches *One Health* et *Global Health* de la propagation des résistances aux antibiotiques. D'après HERNANDO-AMADO et al. 2019.

L'étude du phénomène doit donc intégrer l'étude des écosystèmes naturels et les processus socioculturels qui favorisent l'émergence et la propagation des résistances. Depuis les deux dernières décennies, ce raisonnement a amené à l'émergence de l'approche *One Health* (littéralement une seule santé) des problématiques de santé publique : les problèmes de santé publique (infections émergentes, résistance aux antibiotiques) sont intrinsèquement liés à la santé des écosystèmes anthropisés et naturels. Selon cette approche, la gestion d'un problème de santé publique nécessite de comprendre les différentes interactions entre les environnements anthropisés et naturels, et donc de potentiellement ne

plus seulement agir à l'échelle des patients, mais également à l'échelle d'écosystèmes entiers et d'agir sur les interactions entre écosystèmes.

En outre, du fait de la mondialisation des échanges, il devient clair qu'un problème de santé au niveau local, tel que la présence de bactéries résistantes dans une région, se propage rapidement et peut ainsi devenir mondial. En particulier, il a été démontré que les bactéries résistantes aux antibiotiques peuvent se propager d'un continent à un autre par le biais de transports de marchandises (Le Hello et al. 2011) ou par le biais de courants océaniques (Yu et al. 2018). A l'approche *One Health*, il est donc crucial d'associer une approche de santé globale (*Global Health*), pour développer des stratégies à long terme et à grande échelle de gestion de la résistance (Hernando-Amado et al. 2019, figure 5).

# Section II : Problématique générale de la thèse

La résistance aux antibiotiques est un phénomène global, qui ne se restreint pas à un unique pays ou un unique biome. On retrouve des preuves de transmission à grande échelle : écologique, géographique, et phylogénétique. Résoudre la crise actuelle nécessite donc d'étudier le problème à l'échelle mondiale, et de comprendre les phénomènes qui participent à cette transmission.

Cette thèse vise à mieux comprendre les forces et mécanismes qui définissent la direction et le mouvement des gènes de résistance et à les replacer dans une perspective éco-évolutionniste, dans le but de prédire et de ralentir leur propagation. En particulier, les objectifs de cette thèse sont :

— de fournir une étude de la propagation aux antibiotiques aux échelles les plus larges possibles ;
— de décrire les patrons de propagation de la résistance à différents niveaux ;
— de mesurer la part relative de certains facteurs clés qui pourraient y être associés.

Pour ce faire, les travaux présentés dans cette thèse se focalisent sur la résistance aux aminoglycosides. Ces antibiotiques sont parmi les premiers à avoir été découverts, la streptomycine ayant été isolée de *Streptomyces griseus* en 1943. A cette époque, ils sont les premiers à avoir montré une efficacité dans la lutte contre les infections à bactéries Gram-négatives (en particulier contre la tuberculose). Depuis, leur utilisation s'est étendue à une grande variété d'infections : des plus bénignes comme les otites externes à tympan fermé (néomycine), jusqu'aux plus graves comme les endocardites (streptomycine, gentamicine) ou les septicémies (différents types d'aminoglycosides). Cependant, ces antibiotiques présentent une néphrotoxicité réversible, mais surtout une ototoxicité irréversible qui peut mener jusqu'à la surdité. De plus, au contraire d'autres familles, les aminoglycosides ne peuvent avoir une action systémique que lorsqu'administrés par voie intraveineuse. Ces antibiotiques ont donc été peu à peu écartés des traitements de première intention en médecine humaine à partir des années 1970. Toutefois, leur usage reste encore relativement répandu en médecine vétérinaire et en agriculture. A titre d'example, en 2015, les aminoglycosides représentaient toujours 3.5% des ventes d'antibiotiques à destination des animaux de ferme en Europe, et peuvent avoir une usage systématique dans le cadre de certaines pathologies en élevage, telle que la diarrhée post-sevrage des porcelets (Duijkeren et al. 2019). Ainsi, la résistance aux aminoglycosides reste particulièrement prévalente parmi les bactéries échantillonnées chez des animaux domestiques : au Danemark, 85% des échantillons de *Salmonella spp. serovar Kentucky* isolés chez des dindes ne sont pas susceptibles aux aminoglycosides, et en Italie, jusqu'à 67% des échantillons d'*Escherichia coli* sont résistantes à au moins une aminoglycoside (Duijkeren et al. 2019). Par conséquent, bien que les aminoglycosides sont moins utilisées en médecine humaine, la résistance à cette famille reste un problème majeur, au moins pour la production alimentaire. Par ailleurs, les aminoglycosides sont naturellement produites par certains taxons bactériens (*Streptomyces* et *Micromonospora*, depuis au moins 30000 ans, D'Costa et al. 2011) vivant dans le sol et la matière végétale en décomposition. Par conséquent, de même que pour d'autres familles de résistance beaucoup plus étudiées (telles que les $\beta$-lactamases), la résistance aux aminoglycosides évolue dans de nombreux

contextes écologiques différents, depuis lesquels elle est susceptible de se propager et de générer des risques sanitaires.

Dans le cadre de ma thèse, mes travaux se penchent sur un mécanisme de résistance aux aminoglycosides : les enzymes altérant les aminoglycosides (AME, *aminoglycoside-modifying enzymes*). Ces enzymes empêchent la fixation des aminoglycosides au ribosome bactérien, en leur ajoutant un groupement acétyl, phosphoryl, ou nucléotidyl. Cet ajout de groupement peut se faire à différentes positions de la chaîne carbonée, ce qui implique l'existence d'une très grande diversité d'AME conférant une grande diversité de spectre de résistance. Ainsi, on sait que les AME sont apparues à plusieurs reprises au cours de l'évolution des Eubactéries (Salipante et Hall 2003). Contrairement à d'autres types de résistances (par exemple les pompes d'efflux), les AME sont de simples enzymes cytoplasmiques et ne confèrent de résistance qu'aux aminoglycosides. Elles peuvent donc potentiellement être exprimées et conférer une résistance chez toutes les bactéries, et les aminoglycosides sont les seuls antibiotiques susceptibles de les sélectionner. D'autres mécanismes de résistances aux aminoglycosides sont décrits (modification de la cible ribosomique par exemple, voir partie 1.1.2). Cependant, les AME restent actuellement le principal mécanisme impliqué (Ramirez et Tolmasky 2010).

Le premier chapitre de ma thèse (section III) est consacré à la mise en place d'un outil de détection de plasmides dans les jeux de données génomiques. Cet outil facilite l'étude de l'association entre gènes de résistance et MGE dans les très larges jeux de données.

Le second chapitre de ma thèse (section IV) porte sur la distribution mondiale des AME : au niveau phylogénétique chez les Eubactéries, au niveau écologique dans les écosystèmes continentaux, et au niveau géographique. Les travaux développés dans ce chapitre décrivent le patron de propagation spatiale et temporelle des gènes encodant les AME tant par le biais de la structuration naturelle (continentale et écologique), que par le biais des activités humaines (consommation d'antibiotiques, commerce de biens, et migrations humaines).

Enfin, dans un dernier chapitre (section V), j'aborde la propagation phylogénétique des gènes encodant les AME par le biais du HGT. D'une part, ce chapitre développe une méthode de reconstruction des réseaux de propagation de ces gènes entre différents taxons. D'autre part, l'analyse des réseaux de propagation permet l'identification des partenaires privilégiés du transfert horizontal des AME, ainsi que les facteurs modulant la probabilité d'échange.

# Section III : Caractérisation des éléments génétiques mobiles dans grands jeux de données génomiques

Les plasmides sont des fragments d'ADN extra-chromosomiques qui se répliquent de manière autonome dans les cellules bactériennes. Ils portent souvent des gènes qui peuvent apporter un bénéfice dans des conditions environnementales spécifiques. A ce titre, de nombreux plasmides portent des gènes de résistance aux antibiotiques. Depuis l'article fondateur de Watanabe 1963, les plasmides ont été particulièrement étudiés pour leur rôle dans la propagation des résistances antibiotiques, et notamment par transfert horizontal. A l'heure actuelle, de nombreuses résistances sont essentiellement connues comme médiées par des plasmides.

Dans le cadre de cette thèse, il est donc essentiel de pouvoir prendre en compte l'association entre gènes de résistance et plasmides. L'objectif est de pouvoir le faire sur le jeu de données génomiques le plus large possible, afin d'établir un regard le plus complet possible sur la propagation des résistances antibiotiques. Toutefois, à l'heure où de très nombreux génomes bactériens ont été séquencés, seuls 10% de ceux publiés sur la base de données *NCBI Assembly* ont été complètement assemblés, et leurs éventuels plasmides annotés. Ce jeu de données restreint est déjà une source d'informations précieuses, mais elle est largement biaisée en faveur des quelques taxons les plus séquencés. Son extension à tous les génomes bactériens partiellement assemblés permettrait de prendre en compte une plus large diversité bactérienne, potentiellement échantillonnée dans une plus large diversité de milieux.

Pour ce faire, il est essentiel de pouvoir identifier les contigs (fragments d'ADN partiellement assemblés à partir de *reads*) d'origine plasmidique au sein des génomes partiellement assemblés (à l'état de *drafts* ou de *scaffolds*). Cette thèse n'est pas la première à avoir identifié cette problématique et, depuis plusieurs années, de nombreux outils bioinformatiques ont été développés afin de pallier ce manque. Initialement, le premier outil d'identification de plasmides est *PlasmidFinder* qui comporte une base de données de réplicons provenant de bactéries Gram-positives et d'Entérobactériales, et à laquelle on peut comparer des séquences inconnues grâce à l'outil de recherche de similarité de séquences *BLASTn*. Cet outil, de même que les outils basés sur la recherche d'homologie, a l'avantage d'identifier avec une très grande précision les séquences qui ressemblent à ses réplicons. En revanche, toute la force de cet outil repose dans sa base de données : toute séquence qui s'en éloigne un tant soit peu ne pourra pas être identifiée, et cet outil a donc une sensibilité très limitée.

Face à ce manque, amplifié par l'arrivée d'espèces inconnues dans les bases de données du fait de l'échantillonnage de métagénomes, de nouvelles approches ont été développées. Ces nouvelles approches se basent, pour la plupart, sur des méthodes sans alignement (par exemple, les fréquences de k-mers) de sorte à ne plus dépendre de bases de données, et sur l'apprentissage supervisé, qui augmente le pouvoir de résolution de ces méthodes. Ces nouvelles méthodes élargissent considérablement la gamme de taxons bactériens où des plasmides peuvent être détectés. En revanche, cette sensibilité accrue se fait au prix d'une précision réduite, et d'une fiabilité quasi-nulle sur les contigs les plus courts (en dessous de 1000 paires de bases). De plus, comme on le verra dans ce chapitre, ces méthodes donnent parfois des résultats très peu concordants les uns avec les autres.

Dans ce chapitre, je développe une nouvelle méthode bioinformatique de détection de plasmides, spécifiquement orientée vers les jeux de données génomiques. Cette méthode, nommée *PlasForest*, repose sur une base de données de séquences plasmidiques connues et un classifier basé sur une forêt d'arbres de décision aléatoire. Dans un premier temps, chaque séquence qu'on souhaite classer est comparée à la base de données grâce à *BLASTn*. Dans un second temps, plusieurs caractéristiques de cette séquence sont extraites : d'une part, des statistiques liées à la couverture de cette séquence par celles de la base de données, et d'autre part, le pourcentage de bases G ou C dans la séquence. Ces caractéristiques sont ensuite fournies à l'algorithme de classification, qui en ressort une identification : soit comme séquence chromosomique, soit comme séquence plasmidique. *PlasForest* a été testé sur plusieurs nouveaux jeux de données, et ses performances dépassent largement celles de 4 autres outils d'identification publiés peu de temps auparavant. En particulier, *PlasForest* réussit à concilier forte sensibilité et forte précision, et sa fiabilité est toujours élevée sur les contigs les plus courts. Cet outil est donc utilisé dans la suite de la thèse pour la recherche de plasmides.

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

# PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets

Léa Pradier[1*], Tazzio Tissot[2,3], Anna-Sophie Fiston-Lavier[4] and Stéphanie Bedhomme[1*]

*Correspondence: lea.pradier@cefe.cnrs.fr; stephanie.bedhomme@cefecnrs.fr
[1] Centre d'Ecologie Fonctionnelle et Evolutive, CNRS, Université de Montpellier, Université Paul Valéry Montpellier 3, Ecole Pratique des Hautes Etudes, Institut de Recherche Pour le Développement, 34000 Montpellier, France
Full list of author information is available at the end of the article

## Abstract

**Background:** Plasmids are mobile genetic elements that often carry accessory genes, and are vectors for horizontal transfer between bacterial genomes. Plasmid detection in large genomic datasets is crucial to analyze their spread and quantify their role in bacteria adaptation and particularly in antibiotic resistance propagation. Bioinformatics methods have been developed to detect plasmids. However, they suffer from low sensitivity (i.e., most plasmids remain undetected) or low precision (i.e., these methods identify chromosomes as plasmids), and are overall not adapted to identify plasmids in whole genomes that are not fully assembled (contigs and scaffolds).

**Results:** We developed PlasForest, a homology-based random forest classifier identifying bacterial plasmid sequences in partially assembled genomes. Without knowing the taxonomical origin of the samples, PlasForest identifies contigs as plasmids or chromosomes with a F1 score of 0.950. Notably, it can detect 77.4% of plasmid contigs below 1 kb with 2.8% of false positives and 99.9% of plasmid contigs over 50 kb with 2.2% of false positives.

**Conclusions:** PlasForest outperforms other currently available tools on genomic datasets by being both sensitive and precise. The performance of PlasForest on metagenomic assemblies are currently well below those of other k-mer-based methods, and we discuss how homology-based approaches could improve plasmid detection in such datasets.

**Keywords:** Plasmid identification, Homology, Random forest classifier, Genomic datasets

## Background

Plasmids are extra-chromosomal fragments of DNA that replicate autonomously in the host cell. They often carry genes that can provide a benefit under specific environmental conditions [1]. These mobile genetic elements remain a major biological concern for health and agriculture policies due to their ability to accumulate and spread resistance genes. Indeed, the frequency of plasmids, and of the resistance genes they carry, can increase quickly in populations thanks to their high mobility both within hosts (through

© The Author(s), 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

chromosomal integration) and between hosts (through horizontal gene transfer, hereafter HGT). Due to their high mobility and the function of the genes they carry, plasmids have a great ecological importance in many bacterial communities.

Until recently, the identification and isolation of these mobile genetic elements was limited to a narrow subsample of the bacteria diversity. Most past studies have only focused on specific species-plasmid associations of medical or agronomic interest [2, 3]. On the one hand, the plasmid characteristics determining their role in resistance propagation were established by phenotypic approaches, which require the focal strain to grow on culture media. These approaches could assess the ability of genetic elements to transfer small conjugative plasmids carrying selectable markers into a new recipient, i.e. transfer, replication and expression of selectable markers [4], but were useless for non conjugative plasmids. On the other hand, PCR-based detection methods [5] could detect the number of copies for specific plasmid sequences, but did not allow to understand the ecological characteristics of mobile genetic elements, as for example their host range.

The fast development of sequencing technologies and reduction of sequencing costs led to the rapid increase of available genomic and metagenomic data sets. This material contains a vast amount of information on plasmid diversity, plasmid host-range, resistance conferred to specific host taxa, etc. that could allow to better understand the circulation and spread of plasmids and the genes they carry. Accessing this information requires new tools to process partially assembled datasets to identify plasmids sequences.

Well-defined plasmids can be identified through homology search (e.g., PlasmidFinder [6], PLACNET, PLACNETw [7]). These programs basically look for similarities between a query sequence and a local database. Query and subject sequences are usually quite long (at least several hundreds of bases), so the probability of finding similarities by chance (and thus to wrongly identify a sequence as plasmid) is very low. Homology search is usually very precise and reliable. However, current data limitations can substantially decrease the sensitivity of this method. Indeed, it requires an exhaustive database composed of a wide variety of genomics data, enough to cover many taxa, and even if this is the case, rare plasmids may not be identified. In particular, the range of taxa in which a plasmid can be replicated and maintained (i.e., its host range [5]) can greatly vary: some being restricted to a few close species, and others consisting in a wide range of taxa across the phylogeny. Thus, plasmids with a narrow host range (especially when only present in uncultivated species) could be nearly impossible to identify through homology.

In the absence of an exhaustive database, an accurate identification method requires to define broader associations with plasmids. This can be achieved by the reconstruction of plasmids through homology-based clustering (PLACNET, PLACNETw [7], MOB-recon [8]): the query sequences which are homologous to the same reference plasmid sequences are likely to be part of the same plasmids. This approach can be successful in two cases: (1) if the generated graphs are manually pruned by an expert user (PLACNET, PLACNETw [7], but this strongly impedes the possibility to apply this method to large datasets) and (2) if the query sequences are also compared to complementary databases (e.g. insertion sequences and repeated elements, MOB-recon [8]).

Other algorithms identify plasmids in draft genomes and metagenomes without relying on homologies, but rather on k-mers frequencies: PlasFlow [9]; cBar [10]; PlasmidSeeker [11]; PlasClass [12]; PPR-Meta [13]). Scoring sequences based on their k-mers frequencies is quick, automated, and adaptable to various data, and may result in valuable predictions (e.g., more than 95% of correct predictions for PlasFlow [9]). However, the frequencies of k-mers cannot be estimated with precision in a short contig, even when short k-mers (e.g., 7-mers for PlasFlow [9]) are used. For this reason, k-mer-based methods systematically rule out contigs below 1 kb. To get around such limitations, more recent approaches rather rely on mathematical models of sequence composition, such as one-hot matrices (PPR-Meta, [13]), thereby achieving better performances on short sequences.

Overall, the current plasmid identification methods are limited, in either precision or sensitivity. These limitations are not problematic for certain applications: for example, k-mer-based methods are well adapted to detect plasmid sequences in metagenomics samples containing many uncultivable species and homology-based approaches are well suited if users focus only on a few plasmids of interest. Yet, no tool is currently available to classify, with both high precision and sensitivity, a large data set of sequences as plasmid or chromosome. Such a tool would be very useful for example to monitor the spread of a gene family (e.g., antibiotic resistance genes) using partially assembled genomic datasets. Indeed, 90% of available assemblies on public repositories (up to 100% for species with few sequenced genomes) are partially assembled genomes (contigs or scaffolds). It is thus not possible to rely on the size of the genomic elements and their content in specific sequences (as ribosomal RNA genes or origin of replications) to classify them as chromosome or plasmids and an accurate plasmid identification method is required.

Here, we present PlasForest, a new tool for classifying contigs in partially assembled genomic datasets as plasmid or chromosome. Contrary to classical homology-based tools, our method does not attempt to assign query contigs to specific subject sequences, but rather to sort contigs in plasmid and chromosome sequences through a machine learning algorithm fed with parameters extracted from an homology search against an exhaustive plasmid database, as well as other variables (e.g., contig size and %G + C content). PlasForest can discriminate plasmids from chromosome sequences with an overall F1 score of 0.950 for any bacterial contig in genomic datasets. In particular, PlasForest is able to detect up to 77.4% of plasmid contigs under 1 kb with only 2.8% false-positives and up to 99.9% of plasmid contigs over 50 kb with less than 2.2% false-positives. Compared to other currently available tools, PlasForest has a significantly better capacity to correctly identify plasmids from chromosomes in partially assembled genomes. We implemented this tool in a user-friendly pipeline able to identify plasmids in large datasets in a reasonable amount of time.

## Results

PlasForest is a tool that allows to assign contigs in genomic datasets to plasmid or chromosome by using a random forest classifier on variables extracted from a homology search. We tested the sensitivity of its performances to various changes in the training

process, and we provide a comparison with other classical plasmid identification tools on large genomic datasets.

### Pipeline description

A pipeline executing all analyses required for PlasForest was encoded in Python 3.6. This pipeline performs all the different steps described in Fig. 1. First, it filters query contigs by their annotations, to avoid re-identification of sequences already described as chromosome or plasmid (step 1). If the user wants to re-assign contigs with Plasforest, this step can be skipped with option *-r*. Feature acquisition (step 2) consists of submitting the filtered sequences to BLASTn against a local copy of the plasmid database, and calculation of the overlaps between query and subject sequences. Features are then computed for each query contig on its sequence and the distribution of overlaps. The pipeline thus extracts seven features: average overlap, contig size, G + C content, max overlap, median overlap, number of hits, variance overlap (see Fig. 2A). The classification (step 3) passes features to the random forest classifier, which outputs the predicted identification for each query contig. The output can also include the best hits that were found in the plasmid database with option *-b* and/or the features used for classification with option *-f*. Progression of the pipeline can be displayed if run with option *-v*. Predictions can be

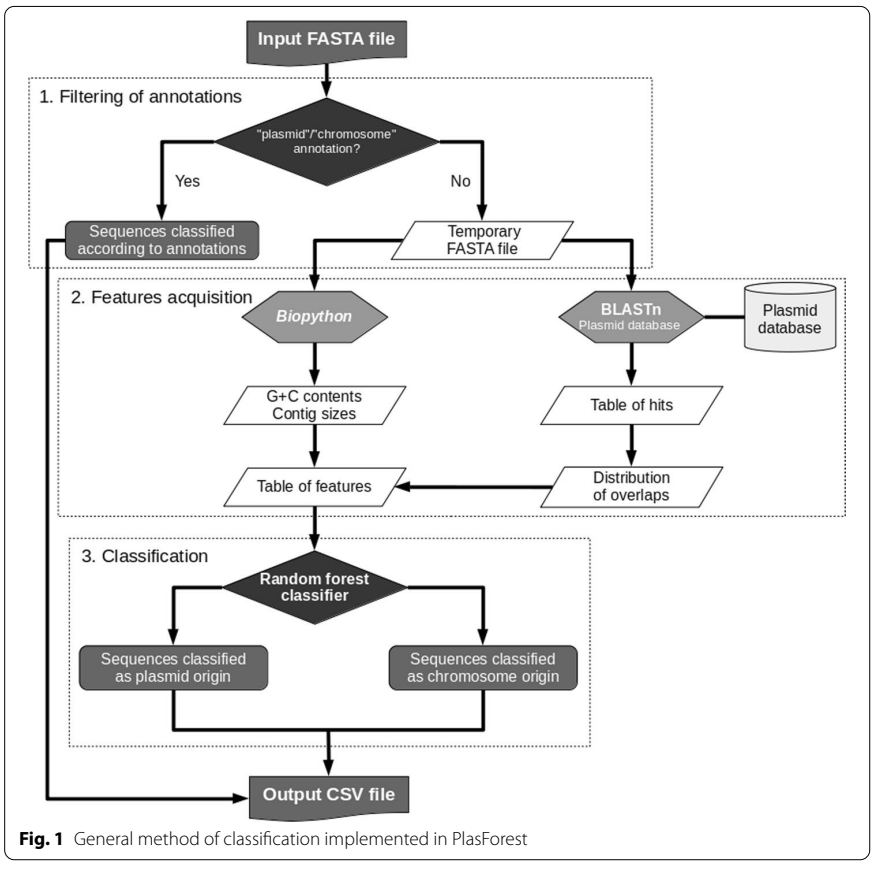


**Fig. 1** General method of classification implemented in PlasForest

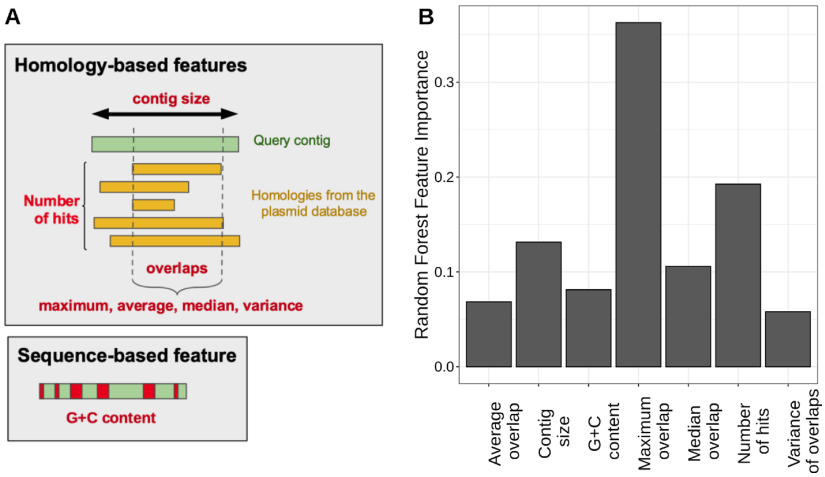


**Fig. 2** Chosen features and their importances in the classification process. **A** Schematic representation of the features extracted from contigs, including homology-based features (number of hits, maximum overlap, average overlap, median overlap, variance of overlaps, contig size) and sequence-based feature (G + C content). **B** Impurity-based feature importance computed with *scikit-learn* library for the seven features kept in the classifier

split to several threads with the multithreading option *-threads* and/or to several batches with the option *-size_of_batch.*

We tested runtime and memory usage for the pipeline on a subset of 1000 contigs randomly drawn from the COMGENOME dataset (see Methods). For a single thread, PlasForest takes on average 490 s and uses a maximum of 2.3 gigabytes of memory (compared to respectively 300 s and 600 megabytes for PlasClass—a k-mer-based tool—on the same dataset). The evolution of these performances with the number of threads is displayed in Additional file 1: Fig. S1.

#### Reliability of the classification method

PlasForest was trained on artificial draft genomes, created by randomly cutting completely assembled genomes into contigs following a size distribution observed in actual draft genomes (Additional file 4: Table S1), that we called balanced training set (Additional file 5: Table S2A). When run on the testing set (Additional file 6: Table S2B) with all seven features (Fig. 2A), PlasForest presented a good ability to discriminate between plasmid and chromosome sequences, with both a high sensitivity (92.7%) and a high precision (97.3%). We tested the sensitivity of PlasForest qualitative changes by evaluating the change in performance when adding or removing features in the training of the classifier and when resampling the training dataset and the plasmid database.

The classifier was trained several times on the balanced training set, by feeding it with different combinations of features. Predictions of the classifiers were made on the testing set. The classification showing the highest MCC (Matthews correlation coefficient) was the one using all the features considered (*average overlap, contig size, G + C content, max overlap, median overlap, number of hits, variance overlap*), and the three variables that

showed the highest feature importance were the *maximal overlap*, the *number of hits*, and *contig size* (Fig. 2B).

Bootstrapping the training set and the plasmid database showed that the performance of PlasForest presents a very low sensitivity to the resampling of the training set (see Fig. 3A). However, the performance of PlasForest decreases substantially when the composition of the plasmid database is resampled (see Fig. 3B). Removing sequences from the plasmid database can therefore significantly diminish the ability of plasForest to correctly assign contigs.

#### Benchmark of plasmid identification methods

We compared the predictions of PlasForest on our testing dataset (see Table 1 and Fig. 4), with those of three classification methods based on genomic signatures (PlasFlow [9], PPR-Meta [13], and PlasClass [12]) and one classical homology-based classification method (MOB-recon [8]). PlasmidSeeker [11] was not included in our analysis, because it requires to know the species from which the genome comes and therefore cannot be used on a wide scale.

PlasForest is overall the most sensitive classifier, being able to predict 92.7% of plasmid contigs. However, for sequences below 2 kb, PlasForest is less sensitive than PlasFlow with down to 77.4% of plasmids correctly predicted. PlasForest had the highest precision, followed by MOB-recon (with respectively 2.7% and 6.1% of sequences incorrectly predicted as plasmids). The precision of PlasForest is lower than that of MOB-recon only for contigs below 1 kb (respectively 2.8% and 0% of sequences incorrectly predicted as plasmids). On these two indices, PlasForest is therefore globally the best software, though other classifiers can have higher sensitivity or precision for short contigs. However, PlasFlow suffers from low precision, especially below 5 kb (up to 70.6% of false positives between 1 and 2 kb), and the sensitivity of MOB-recon can be very low (down to 0.3% of true positives below 1 kb, down to 50% above 50 kb). PlasForest therefore

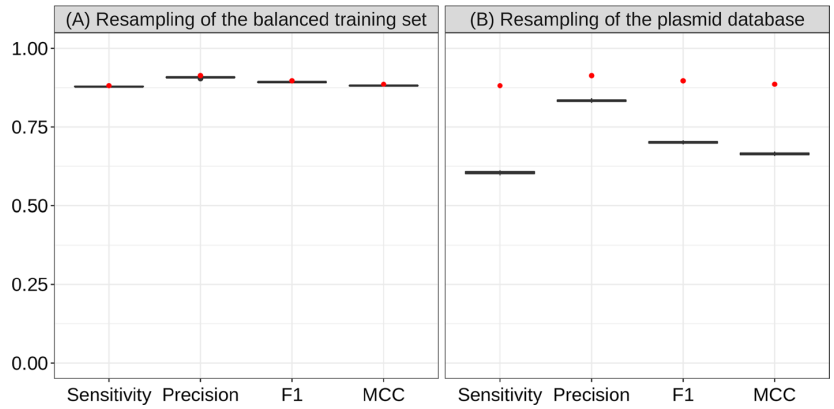


**Fig. 3** Sensitivity of PlasForest to resampling. **A** Performances after 50 resampling of the balanced training set. **B** Performances after 50 resampling of the plasmid database. The initial performances of PlasForest on the testing set are displayed with red dots. The distribution of performances for PlasForest when resampling 50 times either the plasmid database or the balanced training set are displayed in grey boxes

**Table 1** Compared performances of PlasForest and 4 other plasmid identification softwares on the testing dataset

| Contig size | Index | PlasFlow | PlasClass | Mob-Recon | PPR-Meta | PlasForest |
|---|---|---|---|---|---|---|
| [50 bp, 1 kb) | Sensitivity | 0.850 | 0.672 | 0.003 | 0.585 | 0.774 |
| | Precision | 0.373 | 0.473 | 1 | 0.550 | 0.972 |
| | F1 | 0.518 | 0.555 | 0.006 | 0.567 | 0.862 |
| | MCC | 0.203 | 0.309 | 0.047 | 0.361 | 0.819 |
| [1 kb, 2 kb) | Sensitivity | 0.832 | 0.749 | 0.888 | 0.763 | 0.826 |
| | Precision | 0.294 | 0.451 | 0.973 | 0.558 | 0.986 |
| | F1 | 0.434 | 0.563 | 0.928 | 0.644 | 0.899 |
| | MCC | 0.284 | 0.451 | 0.914 | 0.555 | 0.883 |
| [2 kb, 5 kb) | Sensitivity | 0.871 | 0.787 | 0.870 | 0.805 | 0.903 |
| | Precision | 0.308 | 0.464 | 0.946 | 0.552 | 0.975 |
| | F1 | 0.455 | 0.584 | 0.907 | 0.655 | 0.938 |
| | MCC | 0.386 | 0.518 | 0.893 | 0.599 | 0.929 |
| [5 kb, 50 kb) | Sensitivity | 0.899 | 0.816 | 0.729 | 0.889 | 0.976 |
| | Precision | 0.366 | 0.432 | 0.932 | 0.413 | 0.971 |
| | F1 | 0.520 | 0.565 | 0.818 | 0.564 | 0.973 |
| | MCC | 0.518 | 0.545 | 0.811 | 0.557 | 0.971 |
| [50 kb, $+\infty$) | Sensitivity | 0.912 | 0.761 | 0.500 | 0.944 | 0.999 |
| | Precision | 0.374 | 0.423 | 0.919 | 0.259 | 0.978 |
| | F1 | 0.530 | 0.544 | 0.648 | 0.406 | 0.988 |
| | MCC | 0.559 | 0.543 | 0.669 | 0.459 | 0.988 |
| Overall | Sensitivity | 0.883 | 0.785 | 0.670 | 0.831 | 0.927 |
| | Precision | 0.348 | 0.443 | 0.939 | 0.431 | 0.973 |
| | F1 | 0.499 | 0.566 | 0.782 | 0.568 | 0.950 |
| | MCC | 0.483 | 0.531 | 0.775 | 0.540 | 0.945 |

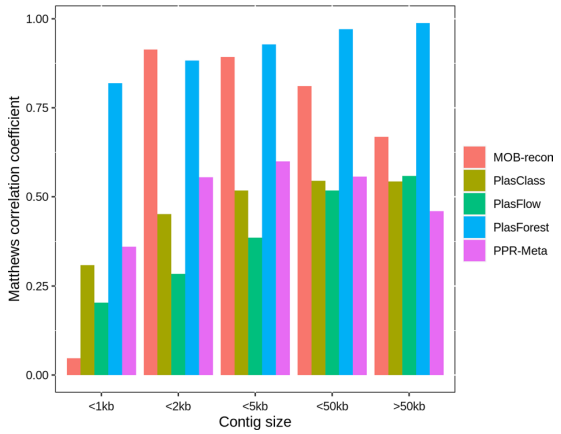


**Fig. 4** Compared performances of PlasForest and 4 other plasmid identification methods on the testing set

outperforms all other programs on the composite indices as well, with the highest MCC (globally 0.945, and even 0.988 for contigs above 50 kb) and the highest F1 score (globally 0.950, and even 0.988 for contigs above 50 kb), except for contigs between 1 and

2 kb where it is outperformed by MOB-recon. Even for very short contigs (under 1 kb) for which k-mers-based methods usually have poor results, PlasForest remains a reliable classifier, with MCC = 0.819 and F1 = 0.862 (respectively MCC = 0.361 and F1 = 0.567 for PPR-Meta, the second-best classifier for these sizes).

#### Applicability of PlasForest to other datasets

We tested further the ability of PlasForest to predict contigs of plasmid origin on three additional datasets. We also compared the results obtained with PlasForest to the results obtained with the four other programs on these three datasets. The COMGENOME dataset gathers artificial contigs, randomly cut with the same size distribution as the testing set, but drawn from complete genomic assemblies that were released after the training and testing sets were constituted (Additional file 7: Table S3). The CONTIG dataset contains partially assembled genomes which were present neither in the training nor in the testing datasets (Additional file 2: List S1), and the METAGENOME dataset contains genomes that were partially assembled from metagenomic short reads (extracted from the datasets of [14] and [15], Additional file 3: List S2). Contrary to the COMGENOME dataset, the genomes used to create the CONTIG and the METAGENOME dataset were not annotated, so we could not measure the accuracy of the predictions of the different programs on these datasets.

On the COMGENOME dataset (see Table 2), PlasForest performances are reduced compared to its performances on the testing set, notably in sensitivity (only 54.3% true positives, compared to 85.7% true positives detected by PlasFlow). Yet, PlasForest remains as precise, with only 11.3% false positives (compared to 9.3% false positives detected by MOB-recon). Overall, on this dataset, it remains the best classifier, with MCC = 0.663 and F1 = 0.674 (respectively MCC = 0.522 for MOB-recon, and F1 = 0.538 for PlasFlow).

On the CONTIG dataset (see Fig. 5A), PlasForest and PlasFlow agree to predict that 14,427 contigs originate from plasmids (80.9% of plasmids predicted by PlasForest). However, 52,999 contigs that PlasFlow detects as plasmids are identified as chromosomes by PlasForest (39.6% of chromosomes predicted by PlasForest). PlasFlow is thus the plasmid identification method whose predictions resemble less those of PlasForest (Cohen's $\kappa = 0.187$). In general, plasmid identification methods based on k-mers disagree with the predictions of PlasForest on this dataset (Cohen's $\kappa = 0.232$ between PlasClass and PlasForest, and Cohen's $\kappa = 0.205$ between PPR-Meta and PlasForest). On the contrary, though PlasForest and MOB-recon only agree to predict that 7350 contigs originate from plasmids (41.2% of plasmids predicted by PlasForest), only 3783 contigs that

**Table 2** Compared predictions of PlasForest and 4 other plasmid identification softwares on the COMGENOME dataset

| Index | PlasFlow | PlasClass | Mob-Recon | PPR-Meta | PlasForest |
|---|---|---|---|---|---|
| Sensitivity | 0.857 | 0.723 | 0.338 | 0.665 | 0.543 |
| Precision | 0.392 | 0.460 | 0.907 | 0.414 | 0.887 |
| F1 | 0.538 | 0.563 | 0.493 | 0.511 | 0.674 |
| MCC | 0.493 | 0.500 | 0.522 | 0.436 | 0.663 |

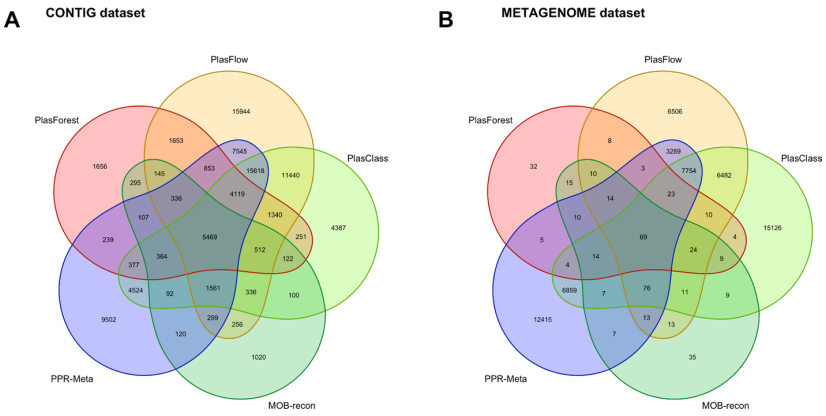


**Fig. 5** Agreement of plasmid identification for PlasForest and 4 other plasmid identification methods on the CONTIG and METAGENOME datasets. **A** Number of contigs identified as plasmids in the CONTIG dataset. **B** Number of contigs identified as plasmids in the METAGENOME dataset. The CONTIG dataset gathers 151,634 contigs collected from 1328 partially assembled genomes. The METAGENOME dataset gathers 143,663 contigs collected from 1000 partially assembled genomes drawn from metagenomic datasets

MOB-recon detects as plasmids are identified as chromosomes by PlasForest (2.8% of chromosomes predicted by PlasForest). Predictions of MOB-recon and PlasForest thus globally agree substantially (Cohen's $\kappa = 0.459$).

On the METAGENOME dataset (see Fig. 5B), PlasForest and MOB-recon globally identify very few contigs as plasmids (respectively 0.18% and 0.24% of contigs). On the contrary, PlasFlow, PlasClass and PPR-Meta all predict a much higher number of contigs as plasmids (up to 25.4% of contigs for PlasClass). However, it is important to note that these three methods have a low level of agreement regarding which contigs are predicted to be plasmids: PPR-Meta and PlasClass agree on only 48.4% of contigs predicted as plasmids by PPR-Meta (Cohen's $\kappa = 0.271$), while this number drops to 36.8% when comparing PPR-Meta and PlasFlow (Cohen's $\kappa = 0.269$).

## Discussion

PlasForest is a homology-based approach, combined with machine learning, which detects plasmids in contig and scaffold genomes. Its operating principle is to seek homologies between query contigs and a large plasmid database, and then to assign a plasmid/chromosome identity to queries with a random forest classifier.

Here we showed that PlasForest is able to deal with large datasets without prior knowledge of the taxonomic background, and that it identifies plasmids with both a high sensitivity and a high precision in unassembled genomes. All the plasmid identification software tested had either a higher false negative error rate or a higher false positive error rate than PlasForest, especially on very short contigs (below 1 kb). PlasForest did not always have the lowest individual error rates, but it optimized the tradeoff between sensitivity and precision, and as a consequence had the highest values of the composite indices F1 and MCC for the vast majority of the size classes. In terms of biological material, PlasForest can be used on draft genomes (in form of contigs and scaffolds) as well as on assembled genomes. PlasForest has been trained and tested on artificial

draft genomes and has good performance on them (overall MCC = 0.945; and even MCC = 0.819 on contigs below 1 kb).

K-mer-based approaches were traditionally the most reliable plasmid identification methods. However, the number of possible k-mers increases exponentially with k and accurately estimating the frequency of all possible k-mers requires large contigs even for small values of k. For example, estimating the frequency of all 256 possible 4-mers in a contig requires thousands of bases, while the size of the shortest known plasmids is below 1 kb [16]. Thus, PlasForest outperforms those methods, especially on short contigs, because the quality of the genomic signature is much less dependent on contig size with homology-based features than with k-mers. K-mer-based approaches are thus valuable only when used on long contigs or in some very specific contexts: for example, PlasmidSeeker [11] achieves up to 100% sensitivity and 99.8% specificity in whole genome sequencing reads by ruling out as chromosomal the k-mers which are shared with a complete, reference assembly. However, the use of PlasmidSeeker on broad diverse datasets is practically infeasible: it requires to know the species from which the genome comes from and that at least one genome has been assembled for this species [11].

The high precision of PlasForest is inherited from its homology-seeking basis. Though homology-based approaches may traditionally show little sensitivity, due to their inability to detect unrelated plasmids, the performances of PlasForest and MOB-recon have significantly improved plasmid identification. One of the reasons why PlasForest is as sensitive as (or even more than) k-mer-based approaches is that it aggregates measures of homologies in a classifier. Thus, it not only considers the presence of homologies, but it also measures the quality and diversity of these homologies and this improves the accuracy of the identification process.

Many bacterial species remain uncultivable but an increasing number of metagenomics tools have allowed access to genomic data in environmental samples without the necessity of obtaining pure cultures. Other identification tools specially designed for such data (e.g., PlasFlow [9]) offer the opportunity of taxonomic assignment of the sequence identified as belonging to plasmids. During the development of PlasForest, we addressed the classification of contigs from genomic (and not metagenomic) data into plasmids or chromosomes. The pipeline of PlasForest offers the possibility to identify which plasmids from the database have the strongest homology with the query. It should be noted though, that the taxonomic assignment can strongly depend on (1) the plasmid host range and (2) horizontal gene transfer events that are yet massively undetermined in bacteria. Any attempt (from PlasForest or from any other method) of taxonomic assignment for broad host range plasmids is thus at least imprecise and can sometimes be impossible [17].

Finally, PlasForest is the best alternative for plasmid identification in genomic assemblies. However, as PlasForest mostly relies on homologies with a plasmid database, the vast majority of plasmids it is able to detect are related to those of its database. It thus showed reduced performances when resampling its database or when testing it on the COMGENOME dataset, which included genomes published after the construction of the training and test sets. Unrelated plasmids, especially with narrow host ranges or when the host is uncultivable, are difficult to detect through this method so far. Thus, in its current state, PlasForest cannot be applied to detect plasmids in metagenomic

assemblies. Both PlasForest and MOB-recon, two identification methods that were initially designed for genomic assemblies, detected very few contigs as plasmids in the METAGENOME dataset. Though it is very difficult to predict the actual frequency of contigs originating from plasmids, it is e.g. 12.7% in the COMGENOME dataset. It is thus far higher than 0.2% so PlasForest and MOB-recon most probably have a very low sensitivity for metagenomic assemblies. However, we argue that investing in homology-based approaches for plasmid detection in metagenomic assemblies is crucial. Indeed, methods based on detection of k-mers probably allow to detect a much higher number of plasmid contigs in unknown species, but they also have a high false-positive rate: they detected up to 25% of plasmid contigs in the METAGENOME dataset. We also showed that their predictions are in high disagreement with one another, therefore rendering them hard to rely on. That is why assembling methods such as metaPlasmidSPAdes [18] also rely on homology search, but these methods only apply to newly sequenced datasets and in general substantially increase the running time of assembly. As novel methods to detect plasmids in metagenomes need to rely on homology search, new datasets will be required. Some sequences from assembled plasmidomes (some being already available on MG-RAST database [19]) could be incorporated in the plasmid database of PlasForest. But most importantly, when plasmids can be mechanically separated from chromosomes in metagenomic samples [20], both chromosomes and plasmids should be sequenced: PlasForest could then be trained on annotated metagenomic datasets. As these datasets become more and more available, it will be possible to update PlasForest such that it will become able to detect plasmids in metagenomes.

## Conclusions

In its current state, among all softwares tested, PlasForest is the best identification method for plasmids in contigs and scaffolds. As shown with the COMGENOME dataset, PlasForest still outcompetes other methods on recently released genomes. We released PlasForest as a user-friendly pipeline, including the trained classifier and the plasmid database, directly available on GitHub (https://github.com/leaemiliepradier/PlasForest). Further releases (at least on a half-year basis) will include a plasmid database updated with the new plasmid sequences submitted to public repositories, and a new classifier trained on this database. This complemented database should be regularly trimmed, in order to keep it of reasonable size without decreasing the performance of PlasForest. Especially in order to identify plasmids in metagenomic data with better accuracy than k-mer-based approaches, plasmidomes should be included in the plasmid database and the classifier should also be trained on annotated metagenomic assemblies.

## Methods

The aim of PlasForest is to combine both the high precision of homology search with the broad sensitivity of signature-based classifiers in order to discriminate contigs of plasmid origins from contigs of chromosomes. We trained a classifier for which decision relies on the homology of sequences with a large database of plasmid sequences. We simulated contigs by randomly cutting assembled genomes, to construct both a dataset to train the classification algorithm (the training set) and a dataset to measure the

classification performance (the testing set). We then compared the classification performance to other plasmid identification tools.

## Data collection

### *Plasmid database*

All bacterial plasmid sequences were downloaded from the NCBI RefSeq Genomes FTP server (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq; September 1st, 2019). This database is composed of 36,450 sequences that we used as reference for homology seeking via BLAST tool [21] (e-value $< 10^{-3}$).

### *Training and testing datasets*

To train the classifier and measure its performance, we randomly sampled 10,152 bacterial genomes classified as 'complete', downloaded from the NCBI Refseq Genomes FTP server (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq).

To mimic the sequence material on which PlasForest will be applied (contigs from unassembled whole genomes), the empirical distribution of contig sizes was established from more than 100,000 Refseq unassembled genomes (see Additional file 4: Table S1 for the chosen distribution). This distribution was then artificially recreated from complete genomes in the training and testing sets by cutting plasmids and chromosomes at random locations and keeping a defined number of each contig size in plasmids and chromosomes. Only contigs larger than 50 bp were kept, since most current sequencing approaches do not produce shorter reads [22]. We ended up with approximately 70% of the generated contigs (552,410 contigs coming from 7,400 genomes) to train PlasForest. Genome annotations were used to identify contigs as plasmids or chromosomes (see Fig. 6A). The remaining 30% of the generated contigs (108,175 contigs coming from 2,752 genomes) were used as a testing set. In these two datasets, plasmid contigs were not at the same frequencies for all the contig sizes (e.g., > 30% under 1 kb and 2% over 100 kb). This could have led to an artificial detection bias based on contig size (e.g., a better identification of small plasmid contigs). Thus, we split the initial training dataset into contig size categories (50 bp to 1 kb, 1 to 2 kb, 2 to 5 kb, 5 to 10 kb, 10 to 50 kb, 50 to 100 kb, and over 100 kb), and randomly removed plasmid or chromosome contigs from each category to keep the fraction of plasmid contigs constant (around 10%) across contig sizes. This new dataset is thereafter called *balanced training set* (Additional file 5: Table S2A). No manipulation of the testing set was performed (Additional file 7: Table S2B).

### *Use case datasets*

We created three other datasets to validate the performance of our approach on several use cases (see Fig. 6B).

The CONTIG dataset was created by using all the 1328 bacterial genomes classified as 'contig' that were released on RefSeq between September 1st and 30th, 2020. This dataset gathered 151,634 contigs (see Additional file 2: List S1 for the list of genome identifiers).

The COMGENOME dataset was created by using all the 653 bacterial genomes classified as 'complete' that were released on RefSeq between September 1st and October 31st, 2020. These genomes were thus released after PlasForest was trained, and none

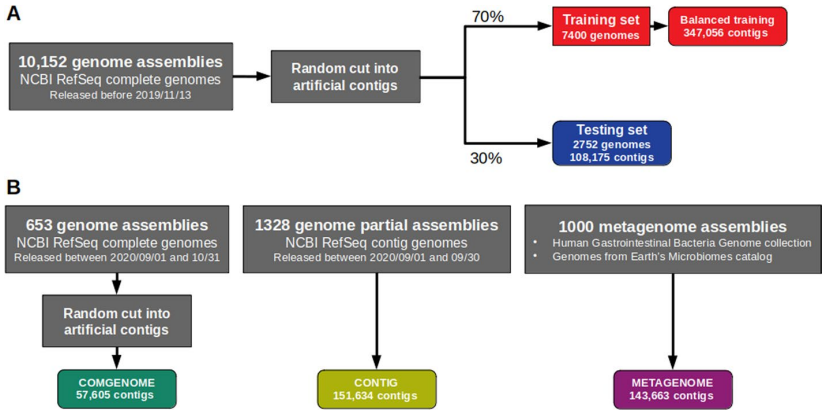


**Fig. 6** Datasets and application of a hold-out method for supervised learning. Schematic representation of the processes that allow to generate the datasets used to build PlasForest and to benchmark its performances. **A** 10,152 bacterial genomes from NCBI Refseq Genomes FTP server were randomly cut into contigs, and were distributed into the following datasets: the *(balanced) training set* contains 70% of the initial 10,152 genomes assemblies and it is used to train the random forest classifier; the *testing set* contains 30% of the genomes. **B** Other genome assemblies were drawn from more recent releases of NCBI Refseq Genomes FTP or from other sources to build the COMGENOME, CONTIG, and METAGENOME datasets. With the testing set, they are used to benchmark the performance of PlasForest compared to other plasmid identification methods

of their sequences were used to create either the training and testing sets or the plasmid database. To simulate draft genome assemblies, the sequences from these genomes were randomly cut into contigs following the same size distribution as used to create the training and testing sets (Additional file 4: Table S1), but no correction was applied to the frequency of contigs of plasmid origin. This process resulted in 57,605 contigs (see Additional file 7: Table S3 for the list of artificial contigs).

The METAGENOME dataset was created by using two subsamples of draft genomes assembled from metagenomic short reads. 500 genomes were randomly sampled from the Human Gastrointestinal Bacteria Genome collection [14], and 500 genomes from the Genomes from Earth's Microbiomes catalog [15], thus bringing together 143,663 contigs (see Additional file: List S2 for the list of genome identifiers).

## Construction of PlasForest

### *Extraction of the features*

All contigs were compared against the plasmid database using BLASTn [21]. Pair-alignments with homologous sequences (hereafter referred to as "hits") were kept if their e-value was below $10^{-3}$. For each contig and homologous sequence, we computed *overlap* as the fraction of the query contig aligning to the homologous sequence hit. The G+C content of all contigs was computed with the function *SeqUtils.GC* from the Biopython library in Python 3.6 [23].

Our aim is not to assemble plasmids (or to assign contigs to precise replicons), but to identify contigs that originate from plasmids. This motivates for a distinct design from other homology-based approaches. By combining both homology search and measures of nucleotide composition, we aim to obtain a strong distinction between plasmids and

chromosomes. We therefore selected features as follows. (1) ***Maximal overlap*** was measured among hits in the subject database, because we expect that query plasmid contigs should form longer alignments with sequences from the plasmid database than query chromosomes. (2) ***Contig size*** was included as short contigs align more often than large contigs with the subject database. (3) The ***number of hits***, the ***average overlap***, the ***median overlap***, and the ***variance of overlaps*** provide other parameters of the distribution of overlaps among hits, that may help distinguish between chromosomes and plasmids. Indeed, due to recombination events, one may expect that query chromosome contigs will align with subject plasmids, but more rarely than query plasmid contigs. (4) Finally, the ***G+C content*** was also included, as the nucleotide composition of plasmids are most often different from those of chromosomes [24]. This set of features used to train the classifier is schematically displayed in Fig. 2A.

#### *Training of the classifier*

We extracted the differences in the features of plasmid contigs and chromosome contigs thanks to a random forest classifier. This approach relies on a multitude of independent decision trees, which allows for a reduction of individual error [25]. The aim was therefore to build a model able to predict, from the extracted features, whether a contig comes from a plasmid or a chromosome. The random forest classifier algorithm was trained with the *RandomForestClassifier* function from *scikit-learn* library [26] in Python 3.6, using the seven features described above. The number of random decision trees was kept to 500, as out-of-bag error estimate (i.e., the internal error of individual decision trees during the training process) did not significantly decrease when using more trees. The global classification method of PlasForest is described in Fig. 1.

#### *Sensitivity of the classifier*

We tested the sensitivity of PlasForest (1) to the composition of the plasmid database and (2) to the composition of the balanced training set, by performing two independent bootstrap analyses. To assess the importance of the composition of the plasmid database, we resampled the plasmid database with 50 different seeds. We then computed new features for each contig of the balanced training set and testing set. A classifier was trained on the balanced training set for each resampled plasmid database, and its performance was measured on the testing set. To test the sensitivity to the composition of the balanced training set, we resampled the balanced training set 50 times, while the testing set did not change. We trained classifiers on the resulting balanced training sets, and measured their performances on the testing set.

### Measure of classification performances

#### *Indices of binary classification performance*

In order to measure the performance of our trained algorithm to correctly identify plasmid sequences and to compare its performance to other available tools, we computed indices derived from the confusion matrix that are commonly used in binary classifications. Sensitivity (sometimes indicated as recall) is the fraction of positive data (in our case, plasmid contigs) which has been correctly identified as positive and allows to measure the false negative error rate. Precision (also indicated as the positive predictive

value) corresponds to the fraction of positive results that are actually true positives and allows to measure the false positive error rate. A good classifier should be able to optimize both sensitivity and precision i.e. in our case, identify as many plasmid contigs as possible without misidentifying chromosome contigs as plasmids. For this reason, we calculated "composite" indices that reflect the overall performance of the classifier. F1 score corresponds to the harmonic mean of sensitivity and precision: it therefore ranges from 0 (i.e., either precision or sensitivity or both are null) to 1 (i.e., there are neither false positives nor false negatives). F1 score does not take into account true negatives. We also calculated *Matthews Correlation Coefficient* (MCC). This metric corresponds to a correlation coefficient between the observed and the predicted classifications and is generally regarded as a balanced measure that can even be used if classes are of very different sizes [27]. Values range between +1 for a perfect prediction, 0 for a random prediction, and -1 for a prediction in total disagreement with the observed data.

#### *Comparison with other softwares*

We ran 4 other plasmid identification softwares on the same datasets as PlasForest, and compared their predictions and performances to those of PlasForest. The version 1.1 of PlasFlow was downloaded from https://github.com/smaegol/PlasFlow. Taxonomic assignments were not taken into account to assess the performance of the classification. To avoid the algorithm assigning sequences as "unclassified", we used a threshold value of 0.5. MOB-suite was downloaded from https://github.com/phac-nml/mob-suite, and MOB-recon was run with default values. MOB-recon clustering algorithm requires draft genome assemblies as inputs, so contigs on which MOB-recon was tested were gathered by their genome of origin. PlasClass was downloaded from https://github.com/Shamir-Lab/PlasClass and the program was run with default values. The virtual machine version of PPR-Meta was downloaded from http://cqb.pku.edu.cn/ZhuLab/PPR_Meta, and the program was run in VirtualBox v. 5.2.42 with default values. When the algorithm assigned sequences as "phage", they were considered as negative predictions.

**Abbreviations**

HGT: Horizontal gene transfer; MCC: Matthews correlation coefficient.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04270-w.

**Additional file 1. Figure S1:** Compared running time and memory use for PlasForest and PlasClass.

**Additional file 2. List S1:** Identifiers of contigs used to create the CONTIG dataset.

**Additional file 3. List S2:** Identifiers of contigs used to create the METAGENOME dataset.

**Additional file 4. Table S1:** Empirical distribution of contig sizes used to cut complete genomes into contigs.

**Additional file 5. Table S2A:** Identifiers and positions of contigs used to create the balanced training set.

**Additional file 6. Table S2B:** Identifiers and positions of contigs used to create the testing set.

**Additional file 7. Table S3:** Identifiers and positions of contigs used to create the COMGENOME dataset.

**Acknowledgements**

We thank Martijn Callens, Emira Cherif, and the two anonymous reviewers for their helpful comments that helped improve the manuscript.

**Authors' contributions**
LP analyzed genomic data and wrote the manuscript. LP and TT designed the current PlasForest pipeline. SB directed the study. SB and ASFL reviewed the manuscript. All authors read and approved the final manuscript.

**Funding**
This work was supported by the ERC HGTCODONUSE (ERC-2015-CoG-682819). The funding body had no involvement in study design, collection, management, analysis and interpretation of data, and in writing the manuscript.

**Availability of data and materials**
PlasForest was released on GitHub (https://github.com/leaemiliepradier/PlasForest) and is available under GPLv3 license. The NCBI accession numbers of sequences used in this study can be retrieved in the Additional file 5: Table S2A for the balanced training set, Additional file 6: Table S2B for the testing set, Additional file 7: Table S3 for the COMGENOME dataset, and List S1 for the CONTIG dataset. The list of NCBI accession numbers of sequences used for the plasmid database is available on the PlasForest GitHub repository (https://github.com/leaemiliepradier/PlasForest/blob/master/list_ids.txt). The sequences used for the METAGENOME dataset can be retrieved from HGG collection (ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/hgg_mags.tar.gz) and GEMs catalog (https://genome.jgi.doe.gov/portal/GEMs/GEMs.home.html), with accession numbers available in Additional file: List S2. All the sequence coordinates necessary to cut complete genomes into contigs (for the balanced training set, the testing set, and the COMGENOME dataset) are available in Additional file 5: Table S2A, Additional file 6: Table S2B and Additional file 7: Table S3.

## 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]Centre d'Ecologie Fonctionnelle et Evolutive, CNRS, Université de Montpellier, Université Paul Valéry Montpellier 3, Ecole Pratique des Hautes Etudes, Institut de Recherche Pour le Développement, 34000 Montpellier, France. [2]Genomics, Bioinformatics and Evolution. Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain. [3]Centre de Recerca Matemàtica, 08193 Cerdanyola del Vallès, Spain. [4]Institut des Sciences de l'Evolution de Montpellier (ISE-M), Equipe Evolution, Vecteurs, Adaptation et Symbiose, UMR 5554, CNRS-Université Montpellier, 34090 Montpellier Cedex 05, France.

Received: 12 November 2020 Accepted: 14 June 2021
Published online: 26 June 2021

**References**
1. Elwell LP, Shipley PL. Plasmid-mediated factors associated with virulence of bacteria to animals. Annu Rev Microbiol. 1980;34:465–96. https://doi.org/10.1146/annurev.mi.34.100180.002341.
2. Johnson TJ, Logue CM, Johnson JR, Kuskowski MA, Sherwood JS, Barnes HJ, et al. Associations between multidrug resistance, plasmid content, and virulence potential among extraintestinal pathogenic and commensal Escherichia coli from humans and poultry. Foodborne Pathog Dis. 2012;9:37–46. https://doi.org/10.1089/fpd.2011.0961.
3. Poolkhet C, Chumsing S, Wajjwalku W, Minato C, Otsu Y, Takai S. Plasmid profiles and prevalence of intermediately virulent rhodococcus equi from pigs in Nakhonpathom Province, Thailand: Identification of a new variant of the 70-kb virulence plasmid, type 18. Vet Med Int. 2010;2010.
4. Costa R, Götz M, Mrotzek N, Lottmann J, Berg G, Smalla K. Effects of site and plant species on rhizosphere community structure as revealed by molecular analysis of microbial guilds. FEMS Microbiol Ecol. 2006;56:236–49. https://doi.org/10.1111/j.1574-6941.2005.00026.x.
5. Heuer H, Binh CTT, Jechalke S, Kopmann C, Zimmerling U, Krögerrecklenfort E, et al. IncP-1ε plasmids are important vectors of antibiotic resistance genes in agricultural systems: diversification driven by class 1 integron gene cassettes. Front Microbiol. 2012. https://doi.org/10.3389/fmicb.2012.00002.
6. Carattoli A, Zankari E, García-Fernández A, Larsen MV, Lund O, Villa L, et al. In silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58:3895–903. https://doi.org/10.1128/AAC.02412-14.
7. Vielva L, De Toro M, Lanza VF, De La Cruz F. PLACNETw: a web-based tool for plasmid reconstruction from bacterial genomes. Bioinformatics. 2017;33:3796–8. https://doi.org/10.1093/bioinformatics/btx462.
8. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genomics. 2018. https://doi.org/10.1099/mgen.0.000206.
9. Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res. 2018;46:35.
10. Zhou F, Xu Y. cBar: A computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics. 2010;26:2051–2. https://doi.org/10.1093/bioinformatics/btq299.

11. Roosaare M, Puustusmaa M, Möls M, Vaher M, Remm M. PlasmidSeeker: Identification of known plasmids from bacterial whole genome sequencing reads. PeerJ. 2018;2018: e4588. https://doi.org/10.7717/peerj.4588.
12. Pellow D, Mizrahi I, Shamir R. PlasClass improves plasmid sequence classification. PLOS Comput Biol. 2020;16: e1007781. https://doi.org/10.1371/journal.pcbi.1007781.
13. Fang Z, Tan J, Wu S, Li M, Xu C, Xie Z, et al. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. 2019;8:1–14. https://doi.org/10.1093/gigascience/giz066.
14. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat Biotechnol. 2019;37:186–92. https://doi.org/10.1038/s41587-018-0009-7.
15. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol. 2020. https://doi.org/10.1038/s41587-020-0718-6.
16. Ciok A, Dziewit L, Grzesiak J, Budzik K, Gorniak D, Zdanowski MK, et al. Identification of miniature plasmids in psychrophilic Arctic bacteria of the genus Variovorax. FEMS Microbiol Ecol. 2016;92:43. https://doi.org/10.1093/femsec/fiw043.
17. Suzuki H, Yano H, Brown CJ, Top EM. Predicting plasmid promiscuity based on genomic signature. J Bacteriol. 2010;192:6045–55. https://doi.org/10.1128/JB.00277-10.
18. Antipov D, Raiko M, Lapidus A, Pevzner PA. Plasmid detection and assembly in genomic and metagenomic data sets. Genome Res. 2019;29:961–8. https://doi.org/10.1101/gr.241299.118.
19. 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;9:386. https://doi.org/10.1186/1471-2105-9-386.
20. Sentchilo V, Mayer AP, Guy L, Miyazaki R, Tringe SG, Barry K, et al. Community-wide plasmid gene mobilization and selection. ISME J. 2013;7:1173–86. https://doi.org/10.1038/ismej.2013.13.
21. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:1–9. https://doi.org/10.1186/1471-2105-10-421.
22. Gweon HS, Shaw LP, Swann J, De Maio N, Abuoun M, Niehus R, et al. The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. Environ Microbiomes. 2019;14:7. https://doi.org/10.1186/s40793-019-0347-1.
23. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3. https://doi.org/10.1093/bioinformatics/btp163.
24. Nishida H. Evolution of genome base composition and genome size in bacteria. Front Microbiol. 2012;3:420. https://doi.org/10.3389/fmicb.2012.00420.
25. Ali J, Khan R, Ahmad N, Maqsood I. Random Forests and Decision Trees. 2012. www.IJCSI.org. Accessed 25 Aug 2020.
26. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python. 2011. http://scikit-learn.sourceforge.net. Accessed 25 Aug 2020.
27. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS ONE. 2017;12:e0177678. https://doi.org/10.1371/journal.pone.0177678.


## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

**Learn more** biomedcentral.com/submissions

BMC

# Section IV : Le rôle de la mobilité dans la propagation des gènes de résistance aux antibiotiques

A l'heure actuelle, il semble évident que la consommation d'antibiotiques est la cause majeure de la propagation des gènes de résistance aux antibiotiques. Des méta-analyses récentes montrent des corrélations assez fortes entre la consommation d'antibiotiques et la présence de bactéries résistantes. Toutefois, la baisse de la consommation dans les pays développés n'est pas toujours suivie par une décrue des infections à bactéries résistantes (Enne et al. 2004). Il semble donc que la quantité d'antibiotiques consommés ne soient pas la seule variable à mettre en cause. De fait, une variété d'autres activités humaines peuvent également moduler le maintien et la propagation des résistances antibiotiques : les infrastructures hygiéniques et sanitaires, la vaccination, la pollution, le mésusage des antibiotiques ou l'emploi d'antibiotiques de mauvaise qualité, etc. De même, plusieurs cas ont été rapportés où des bactéries infectieuses porteuses de résistances antibiotiques ont été propagées d'un pays ou d'un continent à l'autre par le biais du commerce et des voyages.

Enfin, un facteur encore particulièrement négligé reste le rôle des environnements naturels. De fait, les données sur laquelle se base la recherche sur les résistances antibiotiques ont essentiellement été échantillonnées dans les hôpitaux, les élevages, et dans les microbiotes humains et animaux. Pourtant, de nombreux indices laissent penser que les antibiotiques et les résistances aux antibiotiques ont toujours fait partie des écosystèmes naturels. De plus, les écosystèmes naturels ne sont pas isolés des environnements anthropisés : la chasse, l'exploitation forestière, la randonnée, la transhumance, la pollution des cours d'eau, etc. sont tout autant d'exemples de pratiques par lesquelles ces deux types d'écosystèmes peuvent être mis en contact. Les écosystèmes naturels pourraient donc avoir un rôle tant de sources que de vecteurs pour la résistance aux antibiotiques. Bien que certains problème techniques aient pu limiter les études de communautés plus représentatives du milieu naturel en laboratoire (culture et maintien de communautés diversifiées, reproduction de la complexité environnementale), l'exploration de ce type d'environnement devient de plus en plus possible grâce à la métagénomique.

De fait, à ce jour, les études intégratives identifiant des tendances à grande échelle et embrassant toute la complexité de la distribution des résistances antibiotiques font défaut. Cette absence de vision d'ensemble a été identifiée comme une lacune dans notre connaissance des moteurs de la résistance aux antibiotiques. La résistance antibiotique est un problème pour le traitement des infections, et c'est donc sous cet angle que la majorité de la recherche actuelle étudie cette problématique. Cette tendance ne pousse donc à diversifier les modèles d'études et à s'intéresser à des questions évolutives, pourtant

au fondement des problèmes sanitaires.

Dans ce chapitre, je réalise une étude à large échelle de la résistance aux aminoglycosides, en criblant plus de 160,000 génomes bactériens (issus de toute la phylogénie des Eubactéries, de tous les continents, et de tous les écosystèmes continentaux) pour chercher plusieurs familles de gènes encodant des enzymes de modification des aminoglycosides. Grâce à ce jeu de données particulièrement large, je fournis dans un premier temps une description générale de la distribution de ces gènes : l'évolution temporelle de leur prévalence, leur répartition géographique et écologique, leur distribution dans la phylogénie, et leur fréquence d'association à des MGE. Dans un second temps, j'identifie la part relative de plusieurs facteurs, incluant la consommation d'antibiotiques, dans la distribution de ces gènes sur une période de 22 ans en Europe. Globalement, ce chapitre montre que la consommation d'antibiotiques n'est pas le facteur expliquant le plus la répartition des résistances, mais que celle-ci semble être bien davantage liée à différents facteurs de mobilité : leur association à des MGE, les échanges internationaux, et les transferts entre différents biomes.
Le papier correspondant à ce projet est actuellement en cours de revue dans le journal *eLife*.

# Ecology, more than antibiotics consumption, is the major predictor for the global distribution of aminoglycoside-modifying enzymes

Léa Pradier[1], Stéphanie Bedhomme[1]

[1]Centre d'Ecologie Fonctionnelle et Evolutive, CNRS, University of Montpellier, EPHE, IRD, Montpellier, France

## Abstract

Antibiotics consumption and its abuses have been historically and repeatedly pointed out as the major driver of antibiotic resistance emergence and propagation. However, several examples show that resistance may persist despite substantial reductions in antibiotic use, and that other factors are at stake. Here we study the temporal, spatial, and ecological distribution patterns of aminoglycoside resistance, by screening more than 160,000 publicly available genomes for 27 clusters of genes encoding aminoglycoside-modifying enzymes (AMEGs). We find that AMEGs are ubiquitous: about 25% of sequenced bacteria carry AMEGs. These bacteria were sequenced from all the continents and terrestrial biomes, and belong to a wide number of phyla. By focusing on several European countries between 1997 and 2018, we show that aminoglycoside consumption has little impact on the prevalence of AMEG-carrying bacteria, whereas most variation in prevalence is observed among biomes. We further analyze the resemblance of resistome compositions across biomes: soil, wildlife, and human samples appear to be central to understand the exchanges of AMEGs between different ecological contexts. Moreover, the genomic distribution of AMEGs suggests a selection for widening resistance spectra, mostly driven by mobile genetic elements. Together, these results support the idea that interventional strategies based on reducing antibiotic use should be complemented by a stronger control of exchanges, especially between ecosystems.

## 1. Introduction

Antibiotic resistance (AR) is a persistent global public health problem that has increased over the last decades, with resistances spreading faster and faster after antibiotic introduction in clinical use (Witzany et al., 2020). Though first concerns about infections by antibiotic resistant bacteria (ARBs) were formulated in the 1940s, the discovery and development of new antibiotics allowed for treatment substitutions (Podolsky, 2018) during the first decades of antibiotic use. The discovery that AR was frequently acquired by horizontal gene transfer (HGT, Watanabe, 1963) and the subsequent emergence of multi-resistant strains led the international health agencies to start raising the issue, at the end of the 1970s, that antibiotic resistance propagation was threatening to end the antibiotic golden area and jeopardize the huge progress made in the treatment of bacterial infectious diseases. In parallel, the discovery and design of new antibiotics had become more and more difficult (Livermore

et al., 2011). This trend persisted despite the bulk of information provided by recent advances in genomics (da Cunha et al., 2019), thus decreasing the hope for potential treatment substitutions. In 2014, the AMR review estimated that 700,000 deaths per year worldwide were attributable to antibiotic resistances, and predicted 10 million deaths per year by 2050 (O'Neill, 2014, but see de Kraker et al., 2016).

Antibiotics consumption and its abuses have been historically and repeatedly pointed out as the major cause of AR genes propagation (O'Neill, 2014; Podolsky, 2018; Schrijver et al., 2018): the frequency of AR increases in bacteria communities under the selective pressure of antibiotics. To fight this threat, most health agencies are thus focusing their policies on sanitation and mostly on a more reasonable use of antibiotics (World Health Organization, 2015). This perception of the factors driving AR spread and its associated policies still prevail nowadays: e.g. a meta-analysis over 243 studies found a positive correlation between antibiotic consumption and presence of AR (Bell et al., 2014). However, even if antibiotic consumption decreases in most developed countries, AR does not always follow the same path: e.g. a drastic reduction in sulfonamide consumption in the United Kingdom during the 1990s was not followed by a decrease in prevalence of sulfonamide-resistant *Escherichia coli*, which actually remained constant (Enne et al., 2004). Antibiotic consumption therefore does not seem to be the only factor maintaining AR in pathogenic bacteria communities.

Several studies have documented specific examples where antibiotic resistances arrived and spread in new geographic locations through human activities such as food trade, tourism travels or immigration (Berndtson, 2020; Nejidat et al., 2021; Van Boeckel et al., 2019). Although the impact of human activities on the circulation and spread of AR genes is now well documented through the accumulation of specific examples, integrative studies identifying large scale trends are lacking and this absence of general view has been pinpointed as a gap in our knowledge of drivers of antimicrobial resistance (Holmes et al., 2016).

Another strongly overlooked factor is ecology, and a growing number of studies has called for a more comprehensive analysis of AR outside of farms and hospitals (see e.g Bengtsson-Palme et al., 2018). Though data in natural ecosystems remain scarce, AR genes (ARGs) have probably always been natural members of the gene pools of environmental microbial communities (D'Costa et al., 2011). Natural ecosystems may contribute to the spread of AR, both as sources and as vectors of propagation (Bengtsson-Palme et al., 2018; Berendonk et al., 2015; Marti et al., 2014). This realization, combined with an increasing access to genomic data, led to bioinformatic studies which goal was to extend our understanding of AR in ecological contexts that are often overlooked: e.g. β-lactam resistance in dairy industry (Pitta et al., 2016), in slaughterhouses (Lavilla Lerma et al., 2014), in wastewater treatment plants (Karkman et al., 2016), or in natural fresh water (Czekalski et al., 2015). However, even if these studies complement our knowledge on the presence, the frequency, the nature, and the circulation of AR in poorly documented environments (see e.g. Cuadrat et al., 2020; Zhang et al., 2020), descriptions of global distribution patterns and analyses of factors underlying them remain scarce.

In this study, we investigate the relative importance of environmental and genomic factors in

shaping the routes of antibiotic spread on a large scale, focusing on aminoglycoside resistance. Aminoglycosides (AGs) are traditional Gram-negative antibiotics that bind to the 30S subunit of prokaryotic ribosomes and thus inhibit protein synthesis (Mingeot-Leclercq et al., 1999). This antibiotic class was first introduced in 1944 with the successful use of streptomycin against tuberculosis. Several years later, other AGs produced by *Streptomyces spp*. were discovered (kanamycin, spectinomycin, tobramycin, neomycin, apramycin) and introduced in clinics. They were followed in the 1970s by a series of new isolates or derivative synthesized compounds (amikacin, netilmicin, isepamicin, dibekacin, arbekacin, see van Hoek et al., 2011). However, the emergence of resistant strains during the following years, combined with the requirement of administration by injection has reduced the use of AGs in therapies (Murray and Murray, 1991). Nowadays, AGs are only used in humans as a second-line or last-resort treatment for Gram-negative bacteria (Garneau-Tsodikova and Labby, 2016) but they remain frequently used in agriculture and veterinary medicine to treat septicemia and infections of the digestive, respiratory, and urinary tracts in a wide variety of domestic animals (European Medicines Agency, 2017).

The three main mechanisms of AG resistance are (Garneau-Tsodikova and Labby, 2016): (i) decrease in drug uptake (through modifications of membrane permeability or of the peri-membrane ion gradient) and/or increase in drug efflux (through efflux pump activation); (ii) drug inactivating enzymes; and (iii) modification of the drug target by point mutations in the genes coding for the proteins of the ribosomal small subunit (Finken et al., 1993; Sander et al., 1995; Toivonen et al., 1999). Among these mechanisms, aminoglycoside-modifying enzymes (AMEs) are a class of inactivating enzymes that catalyze the transfer of chemical groups on specific residues of the AG molecules, leading to a modified drug which poorly binds to its target (Jana and Deb, 2006). AMEs represent the most common mechanism of AG resistance in clinical isolates and are well characterized biochemically (Ramirez and Tolmasky, 2010). The classical nomenclature of AMEs is based on the group they transfer (acetyltransferases, AACs; nucleotidyltransferases, ANTs; and phosphotransferases, APHs), on the residue modified, and on the resistance profile they confer (Ramirez and Tolmasky, 2010). However, AMEs emerged several times during evolution (see e.g. Salipante and Hall, 2003 for AAC(6') enzymes), so biochemical nomenclatures do not reflect the evolutionary history of any class of AMEs. Finally, many AME genes are carried by mobile genetic elements, which give them the potential to be transmitted both vertically and horizontally (Davies, 1983). They represent today a major threat for the treatment of multidrug-resistant bacteria, notably *Mycobacterium tuberculosis* (Labby and Garneau-Tsodikova, 2013).

Through a computational approach, more than 160,000 publicly available genomes were screened to identify the presence of AME-encoding genes (AMEGs) across the phylogeny of Eubacteria. The present study intended (i) to describe the genomic, geographical and ecological distribution of AMEGs; and based on these data, (ii) to quantify the relative contribution of several key factors (geography, ecology, genomic context, human activities) potentially driving the spread of AMEGs.

## 2. Results

### 2.1 Aminoglycoside resistance is widespread across geography, ecology, and phylogeny.

160,987 publicly available Eubacteria genomes were screened for the presence of AMEGs. The list of genomes and their metadata are listed in Supplementary information, Tab. S1. Published sequences of AMEs and genes coding for AMEs had been previously grouped in 27 clusters of homologous genes (CHGs), each containing sequences of genes and proteins very likely to share a common ancestor (see Methods).

46,053 AMEGs were detected in 38,523 genomes (i.e. about one quarter of the genomes screened). Their distribution in 27 CHGs is listed in Table 1, and all the gene coordinates are listed in Supplementary information, Tab. S2. Our dataset included 54 phyla across the phylogeny of Eubacteria with 89.6% of genomes belonging to the 3 most represented phyla: Proteobacteria, Firmicutes, and Actinobacteriota. In the same way, though we analyzed genomes from 13,879 species, only 10 species made up 43.6% of the dataset. Therefore, though we found AMEGs in 23 phyla, Proteobacteria, Firmicutes, and Actinobacteria contained 97.2% of detected genes. The frequency of the resistance carrying genomes presented very contrasting patterns across CHGs and phyla (Figure 1). For each CHG, the phylogenetic diversity of the species in which it was detected was evaluated by calculating Faith's distances (Faith, 1992). This reveals that the number of species carrying these genes ranges from 2 to 468 (respectively for CHG 12 with Faith's distance $d_{Faith}$=1.4 and CHG 8 with $d_{Faith}$=78.4), but also that for CHGs present in numerous species, these species can belong to a small number of phyla (e.g. CHG 3, $d_{Faith}$=24.5 for 406 species) or be largely spread across the bacteria phylogenetic tree (e.g. CHG 6.1, $d_{Faith}$=66.7 for 418 species). Regarding the phyla, some present a high diversity of CHGs (e.g. Proteobacteria, Actinobacteria, Firmicutes I) whereas others only contain a very small number of different CHGs. This diversity of CHG represented across phyla is however largely explained by the differences in the number of genomes sampled for each phylum (Pearson correlation between the number of genomes sampled and the number of CHGs carried for each phylum, $\boldsymbol{\rho}$=0.820, $p=1.7.10^{-6}$), which is itself linked to the presence of species of medical interest in some phyla.

**Table 1: List of CHGs by biochemical function**

| Cluster | Biochemical function | Number of genes identified |
|---|---|---|
| **CHG 1** | N-acetyltransferases | 619 |
| **CHG 2** | N-acetyltransferases | 711 |
| **CHG 3** | N-acetyltransferases | 4636 |
| **CHG 4** | N-acetyltransferases | 1044 |
| **CHG 5.1** | N-acetyltransferases | 5230 |
| **CHG 5.2** | N-acetyltransferases | 13 |
| **CHG 6.1** | N-acetyltransferases | 11,227 |
| **CHG 6.3** | N-acetyltransferases | 10 |
| **CHG 7** | N-acetyltransferases | 4016 |
| **CHG 8** | nucleotidyltransferases | 4508 |
| **CHG 11** | nucleotidyltransferases | 2990 |
| **CHG 12** | nucleotidyltransferases | 21 |
| **CHG 13** | nucleotidyltransferases | 982 |
| **CHG 14** | phosphotransferases | 675 |
| **CHG 15** | phosphotransferases | 351 |
| **CHG 16** | phosphotransferases | 49 |
| **CHG 17.1** | phosphotransferases | 518 |
| **CHG 17.2** | phosphotransferases | 67 |
| **CHG 19** | N-acetyltransferases | 3124 |
| **CHG 20** | N-acetyltransferases | 871 |
| **CHG 23** | N-acetyltransferases | 2246 |
| **CHG 24** | N-acetyltransferases | 50 |
| **CHG 25** | nucleotidyltransferases | 59 |
| **CHG 26** | phosphotransferases | 20 |
| **CHG 27** | phosphotransferases | 1970 |
| **CHG 28** | phosphotransferases | 23 |
| **CHG 29** | phosphotransferases | 23 |

The location and biome metadata could be recovered for 45,574 genomes and from these data, it was established that AME genes were present in samples coming from all the continents, with the exception of Antarctica. In most regions, the prevalence of AMEG-carrying bacteria (ACBs) ranged between 20% and 40%. However, over 50% of ACBs in Indonesia, Mexico, and Turkey, while only ca. 10% of ACBs in Japan, Eastern Europe, and Eastern Africa (Figure 2A). A quite high spatial heterogeneity in terms of the proportion of each CHG among the ACBs was also revealed (Figure 2B): CHG 6.1 is over-represented in the Southern hemisphere (Africa, South-East Asia, Oceania, Brazil), CHG27 is over-represented in Canada and Mexico and at very low proportion elsewhere, whereas Western Europe, the United States and Japan have a rather balanced representation of all CHGs. ACBs were identified in samples coming from all the biomes investigated. The vast majority of them come from clinical samples (55.3%), human samples (22.1%), and farm samples (12.3%). Despite this large bias in ecological distribution, the prevalence of ACBs per biome varies in a relatively narrow range from 9% of bacteria sampled in domestic animals, to nearly 40% of bacteria sampled in humans (Figure 3). The ecological spread of CHGs is strongly correlated to their phylogenetic spread (Pearson correlation between the number of ecosystems and $d_{Faith}$, $\boldsymbol{\rho}$=0.759, $p=6.8.10^{-6}$). Thus, the CHGs with the largest $d_{Faith}$ are ecologically ubiquitous (CHGs 3, 5.1, 6.1, 7, 8,

and 27), while 8 other CHGs with limited phylogenetic diversity were found in at most 3 biomes (CHGs 5.2, 6.3, 12, 16, 17.2, 24, 25, and 28). Interestingly, clinical samples were found to carry resistance genes of all CHG except two and these two CHGs were specific to agrosystems only (CHG28) and agrosystems and farms (CHG 6.3). Moreover, there are significant negative correlations between the date of first sampling and the phylogenetic and ecological spreads (Fig. 4A, respectively $\rho$=0.712, p=4.4.10$^{-5}$ and $\rho$=0.718, p=3.6.10$^{-5}$): all ecologically ubiquitous CHGs were sampled before 1955, while the CHGs that were never sampled before 2000 could be found in at most 4 biomes.

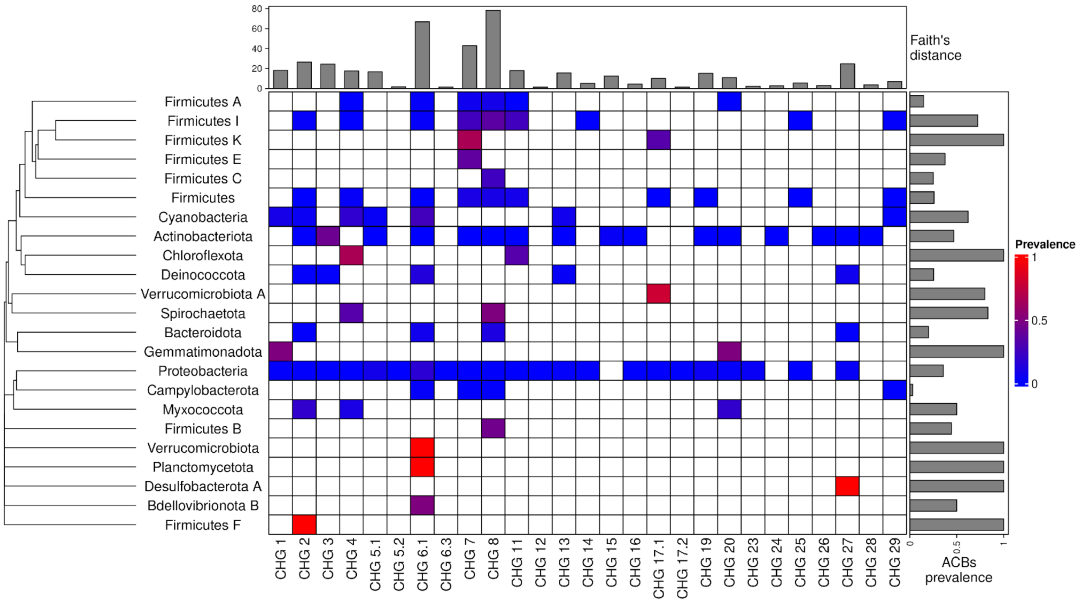


**Figure 1: Prevalence of AMEG carrying bacteria across the phylogeny of Eubacteria.** The phylogenetic tree corresponds to an aggregation of *bac120* phylogeny (Parks et al. 2018) to the phylum level. In the heatmap, blank boxes correspond to the observed absence of a CHG in a phylum. For the other colors, blue to red boxes stand for CHG frequencies from near-zero to one. Gray bars in the top part correspond to the Faith's distance, i.e. the sum of the lengths of all the branches leading to species containing genomes carrying each CHG. Gray bars on the right correspond to the prevalence of ACBs for each phylum, i.e. the proportion of genomes in which at least one AMEG was found.

Our dataset includes bacteria sampled between 1885 and 2019. However, the vast majority of them was sampled recently: 96.3% of genomes were sampled after 1990, and 58.9% after 2010. In this dataset, the first occurrence of an ACB dates from 1905 (Fig. 4A), and the prevalence of ACBs remains constant (about 30%) between 1990 and 2019 (Fig. 4B), though the frequencies of individual CHGs varied much more over time in this same time frame (Fig. 4C), with some CHGs having a regularly increasing (e.g. CHG 5.1, 19) or decreasing (e.g. CHG 6.1) frequency over the whole period or showing one-time frequency peak (e.g. CHG11 in 2003 and 2004). The worldwide dynamics shows more a coexistence across time of a diversity of CHG than sequential substitution of one CHG by another. However, though we also observe coexistence across time at continental levels (Supplementary information, Fig. S1), local time trends suggest that some CHGs might be

progressively replaced: e.g. the frequency of CHGs 3 and 5.1 has increased a lot to the expense of CHG 6.1 in Europe, Southeastern Asia and Oceania since the 1990s.

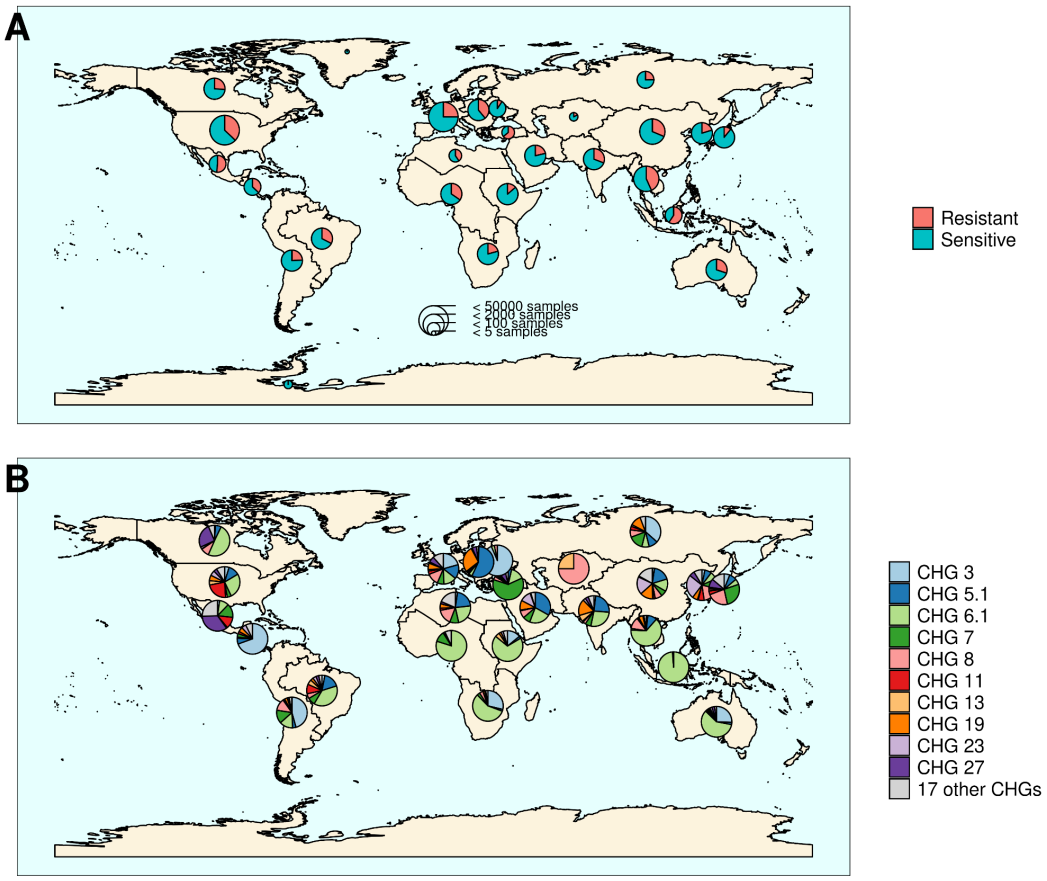


**Figure 2: Distribution of aminoglycoside-resistant bacteria and AME-encoding genes over the world.** (A) Distribution of sampled ACBs in the world. The frequencies of resistant and sensitive bacteria are displayed in red and blue, and the size of pies represents the number of genomes sampled in a given region. (B) Distribution of sampled AMEGs in the world. Pie size is irrelevant to the number of samples.

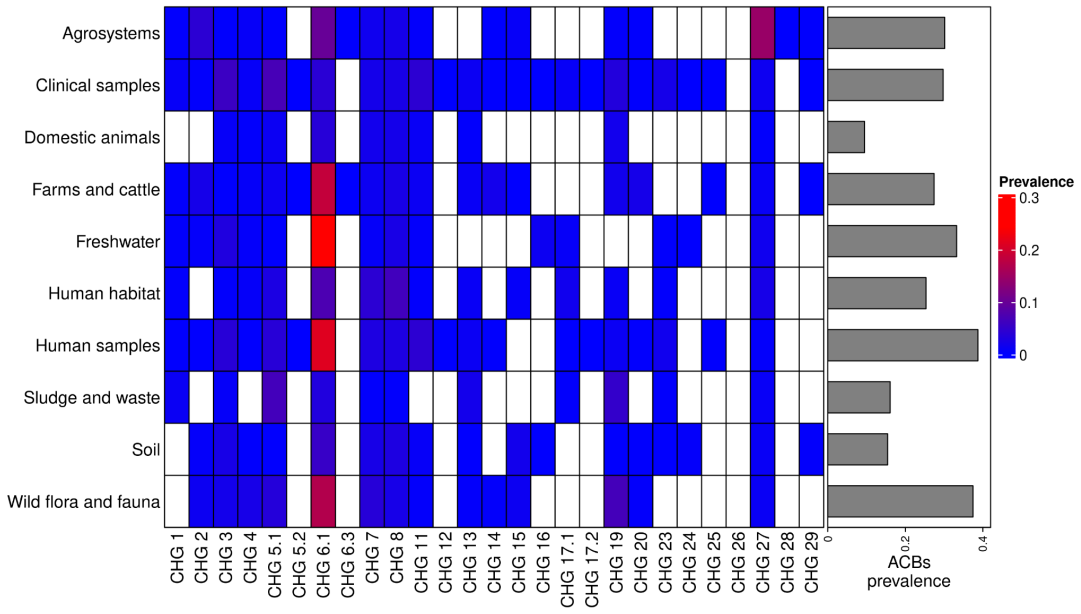


**Figure 3: Prevalence of AMEG carrying bacteria across land biomes.** In the heatmap, blank boxes correspond to the observed absence of a CHG. For the other colors, blue to red boxes stand for CHG frequencies from near-zero to 0.3. Gray bars on the right correspond to the prevalence of ACBs for each biome, i.e. the prevalence of genomes in which at least one AMEG was found.

## 2.2 The European distribution of aminoglycoside resistance is driven by ecology and human exchanges

We investigated the potential role of different factors on the distribution of ACBs in Europe: ecology (structuration in biomes), human exchanges (potential ACB importations through immigration and merchandise imports), and aminoglycoside consumption. This analysis was performed on the timeframe 1997-2018 for which antibiotic consumption data by country and by year was available for Europe. We included in the dataset 16 CHGs for which occurrences were detected in at least 30 genomes sampled in Europe during this timeframe. Most CHGs had very distinct distributions over time and space, so they were analyzed separately.

Within Europe, a Matérn spatial autocorrelation structure was kept in all the models: CHGs tend to cluster between close countries and the autocorrelation structure allowed to control for this. A time autoregressive structure was also kept in all the models, which showed positive autocorrelation for half of CHGs (the others show negative time autocorrelation). Ecology, human exchanges, and antibiotics consumption are kept as explanatory variables for respectively 16 CHGs, 14 CHGs, and 8 CHGs out of 16. The importance of each explanatory variable varies across CHGs, but ecology overall appears to be the most important one: it increased adjusted $R^2_{adj}$ by 15.8% on average (2.1% on average for human exchanges, and 3.0% on average for antibiotics consumption). Interactions between ecology and human exchanges increased $R^2_{adj}$ by 3.4% (for 6 CHGs), and interactions between ecology and antibiotics consumption increased $R^2_{adj}$ by 3.4% (for 3 CHGs) (Figure 5). The effects of all variables and their interactions for individual CHGs are given in Supplementary information, Tab. S3-S18.

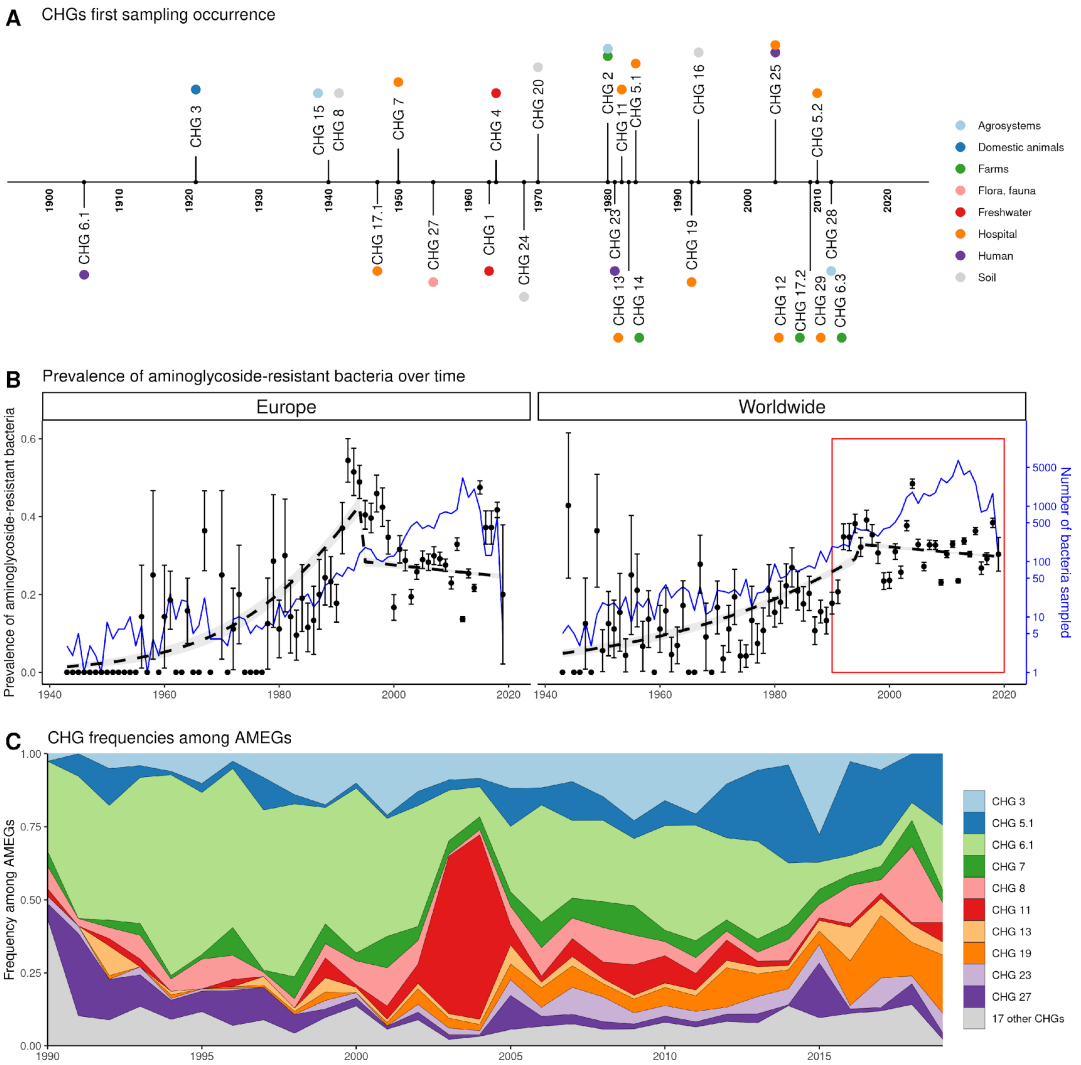


**Figure 4: Global time trends for aminoglycoside resistance.** (A) First sampling occurrence for each CHG in the analyzed dataset. Color points represent the biome in which each CHG was first sampled. CHGs with several color points were sampled in several biomes the first year. (B) Evolution of the worldwide and European prevalences of ACBs between 1943 and 2019. The dots and error bars represent the measured prevalence (± standard error) of ACBs each year. The dotted line represents a binomial regression of the prevalence of ACBs over time, fitted separately before 1995 and after 1995. The blue curve represents the number of bacteria genomes sampled each year (in logarithmic scale). The red rectangle displays the period of time analyzed in panel: (C) Evolution of worldwide CHG frequencies among sampled AMEGs between 1990 and 2019. 17 CHGs for which less than 400 sequences were sampled are grouped and displayed in gray.

The effect of AG consumption is not unidirectional across all the CHGs: when kept in the selected model, its effect is significantly positive for only three CHGs (CHGs 2, 3, and 4), but negative for three others (CHGs 6.1, 17.1, and 19) and nonsignificant for the others. Human exchanges represented in the analysis by our proxy for potential ACBs influx due to trade, has a positive significant effect on the ACBs prevalence for the majority of CHGs (CHGs 4, 6.1, 8, 11, 14, 17.1, 20, and 23), but a negative significant effect for CHGs 3 and 27. Our proxy for potential bacteria influx due to migration has a negative significant effect on the probability to sample ACBs for CHGs

4, 6.1, and 27, and a positive effect for CHG 3. No significant effect of trade and migration were found for other CHGs. Despite the explanatory importance of ecology and the numerous interactions identified, for most CHGs, the probability of sampling ACBs does not significantly differ between most biomes. However, the probability of sampling ACBs in clinical samples is significantly higher for CHGs 1, 3, 5.1, and 19 (respectively significantly lower for CHG 2) than in other biomes. In the same way, CHGs 2, 8, and 27 are significantly less likely to be sampled in farms, while CHGs 3 and 6.1 are significantly more likely to be sampled in humans. CHG 3 is also more likely to be sampled in soil and freshwater, CHG 2 less likely to be sampled in soil, and CHG 6.1 more likely to be sampled in human habitat. We describe the inferred effects of interactions between ecology and antibiotic consumption, and between ecology and human exchanges in Supplementary information, Text S1.

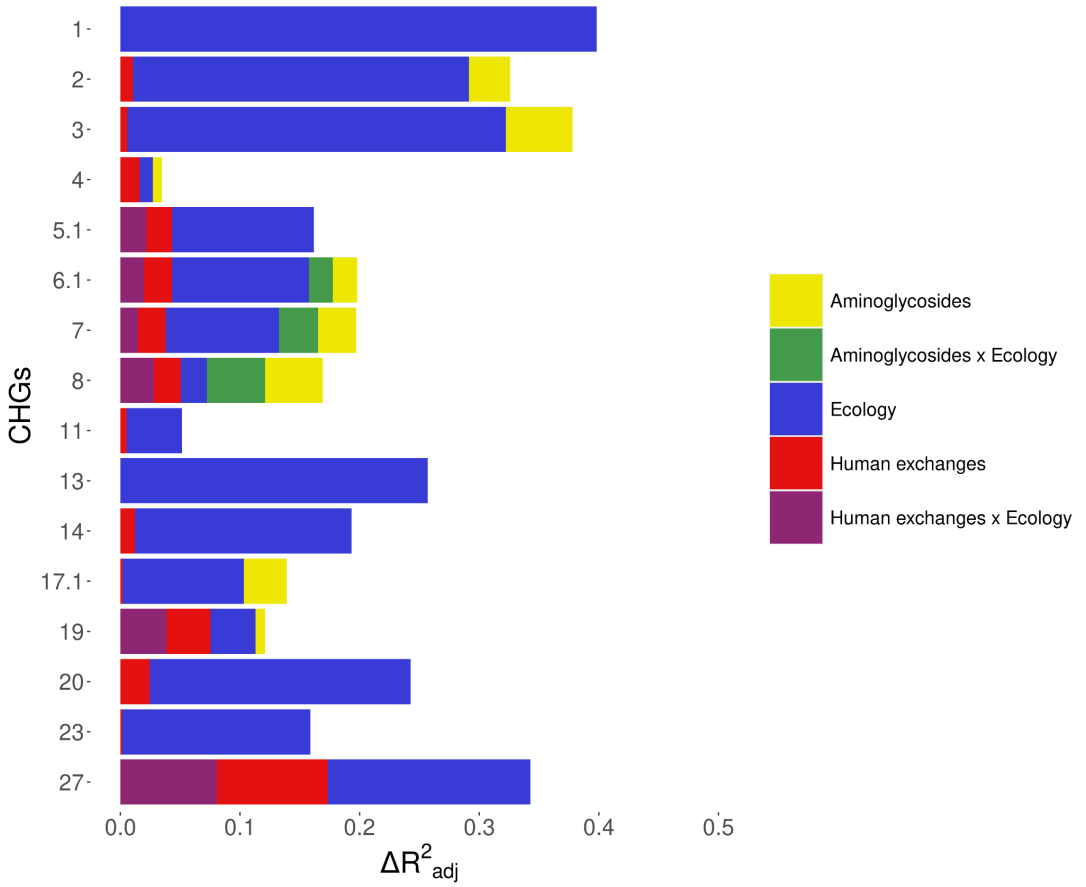


**Figure 5: Relative importance of several factors to explain the distribution of aminoglycoside-resistant bacteria in Europe between 1997 and 2018.** Logistic regressions with a spatial Matérn correlation structure were computed to explain the frequency of 16 CHGs. This figure represents the contribution of each variable in each selected model, as the fraction of adjusted McFadden's pseudo-$R^2$ explained by adding this variable.

For the 14 CHGs sampled in Europe for which we kept human exchanges as explanatory variables, we computed new models based on the decomposition of these effects: we replaced the two effects beneath human exchanges (i.e. international trade and immigration) in the model by only one effect: the proxy for either bacteria influx due to imports of one of the products considered, or bacteria influx due to immigration. We did not find any variable that was able to explain the distribution of all these CHGs, not even for a majority of them (see Supplementary information, Tab. S19). At most, 4 categories of imports were able to impact the distribution of 5 CHGs with a positive effect (beverages, animal feedstuff, meat, sweet foods), and 9 categories were able to explain the distribution of 4 CHGs with a positive effect (cereals, cork, dairy and eggs, fuels, medical products, human migration, oil seeds, tanning products, miscellaneous edible products). However, the distribution of several CHGs could be explained by one or two categories only: cork for CHG 4; rubber and textiles for CHG 5.1; oil seeds for CHG 6.1; animal feedstuff for CHG 7; human migration and sugar for CHG 8; fuels for CHG 11; meat and sweet products for CHG 19; and human migration for CHG 27.

We proceeded to the same analyses for worldwide data on the same timeframe, analyzing only the effects of ecology and human exchanges (see Supplementary information, Tab. S20-S37). Antibiotic consumption data could not be accessed at this spatial scale, but the same trend was found for the other factors: ecology is always the most important explanatory factor, and human exchanges are conserved as an explanatory factor for most CHGs (see Supplementary information, Fig. S2).

### 2.3 The worldwide distribution of aminoglycoside resistance is mainly driven by ecology

Ecology being the most important explanatory factor for the distribution of AMEGs, we investigated the distribution of the different CHGs across ecosystems. Here we approximated an ecosystem as the intersection of a biome and a geographical division (here using the IMAGE24 subdivision from the *rworldmap* library, South, 2011). We described the aminoglycoside resistome composition profile of an ecosystem as the presence or absence of each CHG, in 5-year time frames between 1990 and 2019. For example, in clinical samples of Western Europe, CHG 12 is present over the 2015-2019 period and not before, while CHG 23 is always present except during the 1995-1999 period. We measured differences in resistome composition as Jaccard index, and evaluated their correlation with ecological similarities (being 1 if two ecosystems belong to the same biome, 0 else) and geographical distances between ecosystems. On a global scale, there is no correlation between resistome composition and geography (Mantel test, 9999 permutations, $Z=9.23.10^{9}$, $p=0.585$) but AMEG compositions depend on ecology (Mantel test, 9999 permutations, $Z=174.9$, $p=1.10^{-4}$). Ecosystems are thus more similar in terms of AMEG content if they are from the same biome than if they are geographically close.

A network approach was adopted to get a deeper understanding of the ecology of the resistome composition. We converted the matrix of resistome composition similarities into a network where each vertex consists of a resistome and edges are weighted by Jaccard indices. A minimum spanning network was built that only retained similarity-based edges that could form maximum

similarity paths between resistomes. The resulting network is displayed in Figure 6. Based on the Louvain clustering algorithm, we found three main ecological modules in this network, clustered by their composition in CHGs: (1) clinical, farms, human, and agricultural samples; (2) soil, human habitat, waste, sludge, and freshwater; (3) domestic animals and wild plants and animals. In this network of resistome compositions, wild plants and animals have the highest betweenness centrality (i.e., it is an intermediate in 14 shortest paths in the network), followed by the human habitat (8 shortest paths), soil (7 shortest paths), and human samples (3 shortest paths). All the other biomes have a null betweenness centrality. These four biomes also have the highest closeness centrality in the network. But degree centrality is actually highest for soil (i.e., it shares connections with 7 other biomes), followed by wild plants and animals and human samples (6 connections each), and by agrosystems and farms (5 connections each). Soil, wildlife, and human samples are thus the three biomes for which resistomes are more similar to each other over several time frames.

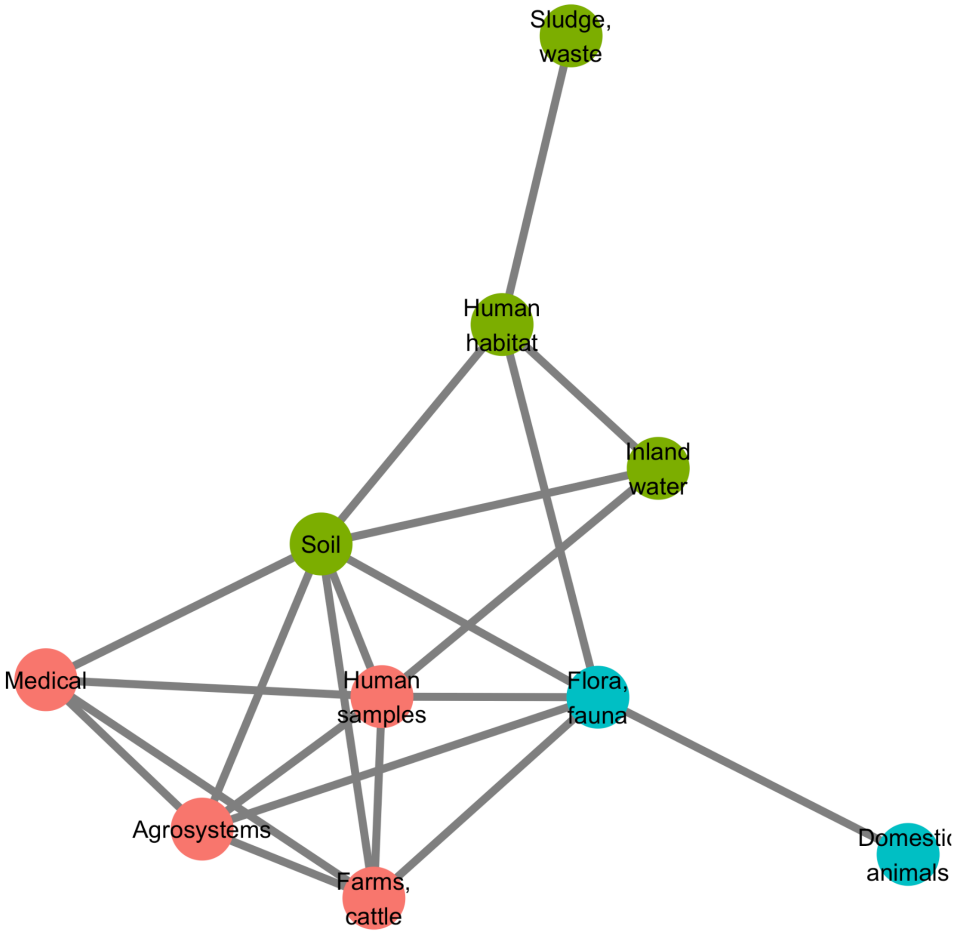


**Figure 6: Minimum spanning network based on shared antibiotic resistance genes between biomes.**

We then investigated whether resistome resemblance within each of the ecological modules defined above is correlated to phylogenetic similarity in bacteria community composition (defined as the whole set of bacteria sampled within a geographical and ecological unit, and measured with the phylogenetic Sørensen index). After correction for the difference in modules sample sizes, it was established that resistome resemblance is significantly, positively correlated with bacteria community composition similarity for both module 1 (95% CI on Pearson correlation coefficient, $\rho \in$[0.075, 0.335]) and module 2 (95% CI on Pearson correlation coefficient, $\rho \in$[0.042, 0.278]), but no significant correlation was found for module 3 (95% CI on Pearson correlation coefficient, $\rho \in$[-0.042, 0.294]).

### 2.4 Resistance gene accumulation is under selection for widening resistance spectra

Among 38,523 ACBs, 6246 (~16%) carried more than one AMEG, and the maximum number of AMEG in a genome was eight. We investigated whether and how this accumulation of AMEGs within genomes lead to a broadening of the resistance spectrum. To do so, we used in separate analyses the classifications provided by two published datasets: either one empirically predicting the resistance phenotypes associated to several ARGs (Feldgarden et al., 2019), or another reviewing and compiling all the antibiotics to which most AMEs were reported to confer resistance (Zárate et al., 2018). These two classifications were used because Feldgarden et al. (2019) probably underestimated each resistance spectrum (due to tests on a limited number of antibiotics), while Zárate et al. (2018) probably overestimated them (due to an exhaustive review from heterogeneous sources). The classification by Feldgarden et al (2019) allowed to infer the resistance spectrum of 11,284 genomes from the 12,982 AMEGs they carried, whereas the one by Zarate et al (2018) allowed to infer the resistance spectrum of 26,006 genomes with 31,910 AMEGs. Using the resistance spectra inferred from Feldgarden et al. (2019), two classical functional dissimilarity indices (functional dispersion, FDis, and Rao's quadratic entropy, RaoQE) indicated that the functional diversity increases, i.e. resistance spectrum widens, with the number of resistance genes carried by a genome (FDis, linear regression, $R^2$=0.756, $p<2.2.10^{-16}$ and RaoQE, linear regression, $R^2$=0.746, $p<2.2.10^{-16}$). More importantly, these increases occur at a higher rate than expected under the hypothesis of a random assortment of AMEGs among genomes, both for FDis (permutation test, 500 perms., $p=1.2.10^{-3}$) and for RaoQE (permutation test, 500 perms., $p=6.7.10^{-3}$). The same trends were also observed using the resistance spectra inferred from Zárate et al. (2018).

Among the 46,053 AMEGs detected in the initial screen, 15,374 were detected to be associated with MGE conferring intergenomic mobility (plasmids, prophages, or ICEs), and 9640 were detected to be associated with a MGE conferring intragenomic mobility (integrons or transposable elements). 7277 genes were detected to be associated with both at the same time. The proportion of genes associated with MGEs is actually very heterogeneous between CHGs, going from

0% for CHGs 16, 24, and 29 to 94.8%, 95.7% and 100% for CHGs 13, 5.1, and 5.2 (Figure 7). We thus investigated the role of MGEs in the accumulation of AMEGs. FDis and RaoQE increase both with the number of MGE-associated AMEGs and with the number of AMEGs not associated with MGEs. Yet, the increase rate is significantly higher than expected under the hypothesis of random assortment for genes associated with MGEs conferring intergenomic mobility (permutation test, 500 perms., $p=1.2.10^{-3}$ for FDis, $p=1.2.10^{-3}$ for RaoQE), whatever the source used to infer resistance spectra. However, for AMEGs associated with MGEs conferring intragenomic mobility or not associated to MGE, the trends differ depending on how the resistance spectra were inferred, and no conclusion can be drawn. The pattern of resistance spectrum widening with the accumulation of AMEGs in a genome is thus stronger and more robust for genes associated with intergenomic mobility MGEs.

Many MGEs are also known to replicate the genes they carry in the same genome. We thus investigated which pairs of AMEGs could be the result of intragenomic replication. The probability for an AMEG to undergo intragenomic replication is at least 10 times higher when associated with MGEs, under the hypothesis of random MGE-gene assortment (permutation test, 500 perms., $p=1.2.10^{-3}$). This probability is similar when considering separately MGEs with intergenomic or intragenomic mobility. MGEs therefore promote the presence of several copies of the same resistance gene in a genome and the broadening of the resistance spectrum by the combination of functionally-distant genes.

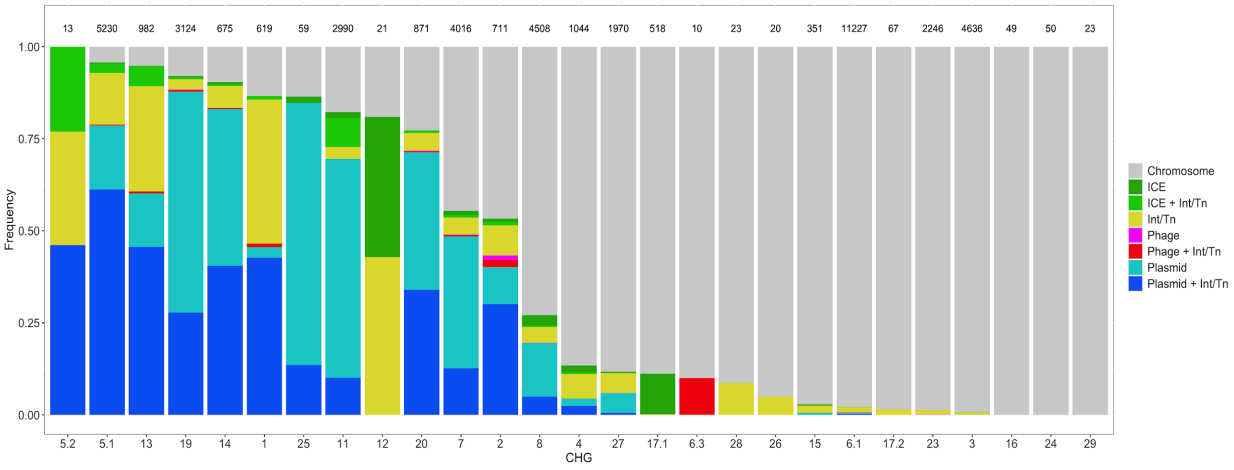


**Figure 7: Distribution of the frequency of genes carried by MGEs per CHG.** CHGs are organized by descending total frequencies of genes associated with MGEs. The number of genes identified for each cluster is indicated above each bar.

## 3. Discussion

### 3.1 AMEGs show a ubiquitous presence in sequenced bacteria genomes

In this study, we defined and screened 27 clusters of genes encoding resistance to aminoglycosides across more than 160,000 genomes. These genomes were widely spread over the phylogeny of Eubacteria (135 phyla, from Actinobacteriota to Verrucomicrobiota), over ecology (from hospitals to fauna and flora), over geography (from Arctic to Antarctic), and over time (from 1885 to 2019). About one quarter of these bacteria were found to carry a gene known to provide resistance to aminoglycosides. This very high prevalence of aminoglycoside resistance is likely biased by the much higher availability of genomic data for human, clinical, and agriculture isolates. Yet, the lowest prevalence of ARBs we found in a biome was actually 9%, which is consistent with previous studies on genomic (Pal et al., 2015; Zhang et al., 2020) and metagenomic (Zhang et al., 2020) datasets sampled from multiple biomes. A large proportion (ca. 40%) of the ARGs found were potentially mobile, i.e. associated with plasmids, integrons, prophages, ICEs or transposons. A previous study reported <20% of potentially mobile ARGs: 29.4% for ARGs conferring multidrug resistance, 15.5% for β-lactam resistance, and 10.5% for aminoglycoside resistance (Zhang et al., 2020). Yet, it is to note that Zhang et al. (2020) only focused on plasmids and integrons, which may potentially have lead to an underestimation of the frequency of potentially mobile ARGs, e.g. associated with ICEs, phages, or transposable elements.

ACBs are widely spread over space, time, and ecology. We were able to detect AMEGs in all the biomes considered, and all over the world. The prevalence of ACBs ranged from 64% in Turkey to 0% in Arctic and Antarctic regions. ACBs were found as early as 1905, i.e. nearly 40 years before the first aminoglycoside was isolated and identified as a potential antibacterial therapeutic agent. This is not a surprise as aminoglycosides are naturally produced by bacteria from the *Streptomyces* and *Micromonospora* genera (Durand et al., 2019) and the selection pressure for the evolution of resistance genes existed long before the clinical use of aminoglycosides, as shown for other antibiotic resistance genes (D'Costa et al., 2011).

ACB prevalence strongly increased between the 1940s and the 1980s, but it has plateaued around 30% since the 1990s (Fig. 4B). The phase of increasing prevalence can easily be explained by the discovery and marketing for clinical use of most AGs. Their massive use must have been a strong selective pressure for the emergence and spread of AMEGs. However, AG consumption has stabilized and/or decreased (at least since 1997 in Europe), so the pre-existing selection pressures have remained stable since the 1990s. Besides, most CHGs coexist over long time periods (Fig. 4C), but they are unevenly distributed across space (Fig. 2B). This apparent coexistence might therefore result from the combination of different local dynamics (Supplementary information, Fig. S1). It is to note that two very distant regions could display the same time trends: e.g., the replacement of CHG 6.1 by CHGs 3 and 5.1 in both Europe and Southern Asia and Oceania since the 1990s, or a surge in the frequency of CHG 11 in both Europe and North America during the 2000s. This suggests that local distributions of AMEGs are driven by local parameters, but that local resistomes are also connected at the global

scale. Finally, the ecological and phylogenetic distribution of AMEGs varies to a great extent, from ubiquitous CHGs to some being restricted to a few biomes and species. Interestingly, ubiquitous CHGs were first sampled earlier than the others, and the extent of ecological and phylogenetic spread for each CHG is positively correlated to the time since its first sampling occurrence. It is yet impossible to determine the causation order. On the one hand, from two CHGs that emerged at the same date, those with a larger ecological and phylogenetic "niche" may have higher chances to be sampled by chance, so they would have been sampled earlier than ecologically and phylogenetically restricted CHGs. On the other hand, CHGs which emerged earlier (and were thus sampled earlier than the others) may have disposed of more time to spread across several biomes and phyla, whereas recently emerged CHGs would be restricted to the biomes and phyla they have had the time to spread to.

### 3.2 Antibiotic resistance prevalence is shaped not only by antibiotic use but also by ecology and human exchanges

Reduction of antibiotic consumption and fighting antibiotic pollution have historically been the most common public health recommendation to control or reduce the spread and rise in frequency of antibiotic resistance (see e.g. Mölstad et al., 2008; Sabuncu et al., 2009). The predominance of this recommendation is also driven by the idea, in evolutionary epidemiology, that selection is the major determinant in the emergence, accumulation, and propagation of resistance (Blanquart et al., 2017; Spicknall et al., 2013). This idea is supported by correlative studies linking antibiotic consumption levels and resistance prevalence, mainly in environments where antibiotic consumption is high such as hospitals or farms (e.g. Goossens et al., 2005).

In this study, one of the goals was to ask whether the impact of antibiotic use on antibiotic resistance genes prevalence could be inferred on a very broad geographical and temporal scale and also across biomes. The local concentrations of antibiotics are likely to strongly vary between biomes, as antibiotics are mainly used in hospitals, outpatients, farms and agrosystems and antibiotics are found as pollutant in other biomes. Data on the amount of antibiotics prescribed in human and veterinary clinics are available but antibiotic pollution data are globally sparse (though local databases have been set up in recent years (see Umweltbundesamt, 2016) and were not available on a sufficiently broad geographical and temporal scale to be integrated in our analysis. However, a large part of the antibiotics consumed are not absorbed by the organism and are excreted, and are thus found in high quantities in urine and feces (Danner et al., 2019; Gothwal and Shashidhar, 2015). Amounts of consumed antibiotics in hospitals and households can thus be found nearly unchanged in effluents (Kümmerer, 2003), and soils could display inhibitory levels of antibiotics due to manuring (Tello et al., 2012). From soil, wastewater treatment plants, and freshwater, indirect contamination can happen towards other biomes (Gothwal and Shashidhar, 2015). Apart from the pollution due to antibiotics manufacturers (Larsson et al., 2018), the presence of man-made antibiotics in the environment is overall highly correlated to antibiotics consumption (see Li et al., 2016 in which the main source of pollution for each antibiotic in a watershed is tracked). Since antibiotic concentrations

in different environments are directly correlated with one another, we can assume that the available antibiotic consumption data are acceptable predictors for the antibiotic concentrations in other biomes.

Our analysis shows that between 1997 and 2018 in Europe, antibiotics consumption is not the main factor explaining AMEGs prevalence patterns. For all AMEG CHGs studied, aminoglycoside consumption was only a minor explanatory factor for the distribution of ARBs, with few positive effects and mainly non-directional effects on the probability to sample ARBs. There are even in this dataset examples of low ARB prevalence occurring in countries with high antibiotics consumption: e.g. over the period, the average prevalence of bacteria carrying CHG 6.1 were 1.5% in France and 18.1% in Austria, whereas these two countries consumed respectively 7.6 and 0.3 tons per 1000km$^2$ of aminoglycosides. How can we explain the discrepancy between our findings and the previously established correlations between antibiotic consumption and antibiotic resistance prevalence as well as the obvious selective advantage conferred by resistance genes across large range of antibiotic concentrations (Gullberg et al., 2011)? There is a strong contrast between the 1960-1990 period during which most aminoglycosides were discovered and introduced for human consumption and AMEGs frequency increased (Fig. 4B) and the period studied in the present analysis (1997-2018 in Europe) during which no new aminoglycoside has been introduced, all CHGs studied had already emerged since at least 10 years, and aminoglycoside consumption was on average stable in Europe. The period studied is thus likely to be a post-emergence period, in which the relative weight of selection as evolutionary force driving the resistance gene frequency variation is likely to be lower than in the previous period. It is also important to integrate the fact that several mechanisms can allow resistant strains to thrive even under decreasing antibiotic selection pressure. (i) Selective pressures on AR are multiple and man-made antimicrobial agents are just one of them. Many ARGs evolved as adaptations to antibiotics produced in natural communities, or as exaptations of metabolic enzymes selected for other functions (transferases, efflux pumps, etc., Martinez, 2012). (ii) ARGs can be maintained in populations by hitchhiking on other genes (e.g. resistance to heavy metals, biocides) carried on the same MGE (Pal et al., 2015). (iii) Carrying an ARG is costly only if the gene is expressed, but ARGs can be silenced in nonselective conditions (Wein et al., 2019) and upregulated in selective conditions only (Jiang et al., 2019; Palmer et al., 2018). The results of our integrated study strongly suggest that reducing antibiotics consumption is a necessary but not a sufficient measure to limit the propagation of antibiotic resistances and reduce their prevalence. Complementary measures based on the reduction of other selection pressures (such as heavy metal pollution maintaining ARG frequency should be implemented (Seiler and Berendonk, 2012). Some studies even argue that the spread and fixation of some AR families may no longer depend on the clinical use of antibiotic agents (Baquero et al., 2015).

### 3.3 Trade and migration matter more

In our dataset, human exchanges explain a significant part of the variation in the prevalence of ACBs for a higher number of CHGs than antibiotics consumption does. Human exchanges were

integrated in our models as imports and immigration weighted by the ARGs frequency in the IMAGE24 region of origin (see Fig. 2B). Over the 1997-2018 period, we observed this trend within Europe, and human exchanges also have an important effect worldwide (Supplementary information, Fig S1 and Tab. S20-S37), although at this larger scale it was impossible to compare their effects to the effect of antibiotic consumption. Numerous examples of direct impact of human activities, outside of antibiotic use, on the emergence, retention, and propagation of AR have been documented: ARBs can be carried over continents by plant (Zurfluh et al., 2015) and animal products (Eltai et al., 2020; Le Hello et al., 2011), exchanged through international trade, as well as immigrants (Nellums et al., 2018) and travelers in general (e.g. Lepelletier et al., 2011). Among these ARBs, several were identified to have caused human infections and epidemics: e.g. the serotype Kentucky ST198 of *Salmonella enterica* (Le Hello et al., 2011); the German Shiga-toxin producing *Escherichia coli* (Buchholz et al., 2011; Rasko et al., 2011). The ARGs carried by imported strains may then spread to local strains and species by horizontal transfer, thus enriching the local resistome or increasing the local frequency of resistant genes. The decomposition of human exchanges in different good categories revealed that no category, either specific (e.g. animal feedstuff imports) or general (e.g. animal products imports), contributes to explain the distribution of all CHGs. This could be due to different CHGs having different geographical origins, to different CHGs being carried by bacteria with different ecological niches, or to contingency in the different transmission pathways. However, imports of meat, beverages, animal feedstuff, and sweet foods explain a significant part of the variation in the distribution of several CHGs. These imports might be responsible for spreading ARGs over large distances. Yet, these factors were identified by correlation only, so whether these imports are transmission pathways for these CHGs cannot be assessed without direct sampling of such products.

The importance of migration of AR genes, through trade and human travels, as a factor explaining the variation in AR gene frequency, revealed by our analysis suggests that reducing the import of ARBs would be an efficient way of limiting AR gene propagation. Procedures of ARG monitoring in traded goods have been set up but are probably insufficient. For example, as reported for the European Union (Schrijver et al., 2018), the modalities for detection and characterization of ARGs in meat are not harmonized between countries, and only focus on clinically relevant species, thereby ignoring the risk of horizontal transmission to and from other unmonitored strains. However, infections by ARBs are often difficult to connect to consumption of given products, and most recommendations insist on hygiene during food processing and cooking (Capita and Alonso-Calleja, 2013).

**3.4 Ecology matters most**

Not only did we find a wide diversity of ARGs outside of hospitals and farms, where antibiotics are the most consumed (see Fig. 3), but differences between biomes explained most of the variance when the frequency of ARBs was modeled over time, space and ecology. Across ecosystems (defined as the intersection of a biome and a geographical unit), we found that the resemblance

between resistomes depended more on the biomes in which they were sampled, than on their geographical location: geographically distant ecosystems belonging to the same biome have more similar resistomes than geographically close ones belonging to different biomes. This is consistent with previous studies: functional metagenomic selection experiments show drastic differences in the resistance profiles between soil and human gut microbiota isolates (Gibson et al., 2014), and less than 10% of ARGs sampled in wastewater treatment plants can also be found in other environments (Munck et al., 2015). Yet, consistent with Forsberg et al. (2012) who found ARGs shared with 100% identity between soil bacteria and human pathogens, many ARGs were still shared between biomes, including between natural and anthropized biomes.

Additionally, the study of certain biomes seems crucial to understand the spread of AR. Indeed, in the network of resistomes, two of the three most central biomes are “fauna and flora” and “soil”. This could be explained by both antibiotics and ARGs transferred from humans and anthropized habitat to natural ecosystems (e.g. through manuring and fecal pollution, see Karkman et al., 2019; Xie et al., 2018). Natural ecosystems have actually been suggested to constitute reservoirs for ARGs originating from hospitals (see e.g. Baquero et al., 2009; Tripathi and Cytryn, 2017). Also, most natural aminoglycosides are produced by two bacteria genera, *Streptomyces* and *Micromonospora* (Durand et al., 2019), that live mostly in soil and decaying vegetation. It is therefore possible that many of the CHGs first evolved in soil and flora bacteria as defense mechanism against the chemical warfare of these two genera and remained at high frequency because of continuous selection pressure. This could be illustrated by the fact that, in our dataset, even though 19 times more genomes were sampled in hospitals than in soil, only 2.5 times more CHGs had their first documented occurrence in clinical samples than in soil. If not all CHGs emerged in soil and wildlife, these two biomes appear at least as a hub for the circulation of ARGs between biomes.

As previously suggested, many different biomes may represent sources and/or reservoirs of AR, before it is transmitted to human-associated bacteria (Bengtsson-Palme et al., 2018). It is thus crucial to identify how AR is passed on between biomes. In the biome network, three major modules of resistome were found: (1) ARGs sampled in hospitals, humans, farms, and agrosystems; (2) ARGs sampled in soil, human habitat, waste, and freshwater; and (3) ARGs sampled in domestic and wild animals and plants. Previous metagenomic studies using abundance-based metrics found that resistome composition is correlated to bacteria community composition: either globally (Pehrsson et al., 2016), or in specific biomes: soil (Forsberg et al., 2014), wastewater treatment plants (Ju et al., 2019), human feces (Pehrsson et al., 2016). In agreement with these studies, while using presence-absence-based indices, we found substantial correlation between resistome similarity and phylogenetic beta-diversity in the first (hospitals, humans, farms, agrosystems) and the second modules (soil, human habitat, waste, and freshwater). This correlation indicates that resistome similarity is at least partially due to species composition similarity, and so to the exchange of ARBs between biomes. Indeed, when bacteria disseminate between ecosystems, both community compositions and resistomes tend to converge. Consequently, our results also indicate that the propagation of resistances is not totally explained by gene exchange by horizontal transfer. It could

result from multiple factors ranging from dissemination of ARBs: between ecosystems or from common sources, by abiotic dissemination, (Allen et al., 2010), fecal transmission (Karkman et al., 2019), transmission by wastewater (Vaz-Moreira et al., 2014), or transmission through the food chain (Founou et al., 2016). The correlation between resistome and species composition was not found in the second module (domestic animals, flora and fauna), even though we controlled for the lower sampling in this module. This surprising result could actually be due to a higher environmental fragmentation in these biomes , which could result in biased sampling, as documented in metagenomic studies (e.g. Lu et al., 2014). Moreover, previous results showed that higher spatial structures increase the likelihood of HGT events to low-abundance strains (Cairns et al., 2018), that might not be sampled and therefore decrease the likelihood to find a correlation between resistome composition and community composition. The absence of correlation between resistome and species composition leads to formulate the hypothesis that horizontal transfer might play a more important role in ARG propagation in this module than in the other two. This hypothesis could be tested using HGT detection methods such as the one developed in Corel et al. (2018).

### 3.5 Combination of AMEGs in a genome is shaped by diversifying selection and MGEs movements.

In our dataset, 40% of AMEGs are carried by MGEs. MGE carriage is known to be a strong determinant of the capacity to propagate within and between genome and has been identified as key element for classifying a given antibiotic resistance in the highest risk category (Martínez et al., 2015). The association with MGE frequently takes the form of an embedded structure: for example, half of transposable elements carrying ARGs were also located on plasmids (Fig. 7). The embedded structures very likely multiply the propagation potential or combine the intra- and inter-genome propagation potential. At a finer scale, the fraction of genes associated to MGE as well as the identity of the associated MGE strongly differ between CHGs, suggesting that these clusters have contrasted propagation probabilities and pathways.

The MGE carriage of resistance genes also means that they can be acquired from different sources and combined within a genome. A second interesting finding is that the combination of AMEGs within genomes is not random: the resistance spectrum widens at a higher rate with the number of AMEGs than expected by chance, which suggests that the combination of AMEGs in a genome is under selection for functional diversification. This broadening of the resistance spectrum is mainly driven by MGEs with inter-genomic mobility (i.e. plasmids, ICEs, and prophages).

Finally, we also found that MGEs increase the likelihood for a genome to carry several copies of the same ARG. It has to be noted here that the copy number accessible from whole genome sequence data is an underestimation of the actual number of copies of the gene for all genomes in which the ARG is carried by a plasmid, because plasmids are usually present in more than one copy. The presence of several copies of a resistance gene, either because of its association with an intra-genomic MGE or its plasmid carriage not only increases the expression level of ARGs (Depardieu et al., 2007; Sandegren and Andersson, 2009), but also allows bacteria to evolve new antibiotic

resistance functions on duplicated sequences (San Millan and Maclean, 2017; Sandegren and Andersson, 2009), thus also potentially participating in the functional diversification.

### 3.6 Limitations

The approach taken here allows to exploit a large amount of publicly available data to gain a broad scale vision of AMEG circulation and propagation. However, it suffers some drawbacks, mainly linked to the fact that it is based on available data and does not result from a dedicated sampling. On the one hand, our screening is likely to underestimate the frequency of resistance carrying genomes because some AMEG families probably remain unknown and AG resistance can be conferred by other resistance mechanisms such as target change, hydrolysis, etc. (Blair et al., 2015). On the other hand, resistance gene carrying genomes do not necessarily produce resistant bacteria. It is indeed nearly impossible to determine with certainty whether each of the genes identified actually confers AG resistance. So far, the genotype-phenotype relationship is poorly understood for ARGs in general (Hughes and Andersson, 2017), and particularly for AMEs: a single amino acid change is likely to change the enzyme's target, and even to remove its resistance function (Feldgarden et al., 2019; Zárate et al., 2018). Besides, some of the genes we identified might actually not be expressed at all, as the presence of cis-regulatory sequences were not screened and their functionality in the genome of residence is difficult to evaluate. However, our screening might still be considered as a decent approximation when working with a high spatio-temporal scale of genomic data.

A second important limitation is that the genomes available are not representative of the species composition of the different biomes and some of the biomes or geographical location are over-represented. This means that the frequencies of resistant bacteria established in this study are useful for comparisons between biomes, geographical location and time periods but cannot be taken as absolute estimates. Along the same line, the absence of detected resistance in certain places in our data set is likely due to a lack of published genomic data from these places: for example, we did not detect any AMEG in genomes of bacteria from the Antarctic, when metagenomic studies have shown that AG resistance have evolved in polar communities (Perron et al., 2015).

Besides, environmental data were insufficiently standardized. Regarding sampling locations, many of them were unknown in NCBI Biosamples metadata; and others could only be determined to the precision of the country, sometimes by looking at unrelated columns. We thus chose to use the country scale as spatial grain, in order to consider as many of them as possible, but at the cost of geographical precision. Moreover, because sampling can be scarce in many geographical areas, countries outside Europe were grouped in larger entities. This choice of spatial scales, constrained by the metadata available, might have prevented to uncover dynamics that occur at finer spatial scales. In the same way, we had to rely on unstandardized data regarding the categorization of the biomes in which bacteria were sampled. Our categorization of biomes was an attempt to reproduce current categorization of metagenomes, but unlike metagenomic datasets, the ecosystems in which bacteria genomes are sampled are usually poorly described: e.g. the distinction between human samples and clinical samples is often very subtle. Some samples may therefore have been assigned to the wrong

biome. And since sampling is greatly biased towards clinical, human, and farms samples, we chose to merge certain classical ecological categories in order to treat the widest diversity of samples. This categorization thus partly differs from other studies.

### 3.7 Conclusion and perspectives

The present study provides a broad picture of the spatial, temporal and ecological distributions of AMEGs as well as their association with MGEs and reveals contrasted patterns for the different gene families. It additionally establishes that the recent temporal variations of ACBs in Europe are explained first by ecology, second by human exchanges and last by antibiotic consumption. This means that selection by man-made antibiotics is not the only evolutionary force explaining the frequency of ARGs and its variation, such that interventional strategies based on prudent uses of antibiotics for humans, animals, and plants are likely to be a necessary but insufficient way to control and limit the spread of antibiotic resistance. The importance of ecology and human exchanges in shaping the patterns of AMEG prevalence is adding to the growing body of evidences that AR must also be tackled from One Health and Global Health perspectives (Hernando-Amado et al., 2020, 2019). It depends not only on clinical therapeutic guidelines, but also on the high interconnectivity of ecosystems, both locally and globally. Thus, though continuing AR monitoring in clinical and farms samples is crucial, current sampling methods hugely bias genomic datasets and insufficient standardization of data limits their exploitation. Understanding the big picture on AR will require a stronger sampling effort in natural ecosystems and as we have shown that resistomes tend to cluster by ecology rather than by geography, emphasis should be put on monitoring the resistome of all biomes with equal intensity. It is important to note that metagenomic datasets currently allow neither to precisely assign ARBs to taxonomy, nor to analyze ARG accumulation in genomes, but that their assembly into draft genomes (van der Walt et al., 2017) could help to quickly widen the range of genomic data available from natural ecosystems in the near future.

Finally, this study highlights that AMEGs are frequently associated with MGE but also shows that this level of association strongly varies between gene families. It additionally reveals the role of MGE in the generation of within genome duplications and even more importantly in functional diversification and resistance spectrum broadening. MGEs are known to be vehicles of HGT and are likely to participate in the spread of AMEGs but the strong correlation between resistome composition and species composition established here within biome groups suggests that AMEGs are also spreading by ACBs exchange between biomes. The relative contributions of HGT and ACB migration to AR propagation as well as the factors that shape and orient them should be investigated on datasets as the one generated in this study in order to understand the antibiotic resistance traffic rules and potentially interfere with them to reduce and slow down their movements.

# 4. Material and methods

## 4.1 Detection of aminoglycoside modifying enzyme genes

The first aim of this study was to document the geographic, historical and ecological extension of gene families that encode AMEs. This was done by screening for these genes across the clade of Eubacteria. Current methods to detect ARGs in genomic datasets can be, for example, based on HMM profiles drawn from biochemical nomenclatures (e.g. AMRFinder, Feldgarden et al., 2019). However, the biochemical characteristics of AMEs do not reflect their evolutionary history and many profiles may thus gather both homologous and paralogous sequences. Such mixed and evolutionary heterogeneous profiles do not allow to track the spread of a gene family as they can gather sequences with different evolutionary origins despite their similar biochemical functions. New HMM profiles based on clusters of homologous sequences only were thus built. While it may detect a wider diversity of resistance genes, it also increased the need for additional filtering steps.

### 4.1.1 Definition of resistance profiles

204 sequences of aminoglycoside modifying enzymes were gathered from examples reviewed in Ramirez and Tolmasky (2010) and Garneau-Tsodikova and Labby (2016). From the sequences contained in these two references, two categories have been excluded for further analyses: sequences for which (i) the protein could not be found from the provided identifier solely, and (ii) two enzymatic functions were assigned (e.g. AAC(6')-APH(2'')). All the other sequences (Supplementary information, Tab. S38) were grouped in 29 clusters of homologous genes (CHGs) using SiLiX (v. 1.2.9, Miele et al., 2011) with the following parameter values: at least 70% overlap and 35% identity to accept BLAST hits for building families. For each CHG, sequences were aligned with Clustal Omega (v. 1.2.4, Sievers and Higgins, 2014), and alignments were submitted to HMMER (v. 3.3, Eddy, 2011) to define profiles. Sequences used to define CHGs, their biochemical classification, and their assignment to CHGs, are listed in Supplementary information, Tab. S38: AACs were split into 14 CHGs, ANTs into 6 CHGs, and APHs into 9 CHGs, and no CHG gathered proteins with distinct enzymatic activities.

### 4.1.2 Bacterial genomes

The genomes included in this study are all the Eubacteria genomes available in the NCBI Refseq database on August 18, 2019 and that could be assigned at the species taxonomic level, to the taxonomy *bac120* (release 89, Parks et al., 2018) with the pipeline GTDB-tk (v. 1.0.2, Chaumeil et al., 2020). This pipeline extracts sequences of 120 conserved genes to assign unannotated genomes to the taxonomy, based on the sequence of these 120 genes. 160,987 genomes were used in this study.

### 4.1.3 Screening resistance profiles into genomes

The genomes were screened for the presence of the 27 CHG profiles defined using the *hmmsearch* command of HMMER (v. 3.3). Around 1,200,000 hits were found in the 160,987 genomes screened.

Results were kept with cutoff values of $10^{-3}$ for e-value and 0.95 for accuracy, and if the length of the predicted proteins was at least 80% of the one of the corresponding profile. When the same protein corresponded to two (or more) overlapping profiles, the result with the lowest e-values was kept. The functions of the proteins screened by HMMER were predicted using InterProScan (v. 5.40, Jones et al., 2014) and proteins with a predicted function incompatible with aminoglycoside modification were filtered out of the dataset based on a keyword list search (see Supplementary information, Text S2 for the list of keywords).

After screening and functional filtering, some CHG profiles did not match any output homologs (i.e. CHGs 9, 10, 21, 22 and 30) and they were thus excluded. The proteins conserved after this functional filtering step were reclustered with SiLiX (with a minimum of 40% identity over 80% overlap) in order to confirm homology. As some sub-clustered CHGs also contained a low number of hits (2 or 3), only clusters that contained at least 10 sequences were kept (<100 sequences excluded). Three CHGs had to be subdivided and the resulting subclusters will be referred to as their initial CHG name, plus another digit (e.g. CHG 6.1). This resulted in 46,053 protein sequences predicted to code for AMEs, belonging to 27 CHGs, and spread across 38,523 genomes.

## 4.2 The genomic context of aminoglycoside modifying enzyme genes

A search for mobile genetic elements in the neighborhood of predicted AMEGs was performed. The software programs and methods used are detailed thereafter. All programs mentioned below were applied to complete genomes for completely assembled genomes, or to each contig containing AMEGs from partially assembled genomes. The aim was to identify MGEs allowing intra-genomic mobility (transposable elements and integrons) and those allowing inter-genomic mobility (plasmids, prophages, and ICEs). Importantly, about one third of the identified AMEGs were located less than 1kb of the extremities of the corresponding contigs, which might have reduced our ability to detect MGEs.

### 4.2.1 Identification of genetic elements allowing inter-genomic mobility

The PlasForest pipeline (v. 1.0.0, Pradier et al., 2021) was used to identify plasmids. Full contigs containing previously identified AMEGs were submitted to this random forest classifier-based software. When a contig was predicted to be of plasmid origin, all the AMEGs it carried were considered as carried by a plasmid. Prophages were then identified with the pipeline PhiSpy (v. 4.0.0, Akhter et al., 2012). AMEGs were considered to be associated with prophages when located within a predicted prophage, or within 1kb upward or downward of it. At last, integrative and conjugative elements (ICEs, sometimes referred to as ‘conjugative transposons’) were identified with the pipeline ICEFinder (Liu et al., 2019). This pipeline can only process annotated genomes, so AMEG-carrying genomes were beforehand annotated with the software Prokka (v. 1.14.5, Seemann, 2014). ARGs were considered to be associated with ICEs when located within a predicted ICE, or within 1kb upward or downward of it.

### 4.2.2 Identification of genetic elements allowing intra-genomic mobility

There are two types of bacterial transposable elements that are likely to carry accessory genes: composite and noncomposite transposons. However, there is currently no standardized method to detect transposable elements in prokaryote genomes (see Goerner-Potvin and Bourque, 2018, with the exception of Red that is not specific to transposable elements, Girgis, 2015). Thus, in order to detect composite transposons, ISfinder database of insertion sequences (Siguier et al., 2006) was downloaded on 2020/05/05 and used as a BLASTn subject database (v. 2.11.0, Camacho et al., 2009). AMEG-carrying contigs were submitted to BLASTn against this database, and only hits larger than minimum IS size (depending on the family, from 600bp for IS200/IS605 to 3000bp for Tn3) were kept. An AMEG was considered as potentially carried by a composite transposon if it was surrounded by at least 2 IS of the same family within 20kb from each other. In order to detect noncomposite transposons, 345,657 transposase protein sequences were extracted from NCBI (list of accession numbers in Supplementary information List. S1) and used as a BLASTp subject database. Contigs carrying AMEGs were translated on 6 frames, then submitted to BLASTp against this database, where only hits larger than 50 amino acids were kept (in order to detect only hits with substantial coverage of transposases). In order to distinguish IS transposases (involved in composite transposon transposition) from other transposases, AMEGs in the neighborhood of IS (previously identified as associated to composite transposons) were excluded from this non-composite transposon screen. Remaining AMEGs were considered as potentially associated with non-composite transposons if they were located within 20kb from a transposase.

At last, integrons were identified with the pipeline IntegronFinder (v. 1.5.1, Cury et al., 2016). AMEGs were considered to be associated with integrons when located within a predicted integron, or within 1kb upward or downward of it.

## 4.3 Environmental contexts of ARG-carrying bacteria

### 4.3.1 Metadata

Through the NCBI BioSamples database, metadata were collected for 90,751 samples (about 55% of the genomes screened for AMEGs).

A classification was attempted for all the samples into the 11 ecological units (biomes) that were defined: clinical environments; freshwater; sludge and waste; wild fauna and flora; domestic animals; farms and farm animals; agrosystems; sea water; human habitat; human samples; soil. To do so, keywords were searched (Supplementary information, Text S3), in a precise order to avoid multiple allocation of the same sample. Some form of sample location was available in the metadata of 75,581 samples: either in the form of geographical coordinates (13,053 samples), location names (exact location, region, or country, 52,898 samples), or institution (university, laboratory, or hospital, 7970 samples). As most locations were not more precise than the country scale, each sampling location was

assigned to a country by reverse search in Google Geocode API. Sea water samples could not be assigned to any country, and were thus not included in analysis of spatial variation in AMEGs prevalence. Due to the large number of geographical areas to consider, countries were gathered into larger regions (IMAGE24 classification, R library *rworldmap*, South, 2011). Samples originating from European Union member countries kept their country-level assignment for analyses focusing on Europe only. Sampling dates recovered from the metadata were considered to the time level of year.
In total, geographical information, ecological information, and sampling dates could be recovered for 45,574 genomes only.

#### 4.3.2 Antibiotic consumption

Aminoglycoside consumption data in the European Union were collected from the ESAC-NET (for human consumption, European Centre for Disease Prevention and Control, 2021) and the ESVAC (for animal consumption, European Medicines Agency, 2021) databases for the period 1997-2018. Antibiotic consumption data are not freely available for countries outside the European Union. As an aim of this study was to test the impact of antibiotic consumption in general, both sources of antibiotic consumption were summed into a single variable, measured in metric tons of aminoglycosides consumed each year by each country. ESVAC consumption data were measured in metric tons, while ECDC consumption data were reported in Defined Daily Doses (DDDs) for 1000 habitants. ECDC consumption data were thus converted to the same unit using 1 gram as a baseline for aminoglycoside DDD, which is the current WHO standard for amikacin (WHO Collaborating Centre for Drug Statistics Methodology, 2020), and each country's population, recovered from the World Bank (2021a). In order to standardize consumption data between countries of heterogeneous sizes, these measures were divided by each country's surface. Then, missing human and animal consumption data on the period 1997-2018 were predicted by interpolating the actual data between two consecutive available data points (linear regression: Antibiotic consumption in metric tons per $1000km^2$ ~ Country x Year, $R^2$=0.959 for ESVAC data, $R^2$=0.928 for ECDC data). Finally, these data were summed to a single variable of AG consumption per country and per year.

#### 4.3.3 Human exchanges

The bilateral trade matrix, including imports between 1997 and 2018, was collected from the United Nations Conference on Trade and Development (United Nations Conference for Trade and Development, 2021). This matrix measured imports in 2018 US dollars. From this matrix, imports of energy, information, and services were excluded, to only keep 27 import categories (categories are listed in the rows of Supplementary information, Tab. S19). The bilateral immigration matrix was collected from the World Bank (2021b). Missing immigration data were predicted by interpolating the actual data (linear regression: Number of migrants ~ Country of origin x Country of destination x

Year, $R^2$=0.883).

Potential AMEG influx from one country to another was calculated as the sum of imports (respectively immigrants) from the origin region, multiplied by the frequency of AMEGs sampled in the origin region. Total AMEG influx in a country due to trade (respectively immigration) was computed by summing influxes from all possible origins.

## 4.4 Inference of resistance spectrum

The aim was to assign a resistance spectrum to each of the AMEG we identified. AR phenotypes in particular can be either determined by empirically testing the resistance phenotype associated to a given genotype as has been done in the AMRFinder database (Feldgarden et al., 2019), or inferred from the systematic compilation of all the antibiotics to which a biochemical family is known to resist (Zárate et al., 2018). Here both datasets were used to infer a minimum and maximum resistance spectrum of the AMEGs detected in Eubacteria genomes. An AME biochemical class was attributed to each AMEG sequence identified thanks to the AMRFinder database. A resistance spectrum was then attributed to each AMEG sequence using the established correspondence between the biochemical classes and the AR phenotypes, either empirically predicted (Feldgarden et al., 2019), or reviewed (Zárate et al., 2018). Even if an assignment of each AMEG sequence to an AMRFinder profile could be done, an assignment of each of them to a resistance spectrum was not possible, either because no precise resistance was indicated in Feldgarden et al. (2019), or because the biochemical classes corresponding to AMRFinder profiles were too broad compared to the ones from Zárate et al. (2018) and thus did not allow correct assignment. An AR phenotype could be assigned to 12,982 AMEGs out of 11,284 genomes using empirically tested data, and to 31,910 AMEGs out of 26,006 genomes using reviewed data.

## 4.5 Statistical analysis

### 4.5.1 Phylogenetic diversity

To evaluate the phylogenetic range across which each CHG is distributed, two measures summing up the phylogenetic distribution were computed using the function *pd* from the R library *picante*: (i) species richness, i.e. the number of species in which a CHG was found; and (ii) Faith's distance (Faith, 1992), i.e. the sum of the lengths of all the branches of the minimum tree spanning all the bacteria carrying a specific CHG.

### 4.5.2 Factors influencing the distribution of aminoglycoside resistance

For each CHG, the aim was to explain the local frequencies of AG-resistant bacteria, i.e. the likelihood to sample a genome carrying at least one gene from a given CHG in a given ecosystem at a given date, during the 1997-2018 period. The function *fitme* of R library spaMM (Rousset and Ferdy, 2014) was used to compute logistic regressions. The random structure of each null model was selected to minimize conditional AIC, among these hypotheses: no structure; a first-order time-autoregressive

process; a Matérn correlation structure based on longitude and latitude of each country's centroid; both a time-autoregressive process and a spatial Matérn correlation structure. Several hypotheses were also considered for the explanatory variables. First, AMEG influx due to trade and AMEG influx due to immigration were highly correlated with each other. In the subsequent statistical analysis, these two variables were thus not considered separately: either both of them were considered to explain the distribution of antibiotic resistance, or none of them were. These two variables were thus considered under the denomination of human exchanges. Then, the following variables were included as fixed effects: ecology (as a categorical variable), AG consumption, human exchanges, and their interactions with ecology (as scaled numerical variables). Including the null model, thirteen different combinations of fixed effects were thus simultaneously considered to explain the distribution of a CHG. For each CHG, the model with the lowest conditional AIC was selected. As antibiotic consumption data were only available for Europe, this model selection was only performed for European data. Due to the temporal variability of sampling in Europe, dates were considered every two years. Another model selection was performed, without considering the antibiotic consumption variable, for worldwide data (with countries being gathered into larger geographical units, see section 4.3.1) with dates being considered every year.

To describe the amount of variance explained by the models and the improvement of the complete models from the null models, adjusted McFadden pseudo-$R^2$ were computed:

$$R^2_{Adj} = 1 - \frac{\log(L) - K}{\log(L_0)}$$

where $L0$ is the likelihood of the null model, $L$ is the likelihood of the considered model, and $K$ is the number of explanatory variables in the model. To measure the unique contribution of each variable, the difference $\Delta R^2_{Adj}$ was computed, i.e. the difference of adjusted McFadden pseudo-$R^2$ value measured in the same model without the focal variable.

For 14 CHGs, model selection kept human exchanges as one of the explanatory variables. For these cases, the effect of the 28 components of the "human exchanges" variable was tested. All these components being highly correlated with each other, "human exchanges" was replaced by only one component at a time. Among the 28 resulting models, only those for which the slope estimates for human exchanges were positive were kept. Finally, models with $AIC < AIC_{min} + 10$ were selected (Burnham and Anderson, 2002).

### 4.5.3 Distribution of resistomes across geography and ecology

In our dataset, an ecosystem was defined as the intersection of a geographical region and a biome. An aminoglycoside resistome (hereafter simplified as "resistome") was defined as the set of AMEGs that could be found in an ecosystem. Several CHGs appeared during the time frame considered (1990-2018). Thus, in order for resistomes not to depend on time, CHG presences and absences over several

time periods were separately considered (from 1990 to 1994, 1995 to 1999, etc.). For example, CHG 1 presence/absence between 1990 and 1994 and its presence/absence between 1995 and 1999 are considered as different variables for resistome composition comparisons, as if they were two independent CHGs. Similarities between resistomes of different ecosystems were computed using Jaccard index, because it is one of the least sensitive to sampling error (see Schroeder and Jenkins, 2018). This index takes a value of 1 when two ecosystems carry exactly the same CHGs, a value of 0 when two ecosystems carry no common CHG.

To test the dependence of resistomes on geography and ecology, distances were considered between ecosystems (i) either as geographical distances between the centroids of considered regions, or (ii) as binary distances between considered biomes (measured as 1 if two ecosystems belong to the same biome, 0 otherwise). Mantel tests were performed to measure a possible correlation between Jaccard index on one side and the geographical or the biome distance on the other side. Correlation with ecological binary distance was found but not with geographical distance, so ecosystems were then grouped into biomes for further analyses.

To test the association strength between biomes, a network based on the previous matrix of Jaccard indices was created. Each biome was represented as a vertex, and each edge was weighed by the resemblance between two resistomes (measured by Jaccard index). This complete network was pruned into a minimum spanning network (Bandelt et al., 1999), in order to remove the edges that were the least likely to correspond to ARG exchange between two resistomes. Modules were established on this network by using the Louvain method (Blondel et al., 2008). Finally, the relationship between AMEG composition and bacteria community composition was studied at the ecosystem level (each ecosystem being defined as the intersection of a biome and a geographic unit). To do so, the correlation between the resistome Jaccard index and the similarity in bacteria community composition, represented by a phylogenetic Sørensen index (Bryant et al., 2008), was computed.

#### 4.5.4 Functional diversity

*Functional dispersion* (FDis) et *Rao's quadratic entropy* (RaoQE) were computed to characterize functional dissimilarity, in terms of resistance spectrum conferred by different AMEGs contained in the same genome. For this, the R library FD was used (Laliberté et al., 2014). To measure how functional dissimilarity changes with AMEG number per genome, a linear regression of either the square root of FDis or the square root of RaoQ and the logarithm of the number of AMEGs carried by a genome was computed. Permutation tests were performed in order to compare the observed pattern with expectations under the null hypothesis of random assortment of AMEGs among genomes. The gene-genome matrix was permuted 500 times using the function *permatswap* from the R library *vegan*, and FDis and RaoQ were calculated on each genome for each iteration. For each permutation, linear regressions were computed, which yielded a distribution of slopes, simulated under the null hypothesis of random assortment of AMEGs among genomes. Observed slopes were then compared to this distribution, to assess whether AMEG accumulation is compatible with this hypothesis. This

analysis was performed for whole genomes, and then distinguishing between AMEGs associated or not associated with MGEs and then further distinguishing between MGEs with intragenomic and intergenomic mobility.

#### 4.5.5 Gene duplications

For each genome that contains at least two AMEGs belonging to the same CHG, the DNA sequences of these genes were recovered, as well as 1kb upstream and downstream, from the NCBI Nucleotide database. Pairwise alignment of these sequences was performed, in order to measure sequence identity between pairs of AMEGs. Because many gene sequences are identical or nearly identical, sequence identity was also computed between pairs of sequences composed of genes and their close genomic contexts. As a conservative threshold, a group of sequences (e.g. a pair, a triplet, or a quadruplet of sequences) was considered to be the result of gene duplication if AMEGs were at least reciprocally 90% identical and their genomic contexts were at least reciprocally 80% identical. Yet, to test the sensitivity of the subsequent analysis to these arbitrary thresholds, other thresholds were used for the identity between AMEGs (80%, 85%, 90% and 95%) and for the identity between genomic contexts (50%, 70%, 80% and 90%). Results did not substantially vary with these thresholds (Supplementary information, Tab. S39-S40). To test if the association of an AMEG sequence with (at least) one MGE makes it more likely to duplicate, the proportions of duplicated groups of AMEGs were computed both among genes associated with MGEs and among genes that were not. Permutations of the AMEGs-MGEs matrix were performed to compute replicates of these proportions under the hypothesis of random physical association between MGEs and AMEGs. The observed odds ratio was then compared to the distribution obtained by permutation.

## Acknowledgments

We thank Christelle Leung and Tazzio Tissot for their helpful comments that helped improve the manuscript, and Alexandre Courtiol and Martijn Callens for stimulating scientific and technical discussions. This work was supported by the ERC HGTCODONUSE (ERC-2015-CoG-682819) to SB.

## Competing interests

The authors declare that they have no competing interests.

## References


Akhter S, Aziz RK, Edwards RA. 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res* **40**:e126–e126. doi:10.1093/NAR/GKS406

Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J, Handelsman J. 2010. Call of the wild: antibiotic resistance genes in natural environments. *Nat Rev Microbiol 2010 84* **8**:251–259. doi:10.1038/nrmicro2312

Bandelt H-J, Forster P, Röhl A. 1999. Median-Joining Networks for Inferring Intraspecific Phylogenies. *Mol Biol Evol* **16**:37–48.

Baquero F, Alvarez-Ortega C, Martinez JL. 2009. Ecology and evolution of antibiotic resistance. *Environ Microbiol Rep* **1**:469–476. doi:10.1111/J.1758-2229.2009.00053.X

Baquero F, Lanza VF, Cantón R, Coque TM. 2015. Public health evolutionary biology of antimicrobial resistance: priorities for intervention. *Evol Appl* **8**:223–239. doi:10.1111/EVA.12235

Bell BG, Schellevis F, Stobberingh E, Goossens H, Pringle M. 2014. A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance. *BMC Infect Dis* **14**:1–25. doi:10.1186/1471-2334-14-13/TABLES/2

Bengtsson-Palme J, Kristiansson E, Larsson DGJ. 2018. Environmental factors influencing the development and spread of antibiotic resistance. *FEMS Microbiol Rev* **42**:68–80. doi:10.1093/femsre/fux053

Berendonk TU, Manaia CM, Merlin C, Fatta-Kassinos D, Cytryn E, Walsh F, Bürgmann H, Sørum H, Norström M, Pons MN, Kreuzinger N, Huovinen P, Stefani S, Schwartz T, Kisand V, Baquero F, Martinez JL. 2015. Tackling antibiotic resistance: the environmental framework. *Nat Rev Microbiol 2015 135* **13**:310–317. doi:10.1038/nrmicro3439

Berndtson AE. 2020. Increasing Globalization and the Movement of Antimicrobial Resistance between Countries. *Surg Infect (Larchmt)* **21**:579–585. doi:10.1089/SUR.2020.145

Blair JMA, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJV. 2015. Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol*. doi:10.1038/nrmicro3380

Blanquart F, Lehtinen S, Fraser C. 2017. An evolutionary model to predict the frequency of antibiotic resistance under seasonal antibiotic use, and an application to Streptococcus pneumoniae. *Proc R Soc B Biol Sci* **284**:20170679. doi:10.1098/rspb.2017.0679

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. *J Stat Mech* **2008**:P10008. doi:10.1088/1742-5468/2008/10/P10008

Bryant JA, Lamanna C, Lè Ne Morlon H, Kerkhoff AJ, Enquist BJ, Green JL. 2008. Microbes on mountainsides: Contrasting elevational patterns of bacterial and plant diversity. *PNAS August* **12**:11505–11511.

Buchholz U, Bernard H, Werber D, Böhmer MM, Remschmidt C, Wilking H, Deleré Y, an der Heiden M, Adlhoch C, Dreesman J, Ehlers J, Ethelberg S, Faber M, Frank C, Fricke G, Greiner M, Höhle M, Ivarsson S, Jark U, Kirchner M, Koch J, Krause G, Luber P, Rosner B, Stark K,

Kühne M. 2011. German Outbreak of Escherichia coli O104:H4 Associated with Sprouts. *http://dx.doi.org/101056/NEJMoa1106482* **365**:1763–1770. doi:10.1056/NEJMOA1106482

Cairns J, Ruokolainen L, Hultman J, Tamminen M, Virta M, Hiltunen T. 2018. Ecology determines how low antibiotic concentration impacts community composition and horizontal transfer of resistance genes. *Commun Biol 2018 11* **1**:1–8. doi:10.1038/s42003-018-0041-7

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**:1–9. doi:10.1186/1471-2105-10-421

Capita R, Alonso-Calleja C. 2013. Antibiotic-Resistant Bacteria: A Challenge for the Food Industry. *Crit Rev Food Sci Nutr* **53**:11–48. doi:10.1080/10408398.2010.519837

Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**:1925–1927. doi:10.1093/BIOINFORMATICS/BTZ848

Corel E, Méheust R, Watson AK, Mcinerney JO, Lopez P, Bapteste E. 2018. Bipartite Network Analysis of Gene Sharings in the Microbial World. *Mol Biol Evol* **35**:899–913. doi:10.1093/MOLBEV/MSY001

Cuadrat RRC, Sorokina M, Andrade BG, Goris T, Dávila AMR. 2020. Global ocean resistome revealed: Exploring antibiotic resistance gene abundance and distribution in TARA Oceans samples. *Gigascience* **9**:1–12. doi:10.1093/GIGASCIENCE/GIAA046

Cury J, Jové T, Touchon M, Néron B, Rocha EP. 2016. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res* **44**:4539–4550. doi:10.1093/NAR/GKW319

Czekalski N, Sigdel R, Birtel J, Matthews B, Bürgmann H. 2015. Does human activity impact the natural antibiotic resistance background? Abundance of antibiotic resistance genes in 21 Swiss lakes. *Environ Int* **81**:45–55. doi:10.1016/J.ENVINT.2015.04.005

D'Costa VM, King CE, Kalan L, Morar M, Sung WWL, Schwarz C, Froese D, Zazula G, Calmels F, Debruyne R, Golding GB, Poinar HN, Wright GD. 2011. Antibiotic resistance is ancient. *Nature*. doi:10.1038/nature10388

da Cunha BR, Fonseca LP, Calado CRC. 2019. Antibiotic Discovery: Where Have We Come from, Where Do We Go? *Antibiot 2019, Vol 8, Page 45* **8**:45. doi:10.3390/ANTIBIOTICS8020045

Danner MC, Robertson A, Behrends V, Reiss J. 2019. Antibiotic pollution in surface fresh waters: Occurrence and effects. *Sci Total Environ* **664**:793–804. doi:10.1016/J.SCITOTENV.2019.01.406

Davies JE. 1983. Resistance to Aminoglycosides: Mechanisms and Frequency. *Rev Infect Dis* **5**:S261–S267. doi:10.1093/CLINIDS/5.SUPPLEMENT_2.S261

de Kraker MEA, Stewardson AJ, Harbarth S. 2016. Will 10 Million People Die a Year due to Antimicrobial Resistance by 2050? *PLOS Med* **13**:e1002184. doi:10.1371/JOURNAL.PMED.1002184

Depardieu F, Podglajen I, Leclercq R, Collatz E, Courvalin P. 2007. Modes and Modulations of Antibiotic Resistance Gene Expression. *Clin Microbiol Rev* **20**:79. doi:10.1128/CMR.00015-06

Durand GA, Raoult D, Dubourg G. 2019. Antibiotic discovery: history, methods and perspectives. *Int*

*J Antimicrob Agents* **53**:371–382. doi:10.1016/J.IJANTIMICAG.2018.11.010

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLOS Comput Biol* **7**:e1002195. doi:10.1371/JOURNAL.PCBI.1002195

Eltai NO, Yassine HM, El-Obeid T, Al-Hadidi SH, Thani AA Al, Alali WQ. 2020. Prevalence of Antibiotic-Resistant Escherichia coli Isolates from Local and Imported Retail Chicken Carcasses. *J Food Prot* **83**:2200–2208. doi:10.4315/JFP-20-113

Enne VI, Bennett PM, Livermore DM, Hall LMC. 2004. Enhancement of host fitness by the sul2-coding plasmid p9123 in the absence of selective pressure. *J Antimicrob Chemother* **53**:958–963. doi:10.1093/JAC/DKH217

European Centre for Disease Prevention and Control. 2021. Antimicrobial consumption database (ESAC-Net). https://www.ecdc.europa.eu/en/antimicrobial-consumption/surveillance-and-disease-data/database

European Medicines Agency. 2021. European database of sales of veterinary antimicrobial agents. https://esvacbi.ema.europa.eu/analytics/saw.dll?PortalPages

European Medicines Agency. 2017. Sales of veterinary antimicrobial agents in 31 European countries in 2017.

Faith DP. 1992. Conservation evaluation and phylogenetic diversity. *Biol Conserv* **61**:1–10. doi:10.1016/0006-3207(92)91201-3

Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu C-H, Mcdermott PF, Tadesse DA, Morales C, Simmons M, Tillman G, Wasilenko J, Folster JP, Klimke W. 2019. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob Agents Chemother* **63**:e00483--19. doi:10.1128/AAC.00483-19

Finken M, Kirschner P, Meier A, Wrede A, Bottger' EC. 1993. Molecular basis of streptomycin resistance in Mycobacterium tuberculosis: alterations of the ribosomal protein S12 gene and point mutations within a functional 16S ribosomal RNA pseudoknot. *Mol Microbiol* **9**:1239–1246.

Forsberg KJ, Patel S, Gibson MK, Lauber CL, Knight R, Fierer N, Dantas G. 2014. Bacterial phylogeny structures soil resistomes across habitats. *Nat 2014 5097502* **509**:612–616. doi:10.1038/nature13377

Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MOA, Dantas G. 2012. The shared antibiotic resistome of soil bacteria and human pathogens. *Science (80- )* **337**:1107–1111. doi:10.1126/SCIENCE.1220761

Founou LL, Founou RC, Essack SY. 2016. Antibiotic resistance in the food chain: A developing country-perspective. *Front Microbiol* **7**:1881. doi:10.3389/FMICB.2016.01881/BIBTEX

Garneau-Tsodikova S, Labby KJ. 2016. Mechanisms of Resistance to Aminoglycoside Antibiotics: Overview and Perspectives. *Medchemcomm* **7**:11. doi:10.1039/C5MD00344J

Gibson MK, Forsberg KJ, Dantas G. 2014. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J 2015 91* **9**:207–216. doi:10.1038/ismej.2014.106

Girgis HZ. 2015. Red: An intelligent, rapid, accurate tool for detecting repeats de-novo on the

genomic scale. *BMC Bioinformatics* **16**:1–19. doi:10.1186/S12859-015-0654-5/TABLES/4

Goerner-Potvin P, Bourque G. 2018. Computational tools to unmask transposable elements. *Nat Rev Genet 2018 1911* **19**:688–704. doi:10.1038/s41576-018-0050-x

Goossens H, Ferech M, Stichele R Vander, Elseviers M. 2005. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet* **365**:579–587. doi:10.1016/S0140-6736(05)17907-0

Gothwal R, Shashidhar T. 2015. Antibiotic Pollution in the Environment: A Review. *CLEAN – Soil, Air, Water* **43**:479–489. doi:10.1002/CLEN.201300989

Gullberg E, Cao S, Berg OG, Ilbäck C, Sandegren L, Hughes D, Andersson DI. 2011. Selection of Resistant Bacteria at Very Low Antibiotic Concentrations. *PLOS Pathog* **7**:e1002158. doi:10.1371/JOURNAL.PPAT.1002158

Hernando-Amado S, Coque TM, Baquero F, Martínez JL. 2020. Antibiotic Resistance: Moving From Individual Health Norms to Social Norms in One Health and Global Health. *Front Microbiol* **11**:1914. doi:10.3389/FMICB.2020.01914/BIBTEX

Hernando-Amado S, Coque TM, Baquero F, Martínez JL. 2019. Defining and combating antibiotic resistance from One Health and Global Health perspectives. *Nat Microbiol* **4**:1432–1442. doi:10.1038/s41564-019-0503-9

Holmes AH, Moore LSP, Sundsfjord A, Steinbakk M, Regmi S, Karkey A, Guerin PJ, Piddock LJ V. 2016. Understanding the mechanisms and drivers of antimicrobial resistance. *Lancet* **387**:176–187. doi:10.1016/S0140-6736(15)00473-0

Jana S, Deb JK. 2006. Molecular understanding of aminoglycoside action and resistance. *Appl Microbiol Biotechnol 2006 702* **70**:140–150. doi:10.1007/S00253-005-0279-0

Jiang Z, Paudyal N, Xu Y, Deng T, Li F, Pan H, Peng X, He Q, Yue M. 2019. Antibiotic Resistance Profiles of Salmonella Recovered From Finishing Pigs and Slaughter Facilities in Henan, China. *Front Microbiol* **0**:1513. doi:10.3389/FMICB.2019.01513

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236–1240. doi:10.1093/BIOINFORMATICS/BTU031

Ju F, Beck K, Yin X, Maccagnan A, McArdell CS, Singer HP, Johnson DR, Zhang T, Bürgmann H. 2019. Wastewater treatment plant resistomes are shaped by bacterial composition, genetic exchange, and upregulated expression in the effluent microbiomes. *ISME J* **13**:346. doi:10.1038/S41396-018-0277-8

Karkman A, Johnson TA, Lyra C, Stedtfeld RD, Tamminen M, Tiedje JM, Virta M. 2016. High-throughput quantification of antibiotic resistance genes from an urban wastewater treatment plant. *FEMS Microbiol Ecol* **92**:14. doi:10.1093/FEMSEC/FIW014

Karkman A, Pärnänen K, Larsson DGJ. 2019. Fecal pollution can explain antibiotic resistance gene abundances in anthropogenically impacted environments. *Nat Commun 2019 101* **10**:1–8. doi:10.1038/s41467-018-07992-3

Kümmerer K. 2003. Promoting resistance by the emission of antibiotics from hospitals and households into effluent. *Clin Microbiol Infect* **9**:1203–1214. doi:10.1111/J.1469-0691.2003.00739.X

Labby KJ, Garneau-Tsodikova S. 2013. Strategies to overcome the action of aminoglycoside-modifying enzymes for treating resistant bacterial infections. *Future Med Chem* **5**:1285–1309. doi:10.4155/FMC.13.80

Laliberté E, Legendre P, Maintainer BS. 2014. Package “FD” Type Package Title Measuring functional diversity (FD) from multiple traits, and other tools for functional ecology.

Larsson DGJ, Andremont A, Bengtsson-Palme J, Brandt KK, de Roda Husman AM, Fagerstedt P, Fick J, Flach CF, Gaze WH, Kuroda M, Kvint K, Laxminarayan R, Manaia CM, Nielsen KM, Plant L, Ploy MC, Segovia C, Simonet P, Smalla K, Snape J, Topp E, van Hengel AJ, Verner-Jeffreys DW, Virta MPJ, Wellington EM, Wernersson AS. 2018. Critical knowledge gaps and research needs related to the environmental dimensions of antibiotic resistance. *Environ Int* **117**:132–138. doi:10.1016/J.ENVINT.2018.04.041

Lavilla Lerma L, Benomar N, Sánchez Valenzuela A, Casado Muñoz M del C, Gálvez A, Abriouel H. 2014. Role of EfrAB efflux pump in biocide tolerance and antibiotic resistance of Enterococcus faecalis and Enterococcus faecium isolated from traditional fermented foods and the effect of EDTA as EfrAB inhibitor. *Food Microbiol* **44**:249–257. doi:10.1016/J.FM.2014.06.009

Le Hello S, Hendriksen RS, Doublet B, Fisher I, Nielsen EM, Whichard JM, Bouchrif B, Fashae K, Granier SA, Jourdan-Da Silva N, Cloeckaert A, Threlfall EJ, Angulo FJ, Aarestrup FM, Wain J, Weill F-X. 2011. International Spread of an Epidemic Population of Salmonella enterica Serotype Kentucky ST198 Resistant to Ciprofloxacin. *J Infect Dis* **204**:675–684. doi:10.1093/INFDIS/JIR409

Lepelletier D, Andremont A, Grandbastien B, Group the NW. 2011. Risk of Highly Resistant Bacteria Importation from Repatriates and Travelers Hospitalized in Foreign Countries: About the French Recommendations to Limit Their Spread. *J Travel Med* **18**:344–351. doi:10.1111/J.1708-8305.2011.00547.X

Li J, Zhang Haibo, Chen Y, Luo Y, Zhang Hua. 2016. Sources identification of antibiotic pollution combining land use information and multivariate statistics. *Environ Monit Assess 2016 1887* **188**:1–12. doi:10.1007/S10661-016-5439-4

Liu M, Li X, Xie Y, Bi D, Sun J, Li J, Tai C, Deng Z, Ou HY. 2019. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res* **47**:D660–D665. doi:10.1093/NAR/GKY1123

Livermore DM, Blaser M, Carrs O, Cassell G, Fishman N, Guidos R, Levy S, Powers J, Norrby R, Tillotson G, Davies R, Projan S, Dawson M, Monnet D, Keogh-Brown M, Hand K, Garner S, Findlay D, Morel C, Wise R, Bax R, Burke F, Chopra I, Czaplewski L, Finch R, Livermore D, Piddock LJ V., White T. 2011. Discovery research: the scientific challenge of finding new antibiotics. *J Antimicrob Chemother* **66**:1941–1944. doi:10.1093/JAC/DKR262

Lu HP, Lai YC, Huang SW, Chen HC, Hsieh CH, Yu HT. 2014. Spatial heterogeneity of gut microbiota reveals multiple bacterial communities with distinct characteristics. *Sci Reports 2014 41* **4**:1–8. doi:10.1038/srep06185

Marti E, Variatza E, Luis Balcazar J. 2014. The role of aquatic ecosystems as reservoirs of antibiotic resistance. *Trends Microbiol* **22**:36–41. doi:10.1016/j.tim.2013.11.001

Martinez JL. 2012. Natural Antibiotic Resistance and Contamination by Antibiotic Resistance Determinants: The Two Ages in the Evolution of Resistance to Antimicrobials. *Front Microbiol* **0**:1. doi:10.3389/FMICB.2012.00001

Martínez JL, Coque TM, Baquero F. 2015. Prioritizing risks of antibiotic resistance genes in all metagenomes. *Nat Rev Microbiol* **13**:396–396. doi:10.1038/nrmicro3399-c2

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinforma 2011 121* **12**:1–9. doi:10.1186/1471-2105-12-116

Mingeot-Leclercq M-P, Glupczynski Y, Tulkens PM. 1999. Aminoglycosides: Activity and Resistance. *Antimicrob Agents Chemother* **43**:727.

Mölstad S, Erntell M, Hanberger H, Melander E, Norman C, Skoog G, Lundborg CS, Söderström A, Torell E, Cars O. 2008. Sustained reduction of antibiotic use and low bacterial resistance: 10-year follow-up of the Swedish Strama programme. *Lancet Infect Dis* **8**:125–132. doi:10.1016/S1473-3099(08)70017-3

Munck C, Albertsen M, Telke A, Ellabaan M, Nielsen PH, Sommer MOA. 2015. Limited dissemination of the wastewater treatment plant core resistome. *Nat Commun 2015 61* **6**:1–10. doi:10.1038/ncomms9452

Murray BE, Murray B. 1991. New Aspects of Antimicrobial Resistance and the Resulting Therapeutic Dilemmas. *J Infect Dis* **163**:1185–1194.

Nejidat A, Diaz-Reck D, Gelfand I, Zaady E. 2021. Persistence and spread of tetracycline resistance genes and microbial community variations in the soil of animal corrals in a semi-arid planted forest. *FEMS Microbiol Ecol* **97**. doi:10.1093/FEMSEC/FIAB106

Nellums LB, Thompson H, Holmes A, Castro-Sánchez E, Otter JA, Norredam M, Friedland JS, Hargreaves S. 2018. Antimicrobial resistance among migrants in Europe: a systematic review and meta-analysis. *Lancet Infect Dis* **18**:796–811. doi:10.1016/S1473-3099(18)30219-6

O'Neill J. 2014. Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations. *London Rev Antimicrob Resist*.

Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ. 2015. Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. *BMC Genomics 2015 161* **16**:1–14. doi:10.1186/S12864-015-2153-5

Palmer AC, Chait R, Kishony R. 2018. Nonoptimal Gene Expression Creates Latent Potential for Antibiotic Resistance. *Mol Biol Evol* **35**:2669–2684. doi:10.1093/MOLBEV/MSY163

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol 2018 3610* **36**:996–1004. doi:10.1038/nbt.4229

Pehrsson EC, Tsukayama P, Patel S, Mejía-Bautista M, Sosa-Soto G, Navarrete KM, Calderon M, Cabrera L, Hoyos-Arango W, Bertoli MT, Berg DE, Gilman RH, Dantas G. 2016. Interconnected microbiomes and resistomes in low-income human habitats. *Nat 2016 5337602* **533**:212–216. doi:10.1038/nature17672

Perron GG, Whyte L, Turnbaugh PJ, Goordial J, Hanage WP, Dantas G, Desai MM. 2015. Functional Characterization of Bacteria Isolated from Ancient Arctic Soil Exposes Diverse Resistance Mechanisms to Modern Antibiotics. *PLoS One* **10**:e0069533. doi:10.1371/JOURNAL.PONE.0069533

Pitta DW, Dou Z, Kumar S, Indugu N, Toth JD, Vecchiarelli B, Bhukya B. 2016. Metagenomic Evidence of the Prevalence and Distribution Patterns of Antimicrobial Resistance Genes in Dairy Agroecosystems. *Foodborne Pathog Dis* **13**:296–302. doi:10.1089/fpd.2015.2092

Podolsky SH. 2018. The evolving response to antibiotic resistance (1945–2018). *Palgrave Commun 2018 41* **4**:1–8. doi:10.1057/s41599-018-0181-x

Pradier L, Tissot T, Fiston-Lavier A-S, Bedhomme S. 2021. PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. *BMC Bioinforma 2021 221* **22**:1–17. doi:10.1186/S12859-021-04270-W

Ramirez MS, Tolmasky ME. 2010. Aminoglycoside modifying enzymes. *Drug Resist Updat* **13**:151–171. doi:10.1016/J.DRUP.2010.08.003

Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin C-S, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-Møller J, Struve C, Petersen AM, Krogfelt KA, Nataro JP, Schadt EE, Waldor MK. 2011. Origins of the E. coli Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany. *https://doi.org/101056/NEJMoa1106920* **365**:709–717. doi:10.1056/NEJMOA1106920

Rousset F, Ferdy J-B. 2014. Testing environmental and genetic effects in the presence of spatial autocorrelation. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography (Cop)* **37**:781–790. doi:10.1111/ecog.00566ï

Sabuncu E, David J, Bernède-Bauduin C, Pépin S, Leroy M, Boëlle PY, Watier L, Guillemot D. 2009. Significant Reduction of Antibiotic Use in the Community after a Nationwide Campaign in France, 2002–2007. *PLOS Med* **6**:e1000084. doi:10.1371/JOURNAL.PMED.1000084

Salipante SJ, Hall BG. 2003. Determining the limits of the evolutionary potential of an antibiotic resistance gene. *Mol Biol Evol* **20**:653–659. doi:10.1093/molbev/msg074

San Millan A, Maclean RC. 2017. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. doi:10.1128/microbiolspec.MTBP-0016-2017

Sandegren L, Andersson DI. 2009. Bacterial gene amplification: implications for the evolution of antibiotic resistance. *Nat Rev Microbiol 2009 78* **7**:578–588. doi:10.1038/nrmicro2174

Sander P, Meier A, Böttger E. 1995. rpsL+: a dominant selectable marker for gene replacement in mycobacteria. *Mol Microbiol* **16**:991–1000. doi:10.1111/J.1365-2958.1995.TB02324.X

Schrijver R, Stijntjes M, Rodríguez-Baño J, Tacconelli E, Babu Rajendran N, Voss A. 2018. Review of antimicrobial resistance surveillance programmes in livestock and meat in EU with focus on humans. *Clin Microbiol Infect* **24**:577–590. doi:10.1016/J.CMI.2017.09.013

Schroeder PJ, Jenkins DG. 2018. How robust are popular beta diversity indices to sampling error? doi:10.1002/ecs2.2100

Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**:2068–2069. doi:10.1093/bioinformatics/btu153

Sievers F, Higgins DG. 2014. Clustal Omega. *Curr Protoc Bioinforma* **48**:3.13.1-3.13.16. doi:10.1002/0471250953.BI0313S48

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* **34**:D32–D36. doi:10.1093/NAR/GKJ014

South A. 2011. rworldmap: A New R package for Mapping Global Data.

Spicknall IH, Foxman B, Marrs CF, Eisenberg JNS. 2013. A Modeling Framework for the Evolution and Spread of Antibiotic Resistance: Literature Review and Model Categorization. *Am J*

*Epidemiol* **178**:508–520. doi:10.1093/AJE/KWT017

Tello A, Austin B, Telfer TC. 2012. Selective pressure of antibiotic pollution on bacteria of importance to public health. *Environ Health Perspect* **120**:1100–1106. doi:10.1289/EHP.1104650

Toivonen JM, Boocock MR, Jacobs HT. 1999. Modelling in Escherichia coli of mutations in mitoribosomal protein S12: novel mutant phenotypes of rpsL. *Mol Microbiol* **31**:1735–1746. doi:10.1046/J.1365-2958.1999.01307.X

Tripathi V, Cytryn E. 2017. Impact of anthropogenic activities on the dissemination of antibiotic resistance across ecological boundaries. *Essays Biochem* **61**:11–21. doi:10.1042/EBC20160054

Umweltbundesamt. 2016. Pharmaceuticals in the environment: Global occurrence and potential cooperative action under the Strategic Approach to International Chemicals Management (SAICM).

United Nations Conference for Trade and Development. 2021. UNCTADstat. https://unctadstat.unctad.org/

Van Boeckel TP, Pires J, Silvester R, Zhao C, Song J, Criscuolo NG, Gilbert M, Bonhoeffer S, Laxminarayan R. 2019. Global trends in antimicrobial resistance in animals in low- And middle-income countries. *Science (80- )* **365**. doi:10.1126/SCIENCE.AAW1944/SUPPL_FILE/AAW1944-VAN-BOECKEL-SM.PDF

van der Walt AJ, van Goethem MW, Ramond J-B, Makhalanyane TP, Reva O, Cowan DA. 2017. Assembling metagenomes, one community at a time. *BMC Genomics 2017 181* **18**:1–13. doi:10.1186/S12864-017-3918-9

van Hoek AHAM, Mevius D, Guerra B, Mullany P, Roberts AP, Aarts HJM. 2011. Acquired Antibiotic Resistance Genes: An Overview. *Front Microbiol* **0**:203. doi:10.3389/FMICB.2011.00203

Vaz-Moreira I, Nunes OC, Manaia CM. 2014. Bacterial diversity and antibiotic resistance in water habitats: searching the links with the human microbiome. *FEMS Microbiol Rev* **38**:761–778. doi:10.1111/1574-6976.12062

Watanabe T. 1963. INFECTIVE HEREDITY OF MULTIPLE DRUG RESISTANCE IN BACTERIA. *Bacteriol Rev* **27**:87–115.

Wein T, Hülter NF, Mizrahi I, Dagan T. 2019. Emergence of plasmid stability under non-selective conditions maintains antibiotic resistance. *Nat Commun 2019 101* **10**:1–13. doi:10.1038/s41467-019-10600-7

WHO Collaborating Centre for Drug Statistics Methodology. 2020. WHOCC - ATC/DDD Index. https://www.whocc.no/atc_ddd_index/?code=J01GB06

Witzany C, Bonhoeffer S, Rolff J. 2020. Is antimicrobial resistance evolution accelerating? *PLoS Pathog* **16**. doi:10.1371/JOURNAL.PPAT.1008905

World Bank. 2021a. World Bank Open Data.

World Bank. 2021b. Global Bilateral Migration Database.

World Health Organization. 2015. Global action plan on antimicrobial resistance.

Xie WY, Shen Q, Zhao FJ. 2018. Antibiotics and antibiotic resistance from animal manures to soil: a

review. *Eur J Soil Sci* **69**:181–195. doi:10.1111/EJSS.12494

Zárate SG, Claure MLD la C, Benito-Arenas R, Revuelta J, Santana AG, Bastida A. 2018. Overcoming Aminoglycoside Enzymatic Resistance: Design of Novel Antibiotics and Inhibitors. *Mol 2018, Vol 23, Page 284* **23**:284. doi:10.3390/MOLECULES23020284

Zhang AN, Hou C-J, Negi M, Li L-G, Zhang T. 2020. Online searching platform for the antibiotic resistome in bacterial tree of life and global habitats. *FEMS Microbiol Ecol* **96**:107. doi:10.1093/FEMSEC/FIAA107

Zurfluh K, Nüesch-Inderbinen M, Morach M, Berner AZ, Hächler H, Stephan R. 2015. Extended-Spectrum-Lactamase-Producing Enterobacteriaceae Isolated from Vegetables Imported from the Dominican Republic, India, Thailand, and Vietnam. doi:10.1128/AEM.00258-15

# Section V : Structuration phylogénétique, écologique, et génomique du transfert horizontal des gènes de résistance

Dans le chapitre précédent, on s'est intéressé à la distribution spatio-temporelle et éco-géographique des gènes de résistance aux aminoglycosides, et on a pu constater que cette distribution s'explique davantage par leur propagation spatiale et leur association à des éléments génétiques mobiles que par la sélection antibiotique. Cependant, le mouvement des gènes de résistance ne se fait pas que dans ce contexte. En effet, la propagation des gènes de résistance entre espèces est l'une des problématiques majeures de la résistance, car elle permet la diffusion rapide des résistances à de nouvelles espèces. Ceci implique que, bien que les gènes de résistance puissent apparaître dans des taxons non pathogènes et être hérités verticalement, ces derniers puissent être ensuite acquis par des espèces pathogènes par HGT. Ce mécanisme a déjà été observé et semble fréquent dans l'histoire évolutive de la résistance : de nombreuses familles de gènes de résistance aux antibiotiques ont été trouvées dans un large éventail de taxons bactériens avec une forte similarité de séquence, ce qui montre que ces gènes ont très probablement été transmis horizontalement.

Bien qu'on sache le HGT répandu, de nombreuses questions restent encore aujourd'hui en suspend quant à la tendance générale de transfert et aux facteurs pouvant le structurer. La réponse est pourtant essentielle pour la mise en place de stratégies à grande échelle de confinement de la résistance. Par exemple, de nombreux gènes de résistance n'ont été observés que récemment chez des espèces pathogènes, mais leur origine reste inconnue. Pour empêcher le transfert de résistances à ces espèces, il est donc crucial d'identifier les taxons et les contextes qui diffusent ces gènes vers les organismes pathogènes, afin que les stratégies de confinement les ciblent en particulier. De même, on sait que la mise en place du HGT dépend de certaines conditions, et qu'il existe notamment certaines barrières (par exemple une distance phylogénétique trop importante) ou certains faciliteurs (par exemple le portage par des MGE). Mais ces barrières ne sont pas spécifiques aux gènes de résistance, et il existe donc encore une incertitude quant à l'influence qu'exercent les facilitateurs et les barrières au HGT sur la circulation des ces gènes.

Ce chapitre réutilise le jeu de données constitué dans le chapitre précédent, afin d'examiner ces différentes questions dans le cas des gènes de résistance aux aminoglycosides. Dans un premier temps, les gènes qui ont été impliqués dans un transfert horizontal récent entre genres différents ont été détectés par une méthode phylogénétique implicite. Dans un second temps, j'ai reconstruit des réseaux dirigés décrivant le transfert horizontal des gènes de résistance d'un taxon à l'autre. D'une part, ces réseaux mettent en évidence que la résistance aux aminoglycosides a été particulièrement propagée

par des taxons d'organismes non pathogènes, et que les taxons pathogènes dont la résistance pose le plus problème actuellement sont en réalité plutôt des receveurs dans le transfert de résistances. D'autre part, l'ensemble des résultats suggèrent que les barrières connues au HGT peuvent en réalité être contournées par l'association des gènes de résistance à certains contextes génomiques, en particulier à des MGE emboîtés (par exemple lorsqu'un plasmide porte un transposon, qui porte lui-même un gène de résistance).

Il est à noter que certaines analyses présentées dans ce manuscrit en cours de construction ne sont pas complètement abouties : *(1)* La manière dont sont établis les *core genomes* , utilisés pour calculer les tableaux d'usage de codons des génomes récepteurs, pourrait être changée, en particulier en utilisant des outils bioinformatiques tel que *roary*. *(2)* Un problème technique dont je n'ai pas encore identifié l'origine m'a empêché jusque-là d'obtenir des mesures de compatibilité d'usage de codons (indice *COUSIN18*) pour une grande partie des gènes transférés ; quand ce problème technique aura été résolu, les analyses incluant ce facteur pourront être conduites sur un jeu de données plus ample. (*3)* La manière dont l'association aux MGE est prise en compte dans les régressions binomiales peut être affinée. *(4)* La mesure de la probabilité des événements de HGT dans la dernière partie des résultats est probablement à perfectionner. La partie des résultats qui porte sur l'impact de la compatibilité d'usage de codons et sur le rôle des MGE, en particulier des MGE emboîtés, sera donc potentiellement révisée, ainsi que la partie de la discussion qui s'y rapporte.

# Deciphering the horizontal traffic rules for aminoglycoside resistance genes through a network analysis

Léa Pradier, Stéphanie Bedhomme

## Abstract

Horizontal gene transfer (HGT) is known to be pervasive across bacterial communities, and had been identified as one of the main routes for the recent spread of antibiotic resistance genes. Although barriers and facilitators of HGT have been described on specific examples, little is known on their global impact on the propagation of antibiotic resistances. Here we studied the horizontal traffic of 27 families of genes encoding aminoglycoside-modifying enzymes. We used an implicit phylogenetic method to detect those which had been recently transferred between genera, and we constructed directed networks to identify their most likely routes of propagation. These networks are heterophilic and show an overrepresentation of HGT from non-pathogenic taxa to human pathogens. Moreover, subsequent analysis confirmed several barriers to HGT: phylogeny, ecology, and codon usage incompatibilities. But it also showed that at least some of these barriers could be bypassed when resistance genes were associated with nested mobile genetic elements (e.g. transposons carried by plasmids). Together, these results suggest that the spread of antibiotic resistance depends much more on non-pathogenic bacteria than previously estimated, and that nested mobile genetic elements can serve as shortcuts for propagation.

## 1. Introduction

Horizontal gene transfer (hereafter HGT) is one of the major mechanisms that shape the evolution of prokaryotic organisms (Soucy et al., 2015). On the one hand, it has a major qualitative impact on the recombination of genomes, allowing them to avoid clonal interference, and providing them sets of genes that permit adaptation to new environments or lifestyles (Marijuán et al., 2010; Schaack et al., 2010) but see Koonin, 2016). On the other hand, numerous gene families are mainly inherited horizontally and can thus quickly spread between distantly related lineages (Gal-Mor & Finlay, 2006; Holden et al., 2004).

There are a number of factors that favor the occurrence of HGT events. Notably, mobile genetic elements with intergenomic mobility, e.g. conjugative plasmids, are known to transfer resistance between genera, phyla, and even major domains (Sørensen et al., 2005). High donor population size (B. Li et al., 2019) and spatial structure (Cairns et al., 2018) can also favor HGT, by increasing the likelihood of encounter between donors and recipients. Finally, HGT is more frequent between closely related taxa (Andam & Gogarten, 2011; Popa et al., 2011; Soucy et al., 2015).

However, even if HGT seems to be ubiquitous and very prevalent among Prokaryotes, core

genomes can always be reconstructed for any taxon, and the composition of the dispensable genome can strongly vary from one strain to another (see e.g. Forsberg et al., 2014) for antibiotic resistance genes). This indicates that genes do not move freely and randomly by HGT and that there are rules and constraints on the horizontal circulation. First, metagenomics studies showed the existence of physical barriers, such as limited transfer between ecologically distant communities (Shapiro et al., 2012 ; Smillie, 2011). But horizontally transmitted genes (hereafter HT genes) may also imply costs for the recipient bacteria (Baltrus, 2013), which limit the success of HGT when strains which recently received HGT are selected against. These costs may involve replication and expression costs (Park & Zhang, 2012), as well as a disruption of the host gene regulatory networks (San Millan et al., 2015). Yet, even the codon usage bias (hereafter CUB) of these genes can generate costs when the recipient genome has a different CUB (Amorós-Moya et al., 2010; Bedhomme et al., 2019; Walsh et al., 2020), and it was suggested that CUB incompatibilities could decrease the retention of HT genes (Callens et al., 2021). In part of the cases, these costs can be counterbalanced by the adaptive functions of newly acquired genes, as horizontally transferred genes are known to be enriched in metabolic genes and genes involved in interspecific interaction warfare, such as antibiotic resistance genes.

One of the current main issues related to HGT is its major role in the spread of antibiotic resistance. Numerous antibiotic resistance gene (hereafter ARG) families have been found in a wide range of bacterial taxa with a high sequence similarity, which shows that these genes were horizontally transmitted (e.g. Gibson et al., 2014). A substantial body of research has consequently focused on tracking the spread of ARG families across the phylogeny (Jiang et al., 2017), and on testing the conditions required for the horizontal spread of ARGs, either in isolated populations (B. Li et al., 2019) or in small experimental communities (Cairns et al., 2018). Yet, most of these studies were limited to plasmid-carried ARGs (e.g. Cairns et al., 2018; B. Li et al., 2019; Q. Li et al., 2019) and/or taxa of clinical interest (mostly human and animal commensals or pathogens, see e.g. Cooper et al., 2017; Jiang et al., 2017; B. Li et al., 2019).

These approaches can inform us on the extent of the horizontal transmission of ARGs, and on the mechanisms that shape it. Apart from experimental approaches are only restricted to small, simple systems, other bioinformatic and statistical analysis of HGT usually rely on network analysis. First, gene-sharing networks are based on the detection of pairs of genes carried by different genomes that share high sequence identity. Gene-sharing networks can inform us on the set of taxa which share the same genes, and thus on the extent to which HGT had occurred, but not on the donor-to-recipient relationships. Second, phylogenetic networks (see e.g. Makarenkov et al., 2006) were developed to take into account the discrepancy between several gene phylogenies that resulted from HGT. Phylogenetic networks provide useful information on the donor-to-recipient relationships, but they require phylogenetic reconstruction at a wide range of loci and thus demand high computational power even when focusing on a single gene family. Finally, directed HGT networks appear as a methodological compromise. Contrary to phylogenetic networks, HGT networks do not require

phylogenomic reconstruction: the construction of HGT networks simply relies on the inference of ancestry for HT genes. Not only does this inference allow to track the propagation of HT genes much more precisely than gene-sharing networks, but topological analyses and the identification of donor-recipient pairs also allows to draw conclusions on the parameters which shape HGT for ARGs. This approach was already applied in past research (see e.g. Cong et al., 2016; Popa et al., 2011, 2016), but never for the specific case of ARGs.

In this study, we tracked the horizontal transmission of several families of aminoglycoside resistance genes with directed HGT networks. Aminoglycosides are traditional Gram-negative antibiotics that bind to the 30S subunit of prokaryotic ribosomes and thus inhibit protein synthesis (Mingeot-Leclercq et al., 1999). This antibiotic class was first introduced in 1944 with the successful use of streptomycin against tuberculosis. Nowadays, AGs are only used in humans as a second-line or last-resort treatment for Gram-negative bacteria (Garneau-Tsodikova & Labby, 2016) but they remain frequently used in agriculture and veterinary medicine to treat septicemia and infections of the digestive, respiratory, and urinary tracts in a wide variety of domestic animals (European Medicines Agency, 2017). In a previous study (Pradier & Bedhomme, 2022), we screened more than 160,000 bacterial genomes for a specific class of aminoglycoside resistance genes (genes encoding aminoglycoside-modifying enzymes, hereafter AMEGs), which we found in nearly 40,000 genomes. The date, location, and biome of sampling were recovered for each genome, when possible, and the potential association with mobile genetic elements (hereafter MGEs) was screened. Here, we further analyze and detect this dataset and detect recent HGT events, using an implicit phylogenetic method. We then reconstruct directed HGT networks for a subset of the gene families analyzed. Topological analyses of these networks allow us to (i) to investigate whether there is a preferential direction in the horizontal transmission of AMEGs; (ii) to assess the relative roles of different factors that may constitute barriers or facilitators to HGT in the specific case of ARGs.

## 2. Material and methods

### 2.1 Dataset

In a previous study (Pradier & Bedhomme, 2022), 160,987 Eubacteria genomes available on the NCBI Assembly database were screened for the presence of 27 clusters of homologous genes (CHGs) coding for aminoglycoside resistance proteins. The CHGs had been determined using Silix (Miele et al., 2011) on a collection of published aminoglycoside-modifying enzymes (Garneau-Tsodikova & Labby, 2016; Ramirez & Tolmasky, 2010) and the genomes were screened using HMMER (Eddy, 2011). The starting point of the present study is thus a collection of 38,523 bacterial genomes in which at least one AMEG has been detected and confirmed. Each genome has been assigned to the bac120 taxonomy thanks to the GTDB-tk pipeline (Chaumeil et al., 2020), which uses a set of 120 ubiquitous proteins, present in ≥90% of genomes, and in a single copy in ≥95% of genomes (Parks et

al., 2017, 2018). Additionally, genome metadata (date, location, biome of sampling) were collected when available and the genomic context of the AMEG was analyzed in terms of association with mobile genetic elements (prophage, plasmid, integrative conjugative elements, transposons).

## 2.2 Detection of horizontally transmitted antibiotic resistance genes

### 2.2.1 Implicit phylogenetic method

The first aim of this paper was to accurately detect AMEGs that had been involved in horizontal transmission (hereafter, HGT events) on a large dataset. An implicit phylogenetic method, inspired by Corel et al. (2018), was used to detect horizontally transmitted genes. The rationale is to test for each pair of resistance genes from the same cluster whether their p-distance is lower than expected if they had been acquired vertically. The set of 120 ubiquitous proteins used for phylogenetic assignment (bac120, Parks et al., 2018) was taken as a reference for vertical inheritance. This phylogeny has a maximal degree of resolution for species, so inference of HGT events could not be performed within species. For each species pair, p-distances were measured between all these 120 proteins, thus defining a reference for the distribution of genetic divergence for vertically transmitted genes. A confidence interval for the 5% quantile of this distribution was determined by a bootstrap procedure (10,000 samples with replacement). For each CHG, proteins were aligned with Clustal Omega (Sievers & Higgins, 2014), and p-distances were computed between the proteins of each pair of species. A putative horizontal transfer event was reported when the p-distance between a pair of aminoglycoside resistance proteins was significantly lower than expected for vertical transmission, i.e. when it was lower than the lower bound of this confidence interval of the p-distances distribution for the ubiquitous set of vertically transmitted genes. For each pair of species, we were thus able to report a set of putative horizontally transmitted genes. This analysis focused on recent HGT events only, and thus pairs of resistance proteins for which p-distance was above 5% were excluded.

### 2.2.2 Assessment of the implicit phylogenetic method

In order to assess the accuracy of the aforementioned method, sequence alignments were simulated under different scenarios. For each CHG, the *pyvolve* library (Spielman & Wilke, 2015) was used to simulate a range of sequence alignments, with size equal to the median protein size of the CHG, and a phylogenetic distribution corresponding to the one of the CHG. The WAG model was considered as the null hypothesis for vertical inheritance, as for the bac120 phylogeny (Parks et al., 2018). This procedure was replicated 1000 times for each CHG. In each sequence alignment simulation, a HGT event was artificially introduced by replacing, for a randomly chosen pair of species, the protein sequence in one species by the one of the other species. The HGT detection method explained above was then applied on this simulated data: p-distances were computed between all protein pairs, and pairs were assigned as putative HT genes depending on their p-distance. For most CHGs, this method gave high precision (i.e. few false positives) for the largest phylogenetic scales (see Supplementary

information, Fig. S1): precision ranges from 63% to 100% depending on the CHG, with a median of 100%, for between-genus HGT events. However, this precision decreased rapidly when considering within-genus HGT events. Indeed, HGT events did reduced the genetic distance between donors and recipients, but they also reduced the genetic distance between recipients and taxa closely related to the donor. These taxa could be detected as involved in HGT events, whereas they may never have been. Intrageneric events were thus discarded for subsequent analysis.

## 2.3 Reconstruction of directed horizontal gene transfer networks

### 2.3.1 Rationale

Implicit phylogenetic methods assume that, for a given pair of taxa, two sequences that were shared by HGT diverged less than those that were vertically inherited. However, this implication is not always reciprocal. As HGT distorts the relationship between genetic and phylogenetic distances, some sequences can be detected as horizontally transmitted simply because they are closely related to ones that were actually involved in HGT events. It is therefore necessary to double check that putative HGT sequence pairs are likely to account for actual HGT events. Moreover, as genetic distances are reciprocal, implicit phylogenetic methods cannot be used to infer donor-recipient relationship. In the past decade, alignment-free distances were proven to be accurate and straightforward methods to measure sequence relatedness, e.g. to reconstruct genome phylogenies (Fan et al., 2015) or to classify viruses (Ahlgren et al., 2017). The rationale of alignment-free methods is that sequences evolving within the same genome tend to have the same nucleotide k-mer composition. Using these methods, it is therefore possible to detect horizontally transferred genomic regions (because their composition substantially differs from the species nucleotide composition), and to potentially identify which genome acted as donor in a HGT event (because the transferred sequence is expected to have a nucleotide k-mer composition which is closer to the donor one than to the receiver one, Bernard et al., 2016; Cong et al., 2016). Among the diversity of metrics used to calculate compositional distances, Tang et al. (2018) identified $d_2^*$ (based on 3-mers with a 1-fold Markov chain, Reinert et al., 2009) as one of the most reliable measures to detect HGT events.

Inspired by these approaches, putative HGT protein pairs were submitted to compositional comparisons, using $d_2^*$ as a measure of relatedness. Consider a gene $g_0$, carried by the genome $G_0$, and involved in *n* putative HGT sequence pairs with the genes $g_1$ to $g_n$, carried by the genomes $G_1$ to $G_n$. If $g_0$ was horizontally transmitted from the genome $G_i$ (with $i \in (1,n)$), then the compositional distance between $g_0$ and $G_i$ must be lower than a threshold $d_2^*{}_{max}$ above which $G_i$ is unlikely to be the donor of $g_0$ (condition (i)). Moreover, the distance between $g_0$ and $G_i$ should be lower than the distance from $g_0$ to $G_0$, otherwise $g_0$ would be very unlikely to have been horizontally acquired (condition (ii)). Among all the genomes that satisfy conditions (i) and (ii), the most likely donor of $g_0$ is assumed to be the genome with the minimum compositional distance to $g_0$. Let the identified donor

be $G_i$, then the $G_i \rightarrow G_0$ event is reconstructed.

This procedure is applied to all the sequences involved in putative HGT protein pairs detected by the implicit phylogenetic method. Once pairs of donor-receiver in HGT events were identified and the direction of transfer established, HGT networks were reconstructed where vertices represent taxonomic units and edges represent horizontal transmission events.

### 2.3.2 Computation of compositional distances

All $d_2^*$ measures were performed with the software CAFE (Lu et al., 2017). For each protein, the corresponding DNA sequence was recovered. In order to correctly assess the composition of each DNA sequence, genomic context upstream and downstream was extracted along with each gene, up to the limit of 5kb contigs. $d_2$-derived metrics have asymptotic normality and can theoretically be tested for significant differences. However, though CAFE provides $d_2^*$ measures with 6 digits, it can be assumed that the precision of differences in $d_2^*$ is actually lower than $10^{-6}$, due to the limited size of compared sequences. A measure for the precision of $d_2^*$ differences, henceforth called $\Delta d_2^*$, is introduced, such that $d_2^*(g_0, G_i) < d_2^*(g_0, G_j)$ is considered true if and only if $d_2^*(g_0, G_i)+\Delta d_2^* < d_2^*(g_0, G_j)$.

### 2.3.3 $d_2^*{}_{max}$

When looking for the donor genome of a gene, then it must first be assessed whether any genome in the dataset considered can actually be a donor for that gene. If the horizontally transferred gene is sequenced straight after the HGT event, it will still carry the composition signature of its donor genome. However, a gene can gradually diverge from its original genomic signature after transmission, and our aim was to be able to retrace inheritance despite this post-transmission divergence. $d_2^*{}_{max}$ is thus defined as a maximum value of $d_2^*$ between a donor genome and the gene that was passed on. This threshold is defined from the distribution of $d_2^*$ among randomly simulated sequences.

To define the composition of sequences, a frequency table was randomly drawn for nucleotide 3-mers (3-mers were considered because they are then used by CAFE to compute $d_2^*$). This table was drawn in two steps: half of the frequencies were drawn from the 64 possible 3-mers, and the remaining half was drawn on a subsample of 3-mers. This way, we ensured that the frequency table was biased towards some 3-mers, since the distribution of nucleotides and codons in bacterial genomes are usually biased (Botzman & Margalit, 2011). However, as far as we know, no global distribution of skewness for bacterial nucleotide composition was established, so we had to explore a wide range of biased frequency tables. Two random sequences are then built from this composition table: a large sequence $G_0$ to simulate a genome, and a smaller sequence $g_0$ to simulate a gene (and in some cases its genomic context). As the smallest genome in our dataset contained 1.9Mb, the size for $G_0$ was fixed to 666,666 3-mers. Contigs that were extracted for genes and their genomic contexts ranged from

150pb to 5kb, so sizes for $g_0$ also ranged from 50 to 1666 3-mers. From $g_0$, n positions are randomly mutated to construct a related gene $g_1$. $d_2^*$ was then measured between $G_0$ and $g_1$. This procedure was independently replicated 500 times for different gene sizes (50, 100, 150, 200, 333, 450, 999, or 1666 3-mers), different composition biases (constituted by oversampling 10, 30, 50, or 64 3-mers), and different divergence rates between $g_0$ and $g_1$ (0.1%, 1%, 10% of positions mutated, being hypothesized that pairs of genes should diverge slower under the hypothesis of vertical inheritance), for a total of 48,000 independent simulations. The 95% quantile of the $d_2^*$ distribution was then calculated to define the threshold of $d_2^*$ above which it is unlikely that a given genome is a credible donor for a focal gene. This gives the following linear regression: $d_2^*_{max}$=0.983-0.062*log(N) +0.075*log(N)*sqrt(p), $R^2$=0.426, with *N* being the gene size and *p* the p-distance between the focal gene $g_1$ and its homologous $g_0$.

### 2.3.4 $\Delta d_2^*$

Two closely related genomes usually have very similar compositional signatures. When this is the case, it might be impossible to distinguish between two $d_2^*$ measures, and therefore to assign any donor genome to a focal gene. $\Delta d_2^*$ is thus defined as the minimum $d_2^*$ difference below which two genomes are equally likely to be donors of a gene. This value is defined from the distribution of $d_2^*$ differences for simulated random sequences.

As in section 2.2.3, a frequency table of nucleotide 3-mers was randomly drawn. Then, from this table, two random sequences were generated: a genome $G_1$ (666,666 3-mers) and a gene $g_1$ (ranging from 50 to 1666 3-mers). Then, to simulate two closely related genomes, the genome $G_2$ is created by randomly mutating *n* positions in $G_1$. A new 3-mer frequency table is calculated from the sequence of $G_2$, and a new gene $g_2$ (same size as $g_1$) is randomly drawn from that second table. $d_2^*$ values were then computed for the following pairs: $G_1$-$g_1$, $G_2$-$g_2$, $G_1$-$g_2$, $G_2$-$g_1$, and $G_1$-$G_2$. We then calculated the following average difference:

$$\delta = ( |d_2^*(G_1,g_1) - d_2^*(G_2,g_1)| + |d_2^*(G_2,g_2) - d_2^*(G_1,g_2)| ) / 2$$

This procedure was independently replicated 500 times for different gene sizes (50, 100, 150, 200, 333, 450, 999, or 1666 3-mers), different composition biases (biased towards 10, 30, 50, or 64 3-mers), and different divergence rates between $G_1$ and $G_2$ (0.01%, 0.1%, 0.2%, 0.5%, 1%, 5%, 10%, or 30% of positions mutated), for a total of 128,000 independent simulations. $\Delta d_2^*$ was then defined as the 5% quantile of $\delta$, i.e. 95% of $d_2^*$ differences should be above this threshold. A linear regression established the following relationship: $\log(\Delta d_2^*)$=-3.090+0.552*$\log(d_2^*(G_1,G_2))$, $R^2$=0.990. For any pair of genomes, the precision level could then be predicted by calculating $d_2^*$ between these two genomes.

### 2.3.5 Approximation of donors for network construction

HGT network reconstruction was performed at the level of genomes: genome $G_i$ was considered as the donor genome for a gene $g_0$ if $d_2^*(g_0, G_i)+\Delta d_2^*$ was lower than any other $d_2^*$ measure. However, it could not be claimed that this method allows to accurately find donor genomes, but only genomes that are (more or less) closely related to the actual donors. Thus, we hypothesize that our method could approximate donor taxa, at least at the level of species. For subsequent analyses, genomes were gathered into higher phylogenetic levels: species or genera. At these levels, taxon $T_i$ was considered as the donor taxon for a gene $g_0$ if $T_i$ contained at least one genome $G_i$ such that $d_2^*(g_0, G_i)+\Delta d_2^*$ was lower than any other $d_2^*$ measure. When several genes were detected as transferred between from one taxon to the other, we could not assess whether numerous HGT events occurred between the two taxa or whether, for example, only one HGT event occurred, but the gene then propagated in the recipient taxon. Thus, topological analysis considered only one edge for each possible direction of transfer between two taxa.

## 2.4 Topological analysis

Several topological indices were measured on all the HGT networks with the R library *igraph*. Assortativity is a preference for a network's vertices to attach to others that are similar in some way: either in degree (degree assortativity), or in a nominal variable (categorical assortativity). Diameter is the length of the longest shortest-path in the network. Reciprocity is a measure of the likelihood of vertices in a directed network to be mutually linked.

Sensitivity analyses were additionally performed on each HGT network. A first analysis consisted in randomly removing a fraction (5%, 10%, or 30%) of genomes from the dataset, and then to reconstruct HGT networks, in order to measure sensibility to low sampling. A second analysis consisted in randomly rewiring a fraction (10%, 30%, or 50%) of edges in each HGT network, in order to measure the difference between reference HGT networks and random networks. Sensitivity to sampling or random rewiring was assessed by testing the difference in several topological measures listed above between reference HGT networks and the distribution of resampled/rewired networks.

## 2.5 Metadata

### 2.5.1 Definition of pathogenicity

In order to track HGT events between pathogenic and non-pathogenic bacteria, four databases listing pathogenic species or taxa were compiled. Taxa further labelled as pathogenic in this study are: (i) the 12 taxa considered priority for the development and production of new antibiotics by WHO because of their high prevalence of antibiotic resistance (World Health Organization, 2017), the 18 taxa estimated to pose the highest national security and public health threats in the USA (National Institute of Allergy and Infectious Diseases, 2018) and (iii) bacteria species involved in pathogenic interactions with human hosts in two databases gathering the characteristics of several hundreds of pathogen-host

interactions (PHIBASE, Urban et al., 2020 and Wardeh et al., 2015 databases). The classifications of species are listed in Supplementary information, Tab. S1.

### 2.5.2 Phylogenetic and ecological distance

Phylogenetic distances between taxa were estimated in the bac120 phylogeny, with the function *distTip* of the R library *ape*. Distances between species were directly measured on the phylogenetic tree. Distances between genera were measured between the most recent common ancestors of the species of each genus. To evaluate the phylogenetic range across which each CHG is distributed, two measures summing up the phylogenetic distribution were computed using the function *pd* from the R library *picante*: (i) species richness, i.e. the number of species in which a CHG was found; and (ii) Faith's distance (Faith, 1992), i.e. the sum of the lengths of all the branches of the minimum tree spanning all the bacteria carrying a specific CHG.

In order to quantify the difference between the ecological niches of different taxa (either species or genera), Jaccard distances were calculated between presence/absence vectors reflecting the occurrence of bacterial taxa in our dataset across 10 terrestrial biomes classified by Pradier and Bedhomme (2022). Jaccard distances, hereafter denominated "ecological distances", were calculated with the function *recluster.dist* of the R library *recluster*. This index, rather than other distances based on prevalences, was chosen because the dataset is biased in favor of clinical samples.

### 2.5.3 Differences of codon usage biases

For each CHG, we evaluated the match in codon usage between all genes identified as members of this cluster and all the species and genera carrying a member of this cluster. First, codon usage tables of the core genome of each taxon were compiled. Many dispensable genes have indeed been acquired by horizontal transfer and their codon usage is likely to not be the product of the coevolution with the translation machinery, contrary to what is expected for the core genes. To delineate core genomes, a maximum of 30 genomes were selected for each species and submitted to Prokka (Seemann, 2014) for detection and annotation of their CDS. CDS with no clear annotation (for which the annotation contained the words "hypothetical" or "putative") were excluded, as well as CDS shorter than 150pb (due to their low reliability in the calculation of codon usage, see Bourret et al., 2019). A core genome was then established for each taxon by gathering all the CDS whose annotations were present in 100% of genomes. From these core genomes, codon usage tables were calculated using the software *COUSIN* (Bourret et al., 2019).

Finally, the similarity of codon usage between AMEGs' nucleotide sequences and potential receiving genomes was measured through the COUSIN18 index calculated by the *COUSIN* software. This index takes values below 0 when the CUB of the query is in the opposite direction to the reference, between 0 and 1 if the CUB of the query is in the same direction but the bias is less pronounced than in the reference and above 1 when the bias is in the same direction and more pronounced than in the

reference.

### 2.6 Statistical analysis

Three types of mixed models were built in this study. First, we computed a binary regression to evaluate the factors which influence whether an AMEG was detected by the implicit phylogenetic method to have been involved in recent HGT events. This model takes association with MGEs and the pathogenicity of the carrier species as two categorical, fixed effects, and the CHG to which the gene belongs as a random effect. This model was computed with the function *glmer* from the R library *lme4*.

Then, we used the *pglmm* function from the R library *phyr* to compute a phylogenetic Poisson GLMM. This model aimed to evaluate the impact of pathogenicity on the out-degree of a species in directed HGT networks. This model takes the *bac120* phylogenetic tree as a correlation structure, and the CHG as a random effect.

Finally, from the HGT networks, we aimed to explain the likelihood of HGT success, i.e the frequency of transmitted genes between donors and recipients. The function *fitme* of R library *spaMM* (Rousset & Ferdy, 2014) was used to compute logistic regressions on the following response variable: $p_{ij}=n_{ij}/n_i$, where $n_i$ is the number of AMEGs carried by the taxon $T_i$, and $n_{ij}$ the number of AMEGs carried by the taxon $T_i$ that were horizontally transmitted from taxon $T_j$. A global model, grouping all CHGs, was performed, with the following random variables: the CHG, taxon $T_i$, and taxon $T_j$. Separate models for each CHG were computed as well to explore the diversity of drivers and barriers to HGT. These separate models included only taxa $T_i$ and $T_j$ as random variables. Variance was considered as Beta-distributed, in order to prevent the possible under- or over-dispersion of residuals. Regarding the model construction of fixed effects, measures of COUSIN18 were missing for a large number of taxa pairs in each network. Models were thus computed with and without COUSIN18. Each model also included the following fixed effects: ecological distance; patristic distance; for each CHG, the frequencies of genes associated with either intergenomic mobility MGEs only, or intragenomic mobility MGEs only, or nested MGEs in the donor taxon $T_j$; as well as the interactions of these frequencies with either patristic distance or COUSIN18.

## 3. Results

### 3.1 The distribution of putative horizontally transmitted AMEGs

We applied an implicit phylogenetic method to search aminoglycoside resistance protein pairs which were likely to have diverged more recently than expected from the bac120 phylogeny. We only focused on recent, large-scale horizontal gene transfer events: we thus searched among 5,973,930 protein pairs with an identity of 95% or higher, and that were carried by species from different genera. We found a total of 27,530 AMEs out of 46,053 that were likely to have been involved in recent HGT

events. The frequency of putative horizontally transmitted AMEGs varies to a great extent across CHGs: from 0% for CHGs 3, 5.2, 6.3, 15, 16, 17.1, 17.2, 24 and 28 to nearly 100% for CHGs 5.1, 13, 14, and 19. We did not find any correlation between this frequency and the phylogenetic spread of CHGs, evaluated either with species richness (Pearson correlation coefficient, $p>0.05$), or with Faith's distance (Pearson correlation coefficient, $p>0.05$).

We computed a mixed model to investigate the determinants for an AMEG to have been involved in a recent HGT event. First, association of AMEGs with mobile genetic elements significantly increases their likelihood to have been involved in HGT (Fig. 1, Supplementary information, Tab. S2). Intergenomic mobility MGEs have a significantly higher effect than intragenomic mobility MGEs (post-hoc Tukey test, $p<10^{-3}$), and nested MGEs have the highest effect on the likelihood to have been involved in HGT (post-hoc Tukey test, $p<10^{-3}$). Moreover, we used several databases to classify species carrying these genes as human pathogens or nonpathogens. We found that the likelihood to be involved in HGT is significantly higher when the gene is carried by a pathogenic species (Fig. 1, Supplementary information, Tab. S2). We additionally performed a mixed ANOVA to describe the phylogenetic distribution of horizontally-transmitted AMEGs. Campylobacterota is the phylum in which the likelihood of being involved in HGT is the significantly highest, followed by Bacteroidota and Firmicutes A. The least susceptible phyla to carry putative HT genes are Actinobacteriota, Cyanobacteria, and Firmicutes I.

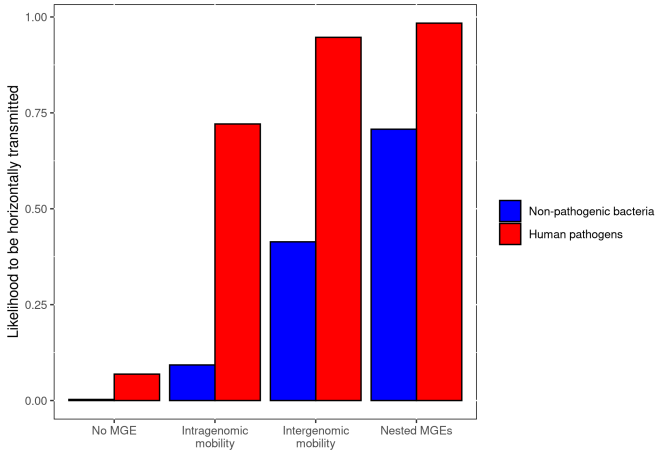


**Figure 1: Effects of genomic contexts and pathogenicity on the likelihood to be involved in horizontal transfer events.** Bars represent the estimates of a mixed binomial regression. Human pathogens are the species which were listed in the WHO, NIAID, PHI-Base and WARDEH databases.

However, pairs of putative horizontally transmitted AMEGs are not oriented, i.e. bacteria carrying these sequences are identified neither as donors nor as recipients for HGT events. We thus selected 6 CHGs out of the 18 in which we identified putative horizontally transmitted AMEGs to perform subsequent analysis: CHGs 1, 2, 13, 14, 20, and 25. These 6 CHGs were chosen because they represent a diversity of configurations encountered among AMEGs CHGs (Table 1): diversity in the frequency of putative HT genes, in their phylogenetic spread, in the frequency at which they were sampled in clinical contexts, and in their associations with MGEs. Additionally, they represent the three main biochemical functions of AMEs: modification of aminoglycosides by transfer of an acetyl (acetyltransferases, AACs), a nucleotidyl (nucleotidyltransferases, ANTs), or a phosphoryl (phosphotransferases, APHs) group.

**Table 1: Summary of the CHGs that were selected for construction of HGT networks**

| CHG | function | f(HGT) | f($MGE_{inter}$) | f($MGE_{intra}$) | f($MGE_{nested}$) | Biomes | Faith distance | f(Clinical) |
|---|---|---|---|---|---|---|---|---|
| **CHG 1** | AAC | 0.892 | 0.029 | 0.391 | 0.446 | 7 | 4.73 | 0.830 |
| **CHG 2** | AAC | 0.556 | 0.121 | 0.082 | 0.331 | 7 | 4.63 | 0.239 |
| **CHG 13** | ANT | 0.987 | 0.147 | 0.286 | 0.515 | 8 | 12.3 | 0.690 |
| **CHG 14** | APH | 0.994 | 0.433 | 0.059 | 0.412 | 5 | 2.01 | 0.196 |
| **CHG 20** | AAC | 0.861 | 0.375 | 0.048 | 0.349 | 6 | 4.87 | 0.190 |
| **CHG 25** | ANT | 0.898 | 0.729 | 0.000 | 0.136 | 3 | 4.30 | 0.500 |

## 3.2 Topological analysis of HGT networks

For these 6 CHGs, we aimed to identify, when possible, the species or genera of origin of the putative horizontally transmitted AMEGs. We used alignment-free methods to reconstruct HGT events, which together form directed HGT networks, at two scales: between species and between genera (Fig. 2). A vast fraction of putative horizontally transmitted AMEGs could be included in these networks (Supplementary information, Tab. S3), and these networks are globally little sensitive to low sampling and significantly different from randomly rewired networks (Supplementary information, Tab. S4-S5). In HGT networks, each taxon (species or genus) is represented by a vertex, and edges account for horizontal transmission events. There is complete correspondence, i.e. edges between the same taxa were reconstructed, between HGT networks at both levels for CHGs 1, 2, 14, 20, and 25, i.e. connections between the same taxa were reconstructed . For CHG 13, 80 connections out of 85 are common to both levels, 3 connections at the species level were not reconstructed at the genera level (*Achromobacter ruhlandii → Pseudomonas aeruginosa*, *Aeromonas salmonicida → Providencia rettgeri D*, *Pseudomonas aeruginosa → Enterobacter himalayensis*), and 2 connections were not reconstructed at the species level (*Escherichia spp. → Klebsiella spp.*, *Pseudomonas A spp. → Pseudomonas E spp.*). Reconstructed HGT events encompass wide phylogenetic distances: only 22%

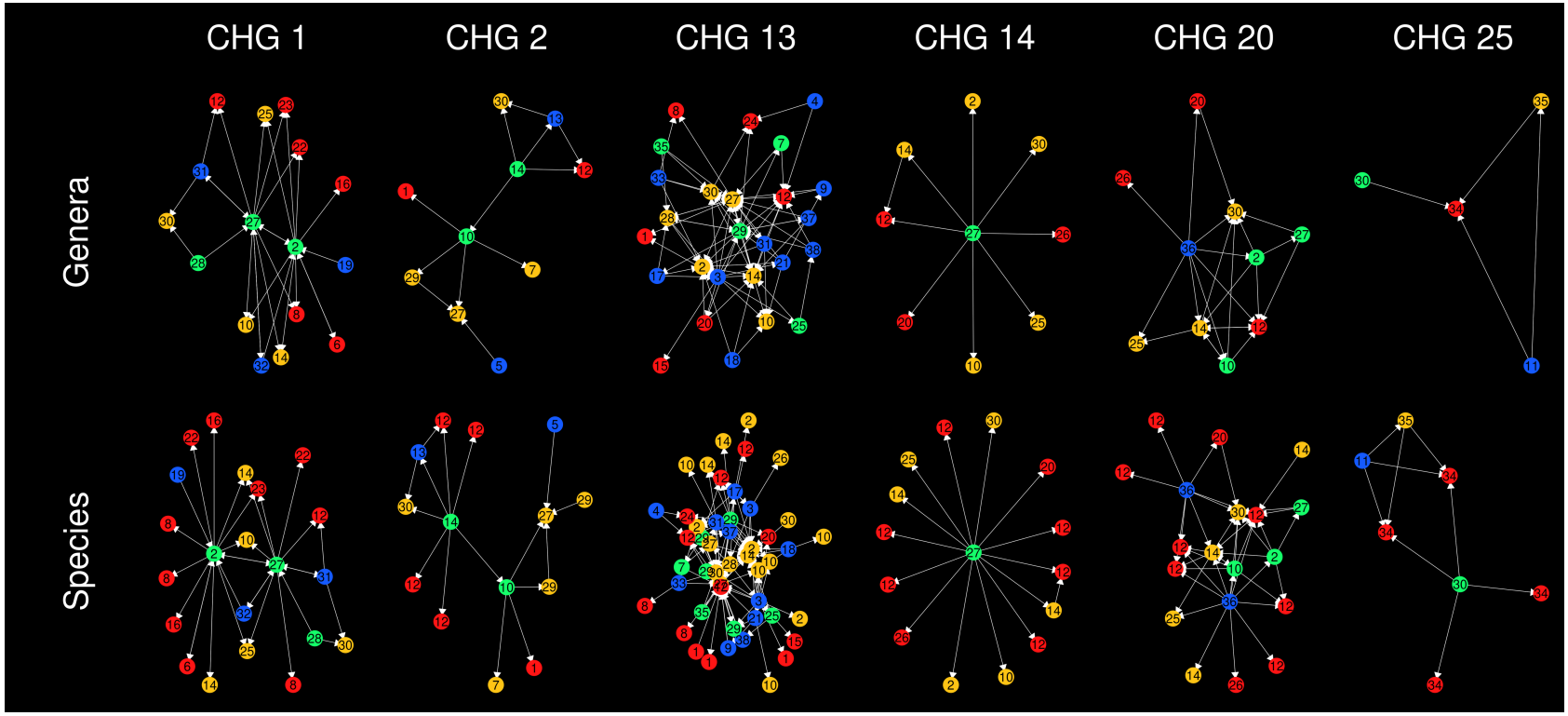


**Figure 2: Horizontal gene transfer networks constructed at different levels.**

Taxa are represented as vertices, and HGT events as directed edges. Blue vertices correspond to taxa which always have an out-degree higher than their in-degree ("donors"), red vertices to those which always have an out-degree lower or equal to their in-degree ("recipients"), green vertices to those which are donor in the focal network but can be recipients in the others, and orange vertices to those which are recipients in the focal network but can be donors in the others. Numbers correspond to the genus to which each vertex belongs. 1: Achromobacter; 2: Acinetobacter; 3: Aeromonas; 4: Alcaligenes; 5: Bosea; 6: Brucella; 7: Burkholderia; 8: Citrobacter; 9: Corynebacterium; 10: Enterobacter; 11: Enterococcus; 12: Escherichia; 13: Gallibacterium; 14: Klebsiella; 15: Klebsiella A; 16: Leclercia; 17: Mannheimia; 18: Metakosakonia; 19: Moraxella A; 20: Morganella; 21: Ochrobactrum; 22: Pantoea; 23: Pantoea B; 24: Pasteurella; 25: Proteus; 26: Providencia; 27: Pseudomonas; 28: Pseudomonas A; 29: Pseudomonas E; 30: Salmonella; 31: Serratia; 32: Shewanella; 33: Sphingobium; 34: Staphylococcus; 35: Stenotrophomonas; 36: Streptomyces; 37: Trueperella; 38: Wohlfahrtiimonas.

of connections occurred within families, whereas more than 10% of events connected different phyla. Yet, this varies a lot depending on the CHG considered (see Fig. 3). Indeed, 50% of CHG 2 transfers occurred within families and all within the same phylum, while 75% of CHG 25 transfers occurred between different phyla and none within the same order.

**Table 2: Topological metrics of HGT networks**

| CHG | Level | Degree assortativity | Diameter | Reciprocity |
|---|---|---|---|---|
| CHG 1 | Species | -0.664 | 4 | 0.258 |
| | Genera | -0.617 | 4 | 0.286 |
| CHG 2 | Species | -0.695 | 2 | 0 |
| | Genera | -0.659 | 2 | 0 |
| CHG 13 | Species | -0.270 | 7 | 0.069 |
| | Genera | -0.336 | 6 | 0.146 |
| CHG 14 | Species | -0.681 | 1 | 0 |
| | Genera | -0.661 | 1 | 0 |
| CHG 20 | Species | -0.437 | 4 | 0.095 |
| | Genera | -0.492 | 3 | 0.174 |
| CHG 25 | Species | -0.400 | 1 | 0 |
| | Genera | -0.577 | 1 | 0 |

**Table 3: Categorical assortativity of HGT networks at the level of species**

| CHG | WHO database | NIAID database | PHIBASE database | WARDEH database |
|---|---|---|---|---|
| CHG 1 | -0.107 | -0.057 | -0.332 | 0.304 |
| CHG 2 | -0.385 | 0 | -0.333 | -0.433 |
| CHG 13 | 0.023 | -0.063 | 0.019 | -0.017 |
| CHG 14 | 0 | 0 | 0 | 0 |
| CHG 20 | 0.033 | -0.048 | -0.018 | 0.018 |
| CHG 25 | -0.154 | -0.154 | -0.125 | NA |

HGT networks form overall short paths: maximum diameter is 6 at the genera level and 7 at the species level for CHG 13, and minimum diameter is 1 at both genera and species level for CHGs 14 and 25. Moreover, HGT networks exhibit negative degree assortativity, varying between -0.66 and -0.26, and very low reciprocity (from 0 for most networks to a maximum of 0.29 for genera exchanging CHG 1). Thus, few taxa with high degree exchange genes with many taxa with low degree. Consistently, out-degree is positively correlated to the inverse of in-degree at the species level (Pearson correlation coefficient, $\rho=0.208$, $p=2.0.10^{-2}$), but we could not find the same relationship at the genera level ($p>0.05$). Thus, at the species level at least, we can identify two main profiles for taxa in HGT networks: few "donor" taxa with high out-degree and low in-degree, which transferred AMEGs to many "recipient" taxa with high in-degree and low out-degree (Figure 2). Some of the taxa identified as donors for one CHG can be identified as recipients for other CHGs (e.g. *Klebsiella pneumoniae* has 7 outcoming edges for CHG 2, but also 13 incoming edges for CHG 13), whereas some taxa can be both donors and recipients for the same CHG (e.g. for CHG 1, *Acinetobacter baumannii* has 12 outcoming and 4 incoming edges). Yet, several species could sometimes be

identified as donors only (*Pseudomonas E psychrophila*, *Aeromonas hydrophila*, *Enterobacter himalayensis*, *Enterococcus faecalis*, *Streptomyces albidoflavus*, *Streptomyces bacilliaris*) or as recipients only (*Escherichia coli*, *Escherichia flexneri*).

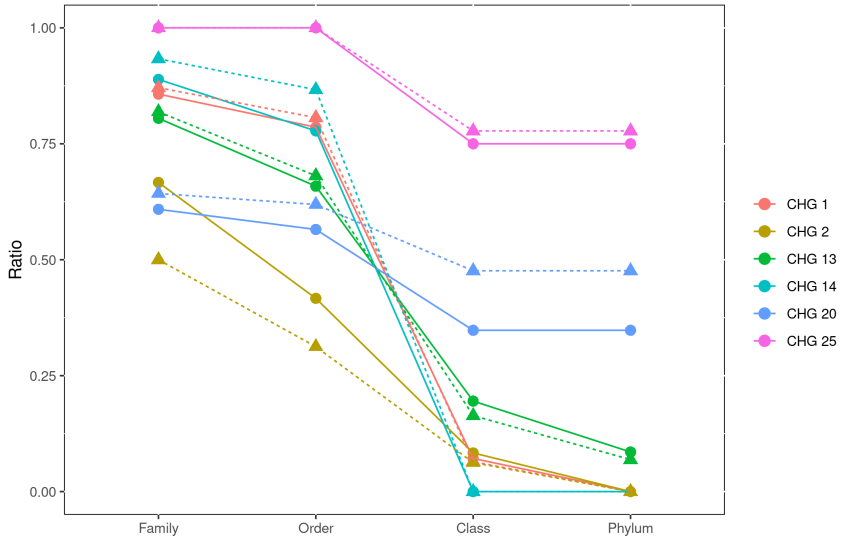


**Figure 3: Inter-taxa connection ratio in HGT networks.** Connection ratios for networks at the level of genera are represented by plain lines and circle points, those at the level of species by dotted lines and triangle points.

We thus aimed to identify the variables that increase the out-degree of a species, i.e. what increases its likelihood of being a source for horizontal transmission of an AMEG. In particular, we attempted to identify whether donors and recipients were structured by pathogenicity. Indeed, whichever classification we use to define pathogenic and non-pathogenic species, all the HGT networks have very low (up to 0.03 for CHG 20 classified with the WHO database) or negative (down to -0.43 for CHG 2 classified with the WARDEH database) categorical assortativity. In particular, while non-pathogenic bacteria represent only 21.8% of vertices in HGT networks, transmission from non-pathogenic to pathogenic bacteria is overrepresented with 35% of HGT events. For subsequent analysis, we classified species into three categories: highly-problematic pathogens (when reported in the WHO and/or NIAID databases, 46% of outcoming and 81% of incoming edges), other pathogens (when reported in the PHIBASE and/or WARDEH but not in the WHO and NIAID databases, 14% of outcoming and 7% of incoming edges), and non-pathogenic bacteria (when reported in no database, 39% of outcoming and 12% of incoming edges, see Fig. 4). A phylogenetic generalized mixed model shows that highly-problematic pathogens have a significantly lower out-degree than non-pathogenic bacteria, although no significant difference was found between other pathogens and highly-

problematic pathogens and between other pathogens and non-pathogenic bacteria (Supplementary information, Tab. S6).

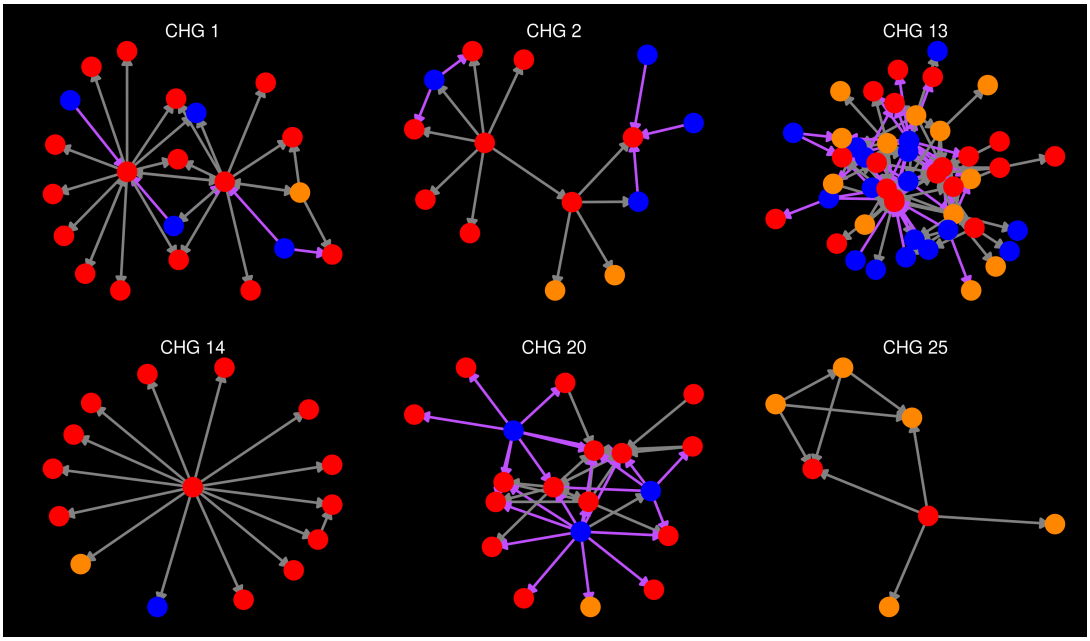


**Figure 4: Pathogenicity in species HGT networks.** Species are represented as vertices of different colors: highly-problematic pathogens in red, other pathogens in orange, and non-pathogenic species in blue. HGT events from a non-pathogenic to a pathogenic species are represented as violet edges, and other HGT events as gray edges.

### 3.3 Factors influencing the likelihood of HGT events

Finally, we investigated the factors that influenced the occurrence and the success of a HGT event between two taxa. To do so, we computed mixed beta-binomial regressions to explain the response variable $p_{ij}$, i.e. the proportion of AMEGs carried by taxon $T_i$ that had an origin in taxon $T_j$. Models were computed on the whole dataset, as well as on CHGs 1, 2, 13, and 20 separately, at both genera and species levels. Models were not computed on CHGs 14 and 25 due to the low number of taxa represented in HGT networks. Besides, due to missing COUSIN data, each model was computed twice, with or without COUSIN as a fixed effect.

We found significant negative effects for both phylogenetic distance and ecological distance on the likelihood of HGT events on the whole dataset (Supplementary information, Tables S7-S10), but also for CHGs 1, 2, and 13 (Tables S11-S22). For the whole dataset and for CHG 1 in particular, at the genera level (Table S11), COUSIN has a positive effect on the likelihood of HGT events in HGT networks: i.e., AMEGs are more likely to be transmitted to taxa with codon usages similar to their ones.

MGEs have diverse effects on the probability of HGT events, depending on the CHGs, and

mostly in interaction with phylogenetic distance and COUSIN. First, association with intergenomic mobility MGEs has a significant positive effect on the whole dataset (Tables S7-S10), but negative effects for CHG 1 (network at the genera level) and 13 (network at the species level). Association with these MGEs also has a significant positive interaction with phylogenetic distance on the whole data set at the species level, for CHG 1 at the genera level, and CHG 13 at the species level: intergenomic mobility MGEs counteract the negative effect of phylogenetic distance on transfer likelihood. AMEGs associated to intergenomic mobility MGEs have a transfer likelihood which increases with phylogenetic distance, which is not the case for AMEGs with another genomic context. However for CHG 2, AMEGs associated to these MGEs have a transfer likelihood which decreases with phylogenetic distance.

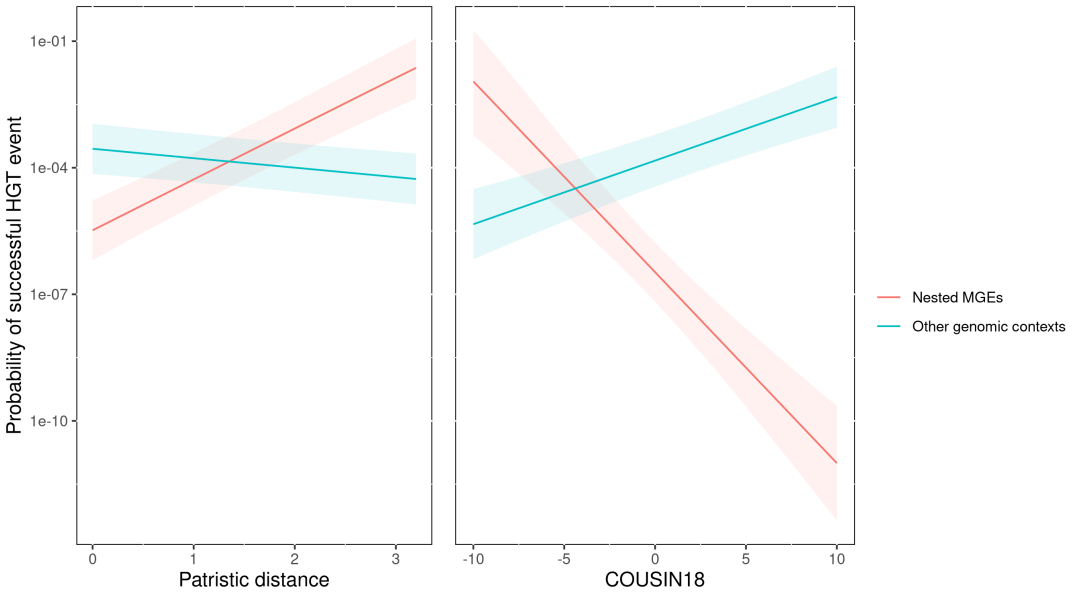


**Figure 5: The effect of nested MGEs on the probability of HGT events.** Curves and areas represent predictions and prediction intervals at the 95% level of the binomial regressions computed on HGT networks at the level of genera. AMEGs that are always associated with nested MGEs in a species are figured by the red curves, and AMEGs associated with other genomic contexts by the blue curves.

Second, regarding intragenomic mobility MGEs, their association to AMEGs decreases the likelihood of HGT events for the whole dataset (at the genera level) and CHG 13 (at the species level), but significantly increased it for CHG 1. For networks constructed at the genus level, intragenomic mobility MGEs also have a significantly positive interaction with phylogenetic distance for the whole dataset and CHG 1, but this interaction is negative for CHG 2. Third, nested MGEs have a negative impact on the likelihood of HGT events, but also a significant positive interaction with phylogenetic distance, and this for the whole dataset and for CHGs 2 and 13. Finally, nested MGEs

have a significant negative interaction with COUSIN for the whole dataset (see Fig. 5), but also for CHGs 1 and 13 at the genera level, i.e. AMEGs associated with nested MGEs are more likely to be transferred to taxa with very dissimilar codon uses.

No other effect was found to be significant. In particular, no significant effect on the likelihood of HGT events was found in models for CHG 20.

## 4. Discussion

In this study, we showed that 27,530 AMEGs out of 46,053 were likely to have been involved in recent, large-scale HGT events. We reconstructed directed networks of HGT events for 6 CHGs and, by analyzing these networks, we provided new insights into the identification of donors and recipients of antibiotic resistance and the factors which favor the emergence of donor-recipient relationships for ARGs.

### 4.1 Non-pathogenic bacteria have an underestimated role on the transmission of AMEs

Most of the putative horizontally transmitted AMEGs we identified were carried by human or animal pathogens and commensals, with highest prevalences in Campylobacteriota, Bacteroidota, and Firmicutes A. This is not surprising since horizontal transmission of ARGs to pathogens has been a huge concern during the recent decades (see e.g. Juhas, 2015). But this is also a direct consequence of a biased dataset, due to the oversampling of bacteria in hospitals and human or animal microbiota compared to other ecological contexts: 88% of bacteria screened by Pradier and Bedhomme (2022) for AMEGs were drawn from clinical, human, or farms samples. With the reconstruction of HGT networks, we were able to show that non-pathogenic bacteria actually have a crucial role in horizontal transmission of AMEGs. Indeed, HGT networks are heterophilic when considering pathogenicity towards humans. In particular, if HGT events occurred depending on the abundance of each bacteria type, as non-pathogenic bacteria represent only 21.8% of species considered in HGT networks, one might expect that only 0.218*0.782=17% of HGT events occur from non-pathogenic to pathogenic bacteria. Here, with 35% of HGT events (up to 57% for CHG 20), non-pathogenic to pathogenic HGT events are overrepresented. Moreover, highly-problematic pathogens have much lower out-degrees than non-pathogenic species.

This overall suggests that non-pathogenic bacteria are major providers of AMEGs to other taxa. For both CHGs 1 and 20, the only donors with null in-degrees are non-pathogenic species, with the exception of *Klebsiella quasipneumoniae* (for which the donor might be simply unresolved). These two CHGs were thus probably spread from non-pathogenic taxa, which passed them on (directly or indirectly) to pathogenic bacteria. Species of the genus *Streptomyces spp.* constitute a striking example in CHG 20, for which they are the most important donors by far. *Streptomyces* is thus both one of the genera (with *Micromonospora*) that naturally produce aminoglycosides (Durand

et al., 2019) and one of the genera from which resistance to these antibiotics is spread. This dual nature of antibiotic producers was already known, as numerous genes conferring resistance to various antibiotics (including AMEs) and encoded by Proteobacteria were found to result from ancient HGT events from *Streptomyces* (Jiang et al., 2017). Here we show that *Streptomyces* has still recently passed on genes to Proteobacteria, and that it has thus conferred aminoglycoside resistance to a wide range of pathogens.

For other CHGs, the HGT circulation patterns are quite different: for CHG 14 and 25, donors with null in-degrees are pathogens; for CHG 2, *Pseudomonas aeruginosa* (a pathogenic species) is a major donor, along with two non-pathogenic species, *Bosea sp900112185* and *Pseudomonas E protegens*; and the CHG 13 HGT networks presents numerous (pathogenic and non-pathogenic) donors with null in-degrees. Our approach does not allow to determine the origin of each CHG, i.e. in which species and which environment a gene conferring aminoglycoside resistance evolved but it is likely that non-pathogenic bacteria still were massive spreaders of these CHGs. Indeed, first of all, there is an overrepresentation of Pseudomonadaceae among donors, which are known to be an ubiquitous family Palleroni, 1981). Pseudomonadaceae live across a wide range of environments (Spiers et al., 2000), and donor Pseudomonadaceae are involved in a wide range of ecological interactions — from pathogens (e.g. *P. aeruginosa*) to symbionts (e.g. *Pseudomonas E protegens*, Ramette et al., 2011). Outside this family, we find either donor species that are also known to live in a wide range of natural ecosystems (e.g. *Sphingobium yanoikuyae*, Gu et al., 2009), or donor species that cannot infect humans and/or domestic animals (e.g. *Moraxella A osloensis* is a symbiont of a slug-parasitic nematode, An et al., 2008). Together, this suggests that AMEGs were spread from a wide range of contexts, and mostly from bacteria that do not (or rarely) infect humans.

On the contrary, highly-problematic human pathogens have the highest in-degrees: e.g. *Klebsiella pneumoniae* for CHG 13, *Salmonella enterica* for CHG 20, *P. aeruginosa* for CHG 2. In two cases only, we have identified highly-problematic human pathogens playing the role of hubs for the horizontal transmission of AMEGs: *P. aeruginosa* and *A. baumannii* have both high in- and out-degrees for CHG 1 and CHG 13. Finally, there is only one case, where a highly-problematic human pathogen is the only donor with null in-degree: *P. aeruginosa* is likely the original spreader of CHG 14.

Antibiotic resistance is known to be an ancient phenomenon, and D'Costa et al. (2011) even found AMEGs in 30,000-year-old Actinobacteriota genomes from the permafrost. AMEs had other biological functions before humans started manufacturing aminoglycosides, notably as a shield in the competition with aminoglycoside-producing bacteria. This competition only occurred in the ecological niche of *Streptomyces* and *Micromonospora*, so successful HGT events could only occur in these environments where AMEGs confer a competitive advantage: mostly in soil and decaying vegetation (Perry & Wright, 2013). It is thus likely that most ancient transfer events occurred from soil bacteria to ecological generalists, and that transmission from ecological generalists to human

pathogens only started recently. In this study, we show that transmission of AMEGs from non-pathogenic (including from soil bacteria) to pathogenic bacteria has also happened recently. Our findings are consistent with other studies which showed that soil and human microbiota shared ARGs with 100% identity (e.g. Forsberg et al., 2012), suggesting that the soil-to-pathogens pipeline is still active.

Yet, one question remains: why are HGT events from non-pathogenic to pathogenic bacteria overrepresented, and not also the other way around? Indeed, it is rare that two taxa exchange AMEGs in both directions, and only 9% of edges we constructed represented reciprocal HGT events at the species level (Tab. 2, Supplementary information, Tab. S27). The topology of HGT networks suggests that AMEGs were often originally spread from bacteria which are not human pathogens, so it may explain why one direction is preferred: from where the propagation of AMEGs emerged to where they likely provide the highest fitness benefits. But once pathogenic taxa have acquired AMEGs, they could pass them on back to non-pathogenic bacteria. AMEGs could provide substantial fitness benefits to non-pathogenic bacteria living in antibiotic-polluted ecosystems (Buelow et al., 2021), allow them to compete with antibiotic-producing bacteria (Martinez, 2012), or simply be fixed in the absence of selection (being co-selected with heavy metal resistance alleles, Pal et al., 2015; being silenced, Wein et al., 2019).

However, an unbalanced distribution of HGT facilitators between pathogenic and non-pathogenic bacteria could explain HGT events are not reciprocal. First, regarding genomic factors, genomes encoding proteins of nonhomologous end-joining (NHEJ) are much more likely to become the recipients of large-scale HGT events than those which are defective in this mechanism (Popa et al., 2011). NHEJ is a DNA repair mechanism that is able to ligate damaged DNA without homology (Shuman & Glickman, 2007), and therefore to capture exogenous DNA into the chromosome. Moreover, most known naturally-competent species are pathogenic (Johnsborg et al., 2007), which could explain why they are more prone than non-pathogenic ones to acquire AMEGs. This said, the association between pathogenicity and NHEJ proteins is not strict (some non-pathogenic species have NHEJ proteins, e.g. for *Streptomyces*, Hoff et al., 2016), such the scarcity of AMEG acquisitions by non-pathogenic bacteria could not be explained only by a scarcity of NHEJ proteins in their genomes. Second, ecological flows of AMEGs could be directed. For example, along a river, non-pathogenic bacteria would be much more frequent upstream of a wastewater treatment plant, and pathogenic bacteria downstream, so that HGT would be much more likely from the former to the latter.

Finally, a possible explanation for the low reciprocity of HGT in the networks we constructed could be a methodological limitation. On the one hand, human pathogens are overrepresented: 88% of bacteria screened by Pradier and Bedhomme (2022) were sampled in hospitals, humans, and farms; and nearly 50% of screened genomes belong to only 10 species of clinical interest. Thus, if non-pathogenic bacteria were rarely recipients of horizontally transferred AMEGs, these events could be even less represented in HGT networks that these taxa are poorly sampled. On the other hand, our

method to orient HGT events is based on the detection of the donor's genomic signature in the transferred gene. If an AMEG has spent little time in a lineage before it is transferred again, then our method will not be able to detect its donor and will thus simplify the transmission network by omitting the intermediate steps. This would be even worse in the case where AMEGs are transferred back to their taxa of origin, because the genomic signature of the original donor would be the same as the one of the recipient. This limitation could be neglected when focusing on ancient HGT, for which transferred genes have enough time to acquire new genomic signatures, but it is unavoidable for recent HGT events.

### 4.2 Mobile genetic elements interact with codon usage and phylogenetic distance to shape horizontal gene transfer

The horizontal transmission of AMEGs mostly occurs from non-pathogenic to pathogenic bacteria. It is thus crucial to understand the mechanisms which favor the occurrence of such events in order to better monitor them. In particular, phylogenetic relatedness and ecological proximity were repeatedly shown to facilitate HGT, and this study is no exception. Popa et al. (2011) showed that the prevalence of recent HGT events correlates to pairwise genome similarity (which does not, nevertheless, prevent the occurrence of intergeneric HGT events). Moreover, whereas within the same environmental type, the resistome composition of a community is highly correlated to its taxonomic composition (Forsberg et al., 2014), resistomes still tend to cluster by ecology (Gibson et al., 2014). Mostly, these barriers were discussed as the result of the large prevalence of MGE-mediated HGT. For example, conjugation requires the contact between donors and recipients, hence it is impossible between two bacteria living in distinct ecosystems. Additionally, MGEs have a limited range of possible hosts. It can vary a lot (from the species to the phylum) but it is always limited, and it is even argued that the risk of HGT could be predicted just from the host range of MGEs (e.g. host range of plasmids, Redondo-Salvo et al., 2020). However, in this study, the overrepresentation of Pseudomonadaceae in HGT networks suggests that the spread of AMEGs might also rely on species which can live in a wide range of environments.

Another factor that was suggested, though never tested on a large dataset, is CUB. Since a gene encoded with least preferred codons has lower chances of being translated efficiently and accurately (Walsh et al., 2020), recipients of a HGT events are less likely to undergo additional costs when the CUB of the transferred gene is similar to their own. Callens et al. (2021) found in *P. aeruginosa* that HT genes had a higher probability of being retained when they had similar CUB. The role of CUB incompatibilities was also supported by our models: HGT is more likely for positive values of COUSIN18.

However, despite HGT is overall more likely between phylogenetically, genomically, and ecologically close taxa, we reconstructed several inter-phyla HGT events that, though they were not a

majority (Fig. 3), had a crucial importance for the spread of ARGs: e.g., most AMEGs from CHG 20 were found in Proteobacteria, but they likely originated in the actinobacteria *Streptomyces*. Therefore, other factors that facilitate HGT are at stake.

In particular, AMEGs have a much higher likelihood to have been involved in HGT events, when they are associated with MGEs, with both intra- or intergenomic mobility. This had been repeatedly reported (Thomas & Nielsen, 2005). However, we also showed that association with nested MGEs, i.e. when an AMEG is associated with both types at the same time, increases the likelihood even more. The impact of nested MGEs on HGT was only highlighted recently: Che et al. (2021) showed computationally and experimentally that the association between conjugative plasmids and insertion sequences increased the horizontal transmission of ARGs.

Yet, nested MGEs could facilitate HGT in two cases, especially when compared to intragenomic mobility MGEs or intergenomic mobility MGEs alone. The first case corresponds to the expansion of the range of potential recipients for AMEGs by nested MGEs. Indeed, carriage by nested MGEs increases the chance for an AMEG of being transferred to more distantly-related recipients (Fig. 5). On the one hand, intragenomic mobility MGEs can mobilize new genes from the chromosome to plasmids, which can broaden their host range: Loftie-Eaton et al. (2016) experimentally observed that the host range of a plasmid had expanded following transposition of a toxin-antitoxin system. On the other hand, nested MGEs can be transferred by conjugation or transduction, whereas transposons and integrons alone can spread to other genomes by transformation only (Domingues et al., 2012). This not only offers two possible routes for HGT, but this also expands the range of potential recipients. Indeed, transformation requires the recipient to be naturally competent (which might not be so common, (Johnsborg et al., 2007): in our dataset, integrons and transposons had a positive impact on the likelihood of HGT events for CHG 1 only, for which more than 80% of carriers belong to 4 naturally-competent genera (*Acinetobacter*, *Pseudomonas*, *Pseudomonas A*, and *Moraxella*). Moreover, several families of naturally-competent bacteria have selection mechanisms and only DNA fragments displaying certain motifs can enter their cells (Mell & Redfield, 2014). Conjugation would allow nested MGEs to bypass the filter of selective uptake in these families.

A second case where nested MGEs can facilitate HGT is when conjugative elements trigger a cost to their carriers. Indeed, even acquisition costs appear to be very common (Prensky et al., 2021), and they globally decrease the chances of successful HGT: either the element is lost or the carrier goes extinct. However, transposons and integrons can quickly mobilize ARGs between plasmids and chromosomes. Thus, as suggested by Che et al. (2021), transitory colonization of recipient bacteria by plasmids could suffice to allow acquisition of new ARGs, without them to undergo the costs associated with these plasmids. This transitory role of plasmid in the interspecific propagation of transposon-carried genes has been demonstrated experimentally: Harrison et al. (2015) experimentally evolved bacteria carrying a plasmid containing a transposon-associated

mercury resistance gene and after several hundreds of generations, the transposon had integrated the chromosome and the plasmid was lost in multiple lineages.

Finally, our models showed that nested MGEs are able to bypass the barrier of CUB incompatibilities, and to have higher chances to transfer genes to taxa for which COUSIN18 is negative (Fig. 5). Even if, to the best of our knowledge, the potential cost of CUB incompatibilities was never studied for genes carried by nested MGEs, it is likely that this cost does not differ from other cases (e.g. with plasmid-carried genes, Amorós-Moya et al., 2010; Bedhomme et al., 2019). Then, why would HGT be favored in those cases? An interpretation could rely on the fact that the cost of CUB incompatibilities can originate from the slow translation of the received gene which gives low amounts of functional proteins (Bulmer, 1991). However, transposons with a copy-and-paste transposition can duplicate the resistance gene, so that the cost of low expression level can be ameliorated by the expression from several copies of the gene.

### 4.3 Conclusions and perspectives

In the present study, we showed that although AMEGs are more prevalent among human pathogens, they likely originated from a wide range of ecological contexts, and mostly from non-pathogenic taxa. This conclusion is already supported by various studies, but it allows us to rethink the role of natural ecosystems in the fight against antibiotic resistance. Antibiotic resistance in natural ecosystems has long been perceived at best as a byproduct of antibiotic pollution, or at worst as a reservoir for ARGs that can be transferred back to pathogenic bacteria (Marti et al., 2014; Wang et al., 2020). But natural ecosystems are also a major source of ARGs, and the management of antibiotic resistance must take into account the connectivity between ecosystems (Bengtsson-Palme et al., 2018). As shown by Pradier and Bedhomme (2022), soil and wildlife are at the center of AMEG exchanges between resistomes. Thus, fighting antibiotic resistance in healthcare will be insufficient if the flow of ARGs from natural to anthropized ecosystems is not put under control. This has already been argued by the advocates of the One Health approach (see e.g. Hernando-Amado et al., 2019, 2020).

However, while there are ecological barriers to HGT, ecological generalists such as *Pseudomonas* might be able to act as a shuttle for ARGs from the soil to the hospitals. The control of antibiotic resistance must therefore also tackle the issue of HGT facilitators. Here we highlighted the role of nested MGEs in increasing the rate at which AMEGs can be transferred. Mechanisms that were previously hypothesized to constitute barriers to HGT, such as codon usage incompatibilities (see e.g. Callens et al., 2021), can be overcome when HGT is mediated by nested MGEs. Further effort in the identification of MGEs associated with ARGs will be required in order to identify the extent of this phenomenon. But potential solutions could involve, rather than fighting antibiotic resistant bacteria, fighting the association between ARGs and MGEs (Kamruzzaman et al., 2017).

## 5. Bibliography


Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). Alignment-free $d_2^*$ oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Research*, *45*(1), 39–53. https://doi.org/10.1093/NAR/GKW1002

Amorós-Moya, D., Bedhomme, S., Hermann, M., & Bravo, I. G. (2010). Evolution in Regulatory Regions Rapidly Compensates the Cost of Nonoptimal Codon Usage. *Molecular Biology and Evolution*, *27*(9), 2141–2151. https://doi.org/10.1093/MOLBEV/MSQ103

An, R., Sreevatsan, S., & Grewal, P. S. (2008). Moraxella osloensis gene expression in the slug host Deroceras reticulatum. *BMC Microbiology*, *8*(1), 1–11. https://doi.org/10.1186/1471-2180-8-19/TABLES/3

Andam, C. P., & Gogarten, J. P. (2011). Biased gene transfer in microbial evolution. *Nature Reviews Microbiology 2011 9:7*, *9*(7), 543–555. https://doi.org/10.1038/nrmicro2593

Baltrus, D. A. (2013). Exploring the costs of horizontal gene transfer. *Trends in Ecology and Evolution*, *28*(8), 489–495. https://doi.org/10.1016/J.TREE.2013.04.002

Bedhomme, S., Amorós-Moya, D., Valero, L. M., Bonifaci, N., Pujana, M. À., Bravo, I. G., & Gonzalez, J. (2019). Evolutionary Changes after Translational Challenges Imposed by Horizontal Gene Transfer. *Genome Biology and Evolution*, *11*(3), 814–831. https://doi.org/10.1093/gbe/evz031

Bengtsson-Palme, J., Kristiansson, E., & Larsson, D. G. J. (2018). Environmental factors influencing the development and spread of antibiotic resistance. *FEMS Microbiology Reviews*, *42*(1), 68–80. https://doi.org/10.1093/femsre/fux053

Bernard, G., Chan, C. X., & Ragan, M. A. (2016). Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Scientific Reports*, *6*(1), 1–12. https://doi.org/10.1038/srep28970

Botzman, M., & Margalit, H. (2011). Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biology*, *12*(10), 1–11. https://doi.org/10.1186/GB-2011-12-10-R109/TABLES/2

Bourret, J., Alizon, S., & Bravo, I. G. (2019). COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. In *Genome biology and evolution* (Vol. 11, Issue 12, pp. 3523–3528). NLM (Medline). https://doi.org/10.1093/gbe/evz262

Buelow, E., Ploy, M. C., & Dagot, C. (2021). Role of pollution on the selection of antibiotic resistance and bacterial pathogens in the environment. *Current Opinion in Microbiology*, *64*, 117–124. https://doi.org/10.1016/J.MIB.2021.10.005

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics*, *129*(3), 897–907. https://doi.org/10.1093/GENETICS/129.3.897

Cairns, J., Ruokolainen, L., Hultman, J., Tamminen, M., Virta, M., & Hiltunen, T. (2018). Ecology determines how low antibiotic concentration impacts community composition and horizontal transfer of resistance genes. *Communications Biology 2018 1:1*, *1*(1), 1–8. https://doi.org/10.1038/s42003-018-0041-7

Callens, M., Scornavacca, C., & Bedhomme, S. (2021). Evolutionary responses to codon usage of horizontally transferred genes in Pseudomonas aeruginosa: gene retention, amelioration and compensatory evolution. *Microbial Genomics*, *7*(6). https://doi.org/10.1099/MGEN.0.000587

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925–1927. https://doi.org/10.1093/BIOINFORMATICS/BTZ848

Che, Y., Yang, Y., Xu, X., Brinda, K., Polz, M. F., Hanage, W. P., & Zhang, T. (2021). Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(6). https://doi.org/10.1073/PNAS.2008731118/-/DCSUPPLEMENTAL

Cong, Y., Chan, Y. B., & Ragan, M. A. (2016). A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. *Scientific Reports*, *6*(1), 1–13. https://doi.org/10.1038/srep30308

Cooper, R. M., Tsimring, L., & Hasty, J. (2017). Inter-species population dynamics enhance microbial horizontal gene transfer and spread of antibiotic resistance. *ELife*, *6*. https://doi.org/10.7554/ELIFE.25950

Corel, E., Méheust, R., Watson, A. K., Mcinerney, J. O., Lopez, P., & Bapteste, E. (2018). Bipartite Network Analysis of Gene Sharings in the Microbial World. *Molecular Biology and Evolution*, *35*(4), 899–913. https://doi.org/10.1093/MOLBEV/MSY001

D'Costa, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W. L., Schwarz, C., Froese, D., Zazula, G., Calmels, F., Debruyne, R., Golding, G. B., Poinar, H. N., & Wright, G. D. (2011). Antibiotic resistance is ancient. In *Nature* (Vol. 477, Issue 7365, pp. 457–461). https://doi.org/10.1038/nature10388

Domingues, S., Harms, K., Fricke, W. F., Johnsen, P. J., da Silva, G. J., & Nielsen, K. M. (2012). Natural Transformation Facilitates Transfer of Transposons, Integrons and Gene Cassettes between Bacterial Species. *PLOS Pathogens*, *8*(8), e1002837. https://doi.org/10.1371/JOURNAL.PPAT.1002837

Durand, G. A., Raoult, D., & Dubourg, G. (2019). Antibiotic discovery: history, methods and perspectives. *International Journal of Antimicrobial Agents*, *53*(4), 371–382. https://doi.org/10.1016/J.IJANTIMICAG.2018.11.010

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/JOURNAL.PCBI.1002195

European Medicines Agency. (2017). *Sales of veterinary antimicrobial agents in 31 European countries in 2017*.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, *61*(1), 1–10. https://doi.org/10.1016/0006-3207(92)91201-3

Fan, H., Ives, A. R., Surget-Groba, Y., & Cannon, C. H. (2015). An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics*, *16*(1), 1–18. https://doi.org/10.1186/S12864-015-1647-5/FIGURES/7

Forsberg, K. J., Patel, S., Gibson, M. K., Lauber, C. L., Knight, R., Fierer, N., & Dantas, G. (2014).

Bacterial phylogeny structures soil resistomes across habitats. *Nature*, *509*(7502), 612–616. https://doi.org/10.1038/nature13377

Forsberg, K. J., Reyes, A., Wang, B., Selleck, E. M., Sommer, M. O. A., & Dantas, G. (2012). The shared antibiotic resistome of soil bacteria and human pathogens. *Science*, *337*(6098), 1107–1111. https://doi.org/10.1126/SCIENCE.1220761

Gal-Mor, O., & Finlay, B. B. (2006). Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cellular Microbiology*, *8*(11), 1707–1719. https://doi.org/10.1111/J.1462-5822.2006.00794.X

Garneau-Tsodikova, S., & Labby, K. J. (2016). Mechanisms of Resistance to Aminoglycoside Antibiotics: Overview and Perspectives. *MedChemComm*, *7*(1), 11. https://doi.org/10.1039/C5MD00344J

Gibson, M. K., Forsberg, K. J., & Dantas, G. (2014). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME Journal*, *9*(1), 207–216. https://doi.org/10.1038/ismej.2014.106

Gu, J. G., Han, B., Duan, S., Zhao, Z., & Wang, Y. (2009). Degradation of the endocrine-disrupting dimethyl phthalate carboxylic ester by Sphingomonas yanoikuyae DOS01 isolated from the South China Sea and the biochemical pathway. *International Biodeterioration & Biodegradation*, *63*(4), 450–455. https://doi.org/10.1016/J.IBIOD.2008.12.004

Harrison, E., Guymer, D., Spiers, A. J., Paterson, S., & Brockhurst, M. A. (2015). Parallel Compensatory Evolution Stabilizes Plasmids across the Parasitism-Mutualism Continuum. *Current Biology*, *25*(15), 2034–2039. https://doi.org/10.1016/J.CUB.2015.06.024

Hernando-Amado, S., Coque, T. M., Baquero, F., & Martínez, J. L. (2019). Defining and combating antibiotic resistance from One Health and Global Health perspectives. *Nature Microbiology*, *4*(9), 1432–1442. https://doi.org/10.1038/s41564-019-0503-9

Hernando-Amado, S., Coque, T. M., Baquero, F., & Martínez, J. L. (2020). Antibiotic Resistance: Moving From Individual Health Norms to Social Norms in One Health and Global Health. *Frontiers in Microbiology*, *11*, 1914. https://doi.org/10.3389/FMICB.2020.01914/BIBTEX

Hoff, G., Bertrand, C., Zhang, L., Piotrowski, E., Chipot, L., Bontemps, C., Confalonieri, F., McGovern, S., Lecointe, F., Thibessard, A., & Leblond, P. (2016). Multiple and variable NHEJ-like genes are involved in resistance to DNA damage in Streptomyces ambofaciens. *Frontiers in Microbiology*, *7*(NOV), 1901. https://doi.org/10.3389/FMICB.2016.01901/BIBTEX

Holden, M. T. G., Feil, E. J., Lindsay, J. A., Peacock, S. J., Day, N. P. J., Enright, M. C., Foster, T. J., Moore, C. E., Hurst, L., Atkin, R., Barron, A., Bason, N., Bentley, S. D., Chillingworth, C., Chillingworth, T., Churcher, C., Clark, L., Corton, C., Cronin, A., … Parkhill, J. (2004). Complete genomes of two clinical Staphylococcus aureus strains: Evidence for the rapid evolution of virulence and drug resistance. *Proceedings of the National Academy of Sciences*, *101*(26), 9786–9791. https://doi.org/10.1073/PNAS.0402521101

Jiang, X., Ellabaan, M. M. H., Charusanti, P., Munck, C., Blin, K., Tong, Y., Weber, T., Sommer, M. O. A., & Lee, S. Y. (2017). Dissemination of antibiotic resistance genes from antibiotic producers to pathogens. *Nature Communications*, *8*(1), 1–7. https://doi.org/10.1038/ncomms15784

Johnsborg, O., Eldholm, V., & Håvarstein, L. S. (2007). Natural genetic transformation: prevalence, mechanisms and function. *Research in Microbiology*, *158*(10), 767–778. https://doi.org/10.1016/J.RESMIC.2007.09.004

Juhas, M. (2015). Horizontal gene transfer in human pathogens. *Critical Reviews in Microbiology*, *41*(1), 101–108. https://doi.org/10.3109/1040841X.2013.804031

Kamruzzaman, M., Shoma, S., Thomas, C. M., Partridge, S. R., & Iredell, J. R. (2017). Plasmid interference for curing antibiotic resistance plasmids in vivo. *PLOS ONE*, *12*(2), e0172913. https://doi.org/10.1371/JOURNAL.PONE.0172913

Koonin, E. V. (2016). Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research*, *5*, 1805. https://doi.org/10.12688/f1000research.8737.1

Li, B., Qiu, Y., Song, Y., Lin, H., & Yin, H. (2019). Dissecting horizontal and vertical gene transfer of antibiotic resistance plasmid in bacterial community using microfluidics. *Environment International*, *131*, 105007. https://doi.org/10.1016/J.ENVINT.2019.105007

Li, Q., Chang, W., Zhang, H., Hu, D., & Wang, X. (2019). The role of plasmids in the multiple antibiotic resistance transfer in ESBLs-producing Escherichia coli isolated from wastewater treatment plants. *Frontiers in Microbiology*, *10*(APR), 633. https://doi.org/10.3389/FMICB.2019.00633/BIBTEX

Loftie-Eaton, W., Yano, H., Burleigh, S., Simmons, R. S., Hughes, J. M., Rogers, L. M., Hunter, S. S., Settles, M. L., Forney, L. J., Ponciano, J. M., & Top, E. M. (2016). Evolutionary Paths That Expand Plasmid Host-Range: Implications for Spread of Antibiotic Resistance. *Molecular Biology and Evolution*, *33*(4), 885–897. https://doi.org/10.1093/MOLBEV/MSV339

Lu, Y. Y., Tang, K., Ren, J., Fuhrman, J. A., Waterman, M. S., & Sun, F. (2017). CAFE: aCcelerated Alignment-FrEe sequence analysis. *Nucleic Acids Research*, *45*(W1), W554–W559. https://doi.org/10.1093/NAR/GKX351

Makarenkov, V., Kevorkov, D., & Legendre, P. (2006). Phylogenetic Network Construction Approaches. *Applied Mycology and Biotechnology*, *6*(C), 61–97. https://doi.org/10.1016/S1874-5334(06)80006-7

Marijuán, P. C., Navarro, J., & del Moral, R. (2010). On prokaryotic intelligence: Strategies for sensing the environment. *Biosystems*, *99*(2), 94–103. https://doi.org/10.1016/J.BIOSYSTEMS.2009.09.004

Marti, E., Variatza, E., & Luis Balcazar, J. (2014). The role of aquatic ecosystems as reservoirs of antibiotic resistance. *Trends in Microbiology*, *22*(1), 36–41. https://doi.org/10.1016/j.tim.2013.11.001

Martinez, J. L. (2012). Natural Antibiotic Resistance and Contamination by Antibiotic Resistance Determinants: The Two Ages in the Evolution of Resistance to Antimicrobials. *Frontiers in Microbiology*, *0*(JAN), 1. https://doi.org/10.3389/FMICB.2012.00001

Mell, J. C., & Redfield, R. J. (2014). Natural competence and the evolution of DNA uptake specificity. *Journal of Bacteriology*, *196*(8), 1471–1483. https://doi.org/10.1128/JB.01293-13/ASSET/8A95EAED-2C41-4B1F-9E9D-

CEBE7D72B5D0/ASSETS/GRAPHIC/ZJB9990931010006.JPEG

Miele, V., Penel, S., & Duret, L. (2011). Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics 2011 12:1*, *12*(1), 1–9. https://doi.org/10.1186/1471-2105-12-116

Mingeot-Leclercq, M.-P., Glupczynski, Y., & Tulkens, P. M. (1999). Aminoglycosides: Activity and Resistance. *Antimicrobial Agents and Chemotherapy*, *43*(4), 727. /pmc/articles/PMC89199/

National Institute of Allergy and Infectious Diseases. (2018). *NIAID Emerging Infectious Diseases/Pathogens*. https://www.niaid.nih.gov/research/emerging-infectious-diseases-pathogens

Pal, C., Bengtsson-Palme, J., Kristiansson, E., & Larsson, D. G. J. (2015). Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. *BMC Genomics 2015 16:1*, *16*(1), 1–14. https://doi.org/10.1186/S12864-015-2153-5

Palleroni, N. J. (1981). Introduction to the Family Pseudomonadaceae. In M. P. Starr, H. Stolp, H. G. Trüper, A. Balows, & H. G. Schlegel (Eds.), *The Prokaryotes* (pp. 655–665). Springer. https://doi.org/10.1007/978-3-662-13187-9_58

Park, C., & Zhang, J. (2012). High Expression Hampers Horizontal Gene Transfer. *Genome Biology and Evolution*, *4*(4), 523–532. https://doi.org/10.1093/GBE/EVS030

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

Parks, D. H., 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

Perry, J. A., & Wright, G. D. (2013). The antibiotic resistance "mobilome": Searching for the link between environment and clinic. *Frontiers in Microbiology*, *4*(MAY), 138. https://doi.org/10.3389/FMICB.2013.00138/BIBTEX

Popa, O., Hazkani-Covo, E., Landan, G., Martin, W., & Dagan, T. (2011). Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Research*, *21*(4), 599–609. https://doi.org/10.1101/GR.115592.110

Popa, O., Landan, G., & Dagan, T. (2016). Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. *The ISME Journal 2017 11:2*, *11*(2), 543–554. https://doi.org/10.1038/ismej.2016.116

Pradier, L., & Bedhomme, S. (2022). Ecology, more than antibiotics consumption, is the major predictor for the global distribution of aminoglycoside-modifying enzymes. *BioRxiv*, 2022.01.07.475340. https://doi.org/10.1101/2022.01.07.475340

Prensky, H., Gomez-Simmonds, A., Uhlemann, A.-C., & Lopatkin, A. J. (2021). Conjugation dynamics depend on both the plasmid acquisition cost and the fitness cost. *Molecular Systems Biology*, *17*(3), e9913. https://doi.org/10.15252/MSB.20209913

Ramette, A., Frapolli, M., Saux, M. F. Le, Gruffaz, C., Meyer, J. M., Défago, G., Sutra, L., & Moënne-Loccoz, Y. (2011). Pseudomonas protegens sp. nov., widespread plant-protecting bacteria producing the biocontrol compounds 2,4-diacetylphloroglucinol and pyoluteorin. *Systematic and Applied Microbiology*, *34*(3), 180–188. https://doi.org/10.1016/J.SYAPM.2010.10.005

Ramirez, M. S., & Tolmasky, M. E. (2010). Aminoglycoside modifying enzymes. *Drug Resistance Updates*, *13*(6), 151–171. https://doi.org/10.1016/J.DRUP.2010.08.003

Redondo-Salvo, S., Fernández-López, R., Ruiz, R., Vielva, L., de Toro, M., Rocha, E. P. C., Garcillán-Barcia, M. P., & de la Cruz, F. (2020). Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nature Communications 2020 11:1*, *11*(1), 1–13. https://doi.org/10.1038/s41467-020-17278-2

Reinert, G., Chew, D., Sun, F., & Waterman, M. S. (2009). Alignment-Free Sequence Comparison (I): Statistics and Power. *Journal of Computational Biology*, *16*(12), 1615–1634. www.liebertonline.com

Rousset, F., & Ferdy, J.-B. (2014). Testing environmental and genetic effects in the presence of spatial autocorrelation. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography*, *37*(8), 781–790. https://doi.org/10.1111/ecog.00566ï

San Millan, A., Toll-Riera, M., Qi, Q., & MacLean, R. C. (2015). Interactions between horizontally acquired genes create a fitness cost in Pseudomonas aeruginosa. *Nature Communications*, *6*(1), 1–8. https://doi.org/10.1038/ncomms7845

Schaack, S., Gilbert, C., & Feschotte, C. (2010). Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution*, *25*(9), 537–546. https://doi.org/10.1016/J.TREE.2010.06.001

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Shuman, S., & Glickman, M. S. (2007). Bacterial DNA repair by non-homologous end joining. *Nature Reviews Microbiology*, *5*(11), 852–861. https://doi.org/10.1038/nrmicro1768

Sievers, F., & Higgins, D. G. (2014). Clustal Omega. *Current Protocols in Bioinformatics*, *48*(1), 3.13.1-3.13.16. https://doi.org/10.1002/0471250953.BI0313S48

Sørensen, S. J., Bailey, M., Hansen, L. H., Kroer, N., & Wuertz, S. (2005). Studying plasmid horizontal transfer in situ: a critical review. *Nature Reviews Microbiology*, *3*(9), 700–710. https://doi.org/10.1038/nrmicro1232

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, *16*(8), 472–482. https://doi.org/10.1038/nrg3962

Spielman, S. J., & Wilke, C. O. (2015). Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. *PLOS ONE*, *10*(9), e0139047. https://doi.org/10.1371/JOURNAL.PONE.0139047

Spiers, A. J., Buckling, A., & Rainey, P. B. (2000). The causes of Pseudomonas diversity. *Microbiology*, *146*(10), 2345–2350. https://doi.org/10.1099/00221287-146-10-2345/CITE/REFWORKS

Tang, K., Lu, Y. Y., & Sun, F. (2018). Background adjusted alignment-free dissimilarity measures improve the detection of horizontal gene transfer. *Frontiers in Microbiology*, *9*(APR), 711. https://doi.org/10.3389/FMICB.2018.00711/BIBTEX

Thomas, C. M., & Nielsen, K. M. (2005). Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nature Reviews Microbiology*, *3*(9), 711–721. https://doi.org/10.1038/nrmicro1234

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., De Silva, N., Martinez, M. C., Pedro, H., Yates, A. D., Hassani-Pak, K., & Hammond-Kosack, K. E. (2020). PHI-base: the pathogen–host interactions database. *Nucleic Acids Research*, *48*(D1), D613–D620. https://doi.org/10.1093/NAR/GKZ904

Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A., & Clark, P. L. (2020). Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(7), 3528–3534. https://doi.org/10.1073/PNAS.1907126117/-/DCSUPPLEMENTAL

Wang, Z., Chen, Q., Zhang, J., Guan, T., Chen, Y., & Shi, W. (2020). Critical roles of cyanobacteria as reservoir and source for antibiotic resistance genes. *Environment International*, *144*, 106034. https://doi.org/10.1016/J.ENVINT.2020.106034

Wardeh, M., Risley, C., Mcintyre, M. K., Setzkorn, C., & Baylis, M. (2015). Database of host-pathogen and related species interactions, and their global distribution. *Scientific Data*, *2*(1), 1–11. https://doi.org/10.1038/sdata.2015.49

Wein, T., Hülter, N. F., Mizrahi, I., & Dagan, T. (2019). Emergence of plasmid stability under non-selective conditions maintains antibiotic resistance. *Nature Communications*, *10*(1), 1–13. https://doi.org/10.1038/s41467-019-10600-7

World Health Organization. (2017). *WHO publishes list of bacteria for which new antibiotics are urgently needed*. https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed

# Section VI : Discussion

Les travaux exposés dans cette thèse s'appuient sur une approche écologique et évolutive pour étudier les facteurs déterminant la propagation des résistances antibiotiques via l'exemple des AMEGs. En particulier, deux types de facteurs ont été abordés, et leurs résultats seront discutés dans cette section : d'une part, les paramètres environnementaux qui peuvent amener à la sélection et/ou à la circulation des gènes et bactéries résistantes ; d'autre part, les paramètres phylogénétiques et génomiques qui favorisent la circulation des gènes de résistances d'une bactérie à l'autre, voire d'un taxon à l'autre. Ensemble, l'étude de ces deux catégories de facteurs permet de dessiner de nouveaux angles d'attaque dans le cadre de l'étude et de la gestion des résistances antibiotiques. Enfin, les différentes limitations de cette thèse et de possibles travaux de recherche dans sa continuité seront exposés en perspective.

## 4 Structuration environnementale de la résistance

### 4.1 Les antibiotiques anthropogéniques parmi d'autres forces

La sélection est la force évolutive sur laquelle la littérature, que ce soit en biologie évolutive ou en santé publique, s'est le plus concentrée jusqu'à présent pour étudier la résistance antibiotique. Cette tendance repose sur trois postulats. *(1)* Chaque gène de résistance confère à lui seul un phénotype résistant à un antibiotique, si ce n'est plusieurs. *(2)* Les antibiotiques, dans les concentrations auxquelles les bactéries sont exposées, sont suffisants pour conférer un avantage sélectif aux bactéries résistantes. *(3)* En l'absence d'antibiotique (ou lorsque ces derniers sont en concentration insuffisante), le coût de la résistance est trop important pour que les bactéries résistantes restent compétitives par rapport aux génotypes sensibles.

Les données de la section IV permettent de reconsidérer certains de ces postulats. En effet, depuis 1943 (année où la streptomycine a commencé à être testée), la prévalence des bactéries portant au moins un gène encodant une AME a très fortement augmenté, passant de rares jusqu'à atteindre quasiment la moitié des génomes séquencés. Ceci soutient effectivement l'idée que l'utilisation des aminoglycosides par l'Homme est une condition suffisante à la sélection des AME, et donc au succès de leur propagation. Par la suite, depuis les années 1990, la prévalence des génomes encodant des AME tend à plafonner autour de 30 %. Différentes interprétations complémentaires entres elles, peuvent être données à ce résultat. En premier lieu, la stabilisation des prévalences pourrait résulter de la stabilisation des consommations d'aminoglycosides dans les pays industrialisés (dans la mesure où l'échantillonnage a été plus important dans ces régions). Notamment, à partir des années 1980, les aminoglycosides ont de plus en plus été remplacées par d'autres catégories d'antibiotiques à spectre plus large et à toxicité moindre (Krause et al. 2016). Toutefois, comme montré par les modèles dans la section IV, les variations de consommation d'aminoglycosides d'un pays européen à l'autre ne semblent expliquer que peu ou

pas les variations de prévalence de bactéries résistantes. Un phénomène similaire avait été observé en ce qui concerne la résistance aux sulfonamides, dont la prévalence n'avait pas diminué dans les années 1990 au Royaume-Uni malgré une baisse drastique des prescriptions (Enne et al. 2004). Une seconde interprétation, pourrait donc être que les aminoglycosides se sont accumulés dans différents environnements, dans lesquels des réservoirs de bactéries résistantes auraient pu être sélectionnées et qui seraient susceptibles de ré-ensemencer les hôpitaux. Cependant, bien qu'ils ne soient pas actuellement considérés comme biodégradables (au contraire des sulfonamides par exemple), les aminoglycosides se dégradent rapidement à température ambiante en présence de glucose ou par photolyse en association à la matière organique (Reis et al. 2020), et sont donc peu susceptibles de sélectionner des bactéries sur une longue période de temps. Enfin, une dernière interprétation repose sur la fin de la commercialisation de nouveaux aminoglycosides. En effet, aucune nouvelle aminoglycoside n'a été mise sur le marché entre l'amikacine en 1981 et la plazomicine en 2018 (qui n'est d'ailleurs indiquée que pour les infections urinaires en traitement de seconde intention, Serio et al. 2018). Il est donc possible qu'aucune nouvelle pression de sélection ne soit venue renforcer le réservoir de bactéries résistantes.

Dans l'ensemble, s'il est aisé de comprendre pourquoi la prévalence des AME a cessé d'augmenter, il peut être plus difficile de comprendre pourquoi cette résistance reste toujours aussi prévalente. Des modèles théoriques ont suggéré que même de très faibles consommations d'antibiotiques peuvent avoir des effets positifs sur la prévalence des résistances (Aldrin et al. 2013), et qu'un traitement à faible dose peut augmenter davantage le risque de résistance que de fortes doses (Lipsitch 2001). Toutefois, on peut douter que ces phénomènes soient impliqués dans un maintien des bactéries résistantes à des prévalences si élevées, et il pourrait être plus productif de chercher à resituer le rôle des antibiotiques dans le maintien et la propagation de la résistance. Notamment, si on sait qu'un gène de résistance peut être coûteux en l'absence de pression de sélection (Andersson, Maisnier Patin et al. 2007), il est beaucoup plus difficile d'affirmer que les bactéries résistantes s'éteindront ou perdront leur résistance, au moins à court terme (Andersson et Hughes 2010). En effet, les gènes de résistance ne sont pas tous coûteux (Björkman et al. 1998). De plus, comme déjà discuté dans la section IV, il existe d'autres pressions de sélection favorisant le maintien des gènes de résistance : que ce soient les antibiotiques produits dans les environnements naturels (Clardy et al. 2009), ou la co-sélection de résistances aux métaux lourds (Baker-Austin et al. 2006). Enfin, une évolution compensatoire peut réduire (voire supprimer) le coût des gènes de résistance, par exemple en réduisant l'expression de ces gènes en l'absence de pression (Wein et al. 2019). L'évolution compensatoire peut même, tout en réduisant le coût de la résistance, augmenter l'intensité de la résistance elle-même (Lindgren et al. 2005).

Enfin, la section IV montre également que les dynamiques temporelles des différents clusters d'AME peuvent varier fortement d'une région à l'autre. Les pressions de sélection s'exerçant sur les gènes de résistance sont donc probablement distribuées de manière hétérogène en fonction des continents et des écosystèmes. Une baisse de la prévalence de la résistance dans une région, liée à des pressions de sélection plus faibles, pourrait donc être également compensée par un afflux de bactéries migrant

depuis des communautés où la résistance est adaptative.

## 4.2 La connectivité écologique et géographique

De fait, la plupart des clusters d'AME étudiés dans le cadre de cette thèse ont été détectés dans un très grand nombre de biomes (tant naturels qu'anthropisés), et sur plusieurs continents. Ceci témoigne de l'existence de transferts de gènes de résistances et/ou de bactéries résistantes d'un biome à l'autre et d'un continent à l'autre. Il est donc crucial de comprendre les modalités de cette connectivité, de sorte à mettre en place des stratégies de confinement.

### 4.2.1 Mondialisation et connectivité géographique

Si, comme suggéré par la section IV, la composition des résistomes est peu structurée géographiquement, cela signifie que la connectivité géographique des résistomes est très forte. Au grain spatial utilisé (celui de très larges régions), deux facteurs peuvent être impliqués pour expliquer une telle connectivité. D'une part, il a été suggéré que des phénomènes climatiques majeurs, tels que le courant El Niño, pourraient transporter des organismes pathogènes d'un continent à l'autre (Martinez-Urtaza et al. 2016). Toutefois, ce type d'événement est localisé dans le temps et dans l'espace, et certains continents ne seraient connectés entre eux que de manière très indirecte. D'autre part, la mondialisation des échanges a des effets observés sur la propagation des bactéries résistantes. Des épidémies d'infection à bactéries résistantes ont pu être retracées à des échanges internationaux, tel que le commerce de volailles qui a propagé une souche de *S. enterica* résistante à la ciprofloxacine au travers des Etats-Unis et de l'Europe (Le Hello et al. 2011) et un plasmide conférant une résistance à la colistine au travers de l'Europe (Skov et Monnet 2016). L'importance des échanges humains intercontinentaux dans la propagation des résistances est confirmée par la section IV, où la migration et le commerce de marchandises expliquent tous deux une part non négligeable de la prévalence de la plupart des clusters d'AME (figure 6).

En revanche, au regard des travaux présentés ici et de la littérature, il reste encore particulièrement compliqué de trouver une (ou plusieurs) voie(s) préférentielle(s) pour la circulation intercontinentale de la résistance. Dans la section IV, aucune des différentes catégories de marchandises ou la migration n'explique mieux que les autres la propagation de l'ensemble des AME. Même lorsque la répartition de certains clusters est bien expliquée par l'importation d'un faible nombre de catégories, les types de biens impliqués ne sont pas les mêmes d'un cluster à l'autre. De même, bien que nombreux, les exemples de gènes ou de souches dont la circulation a été reliée à un mode de transport identifié de manière précise représentent une faible proportion de l'ensemble des résistances connues. Ces dernières années, les aliments (d'origine animale ou végétale) ont été particulièrement étudiés comme de potentiels vecteurs de résistance (George 2018). Cependant, la section IV met en évidence que d'autres types de marchandises pourraient être surveillées : par exemple, la propagation du cluster 11 d'AME est principalement corrélée aux importations de carburants. A ce stade, il ne s'agit que d'une corrélation, et

seuls des prélèvements directs permettront de confirmer si ces marchandises (ou leur conditionnement pour le transport, par exemple) sont susceptibles de porter et transporter des bactéries ou des gènes.

### 4.2.2 Flux naturels et anthropiques entre écosystèmes

Là où on peut voir une forte diffusion géographique des résistomes, la section IV montre cependant qu'il existe des similarités de résistomes entre écosystèmes ce qui suggère qu'il existe également une connectivitée écologique. Dans l'ensemble, les travaux de cette thèse remettent en question l'idée selon laquelle la résistance aux antibiotiques émerge principalement dans les environnements cliniques. D'une part, section IV montre que les biomes centraux dans le réseau connectant les résistomes sont le sol, la faune et la flore, et l'humain. D'autre part, la section V montre que les espèces ou genres identifiés comme propagateurs des clusters d'AME sont associés à des contextes écologiques variés : par exemple, le sol et l'humus pour *Streptomyces*, ou encore comme symbiotes de nématodes parasites de gastéropodes pour *Moraxella A osloensis*.

La mise en contact de ces contextes écologiques avec l'environnement humain nécessite donc des transferts de matière. Si les bactéries peuvent être naturellement disséminées (par exemple par les courants aériens, (Allen et al. 2010), il est clair que l'impact des activités humaines sur la connectivité entre écosystèmes est à interroger. En particulier, la colonisation d'espaces toujours croissants, via l'urbanisation ou le défrichement, peut mettre en contact les bactéries de l'environnement humain avec celles des environnements naturels. Il a été montré que ce type de route peut être emprunté par les maladies émergentes (Hassell et al. 2017), et il ne serait pas surprenant qu'il en soit de même pour les gènes de résistances. En outre, cette colonisation tend également à homogénéiser les paramètres des écosystèmes dans lesquelles les bactéries résistantes vivent (voir par exemple pour le sol, Herrmann et al. 2020). Il existe des preuves que les bactéries s'adaptent à ces nouvelles conditions environnementales (Epp Schmidt, Kotze et al. 2019) mais, contrairement aux archées et aux champignons, à l'heure actuelle il ne semble pas que les communautés bactériennes convergent (Epp Schmidt, Pouyat et al. 2017). Ces mises en contact peuvent ainsi favoriser le transfert des résistances des milieux naturels vers l'environnement humain.

A l'inverse, la résistance peut passer des environnements anthropisés vers les milieux naturels. Dans ce cadre, il semble que le phénomène concerne principalement le transport de bactéries par l'eau. Mais les activités humaines ou leur organisation peuvent, là aussi, fortement faciliter la propagation des résistances. En milieu agricole, du fait de l'amendement des sols agricoles au fumier, des bactéries fécales sont ensuite être transportées par le ruissellement des eaux de pluie (Barrios et al. 2020). En milieu urbain, le ruissellement crée également des bassins de forte concentration en bactéries résistantes (Almakki et al. 2019), qui sont ensuite être déversés vers d'autres écosystèmes lors d'événements de plus grand ampleur (Lee, Suits et al. 2020).

### 4.2.3 La modulation écologique du transfert horizontal de gènes

Les résultats de la section IV suggèrent qu'il existe une forte connectivité entre biomes, mais ceux de la section V indiquent également que le transfert horizontal est plus probable entre espèces dont les niches se recouvrent. En effet, il est plus facile d'échanger du matériel génétique lorsque le donneur et le receveur sont physiquement proches. Cette proximité physique est obligatoire pour les transferts par vecteurs conjugatifs et transductifs. Cependant, il est également à noter que des gènes de résistance peuvent se disséminer à plus grande distance librement d'une bactérie à l'autre sous forme d'ADN libre par transformation. En effet, à l'extérieur d'une bactérie, le matériel génétique peut être protégé de la dégradation par son association à des macromolécules (polysaccharides, He, Zhou et al. 2019, biofilms, Guo, Yang et al. 2018). Ainsi, certains gènes de résistance peuvent rester intacts pendant plusieurs années dans les sols (Pietramellara et al. 2009). Il est donc possible que le HGT se produise entre deux espèces ne partageant pas du tout la même niche. Cependant, la transformation nécéssite la compétence naturelle du receveur, ce qui n'est pas le cas de la plupart des espèces (Attaiech et al. 2016). De plus, même si de nombreux systèmes bactériens semblent avoir un mécanisme d'importation d'ADN, ce dernier est finement régulé et même « silencer » en permanence chez certaines espèces. Il semble donc légitime de conclure que le HGT dépend principalement de vecteurs conjugatifs ou transductifs.

Toutefois, la proximité des niches peut également moduler le HGT d'autres manières. En effet, si deux espèces partagent la même niche, elles tendent à entrer en compétition pour l'exploitation des mêmes ressources, ce qui peut entraîner l'exclusion de certaines souches. Cela n'empêche pas le HGT par transformation, mais peut considérablement réduire le potentiel de conjugaison ou de transduction. Au contraire, la coopération entre souches bactériennes pour la formation de biofilms peut faciliter le HGT : d'une part en protégeant l'ADN échangé, et d'autre part en stabilisant les communautés entre lesquelles les gènes sont échangés (Madsen et al. 2012). Ce mécanisme souvent négligé est même parfois qualifié de point chaud de HGT ce qui amène certains à le considérer comme un mécanisme à part entière du HGT (Abe et al. 2020).

Enfin, des paramètres abiotiques de l'environnement peuvent également jouer dans la mise en place du HGT. En premier lieu, les environnements spatialement structurés sont des points chauds de HGT. En créant des alcôves de forte densité bactérienne, la structuration spatiale facilite la mise en contact de souches différentes et donc la mise en place de la conjugaison (Cairns et al. 2018). De plus, les bactéries peuvent s'attacher à des particules dans les environnements à forte granulométrie, ce qui leur permet, lors d'un ruissellement, une dissémination plus importante vers d'autres environnements (Soupir et al. 2010). Par ailleurs, différents paramètres abiotiques de l'environnement peuvent exercer des stress chimiques ou physiques sur les bactéries, et ainsi stimuler le HGT. En particulier, les concentrations sub-inhibitrices d'antibiotiques sont connues pour augmenter la fréquence des bactéries portant des gènes de résistance (Andersson et Hughes 2014). De même, cet effet a aussi été observé pour la chloration et la désinfection des eaux aux rayons ultra-violets (Su et al. 2014).

Bien qu'il n'ait pas été possible de prendre en compte ces différents paramètres dans la section V, ces éléments montrent qu'il est essentiel d'étudier les liens entre écologie et HGT, car les paramètres biotiques et abiotiques des écosystèmes peuvent moduler la probabilité de rencontre et de transfert des bactéries portant des gènes de résistance.

### 4.2.4 Les bactéries généralistes, des navettes pour la résistance ?

Dans le cadre de la connectivité écologique et géographique, certains taxons bactériens peuvent participer à une circulation accrue de la résistance. Par exemple, certaines espèces, telles que *Campylobacter jejuni, Escherichia coli, ou Salmonella enterica*, sont particulièrement surveillées du fait de leur risque de transmission de l'animal à l'humain par la consommation de viande (Schrijver et al. 2018). De manière générale, plus la niche d'un organisme est large, plus cet organisme est susceptible de coloniser de nombreux écosystèmes. Les bactéries généralistes sont donc celles qui présentent le risque théorique le plus élevé de propager les gènes de résistance entre les écosystèmes. A l'heure actuelle, les bactéries les plus surveillées face au risque de propagation de la résistance sont les *Enterobacteriaceae*, du fait de leur capacité à coloniser les microbiotes humains et animaux. Cette famille fait même partie des taxons pathogènes les plus problématiques, pour lesquels l'OMS juge qu'il est critique de développer de nouveaux antibiotiques (*Global action plan on antimicrobial resistance* 2021).

Toutefois, la section V met en évidence que d'autres taxons semblent particulièrement à risque de servir de navette aux résistances antibiotiques. En particulier, les *Pseudomonadaceae* couvrent une très large gamme de niches écologiques, des organismes pathogènes jusqu'aux symbiotes mutualistes, et la niche de certaines espèces telles que *P. aeruginosa* peuvent couvrir à elles seules une très large gamme d'écosystèmes (Wiehlmann et al. 2015). Cette famille représente donc un archétype de bactéries généralistes. Les résultats de la section V tendent à soutenir l'idée que les organismes généralistes ne sont pas seulement plus à risque de migrer entre écosystèmes, mais également qu'ils sont plus susceptibles de devenir résistants, et de transférer ces résistances à d'autres taxons. En effet, dans les réseaux de transfert d'AME qui ont été reconstruits, les *Pseudomonadaceae* sont largement sur-représentées, et comportent à la fois beaucoup de donneurs et de receveurs.

D'une manière générale, on peut supposer que les bactéries généralistes sont susceptibles de participer à davantage de communautés différentes, de rencontrer davantage d'espèces différentes, et donc potentiellement d'avoir davantage d'opportunités d'acquérir ou de transmettre horizontalement des gènes de résistance que les bactéries spécialistes. Ceci tend à être confirmé par le fait que la coévolution bactérie-plasmide sélectionne des plasmides à plus large spectre d'hôte lorsque les communautés comportent davantage d'espèces (Kottara et al. 2016). De même, les bactériophages généralistes ont plus de succès que les spécialistes lorsqu'une plus grande diversité d'hôtes potentiels sont disponibles (Sant et al. 2021). Le transfert interspécifique de gènes impliquant des souches ubiquitaires est donc facilité d'une part parce qu'ils rencontrent davantage de bactéries résistantes, et d'autre part parce que les MGE qu'ils portent ont un spectre d'hôte plus large.

A l'heure actuelle, certaines *Pseudomonadaceae* sont surveillées par l'OMS au même titre que les *Enterobacteriaceae* (*Global action plan on antimicrobial resistance* 2021) : leur pathogénicité et leur résistance à très forte prévalence. Cependant, les résultats de cette thèse suggèrent également que les bactéries généralistes devraient aussi être surveillées au titre de leur potentiel de navette pour les gènes de résistance entre différents biomes.

## 4.3 Origines et circulation des résistances antibiotiques

### 4.3.1 Les environnements naturels sont-ils des réservoirs de résistance ?

L'augmentation de la prévalence des résistances antibiotiques dans les écosystèmes naturels depuis le début de leur utilisation par l'Homme suggère un fort impact des activités humaines, soit comme le résultat de la circulation de bactéries résistantes et de leurs gènes de résistance, soit du fait de la sélection par la pollution antibiotique. Il a donc été suggéré que les écosystèmes naturels pourraient constituer des réservoirs de résistance (Marti et al. 2014). Les résultats de cette thèse soutiennent l'idée que ces environnements sont aussi des sources potentiellement importantes de résistance. En effet, le réseau de résistomes reconstruit dans la section IV montre clairement une centralité particulièrement forte pour le sol d'une part, et la faune et la flore sauvages d'autre part. S'agissant d'un réseau de similarités, il est impossible d'orienter les potentiels échanges entre biomes, mais cela démontre qu'au moins certains transferts ont ces deux biomes pour origine. De même, la section V démontre un rôle important des bactéries vivant dans ces milieux pour la diffusion des gènes de résistance, si ce n'est pour leur apparition.

L'émergence des résistances antibiotiques dans les milieux naturels est déjà bien connue (D'Costa et al. 2011). En effet, la présence de résistance antibiotique dans des milieux naturels est attestée avant l'utilisation anthropique des antibiotiques. On infère donc que comme ces résistances ont émergées en dehors des environnements hospitaliers, leurs origines se trouve dans le milieu où elles ont été primairement échantillonnées. Cependant, si on peut dire que les résistances environnementales ont un impact sur le milieu médicale (en tant que réservoirs et aussi potentiellement de sources), il reste encore difficile de quantifier la part relative de la transmission de ces dernières par rapport aux transferts intra-milieux hospitaliers sur la prévalence des infections à bactéries résistantes chez l'humain. Toutefois, une différence notable peut se faire en retraçant les chaînes de transmission. En effet, en milieu hospitalier, des résistances peuvent apparaître directement chez des bactéries pathogènes. Dans le cas de résistances environnementales, il est nécessaire que les gènes concernés soient transférés à des commensaux ou à des pathogènes, d'humains ou d'animaux domestiques (Bengtsson-Palme et al. 2018). Les réseaux reconstruits dans la section V sont clairement hétérophiles, et montrent une sur-représentation des transferts vers les bactéries pathogènes de l'humain, ce qui montre que les résistances environnementales peuvent clairement poser un risque de santé publique, au moins dans le cas des AME.

Cependant, à ce stade, il n'est pas possible d'affirmer que la majorité des résistances (ou même des

AME) émerge dans les environnements naturels. De même, l'apparition des clusters d'AME n'ayant pas été datée, il reste encore impossible de déterminer si les résistances diffusées depuis ces écosystèmes y sont bien apparues ou d'identifier le rôle des activités humaines sur l'émergence des résistances dans ces environnements. En effet, depuis les dernières décennies, les pollutions antibiotiques (tout comme d'autres types de pollutions) ont probablement favorisé une accélération des émergences. Il reste en tout cas clair que les écosystèmes naturels et les taxons de bactéries non pathogènes doivent faire l'objet d'une surveillance accrue.

### 4.3.2 Dans quelle(s) direction(s) les résistances se propagent-elles ?

Le point précédent ouvre l'idée contre-intuitive que la résistance pourrait non seulement émerger dans les écosystèmes peu anthropisés, mais aussi se propager vers des écosystèmes où la pression antibiotique serait plus importante. En effet, le consensus actuel suppose une apparition préférentielle des résistances dans des écosystèmes sous forte sélection (souvent considérée sous le prisme unique de l'exposition aux antibiotiques), et une propagation vers des écosystèmes où la pression est moins forte. Ce travail de thèse ne remet pas en cause le rôle essentiel des antibiotiques dans la propagation des résistances au cours des dernières décennies, mais appelle à nuancer le consensus actuel par un recours à une analyse plus intégrative de la résistance et des mécanismes de propagation de cette dernière. En effet, même s'ils sont reconstruits à partir d'un jeu de données fortement biaisé en faveur des bactéries pathogènes et des environnements cliniques, les réseaux de propagation des AME démontrent un rôle important des espèces non pathogènes. De plus, bien que les hôpitaux, les microbiotes humains et animaux, et les agrosystèmes soient les biomes où les AME sont les plus prévalentes, ce ne sont pas les biomes les plus centraux.

A l'heure actuelle, la propagation écologique des résistances antibiotiques est décrite selon le modèle suivant. La résistance peut diffuser :

- des environnements cliniques vers les environnements domestiques suite à l'hospitalisation des humains et des animaux ;
- des animaux domestiques vers les humains par contact ou par consommation de viande ;
- des humains et des animaux domestiques vers les sols, les agrosystèmes, et les stations d'épuration par pollution fécale ;
- des stations d'épuration vers les circuits d'eau domestiques, et vers les cours d'eau.

S'il est clair que ces processus correspondent probablement à des transferts massifs de gènes et de bactéries entre écosystèmes, le rôle des sols et de la biodiversité sauvage est sous-estimé par les programmes de surveillance. D'autres mécanismes de propagation des résistances peuvent donc être suggérés (figure 6). Par exemple, des bactéries résistantes peuvent passer du sol vers les plantes cultivées (Cerqueira et al. 2019), qui peuvent ensuite nourrir des animaux d'élevage et les contaminer (Friedman 2015). Le même processus peut se produire dans les environnements naturels où les sols et les végétaux contaminent des animaux herbivores et carnivores, qui sont ensuite susceptibles de rentrer en contact

avec des humains ou des animaux domestiques (par exemple par le biais de la chasse, ou de vecteurs arthropodes, VITTECOQ et al. 2016). Ces voies de transmission ne sont probablement pas les seules, et leur importance quantitative dans la propagation des résistances reste encore à mesurer (DOLEJSKA et LITERAK 2019). Toutefois, à la lumière des résultats de cette thèse, il semble que les flux de propagations des résistances ne sont pas linéaires mais plutôt organisées en réseau.

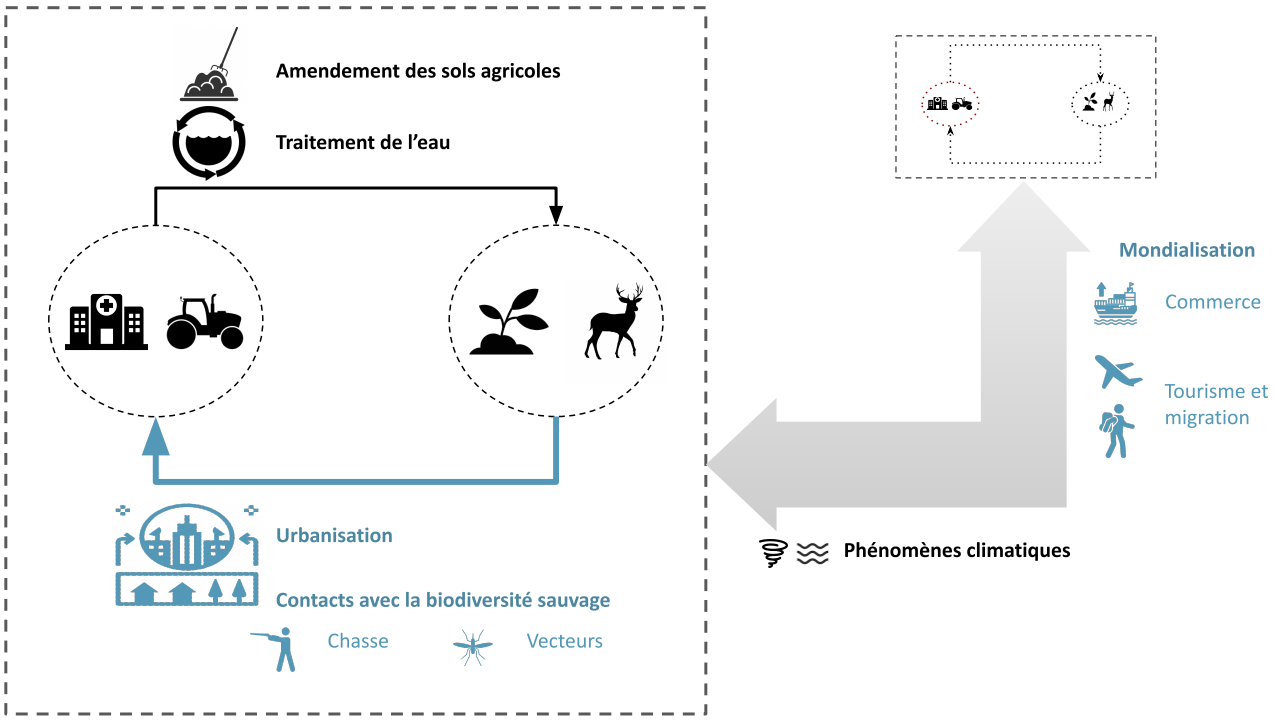


**FIGURE 6 –** Propagation écologique et géographique des résistances antibiotiques.
Les éléments représentés en bleu ont été étudiés ou mis en évidence dans le cadre de cette thèse.

# 5 Structuration génomique et phylogénétique de la résistance

## 5.1 Une résistance ubiquitaire chez les bactéries

### 5.1.1 Les AME sont présents dans toute la phylogénie des Eubactéries

Environ un quart des génomes analysés au cours de cette thèse présentent des gènes encodant des AME. Il s'agit d'une très forte proportion, qui cache cependant une certaine hétérogénéité phylogénétique, notamment du fait d'une forte surreprésentation des organismes-modèles ou des espèces d'intérêt médical (par exemple, *Escherichia flexneri* représente 8.4% des génomes analysés). Toutefois, ces gènes ont tout de même été retrouvés dans 23 phyla particulièrement diversifiés. Sans surprise, ces gènes se retrouvent particulièrement dans des taxons particulièrement associés aux microbiotes

humains (*Proteobacteria, Actinobacteriota, et Firmicutes*), mais également dans des taxons dont l'impact est relativement peu considéré. En particulier, les AME sont présentes dans plusieurs phyla vivant exclusivement dans l'eau ou le sol : les *Cyanobacteria*, les *Gemmatimonadota*, les *Bdellovibriota B*, les *Planctomycetota*, les *Myxococcota*, ou encore les *Verrucomicrobiota A*. La distribution des AME dans des taxons telluriques est compatible avec les origines connues des AME, telles que détectées dans la section V par exemple. La présence dans des taxons aquicoles pourrait être compatible avec un ruissellement de bactéries depuis les sols vers les cours d'eau, bien que certains de ces taxons, tels que les *Cyanobacteria*, sont aussi connus comme réservoirs et sources de résistances antibiotiques (Wang, Chen et al. 2020). Il est aussi intéressant de constater qu'on a pu découvrir des résistances y compris chez les *Gemmatimonadota*, qui n'ont été découvertes qu'en 2003 (Zhang, Sekiguchi et al. 2003). Il est donc possible d'envisager que l'étendue phylogénétique des résistances antibiotiques pourrait aller croissant avec la découverte de nouveaux taxons bactériens, de même que l'on peut trouver de nouveaux gènes de résistance lors d'études métagénomiques dans les environnements naturels (Willms et al. 2019).

De fait, la répartition phylogénétique des AME semble aller de pair avec leurs répartitions écologique, comme démontré dans la section IV. Le génome encodant une AME le plus ancien du jeu de données (échantillonné en 1905) appartient à l'espèce *Salmonella enterica*, et a été retrouvé chez l'Homme. Ceci pourrait témoigner d'une émergence précoce, antérieure à l'utilisation des antibiotiques, de certaines AME chez les *Enterobacteriaceae*. Pour autant, cette première identification chez *S. enterica* est probablement liée à un biais d'échantillonnage. De plus, les résultats de la section V et le reste de la littérature tendent plutôt à démontrer une barrière écologique au transfert horizontal des AME. Tout comme discuté dans la partie 4.2.4, le scénario le plus plausible est donc que les AME identifiées chez des taxons exclusivement telluriques ou aquicoles aient été échangées avec des espèces généralistes vivant dans les mêmes milieux, et que ces dernières aient ensuite migré vers d'autres environnements.

### 5.1.2 Une distribution différente en fonction des familles de gènes

Bien que les AME sont largement réparties dans la phylogénie, la distribution varie énormément en fonction des clusters identifiés. L'envergure de la distribution peut varier d'un unique genre (*Pseudomonas* pour le cluster 12) jusqu'à plus de 12 phyla (pour le cluster 6.1). Les résultats de la section IV indiquent que cette envergure est corrélée à la date de première apparition dans le jeu de données. Une envergure phylogénétique plus large pourrait donc simplement correspondre à un temps de diffusion plus long. De manière similaire, il a été suggéré que la β-lactamase TEM-1 est plus répandue que les autres chez les *Enterobacteriaceae* car il s'agit d'une des premières à être apparues (Medeiros 1997).

Cependant, d'autres caractéristiques pourraient expliquer le succès de propagation propre à chaque cluster. Comme argumenté par Bengtsson-Palme et al. 2018, dès lors qu'un gène de résistance est acquis par une communauté, il est très difficile de l'en déloger. Le sol et la biodiversité sauvage ont des résistomes centraux par rapport aux autres biomes, et plusieurs taxons y vivant semblent être de gros diffuseurs d'AME : ces deux environnements pourraient donc être des points de départ facilitant la pro-

pagation écologique. De plus, les résultats de cette thèse suggèrent que certaines communautés bactériennes, en particulier celles associées à l'humain ou aux animaux domestiques, sont très fortement interconnectées géographiquement. Par conséquent, une famille de résistance acquise précocement par une bactérie capable de vivre dans l'environnement humain, a beaucoup plus de chances de se propager géographiquement. Ces chances devraient encore augmenter quand les résistances sont portées par des bactéries capables de vivre dans plusieurs types d'écosystèmes (figure 7).

En outre, le contexte génomique préférentiellement associé à chaque famille peut aussi participer à expliquer la diffusion phylogénétique de chaque famille. En effet, les MGE à mobilité intergénomique peuvent avoir un spectre d'hôte relativement large (encore plus pour les plasmides conjugatifs et les ICE, Cury et al. 2017). Dans ce cadre, la diffusion hospitalière de certaines résistances est étroitement associée à la diffusion des plasmides qui les portent (San Millan 2018). De plus, l'association à des MGE permet aussi potentiellement l'association à d'autres gènes, donnant potentiellement lieu à un phénomène de co-sélection. La section V démontre effectivement que les AME associés à des MGE ont plus de chances d'être transmis horizontalement, et que les MGE emboîtés (l'association de deux MGE, l'un à mobilité intergénomique et l'autre à mobilité intragénomique) favorisent le transfert horizontal à plus grande distance phylogénétique (figure 7).

### 5.1.3 Le transfert horizontal des résistances se fait-il à large échelle ?

Malgré la répartition phylogénétique très large des AME, certains clusters présentent une très faible proportion de gènes horizontalement transmis. C'est notamment le cas du cluster 6.1 qui, bien que très largement distribué dans la phylogénie, n'a été détecté comme horizontalement transmis que pour moins de 5 % des gènes dans la section V. Au vu de la distribution phylogénétique de ces clusters, il est peu probable que ces clusters aient émergé chez l'ancêtre commun le plus récent de tous les taxons qui les portent (car cela supposerait de très nombreuses disparitions du gène). Cependant, la section V ne traite que des événements de HGT récents (après lesquels il y a moins de 5% de divergence génétique) et à grande échelle phylogénétique (entre taxons n'appartenant pas au même genre). Il est donc possible que les clusters concernés par une faible détection de ces événements se soient tout de même propagés horizontalement de manière plus ancienne, et plus récemment à plus courte échelle (au sein du même genre ou de la même espèce). En effet, des événements anciens ont pu être détectés entre les actinobactéries et les protéobactéries (voir par exemple Jiang, Ellabaan et al. 2017). De même, ces clusters peuvent être très peu mobiles (0.7 % des gènes du cluster 6.1 sont associés à des MGE à mobilité intergénomique), ce qui suggère qu'ils peuvent surtout se propager à des souches phylogénétiquement proches.

De manière générale, le HGT est connu pour se faire plutôt à courte échelle phylogénétique (Popa et al. 2011). Certains clusters d'AME semblent suivre cette tendance, mais l'exemple du cluster 6.1 laisse supposer que quelques rares événements à plus large échelle peuvent suffire à rendre ubiquitaire une famille de résistances. Au contraire, pour d'autres clusters, la forte prévalence de gènes horizontalement

transmis à large échelle suggère que les AME (et possiblement les gènes de résistance en général) ont une mobilité forte, comparable à ce qui a pu être observé pour d'autres types de gènes (voir par exemple Stevenson et al. 2017 sur la résistance au mercure) . Cette mobilité peut supposer un succès plus facile des événements de transfert horizontal (soit que les AME sont moins coûteux à acquérir, soit qu'ils apportent un bénéfice adaptatif plus important), ou tout simplement une plus forte association à des MGE. Une interprétation possible est que les premières décennies de consommation d'antibiotiques auraient créé un terrain favorable pour que l'acquisition de gènes de résistance soit plus avantageuse que pour n'importe quel autre type de gène, et que cette meilleure capacité de transfert sélectionne les MGE associés à des gènes de résistance (Jansen et Aktipis 2014). Même si les antibiotiques ne semblent plus jouer un rôle aussi important qu'auparavant, il est rare que des bactéries perdent leur résistance (Andersson et Hughes 2010), et la distribution phylogénétique actuelle pourrait simplement être le résultat d'une forte mobilité acquise par le passé.

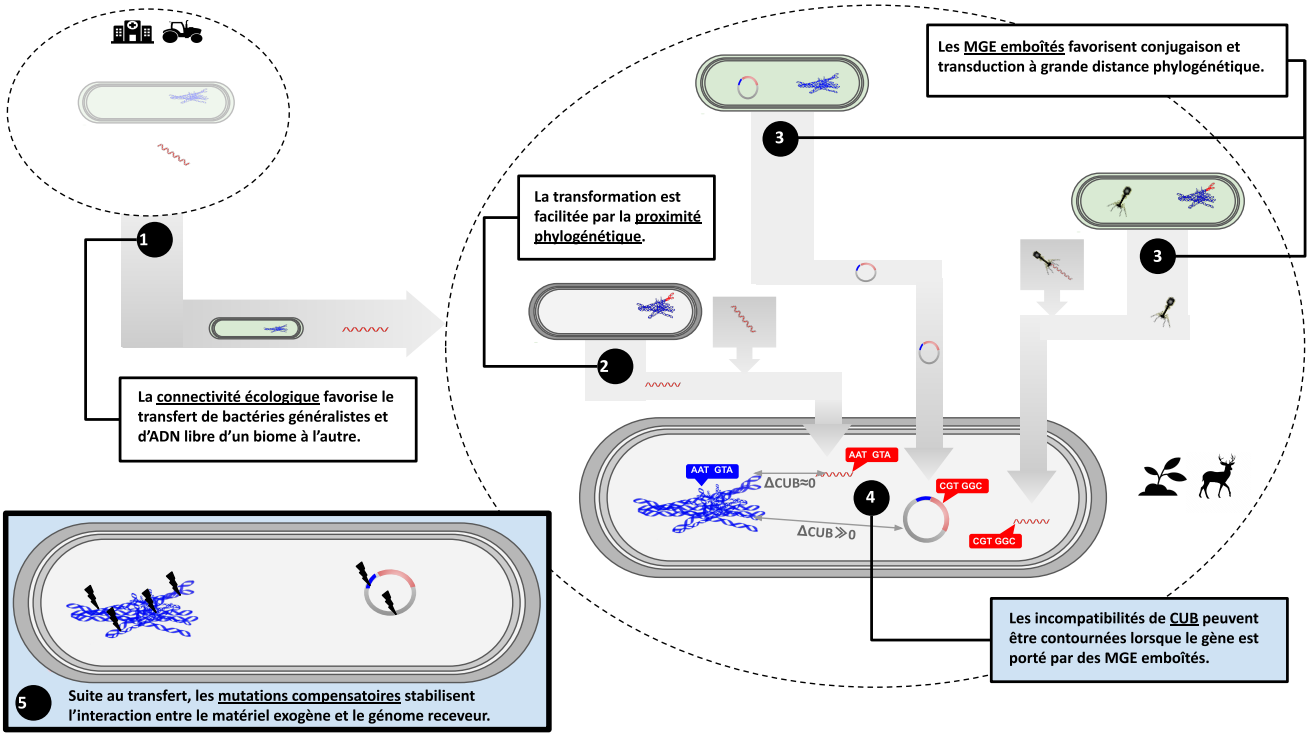


**Figure 7 –** Les facilitateurs du transfert horizontal de gènes.
Les légendes sur cadre bleu (points 4 et 5) correspondent à des facilitateurs post-transfert.

## 5.2 Le rôle structurant des éléments génétiques mobiles

Tout du long de cette thèse, j'ai pu montrer que les MGE sont impliqués dans une diversité de mécanismes participant à la propagation des gènes de résistance. Ces mécanismes sont résumés dans la figure 8.

### 5.2.1 Diversification fonctionnelle et émergence de la multirésistance

Le rôle des MGE dans la propagation des résistances antibiotiques a été particulièrement étudié durant les 50 dernières années. C'est grâce à leur découverte qu'on a pu expliquer le transfert horizontal des résistances antibiotiques, et ils sont régulièrement identifiés comme marqueurs de propagation de certaines familles de résistance. Cette thèse ne fait pas exception, et soutient notamment l'idée que les MGE pourraient avoir un rôle déterminant dans l'émergence des souches avec un large spectre de résistance.

D'une part, une bactérie capable de recevoir de nombreux transferts horizontaux est donc potentiellement capable de recevoir de nombreux gènes de résistance différents. Par exemple, le genre *Acinetobacter* accumule régulièrement des gènes de résistance par prédation (Cooper et al. 2017). Comme démontré par la section V et un grand pan de la littérature, l'acquisition de nouveaux gènes de résistance aux antibiotiques peut être facilitée lorsque ceux-ci sont portés par des MGE. De plus, la section IV montre que l'accumulation de ces gènes portés par des MGE est associée à un élargissement significatif du spectre de résistance de leurs hôtes. Toutefois, il est à noter que la manière dont les MGE peuvent s'accumuler au sein d'un même hôte est encore mal compris (Rocha et Bikard 2022) : par exemple, certains plasmides peuvent se réprimer les uns les autres (Gama et al. 2017), tandis que d'autres peuvent coopérer pour stabiliser l'interaction avec leur hôte (San Millan, Peña-Miller et al. 2014).

D'autre part, les MGE favorisent la duplication des gènes de résistance. Plusieurs mécanismes peuvent ici entrer en ligne de compte : que ce soit la réplication indépendante des plasmides, ou bien la transposition de type "copier-coller" de certains transposons et ICE. Les transposons et intégrons ont un effet duplicatif particulièrement fort lorsqu'ils s'emboîtent dans les plasmides, ICE, ou phages, puisqu'ils permettent la potentielle mobilisation de gènes de résistance à la fois sur ces éléments et sur le chromosome. Cette duplication autoriserait une bactérie à donner une copie de son gène de résistance tout en gardant la version originale de ce dernier. Dans ce cas, le HGT ne serait pas un acte altruiste à l'origine d'une sélection pour un partage préférentiel aux apparentés (une hypothèse défendue par certains auteurs, tels que Norman et al. 2009). Quel que soit le mécanisme impliqué, la duplication d'un gène peut alléger la sélection sur le duplicat, et permettre par dérive l'acquisition de nouvelles fonctions de résistance (San Millan et Maclean 2017 ; Sandegren et Andersson 2009). Là encore, cela pourrait expliquer pourquoi l'accumulation d'AME portés par des MGE est associé à une plus large diversification du spectre de résistance.

### 5.2.2 Des facilitateurs du transfert horizontal

L'association des gènes à des MGE est aujourd'hui largement connu pour être un déterminant important de la capacité à se propager au sein et entre les génomes. C'est un élément-clé pour considérer qu'un gène de résistance pose un risque majeur pour la santé publique (Martínez et al. 2015). Une nouveauté apportée par cette thèse est le rôle prépondérant des structures emboîtées de MGE dans

la propagation des gènes de résistance. Dans le jeu de données décrit dans la section IV, la moitié des transposons et intégrons porteurs d'AME sont également localisés sur des plasmides, des ICE, ou des prophages. En outre, la section V montre que cette structure est associée à une plus forte probabilité de transfert horizontal, notamment vers des taxons phylogénétiquement éloignés ou présentant un usage de codons très dissimilaire. Comme déjà discuté dans cette partie, les MGE emboîtés ont l'avantage de permettre la conjugaison ou la transduction de gènes portés par des intégrons ou des transposons. De même, comme suggéré par Che et al. 2021, les transposons peuvent être facilement déplacés d'un plasmide vers le chromosome, ce qui permet un transfert de gène pérenne même si l'élément conjugatif est perdu (figure 7).

En outre, il a été montré que les séquences d'insertion participent à la cointégration des plasmides (Machida et al. 1982). Ce processus peut permettre à la fois aux plasmides d'acquérir ou accumuler des gènes de résistance (Desmet et al. 2018), mais également pour un gène de se copier d'un plasmide à un autre (Clewell et al. 1975). Le processus de cointégration, qui implique donc des transposons et des plasmides, participe parfois à transférer des gènes de résistance vers des plasmides à plus large spectre d'hôte (Bonafede et al. 1997 ; Li et al. 2020), ce qui contribue à augmenter l'étendue phylogénétique à laquelle ils peuvent se transmettre.

Il est à noter que tous les plasmides porteurs d'AME identifiés ne sont pas nécessairement conjugatifs, dans la mesure où l'outil *PlasForest* développé en section III n'est pas capable de les discriminer. Pour autant, des plasmides non-conjugatifs peuvent être mobilisés par plusieurs mécanismes (Ramsay et Firth 2017). Par exemple, la relaxase mob est capable de fusionner des plasmides non-conjugatifs avec des plasmides conjugatifs, les permettant d'être transférés ensemble (Gennaro et al. 1987). De même, certains plasmides peuvent parasiter le système de mobilisation de plasmides conjugatifs corésidents en imitant leur séquence *oriT* (origine du transfert, O'Brien et al. 2015). Cependant, des analyses complémentaires seraient nécessaires afin de déterminer la part de plasmides conjugatifs et de plasmides non-conjugatifs mobilisables parmi ceux qui portent des AME.

## 5.3 Les préférences d'usage de codons : une barrière au transfert horizontal des résistances ?

Le succès du transfert horizontal dépend de plusieurs facteurs : dans un premier temps, de l'entrée de matériel génétique dans l'hôte ; et dans un second temps, de la compatibilité du nouveau gène avec l'hôte et des potentiels coûts associés. Plusieurs mécanismes peuvent ensuite être envisagés comme facilitateurs ou barrières à ces deux étapes. Bien que plusieurs mécanismes aient été décrits en détail pour l'entrée du matériel génétique (partie 2.3.1), le devenir du matériel génétique exogène est le plus souvent analysé en termes de coûts de réplication et d'expression. Dans le cadre de cette thèse, un autre mécanisme a été abordé : l'incompatibilité de CUB.

### 5.3.1 Une barrière entre sélection diversifiante et sélection stabilisante

Dans la mesure où un gène récemment transféré n'a pas été co-sélectionné avec le reste du génome, il présente parfois une rupture de co-adaptation avec la machinerie de traduction. Un gène encodé par des codons rares a moins de chances d'être traduit rapidement et fidèlement. Ainsi, des substitutions synonymes sont susceptibles à elles seules de perturber le repliement co-traductionnel des protéines et de réduire la valeur sélective de la cellule (Walsh et al. 2020). Il existe une large littérature traitant de l'efficacité translationnelle comme une pression de sélection sur l'usage de codons, et des expériences ont suggéré que les incompatibilités de CUB lors des événements de HGT peuvent réduire les chances de succès de l'événement (Amorós-Moya et al. 2010 ; Bedhomme, Amorós-Moya et al. 2019). Les résultats de la section V soutiennent cette hypothèse : les AME sont plus souvent transférés vers des génomes présentant un CUB similaire au leur.

En revanche, l'incompatibilité de CUB peut également provenir d'autres sources que la sélection pour l'efficacité translationnelle. Ces autres sources de sélection font l'objet d'une revue à laquelle j'ai contribuée (Callens, Pradier et al. 2021, voir section 11). D'une part, certains codons seraient évités de sorte à ne pas former accidentellement des motifs impliquées dans la régulation de l'expression (séquences promotrices, codons stop, etc.). Ainsi, si un gène exogène présentait certains de ces motifs, cela entraînerait la production de protéines tronquées et/ou de protéines non fonctionnelles. D'autre part, il pourrait y avoir une co-sélection entre l'usage de codons et le système antiviral de restriction-modification. Un gène exogène présentant l'un des motifs reconnus par ce système serait clivé et éliminé. Enfin, Callens, Pradier et al. 2021 formulent l'hypothèse que l'usage de codons tend à éviter les motifs préférentiels d'insertion des éléments transposables. Si cette hypothèse est confirmée, cela signifierait qu'un gène avec un usage de codon dissimilaire à celui de son hôte a plus de chances d'être soumis à l'intégration d'un élément transposable, ce qui empêcherait son expression. Toutefois, les motifs évités n'ont pas été recherchés dans la section V, et à ce jour, l'impact de ces sources d'incompatibilité sur le HGT n'a pas encore été mesuré.

Globalement, l'établissement de fortes différences entre les fréquences d'usage des codons synonymes forme une barrière à l'acquisition de nouveau matériel génétique, ce qui peut être rapproché de l'équilibre évolutif entre sélection diversifiante et sélection stabilisante. En effet, si l'acquisition de nouveaux gènes présente de nombreux avantages potentiels en termes d'évolvabilité, elle peut également perturber l'ensemble du phénotype. Une hypothèse qui pourrait être formulée est que cette perturbation soit l'une des raisons pour laquelle certaines familles de bactéries naturellement compétentes ont évolué une sélectivité à l'acquisition d'ADN exogène (Mell et Redfield 2014). Ces bactéries ne reçoivent que des gènes dont la séquence a un contenu nucléotidique similaire au leur, ce qui augmente ainsi les chances que ces gènes aient évolué dans un contexte génomique similaire et qu'ils ne perturbent pas l'expression de leur phénotype. Ainsi, ces bactéries sont moins à risque de recevoir par transformation des gènes au CUB incompatible au leur. De ce point de vue, il est intéressant de constater que, dans la section V, l'incompatibilité de CUB s'observe principalement pour les AME associés à aucun MGE, et qui

ne peuvent donc être transférés que par transformation.

### 5.3.2 Une barrière qui peut être contournée

Puisque les incompatibilités de CUB forment une barrière au HGT, certains taxons qui présentent un fort biais d'usage de codons devraient être très peu susceptibles de recevoir du matériel génétique exogène. En particulier, l'acquisition de gènes de résistance par HGT devrait être rare pour ces taxons. Pourtant, les CUB des bactéries généralistes, qui sont les espèces les plus à risque de propager des gènes de résistance entre les écosystèmes, sont globalement les plus forts de toutes les bactéries (Botzman et Margalit 2011). De fait, la répartition actuelle des gènes de résistance implique qu'il existe des moyens de contourner la barrière de l'incompatibilité de CUB.

Par exemple, Bedhomme, Amorós-Moya et al. 2019 ont montré expérimentalement que, suite à l'acquisition d'un gène de résistance présentant une incompatibilité de CUB, le coût de cette incompatibilité tend à diminuer au cours des générations du fait de mutations compensatoires, notamment dans certains gènes du métabolisme. De même, une étude rétrospective chez *Pseudomonas aeruginosa* a montré que les gènes exogènes reçus ont parfois évolué pour se rapprocher du CUB de leur hôte, ou que certains gènes d'ARNt avaient pu être dupliqués en réponse à l'acquisition de gènes présentant une incompatibilité de CUB (Callens, Scornavacca et al. 2021). Si le temps de rétention d'un gène exogène semble effectivement être corrélé à la proximité entre le CUB d'un gène et de celui de son hôte (Callens, Scornavacca et al. 2021), l'évolution compensatoire des gènes reçus et des génomes receveurs sont des moyens de stabiliser l'acquisition de matériel génétique exogène.

Dans ce cadre, la section V montre que les AME associés à des MGE emboîtés sont plus susceptibles d'être transférés à des taxons avec qui ils présentent des incompatibilités de CUB (figure 7). Bien que ce contexte génomique particulier n'est pas associé à une évolvabilité particulièrement élevée, on peut formuler deux hypothèses pour expliquer leur meilleure capacité à contourner l'incompatibilité de CUB. D'une part, comme discuté dans la partie 5.2.1, les MGE emboîtés ont une capacité de duplication plus élevée que les autres. Dans les cas où le coût de l'incompatibilité de CUB est dû à une trop faible production de protéines fonctionnelles, la duplication du gène concerné pourrait (au moins partiellement) compenser ce coût. D'autre part, si ce transfert s'accompagne de l'acquisition de nouveaux transposons, ceux-ci pourraient parfois se transposer au chromosome, et ainsi générer des mutations (potentiellement compensatoires) par insertion.

# 6 Vers de nouvelles approches de gestion de la résistance antibiotique

Les résultats de cette thèse ouvrent de nouvelles perspectives pour le suivi et gestion du problème de la résistance aux antibiotiques. Certains exemples de stratégies potentielles à mettre en place sont

résumées dans la figure 9.

## 6.1 Dépasser une conception adaptationniste de la résistance

L'expansion des bactéries résistantes aux antibiotiques depuis les dernières décennies est très clairement le résultat de leur sélection par la consommation d'antibiotiques. Cependant, comme discuté dans la partie 4.1, cela ne signifie pas que les gènes qui confèrent cette résistance sont uniquement des adaptations à la consommation d'antibiotiques par l'Homme. En effet, nombre de ces gènes ont porté d'autres fonctions par le passé : la résistance aux antibiotiques produits par d'autres bactéries, mais pas seulement (Martinez 2012). Ces derniers ne sont alors que des exaptations. De plus, les résultats de cette thèse suggèrent fortement que ces gènes ne sont pas nécessairement maladaptatifs en dehors de l'environnement qui a sélectionné leur expansion (c'est-à-dire lorsque les consommations d'antibiotiques décroissent). Les stratégies de confinement de la résistance antibiotique gagneraient donc à dépasser la conception adaptationniste de la résistance aux antibiotiques.

Parmi les modèles alternatifs qui peuvent être proposés, un premier consisterait en une approche neutraliste de la résistance aux antibiotiques. En effet, comme discuté dans la partie 4.1, le coût conféré par les gènes de résistance, en l'absence d'antibiotique, peut être assez rapidement réduit (voire annulé) par évolution compensatoire. Même si cette évolution compensatoire devait avoir lieu après chaque événement de HGT, il reste que le temps nécessaire à cette évolution est très court chez les bactéries par rapport aux décennies d'utilisation des antibiotiques. Selon ce modèle, les bactéries résistantes et les gènes de résistance se maintiendraient dans les communautés par dérive. Leur relative abondance serait simplement le résultat de leur sélection passée.

Un autre modèle possible pourrait se centrer sur la migration de bactéries et de gènes entre communautés et entre continents. Comme cette thèse et un corpus grandissant de recherches le montrent, la mondialisation et la modification des équilibres écologiques par l'Homme a de lourdes conséquences sur la répartition des résistances. D'une part, si la sélection sur les bactéries résistantes a baissé depuis les dernières décennies dans les pays développés, la consommation d'antibiotique mondiale tend toujours à augmenter, notamment dans les pays en développement (Klein et al. 2018). D'autre part, les résistances peuvent être adaptatives et/ou sélectionnées dans les écosystèmes naturels. Ces pays et écosystèmes pourraient donc constituer des réservoirs ou des sources pour le reste du monde, vers lequel la résistance ne ferait que migrer sans nécessairement y être sélectionnée.

Si les résultats de cette thèse soutiennent davantage ces deux modèles, il est clair que de nouvelles approches doivent être proposées et testées afin de construire une meilleure compréhension du phénomène de la résistance à l'heure actuelle.

## 6.2 Intégrer la gestion écologique dans la santé publique

### 6.2.1 Vers une surveillance accrue des écosystèmes naturels

A l'heure actuelle, le suivi des résistances antibiotiques concerne essentiellement les organismes susceptibles d'infecter l'humain. Par exemple, le rapport exécutif de l'ECDC pour 2020 ne rapporte les résultats que pour un très faible nombre de taxons : *Acinetobacter spp., Escherichia coli, Staphylococcus aureus, Klebsiella pneumoniae, Pseudomonas aeruginosa, et Enterococcus faecium* (WHO Regional Office for Europe 2021). Comme déjà discuté dans la partie 5.1 et plus particulièrement la partie 4.2.4, le suivi de ces taxons est utile, car il s'agit de pathogènes humains, qui accumulent les résistances, et sont susceptibles de les déplacer d'un écosystème à l'autre. Toutefois, il est à noter que ce suivi n'offre qu'une image incomplète de la circulation des résistances antibiotiques. Les résultats de cette thèse suggèrent que ce tableau devrait être élargi.

D'une part, l'existence du transfert horizontal de la résistance et sa pervasivité à l'échelle mondiale devrait être un argument suffisant pour élargir la surveillance à un plus grand nombre d'espèces. Même une espèce incapable d'infecter l'humain (ou des animaux domestiques) est susceptible de porter des gènes de résistance, et donc de les transférer à des espèces pathogènes. Par conséquent, suivre la progression des résistances chez les bactéries pathogènes uniquement revient à constater l'ampleur des dégâts sur la santé humaine. Un suivi de la résistance chez des taxons non pathogènes permettrait d'anticiper ces dégâts, et potentiellement d'agir en conséquence pour en limiter l'ampleur. De manière plus générale, le phénotypage des bactéries pathogènes pourrait être complémenté par une surveillance génétique, en focalisant la surveillance sur les gènes de résistance et non plus seulement leur association à des bactéries pathogènes. L'élargissement à des bactéries non pathogènes pour l'humain commence à se faire au sein du monde académique. Par exemple, Mader et al. 2022 proposent la mise en place d'un réseau de surveillance des résistances en contexte vétérinaire en proposant de commencer par la surveillance d'un petit nombre de gènes de résistance. Ces approches trop minoritaires gagneraient à être plébiscitées et élargies car encore restreintes à des espèces animales et bactériennes déjà surveillées.

D'autre part, comme déjà largement discuté, les écosystèmes naturels constituent des sources et réservoirs importants de résistance. Les résistances introduites dans les élevages ou en milieu hospitalier peuvent avoir émergé plusieurs mois, années ou décennies auparavant dans les sols ou les cours d'eau. Pourtant, les échantillonnages en hôpitaux, élevages, et microbiotes humains sont globalement surreprésentés, et nous sommes donc collectivement aveugles à une partie importante du processus d'émergence et de circulation des résistances. Une manière relativement simple d'améliorer la surveillance des écosystèmes naturels serait de s'appuyer sur la présence de gènes de résistance dans les métagénomes. Toutefois, de nombreux gènes de résistance encore inconnus pourront être découverts, et déterminer précisément lesquels d'entre eux seront transférés aux pathogènes humains et constitueront un problème pour la santé humaine est probablement au-delà de nos capacités (Baquero, Martínez et al. 2021). Bien qu'essentielle à notre compréhension du phénomène, la surveillance des écosystèmes naturels devra donc être associée à une poursuite des études sur les déterminants du transfert horizontal.

### 6.2.2 Vers une réduction de la connectivité écologique

Bien que notre compréhension du phénomène dans les écosystèmes naturels soit encore largement incomplète, les travaux de cette thèse confirment que la connectivité écologique est un des facteurs majeurs et encore sous-estimés dans la propagation des résistances. La gestion des écosystèmes est donc un des leviers majeurs sur lesquels la lutte contre la résistance antibiotique peut s'appuyer.

Un premier axe consiste à limiter la circulation bactérienne des environnements anthropisés vers les environnements naturels. Il est déjà bien connu que les infrastructures d'hygiène et d'assainissement contribuent déjà à limiter la propagation. Ainsi, le matériel médical est stérilisé, de même que l'eau courante, mais aussi les aliments dans certains pays (Johnson 2015), de sorte à réduire la circulation des bactéries vers les humains. Mais, de même que les eaux usées peuvent également être stérilisées avant leur rejet vers les fleuves, il serait possible de traiter d'autres types de rejets afin de contenir la circulation des résistances. En particulier, l'épandage du fumier et du lisier est une source importante de pollution fécale pour les sols et les cours d'eau. La stérilisation de ces matériaux par ozonation ou rayonnement ultra-violet serait une manière de limiter, sans infrastructures coûteuses, la propagation des bactéries résistantes, mais devrait être généralisée à un grand nombre d'exploitations agricoles pour avoir un impact notable. En revanche, ces méthodes n'assurent pas la destruction complète du matériel génétique et facilitent même dans certains cas les événements de HGT (Su et al. 2014).

Le second axe de la stratégie de confinement de la résistance vise à limiter la circulation des résistances antibiotiques des écosystèmes naturels vers les nôtres. Ce confinement implique donc d'une part de réduire le nombre de bactéries résistantes dans ces écosystèmes, et d'autre part de restreindre nos contacts avec ces écosystèmes. Par exemple, cela peut passer par des campagnes de vaccination des animaux sauvages, de sorte à supprimer les réservoirs potentiels de bactéries résistantes. De même, une régulation de l'expansion urbaine et de l'exploitation des ressources animales et végétales non domestiques (chasse, exploitation forestière) pourrait ralentir la transmission des résistances sauvages vers les écosystèmes anthropisés. Toutefois, cet axe est beaucoup plus dur à mettre en place, car il impliquerait des coûts économiques et humains particulièrement élevés et/ou des modifications structurelles de nos systèmes économiques et de nos modes de vie.

### 6.2.3 Vers la décontamination des écosystèmes ?

Finalement, si jamais nous parvenons un jour à stopper la circulation des résistances antibiotiques, il reste que l'ensemble des écosystèmes semble, à ce jour, contaminés par des gènes de résistance. Comme évoqué plus haut, le coût des gènes de résistance n'est pas forcément suffisant pour provoquer leur disparition en l'absence d'antibiotiques. Une éradication complète ou partielle du problème nécessitera donc sûrement la mise en place de stratégies de décontamination. La première et la plus simple d'entre elles est la recolonisation des écosystèmes par des bactéries sensibles (Baquero, Lanza et al. 2015), comme ce qui se fait déjà à une échelle moindre avec l'usage de probiotiques. Cela diminuerait

mécaniquement la fréquence des bactéries résistantes, mais augmenterait également la compétition avec les bactéries résistantes. Cependant, cette stratégie peut engendrer des coûts importants pour des raisons insoupçonnées (beaucoup d'exemple d'ingénierie écologique ayant été désatreuses pour écosystèmes. Par conséquent, avant de considérer l'application de cette stratégie *in natura*, il faudrait renforcer notre compréhension fondamentale du système biologique qu'est la résistance et des imbrications de ce derniers avec la biosphère via l'intégration d'une vision multi-échelle de la résistance (Gil-Gil et al. 2021). De plus, couplé à des approches théoriques, cette stratégie doit passer préalablement par la mise en place de tests expérimentaux sur le risque de transmission des gènes de résistance vers ces bactéries sensibles.

Une deuxième stratégie, qui ne pourrait être appliquée qu'à des échelles très locales, serait d'utiliser les compromis adaptatifs des bactéries afin de rendre leur résistance maladaptative. Par exemple, l'utilisation de cocktails de bactériophages dans les hôpitaux pourrait sélectionner les bactéries pour une résistance à l'infection, aux dépens de leur résistance aux antibiotiques. Enfin, une stratégie qui a été formulée est le forçage génétique, afin de provoquer la disparition des gènes de résistance. A cet effet, plusieurs méthodes ont été formulées et/ou testées in vitro : des systèmes basés sur la technologie *CRISPR-Cas* sont capables d'éditer le génome de bactéries résistantes afin de les rendre sensibles de nouveau (Valderrama et al. 2019), et des systèmes basés sur l'interférence entre plasmides pourraient théoriquement supprimer les plasmides porteurs de résistances (Kamruzzaman et al. 2017).

## 6.3 Un problème global nécessitant une coordination globale

La circulation des résistances antibiotiques se produit non seulement entre écosystèmes, mais également entre continents. Si une famille de gènes pose problème dans une seule région du monde, ce n'est qu'une question de temps avant qu'elle devienne un problème de santé publique mondiale. De ce fait, une action globale et coordonnée est nécessaire : pour la surveillance de la propagation des résistances, pour le financement de la recherche et le partage des connaissances, et pour la mise en place de solutions. En particulier, de même que le confinement écologique de la résistance, la question du confinement géographique peut également se poser, bien qu'à ma connaissance, aucune législation ou recommandation dans ce sens n'ait été formulée. Une stratégie potentielle serait de surveiller la circulation des résistances dans les importations de marchandises, en procédant à des prélèvements aléatoires, et éventuellement de décider de quarantaines en cas de résistance avérée.

Toutefois, n'importe quelle intervention nécessite une coordination globale de l'ensemble des États et des organisations inter-gouvernementales. De nombreuses réglementations ont été suggérées (par exemple concernant l'usage des antibiotiques) et des feuilles de route ont été éditées par l'OMS ou des agences de santé locales, mais globalement, les effets ne suivent pas. En effet, si les plans d'action nationaux présentent effectivement les mêmes objectifs que le plan d'action global de l'OMS, les politiques implémentées correspondent assez peu à celles qui sont recommandées (Munkholm et Rubin 2020). En

réalité, aucun accord contraignant international (traité ou protocole) n'existe concernant la lutte contre la résistance aux antibiotiques. Un tel accord serait particulièrement intéressant car il permettrait une coordination économique : il permettrait par exemple d'éviter la compétition économique entre les producteurs de bétail qui restreignent l'usage des antibiotiques, et ceux qui ne le font pas (Ruckert et al. 2020). Cependant, un tel accord nécessiterait pour les pays signataires un soutien politique de la part des acteurs nationaux (Hoffman et Behdinan 2016), ce qui pourrait être entravé par une pression du secteur privé contre les régulations.

Cette absence de coordination globale soulève la question des inégalités entre pays développés et pays en développement. En effet, c'est dans les pays à faible revenu que la pression infectieuse est la plus importante, et ces pays ont le moins bon accès à des antibiotiques de qualité. C'est donc aux populations de ces pays qu'une action globale et coordonnée bénéficierait le plus. En outre, il ne s'agit pas que d'une question de réduction des inégalités, car ces pays sont aussi les plus susceptibles de contribuer à la propagation des résistances (Millar 2019).

# 7 Limitations méthodologiques

L'ambition de cette thèse est d'apporter une approche globale et intégrative pour l'étude de la propagation des résistances antibiotiques, en s'appuyant sur l'exemple des résistances aux aminoglycosides. En particulier, j'ai souhaité améliorer notre compréhension de la circulation écologique, géographique, et phylogénétique des gènes de résistance. Cependant, plusieurs limitations techniques peuvent réduire la portée des résultats de cette thèse.

## 7.1 Les limites de la détection bioinformatique

Les résultats de cette thèse reposent sur des méthodes bioinformatiques : pour détecter des gènes de résistance, pour détecter des éléments génétiques mobiles, pour inférer des événements de HGT, etc. Ces méthodes ont l'avantage de permettre la récolte rapide d'un grand nombre de données, et donc d'inférer des tendances à large échelle. Cependant, elles ne sont pas exemptes de limites.

En particulier, cette thèse repose sur un criblage par homologie des gènes encodant les AME. Comme montré sur l'exemple des plasmides dans la section III, la recherche d'homologie est en général très précise, et il y a donc peu de doute sur le fait que les clusters formés soient bien apparentés aux AME. Les fonctions de ces gènes ont été inférées de manière indirecte, sur la base de cette homologie. Pourtant, comme démontré par Feldgarden et al. 2019, le lien entre génotype et phénotype est particulièrement ténu dans le cas de la résistance aux aminoglycosides : une simple substitution peut suffire à changer la fonction de l'enzyme. De plus, il n'a pas été possible de collecter ni d'analyser les transcriptomes associés à chacun des génomes, et donc de déterminer le niveau d'expression de chaque gène. Comme discuté dans la section IV, l'expression de nombreux gènes de résistance est réprimées, par exemple

pour réduire leur coût d'expression en l'absence d'antibiotique. Il n'est donc pas possible d'affirmer avec certitude que l'ensemble des homologues identifiés confèrent systématiquement un phénotype de résistance aux aminoglycosides aux bactéries qui les portent, mais que ces bactéries portent de potentiels candidats à la résistance. De plus, il n'est pas possible d'affirmer que cette thèse étudie l'ensemble de la résistance aux aminoglycosides : d'une part car d'autres types de gènes peuvent conférer cette résistance, et d'autre part car certaines AME ont pu être laissées de côté par le criblage.

Cette limite des méthodes basées sur la recherche d'homologie peut avoir eu plusieurs conséquences sur la détection des MGE et des événements de HGT. En effet, l'exemple de *PlasForest* (section III) montre que, si la recherche d'homologies est effectivement très précise, elle est relativement peu sensible en dehors de son domaine de compétence. De fait, *PlasForest* est très peu capable de détecter des plasmides dans des métagénomes assemblés, ce qui suggère que des plasmides présents chez des espèces peu séquencées ont peu de chances d'avoir été identifiés dans le cadre de cette thèse. Par extension, et dans la mesure où toutes les recherches de MGE ont reposé sur la recherche d'homologies (*ISfinder, IntegronFinder, ICEFinder, Phispy*), cela pourrait expliquer pourquoi certains MGE, tels que les ICE ou les prophages, sont sous-représentés par rapport aux autres catégories dans le jeu de données. Au contraire, les méthodes sans alignement (par exemple, par recherche de k-mers) présentent une sensibilité beaucoup plus forte, au détriment de la précision. Cette sensibilité a permis l'identification de paires d'espèces entre lesquelles a eu lieu un événement de HGT, parmi des événements potentiels détectés par une méthode phylogénétique implicite, ainsi que l'orientation des transferts. Cependant la précision sur ces deux étapes d'analyse basées sur les distributions de k-mers est moyenne. Il a donc été nécessaire de faire usage de simulations afin d'établir un attendu sur la proximité nucléotidique minimale permettant de différencier entre plusieurs donneurs potentiels.

## 7.2 Des jeux de données biaisés et incomplets

Les conclusions de cette thèse reposent également sur la mise en relation, par des méthodes statistiques, des données génomiques avec d'autres jeux de données. Cependant, à toutes les étapes des analyses, j'ai pu faire le constat d'un problème de qualité des données. Plus précisément, de nombreuses données ont manqué ou n'étaient pas suffisamment précises pour pouvoir être exploitées, et les données disponibles étaient fortement biaisées.

Premièrement, une surabondance des pathogènes et commensaux de l'Homme et des animaux (liée à une surreprésentation des génomes échantillonnés dans un cadre clinique) empêche toute analyse fine de la propagation des résistances dans les écosystèmes naturels. De fait, ces écosystèmes ont dû être réunis en quelques catégories écologiques (sols, eau douce, faune et flore) qui ne correspondent pas aux classifications habituelles des biomes continentaux et qui ne permettent pas de juger précisément du degré d'anthropisation de ces écosystèmes. De plus, les métadonnées récupérées depuis la base de données *NCBI BioSamples* étaient très peu standardisées (avec un total de plus de 900 variables

différentes, pas toutes utilisées pour chaque génome), tant pour l'écologie, que pour les dates et lieux d'échantillonnage, ce qui a limité la plupart des analyses statistiques à environ un quart des génomes initialement criblés.

Un second point concerne la disponibilité des données publiées par les organismes publics. D'une part, les données relatives à la consommation d'aminoglycosides n'étaient publiquement disponibles que pour l'Europe, et sur des périodes extrêmement variables d'un pays à l'autre. A titre de comparaison, les CDC américains n'ont rendu publiques que les données de consommation de quelques autres catégories d'antibiotiques, et ce pour une période démarrant en 2011. De plus, il peut être constaté un manque crucial de données concernant les concentrations d'antibiotiques dans l'environnement : seules des données parcellaires, publiées dans des revues académiques, sont disponibles (Le Page et al. 2017) et les travaux de compilation restent encore trop rares (Gothwal et Shashidhar 2015). D'autre part, les données relatives aux échanges internationaux sont peu accessibles ou peu exploitables. Par exemple, la matrice bilatérale des échanges commerciaux publiée par l'UNCTAD comptabilise les échanges en dollars américains, et pas en tonnes. En outre, la matrice bilatérale des migrations humaines publiée par la Banque mondiale comptabilise le nombre de personnes migrant d'un pays à l'autre par décennie uniquement. Dans l'ensemble, le manque de données ou la faible exploitabilité des données disponibles oblige le recours aux extrapolations, ce qui limite la fiabilité des résultats obtenus.

# 8 Perspectives

## 8.1 L'apport des données métagénomiques à l'étude de la propagation des résistances antibiotiques

D'une part, comme discuté dans les partie 4.2 et partie 6.2, l'étude de la propagation des résistances antibiotiques pâtit d'un manque de données en dehors des environnements cliniques, et en particulier dans les écosystèmes naturels. D'autre part, les données de génome complet utilisées dans cette thèse ne peuvent être obtenues que pour des espèces cultivables. De nombreuses espèces bactériennes ne sont pas cultivables et leur contribution à la circulation des gènes de résistance est donc complètement ignorée. Intégrer des données métagénomiques devrait permettre s'affranchir au moins partiellement de ces deux limites. En effet, les métagénomes permettent d'intégrer toutes les bactéries sans discrimination envers certains taxons. Ces données apporteraient, en plus d'une exploration plus large et de jeux de données moins biaisés, la possibilité de déterminer avec plus de précision les prévalences de résistance et in fine une estimation des flux de résistance entre différents écosystèmes.

Cependant, bien que la métagénomique apporte un bénéfice quantitatif et augmente la sensibilité des jeux de données, cette dernière ne peut pas à elle seule nous éclairer sur la circulation de la résistance. En effet, bien que très sensibles, ces méthodes sont peu précises et bien qu'elles puissent être utilisées pour apporter des réponses à des questions épidémiologiques, elles ne permettent pas l'étude

de phénomènes complexes sur le système. Par exemple, la plupart des études de métagénomes peuvent analyser la composition en unités taxonomiques d'une communauté, en se basant sur les séquences d'ARN 16S. Ces analyses peuvent ensuite être couplées à des recherches de gènes de résistances. Cependant, la composition taxonomique d'une communauté dans laquelle on trouve des gènes de résistance n'indique pas quels taxons portent ces résistances. En revanche, et bien que ces méthodes ne permettent pas nécessairement d'identifier les origines des MGE à large spectre d'hôte, des traitements bioinformatiques complémentaires de ces données permettraient d'avoir une résolution plus fine pour l'analyse de la propagation des gènes de résistance. Depuis la dernière décennie, plusieurs méthodes de regroupement ont été développées, la plupart basées sur des méthodes sans alignement, pour délimiter des génomes parmi les contigs de métagénomes assemblés (voir par exemple Alneberg et al. 2014 ; Nissen et al. 2021). Ces méthodes ne permettent pas une reconstitution précise de tous les génomes composant un métagénome, mais augmentent le niveau de résolution du système. De fait, plusieurs bases de données de métagénomes regroupés en génomes existent déjà (Nayfach et al. 2020). Une assignation taxonomique précise de ces génomes est ensuite beaucoup plus simple : en effet, les 120 gènes utilisés pour définir la taxonomie *bac120* se retrouvent également dans les métagénomes (Parks et al. 2017), et l'outil d'assignation *GTDB-tk* (Chaumeil et al. 2020) peut leur être appliqué. Cependant, les outils actuellement à disposition pour la détection de MGE dans les métagénomes restent encore trop peu précis (voir section III).

Par ailleurs, une étude intégrative des facteurs écologiques favorisant la propagation des résistances antibiotiques nécessitera un degré de résolution écologique plus fin. A cet effet, la caractérisation en biomes des écosystèmes où les métagénomes sont échantillonnés, ainsi que la mesure de leurs paramètres abiotiques (salinité, température, pH, granulométrie, présence de polluants, etc.), pourraient permettre d'identifier le rôle de chacun de ces paramètres sur la sélection et la circulation des gènes de résistance. De plus, il sera tout aussi utile de mesurer la diversité de composition des communautés microbiennes, et la présence de certains macro-organismes susceptibles de transporter des bactéries résistantes d'un écosystème à l'autre. L'étude *in silico* d'écosystèmes pourrait aider à réduire en amont le nombre de paramètres à considérer dans les prélèvements.

## 8.2 Le rôle du système CRISPR-Cas dans le transfert horizontal de gènes

Comme discuté dans les partie 5.3 et partie 5.2.2, il existe plusieurs paramètres génomiques qui peuvent moduler le transfert horizontal des gènes. A ce titre, la présence du système *CRISPR-Cas* chez certaines espèces bactériennes pourrait constituer une barrière à l'acquisition de nouveaux gènes. Le système *CRISPR-Cas* a été initialement détecté chez *Streptococcus thermophilus* en tant que mécanisme de résistance aux infections à bactériophages virulents (Barrangou et al. 2007). Ce système repose sur l'action conjointe du locus *CRISPR* (*clustered regularly interspaced short palindromic repeats*) et des endonucléases Cas (*CRISPR-associated proteins*). Son efficacité est entretenue par l'intégration indépendante de nombreuses séquences d'origine virale (*spacers*) dans le génome bactérien. Il est important

de noter que la création de nouveaux génotypes de résistance suit un processus particulier dans le système *CRISPR-Cas* : cette mutation n'est pas aléatoire, mais dirigée vers le virus qui infecte la population (Koonin et Wolf 2009). Il s'agit donc d'un système immunitaire adaptatif héréditaire.

Le système *CRISPR-Cas* est donc capable de limiter le transfert horizontal de gènes par transduction. Cependant, il a été montré que *CRISPR-Cas* peut aussi mener au clivage de l'ADN plasmidique (Garneau et al. 2010), et donc limiter l'acquisition de nouveaux gènes par conjugaison (Jiang, Maniv et al. 2013). De nombreux *spacers* correspondant à des locus d'ICE ont également été retrouvés chez certaines bactéries (Wheatley et MacLean 2020). Ce système est même en mesure d'empêcher l'acquisition d'ADN exogène par transformation chez *Neisseria meningitidis* (Zhang, Heidrich et al. 2013). De manière générale, les génomes qui encodent le système *CRISPR-Cas* sont plus petits (Wheatley et MacLean 2020), ce qui pourrait être dû à la forte limite qu'ils imposent au HGT.

Il serait donc intéressant de réanalyser le jeu de données constitué par cette thèse à la lumière de la présence ou de l'absence du système *CRISPR-Cas* dans chaque génome. Le pipeline CRISPRCasFinder (Couvin et al. 2018) permettrait une recherche rapide des locus *CRISPR* et *Cas* dans les 160987 génomes déjà criblés pour les gènes encodant les AME. Des analyses statistiques subséquentes permettraient de déterminer si la distribution des AME est corrélée à l'absence du système *CRISPR-Cas*, si ce système limite l'accumulation et la diversité de ces gènes, et s'il constitue une barrière à leur acquisition par HGT.

## 8.3 Modéliser la propagation d'un gène de résistance dans une métacommunauté

Les travaux de cette thèse montrent dans leur ensemble que la propagation des résistances antibiotiques est un phénomène complexe, qui dépasse les frontières de la sélection et du transfert horizontal seuls. De ce fait, ce champ bénéficierait de l'utilisation d'approches théoriques issues de la biologie des systèmes (en particulier, des écosystèmes). Par exemple, au vu des similarités entre certains résistomes, j'ai fait l'hypothèse que cette similarité pouvait être expliquée par des flux de gènes et de bactéries entre écosystèmes. De même, j'ai évoqué le rôle des espèces généralistes et de la mondialisation des échanges dans la propagation des gènes de résistance entre écosystèmes et entre continents. Ces hypothèses restent toutefois à tester. Or, il est actuellement compliqué de maintenir des communautés complexes en laboratoire, et de contrôler finement les paramètres d'un écosystème (structuration spatiale, modulation des ressources, etc.). La modélisation d'une méta-communauté donnerait donc la possibilité de tester *in silico* si les différents paramètres évoqués sont à même d'expliquer la distribution actuelle des gènes de résistance.

Cette approche consisterait à simuler des populations, leurs environnements, et leurs interactions, par exemple par un modèle individu-centré où chaque individu représenterait une bactérie ou un clone bactérien. La modélisation de certains paramètres locaux, tels que la concentration locale en antibiotiques, la compétition, ou la structuration spatiale, permettraient de prendre en compte les variations

locales dans la sélection pour les gènes de résistance et la probabilité de leur transfert horizontal entre souches. De plus, pour prendre en compte la propagation des bactéries et des gènes entre communautés, certains processus de flux de matière pourraient également être incorporés au modèle : transports ponctuels (par des animaux ou des activités humaines) ou flux continu (par le vent et les courants). En outre, la modélisation d'une diversité de souches pourrait prendre en compte la variabilité de capacité des bactéries à accumuler les gènes de résistance et/ou les transporter d'un écosystème à l'autre. Pour ce faire, ces caractéristiques pourraient inclure les variables suivantes : résistance innée aux antibiotiques, taux de croissance intrinsèque, capacité à vivre dans plusieurs écosystèmes, capacité à recevoir des gènes exogènes par HGT, et capacité à se fixer aux vecteurs du transport entre écosystèmes.

Un tel modèle permettrait de tester les différentes hypothèses soulevées dans cette thèse concernant l'émergence et la circulation des gènes de résistance au sein d'une communauté, mais également leur propagation dans le réseau d'une métacommunauté. En particulier, plusieurs scénarios pourraient être testés : par exemple, avec une distribution inégale des antibiotiques entre communautés, en présence ou en l'absence d'espèces généralistes, en incluant ou non des gènes de résistance dans les communautés initiales. Ainsi, la prévalence totale des gènes de résistance ou leur répartition entre communautés et entre souches pourraient être comparées d'une part d'un scénario à l'autre, et d'autre part d'un scénario aux données observées. Cette comparaison permettrait la confirmation ou l'infirmation de certaines hypothèses, mais aussi de faire des prédictions sur la distribution des résistances dans des contextes encore mal connus .

# Section VII : Bibliographie


Abass, Abdulai, Frederick Adzitey et Nurul Huda (déc. 2020). « Escherichia coli of Ready-to-Eat (RTE) Meats Origin Showed Resistance to Antibiotics Used by Farmers ». *Antibiotics* 9.12, p. 869.

Abe, Kimihiro, Nobuhiko Nomura et Satoru Suzuki (mai 2020). « Biofilms : hot spots of horizontal gene transfer (HGT) in aquatic environments, with a focus on a new HGT mechanism ». *FEMS Microbiology Ecology* 96.5, p. 31.

Aldrin, Magne, Ragnhild Raastad, Ingunn Fride Tvete, Dag Berild, Arnoldo Frigessi, Truls Leegaard, Dominique L. Monnet, Mette Walberg et Fredrik Müller (avr. 2013). « Antibiotic resistance in hospitals : a ward-specific random effect model in a low antibiotic consumption environment ». *Statistics in Medicine* 32.8, p. 1407-1418.

Allen, Heather K., Justin Donato, Helena Huimi Wang, Karen A. Cloud-Hansen, Julian Davies et Jo Handelsman (mars 2010). « Call of the wild : antibiotic resistance genes in natural environments ». *Nature Reviews Microbiology* 8.4, p. 251-259.

Almakki, Ayad, Estelle Jumas-Bilak, Hélène Marchandin et Patricia Licznar-Fajardo (juin 2019). « Antibiotic resistance in urban runoff ». *Science of The Total Environment* 667, p. 64-76.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino De Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson et Christopher Quince (sept. 2014). « Binning metagenomic contigs by coverage and composition ». *Nature Methods* 11.11, p. 1144-1146.

Aminov, Rustam I. (déc. 2009). « The role of antibiotics and antibiotic resistance in nature ». *Environmental Microbiology* 11.12, p. 2970-2988.

Aminov, Rustam I. et Roderick I. Mackie (juin 2007). « Evolution and ecology of antibiotic resistance genes ». *FEMS Microbiology Letters* 271.2, p. 147-161.

Amorós-Moya, Dolors, Stéphanie Bedhomme, Marita Hermann et Ignacio G. Bravo (sept. 2010). « Evolution in Regulatory Regions Rapidly Compensates the Cost of Nonoptimal Codon Usage ». *Molecular Biology and Evolution* 27.9, p. 2141-2151.

Andersson, Dan I, Sophie Maisnier Patin, Annika I Nilsson et Elisabeth Kugelberg (avr. 2007). « The Biological Cost of Antibiotic Resistance ». *Enzyme-Mediated Resistance to Antibiotics*. Sous la dir. de Robert A. Bonomo et Marcelo Tolmasky. John Wiley & Sons, Ltd, p. 339-348.

Andersson, Dan I. et Diarmaid Hughes (mars 2010). « Antibiotic resistance and its cost : is it possible to reverse resistance ? » *Nature Reviews Microbiology* 8.4, p. 260-271.

— (mai 2014). « Microbiological effects of sublethal levels of antibiotics ». *Nature Reviews Microbiology* 12.7, p. 465-478.

Årdal, Christine et al. (jan. 2016). « International cooperation to improve access to and sustain effectiveness of antimicrobials ». *The Lancet* 387.10015, p. 296-307.

Attaiech, Laetitia, Aïda Boughammoura, Céline Brochier-Armanet, Omran Allatif, Flora Peillard-Fiorente, Ross A. Edwards, Ayat R. Omar, Andrew M. MacMillan, Mark Glover et Xavier

Charpentier (août 2016). « Silencing of natural transformation by an RNA chaperone and a multitarget small RNA ». *Proceedings of the National Academy of Sciences of the United States of America* 113.31, p. 8813-8818.

Baker-Austin, Craig, Meredith S. Wright, Ramunas Stepanauskas et J. V. McArthur (avr. 2006). « Co-selection of antibiotic and metal resistance ». *Trends in Microbiology* 14.4, p. 176-182.

Baltrus, David A. (août 2013). « Exploring the costs of horizontal gene transfer ». *Trends in Ecology and Evolution* 28.8, p. 489-495.

Baquero, F., J. L. Martínez, V. F. Lanza, J. Rodríguez-Beltrán, J. C. Galán, A. San Millán, R. Cantón et T. M. Coque (juin 2021). « Evolutionary Pathways and Trajectories in Antibiotic Resistance ». *Clinical Microbiology Reviews*.

Baquero, Fernando, Val F. Lanza, Rafael Cantón et Teresa M. Coque (mars 2015). « Public health evolutionary biology of antimicrobial resistance : priorities for intervention ». *Evolutionary Applications* 8.3, p. 223-239.

Barrangou, Rodolphe, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A. Romero et Philippe Horvath (mars 2007). « CRISPR provides acquired resistance against viruses in prokaryotes ». *Science* 315.5819, p. 1709-1712.

Barrios, Renys E., Himanshu K. Khuntia, Shannon L. Bartelt-Hunt, John E. Gilley, Amy M. Schmidt, Daniel D. Snow et Xu Li (avr. 2020). « Fate and transport of antibiotics and antibiotic resistance genes in runoff and soil as affected by the timing of swine manure slurry application ». *Science of The Total Environment* 712, p. 136505.

Bartell, Jennifer A., David R. Cameron, Biljana Mojsoska, Janus Anders Juul Haagensen, Tacjana Pressler, Lea M. Sommer, Kim Lewis, Søren Molin et Helle Krogh Johansen (déc. 2020). « Bacterial persisters in long-term infection : Emergence and fitness in a complex host environment ». *PLOS Pathogens* 16.12, e1009112.

Batten, L. W. (1955). « Discussion on the use and abuse of antibiotics ». *Proceedings of the Royal Society of Medicine* 48, p. 355-364.

Bedhomme, S., D. Perez Pantoja et I. G. Bravo (avr. 2017). « Plasmid and clonal interference during post horizontal gene transfer evolution ». *Molecular Ecology* 26.7, p. 1832-1847.

Bedhomme, Stephanie, Dolors Amorós-Moya, Luz M. Valero, Nùria Bonifaci, Miquel Àngel Pujana, Ignacio G. Bravo et Josefa Gonzalez (mars 2019). « Evolutionary Changes after Translational Challenges Imposed by Horizontal Gene Transfer ». *Genome Biology and Evolution* 11.3, p. 814-831.

Bengtsson-Palme, Johan, Erik Kristiansson et D. G.Joakim Larsson (2018). « Environmental factors influencing the development and spread of antibiotic resistance ». *FEMS Microbiology Reviews* 42.1, p. 68-80.

Benveniste, R. et J. Davies (1973). « Structure-Activity Relationships Among the Aminoglycoside Antibiotics : Role of Hydroxyl and Amino Groups ». *Antimicrobial Agents and Chemotherapy* 4.4, p. 402-409.

Bershtein, Shimon, Adrian W.R. Serohijos, Sanchari Bhattacharyya, Michael Manhart, Jeong Mo Choi, Wanmeng Mu, Jingwen Zhou et Eugene I. Shakhnovich (2015). « Protein Homeostasis Imposes

a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria ». *PLOS Genetics* 11.10, e1005612.

Björkman, Johanna, Diarmaid Hughes et Dan I. Andersson (mars 1998). « Virulence of antibiotic-resistant Salmonella typhimurium ». *Proceedings of the National Academy of Sciences* 95.7, p. 3949-3953.

Blair, Jessica M.A., Mark A. Webber, Alison J. Baylay, David O. Ogbolu et Laura J.V. Piddock (jan. 2015). *Molecular mechanisms of antibiotic resistance*.

Blázquez, Jesús et José María Gómez-Gómez (avr. 2014). « Evolution of Antibiotic Resistance by Hypermutation ». *Evolutionary Biology of Bacterial and Fungal Pathogens*, p. 319-331.

Bonafede, Mary E., Lenore L. Carias et Louis B. Rice (1997). « Enterococcal transposon Tn5384 : evolution of a composite transposon through cointegration of enterococcal and staphylococcal plasmids ». *Antimicrobial Agents and Chemotherapy* 41.9, p. 1854-1858.

Bondarczuk, Kinga, Anna Markowicz et Zofia Piotrowska-Seget (fév. 2016). « The urgent need for risk assessment on the antibiotic resistance spread via sewage sludge land application ». *Environment International* 87, p. 49-55.

Bonomo, Jeanne et Ryan T. Gill (avr. 2005). « Amino acid content of recombinant proteins influences the metabolic burden response ». *Biotechnology and Bioengineering* 90.1, p. 116-126.

Botzman, Maya et Hanah Margalit (oct. 2011). « Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles ». *Genome Biology* 12.10, p. 1-11.

Brückner, Reinhold et Fritz Titgemeyer (avr. 2002). « Carbon catabolite repression in bacteria : choice of the carbon source and autoregulatory limitation of sugar utilization ». *FEMS Microbiology Letters* 209.2, p. 141-148.

Bud, R. (2007). *Penicillin : Triumph and tragedy*. New York : Oxford University Press.

Bundschuh, Mirco, Torsten Hahn, Mark O. Gessner et Ralf Schulz (jan. 2009). « Antibiotics as a chemical stressor affecting an aquatic decomposer–detritivore system ». *Environmental Toxicology and Chemistry* 28.1, p. 197-203.

Cairns, Johannes, Lasse Ruokolainen, Jenni Hultman, Manu Tamminen, Marko Virta et Teppo Hiltunen (avr. 2018). « Ecology determines how low antibiotic concentration impacts community composition and horizontal transfer of resistance genes ». *Communications Biology* 1.1, p. 1-8.

Callens, Martijn, Léa Pradier, Michael Finnegan, Caroline Rose et Stéphanie Bedhomme (sept. 2021). « Read between the Lines : Diversity of Nontranslational Selection Pressures on Local Codon Usage ». *Genome Biology and Evolution* 13.9.

Callens, Martijn, Celine Scornavacca et Stéphanie Bedhomme (2021). « Evolutionary responses to codon usage of horizontally transferred genes in Pseudomonas aeruginosa : gene retention, amelioration and compensatory evolution ». *Microbial Genomics* 7.6.

Centers for Disease Control and Prevention (2013). *Antibiotic resistance threats in the United States*. Rapp. tech.

Cerqueira, Francisco, Víctor Matamoros, Josep M. Bayona, Thomas U. Berendonk, Goffe Elsinga, Luc M. Hornstra et Benjamin Piña (oct. 2019). « Antibiotic resistance gene distribution in agricultural fields and crops. A soil-to-food analysis ». *Environmental Research* 177, p. 108608.

Chao, Lin et Edward C. Cox (jan. 1983). « Competition Between High and Low Mutating Strains of Escherichia coli ». *Evolution* 37.1, p. 125.

Chaumeil, Pierre Alain, Aaron J. Mussig, Philip Hugenholtz et Donovan H. Parks (mars 2020). « GTDB-Tk : a toolkit to classify genomes with the Genome Taxonomy Database ». *Bioinformatics* 36.6, p. 1925-1927.

Che, You, Yu Yang, Xiaoqing Xu, Karel Brinda, Martin F. Polz, William P. Hanage et Tong Zhang (fév. 2021). « Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes ». *Proceedings of the National Academy of Sciences of the United States of America* 118.6, e2008731118.

Chen, Qing Lin, Xin Li An, Bang Xiao Zheng, Michael Gillings, Josep Peñuelas, Li Cui, Jian Qiang Su et Yong Guan Zhu (juill. 2019). « Loss of soil microbial diversity exacerbates spread of antibiotic resistance ». *Soil Ecology Letters* 1.1, p. 3-13.

Chen, Ying, Ido Golding, Satoshi Sawai, Ling Guo et Edward C. Cox (2005). « Population Fitness and the Regulation of Escherichia coli Genes by Bacterial Viruses ». *PLOS Biology* 3.7, e229.

Chopra, Ian (juin 2002). « New developments in tetracycline antibiotics : glycylcyclines and tetracycline efflux pump inhibitors ». *Drug Resistance Updates* 5.3-4, p. 119-125.

Clardy, Jon, Michael A. Fischbach et Cameron R. Currie (juin 2009). « The natural history of antibiotics ». *Current Biology* 19.11, R437-R441.

Clewell, D. B., Y. Yagi et B. Bauer (mai 1975). « Plasmid-determined tetracycline resistance in Streptococcus faecalis : evidence for gene amplification during growth in presence of tetracycline ». *Proceedings of the National Academy of Sciences* 72.5, p. 1720-1724.

Conlan, Sean et al. (sept. 2014). « Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae ». *Science Translational Medicine* 6.254.

Cooper, Robert M., Lev Tsimring et Jeff Hasty (nov. 2017). « Inter-species population dynamics enhance microbial horizontal gene transfer and spread of antibiotic resistance ». *eLife* 6.

Cottell, Jennifer L., Mark A. Webber, Nick G. Coldham, Dafydd L. Taylor, Anna M. Cerdeño-Tárraga, Heidi Hauser, Nicholas R. Thomson, Martin J. Woodward et Laura J.V. Piddock (avr. 2011). « Complete Sequence and Molecular Epidemiology of IncK Epidemic Plasmid Encoding blaCTX-M-14 ». *Emerging Infectious Diseases* 17.4, p. 645.

Couvin, David, Aude Bernheim, Claire Toffano-Nioche, Marie Touchon, Juraj Michalik, Bertrand Néron, Eduardo P.C. Rocha, Gilles Vergnaud, Daniel Gautheret et Christine Pourcel (juill. 2018). « CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins ». *Nucleic Acids Research* 46.W1, W246-W251.

Cox, Georgina et Gerard D. Wright (août 2013). « Intrinsic antibiotic resistance : Mechanisms, origins, challenges and solutions ». *International Journal of Medical Microbiology* 303.6-7, p. 287-292.

Cury, Jean, Marie Touchon et Eduardo P. C. Rocha (juill. 2017). « Integrative and conjugative elements and their hosts : composition, distribution and organization ». *Nucleic Acids Research* 45.15, p. 8943-8956.

Cycoń, Mariusz, Agnieszka Mrozik et Zofia Piotrowska-Seget (2019). « Antibiotics in the soil environment—degradation and their impact on microbial activity and diversity ». *Frontiers in Microbiology* 10.3, p. 338.

D'Costa, Vanessa M. et al. (sept. 2011). *Antibiotic resistance is ancient*.

Dagan, Tal, Yael Artzy-Randrup et William Martin (juill. 2008). « Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution ». *Proceedings of the National Academy of Sciences* 105.29, p. 10039-10044.

Davies, Julian (juill. 2006). « Are antibiotics naturally antibiotics ? » *Journal of Industrial Microbiology and Biotechnology* 33.7, p. 496-499.

Davies, S. C. (2013). *Annual report of the chief medical officer*. Rapp. tech. London : Department of Health.

Desmet, Stefanie, Suruchi Nepal, Jan Maarten van Dijl, Marc Van Ranst, Monika A. Chlebowicz, John W. Rossen, Jeroen K.J. Van Houdt, Piet Maes, Katrien Lagrou et Erik Bathoorn (mars 2018). « Antibiotic resistance plasmids cointegrated into a megaplasmid harboring the bla OXA-427 carbapenemase gene ». *Antimicrobial Agents and Chemotherapy* 62.3.

Ding, Feng, Petrus Tang, Mei Hua Hsu, Peng Cui, Songnian Hu, Jun Yu et Cheng Hsun Chiu (avr. 2009). « Genome evolution driven by host adaptations results in a more virulent and antimicrobial-resistant Streptococcus pneumoniae serotype 14 ». *BMC Genomics* 10.1, p. 1-13.

Dolejska, Monika et Ivan Literak (2019). « Wildlife Is Overlooked in the Epidemiology of Medically Important Antibiotic-Resistant Bacteria ». *Antimicrobial Agents and Chemotherapy* 63.8.

Done, Hansa Y., Arjun K. Venkatesan et Rolf U. Halden (mai 2015). « Does the Recent Growth of Aquaculture Create Antibiotic Resistance Threats Different from those Associated with Land Animal Production in Agriculture ? » *The AAPS journal* 17.3, p. 513-524.

Douglass, J. et L. M. Steyn (1993). « A ribosomal gene mutation in streptomycin-resistant Mycobacterium tuberculosis isolates ». *Journal of Infectious Diseases* 167.6, p. 1505-1506.

Driessen, Bert, Louis Freson et Johan Buyse (nov. 2020). « Fasting Finisher Pigs before Slaughter Influences Pork Safety, Pork Quality and Animal Welfare ». *Animals* 10.12, p. 2206.

Dugatkin, Lee Alan, Michael Perlin, J. Scott Lucas et Ronald Atlas (jan. 2005). « Group-beneficial traits, frequency-dependent selection and genotypic diversity : an antibiotic resistance paradigm ». *Proceedings of the Royal Society B : Biological Sciences* 272.1558, p. 79-83.

Duijkeren, Engeline van et al. (avr. 2019). « The use of aminoglycosides in animals within the EU : development of resistance in animals and possible impact on human and animal health : a review ». *Journal of Antimicrobial Chemotherapy* 74.9, p. 2480-2496.

Dunning Hotopp, Julie C. (avr. 2011). « Horizontal gene transfer between bacteria and animals ». *Trends in Genetics* 27.4, p. 157-163.

Ebmeyer, Stefan, Erik Kristiansson et D. G.Joakim Larsson (jan. 2021). « A framework for identifying the recent origins of mobile antibiotic resistance genes ». *Communications Biology* 4.1, p. 1-10.

Eggen, Trine, Tone Normann Asp, Kari Grave et Victor Hormazabal (sept. 2011). « Uptake and translocation of metformin, ciprofloxacin and narasin in forage- and crop plants ». *Chemosphere* 85.1, p. 26-33.

Ehrlich, Paul et S. Hata (1910). *Die experimentelle Chemotherapie der Spirillosen*. Springer Berlin Heidelberg.

Enne, Virve I., Peter M. Bennett, David M. Livermore et Lucinda M.C. Hall (juin 2004). « Enhancement of host fitness by the sul2-coding plasmid p9123 in the absence of selective pressure ». *Journal of Antimicrobial Chemotherapy* 53.6, p. 958-963.

Epp Schmidt, Dietrich J., David Johan Kotze et al. (oct. 2019). « Metagenomics Reveals Bacterial and Archaeal Adaptation to Urban Land-Use : N Catabolism, Methanogenesis, and Nutrient Acquisition ». *Frontiers in Microbiology* 10, p. 2330.

Epp Schmidt, Dietrich J., Richard Pouyat, Katalin Szlavecz, Heikki Setälä, D. Johan Kotze, Ian Yesilonis, Sarel Cilliers, Erzsébet Hornung, Miklós Dombos et Stephanie A. Yarwood (avr. 2017). « Urbanization erodes ectomycorrhizal fungal diversity and may cause microbial communities to converge ». *Nature Ecology & Evolution* 1.5, p. 1-9.

Fajardo, Alicia et José L. Martínez (avr. 2008). « Antibiotics as signals that trigger specific bacterial responses ». *Current Opinion in Microbiology* 11.2, p. 161-167.

Feldgarden, Michael et al. (2019). « Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates ». *Antimicrobial Agents and Chemotherapy* 63, e00483-19.

Fernández-Oliva, Alberto, Paula Ortega-González et Cristina Risco (mars 2019). « Targeting host lipid flows : Exploring new antiviral and antibiotic strategies ». *Cellular Microbiology* 21.3, e12996.

Finley, Rita L., Peter Collignon, D. G.Joakim Larsson, Scott A. Mcewen, Xian Zhi Li, William H. Gaze, Richard Reid-Smith, Mohammed Timinouni, David W. Graham et Edward Topp (sept. 2013). « The Scourge of Antibiotic Resistance : The Important Role of the Environment ». *Clinical Infectious Diseases* 57.5, p. 704-710.

Fleming, Alexander (1945). « Penicillin ». *Nobel Lecture*. Elsevier Publishing Company.

Floyd, Jody L., Kenneth P. Smith, Sanath H. Kumar, Jared T. Floyd et Manuel F. Varela (déc. 2010). « LmrS is a multidrug efflux pump of the major facilitator superfamily from Staphylococcus aureus ». *Antimicrobial Agents and Chemotherapy* 54.12, p. 5406-5412.

Fondi, Marco et Renato Fani (déc. 2010). « The horizontal flow of the plasmid resistome : clues from inter-generic similarity networks ». *Environmental Microbiology* 12.12, p. 3228-3242.

Forsberg, Kevin J, Alejandro Reyes, Bin Wang, Elizabeth M Selleck, Morten O A Sommer et Gautam Dantas (2012). « The shared antibiotic resistome of soil bacteria and human pathogens ». *Science* 337.6098, p. 1107-1111.

FRIEDMAN, Mendel (avr. 2015). « Antibiotic-Resistant Bacteria : Prevalence in Food and Inactivation by Food-Compatible Compounds and Plant Extracts ». *Journal of Agricultural and Food Chemistry* 63.15, p. 3805-3822.

FROST, Laura S., Raphael LEPLAE, Anne O. SUMMERS et Ariane TOUSSAINT (sept. 2005). « Mobile genetic elements : the agents of open source evolution ». *Nature Reviews Microbiology* 3.9, p. 722-732.

FURUYA, E. Yoko et Franklin D. LOWY (jan. 2006). « Antimicrobial-resistant bacteria in the community setting ». *Nature Reviews Microbiology* 4.1, p. 36-45.

GALTIER, N., G. PIGANEAU, D. MOUCHIROUD et L. DURET (oct. 2001). « GC-Content Evolution in Mammalian Genomes : The Biased Gene Conversion Hypothesis ». *Genetics* 159.2, p. 907-911.

GAMA, João Alves, Rita ZILHÃO et Francisco DIONISIO (sept. 2017). « Multiple plasmid interference – Pledging allegiance to my enemy's enemy ». *Plasmid* 93, p. 17-23.

GAO, Fang Zhou, Liang Ying HE, Lu Xi HE, Hai Yan ZOU, Min ZHANG, Dai Ling WU, You Sheng LIU, Yi Jing SHI, Hong BAI et Guang Guo YING (nov. 2020). « Untreated swine wastes changed antibiotic resistance and microbial community in the soils and impacted abundances of antibiotic resistance genes in the vegetables ». *Science of The Total Environment* 741, p. 140482.

GARNEAU, Josiane E., Marie Ève DUPUIS, Manuela VILLION, Dennis A. ROMERO, Rodolphe BARRANGOU, Patrick BOYAVAL, Christophe FREMAUX, Philippe HORVATH, Alfonso H. MAGADÁN et Sylvain MOINEAU (nov. 2010). « The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA ». *Nature* 468.7320, p. 67-71.

GARRETT, L. (1994). *The coming plague : newly emerging diseases in a world out of balance*. New York : Farrar, Straus et Giroux.

GENNARO, M. L., J. KORNBLUM et R. P. NOVICK (1987). « A site-specific recombination function in Staphylococcus aureus plasmids ». *Journal of Bacteriology* 169.6, p. 2601-2610.

GEORGE, Anna (juin 2018). « Antimicrobial resistance, trade, food safety and security ». *One Health* 5, p. 6.

GESER, Nadine, Roger STEPHAN et Herbert HÄCHLER (mars 2012). « Occurrence and characteristics of extended-spectrum $\beta$-lactamase (ESBL) producing Enterobacteriaceae in food producing animals, minced meat and raw milk ». *BMC Veterinary Research* 8.1, p. 1-9.

GIBBS, C. W., J. T. GIBSON et D. S. NEWTON (oct. 1973). « Drug utilization review of actual versus preferred pediatric antibiotic therapy ». *American Journal of Hospital Pharmacy* 30.10, p. 892-897.

GIL-GIL, Teresa, Luz Edith OCHOA-SÁNCHEZ, Fernando BAQUERO et José Luis MARTÍNEZ (jan. 2021). « Antibiotic resistance : Time of synthesis in a post-genomic age ». *Computational and Structural Biotechnology Journal* 19, p. 3110-3124.

GIRAUD, A., I. MATIC, O. TENAILLON, A. CLARA, M. RADMAN, M. FONS et F. TADDEI (mars 2001). « Costs and benefits of high mutation rates : Adaptive evolution of bacteria in the mouse gut ». *Science* 291.5513, p. 2606-2608.

*Global action plan on antimicrobial resistance* (2021).

Gorrie, Claire L. et al. (juill. 2018). « Antimicrobial-Resistant Klebsiella pneumoniae Carriage and Infection in Specialized Geriatric Care Wards Linked to Acquisition in the Referring Hospital ». *Clinical Infectious Diseases* 67.2, p. 161-170.

Gothwal, Ritu et Thhatikkonda Shashidhar (avr. 2015). « Antibiotic Pollution in the Environment : A Review ». *CLEAN – Soil, Air, Water* 43.4, p. 479-489.

Griffith, Fred (1928). « The Significance of Pneumococcal Types ». *Epidemiology & Infection* 27.2, p. 113-159.

Grinten, Esther van der, Mariël G. Pikkemaat, Evert Jan van den Brandhof, Gerard J. Stroomberg et Michiel H.S. Kraak (juin 2010). « Comparing the sensitivity of algal, cyanobacterial and bacterial bioassays to different groups of antibiotics ». *Chemosphere* 80.1, p. 1-6.

Guerrero-Ramos, Emilia, Jorge Cordero, Diana Molina-González, Patrícia Poeta, Gilberto Igrejas, Carlos Alonso-Calleja et Rosa Capita (fév. 2016). « Antimicrobial resistance and virulence genes in enterococci from wild game meat in Spain ». *Food Microbiology* 53, p. 156-164.

Gullberg, Erik, Sha Cao, Otto G. Berg, Carolina Ilbäck, Linus Sandegren, Diarmaid Hughes et Dan I. Andersson (juill. 2011). « Selection of Resistant Bacteria at Very Low Antibiotic Concentrations ». *PLOS Pathogens* 7.7, e1002158.

Guo, Jiahua, Alistair Boxall et Katherine Selby (déc. 2015). « Do Pharmaceuticals Pose a Threat to Primary Producers ? » *http://dx.doi.org/10.1080/10643389.2015.1061873* 45.23, p. 2565-2610.

Guo, Xing pan, Yi Yang et al. (fév. 2018). « Biofilms as a sink for antibiotic resistance genes (ARGs) in the Yangtze Estuary ». *Water Research* 129, p. 277-286.

Gurney, James, Léa Pradier, Joanne S. Griffin, Claire Gougat-Barbera, Benjamin K. Chan, Paul E. Turner, Oliver Kaltz et Michael E. Hochberg (jan. 2020). « Phage steering of antibiotic-resistance evolution in the bacterial pathogen, Pseudomonas aeruginosa ». *Evolution, Medicine, and Public Health* 2020.1, p. 148-157.

Hall, Clayton W. et Thien Fah Mah (mai 2017). « Molecular mechanisms of biofilm-based antibiotic resistance and tolerance in pathogenic bacteria ». *FEMS Microbiology Reviews* 41.3, p. 276-301.

Harrison, Ellie, David Guymer, Andrew J. Spiers, Steve Paterson et Michael A. Brockhurst (août 2015). « Parallel Compensatory Evolution Stabilizes Plasmids across the Parasitism-Mutualism Continuum ». *Current Biology* 25.15, p. 2034-2039.

Hassell, James M., Michael Begon, Melissa J. Ward et Eric M. Fèvre (jan. 2017). « Urbanization and Disease Emergence : Dynamics at the Wildlife–Livestock–Human Interface ». *Trends in Ecology & Evolution* 32.1, p. 55-67.

He, Pinjing, Yizhou Zhou, Liming Shao, Jinghua Huang, Zhan Yang et Fan Lü (déc. 2019). « The discrepant mobility of antibiotic resistant genes : Evidence from their spatial distribution in sewage sludge flocs ». *Science of The Total Environment* 697, p. 134176.

He, Ya, Qingbin Yuan, Jacques Mathieu, Lauren Stadler, Naomi Senehi, Ruonan Sun et Pedro J.J. Alvarez (fév. 2020). « Antibiotic resistance genes from livestock waste : occurrence, dissemination, and treatment ». *npj Clean Water* 3.1, p. 1-11.

Hernando-Amado, Sara, Teresa M. Coque, Fernando Baquero et José L. Martínez (août 2019). « Defining and combating antibiotic resistance from One Health and Global Health perspectives ». *Nature Microbiology* 4.9, p. 1432-1442.

Herrmann, Dustin L, Laura A Schifman et William D Shuster (oct. 2020). « Urbanization drives convergence in soil profile texture and carbon content ». *Environmental Research Letters* 15.11, p. 114001.

Hershberg, Ruth et Dmitri A. Petrov (sept. 2010). « Evidence That Mutation Is Universally Biased towards AT in Bacteria ». *PLOS Genetics* 6.9, e1001115.

Hoffman, Steven J. et Asha Behdinan (2016). « Towards an International Treaty on Antimicrobial Resistance ». *Ottawa Law Review* 47.2.

Hussar, A. E. et H. L. Holley (1954). *Antibiotics and antibiotic therapy*. New York : The MacMillan Compagny.

Hyun, Jason C., Jonathan M. Monk et Bernhard O. Palsson (déc. 2022). « Comparative pangenomics : analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity ». *BMC Genomics* 23.1, p. 1-18.

Jansen, Gunther et C Athena Aktipis (2014). « Resistance Is Mobile : The Accelerating Evolution of Mobile Genetic Elements Encoding Resistance ». *Journal of Evolutionary Medicine* 2.

Jiang, Wenyan, Inbal Maniv, Fawaz Arain, Yaying Wang, Bruce R. Levin et Luciano A. Marraffini (sept. 2013). « Dealing with the Evolutionary Downside of CRISPR Immunity : Bacteria and Beneficial Plasmids ». *PLOS Genetics* 9.9, e1003844.

Jiang, Xinglin, Mostafa M.Hashim Ellabaan, Pep Charusanti, Christian Munck, Kai Blin, Yaojun Tong, Tilmann Weber, Morten O.A. Sommer et Sang Yup Lee (juin 2017). « Dissemination of antibiotic resistance genes from antibiotic producers to pathogens ». *Nature Communications* 8.1, p. 1-7.

Jjemba, Patrick K. (jan. 2006). « Excretion and ecotoxicity of pharmaceutical and personal care products in the environment ». *Ecotoxicology and Environmental Safety* 63.1, p. 113-130.

Johnsborg, Ola, Vegard Eldholm et Leiv Sigve Håvarstein (déc. 2007). « Natural genetic transformation : prevalence, mechanisms and function ». *Research in Microbiology* 158.10, p. 767-778.

Johnson, Renée (2015). *U.S.-EU Poultry Dispute on the Use of Pathogen Reduction Treatments (PRTs)*. Rapp. tech.

Johnson, Timothy J. et Lisa K. Nolan (déc. 2009). «  Pathogenomics of the Virulence Plasmids of Escherichia coli  ». *Microbiology and Molecular Biology Reviews* 73.4, p. 750-774.

Jorth, Peter et al. (sept. 2015). « Regional Isolation Drives Bacterial Diversification within Cystic Fibrosis Lungs ». *Cell Host & Microbe* 18.3, p. 307-319.

Kamruzzaman, Muhammad, Shereen Shoma, Christopher M. Thomas, Sally R. Partridge et Jonathan R. Iredell (fév. 2017). « Plasmid interference for curing antibiotic resistance plasmids in vivo ». *PLOS ONE* 12.2, e0172913.

Klein, Eili Y., Thomas P. Van Boeckel, Elena M. Martinez, Suraj Pant, Sumanth Gandra, Simon A. Levin, Herman Goossens et Ramanan Laxminarayan (avr. 2018). « Global increase and geographic conver-

gence in antibiotic consumption between 2000 and 2015 ». *Proceedings of the National Academy of Sciences of the United States of America* 115.15, E3463-E3470.

KLÜMPER, Uli, Mario RECKER, Lihong ZHANG, Xiaole YIN, Tong ZHANG, Angus BUCKLING et William H. GAZE (août 2019). « Selection for antimicrobial resistance is reduced when embedded in a natural microbial community ». *The ISME Journal* 13.12, p. 2927-2937.

KNAPP, Charles W., Jan DOLFING, Phillip A.I. EHLERT et David W. GRAHAM (jan. 2010). « Evidence of increasing antibiotic resistance gene abundances in archived soils since 1940 ». *Environmental Science and Technology* 44.2, p. 580-587.

KNÖPPEL, Anna, Joakim NÄSVALL et Dan I. ANDERSSON (nov. 2017). « Evolution of antibiotic resistance without antibiotic exposure ». *Antimicrobial Agents and Chemotherapy* 61.11.

KOBAYASHI, Takeshi, Lisa NONAKA, Fumito MARUYAMA et Satoru SUZUKI (sept. 2007). « Molecular evidence for the ancient origin of the ribosomal protection protein that mediates tetracycline resistance in bacteria ». *Journal of Molecular Evolution* 65.3, p. 228-235.

KOONIN, Eugene V, Purificacion LOPEZ-GARCIA et Olga ZHAXYBAYEVA (juill. 2016). « Horizontal gene transfer : essentiality and evolvability in prokaryotes, and roles in evolutionary transitions ». *F1000Research* 5, p. 1805.

KOONIN, Eugene V et Yuri I WOLF (2009). « Is evolution Darwinian or/and Lamarckian ? » *Biology Direct* 4.42, p. 1-14.

KOTTARA, Anastasia, James P.J. HALL, Ellie HARRISON et Michael A. BROCKHURST (avr. 2016). « Multi-host environments select for host-generalist conjugative plasmids ». *BMC Evolutionary Biology* 16.1, p. 1-4.

KOTWANI, Anita, Jyoti JOSHI et Deeksha KALONI (juill. 2021). « Pharmaceutical effluent : a critical link in the interconnected ecosystem promoting antimicrobial resistance ». *Environmental Science and Pollution Research* 28.25, p. 32111-32124.

KRAKER, Marlieke E. A. de, Andrew J. STEWARDSON et Stephan HARBARTH (nov. 2016). « Will 10 Million People Die a Year due to Antimicrobial Resistance by 2050 ? » *PLOS Medicine* 13.11, e1002184.

KRAUSE, Kevin M, Alisa W SERIO, Timothy R KANE et Lynn E CONNOLLY (2016). « Aminoglycosides : An Overview ». *Cold Spring Harbor Perspectives in Medicine* 6.6, a027029.

KÜMMERER, Klaus (avr. 2009). « Antibiotics in the aquatic environment – A review – Part I ». *Chemosphere* 75.4, p. 417-434.

LAROUCHE, Élodie, Mylène GÉNÉREUX, Marie Ève TREMBLAY, Mohamed RHOUMA, Marc Olivier GASSER, Sylvain QUESSY et Caroline CÔTÉ (2020). « Impact of liquid hog manure applications on antibiotic resistance genes concentration in soil and drainage water in field crops ». *Canadian Journal of Microbiology* 66.10, p. 549-561.

LASSALLE, Florent, Séverine PÉRIAN, Thomas BATAILLON, Xavier NESME, Laurent DURET et Vincent DAUBIN (2015). « GC-Content Evolution in Bacterial Genomes : The Biased Gene Conversion Hypothesis Expands ». *PLOS Genetics* 11.2, e1004941.

LAWRENCE, Jeffrey G. et Howard OCHMAN (1997). « Amelioration of Bacterial Genomes : Rates of Change and Exchange ». *Journal of Molecular Evolution 1997 44 :4* 44.4, p. 383-397.

LE HELLO, Simon et al. (sept. 2011). « International Spread of an Epidemic Population of Salmonella enterica Serotype Kentucky ST198 Resistant to Ciprofloxacin ». *The Journal of Infectious Diseases* 204.5, p. 675-684.

LE PAGE, Gareth, Lina GUNNARSSON, Jason SNAPE et Charles R. TYLER (déc. 2017). « Integrating human and environmental health in antibiotic risk assessment : A critical analysis of protection goals, species sensitivity and antimicrobial resistance ». *Environment International* 109, p. 155-169.

LEDERBERG, J. et S. C. OAKS (1992). *Emerging infections : Microbial threats to health in the United States.* Washington D.C. : National Academy Press.

LEDERBERG, Joshua (août 1988). « Medical Science, Infectious Disease, and the Unity of Humankind ». *JAMA* 260.5, p. 684-685.

LEE, Hyo Ju, Seung Hak CHO, Dasom SHIN et Hui Seung KANG (oct. 2018). « Prevalence of Antibiotic Residues and Antibiotic Resistance in Isolates of Chicken Meat in Korea ». *Korean Journal for Food Science of Animal Resources* 38.5, p. 1055.

LEE, Seungjun, Michael SUITS, David WITUSZYNSKI, Ryan WINSTON, Jay MARTIN et Jiyoung LEE (juin 2020). « Residential urban stormwater runoff : A comprehensive profile of microbiome and antibiotic resistance ». *Science of The Total Environment* 723, p. 138033.

LEI, Kai et al. (sept. 2019). « Spatial and seasonal variations of antibiotics in river waters in the Haihe River Catchment in China and ecotoxicological risk assessment ». *Environment International* 130, p. 104919.

LEÓN-SAMPEDRO, Ricardo et al. (mars 2021). « Pervasive transmission of a carbapenem resistance plasmid in the gut microbiota of hospitalized patients ». *Nature Microbiology* 6.5, p. 606-616.

LESCH, John E. (2007). *The First Miracle Drugs : How the Sulfa Drugs Transformed Medicine*.

LEVIN-REISMAN, Irit, Irine RONIN, Orit GEFEN, Ilan BRANISS, Noam SHORESH et Nathalie Q. BALABAN (fév. 2017). « Antibiotic tolerance facilitates the evolution of resistance ». *Science* 355.6327, p. 826-830.

LEWIS, Kim (avr. 2013). « Platforms for antibiotic discovery ». *Nature Reviews Drug Discovery* 12.5, p. 371-387.

LI, Ruichao, Miaomiao XIE, Lizhang LIU, Yonglu HUANG, Xiaoyan WU, Zhiqiang WANG, Edward Wai Chi CHAN et Sheng CHEN (jan. 2020). « Characterisation of a cointegrate plasmid harbouring blaNDM-1 in a clinical Salmonella Lomita strain ». *International Journal of Antimicrobial Agents* 55.1, p. 105817.

LIM, Daniel et Natalie C.J. STRYNADKA (oct. 2002). « Structural basis for the $\beta$ lactam resistance of PBP2a from methicillin-resistant Staphylococcus aureus ». *Nature Structural Biology* 9.11, p. 870-876.

LINDGREN, Patricia Komp, Linda L. MARCUSSON, Dorthe SANDVANG, Niels FRIMODT-MØLLER et Diarmaid HUGHES (juin 2005). « Biological cost of single and multiple norfloxacin resistance mutations in Escherichia coli implicated in urinary tract infections ». *Antimicrobial Agents and Chemotherapy* 49.6, p. 2343-2351.

Lipsitch, Marc (2001). « Measuring and Interpreting Associations between Antibiotic Use and Penicillin Resistance in Streptococcus pneumoniae ». *Clinical Infectious Diseases* 32, p. 1044-54.

Livanos, Alexandra E., Erik J. Snider, Susan Whittier, David H. Chong, Timothy C. Wang, Julian A. Abrams et Daniel E. Freedberg (août 2018). « Rapid gastrointestinal loss of Clostridial Clusters IV and XIVa in the ICU associates with an expansion of gut pathogens ». *PLOS ONE* 13.8, e0200322.

Livermore, David M. (2008). « Defining an extended-spectrum $\beta$-lactamase ». *Clinical Microbiology and Infection* 14.SUPPL. 1, p. 3-10.

Loftie-Eaton, Wesley, Kelsie Bashford, Hannah Quinn, Kieran Dong, Jack Millstein, Samuel Hunter, Maureen K. Thomason, Houra Merrikh, Jose M. Ponciano et Eva M. Top (août 2017). « Compensatory mutations improve general permissiveness to antibiotic resistance plasmids ». *Nature Ecology & Evolution* 1.9, p. 1354-1363.

Long, Katherine S., Jacob Poehlsgaard, Corinna Kehrenberg, Stefan Schwarz et Birte Vester (juill. 2006). « The Cfr rRNA methyltransferase confers resistance to phenicols, lincosamides, oxazolidinones, pleuromutilins, and streptogramin A antibiotics ». *Antimicrobial Agents and Chemotherapy* 50.7, p. 2500-2505.

Lopatkin, Allison J., Sarah C. Bening, Abigail L. Manson, Jonathan M. Stokes, Michael A. Kohanski, Ahmed H. Badran, Ashlee M. Earl, Nicole J. Cheney, Jason H. Yang et James J. Collins (fév. 2021). « Clinically relevant mutations in core metabolic genes confer antibiotic resistance ». *Science* 371.6531.

López-Serna, Rebeca, Anna Jurado, Enric Vázquez-Suñé, Jesus Carrera, Mira Petrović et Damià Barceló (mars 2013). « Occurrence of 95 pharmaceuticals and transformation products in urban groundwaters underlying the metropolis of Barcelona, Spain ». *Environmental Pollution* 174, p. 305-315.

Lukačišinová, Marta, Booshini Fernando et Tobias Bollenbach (juin 2020). « Highly parallel lab evolution reveals that epistasis can curb the evolution of antibiotic resistance ». *Nature Communications* 11.1, p. 1-14.

Machado, Elisabete, Rafael Cantón, Fernando Baquero, Juan Carlos Galán, Azucena Rollán, Luísa Peixe et Teresa M. Coque (mai 2005). « Integron Content of Extended-Spectrum-$\beta$-Lactamase-Producing Escherichia coli Strains over 12 Years in a Single Hospital in Madrid, Spain ». *Antimicrobial Agents and Chemotherapy* 49.5, p. 1823.

Machida, Y., C. Machida, H. Ohtsubo et E. Ohtsubo (jan. 1982). « Factors determining frequency of plasmid cointegration mediated by insertion sequence IS1 ». *Proceedings of the National Academy of Sciences* 79.2, p. 277-281.

MacLean, R. Craig et Tom Vogwill (jan. 2015). « Limits to compensatory adaptation and the persistence of antibiotic resistance in pathogenic bacteria ». *Evolution, Medicine, and Public Health* 2015.1, p. 4-12.

Mader, Rodolphe et al. (jan. 2022). « Defining the scope of the European Antimicrobial Resistance Surveillance network in Veterinary medicine (EARS-Vet) : a bottom-up and One Health approach ». *Journal of Antimicrobial Chemotherapy*.

Madsen, Jonas Stenløkke, Mette Burmølle, Lars Hestbjerg Hansen et Søren Johannes Sørensen (juill. 2012). « The interconnection between biofilm formation and horizontal gene transfer ». *FEMS Immunology & Medical Microbiology* 65.2, p. 183-195.

Mahrt, Niels, Alexandra Tietze, Sven Künzel, Sören Franzenburg, Camilo Barbosa, Gunther Jansen et Hinrich Schulenburg (juill. 2021). « Bottleneck size and selection level reproducibly impact evolution of antibiotic resistance ». *Nature Ecology & Evolution* 5.9, p. 1233-1242.

Marijuán, Pedro C., Jorge Navarro et Raquel del Moral (fév. 2010). « On prokaryotic intelligence : Strategies for sensing the environment ». *Biosystems* 99.2, p. 94-103.

Marti, Elisabet, Eleni Variatza et Jose Luis Balcazar (jan. 2014). « The role of aquatic ecosystems as reservoirs of antibiotic resistance ». *Trends in Microbiology* 22.1, p. 36-41.

Martinez, Jose L (2012). « Natural Antibiotic Resistance and Contamination by Antibiotic Resistance Determinants : The Two Ages in the Evolution of Resistance to Antimicrobials ». *Frontiers in Microbiology* 0.1, p. 1.

Martinez, Jose Luis (nov. 2009). « Environmental pollution by antibiotics and by antibiotic resistance determinants ». *Environmental Pollution* 157.11, p. 2893-2902.

Martínez, José L. (juill. 2008). « Antibiotics and antibiotic resistance genes in natural environments ». *Science* 321.5887, p. 365-367.

Martínez, José L., Teresa M. Coque et Fernando Baquero (avr. 2015). « Prioritizing risks of antibiotic resistance genes in all metagenomes ». *Nature Reviews Microbiology* 13.6, p. 396-396.

Martínez, José Luis (avr. 2007). « Evolution of Bacterial Opportunistic Pathogens ». *Evolutionary Biology of Bacterial and Fungal Pathogens*. Sous la dir. de Fernando Baquero, César Nombela, Gail H. Cassell et José A. Gutiérrez-Fuentes. John Wiley & Sons, Ltd, p. 85-91.

Martinez-Urtaza, Jaime, Joaquin Trinanes, Narjol Gonzalez-Escalona et Craig Baker-Austin (fév. 2016). « Is El Niño a long-distance corridor for waterborne disease ? » *Nature Microbiology* 1.3, p. 1-3.

Mashak, Zohreh, Sedigheh Jafariaskari, Iman Alavi, Mohammadhossein Sakhaei, Shahreza et Farhad Safarpoor Dehkordi (2020). « Phenotypic and Genotypic Assessment of Antibiotic Resistance and Genotyping of vacA, cagA, iceA, oipA, cagE, and babA2 Alleles of Helicobacter pylori Bacteria Isolated from Raw Meat ». *Infection and Drug Resistance* 13, p. 257.

Massot, Méril, Pierre Châtre, Bénédicte Condamine, Véronique Métayer, Olivier Clermont, Jean Yves Madec, Erick Denamur et Marisa Haenni (nov. 2021). « Interplay between Bacterial Clones and Plasmids in the Spread of Antibiotic Resistance Genes in the Gut : Lessons from a Temporal Study in Veal Calves ». *Applied and Environmental Microbiology* 87.24.

Mateus-Vargas, Rafael H., Viktoria Atanassova, Felix Reich et Günter Klein (mai 2017). « Antimicrobial susceptibility and genetic characterization of Escherichia coli recovered from frozen game meat ». *Food Microbiology* 63, p. 164-169.

Mathers, Amy J., Heather L. Cox, Brandon Kitchel, Hugo Bonatti, Ann Karen C. Brassinga, Joanne Carroll, W. Michael Scheld, Kevin C. Hazen et Costi D. Sifri (2011). « Molecular dissection of an out-

break of carbapenem-resistant Enterobacteriaceae reveals intergenus KPC carbapenemase transmission through a promiscuous plasmid ». *mBio* 2.6.

MAZEL, Didier (août 2006). « Integrons : agents of bacterial evolution ». *Nature Reviews Microbiology* 4.8, p. 608-620.

MCCLINTOCK, B. (juin 1950). « The origin and behavior of mutable loci in maize ». *Proceedings of the National Academy of Sciences* 36.6, p. 344-355.

— (jan. 1951). « CHROMOSOME ORGANIZATION AND GENIC EXPRESSION ». *Cold Spring Harbor Symposia on Quantitative Biology* 16, p. 13-47.

MCCUTCHEON, A. R., S. M. ELLIS, R. E.W. HANCOCK et G. H.N. TOWERS (oct. 1992). « Antibiotic screening of medicinal plants of the British Columbian native peoples ». *Journal of Ethnopharmacology* 37.3, p. 213-223.

MCKENNA, M. (2007). *Big chicken : the incredible story of how antibiotics created modern agriculture and changed the way the world eats.* Washington D.C. : National Geographic Partners, LLC.

MEDEIROS, Antone A (1997). « Evolution and Dissemination of beta-Lactamases Accelerated by Generations of beta-Lactam Antibiotics ». *Clinical Infectious Diseases* 24.1, p. 19-45.

MELL, Joshua Chang et Rosemary J. REDFIELD (2014). « Natural competence and the evolution of DNA uptake specificity ». *Journal of Bacteriology* 196.8, p. 1471-1483.

MILLAR, Michael (sept. 2019). « Inequality and antibiotic resistance : A contractualist perspective ». *Bioethics* 33.7, p. 749-755.

MITCHELL, Marisa E.V., Robyn ALDERS, Fred UNGER, Hung NGUYEN-VIET, Trang Thi Huyen LE et Jenny Ann TORIBIO (fév. 2020). « The challenges of investigating antimicrobial resistance in Vietnam - What benefits does a One Health approach offer the animal and human health sectors ? » *BMC Public Health* 20.1, p. 1-12.

MUELLER, Elizabeth A., Alexander J.F. EGAN, Eefjan BREUKINK, Waldemar VOLLMER et Petra Anne LEVIN (2019). « Plasticity of Escherichia coli cell wall metabolism promotes fitness and antibiotic resistance across environmental conditions ». *eLife* 8.

MUELLER, Tanja et Per Olof ÖSTERGREN (août 2016). « The correlation between regulatory conditions and antibiotic consumption within the WHO European Region ». *Health Policy* 120.8, p. 882-889.

MUNKHOLM, Louise et Olivier RUBIN (déc. 2020). « The global governance of antimicrobial resistance : a cross-country study of alignment between the global action plan and national action plans ». *Globalization and Health* 16.1, p. 1-11.

MURRAY, Christopher JL et al. (jan. 2022). « Global burden of bacterial antimicrobial resistance in 2019 : a systematic analysis ». *The Lancet* 399.10325, p. 629-655.

MWANGI, Michael M. et al. (mai 2007). « Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing ». *Proceedings of the National Academy of Sciences* 104.22, p. 9451-9456.

NA, Guangshui, Xiaodan FANG, Yaqi CAI, Linke GE, Humin ZONG, Xiutang YUAN, Ziwei YAO et Zhifeng ZHANG (avr. 2013). « Occurrence, distribution, and bioaccumulation of antibiotics in coastal environment of Dalian, China ». *Marine Pollution Bulletin* 69.1-2, p. 233-237.

NAYFACH, Stephen et al. (nov. 2020). « A genomic catalog of Earth's microbiomes ». *Nature Biotechnology*, p. 1-11.

NISSEN, Jakob Nybo et al. (jan. 2021). « Improved metagenome binning and assembly using deep variational autoencoders ». *Nature Biotechnology* 39.5, p. 555-560.

NORMAN, Anders, Lars H. HANSEN et Soren J. SØRENSEN (août 2009). « Conjugative plasmids : vessels of the communal gene pool ». *Philosophical Transactions of the Royal Society B : Biological Sciences* 364.1527, p. 2275-2289.

NOVAK, Sebastian, Krishnendu CHATTERJEE et Martin A. NOWAK (oct. 2013). « Density games ». *Journal of Theoretical Biology* 334, p. 26-34.

O'BRIEN, Frances G., Karina Yui ETO, Riley J.T. MURPHY, Heather M. FAIRHURST, Geoffrey W. COOMBS, Warren B. GRUBB et Joshua P. RAMSAY (sept. 2015). « Origin-of-transfer sequences facilitate mobilisation of non-conjugative antimicrobial-resistance plasmids in Staphylococcus aureus ». *Nucleic Acids Research* 43.16, p. 7971-7983.

O'NEILL, Jim (2014). « Antimicrobial Resistance : Tackling a crisis for the health and wealth of nations ». *London : Review on Antimicrobial Resistance*.

OCHMAN, Howard, Jeffrey G. LAWRENCE et Eduardo A. GROLSMAN (mai 2000). « Lateral gene transfer and the nature of bacterial innovation ». *Nature* 405.6784, p. 299-304.

PARK, Chungoo et Jianzhi ZHANG (jan. 2012). « High Expression Hampers Horizontal Gene Transfer ». *Genome Biology and Evolution* 4.4, p. 523-532.

PARK, Jeongim, Myung-Hyun KIM, Kyungho CHOI, Young-Hee KIM et Min-Young KIM (2007). *Environmental Risk Assessment of Pharmaceuticals : Model Application for Estimating Pharmaceutical Exposures in the Han River Basin*. Rapp. tech. Korea Environment Institute.

PARKS, Donovan H., Christian RINKE, Maria CHUVOCHINA, Pierre Alain CHAUMEIL, Ben J. WOODCROFT, Paul N. EVANS, Philip HUGENHOLTZ et Gene W. TYSON (sept. 2017). « Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life ». *Nature Microbiology* 2.11, p. 1533-1542.

PATERSON, David L. et Robert A. BONOMO (oct. 2005). « Extended-spectrum $\beta$-lactamases : A clinical update ». *Clinical Microbiology Reviews* 18.4, p. 657-686.

PERRON, Gabriel G., Michael ZASLOFF et Graham BELL (jan. 2005). « Experimental evolution of resistance to an antimicrobial peptide ». *Proceedings of the Royal Society B : Biological Sciences* 273.1583, p. 251-256.

PETROVA, Mayya, Zhosephine GORLENKO et Sofia MINDLIN (juill. 2009). « Molecular structure and translocation of a multiple antibiotic resistance region of a Psychrobacter psychrophilus permafrost strain ». *FEMS Microbiology Letters* 296.2, p. 190-197.

Pietramellara, G., J. Ascher, F. Borgogni, M. T. Ceccherini, G. Guerri et P. Nannipieri (fév. 2009). « Extracellular DNA in soil and sediment : Fate and ecological relevance ». *Biology and Fertility of Soils* 45.3, p. 219-235.

Podolsky, Scott H. (2015). *The antibiotic era : reform, resistance, and the pursuit of a rational therapeutics*. Baltimore : Johns Hopkins University Press.

— (oct. 2018). « The evolving response to antibiotic resistance (1945–2018) ». *Palgrave Communications* 4.1, p. 1-8.

Popa, Ovidiu, Einat Hazkani-Covo, Giddy Landan, William Martin et Tal Dagan (avr. 2011). « Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes ». *Genome Research* 21.4, p. 599-609.

Poros-Gluchowska, Joanna, Malgorzata Kloszewska et Zdzlislaw Markiewicz (2003). « Amplicillin resistance in Listeria monocytogenes acquired as a result of transposon mutagenesis ». *Acta Microbiologica Polonica* 52.2, p. 131-142.

Puangseree, Jiratchaya, Saharuetai Jeamsripong, Rangsiya Prathan, Chanika Pungpian et Rungtip Chuanchuen (juin 2021). « Resistance to widely-used disinfectants and heavy metals and cross resistance to antibiotics in Escherichia coli isolated from pigs, pork and pig carcass ». *Food Control* 124, p. 107892.

Ramirez, Maria S. et Marcelo E. Tolmasky (déc. 2010). « Aminoglycoside modifying enzymes ». *Drug Resistance Updates* 13.6, p. 151-171.

Ramsay, Joshua P. et Neville Firth (août 2017). « Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements ». *Current Opinion in Microbiology* 38, p. 1-9.

Raymann, Kasie, Zack Shaffer et Nancy A. Moran (mars 2017). « Antibiotic exposure perturbs the gut microbiota and elevates mortality in honeybees ». *PLOS Biology* 15.3, e2001861.

Reding-Roman, Carlos, Mark Hewlett, Sarah Duxbury, Fabio Gori, Ivana Gudelj et Robert Beardmore (jan. 2017). « The unconstrained evolution of fast and efficient antibiotic-resistant bacterial genomes ». *Nature Ecology & Evolution* 1.3, p. 1-11.

Redondo-Salvo, Santiago, Raúl Fernández-López, Raúl Ruiz, Luis Vielva, María de Toro, Eduardo P.C. Rocha, M. Pilar Garcillán-Barcia et Fernando de la Cruz (juill. 2020). « Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids ». *Nature Communications* 11.1, p. 1-13.

Reis, Ana C., Boris A. Kolvenbach, Olga C. Nunes et Philippe F.X. Corvini (jan. 2020). « Biodegradation of antibiotics : The new resistance determinants – part I ». *New Biotechnology* 54, p. 34-51.

Roberts, A. W. et J. A. Visconti (oct. 1972). « The Rational and Irrational Use of Systemic Antimicrobial Drugs ». *American Journal of Hospital Pharmacy* 29.10, p. 828-834.

Rocha, Eduardo P C et David Bikard (jan. 2022). « Microbial defenses against mobile genetic elements and viruses : Who defends whom from what ? » *PLOS Biology* 20.1, e3001514.

Rocha, Eduardo P.C. (nov. 2004). « Codon usage bias from tRNA's point of view : Redundancy, specialization, and efficient decoding for translation optimization ». *Genome Research* 14.11, p. 2279-2286.

Rodriguez-Beltran, Jeronimo, J. Carlos R. Hernandez-Beltran, Javier Delafuente, Jose A. Escudero, Ayari Fuentes-Hernandez, R. Craig MacLean, Rafael Peña-Miller et Alvaro San Millan (avr. 2018). « Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation ». *Nature Ecology & Evolution* 2.5, p. 873-881.

Rodríguez-Beltrán, Jerónimo, Vidar Sørum, Macarena Toll-Riera, Carmen de la Vega, Rafael Peña-Miller et Álvaro San Millán (juill. 2020). « Genetic dominance governs the evolution and spread of mobile genetic elements in bacteria ». *Proceedings of the National Academy of Sciences of the United States of America* 117.27, p. 15755-15762.

Rosier, Florence (nov. 2021). « La pandémie de Covid-19 a réduit de 18% notre consommation d'antibiotiques.l ». *Le Monde*.

Ruckert, Arne, Patrick Fafard, Suzanne Hindmarch, Andrew Morris, Corinne Packer, David Patrick, Scott Weese, Kumanan Wilson, Alex Wong et Ronald Labonté (sept. 2020). « Governing antimicrobial resistance : a narrative review of global governance mechanisms ». *Journal of Public Health Policy* 41.4, p. 515-528.

Sabuncu, Elifsu, Julie David, Claire Bernède-Bauduin, Sophie Pépin, Michel Leroy, Pierre Yves Boëlle, Laurence Watier et Didier Guillemot (juin 2009). « Significant Reduction of Antibiotic Use in the Community after a Nationwide Campaign in France, 2002–2007 ». *PLOS Medicine* 6.6, e1000084.

Salipante, Stephen J et Barry G Hall (2003). « Determining the limits of the evolutionary potential of an antibiotic resistance gene ». *Molecular Biology and Evolution* 20.4, p. 653-659.

San Millan, A., R. Peña-Miller, M. Toll-Riera, Z. V. Halbert, A. R. McLean, B. S. Cooper et R. C. Maclean (oct. 2014). « Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids ». *Nature Communications* 5.1, p. 1-11.

San Millan, Alvaro (déc. 2018). « Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context ». *Trends in Microbiology* 26.12, p. 978-985.

San Millan, Alvaro et R Craig Maclean (2017). « Fitness Costs of Plasmids : a Limit to Plasmid Transmission ». *Microbiology Spectrum* 5.5, MTBP-0016-2017.

Sandegren, Linus et Dan I. Andersson (2009). « Bacterial gene amplification : implications for the evolution of antibiotic resistance ». *Nature Reviews Microbiology* 7.8, p. 578-588.

Sant, Duhita G., Laura C. Woods, Jeremy J. Barr et Michael J. McDonald (jan. 2021). « Host diversity slows bacteriophage adaptation by selecting generalists over specialists ». *Nature Ecology & Evolution* 5.3, p. 350-359.

Santiago-Rodriguez, Tasha M., Gino Fornaciari, Stefania Luciani, Scot E. Dowd, Gary A. Toranzos, Isolina Marota, Raul J. Cano et Brenda A. Wilson (sept. 2015). « Gut Microbiome of an 11th Century A.D. Pre-Columbian Andean Mummy ». *PLOS ONE* 10.9, e0138135.

Schaack, Sarah, Clément Gilbert et Cédric Feschotte (sept. 2010). « Promiscuous DNA : horizontal transfer of transposable elements and why it matters for eukaryotic evolution ». *Trends in Ecology & Evolution* 25.9, p. 537-546.

Schrijver, R., M. Stijntjes, J. Rodríguez-Baño, E. Tacconelli, N. Babu Rajendran et A. Voss (juin 2018). « Review of antimicrobial resistance surveillance programmes in livestock and meat in EU with focus on humans ». *Clinical Microbiology and Infection* 24.6, p. 577-590.

Serio, Alisa W., Tiffany Keepers, Logan Andrews et Kevin M. Krause (fév. 2018). « Aminoglycoside Revival : Review of a Historically Important Class of Antimicrobials Undergoing Rejuvenation ». *EcoSal Plus* 8.1.

Shallcross, Laura J., Simon J. Howard, Tom Fowler et Sally C. Davies (2015). « Tackling the threat of antimicrobial resistance : from policy to sustainable action ». *Philosophical Transactions of the Royal Society B : Biological Sciences* 370.1670.

Shapiro, Julie Teresa et al. (oct. 2020). « Metapopulation ecology links antibiotic resistance, consumption, and patient transfers in a network of hospital wards ». *eLife* 9, p. 1-42.

Sharp, Paul M., Elizabeth Bailes, Russell J. Grocock, John F. Peden et R. Elizabeth Sockett (mars 2005). « Variation in the strength of selected codon usage bias among bacteria ». *Nucleic Acids Research* 33.4, p. 1141-1153.

Siefert, Janet L. (2009). « Defining the Mobilome ». *Methods in molecular biology (Clifton, N.J.)* 532, p. 13-27.

Simmons, Henry E. et Paul D. Stolley (mars 1974). « This is Medical Progress ?: Trends and Consequences of Antibiotic Use in the United States ». *JAMA* 227.9, p. 1023-1028.

Skov, Robert L. et D. L. Monnet (mars 2016). « Plasmid-mediated colistin resistance (mcr-1 gene) : Three months later, the story unfolds ». *Eurosurveillance* 21.9, p. 30155.

Smillie, Chris S, Mark B Smith, Jonathan Friedman, Otto X Cordero, Lawrence A David et Eric J Alm (2011). « Ecology drives a global network of gene exchange connecting the human microbiome ».

Sniegowski, Paul D, Philip J Gerrish, Toby Johnson et Aaron Shaver (2000). « The evolution of mutation rates : separating causes from consequences ». *BioEssays* 22.12, p. 1057-1066.

Solar, Gloria del, Rafael Giraldo, María Jesús Ruiz-Echevarría, Manuel Espinosa et Ramón Díaz-Orejas (juin 1998). « Replication and Control of Circular Bacterial Plasmids ». *Microbiology and Molecular Biology Reviews* 62.2, p. 434-464.

Sorek, Rotem, Yiwen Zhu, Christopher J. Creevey, M. Pilar Francino, Peer Bork et Edward M. Rubin (nov. 2007). « Genome-wide experimental determination of barriers to horizontal gene transfer ». *Science* 318.5855, p. 1449-1452.

Sørensen, Thomas Lund et Dominique Monnet (juin 2000). « Control of Antibiotic Use in the Community : The Danish Experience ». *Infection Control & Hospital Epidemiology* 21.6, p. 387-389.

Soupir, Michelle L, Saied Mostaghimi, Theo Dillaha et Virginia Tech (mai 2010). « Attachment of Escherichia coli and Enterococci to Particles in Runoff ». *Journal of Environmental Quality* 39.3, p. 1019-1027.

Stevenson, Cagla, James PJ Hall, Ellie Harrison, AJamie Wood et Michael A Brockhurst (2017). « Gene mobility promotes the spread of resistance in bacterial populations ». *The ISME journal* 11.8, p. 1930-1932.

Stoletzki, Nina et Adam Eyre-Walker (fév. 2007). « Synonymous Codon Usage in Escherichia coli : Selection for Translational Accuracy ». *Molecular Biology and Evolution* 24.2, p. 374-381.

Su, Hao Chang, Guang Guo Ying, Liang Ying He, You Sheng Liu, Rui Quan Zhang et Ran Tao (jan. 2014). « Antibiotic resistance, plasmid-mediated quinolone resistance (PMQR) genes and ampC gene in two typical municipal wastewater treatment plants ». *Environmental Science : Processes & Impacts* 16.2, p. 324-332.

Taddei, F., M. Radman, J. Maynard-Smith, B. Toupance, P. H. Gouyon et B. Godelle (1997). « Role of mutator alleles in adaptive evolution ». *Nature* 387.6634, p. 700-702.

Tamber, Sandeep et Robert E W Hancock (2003). « ON THE MECHANISM OF SOLUTE UPTAKE IN PSEUDOMONAS ». *Frontiers in Bioscience* 8, p. 472.

Tello, Alfredo, Brian Austin et Trevor C. Telfer (août 2012). « Selective pressure of antibiotic pollution on bacteria of importance to public health ». *Environmental Health Perspectives* 120.8, p. 1100-1106.

Teng, Ching Hao, Pin Chieh Wu, Sen Lin Tang, Yi Chen Chen, Ming Fang Cheng, Ping Chih Huang, Wen Chien Ko et Jiun Ling Wang (mars 2021). « A Large Spatial Survey of Colistin-Resistant Gene mcr-1-Carrying E. coli in Rivers across Taiwan ». *Microorganisms* 9.4, p. 722.

Thoms, Ulrike (sept. 2012). « Between Promise and Threat : Antibiotics in Foods in West Germany 1950-1980 ». *NTM International Journal of History and Ethics of Natural Sciences, Technology and Medicine* 20.3, p. 181-214.

Tofteland, Ståle, Umaer Naseer, Jan Helge Lislevand, Arnfinn Sundsfjord et Ørjan Samuelsen (mars 2013). « A Long-Term Low-Frequency Hospital Outbreak of KPC-Producing Klebsiella pneumoniae Involving Intergenus Plasmid Diffusion and a Persisting Environmental Reservoir ». *PLOS ONE* 8.3, e59015.

Tong, Lei, Liting Qin, Chuan Guan, Mesmire Emade Wilson, Xianju Li, Dandan Cheng, Jie Ma, Hui Liu et Fujun Gong (sept. 2020). « Antibiotic resistance gene profiling in response to antibiotic usage and environmental factors in the surface water and groundwater of Honghu Lake, China ». *Environmental Science and Pollution Research* 27.25, p. 31995-32005.

Valcárcel, Y., S. González Alonso, J. L. Rodríguez-Gil, A. Gil et M. Catalá (sept. 2011). « Detection of pharmaceutically active compounds in the rivers and tap water of the Madrid Region (Spain) and potential ecotoxicological risk ». *Chemosphere* 84.10, p. 1336-1348.

Valderrama, J. Andrés, Surashree S. Kulkarni, Victor Nizet et Ethan Bier (déc. 2019). « A bacterial gene-drive system efficiently edits and inactivates a high copy number antibiotic resistance locus ». *Nature Communications* 10.1, p. 1-8.

Vittecoq, Marion et al. (avr. 2016). « Antimicrobial resistance in wildlife ». *Journal of Applied Ecology* 53.2, p. 519-529.

Walsh, Ian M., Micayla A. Bowman, Iker F. Soto Santarriaga, Anabel Rodriguez et Patricia L. Clark (fév. 2020). « Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness ». *Proceedings of the National Academy of Sciences of the United States of America* 117.7, p. 3528-3534.

WANG, G., T. J.M. WILSON, Q. JIANG et D. E. TAYLOR (2001). « Spontaneous mutations that confer antibiotic resistance in Helicobacter pylori ». *Antimicrobial Agents and Chemotherapy* 45.3, p. 727-733.

WANG, Zhiyuan, Qiuwen CHEN, Jianyun ZHANG, Tiesheng GUAN, Yuchen CHEN et Wenqing SHI (nov. 2020). « Critical roles of cyanobacteria as reservoir and source for antibiotic resistance genes ». *Environment International* 144, p. 106034.

WATANABE, Tsutomu (1963). « INFECTIVE HEREDITY OF MULTIPLE DRUG RESISTANCE IN BACTERIA ». *Bacteriology Reviews* 27.1, p. 87-115.

WEIN, Tanita, Nils F. HÜLTER, Itzhak MIZRAHI et Tal DAGAN (juin 2019). « Emergence of plasmid stability under non-selective conditions maintains antibiotic resistance ». *Nature Communications* 10.1, p. 1-13.

WEINGARTEN, Rebecca A. et al. (jan. 2018). « Genomic analysis of hospital plumbing reveals diverse reservoir of bacterial plasmids conferring carbapenem resistance ». *mBio* 9.1.

WEINREICH, Daniel M. et Lin CHAO (juin 2005). « RAPID EVOLUTIONARY ESCAPE BY LARGE POPULATIONS FROM LOCAL FITNESS PEAKS IS LIKELY IN NATURE ». *Evolution* 59.6, p. 1175-1182.

WHEATLEY, Rachel M. et R. Craig MACLEAN (déc. 2020). « CRISPR-Cas systems restrict horizontal gene transfer in Pseudomonas aeruginosa ». *The ISME Journal* 15.5, p. 1420-1433.

WHO (2015). « WHO Library Cataloguing-in-Publication Data Global Action Plan on Antimicrobial Resistance ». *Microbe Magazine* 10.9, p. 354-355.

WHO REGIONAL OFFICE FOR EUROPE (2021). *Surveillance of antimicrobial resistance in Europe, 2020 data. Executive Summary.* Rapp. tech. Copenhagen.

WIEHLMANN, Lutz, Nina CRAMER et Burkhard TÜMMLER (déc. 2015). « Habitat-associated skew of clone abundance in the Pseudomonas aeruginosa population ». *Environmental Microbiology Reports* 7.6, p. 955-960.

WILLMS, Inka Marie, Aysha KAMRAN, Nils Frederik ASSMANN, Denis KRONE, Simon Henning BOLZ, Fabian FIEDLER et Heiko NACKE (2019). « Discovery of novel antibiotic resistance determinants in forest and grassland soil metagenomes ». *Frontiers in Microbiology* 10.3, p. 460.

WITZANY, Christopher, Sebastian BONHOEFFER et Jens ROLFF (oct. 2020). « Is antimicrobial resistance evolution accelerating ? » *PLOS Pathogens* 16.10, e1008905.

WORLD BANK (mars 2017). *Drug-Resistant Infections : A Threat to Our Economic Future*. Rapp. tech. World Bank.

WORLD HEALTH ORGANIZATION (1994). *WHO scientific working group on monitoring and management of bacterial resistance to antimicrobial agents*. Rapp. tech. New York : World Health Organization.

— (1997). *The medical impact of the use of antimicrobials in food animals : Report of a WHO Meeting.* Rapp. tech. Geneva : World Health Organization.

— (1998). *The current status of antimicrobial resistance in Europe : report of a WHO workshop held in collaboration with the Italian Associazione Culturale Microbiologia Medica*. Rapp. tech. Geneva : World Health Organization.

World Health Organization (2001). *WHO global strategy for containment of antimicrobial resistance*. Rapp. tech. Geneva : World Health Organization.

— (2014). *Antimicrobial resistance global report on surveillance*. Rapp. tech. Geneva : World Health Organization.

Wright, Gerard D. (oct. 2010). « Antibiotic resistance in the environment : a link to the clinic ? » *Current Opinion in Microbiology* 13.5, p. 589-594.

Wright, Sewall (mars 1931). « Evolution in Mendelian Populations ». *Genetics* 16.2, p. 97.

Yang, Fengxia, Bingjun Han, Yanru Gu et Keqiang Zhang (sept. 2020). « Swine liquid manure : a hotspot of mobile genetic elements and antibiotic resistance genes ». *Scientific Reports* 10.1, p. 1-10.

Yu, Pingfeng, Avery Zaleski, Qilin Li, Ya He, Kris Mapili, Amy Pruden, Pedro J.J. Alvarez et Lauren B. Stadler (août 2018). « Elevated Levels of Pathogenic Indicator Bacteria and Antibiotic Resistance Genes after Hurricane Harvey's Flooding in Houston ». *Environmental Science and Technology Letters* 5.8, p. 481-486.

Zhang, Hui, Yuji Sekiguchi, Satoshi Hanada, Philip Hugenholtz, Hongik Kim, Yoichi Kamagata et Kazunori Nakamura (juill. 2003). « Gemmatimonas aurantiaca gen. nov., sp. nov., a Gram-negative, aerobic, polyphosphate-accumulating micro-organism, the first cultured representative of the new bacterial phylum Gemmatimonadetes phyl. nov ». *International Journal of Systematic and Evolutionary Microbiology* 53.4, p. 1155-1163.

Zhang, Rong Min, Xiao Liu, Shao Lin Wang, Liang Xing Fang, Jian Sun, Ya Hong Liu et Xiao Ping Liao (mars 2021). « Distribution patterns of antibiotic resistance genes and their bacterial hosts in pig farm wastewater treatment systems and soil fertilized with pig manure ». *Science of The Total Environment* 758, p. 143654.

Zhang, Yan, Nadja Heidrich, Biju Joseph Ampattu, Carl W. Gunderson, H. Steven Seifert, Christoph Schoen, Jörg Vogel et Erik J. Sontheimer (mai 2013). « Processing-Independent CRISPR RNAs Limit Natural Transformation in Neisseria meningitidis ». *Molecular Cell* 50.4, p. 488-503.

Zhang, Yu, Yunhao Zheng, Zhiping Zhu, Yongxing Chen et Hongmin Dong (mars 2021a). « Dispersion of Antibiotic Resistance Genes (ARGs) from stored swine manure biogas digestate to the atmosphere ». *Science of The Total Environment* 761, p. 144108.

— (mars 2021b). « Dispersion of Antibiotic Resistance Genes (ARGs) from stored swine manure biogas digestate to the atmosphere ». *Science of The Total Environment* 761, p. 144108.

Zheng, Jinshui, Ziyu Guan, Shiyun Cao, Donghai Peng, Lifang Ruan, Daohong Jiang et Ming Sun (jan. 2015). « Plasmids are vectors for redundant chromosomal genes in the group ». *BMC Genomics* 16.1, p. 1-10.

# Section VIII : Liste des abréviations

**ACB** : *AMEG-carrying bacteria*, bactérie dont le génome comporte des gènes de résistance aux aminoglycosides

**AG** : aminoglycoside

**AME** : *aminoglycoside modifying enzymes*, enzymes d'altération des aminoglycosides

**AMEG** : *AME-encoding genes*, gènes encodant des enzymes d'altération des aminoglycosides

**AR** : *antibiotic resistance*, résistance antibiotique

- **ARG** : gène de résistance aux antibiotiques
- **ARB** : bactérie résistante aux antibiotiques

**CHG** : *cluster of homologous genes*, groupe de gènes homologues, partageant une similarité de séquence d'au moins 40% sur 80% de leur longueur

**CRISPR-Cas** : *Clustered Regularly Interspaced Short Palindromic Repeats* et *CRISPR-associated proteins* ; courtes répétitions en palindrome regroupées et régulièrement espacées, et protéines associées ; système immunitaire adaptatif bactérien

**CUB** : *codon usage bias*, biais d'usage de codons

**HGT** : *horizontal gene transfer*, transfert horizontal de gènes

**ICE** : *Integrative and conjugative elements* (ICEs, éléments intégratifs et conjugatifs (parfois appelés transposons conjugatifs)

**MGE** : *mobile genetic elements*, éléments génétiques mobiles

**MST** : *minimum spanning tree*, arbre couvrant minimal ; réseau comportant $N-1$ arêtes pour $N$ noeuds

**MSN** : *minimum spanning network*, réseau couvrant minimal ; réseau ne pouvant être réduit à un MST du fait de la présence de certains chemins de même poids

**NHEJ** : *non-homologous end joining*, mécanisme de réparation de l'ADN par jonction d'extrémités non-homologues

**OriT** : origine de transfert des plasmides

**OriR** : origine de réplication des plasmides

# Section IX : Annexes

## 9 Annexe 1 : Informations supplémentaires de la section IV

Les informations supplémentaires de la section IV sont pour partie exposées dans la section suivante : Textes S1–S3, Tables S3–S37, Table S40, Figures S1–S2.

Le reste du matériel de ce chapitre (Tables S1–S2, Table S38, Liste S1) n'ont pas pu être affichés dans cette thèse. Ils sont cependant disponibles en ligne sur le serveur *bioRXiV* à l'adresse suivante : https://www.biorxiv.org/content/10.1101/2022.01.07.475340v1.supplementary-material.

**Text S1: Description of interaction effects on the prevalence of ACBs in European data between 1997 and 2018**

Aminoglycoside consumption can have a positive significant effect on the probability to sample ACBs in clinical samples (for CHG 6.1), soil (for CHG 6.1), farms samples (for CHGs 6.1 and 8), and domestic animals (for CHG 8), but also a negative significant effect in clinical samples (for CHG 17.1). Human exchanges also have ecology-specific impacts for 3 CHGs only. For CHG 6.1, migration has a positive effect on the probability to sample ACBs in clinical, farms, human, and soil samples but a negative effect in freshwater samples, whereas trade has opposite effects in all these biomes. For CHG 8, migration has a negative effect in soil samples, and trade has a negative effect in farms samples. For CHG 27, migration has a negative effect in clinical, farms, and human samples, while trade has a positive effect in farms.

**Text S2: Keywords used to filter sequences based on InterProScan annotations**

Amino acid sequences that were screened in genomes with HMMER AME profiles, were submitted to InterProScan in order to infer their putative functions. These sequences were then filtered based on keywords among their inferred functions. On the one hand, sequences that may have an enzymatic function similar to the ones of AMEs were screened with the following keywords: “acetyltransferase”, “adenylyltransferase”, “adenyltransferase”, “phosphotransferase”, “phosphoryltransferase”, and “nucleotidyltransferase”. On the other hand, sequences whose functions might be involved in aminoglycoside resistance were screened with the following keywords: “aminoglycoside”, “aminoside”, “mycin”, “micin”, and “amikacin”.

**Text S3: Keywords used to classify BioSamples into ecological biomes**

Among the BioSamples metadata columns, a number of columns were isolated to search for information enabling to characterize a sampling biome for each genome. On the one hand, were selected the column whose names contained one of the following keywords: “biome”, “habitat”, “env”, “host”. On the other hand, other columns were selected if they contained the keyword “source” or “isolat”, but none of the following keywords: “date”, “year”, “by”, “id”, “preservation”, “number”, “energy”, “carbon”, “method”, “comment”, “note”, “resource”, “history”, “annot”, or “lab”. Samples were then classified into biomes based on the contents of these columns.

Samples were classified as from clinical origin:

- if one of these columns (or columns regarding the sampling location) contained the keyword “hospit”, “clinic”, or “medical”;
- if one of these columns contained the keyword “human” or “homo” and at least one of the following keywords: “disease”, “blood”, “oral”, “feces”, “infection”, “arthr”, “hemo”, “bronch”, “pulmo”, “respi”, “urine”, “failure”, “pus”, “throat”, “thora”, “semen”, “wound”, “skin”, “gut”, “intestin”, “septicemia”, “sputum”, “fibrosis”, “swab”, “fluid”, “itis”, “bile”, “pneumonia”, “sputamentum”, “sick”, “gastr”, “aspirate”, “fecal”, “sudate”, “groin”, “emia”, “faecal”, “nares”, “osis”, “excreted”, “sepsis”, “patient”, “vagin”, “rect”, “surg”, “drainage”, “trach”, “lung”, “nasal”, “tissue”, “cornea”, “nosocomial”, “ICU”, “fever”, “bronc”, “absc”, “phary”;
- if the host was not specified but one of these columns contained any of the keywords: “melioidosis”, “tract infection”, “burn”, “urine”, “feces”, “blood”, “stool”, “sputum”

Samples were classified as originating from human habitats if they were not already assigned to clinical samples, and if their columns contained any of the following: “habitat”, “rural”, “bath”, “potable”, “tap water”, “toilet”, “spacecraft”, “food”.

Samples were classified as originating from domestic animals if they were not already assigned to clinical or human habitat samples, and if their columns contained any of the following: “canis”, “canine”, “feline”, “dog” (but not “hot-dog”) or “prairie dog”), “felis”, “rattus”, “cavia”, “cricet”, “chinchilla”, “mustela”, “mus”, “serinus”, “tortoise”, “parrot”, “carassius”, “mouse”.

Samples were classified as originating from farms if they were not already assigned to any of the previous origins, and if their columns contained any of the following: “bos”, “livestock”, “ovis”, “ovine”, “sheep”, “capra”, “meat”, “anas”, “goose”, “duck”, “cow”, “ovine”, “fish farm”, “slaughterhouse”, “poultry”, “gallus”, “pig”, “pork”, “beef”, “meleagris”, “oyster” (if also containing “cult”), “mellifera”, “milk”, “hen”, “chicken”, “lamb”, “egg”, “cattle”, “turkey”, “swine”, “porcine”, “goat”, “equine”, “horse”, “rabbit”, “calf”, “sus scrofa”, “equus”, “oryctolagus”, “manure”, “dairy”, “chease”, “sus” (if also containing “disease”), “farm”.

Samples were classified as originating from agrosystems if they were not already assigned to any of the previous origins, and if their columns contained any of the following: “wheat”, “salad”, “triticum”, “maize”, “zea mays”, “rice”, “oryza”, “field”, “plantation”, “olive”, “coffee”, “coffea”, “bean”, “banana”, “tomato”, “potato”, “solanum”, “musa”, “avena”, “oats”, “soy”, “brassica”, “hordeum”, “barley”, “rye”, “sorghum”, “millet”, “phaseolus”, “saccharum, “quinoa”, “cicer”, “pisum”, “sesam”, “cassava”, “sugar”, “manihot”, “ipomo”, “coconut”, “cotton”, “oil palm”, “helianthus”, “berry”, “onion”, “grape”, “walnut”, “peanut”, “prunus”, “citrus”, “greenhouse”, “lentil”, “tea”, “camellia”, “compost”, “tobacco”, “nicotiana”, “lettuce”, “vegetable”, “carrot”, “cucumber”, “pepper”, “pea”, “cilantro”.

Samples were classified as originating from wild plants and animals if they were not already assigned to any of the previous origins, and if their columns contained any of the following: “tapir”, “panthera”, “forest”, “grass”, “prairie dog”, “snake”, “ursus”, “ursi”, “migratory”, “crow”, “pika”, “gull”, “bird”, “clover”, “grazing”, “monkey”, “elephant”, “corvus”, “zebra”, “lily”, “vulture”.

Samples were classified as originating from freshwater if they were not already assigned to any of the previous origins, and if their columns contained any of the following: “river”, “marsh”, “pond”, “lake”, “swamp”, “bog”, “glacier”, “freshwater”, “fresh water”, “aquifer”, “permafrost”, “ground water”, “wetland”, “spring”, “catfish”, “creek”, “mineral water”, “environmental (streams)”.

Samples were classified as originating from sea water if they were not already assigned to any of the previous origins, and if their columns (or columns regarding the sampling location) contained any of the following: “ocean”, “seawater”, “sea”, “sea water”, “hydrotherm”, “continental shelf”, “saltwater”, “salt water”, “algae”, “tidal”, “tuna”, “seafood”, “shrimp”. These samples were later discarded because their sampling location could not be assigned to countries.

Samples were classified as originating from waste and sludge if they were not already assigned to any of the previous origins, and if their columns contained any of the following: “wastewater”, “waste water”, “sewage”, “sludge”.

Samples were classified as originating from soil if they were not already assigned to any of the previous origins, and if their columns contained any of the following: “soil”, “terrestrial”, “sand”, “humus”.

Finally, among the remaining samples that were assigned to none of the previous origins, samples whose columns contained the keywords “homo sapiens” or “human” were considered as originating from humans.

**Table S3: Summary of the selected model for CHG 1.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = -0.205, ν = 0.196, ρ = 0.36

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -6.11 | 1.54 | -3.98 | $6.97 \times 10^{-5}$ *** |
| Domestic animals | -34.2 | $5.96 \times 10^{6}$ | $-5.73 \times 10^{-6}$ | 1 |
| Farms | -5.34 | 0.576 | -9.26 | 0 *** |
| Flora, fauna | -33.3 | $8.04 \times 10^{6}$ | $-4.14 \times 10^{-6}$ | 1 |
| Human samples | -32.8 | $2.06 \times 10^{6}$ | $-1.59 \times 10^{-5}$ | 1 |
| Human habitat | -29.7 | $1.51 \times 10^{7}$ | $-1.97 \times 10^{-6}$ | 1 |
| Freshwater | -1.93 | 1.27 | -1.52 | 0.129 |
| Clinical samples | -4.99 | 0.587 | -8.51 | 0 *** |
| Sludge, waste | -35.7 | $4.48 \times 10^{6}$ | $-7.96 \times 10^{-6}$ | 1 |
| Soil | -2.82 | 1.24 | -2.28 | $2.28 \times 10^{-2}$ * |
| Trade | $5.83 \times 10^{-2}$ | 0.308 | 0.189 | 0.85 |
| Migration | 0.829 | 0.589 | 1.41 | 0.159 |
| Aminoglycosides | 1.63 | 0.589 | 2.76 | $5.71 \times 10^{-3}$ ** |

**Table S4: Summary of the selected model for CHG 2.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.597, ν = 0.192, ρ = 10.3

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -9.34 | 1.12 | -8.33 | $1.11 \times 10^{-16}$ *** |
| Domestic animals | -28.4 | $6.23 \times 10^{6}$ | $-4.56 \times 10^{-6}$ | 1 |
| Farms | -28.5 | $1.61 \times 10^{6}$ | $-1.77 \times 10^{-5}$ | 1 |
| Flora, fauna | -29.6 | $8.95 \times 10^{6}$ | $-3.3 \times 10^{-6}$ | 1 |
| Human samples | -28.1 | $2.08 \times 10^{6}$ | $-1.35 \times 10^{-5}$ | 1 |
| Human habitat | -27.7 | $1.33 \times 10^{7}$ | $-2.08 \times 10^{-6}$ | 1 |
| Freshwater | -30 | $8.93 \times 10^{6}$ | $-3.36 \times 10^{-6}$ | 1 |
| Clinical samples | 1.93 | 0.756 | 2.55 | $1.08 \times 10^{-2}$ * |
| Sludge, waste | -27.9 | $4.4 \times 10^{6}$ | $-6.35 \times 10^{-6}$ | 1 |
| Soil | -28.9 | $6.17 \times 10^{6}$ | $-4.69 \times 10^{-6}$ | 1 |

**Table S5: Summary of the selected model for CHG 3.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 7.59 \times 10^{-8}$, $\nu = 16.7$, $\rho = 8.47$

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| *Intercept* | -13.7 | 1.52 | -9.02 | 0 *** |
| *Domestic animals* | -26.9 | $7.9 \times 10^{6}$ | $-3.4 \times 10^{-6}$ | 1 |
| *Farms* | -25.5 | $1.19 \times 10^{5}$ | $-2.14 \times 10^{-4}$ | 1 |
| *Flora, fauna* | -25.3 | $1.37 \times 10^{6}$ | $-1.86 \times 10^{-5}$ | 1 |
| *Human samples* | 3.52 | 1.19 | 2.96 | $3.05 \times 10^{-3}$ ** |
| *Human habitat* | -22.6 | $1.15 \times 10^{7}$ | $-1.97 \times 10^{-6}$ | 1 |
| *Freshwater* | 4.46 | 1.79 | 2.5 | $1.26 \times 10^{-2}$ * |
| *Clinical samples* | 6.45 | 1.07 | 6.05 | $1.41 \times 10^{-9}$ *** |
| *Sludge, waste* | -28 | $4.51 \times 10^{6}$ | $-6.2 \times 10^{-6}$ | 1 |
| *Soil* | 4.41 | 1.48 | 2.98 | $2.85 \times 10^{-3}$ ** |
| *Trade* | -0.952 | 0.307 | -3.1 | $1.91 \times 10^{-3}$ ** |
| *Migration* | 1.73 | 0.582 | 2.97 | $2.95 \times 10^{-3}$ ** |
| *Aminoglycosides* | 1.48 | 0.124 | 12 | 0 *** |

**Table S6: Summary of the selected model for CHG 4.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 3.56 \times 10^{-4}$, $\nu = 6.04$, $\rho = 1.33$

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -36 | $3.42 \times 10^{6}$ | $-1.05 \times 10^{-5}$ | 1 |
| Domestic animals | -2.92 | $6.77 \times 10^{6}$ | $-4.32 \times 10^{-7}$ | 1 |
| Farms | 29.7 | $3.42 \times 10^{6}$ | $8.68 \times 10^{-6}$ | 1 |
| Flora, fauna | 31.6 | $3.42 \times 10^{6}$ | $9.24 \times 10^{-6}$ | 1 |
| Human samples | 29.1 | $3.42 \times 10^{6}$ | $8.51 \times 10^{-6}$ | 1 |
| Human habitat | -2.03 | $1.12 \times 10^{7}$ | $-1.81 \times 10^{-7}$ | 1 |
| Freshwater | -3.41 | $9.15 \times 10^{6}$ | $-3.73 \times 10^{-7}$ | 1 |
| Clinical samples | 29.8 | $3.42 \times 10^{6}$ | $8.71 \times 10^{-6}$ | 1 |
| Sludge, waste | -2 | $5.54 \times 10^{6}$ | $-3.61 \times 10^{-7}$ | 1 |
| Soil | -2.63 | $6.6 \times 10^{6}$ | $-3.99 \times 10^{-7}$ | 1 |
| Trade | 1.33 | 0.225 | 5.9 | $3.72 \times 10^{-9}$ *** |
| Migration | -1.38 | 0.339 | -4.06 | $4.81 \times 10^{-5}$ *** |
| Aminoglycosides | 0.568 | 0.214 | 2.65 | $8.03 \times 10^{-3}$ ** |

**Table S7: Summary of the selected model for CHG 5.1.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.557, ν = 16.7, ρ = 3.6

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -6.3 | 1.72 | -3.66 | $2.47 \times 10^{-4}$ *** |
| Domestic animals | -65.2 | $1.66 \times 10^{6}$ | $-3.92 \times 10^{-5}$ | 1 |
| Farms | -2.36 | 1.95 | -1.21 | 0.225 |
| Flora, fauna | -3.09 | 6.53 | -0.472 | 0.637 |
| Human samples | 2.46 | 1.67 | 1.48 | 0.139 |
| Human habitat | $-7.11 \times 10^{2}$ | $1.29 \times 10^{6}$ | $-5.5 \times 10^{-4}$ | 1 |
| Freshwater | -30.8 | $8.85 \times 10^{6}$ | $-3.48 \times 10^{-6}$ | 1 |
| Clinical samples | 3.74 | 1.65 | 2.27 | $2.34 \times 10^{-2}$ * |
| Sludge, waste | 1.77 | 2.31 | 0.766 | 0.444 |
| Soil | -30.3 | $7.39 \times 10^{6}$ | $-4.1 \times 10^{-6}$ | 1 |
| Trade | 0.802 | 1.69 | 0.474 | 0.635 |
| Migration | -2.15 | 2.76 | -0.781 | 0.435 |
| Domestic animals × Trade | 14.9 | $5.95 \times 10^{5}$ | $2.51 \times 10^{-5}$ | 1 |
| Farms × Trade | 3 | 1.88 | 1.6 | 0.11 |
| Flora, fauna × Trade | -16.4 | 10.8 | -1.51 | 0.13 |
| Human samples × Trade | -0.366 | 1.7 | -0.215 | 0.83 |
| Human habitat × Trade | $8.32 \times 10^{2}$ | $1.5 \times 10^{6}$ | $5.54 \times 10^{-4}$ | 1 |
| Freshwater × Trade | -0.792 | $1.19 \times 10^{7}$ | $-6.64 \times 10^{-8}$ | 1 |
| Clinical samples × Trade | -0.807 | 1.69 | -0.479 | 0.632 |
| Sludge, waste × Trade | -1.31 | 1.84 | -0.712 | 0.477 |
| Soil × Trade | -1.1 | $8.88 \times 10^{6}$ | $-1.23 \times 10^{-7}$ | 1 |
| Domestic animals × Migration | 3.86 | $7.45 \times 10^{5}$ | $5.18 \times 10^{-6}$ | 1 |
| Farms × Migration | -1.98 | 3 | -0.662 | 0.508 |
| Flora, fauna × Migration | 3.28 | 3.13 | 1.05 | 0.295 |
| Human samples × Migration | 1.75 | 2.76 | 0.632 | 0.528 |
| Human habitat × Migration | $-1.17 \times 10^{3}$ | $2.11 \times 10^{6}$ | $-5.54 \times 10^{-4}$ | 1 |
| Freshwater × Migration | 2.1 | $1.13 \times 10^{7}$ | 1.86 × 10-7 | 1 |
| Clinical samples × Migration | 1.94 | 2.75 | 0.703 | 0.482 |
| Sludge, waste × Migration | 3.16 | 3.11 | 1.01 | 0.31 |
| Soil × Migration | 2.03 | $8.93 \times 10^{6}$ | $2.27 \times 10^{-7}$ | 1 |

**Table S8: Summary of the selected model for CHG 6.1.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 0.971$, $\nu = 0.21$, $\rho = 9.39 \times 10^{-2}$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -5 | 1.82 | -2.75 | $5.93 \times 10^{-3}$ ** |
| Domestic animals | $-2.14 \times 10^{2}$ | $1.64 \times 10^{6}$ | $-1.31 \times 10^{-4}$ | 1 |
| Farms | 0.811 | 0.62 | 1.31 | 0.191 |
| Flora, fauna | -33 | $1.02 \times 10^{7}$ | $-3.25 \times 10^{-6}$ | 1 |
| Human samples | 1.35 | 0.636 | 2.13 | $3.34 \times 10^{-2}$ * |
| Human habitat | 2.55 | 0.999 | 2.55 | $1.07 \times 10^{-2}$ * |
| Freshwater | -3.14 | 2.09 | -1.5 | 0.132 |
| Clinical samples | -0.795 | 0.62 | -1.28 | 0.2 |
| Sludge, waste | -3.44 | 6.39 | -0.537 | 0.591 |
| Soil | 0.721 | 0.88 | 0.82 | 0.412 |
| Trade | 1.5 | 0.648 | 2.31 | $2.11 \times 10^{-2}$ * |
| Migration | -1.06 | 0.501 | -2.12 | $3.4 \times 10^{-2}$ * |
| Aminoglycosides | -2.55 | 0.901 | -2.83 | $4.67 \times 10^{-3}$ ** |
| Domestic animals × Trade | 59.7 | $7.05 \times 10^{5}$ | $8.48 \times 10^{-5}$ | 1 |
| Farms × Trade | -2.49 | 0.618 | -4.03 | $5.66 \times 10^{-5}$ *** |
| Flora, fauna × Trade | -0.179 | $1.33 \times 10^{7}$ | $-1.35 \times 10^{-8}$ | 1 |
| Human samples × Trade | -2.12 | 0.602 | -3.52 | $4.36 \times 10^{-4}$ *** |
| Human habitat × Trade | -0.994 | 1.09 | -0.916 | 0.36 |
| Freshwater × Trade | 3.51 | 1.68 | 2.09 | $3.64 \times 10^{-2}$ * |
| Clinical samples × Trade | -2.07 | 0.612 | -3.38 | $7.23 \times 10^{-4}$ *** |
| Sludge, waste × Trade | -7.3 | 7.41 | -0.985 | 0.324 |
| Soil × Trade | -3.01 | 0.829 | -3.62 | $2.89 \times 10^{-4}$ *** |
| Domestic animals × Migration | 51.9 | $1.18 \times 10^{6}$ | $4.41 \times 10^{-5}$ | 1 |
| Farms × Migration | 2.03 | 0.49 | 4.15 | $3.29 \times 10^{-5}$ *** |
| Flora, fauna × Migration | 0.553 | $9.46 \times 10^{6}$ | $5.84 \times 10^{-8}$ | 1 |
| Human samples × Migration | 2.02 | 0.518 | 3.91 | $9.16 \times 10^{-5}$ *** |
| Human habitat × Migration | 0.725 | 1.36 | 0.532 | 0.595 |
| Freshwater × Migration | -3.87 | 1.56 | -2.47 | $1.35 \times 10^{-2}$ * |
| Clinical samples × Migration | 1.61 | 0.507 | 3.17 | $1.51 \times 10^{-3}$ ** |
| Sludge, waste × Migration | 2.33 | 1.87 | 1.25 | 0.212 |
| Soil × Migration | 2.2 | 0.944 | 2.33 | $1.97 \times 10^{-2}$ * |
| Domestic animals × Aminoglycosides | $-1.43 \times 10^{2}$ | $2.2 \times 10^{6}$ | $-6.51 \times 10^{-5}$ | 1 |
| Farms × Aminoglycosides | 2.57 | 0.92 | 2.79 | $5.28 \times 10^{-3}$ ** |
| Flora, fauna × Aminoglycosides | 1.52 | $1.46 \times 10^{7}$ | $1.04 \times 10^{-7}$ | 1 |
| Human samples × Aminoglycosides | 1.63 | 0.943 | 1.73 | $8.38 \times 10^{-2}$ * |
| Human habitat × Aminoglycosides | 1.13 | 1.65 | 0.683 | 0.495 |
| Freshwater × Aminoglycosides | -3.19 | 2.49 | -1.28 | 0.2 |
| Clinical samples × Aminoglycosides | 2.42 | 0.917 | 2.64 | $8.23 \times 10^{-3}$ ** |
| Sludge, waste × Aminoglycosides | -0.529 | 4.7 | -0.113 | 0.91 |

| | | | | |
|---|---|---|---|---|
| Soil × Aminoglycosides | 4.3 | 1 | 4.29 | $1.78 \times 10^{-5}$ *** |

**Table S9: Summary of the selected model for CHG 7.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = -0.429, ν = 16.7, ρ = 1.45

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -5.57 | 0.764 | -7.28 | $3.26 \times 10^{-13}$ *** |
| Domestic animals | -16.8 | 30.6 | -0.548 | 0.583 |
| Farms | $1.01 \times 10^{-2}$ | 0.76 | $1.33 \times 10^{-2}$ | 0.989 |
| Flora, fauna | $-6.75 \times 10^{2}$ | $8 \times 10^{6}$ | $-8.44 \times 10^{-5}$ | 1 |
| Human samples | 0.76 | 0.782 | 0.971 | 0.332 |
| Human habitat | $-3.86 \times 10^{2}$ | $3.6 \times 10^{6}$ | $-1.07 \times 10^{-4}$ | 1 |
| Freshwater | $-2.19 \times 10^{2}$ | $8.67 \times 10^{6}$ | $-2.52 \times 10^{-5}$ | 1 |
| Clinical samples | -0.126 | 0.714 | -0.176 | 0.86 |
| Sludge, waste | $-2.14 \times 10^{2}$ | $1.19 \times 10^{7}$ | $-1.8 \times 10^{-5}$ | 1 |
| Soil | -4.54 | 6.36 | -0.713 | 0.476 |
| Trade | $9.54 \times 10^{-2}$ | 0.339 | 0.281 | 0.779 |
| Migration | 0.34 | 0.519 | 0.656 | 0.512 |
| Aminoglycosides | 0.679 | 0.381 | 1.78 | $7.47 \times 10^{-2}$ * |
| Domestic animals × Trade | 5.97 | 10 | 0.597 | 0.551 |
| Farms × Trade | 0.465 | 0.363 | 1.28 | 0.2 |
| Flora, fauna × Trade | $-3.71 \times 10^{2}$ | $9.92 \times 10^{6}$ | $-3.74 \times 10^{-5}$ | 1 |
| Human samples × Trade | 0.528 | 0.363 | 1.45 | 0.146 |
| Human habitat × Trade | 68.6 | $6.51 \times 10^{5}$ | $1.05 \times 10^{-4}$ | 1 |
| Freshwater × Trade | 4.6 | $7.92 \times 10^{6}$ | $5.81 \times 10^{-7}$ | 1 |
| Clinical samples × Trade | 0.384 | 0.349 | 1.1 | 0.271 |
| Sludge, waste × Trade | 5.08 | $8.1 \times 10^{6}$ | $6.28 \times 10^{-7}$ | 1 |
| Soil × Trade | 1.55 | 2.04 | 0.761 | 0.447 |
| Domestic animals × Migration | 3.65 | 7.53 | 0.484 | 0.628 |
| Farms × Migration | -0.68 | 0.576 | -1.18 | 0.238 |
| Flora, fauna × Migration | $-2.61 \times 10^{2}$ | $7.39 \times 10^{6}$ | $-3.53 \times 10^{-5}$ | 1 |
| Human samples × Migration | -0.237 | 0.548 | -0.433 | 0.665 |
| Human habitat × Migration | $1.79 \times 10^{2}$ | $1.67 \times 10^{6}$ | $1.07 \times 10^{-4}$ | 1 |
| Freshwater × Migration | 3.51 | $9.19 \times 10^{6}$ | $3.82 \times 10^{-7}$ | 1 |
| Clinical samples × Migration | 0.558 | 0.51 | 1.1 | 0.273 |
| Sludge, waste × Migration | 2.89 | $1.2 \times 10^{7}$ | $2.41 \times 10^{-7}$ | 1 |
| Soil × Migration | $-3.87 \times 10^{-2}$ | 1.97 | $-1.96 \times 10^{-2}$ | 0.984 |
| Domestic animals × Aminoglycosides | 6.77 | 10.6 | 0.639 | 0.523 |
| Farms × Aminoglycosides | 0.176 | 0.497 | 0.355 | 0.723 |
| Flora, fauna × Aminoglycosides | 22.8 | $8.37 \times 10^{6}$ | $2.72 \times 10^{-6}$ | 1 |
| Human samples × Aminoglycosides | 0.134 | 0.426 | 0.314 | 0.754 |
| Human habitat × Aminoglycosides | $-2.8 \times 10^{2}$ | $2.68 \times 10^{6}$ | $-1.04 \times 10^{-4}$ | 1 |
| Freshwater × Aminoglycosides | 3.51 | $9.83 \times 10^{6}$ | $3.57 \times 10^{-7}$ | 1 |
| Clinical samples × Aminoglycosides | -2.15 | 0.524 | -4.1 | $4.11 \times 10^{-5}$ *** |
| Sludge, waste × Aminoglycosides | 3.47 | $1.88 \times 10^{7}$ | $1.85 \times 10^{-7}$ | 1 |
| Soil × Aminoglycosides | 1.26 | 2.2 | 0.574 | 0.566 |

**Table S10: Summary of the selected model for CHG 8.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 4.44 \times 10^{-2}$, $\nu = 0.701$, $\rho = 0.205$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -4.95 | 0.639 | -7.74 | $1.01 \times 10^{-14}$ *** |
| Domestic animals | -13.1 | 7.71 | -1.69 | $9.04 \times 10^{-2}$ * |
| Farms | -1.41 | 0.645 | -2.19 | $2.87 \times 10^{-2}$ * |
| Flora, fauna | -3.97 | 6.36 | -0.624 | 0.532 |
| Human samples | 0.331 | 0.55 | 0.601 | 0.548 |
| Human habitat | -2.18 | 2.56 | -0.853 | 0.394 |
| Freshwater | -20.6 | $3.62 \times 10^{4}$ | $-5.68 \times 10^{-4}$ | 1 |
| Clinical samples | -0.764 | 0.516 | -1.48 | 0.138 |
| Sludge, waste | -2.86 | 4.7 | -0.608 | 0.543 |
| Soil | $6.16 \times 10^{-2}$ | 0.935 | $6.59 \times 10^{-2}$ | 0.947 |
| Trade | 1.42 | 0.638 | 2.23 | $2.57 \times 10^{-2}$ * |
| Migration | -0.434 | 0.801 | -0.542 | 0.588 |
| Aminoglycosides | -0.695 | 0.929 | -0.747 | 0.455 |
| Domestic animals × Trade | 6.73 | 3.94 | 1.71 | $8.78 \times 10^{-2}$ * |
| Farms × Trade | 0.151 | 0.697 | 0.216 | 0.829 |
| Flora, fauna × Trade | 1.84 | 4.92 | 0.375 | 0.708 |
| Human samples × Trade | -0.93 | 0.655 | -1.42 | 0.156 |
| Human habitat × Trade | 6.31 | 3.94 | 1.6 | 0.109 |
| Freshwater × Trade | -1.38 | $3.75 \times 10^{4}$ | $-3.67 \times 10^{-5}$ | 1 |
| Clinical samples × Trade | -0.883 | 0.641 | -1.38 | 0.168 |
| Sludge, waste × Trade | 2.59 | 13.2 | 0.196 | 0.844 |
| Soil × Trade | -4.7 | 1.9 | -2.47 | $1.34 \times 10^{-2}$ * |
| Domestic animals × Migration | -16.4 | 9.02 | -1.82 | $6.84 \times 10^{-2}$ * |
| Farms × Migration | -1.99 | 1 | -1.99 | $4.68 \times 10^{-2}$ * |
| Flora, fauna × Migration | -4.68 | 7.69 | -0.609 | 0.543 |
| Human samples × Migration | 0.248 | 0.834 | 0.298 | 0.766 |
| Human habitat × Migration | -9.79 | 5.03 | -1.95 | $5.17 \times 10^{-2}$ * |
| Freshwater × Migration | 0.484 | $5.19 \times 10^{4}$ | $9.33 \times 10^{-6}$ | 1 |
| Clinical samples × Migration | 0.99 | 0.805 | 1.23 | 0.219 |
| Sludge, waste × Migration | $3.63 \times 10^{-2}$ | 4.21 | $8.62 \times 10^{-3}$ | 0.993 |
| Soil × Migration | 1.47 | 0.884 | 1.67 | $9.57 \times 10^{-2}$ * |
| Domestic animals × Aminoglycosides | 6.28 | 3.18 | 1.98 | $4.82 \times 10^{-2}$ * |
| Farms × Aminoglycosides | 2.06 | 0.954 | 2.15 | $3.12 \times 10^{-2}$ * |
| Flora, fauna × Aminoglycosides | -4.24 | 7.09 | -0.597 | 0.55 |
| Human samples × Aminoglycosides | 1.37 | 0.935 | 1.46 | 0.144 |
| Human habitat × Aminoglycosides | -0.23 | 1.62 | -0.142 | 0.887 |
| Freshwater × Aminoglycosides | 0.649 | $3.07 \times 10^{4}$ | $2.12 \times 10^{-5}$ | 1 |
| Clinical samples × Aminoglycosides | -0.729 | 0.963 | -0.757 | 0.449 |
| Sludge, waste × Aminoglycosides | -3.34 | 17.6 | -0.19 | 0.849 |
| Soil × Aminoglycosides | 0.481 | 1.06 | 0.452 | 0.651 |

**Table S11: Summary of the selected model for CHG 11.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.368, ν = 16.7, ρ = 1.49

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -35.7 | $3.84 \times 10^{6}$ | $-9.28 \times 10^{-6}$ | 1 |
| Domestic animals | -48.9 | $6.94 \times 10^{6}$ | $-7.05 \times 10^{-6}$ | 1 |
| Farms | 29.5 | $3.84 \times 10^{6}$ | $7.68 \times 10^{-6}$ | 1 |
| Flora, fauna | -50.1 | $8.65 \times 10^{6}$ | $-5.79 \times 10^{-6}$ | 1 |
| Human samples | 30.9 | $3.84 \times 10^{6}$ | $8.05 \times 10^{-6}$ | 1 |
| Human habitat | -49.1 | $1.1 \times 10^{7}$ | $-4.44 \times 10^{-6}$ | 1 |
| Freshwater | -52.4 | $9.23 \times 10^{6}$ | $-5.68 \times 10^{-6}$ | 1 |
| Clinical samples | 29.7 | $3.84 \times 10^{6}$ | $7.73 \times 10^{-6}$ | 1 |
| Sludge, waste | -48.2 | $5.81 \times 10^{6}$ | $-8.29 \times 10^{-6}$ | 1 |
| Soil | -49.9 | $6.78 \times 10^{6}$ | $-7.35 \times 10^{-6}$ | 1 |
| Trade | 0.634 | 0.189 | 3.35 | $8 \times 10^{-4}$ *** |
| Migration | -0.342 | 0.329 | -1.04 | 0.299 |

**Table S12: Summary of the selected model for CHG 13.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.874, ν = 16.7, ρ = 2.75

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -40.6 | $2.78 \times 10^{6}$ | $-1.46 \times 10^{-5}$ | 1 |
| Domestic animals | -76.2 | $6.41 \times 10^{6}$ | $-1.19 \times 10^{-5}$ | 1 |
| Farms | 33.5 | $2.78 \times 10^{6}$ | $1.2 \times 10^{-5}$ | 1 |
| Flora, fauna | -78.9 | $8.23 \times 10^{6}$ | $-9.59 \times 10^{-6}$ | 1 |
| Human samples | 34.6 | $2.78 \times 10^{6}$ | $1.24 \times 10^{-5}$ | 1 |
| Human habitat | -76.3 | $1.07 \times 10^{7}$ | $-7.11 \times 10^{-6}$ | 1 |
| Freshwater | -91.7 | $8.84 \times 10^{6}$ | $-1.04 \times 10^{-5}$ | 1 |
| Clinical samples | 34.9 | $2.78 \times 10^{6}$ | $1.25 \times 10^{-5}$ | 1 |
| Sludge, waste | -75.8 | $5.17 \times 10^{6}$ | $-1.46 \times 10^{-5}$ | 1 |
| Soil | -82.4 | $6.25 \times 10^{6}$ | $-1.32 \times 10^{-5}$ | 1 |

**Table S13: Summary of the selected model for CHG 14.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 5.13 \times 10^{-4}$, $\nu = 5.24 \times 10^{-3}$, $\rho = 0.724$

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -39 | $3.05 \times 10^{6}$ | $-1.28 \times 10^{-5}$ | 1 |
| Domestic animals | -16.3 | $6.53 \times 10^{6}$ | $-2.49 \times 10^{-6}$ | 1 |
| Farms | 33.8 | $3.05 \times 10^{6}$ | $1.11 \times 10^{-5}$ | 1 |
| Flora, fauna | -17 | $8.33 \times 10^{6}$ | $-2.04 \times 10^{-6}$ | 1 |
| Human samples | -16.1 | $3.6 \times 10^{6}$ | $-4.46 \times 10^{-6}$ | 1 |
| Human habitat | -17.9 | $1.08 \times 10^{7}$ | $-1.66 \times 10^{-6}$ | 1 |
| Freshwater | -15.8 | $8.93 \times 10^{6}$ | $-1.77 \times 10^{-6}$ | 1 |
| Clinical samples | 30.8 | $3.05 \times 10^{6}$ | $1.01 \times 10^{-5}$ | 1 |
| Sludge, waste | -18.1 | $5.32 \times 10^{6}$ | $-3.4 \times 10^{-6}$ | 1 |
| Soil | -16.8 | $6.37 \times 10^{6}$ | $-2.63 \times 10^{-6}$ | 1 |
| Trade | 0.893 | 0.135 | 6.6 | $4.19 \times 10^{-11}$ *** |
| Migration | 0.173 | 0.242 | 0.717 | 0.473 |

**Table S14: Summary of the selected model for CHG 17.1.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 4.14 \times 10^{-3}$, $\nu = 16.7$, $\rho = 0.379$

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -39.9 | $2.83 \times 10^{6}$ | $-1.41 \times 10^{-5}$ | 1 |
| Domestic animals | $-1.95 \times 10^{2}$ | $6.43 \times 10^{6}$ | $-3.03 \times 10^{-5}$ | 1 |
| Farms | $-1.97 \times 10^{2}$ | $3.2 \times 10^{6}$ | $-6.14 \times 10^{-5}$ | 1 |
| Flora, fauna | $-1.97 \times 10^{2}$ | $8.25 \times 10^{6}$ | $-2.39 \times 10^{-5}$ | 1 |
| Human samples | 29.2 | $2.83 \times 10^{6}$ | $1.03 \times 10^{-5}$ | 1 |
| Human habitat | $-1.97 \times 10^{2}$ | $1.07 \times 10^{7}$ | $-1.83 \times 10^{-5}$ | 1 |
| Freshwater | 31.7 | $2.83 \times 10^{6}$ | $1.12 \times 10^{-5}$ | 1 |
| Clinical samples | 31.5 | $2.83 \times 10^{6}$ | $1.11 \times 10^{-5}$ | 1 |
| Sludge, waste | 32 | $2.83 \times 10^{6}$ | $1.13 \times 10^{-5}$ | 1 |
| Soil | $-1.98 \times 10^{2}$ | $6.27 \times 10^{6}$ | $-3.15 \times 10^{-5}$ | 1 |
| Trade | 0.805 | 0.29 | 2.77 | $5.52 \times 10^{-3}$ ** |
| Migration | $8.64 \times 10^{-2}$ | 0.389 | 0.222 | 0.824 |
| Aminoglycosides | -3.95 | 0.782 | -5.05 | $4.43 \times 10^{-7}$ *** |

**Table S15: Summary of the selected model for CHG 19.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.842, ν = 16.7, ρ = 3.23

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -6.95 | 1.38 | -5.03 | $4.87 \times 10^{-7}$ *** |
| Domestic animals | -15.6 | 11.9 | -1.31 | 0.19 |
| Farms | 1.32 | 1.27 | 1.04 | 0.297 |
| Flora, fauna | -2.06 | 5.33 | -0.386 | 0.699 |
| Human samples | 1.71 | 1.28 | 1.33 | 0.183 |
| Human habitat | $-1.33 \times 10^{2}$ | $1.98 \times 10^{6}$ | $-6.71 \times 10^{-5}$ | 1 |
| Freshwater | -30.1 | $8.69 \times 10^{6}$ | $-3.46 \times 10^{-6}$ | 1 |
| Clinical samples | 3.24 | 1.21 | 2.68 | $7.41 \times 10^{-3}$ ** |
| Sludge, waste | 2.1 | 1.49 | 1.41 | 0.159 |
| Soil | -29.8 | $7.58 \times 10^{6}$ | $-3.94 \times 10^{-6}$ | 1 |
| Trade | 0.169 | 0.697 | 0.243 | 0.808 |
| Migration | 0.997 | 1.27 | 0.785 | 0.433 |
| Aminoglycosides | -0.643 | 0.199 | -3.23 | $1.25 \times 10^{-3}$ ** |
| Domestic animals × Trade | -1.5 | 0.956 | -1.57 | 0.117 |
| Farms × Trade | 0.362 | 0.751 | 0.482 | 0.63 |
| Flora, fauna × Trade | -12.3 | 7.89 | -1.56 | 0.12 |
| Human samples × Trade | 0.329 | 0.721 | 0.456 | 0.648 |
| Human habitat × Trade | $1.17 \times 10^{2}$ | $1.39 \times 10^{6}$ | $8.39 \times 10^{-5}$ | 1 |
| Freshwater × Trade | -0.456 | $9.8 \times 10^{6}$ | $-4.65 \times 10^{-8}$ | 1 |
| Clinical samples × Trade | -0.497 | 0.697 | -0.713 | 0.476 |
| Sludge, waste × Trade | 0.779 | 0.885 | 0.881 | 0.378 |
| Soil × Trade | -0.613 | $7.35 \times 10^{6}$ | $-8.35 \times 10^{-8}$ | 1 |
| Domestic animals × Migration | 16.1 | 9.02 | 1.79 | $7.38 \times 10^{-2}$ * |
| Farms × Migration | -2.05 | 1.34 | -1.53 | 0.126 |
| Flora, fauna × Migration | -0.988 | 1.71 | -0.578 | 0.563 |
| Human samples × Migration | -0.633 | 1.32 | -0.48 | 0.631 |
| Human habitat × Migration | $-1.59 \times 10^{2}$ | $1.99 \times 10^{6}$ | $-7.98 \times 10^{-5}$ | 1 |
| Freshwater × Migration | -1.26 | $1.03 \times 10^{7}$ | $-1.22 \times 10^{-7}$ | 1 |
| Clinical samples × Migration | -1.02 | 1.26 | -0.809 | 0.419 |
| Sludge, waste × Migration | -0.713 | 1.44 | -0.493 | 0.622 |
| Soil × Migration | -1.08 | $7.66 \times 10^{6}$ | $-1.41 \times 10^{-7}$ | 1 |

**Table S16: Summary of the selected model for CHG 20.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = -0.922, ν = 0.65, ρ = 10.3

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -39.1 | $3.01 \times 10^{6}$ | $-1.3 \times 10^{-5}$ | 1 |
| Domestic animals | 1.73 | $6.89 \times 10^{6}$ | $2.51 \times 10^{-7}$ | 1 |
| Farms | 33.7 | $3.01 \times 10^{6}$ | $1.12 \times 10^{-5}$ | 1 |
| Flora, fauna | 1.58 | $9.19 \times 10^{6}$ | $1.71 \times 10^{-7}$ | 1 |
| Human samples | 2.48 | $3.64 \times 10^{6}$ | $6.81 \times 10^{-7}$ | 1 |
| Human habitat | 0.782 | $1.23 \times 10^{7}$ | $6.35 \times 10^{-8}$ | 1 |
| Freshwater | 1.97 | $9.04 \times 10^{6}$ | $2.18 \times 10^{-7}$ | 1 |
| Clinical samples | 30.9 | $3.01 \times 10^{6}$ | $1.03 \times 10^{-5}$ | 1 |
| Sludge, waste | -0.653 | $5.3 \times 10^{6}$ | $-1.23 \times 10^{-7}$ | 1 |
| Soil | 1.69 | $7.3 \times 10^{6}$ | $2.31 \times 10^{-7}$ | 1 |
| Trade | 1.04 | 0.198 | 5.26 | $1.44 \times 10^{-7}$ *** |
| Migration | 0.117 | 0.343 | 0.342 | 0.732 |

**Table S17: Summary of the selected model for CHG 23.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = -0.738$, $\nu = 4.83 \times 10^{-2}$, $\rho = 6.65$

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -41.4 | $3.52 \times 10^{6}$ | $-1.17 \times 10^{-5}$ | 1 |
| Domestic animals | -8.59 | $6.76 \times 10^{6}$ | $-1.27 \times 10^{-6}$ | 1 |
| Farms | -13.3 | $3.82 \times 10^{6}$ | $-3.48 \times 10^{-6}$ | 1 |
| Flora, fauna | -16.6 | $8.51 \times 10^{6}$ | $-1.95 \times 10^{-6}$ | 1 |
| Human samples | 29.3 | $3.52 \times 10^{6}$ | $8.33 \times 10^{-6}$ | 1 |
| Human habitat | -16.2 | $1.09 \times 10^{7}$ | $-1.48 \times 10^{-6}$ | 1 |
| Freshwater | -37.4 | $9.1 \times 10^{6}$ | $-4.11 \times 10^{-6}$ | 1 |
| Clinical samples | 32.1 | $3.52 \times 10^{6}$ | $9.12 \times 10^{-6}$ | 1 |
| Sludge, waste | -11.9 | $5.6 \times 10^{6}$ | $-2.12 \times 10^{-6}$ | 1 |
| Soil | -21.5 | $6.61 \times 10^{6}$ | $-3.26 \times 10^{-6}$ | 1 |
| Trade | 1.12 | 0.441 | 2.54 | $1.12 \times 10^{-2}$ * |
| Migration | -0.395 | 0.283 | -1.4 | 0.163 |

**Table S18: Summary of the selected model for CHG 27.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 1.31 \times 10^{-7}$, $\nu = 0.33$, $\rho = 7.09 \times 10^{-2}$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -7.25 | 2.95 | -2.46 | $1.39 \times 10^{-2}$ * |
| Domestic animals | -30.7 | $8.31 \times 10^{6}$ | $-3.69 \times 10^{-6}$ | 1 |
| Farms | -6.4 | 1.97 | -3.26 | $1.13 \times 10^{-3}$ ** |
| Flora, fauna | -31.2 | $1.54 \times 10^{7}$ | $-2.02 \times 10^{-6}$ | 1 |
| Human samples | -0.577 | 0.691 | -0.836 | 0.403 |
| Human habitat | -22.9 | $5.58 \times 10^{7}$ | $-4.09 \times 10^{-7}$ | 1 |
| Freshwater | -2.88 | 5.5 | -0.524 | 0.6 |
| Clinical samples | -0.555 | 0.478 | -1.16 | 0.246 |
| Sludge, waste | -30 | $4.52 \times 10^{6}$ | $-6.64 \times 10^{-6}$ | 1 |
| Soil | -0.701 | 1.04 | -0.676 | 0.499 |
| Trade | -1.67 | 0.216 | -7.75 | $9.1 \times 10^{-15}$ *** |
| Migration | -2.77 | 0.74 | -3.74 | $1.85 \times 10^{-4}$ *** |
| Domestic animals × Trade | -1 | $9.69 \times 10^{6}$ | $-1.03 \times 10^{-7}$ | 1 |
| Farms × Trade | 3.06 | 0.886 | 3.46 | $5.44 \times 10^{-4}$ *** |
| Flora, fauna × Trade | -0.336 | $8.97 \times 10^{6}$ | $-3.74 \times 10^{-8}$ | 1 |
| Human samples × Trade | 0.183 | 0.914 | 0.2 | 0.841 |
| Human habitat × Trade | -5.51 | $5.58 \times 10^{7}$ | $-9.87 \times 10^{-8}$ | 1 |
| Freshwater × Trade | -4.57 | 11.2 | -0.408 | 0.683 |
| Clinical samples × Trade | -1.17 | 0.64 | -1.83 | $6.69 \times 10^{-2}$ * |
| Sludge, waste × Trade | -0.979 | $5.62 \times 10^{6}$ | $-1.74 \times 10^{-7}$ | 1 |
| Soil × Trade | 0.215 | 0.642 | 0.335 | 0.737 |
| Domestic animals × Migration | -1.04 | $1.34 \times 10^{7}$ | $-7.8 \times 10^{-8}$ | 1 |
| Farms × Migration | -10.1 | 2.73 | -3.68 | $2.31 \times 10^{-4}$ *** |
| Flora, fauna × Migration | -1.17 | $2.83 \times 10^{7}$ | $-4.13 \times 10^{-8}$ | 1 |
| Human samples × Migration | -2.9 | 0.894 | -3.25 | $1.16 \times 10^{-3}$ ** |
| Human habitat × Migration | 13 | $1.06 \times 10^{8}$ | $1.22 \times 10^{-7}$ | 1 |
| Freshwater × Migration | -4.33 | 7.94 | -0.546 | 0.585 |
| Clinical samples × Migration | -2.16 | 0.334 | -6.48 | $9.4 \times 10^{-11}$ *** |
| Sludge, waste × Migration | -2.77 | $5.52 \times 10^{6}$ | $-5.01 \times 10^{-7}$ | 1 |
| Soil × Migration | -1.46 | 0.87 | -1.68 | $9.39 \times 10^{-2}$ * |

**Table S19: Human exchanges variables influencing the prevalence of ARBs**

| | CHG 2 | CHG 3 | CHG 4 | CHG 5.1 | CHG 6.1 | CHG 7 | CHG 8 | CHG 11 | CHG 14 | CHG 17.1 | CHG 19 | CHG 20 | CHG 23 | CHG 27 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Animal and vegetable oils, fats and waxes | | | | | | | | | × | × | | | × | |
| Beverages | × | | | | | | | | × | × | | × | × | |
| Cereals and cereal preparations | | × | | | | | | | × | × | | × | | |
| Coal, coke and briquettes | | | | | | | | | × | × | | × | | |
| Coffee, tea, cocoa, spices, and manufactures thereof | | | | | | | | | × | × | | | | |
| Cork and wood | | | × | | | | | | × | × | | × | | |
| Crude materials, inedible, except fuels | | | | | | | | | | × | | | | |
| Crude rubber (including synthetic and reclaimed) | | | | × | | | | | | × | | | × | |
| Dairy products and birds' eggs | | | | | | | | | × | × | | × | × | |
| Dyeing, tanning and colouring materials | | × | | | | | | | × | × | | | × | |
| Feedstuff for animals (excluding unmilled cereals) | × | | | | | | | | × | × | | × | | |
| Fish, crustaceans, molluscs and preparations thereof | | | | | | | | | × | × | | × | | |
| Food and live animals | | | | | | | | | × | × | | × | | |
| Hides, skins and furskins, raw | | | | | | | | | | × | | | × | |
| Live animals other than animals of division 03 | | × | | | | | | | | × | | | | |
| Manufactured goods | | | | | | | | | × | × | | × | | |
| Meat and meat preparations | × | | | | | | | | × | × | × | × | | |
| Medicinal and pharmaceutical products | × | | | | | | | | × | × | | | × | |
| Mineral fuels, lubricants and related materials | × | | | | | | | × | × | × | | | | |
| Miscellaneous edible products and preparations | × | | | | | | | | × | × | | | × | |
| Oil seeds and oleaginous fruits | | × | | | × | × | | | | × | | | × | |
| Petroleum, petroleum products and related materials | × | | | | | | | | × | × | | | | |
| Plastics | | | | | | | | | × | × | | | × | |
| Sugar, sugar preparations and honey | × | | | | | | × | | × | × | × | | | |
| Textiles fibres and their wastes | | | × | × | | | | | | × | | | | |
| Vegetables and fruits | | | | | | | | | × | × | | × | | |
| Human migration | × | | | | | | × | | | × | | | | × |

**Table S20: Summary of the selected model for CHG 1 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.264, ν = 16.7, ρ = 0.729

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -7.91 | 0.833 | -9.51 | 0 *** |
| Domestic animals | -29.8 | $4.08 \times 10^{6}$ | $-7.32 \times 10^{-6}$ | 1 |
| Farms | 0.411 | 0.76 | 0.541 | 0.589 |
| Flora, fauna | -29.9 | $3.97 \times 10^{6}$ | $-7.54 \times 10^{-6}$ | 1 |
| Human samples | -0.307 | 0.815 | -0.376 | 0.707 |
| Human habitat | 0.279 | 1.24 | 0.225 | 0.822 |
| Freshwater | -0.622 | 1.23 | -0.504 | 0.614 |
| Clinical samples | 1.65 | 0.721 | 2.28 | $2.23 \times 10^{-2}$ * |
| Sludge, waste | 2.2 | 0.933 | 2.35 | $1.86 \times 10^{-2}$ * |
| Soil | -30.1 | $2.02 \times 10^{6}$ | $-1.49 \times 10^{-5}$ | 1 |
| Trade | -0.136 | $7.04 \times 10^{-2}$ | -1.93 | $5.36 \times 10^{-2}$ * |
| Migration | -0.162 | $7.78 \times 10^{-2}$ | -2.09 | $3.68 \times 10^{-2}$ * |

**Table S21: Summary of the selected model for CHG 2 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = -0.167$, $\nu = 9.1 \times 10^{-3}$, $\rho = 1.48$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -5.81 | 0.556 | -10.4 | 0 *** |
| Domestic animals | -32 | $4.43 \times 10^{6}$ | $-7.22 \times 10^{-6}$ | 1 |
| Farms | -0.117 | 0.277 | -0.423 | 0.673 |
| Flora, fauna | -20.2 | 23.9 | -0.846 | 0.397 |
| Human samples | -2.84 | 0.657 | -4.32 | $1.57 \times 10^{-5}$ *** |
| Human habitat | -31.6 | $4.08 \times 10^{6}$ | $-7.74 \times 10^{-6}$ | 1 |
| Freshwater | -3.6 | 1.68 | -2.15 | $3.17 \times 10^{-2}$ * |
| Clinical samples | -2.04 | 0.282 | -7.26 | $3.98 \times 10^{-13}$ *** |
| Sludge, waste | -31.1 | $3.88 \times 10^{6}$ | $-8 \times 10^{-6}$ | 1 |
| Soil | -1.4 | 0.588 | -2.38 | $1.73 \times 10^{-2}$ * |
| Trade | -1.03 | 0.317 | -3.26 | $1.11 \times 10^{-3}$ ** |
| Migration | 1.09 | 0.168 | 6.5 | $8.08 \times 10^{-11}$ *** |
| Domestic animals × Trade | 0.743 | $2.22 \times 10^{6}$ | $3.35 \times 10^{-7}$ | 1 |
| Farms × Trade | 0.95 | 0.307 | 3.1 | $1.97 \times 10^{-3}$ ** |
| Flora, fauna × Trade | -83.2 | 87.9 | -0.947 | 0.343 |
| Human samples × Trade | -0.455 | 1.78 | -0.256 | 0.798 |
| Human habitat × Trade | 0.741 | $1.17 \times 10^{6}$ | $6.33 \times 10^{-7}$ | 1 |
| Freshwater × Trade | -7.78 | 5.67 | -1.37 | 0.17 |
| Clinical samples × Trade | 0.958 | 0.31 | 3.09 | $2.03 \times 10^{-3}$ ** |
| Sludge, waste × Trade | 0.618 | $3.11 \times 10^{6}$ | $1.99 \times 10^{-7}$ | 1 |
| Soil × Trade | 0.781 | 0.566 | 1.38 | 0.167 |
| Domestic animals × Migration | -1.01 | $2.19 \times 10^{6}$ | $-4.62 \times 10^{-7}$ | 1 |
| Farms × Migration | -1.33 | 0.216 | -6.15 | $7.85 \times 10^{-10}$ *** |
| Flora, fauna × Migration | -0.18 | 2.61 | $-6.88 \times 10^{-2}$ | 0.945 |
| Human samples × Migration | -2.09 | 1.46 | -1.43 | 0.154 |
| Human habitat × Migration | -1.46 | $4.82 \times 10^{6}$ | $-3.03 \times 10^{-7}$ | 1 |
| Freshwater × Migration | -3.82 | 4.09 | -0.932 | 0.351 |
| Clinical samples × Migration | -1.61 | 0.313 | -5.13 | $2.84 \times 10^{-7}$ *** |
| Sludge, waste × Migration | -1.28 | $3.76 \times 10^{6}$ | $-3.41 \times 10^{-7}$ | 1 |
| Soil × Migration | -1.21 | 0.856 | -1.41 | 0.158 |

**Table S22: Summary of the selected model for CHG 3 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = -0.208, ν = 1.53, ρ = 2.18

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -8.37 | 0.984 | -8.51 | 0 *** |
| Domestic animals | $-1.94 \times 10^{3}$ | $2.9 \times 10^{5}$ | $-6.71 \times 10^{-3}$ | 0.995 |
| Farms | -1.59 | 1.44 | -1.1 | 0.27 |
| Flora, fauna | 2.83 | 1.12 | 2.54 | $1.12 \times 10^{-2}$ * |
| Human samples | 2.92 | 0.857 | 3.4 | $6.64 \times 10^{-4}$ *** |
| Human habitat | 0.778 | 1.57 | 0.495 | 0.621 |
| Freshwater | $3.3 \times 10^{-2}$ | 1.15 | $2.85 \times 10^{-2}$ | 0.977 |
| Clinical samples | 3.48 | 0.853 | 4.08 | $4.59 \times 10^{-5}$ *** |
| Sludge, waste | -1.08 | 2.33 | -0.464 | 0.643 |
| Soil | 2.36 | 0.888 | 2.66 | $7.88 \times 10^{-3}$ ** |
| Trade | -0.7 | 0.429 | -1.63 | 0.103 |
| Migration | -1.41 | 1.63 | -0.864 | 0.387 |
| Domestic animals × Trade | $2.14 \times 10^{2}$ | $3.2 \times 10^{4}$ | $6.67 \times 10^{-3}$ | 0.995 |
| Farms × Trade | -3.17 | 2.17 | -1.46 | 0.143 |
| Flora, fauna × Trade | -1.18 | 1.49 | -0.793 | 0.428 |
| Human samples × Trade | -0.76 | 0.449 | -1.69 | $9.08 \times 10^{-2}$ * |
| Human habitat × Trade | $9.91 \times 10^{-2}$ | 0.992 | 0.1 | 0.92 |
| Freshwater × Trade | -0.7 | 0.512 | -1.37 | 0.172 |
| Clinical samples × Trade | 0.495 | 0.429 | 1.15 | 0.249 |
| Sludge, waste × Trade | -4.15 | 3.48 | -1.19 | 0.233 |
| Soil × Trade | 0.743 | 0.453 | 1.64 | 0.1 |
| Domestic animals × Migration | $-2.94 \times 10^{3}$ | $4.37 \times 10^{5}$ | $-6.72 \times 10^{-3}$ | 0.995 |
| Farms × Migration | -0.703 | 2.21 | -0.318 | 0.75 |
| Flora, fauna × Migration | 0.675 | 1.72 | 0.392 | 0.695 |
| Human samples × Migration | 1.55 | 1.63 | 0.95 | 0.342 |
| Human habitat × Migration | $-5.3 \times 10^{-2}$ | 2.47 | $-2.14 \times 10^{-2}$ | 0.983 |
| Freshwater × Migration | -6.51 | 2.14 | -3.04 | $2.4 \times 10^{-3}$ ** |
| Clinical samples × Migration | 1.44 | 1.63 | 0.882 | 0.378 |
| Sludge, waste × Migration | -4.59 | 5.36 | -0.857 | 0.391 |
| Soil × Migration | $7.08 \times 10^{-2}$ | 1.67 | $4.25 \times 10^{-2}$ | 0.966 |

**Table S23: Summary of the selected model for CHG 4 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 0.216$, $\nu = 5.49 \times 10^{-3}$, $\rho = 2.12$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -6.84 | 0.698 | -9.81 | 0 *** |
| Domestic animals | -30.7 | $4.59 \times 10^{6}$ | $-6.69 \times 10^{-6}$ | 1 |
| Farms | 0.109 | 0.49 | 0.222 | 0.824 |
| Flora, fauna | 2.27 | 0.598 | 3.79 | $1.48 \times 10^{-4}$ *** |
| Human samples | -1.99 | 0.956 | -2.08 | $3.71 \times 10^{-2}$ * |
| Human habitat | -3.12 | 1.62 | -1.93 | $5.37 \times 10^{-2}$ * |
| Freshwater | -3.28 | 3.4 | -0.965 | 0.335 |
| Clinical samples | -0.106 | 0.459 | -0.231 | 0.817 |
| Sludge, waste | -31.2 | $5.99 \times 10^{6}$ | $-5.21 \times 10^{-6}$ | 1 |
| Soil | -1.45 | 1.24 | -1.17 | 0.243 |
| Trade | 0.328 | 0.437 | 0.752 | 0.452 |
| Migration | -1.8 | 0.983 | -1.83 | $6.7 \times 10^{-2}$ * |
| Domestic animals × Trade | -1.12 | $6.93 \times 10^{6}$ | $-1.62 \times 10^{-7}$ | 1 |
| Farms × Trade | -1.59 | 0.527 | -3.01 | $2.6 \times 10^{-3}$ ** |
| Flora, fauna × Trade | -2.47 | 1.16 | -2.13 | $3.32 \times 10^{-2}$ * |
| Human samples × Trade | -1.03 | 0.741 | -1.39 | 0.163 |
| Human habitat × Trade | 0.215 | 0.973 | 0.221 | 0.825 |
| Freshwater × Trade | -16.4 | 14.4 | -1.14 | 0.254 |
| Clinical samples × Trade | -0.478 | 0.44 | -1.09 | 0.277 |
| Sludge, waste × Trade | -0.682 | $5.94 \times 10^{6}$ | $-1.15 \times 10^{-7}$ | 1 |
| Soil × Trade | -0.349 | 1.21 | -0.288 | 0.774 |
| Domestic animals × Migration | 2.23 | $7.94 \times 10^{6}$ | $2.81 \times 10^{-7}$ | 1 |
| Farms × Migration | 0.544 | 1.11 | 0.488 | 0.625 |
| Flora, fauna × Migration | 1.62 | 1.24 | 1.31 | 0.191 |
| Human samples × Migration | -3.29 | 3.33 | -0.986 | 0.324 |
| Human habitat × Migration | -9.21 | 6.46 | -1.43 | 0.154 |
| Freshwater × Migration | -1.16 | 6.12 | -0.189 | 0.85 |
| Clinical samples × Migration | 1.42 | 0.985 | 1.44 | 0.15 |
| Sludge, waste × Migration | 1.82 | $6.44 \times 10^{6}$ | $2.83 \times 10^{-7}$ | 1 |
| Soil × Migration | 1.22 | 1.8 | 0.68 | 0.496 |

**Table S24: Summary of the selected model for CHG 5.1 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 0.755$, $\nu = 1.27 \times 10^{-2}$, $\rho = 1.17$

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | $-1.43 \times 10^{2}$ | $6.01 \times 10^{3}$ | $-2.38 \times 10^{-2}$ | 0.981 |
| Domestic animals | $1.33 \times 10^{2}$ | $6.01 \times 10^{3}$ | $2.22 \times 10^{-2}$ | 0.982 |
| Farms | $1.35 \times 10^{2}$ | $6.01 \times 10^{3}$ | $2.25 \times 10^{-2}$ | 0.982 |
| Flora, fauna | $1.38 \times 10^{2}$ | $6.01 \times 10^{3}$ | $2.3 \times 10^{-2}$ | 0.982 |
| Human samples | $1.37 \times 10^{2}$ | $6.01 \times 10^{3}$ | $2.28 \times 10^{-2}$ | 0.982 |
| Human habitat | $1.37 \times 10^{2}$ | $6.01 \times 10^{3}$ | $2.29 \times 10^{-2}$ | 0.982 |
| Freshwater | $-1.55 \times 10^{2}$ | $6.83 \times 10^{4}$ | $-2.28 \times 10^{-3}$ | 0.998 |
| Clinical samples | $1.39 \times 10^{2}$ | $6.01 \times 10^{3}$ | $2.31 \times 10^{-2}$ | 0.982 |
| Sludge, waste | $1.39 \times 10^{2}$ | $6.01 \times 10^{3}$ | $2.31 \times 10^{-2}$ | 0.982 |
| Soil | $1.32 \times 10^{2}$ | $6.01 \times 10^{3}$ | $2.2 \times 10^{-2}$ | 0.982 |
| Trade | 32.5 | $1.43 \times 10^{3}$ | $2.27 \times 10^{-2}$ | 0.982 |
| Migration | -7.17 | $3.77 \times 10^{2}$ | $-1.9 \times 10^{-2}$ | 0.985 |
| Domestic animals × Trade | -31.3 | $1.43 \times 10^{3}$ | $-2.18 \times 10^{-2}$ | 0.983 |
| Farms × Trade | -31.8 | $1.43 \times 10^{3}$ | $-2.22 \times 10^{-2}$ | 0.982 |
| Flora, fauna × Trade | -33.2 | $1.43 \times 10^{3}$ | $-2.32 \times 10^{-2}$ | 0.982 |
| Human samples × Trade | -31.7 | $1.43 \times 10^{3}$ | $-2.21 \times 10^{-2}$ | 0.982 |
| Human habitat × Trade | -32.6 | $1.43 \times 10^{3}$ | $-2.28 \times 10^{-2}$ | 0.982 |
| Freshwater × Trade | 33.7 | $1.54 \times 10^{4}$ | $2.19 \times 10^{-3}$ | 0.998 |
| Clinical samples × Trade | -32.9 | $1.43 \times 10^{3}$ | $-2.29 \times 10^{-2}$ | 0.982 |
| Sludge, waste × Trade | -32.6 | $1.43 \times 10^{3}$ | $-2.27 \times 10^{-2}$ | 0.982 |
| Soil × Trade | -32 | $1.43 \times 10^{3}$ | $-2.24 \times 10^{-2}$ | 0.982 |
| Domestic animals × Migration | 6.77 | $3.77 \times 10^{2}$ | $1.8 \times 10^{-2}$ | 0.986 |
| Farms × Migration | 6.55 | $3.77 \times 10^{2}$ | $1.74 \times 10^{-2}$ | 0.986 |
| Flora, fauna × Migration | 6.58 | $3.77 \times 10^{2}$ | $1.75 \times 10^{-2}$ | 0.986 |
| Human samples × Migration | 6.48 | $3.77 \times 10^{2}$ | $1.72 \times 10^{-2}$ | 0.986 |
| Human habitat × Migration | 6.99 | $3.77 \times 10^{2}$ | $1.85 \times 10^{-2}$ | 0.985 |
| Freshwater × Migration | $-4.24 \times 10^{2}$ | $1.01 \times 10^{5}$ | $-4.18 \times 10^{-3}$ | 0.997 |
| Clinical samples × Migration | 7.33 | $3.77 \times 10^{2}$ | $1.94 \times 10^{-2}$ | 0.984 |
| Sludge, waste × Migration | 6.91 | $3.77 \times 10^{2}$ | $1.83 \times 10^{-2}$ | 0.985 |
| Soil × Migration | 1.92 | $3.77 \times 10^{2}$ | $5.09 \times 10^{-3}$ | 0.996 |

**Table S25: Summary of the selected model for CHG 6.1 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 0.747$, $\nu = 0.415$, $\rho = 4.37 \times 10^{-2}$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -3.24 | 0.718 | -4.51 | $6.49 \times 10^{-6}$ *** |
| Domestic animals | -1.2 | 0.417 | -2.88 | $3.94 \times 10^{-3}$ ** |
| Farms | 1.03 | 0.134 | 7.71 | $1.28 \times 10^{-14}$ *** |
| Flora, fauna | 1.31 | 0.247 | 5.3 | $1.13 \times 10^{-7}$ *** |
| Human samples | 0.5 | 0.133 | 3.75 | $1.77 \times 10^{-4}$ *** |
| Human habitat | 0.329 | 0.269 | 1.22 | 0.221 |
| Freshwater | -0.541 | 0.287 | -1.88 | $5.96 \times 10^{-2}$ * |
| Clinical samples | -1.25 | 0.133 | -9.34 | 0 *** |
| Sludge, waste | -0.25 | 0.419 | -0.597 | 0.551 |
| Soil | -1.56 | 0.2 | -7.81 | $5.55 \times 10^{-15}$ *** |
| Trade | -0.131 | $8.09 \times 10^{-2}$ | -1.62 | 0.105 |
| Migration | -0.679 | 0.121 | -5.63 | $1.77 \times 10^{-8}$ *** |
| Domestic animals × Trade | -0.293 | 0.326 | -0.897 | 0.37 |
| Farms × Trade | -0.531 | $8.33 \times 10^{-2}$ | -6.38 | $1.73 \times 10^{-10}$ *** |
| Flora, fauna × Trade | -2.51 | 0.492 | -5.09 | $3.5 \times 10^{-7}$ *** |
| Human samples × Trade | -0.981 | $9.27 \times 10^{-2}$ | -10.6 | 0 *** |
| Human habitat × Trade | -0.654 | 0.227 | -2.87 | $4.04 \times 10^{-3}$ ** |
| Freshwater × Trade | -0.783 | 0.131 | -5.96 | $2.6 \times 10^{-9}$ *** |
| Clinical samples × Trade | -0.224 | $8.33 \times 10^{-2}$ | -2.68 | $7.3 \times 10^{-3}$ ** |
| Sludge, waste × Trade | -0.622 | 0.415 | -1.5 | 0.134 |
| Soil × Trade | 0.19 | 0.128 | 1.48 | 0.14 |
| Domestic animals × Migration | 0.102 | 0.342 | 0.3 | 0.764 |
| Farms × Migration | 0.179 | 0.116 | 1.55 | 0.121 |
| Flora, fauna × Migration | 0.26 | 0.25 | 1.04 | 0.299 |
| Human samples × Migration | 0.867 | 0.116 | 7.47 | $8.17 \times 10^{-14}$ *** |
| Human habitat × Migration | 0.833 | 0.222 | 3.76 | $1.72 \times 10^{-4}$ *** |
| Freshwater × Migration | -3.19 | 0.416 | -7.68 | $1.63 \times 10^{-14}$ *** |
| Clinical samples × Migration | 0.362 | 0.114 | 3.18 | $1.49 \times 10^{-3}$ ** |
| Sludge, waste × Migration | 0.692 | 0.251 | 2.76 | $5.72 \times 10^{-3}$ ** |
| Soil × Migration | -0.409 | 0.239 | -1.71 | $8.77 \times 10^{-2}$ * |

**Table S26: Summary of the selected model for CHG 7 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = -0.599, ν = 16.7, ρ = 1.51

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -4.9 | 0.437 | -11.2 | 0 *** |
| Domestic animals | 0.697 | 0.575 | 1.21 | 0.225 |
| Farms | 0.253 | 0.308 | 0.822 | 0.411 |
| Flora, fauna | 1.23 | 0.487 | 2.52 | $1.17 \times 10^{-2}$ * |
| Human samples | 0.697 | 0.292 | 2.38 | $1.71 \times 10^{-2}$ * |
| Human habitat | 0.398 | 0.501 | 0.796 | 0.426 |
| Freshwater | -0.776 | 0.592 | -1.31 | 0.19 |
| Clinical samples | 0.785 | 0.278 | 2.83 | $4.71 \times 10^{-3}$ ** |
| Sludge, waste | -1.03 | 1.6 | -0.642 | 0.521 |
| Soil | 0.321 | 0.451 | 0.711 | 0.477 |
| Trade | -0.111 | 0.257 | -0.434 | 0.664 |
| Migration | 0.182 | 0.145 | 1.25 | 0.21 |
| Domestic animals × Trade | $6.37 \times 10^{-2}$ | 0.43 | 0.148 | 0.882 |
| Farms × Trade | 0.229 | 0.267 | 0.857 | 0.391 |
| Flora, fauna × Trade | -1.88 | 1.16 | -1.62 | 0.105 |
| Human samples × Trade | 0.554 | 0.258 | 2.15 | $3.17 \times 10^{-2}$ * |
| Human habitat × Trade | -0.434 | 0.756 | -0.575 | 0.565 |
| Freshwater × Trade | -0.488 | 0.884 | -0.552 | 0.581 |
| Clinical samples × Trade | $-2.93 \times 10^{-3}$ | 0.259 | $-1.13 \times 10^{-2}$ | 0.991 |
| Sludge, waste × Trade | -3.69 | 4.33 | -0.852 | 0.394 |
| Soil × Trade | 0.593 | 0.287 | 2.06 | $3.9 \times 10^{-2}$ * |
| Domestic animals × Migration | -0.343 | 0.268 | -1.28 | 0.2 |
| Farms × Migration | -0.714 | 0.188 | -3.79 | $1.5 \times 10^{-4}$ *** |
| Flora, fauna × Migration | -0.519 | 0.578 | -0.897 | 0.37 |
| Human samples × Migration | -0.383 | 0.154 | -2.49 | $1.28 \times 10^{-2}$ * |
| Human habitat × Migration | 0.946 | 0.479 | 1.98 | $4.83 \times 10^{-2}$ * |
| Freshwater × Migration | -0.174 | 0.657 | -0.265 | 0.791 |
| Clinical samples × Migration | -0.376 | 0.147 | -2.55 | $1.09 \times 10^{-2}$ * |
| Sludge, waste × Migration | $1.43 \times 10^{-2}$ | 1.77 | $8.05 \times 10^{-3}$ | 0.994 |
| Soil × Migration | -1.93 | 0.747 | -2.58 | $9.84 \times 10^{-3}$ ** |

**Table S27: Summary of the selected model for CHG 8 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.28, ν = 16.7, ρ = 0.456

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -5.16 | 0.519 | -9.94 | 0 *** |
| Domestic animals | 0.615 | 0.519 | 1.18 | 0.236 |
| Farms | 0.941 | 0.274 | 3.43 | $5.98 \times 10^{-4}$ *** |
| Flora, fauna | 0.265 | 0.538 | 0.493 | 0.622 |
| Human samples | $-2.76 \times 10^{-2}$ | 0.28 | $-9.88 \times 10^{-2}$ | 0.921 |
| Human habitat | 0.471 | 0.396 | 1.19 | 0.234 |
| Freshwater | 0.299 | 0.569 | 0.526 | 0.599 |
| Clinical samples | 0.303 | 0.264 | 1.15 | 0.251 |
| Sludge, waste | -3.42 | 2.59 | -1.32 | 0.186 |
| Soil | 0.759 | 0.332 | 2.29 | $2.22 \times 10^{-2}$ * |
| Trade | $-4.04 \times 10^{-3}$ | 0.223 | $-1.81 \times 10^{-2}$ | 0.986 |
| Migration | -0.106 | 0.221 | -0.48 | 0.631 |
| Domestic animals × Trade | -0.799 | 0.616 | -1.3 | 0.195 |
| Farms × Trade | -0.482 | 0.24 | -2.01 | $4.45 \times 10^{-2}$ * |
| Flora, fauna × Trade | $3.12 \times 10^{-2}$ | 0.503 | $6.21 \times 10^{-2}$ | 0.95 |
| Human samples × Trade | 0.299 | 0.227 | 1.32 | 0.188 |
| Human habitat × Trade | 0.634 | 0.275 | 2.31 | $2.12 \times 10^{-2}$ * |
| Freshwater × Trade | -0.781 | 0.401 | -1.95 | $5.14 \times 10^{-2}$ * |
| Clinical samples × Trade | -0.201 | 0.225 | -0.895 | 0.371 |
| Sludge, waste × Trade | 0.237 | 1.39 | 0.171 | 0.864 |
| Soil × Trade | $2.5 \times 10^{-2}$ | 0.277 | $9.03 \times 10^{-2}$ | 0.928 |
| Domestic animals × Migration | -0.199 | 0.417 | -0.478 | 0.633 |
| Farms × Migration | -0.929 | 0.251 | -3.7 | $2.17 \times 10^{-4}$ *** |
| Flora, fauna × Migration | $-6.91 \times 10^{-2}$ | 0.428 | -0.161 | 0.872 |
| Human samples × Migration | $-5.02 \times 10^{-2}$ | 0.223 | -0.225 | 0.822 |
| Human habitat × Migration | -0.292 | 0.278 | -1.05 | 0.293 |
| Freshwater × Migration | -2.76 | 1.03 | -2.68 | $7.42 \times 10^{-3}$ ** |
| Clinical samples × Migration | -0.371 | 0.22 | -1.69 | $9.12 \times 10^{-2}$ * |
| Sludge, waste × Migration | 0.144 | 0.992 | 0.145 | 0.885 |
| Soil × Migration | -0.265 | 0.285 | -0.928 | 0.353 |

**Table S28: Summary of the selected model for CHG 11 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.671, ν = 16.7, ρ = 0.345

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -10.1 | 2.03 | -4.96 | $7.17 \times 10^{-7}$ *** |
| Domestic animals | -2.74 | 4.7 | -0.581 | 0.561 |
| Farms | 2.9 | 1.84 | 1.57 | 0.116 |
| Flora, fauna | 1.18 | 2.48 | 0.477 | 0.633 |
| Human samples | 3.56 | 1.84 | 1.94 | $5.27 \times 10^{-2}$ * |
| Human habitat | -16.9 | 17.1 | -0.99 | 0.322 |
| Freshwater | 1.77 | 1.95 | 0.908 | 0.364 |
| Clinical samples | 3.51 | 1.84 | 1.91 | $5.62 \times 10^{-2}$ * |
| Sludge, waste | -28 | $5.67 \times 10^{6}$ | $-4.93 \times 10^{-6}$ | 1 |
| Soil | 2.39 | 1.9 | 1.26 | 0.208 |
| Trade | -0.168 | 0.4 | -0.42 | 0.675 |
| Migration | -4.12 | 4.49 | -0.918 | 0.359 |
| Domestic animals × Trade | -7.07 | 5.47 | -1.29 | 0.196 |
| Farms × Trade | -0.644 | 0.429 | -1.5 | 0.133 |
| Flora, fauna × Trade | 0.665 | 0.764 | 0.871 | 0.384 |
| Human samples × Trade | 0.281 | 0.4 | 0.702 | 0.482 |
| Human habitat × Trade | -2.31 | 5.16 | -0.448 | 0.654 |
| Freshwater × Trade | -0.69 | 0.736 | -0.937 | 0.349 |
| Clinical samples × Trade | $-8.78 \times 10^{-2}$ | 0.399 | -0.22 | 0.826 |
| Sludge, waste × Trade | -0.4 | $5.77 \times 10^{6}$ | $-6.94 \times 10^{-8}$ | 1 |
| Soil × Trade | $7.43 \times 10^{-2}$ | 0.453 | 0.164 | 0.87 |
| Domestic animals × Migration | -5.16 | 12.2 | -0.422 | 0.673 |
| Farms × Migration | 4.43 | 4.49 | 0.986 | 0.324 |
| Flora, fauna × Migration | 2.4 | 5.45 | 0.441 | 0.659 |
| Human samples × Migration | 4.18 | 4.49 | 0.93 | 0.352 |
| Human habitat × Migration | -35.7 | 35.5 | -1.01 | 0.315 |
| Freshwater × Migration | 4.95 | 4.52 | 1.09 | 0.274 |
| Clinical samples × Migration | 3.65 | 4.49 | 0.814 | 0.416 |
| Sludge, waste × Migration | 4.59 | $7.67 \times 10^{6}$ | $5.98 \times 10^{-7}$ | 1 |
| Soil × Migration | 3.66 | 4.58 | 0.8 | 0.424 |

**Table S29: Summary of the selected model for CHG 13 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 0.914$, $\nu = 6.33 \times 10^{-3}$, $\rho = 0.303$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -24.5 | $3.48 \times 10^{3}$ | $-7.03 \times 10^{-3}$ | 0.994 |
| Domestic animals | $-2.15 \times 10^{2}$ | $2.63 \times 10^{6}$ | $-8.16 \times 10^{-5}$ | 1 |
| Farms | 17.3 | $3.48 \times 10^{3}$ | $4.97 \times 10^{-3}$ | 0.996 |
| Flora, fauna | 17.1 | $3.48 \times 10^{3}$ | $4.91 \times 10^{-3}$ | 0.996 |
| Human samples | 18.4 | $3.48 \times 10^{3}$ | $5.29 \times 10^{-3}$ | 0.996 |
| Human habitat | 3.44 | $3.48 \times 10^{3}$ | $9.88 \times 10^{-4}$ | 0.999 |
| Freshwater | $-2.74 \times 10^{2}$ | $2.43 \times 10^{6}$ | $-1.13 \times 10^{-4}$ | 1 |
| Clinical samples | 18.6 | $3.48 \times 10^{3}$ | $5.34 \times 10^{-3}$ | 0.996 |
| Sludge, waste | 19.9 | $3.48 \times 10^{3}$ | $5.73 \times 10^{-3}$ | 0.995 |
| Soil | 7.25 | $3.48 \times 10^{3}$ | $2.08 \times 10^{-3}$ | 0.998 |
| Trade | -0.847 | $3.85 \times 10^{3}$ | $-2.2 \times 10^{-4}$ | 1 |
| Migration | -0.154 | $2.48 \times 10^{3}$ | $-6.2 \times 10^{-5}$ | 1 |
| Domestic animals × Trade | $-3.48 \times 10^{2}$ | $2.26 \times 10^{6}$ | $-1.54 \times 10^{-4}$ | 1 |
| Farms × Trade | -0.535 | $3.85 \times 10^{3}$ | $-1.39 \times 10^{-4}$ | 1 |
| Flora, fauna × Trade | -1.36 | $3.85 \times 10^{3}$ | $-3.54 \times 10^{-4}$ | 1 |
| Human samples × Trade | 0.598 | $3.85 \times 10^{3}$ | $1.55 \times 10^{-4}$ | 1 |
| Human habitat × Trade | 0.8 | $3.85 \times 10^{3}$ | $2.08 \times 10^{-4}$ | 1 |
| Freshwater × Trade | 4.81 | $2.59 \times 10^{6}$ | $1.86 \times 10^{-6}$ | 1 |
| Clinical samples × Trade | $9.19 \times 10^{-2}$ | $3.85 \times 10^{3}$ | $2.39 \times 10^{-5}$ | 1 |
| Sludge, waste × Trade | -0.213 | $3.85 \times 10^{3}$ | $-5.53 \times 10^{-5}$ | 1 |
| Soil × Trade | 1.3 | $3.85 \times 10^{3}$ | $3.37 \times 10^{-4}$ | 1 |
| Domestic animals × Migration | $1.36 \times 10^{2}$ | $1.69 \times 10^{6}$ | $8.07 \times 10^{-5}$ | 1 |
| Farms × Migration | -3.3 | $2.48 \times 10^{3}$ | $-1.33 \times 10^{-3}$ | 0.999 |
| Flora, fauna × Migration | -3.74 | $2.48 \times 10^{3}$ | $-1.51 \times 10^{-3}$ | 0.999 |
| Human samples × Migration | 0.108 | $2.48 \times 10^{3}$ | $4.36 \times 10^{-5}$ | 1 |
| Human habitat × Migration | -48.6 | $2.48 \times 10^{3}$ | $-1.96 \times 10^{-2}$ | 0.984 |
| Freshwater × Migration | 1.16 | $2.65 \times 10^{6}$ | $4.39 \times 10^{-7}$ | 1 |
| Clinical samples × Migration | 0.192 | $2.48 \times 10^{3}$ | $7.71 \times 10^{-5}$ | 1 |
| Sludge, waste × Migration | -0.161 | $2.48 \times 10^{3}$ | $-6.47 \times 10^{-5}$ | 1 |
| Soil × Migration | -31.6 | $2.48 \times 10^{3}$ | $-1.27 \times 10^{-2}$ | 0.99 |

**Table S30: Summary of the selected model for CHG 14 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.294, ν = 16.7, ρ = 0.922

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -8.94 | 1.92 | -4.66 | $3.23 \times 10^{-6}$ *** |
| Domestic animals | -29 | $4.39 \times 10^{6}$ | $-6.6 \times 10^{-6}$ | 1 |
| Farms | 2.93 | 1.87 | 1.57 | 0.117 |
| Flora, fauna | 0.685 | 3.3 | 0.207 | 0.836 |
| Human samples | -5.35 | 3.29 | -1.63 | 0.104 |
| Human habitat | -28.4 | $3.88 \times 10^{6}$ | $-7.33 \times 10^{-6}$ | 1 |
| Freshwater | -28.9 | $2.55 \times 10^{6}$ | $-1.13 \times 10^{-5}$ | 1 |
| Clinical samples | 1.11 | 1.87 | 0.591 | 0.555 |
| Sludge, waste | -29.2 | $5.48 \times 10^{6}$ | $-5.32 \times 10^{-6}$ | 1 |
| Soil | -28.4 | $2.13 \times 10^{6}$ | $-1.34 \times 10^{-5}$ | 1 |
| Trade | -0.984 | 1.55 | -0.635 | 0.526 |
| Migration | -1.73 | 6.3 | -0.275 | 0.783 |
| Domestic animals × Trade | 0.468 | $2.76 \times 10^{6}$ | $1.69 \times 10^{-7}$ | 1 |
| Farms × Trade | 0.83 | 1.55 | 0.537 | 0.591 |
| Flora, fauna × Trade | -0.228 | 3.42 | $-6.67 \times 10^{-2}$ | 0.947 |
| Human samples × Trade | 1.27 | 1.57 | 0.806 | 0.42 |
| Human habitat × Trade | 0.426 | $2.01 \times 10^{6}$ | $2.12 \times 10^{-7}$ | 1 |
| Freshwater × Trade | 0.504 | $1.44 \times 10^{6}$ | $3.5 \times 10^{-7}$ | 1 |
| Clinical samples × Trade | 0.269 | 1.56 | 0.173 | 0.862 |
| Sludge, waste × Trade | 0.452 | $2.4 \times 10^{6}$ | $1.88 \times 10^{-7}$ | 1 |
| Soil × Trade | 0.197 | $2 \times 10^{6}$ | $9.86 \times 10^{-8}$ | 1 |
| Domestic animals × Migration | 1.63 | $2.67 \times 10^{6}$ | $6.11 \times 10^{-7}$ | 1 |
| Farms × Migration | 1.76 | 6.3 | 0.28 | 0.78 |
| Flora, fauna × Migration | -0.181 | 12 | $-1.5 \times 10^{-2}$ | 0.988 |
| Human samples × Migration | -21 | 10.3 | -2.05 | $4.08 \times 10^{-2}$ * |
| Human habitat × Migration | 1.38 | $5.88 \times 10^{6}$ | $2.34 \times 10^{-7}$ | 1 |
| Freshwater × Migration | 1.36 | $3.21 \times 10^{6}$ | $4.25 \times 10^{-7}$ | 1 |
| Clinical samples × Migration | 1.4 | 6.3 | 0.223 | 0.824 |
| Sludge, waste × Migration | 1.65 | $4.24 \times 10^{6}$ | $3.9 \times 10^{-7}$ | 1 |
| Soil × Migration | 1.62 | $3.35 \times 10^{6}$ | $4.83 \times 10^{-7}$ | 1 |

**Table S31: Summary of the selected model for CHG 15 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 2.17 \times 10^{-2}$, $\nu = 16.7$, $\rho = 0.724$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -5.96 | 0.603 | -9.88 | 0 *** |
| Domestic animals | -31.3 | $3.91 \times 10^{6}$ | $-8.01 \times 10^{-6}$ | 1 |
| Farms | -1.63 | 0.648 | -2.51 | $1.2 \times 10^{-2}$ * |
| Flora, fauna | 0.809 | 0.701 | 1.15 | 0.248 |
| Human samples | -31.3 | $8.32 \times 10^{5}$ | $-3.76 \times 10^{-5}$ | 1 |
| Human habitat | $-9.06 \times 10^{-2}$ | 0.902 | -0.1 | 0.92 |
| Freshwater | -31.2 | $2.34 \times 10^{6}$ | $-1.33 \times 10^{-5}$ | 1 |
| Clinical samples | -4.82 | 1.15 | -4.19 | $2.74 \times 10^{-5}$ *** |
| Sludge, waste | -31.1 | $3.57 \times 10^{6}$ | $-8.7 \times 10^{-6}$ | 1 |
| Soil | 1.26 | 0.528 | 2.39 | $1.71 \times 10^{-2}$ * |
| Trade | -0.221 | 0.229 | -0.968 | 0.333 |
| Migration | -0.24 | 0.18 | -1.34 | 0.181 |

**Table S32: Summary of the selected model for CHG 17.1 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 0.704$, $\nu = 5.7 \times 10^{-3}$, $\rho = 0.178$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -36.9 | $3.13 \times 10^{6}$ | $-1.18 \times 10^{-5}$ | 1 |
| Domestic animals | -7.7 | $5 \times 10^{6}$ | $-1.54 \times 10^{-6}$ | 1 |
| Farms | -7.01 | $3.25 \times 10^{6}$ | $-2.16 \times 10^{-6}$ | 1 |
| Flora, fauna | -7.1 | $4.95 \times 10^{6}$ | $-1.43 \times 10^{-6}$ | 1 |
| Human samples | 25.1 | $3.13 \times 10^{6}$ | $8.03 \times 10^{-6}$ | 1 |
| Human habitat | 29.3 | $3.13 \times 10^{6}$ | $9.37 \times 10^{-6}$ | 1 |
| Freshwater | 27.9 | $3.13 \times 10^{6}$ | $8.93 \times 10^{-6}$ | 1 |
| Clinical samples | 27.7 | $3.13 \times 10^{6}$ | $8.87 \times 10^{-6}$ | 1 |
| Sludge, waste | 27.7 | $3.13 \times 10^{6}$ | $8.87 \times 10^{-6}$ | 1 |
| Soil | -6.69 | $3.68 \times 10^{6}$ | $-1.82 \times 10^{-6}$ | 1 |
| Trade | 0.163 | 0.134 | 1.22 | 0.224 |
| Migration | -2.89 | 0.862 | -3.35 | $8.09 \times 10^{-4}$ *** |

**Table S33: Summary of the selected model for CHG 19 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 0.931$, $\nu = 1.88 \times 10^{-2}$, $\rho = 1.89$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -15.4 | 10.6 | -1.45 | 0.148 |
| Domestic animals | 8.65 | 10.6 | 0.816 | 0.414 |
| Farms | 9.64 | 10.6 | 0.912 | 0.362 |
| Flora, fauna | 10.6 | 10.6 | 1 | 0.316 |
| Human samples | 9.07 | 10.6 | 0.858 | 0.391 |
| Human habitat | 7.88 | 10.6 | 0.741 | 0.458 |
| Freshwater | -23.1 | $2.68 \times 10^{6}$ | $-8.61 \times 10^{-6}$ | 1 |
| Clinical samples | 10.2 | 10.6 | 0.965 | 0.335 |
| Sludge, waste | 10.5 | 10.6 | 0.989 | 0.323 |
| Soil | 0.975 | 12.8 | $7.62 \times 10^{-2}$ | 0.939 |
| Trade | 1.21 | 2.44 | 0.495 | 0.621 |
| Migration | 0.582 | 0.364 | 1.6 | 0.11 |
| Domestic animals × Trade | -0.813 | 2.47 | -0.329 | 0.742 |
| Farms × Trade | -1.61 | 2.45 | -0.66 | 0.509 |
| Flora, fauna × Trade | -1.68 | 2.47 | -0.679 | 0.497 |
| Human samples × Trade | -1.27 | 2.44 | -0.519 | 0.604 |
| Human habitat × Trade | -0.933 | 2.52 | -0.371 | 0.711 |
| Freshwater × Trade | -1.93 | $2.74 \times 10^{6}$ | $-7.06 \times 10^{-7}$ | 1 |
| Clinical samples × Trade | -1.42 | 2.44 | -0.58 | 0.562 |
| Sludge, waste × Trade | -1.26 | 2.45 | -0.514 | 0.607 |
| Soil × Trade | -4 | 4.5 | -0.889 | 0.374 |
| Domestic animals × Migration | -0.717 | 0.417 | -1.72 | $8.57 \times 10^{-2}$ * |
| Farms × Migration | -1.27 | 0.409 | -3.11 | $1.86 \times 10^{-3}$ ** |
| Flora, fauna × Migration | -1.58 | 0.614 | -2.57 | $1.02 \times 10^{-2}$ * |
| Human samples × Migration | -0.499 | 0.374 | -1.34 | 0.182 |
| Human habitat × Migration | -0.496 | 0.601 | -0.826 | 0.409 |
| Freshwater × Migration | -0.385 | $2.94 \times 10^{6}$ | $-1.31 \times 10^{-7}$ | 1 |
| Clinical samples × Migration | -0.702 | 0.365 | -1.93 | $5.41 \times 10^{-2}$ * |
| Sludge, waste × Migration | -0.682 | 0.411 | -1.66 | $9.68 \times 10^{-2}$ * |
| Soil × Migration | -12.4 | 12.9 | -0.96 | 0.337 |

**Table S34: Summary of the selected model for CHG 20 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.283, ν = 16.7, ρ = 0.823

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -11 | 2.33 | -4.69 | $2.67 \times 10^{-6}$ *** |
| Domestic animals | -26.9 | $4.38 \times 10^{6}$ | $-6.14 \times 10^{-6}$ | 1 |
| Farms | 4.94 | 2.27 | 2.17 | $2.96 \times 10^{-2}$ * |
| Flora, fauna | 2.49 | 4.18 | 0.595 | 0.552 |
| Human samples | $7.31 \times 10^{-2}$ | 2.67 | $2.74 \times 10^{-2}$ | 0.978 |
| Human habitat | -26.4 | $3.88 \times 10^{6}$ | $-6.8 \times 10^{-6}$ | 1 |
| Freshwater | -26.9 | $2.53 \times 10^{6}$ | $-1.06 \times 10^{-5}$ | 1 |
| Clinical samples | 3.34 | 2.28 | 1.46 | 0.144 |
| Sludge, waste | -27.1 | $5.52 \times 10^{6}$ | $-4.9 \times 10^{-6}$ | 1 |
| Soil | -26.6 | $2.1 \times 10^{6}$ | $-1.27 \times 10^{-5}$ | 1 |
| Trade | 0.289 | 0.493 | 0.586 | 0.558 |
| Migration | -10.8 | 7.07 | -1.52 | 0.128 |
| Domestic animals × Trade | -0.844 | $2.89 \times 10^{6}$ | $-2.92 \times 10^{-7}$ | 1 |
| Farms × Trade | -0.413 | 0.492 | -0.84 | 0.401 |
| Flora, fauna × Trade | -1.65 | 3.5 | -0.471 | 0.638 |
| Human samples × Trade | -0.289 | 0.533 | -0.543 | 0.587 |
| Human habitat × Trade | -0.84 | $2.03 \times 10^{6}$ | $-4.14 \times 10^{-7}$ | 1 |
| Freshwater × Trade | -0.772 | $1.45 \times 10^{6}$ | $-5.33 \times 10^{-7}$ | 1 |
| Clinical samples × Trade | -1.03 | 0.531 | -1.93 | $5.31 \times 10^{-2}$ * |
| Sludge, waste × Trade | -0.85 | $2.45 \times 10^{6}$ | $-3.47 \times 10^{-7}$ | 1 |
| Soil × Trade | -1.06 | $2.08 \times 10^{6}$ | $-5.1 \times 10^{-7}$ | 1 |
| Domestic animals × Migration | 10.6 | $2.74 \times 10^{6}$ | $3.86 \times 10^{-6}$ | 1 |
| Farms × Migration | 10.8 | 7.07 | 1.53 | 0.127 |
| Flora, fauna × Migration | 7.83 | 14.1 | 0.556 | 0.578 |
| Human samples × Migration | -1.99 | 8.09 | -0.246 | 0.806 |
| Human habitat × Migration | 10.2 | $5.3 \times 10^{6}$ | $1.92 \times 10^{-6}$ | 1 |
| Freshwater × Migration | 10.2 | $3.1 \times 10^{6}$ | $3.28 \times 10^{-6}$ | 1 |
| Clinical samples × Migration | 10.2 | 7.07 | 1.45 | 0.148 |
| Sludge, waste × Migration | 10.5 | $4.13 \times 10^{6}$ | $2.55 \times 10^{-6}$ | 1 |
| Soil × Migration | 10.5 | $3.27 \times 10^{6}$ | $3.2 \times 10^{-6}$ | 1 |

**Table S35: Summary of the selected model for CHG 23 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

φ = 0.521, ν = 1.22, ρ = 0.534

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -25.7 | $4.1 \times 10^{3}$ | $-6.27 \times 10^{-3}$ | 0.995 |
| Domestic animals | $-4.01 \times 10^{2}$ | $4.8 \times 10^{6}$ | $-8.36 \times 10^{-5}$ | 1 |
| Farms | $-4.01 \times 10^{2}$ | $1.09 \times 10^{6}$ | $-3.66 \times 10^{-4}$ | 1 |
| Flora, fauna | $-4 \times 10^{2}$ | $4.13 \times 10^{6}$ | $-9.69 \times 10^{-5}$ | 1 |
| Human samples | 18.6 | $4.1 \times 10^{3}$ | $4.54 \times 10^{-3}$ | 0.996 |
| Human habitat | -29.7 | $4.1 \times 10^{3}$ | $-7.23 \times 10^{-3}$ | 0.994 |
| Freshwater | 17 | $4.1 \times 10^{3}$ | $4.13 \times 10^{-3}$ | 0.997 |
| Clinical samples | 19.8 | $4.1 \times 10^{3}$ | $4.84 \times 10^{-3}$ | 0.996 |
| Sludge, waste | $-2.03 \times 10^{2}$ | $6.25 \times 10^{4}$ | $-3.24 \times 10^{-3}$ | 0.997 |
| Soil | 12.9 | $4.1 \times 10^{3}$ | $3.13 \times 10^{-3}$ | 0.997 |
| Trade | -0.717 | $3.22 \times 10^{3}$ | $-2.23 \times 10^{-4}$ | 1 |
| Migration | 0.102 | $2.9 \times 10^{3}$ | $3.53 \times 10^{-5}$ | 1 |
| Domestic animals × Trade | 4.19 | $3.26 \times 10^{6}$ | $1.29 \times 10^{-6}$ | 1 |
| Farms × Trade | 2.46 | $5.65 \times 10^{5}$ | $4.35 \times 10^{-6}$ | 1 |
| Flora, fauna × Trade | 1.5 | $2.35 \times 10^{6}$ | $6.38 \times 10^{-7}$ | 1 |
| Human samples × Trade | 0.91 | $3.22 \times 10^{3}$ | $2.83 \times 10^{-4}$ | 1 |
| Human habitat × Trade | 2.83 | $3.22 \times 10^{3}$ | $8.79 \times 10^{-4}$ | 0.999 |
| Freshwater × Trade | -0.351 | $3.22 \times 10^{3}$ | $-1.09 \times 10^{-4}$ | 1 |
| Clinical samples × Trade | 0.423 | $3.22 \times 10^{3}$ | $1.31 \times 10^{-4}$ | 1 |
| Sludge, waste × Trade | $-2.58 \times 10^{2}$ | $1.57 \times 10^{5}$ | $-1.64 \times 10^{-3}$ | 0.999 |
| Soil × Trade | -6.99 | $3.22 \times 10^{3}$ | $-2.17 \times 10^{-3}$ | 0.998 |
| Domestic animals × Migration | -0.821 | $2.08 \times 10^{6}$ | $-3.95 \times 10^{-7}$ | 1 |
| Farms × Migration | -0.64 | $3.4 \times 10^{5}$ | $-1.88 \times 10^{-6}$ | 1 |
| Flora, fauna × Migration | $6.75 \times 10^{-2}$ | $2.53 \times 10^{6}$ | $2.67 \times 10^{-8}$ | 1 |
| Human samples × Migration | -2.87 | $2.9 \times 10^{3}$ | $-9.9 \times 10^{-4}$ | 0.999 |
| Human habitat × Migration | $-1.63 \times 10^{2}$ | $2.91 \times 10^{3}$ | $-5.62 \times 10^{-2}$ | 0.955 |
| Freshwater × Migration | -1.54 | $2.9 \times 10^{3}$ | $-5.29 \times 10^{-4}$ | 1 |
| Clinical samples × Migration | -0.195 | $2.9 \times 10^{3}$ | $-6.71 \times 10^{-5}$ | 1 |
| Sludge, waste × Migration | $-1.94 \times 10^{2}$ | $3.11 \times 10^{5}$ | $-6.24 \times 10^{-4}$ | 1 |
| Soil × Migration | -5.92 | $2.9 \times 10^{3}$ | $-2.04 \times 10^{-3}$ | 0.998 |

**Table S36: Summary of the selected model for CHG 25 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = -0.469$, $\nu = 5.01 \times 10^{-3}$, $\rho = 2.54 \times 10^{-2}$

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -39.6 | 25.7 | -1.54 | 0.123 |
| Trade | -28.9 | 36.1 | -0.801 | 0.423 |
| Migration | $-1.23 \times 10^{2}$ | $1.43 \times 10^{2}$ | -0.856 | 0.392 |

**Table S37: Summary of the selected model for CHG 27 for worldwide data.** φ correspond to the estimated time autoregression coefficient. ν and ρ are respectively the estimated smoothness and scaling parameters of the Matérn correlation function. *: p<0.05, **: p<0.01, ***: p<0.001.

$\varphi = 1.14 \times 10^{-2}$, $\nu = 8.21 \times 10^{-2}$, $\rho = 1.29 \times 10^{-2}$

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -4.61 | 0.549 | -8.39 | 0 *** |
| Domestic animals | -32.5 | $4.54 \times 10^{6}$ | $-7.15 \times 10^{-6}$ | 1 |
| Farms | -2.1 | 0.509 | -4.12 | $3.76 \times 10^{-5}$ *** |
| Flora, fauna | -0.698 | 0.85 | -0.821 | 0.412 |
| Human samples | -1.45 | 0.317 | -4.58 | $4.66 \times 10^{-6}$ *** |
| Human habitat | -3.63 | 3.63 | -0.999 | 0.318 |
| Freshwater | -0.198 | 0.367 | -0.541 | 0.589 |
| Clinical samples | -0.395 | 0.209 | -1.89 | $5.84 \times 10^{-2}$ * |
| Sludge, waste | -0.156 | 0.659 | -0.236 | 0.813 |
| Soil | -0.437 | 0.389 | -1.12 | 0.261 |
| Trade | $-5.93 \times 10^{-2}$ | $8.65 \times 10^{-2}$ | -0.686 | 0.493 |
| Migration | 0.651 | $8.8 \times 10^{-2}$ | 7.39 | $1.45 \times 10^{-13}$ *** |
| Domestic animals × Trade | -0.464 | $3.5 \times 10^{6}$ | $-1.33 \times 10^{-7}$ | 1 |
| Farms × Trade | $6.4 \times 10^{-2}$ | 0.119 | 0.539 | 0.59 |
| Flora, fauna × Trade | -0.198 | 0.447 | -0.444 | 0.657 |
| Human samples × Trade | -0.483 | 0.268 | -1.8 | $7.16 \times 10^{-2}$ * |
| Human habitat × Trade | $2.75 \times 10^{-2}$ | 0.319 | $8.63 \times 10^{-2}$ | 0.931 |
| Freshwater × Trade | -1.38 | 0.716 | -1.92 | $5.47 \times 10^{-2}$ * |
| Clinical samples × Trade | -1.09 | 0.175 | -6.26 | $3.79 \times 10^{-10}$ *** |
| Sludge, waste × Trade | -0.256 | 0.284 | -0.901 | 0.368 |
| Soil × Trade | -0.323 | 0.322 | -1 | 0.316 |
| Domestic animals × Migration | -0.421 | $3.61 \times 10^{6}$ | $-1.17 \times 10^{-7}$ | 1 |
| Farms × Migration | -2.99 | 1.08 | -2.76 | $5.82 \times 10^{-3}$ ** |
| Flora, fauna × Migration | -0.776 | 0.7 | -1.11 | 0.268 |
| Human samples × Migration | -1.02 | 0.312 | -3.27 | $1.08 \times 10^{-3}$ ** |
| Human habitat × Migration | -10.2 | 8.7 | -1.17 | 0.243 |
| Freshwater × Migration | -1.05 | 0.426 | -2.47 | $1.33 \times 10^{-2}$ * |
| Clinical samples × Migration | -0.883 | 0.148 | -5.95 | $2.65 \times 10^{-9}$ *** |
| Sludge, waste × Migration | -0.665 | 0.477 | -1.4 | 0.163 |
| Soil × Migration | $-7.66 \times 10^{-3}$ | 0.306 | $-2.51 \times 10^{-2}$ | 0.98 |

**Table S39: Impact of the association of AMEGs with mobile genetic elements with their duplication risk**

| Minimum identity between genomic contexts | Minimum identity between genes | Odds ratio | 500-permutation p-value |
|---|---|---|---|
| 50% | 80% | 14.2 | 0.001 |
| 50% | 85% | 11.9 | 0.001 |
| 50% | 90% | 10.7 | 0.001 |
| 50% | 95% | 10.4 | 0.001 |
| 70% | 80% | 14.2 | 0.001 |
| 70% | 85% | 11.9 | 0.001 |
| 70% | 90% | 10.7 | 0.001 |
| 70% | 95% | 10.4 | 0.001 |
| 80% | 80% | 14.2 | 0.001 |
| 0.001 | 85% | 11.9 | 0.001 |
| 0.001 | 90% | 10.7 | 0.001 |
| 80% | 95% | 10.4 | 0.001 |
| 90% | 80% | 14.2 | 0.001 |
| 90% | 85% | 11.9 | 0.001 |
| 90% | 90% | 10.7 | 0.001 |
| 90% | 95% | 10.4 | 0.001 |

**Table S40: Impact of the association of AMEGs with different types of mobile genetic elements with their duplication risk**

| Minimum identity between genomic contexts | Minimum identity between genes | Type of mobile genetic element | Odds ratio | 500-permutation p-value |
|---|---|---|---|---|
| 50% | 80% | Intergenomic mobility | 18.5 | 0.001 |
| 50% | 85% | Intergenomic mobility | 14.7 | 0.001 |
| 50% | 90% | Intergenomic mobility | 13.7 | 0.001 |
| 50% | 95% | Intergenomic mobility | 13.3 | 0.001 |
| 70% | 80% | Intergenomic mobility | 18.5 | 0.001 |
| 70% | 85% | Intergenomic mobility | 14.7 | 0.001 |
| 70% | 90% | Intergenomic mobility | 13.7 | 0.001 |
| 70% | 95% | Intergenomic mobility | 13.3 | 0.001 |
| 80% | 80% | Intergenomic mobility | 18.5 | 0.001 |
| 80% | 85% | Intergenomic mobility | 14.7 | 0.001 |
| 80% | 90% | Intergenomic mobility | 13.7 | 0.001 |
| 80% | 95% | Intergenomic mobility | 13.3 | 0.001 |
| 90% | 80% | Intergenomic mobility | 18.5 | 0.001 |
| 90% | 85% | Intergenomic mobility | 14.7 | 0.001 |
| 90% | 90% | Intergenomic mobility | 13.7 | 0.001 |
| 90% | 95% | Intergenomic mobility | 13.3 | 0.001 |
| 50% | 80% | Intragenomic mobility | 12.0 | 0.001 |
| 50% | 85% | Intragenomic mobility | 12.0 | 0.002 |
| 50% | 90% | Intragenomic mobility | 12.0 | 0.002 |
| 50% | 95% | Intragenomic mobility | 10.8 | 0.003 |
| 70% | 80% | Intragenomic mobility | 12.0 | 0.001 |
| 70% | 85% | Intragenomic mobility | 12.0 | 0.002 |
| 70% | 90% | Intragenomic mobility | 12.0 | 0.002 |
| 70% | 95% | Intragenomic mobility | 10.8 | 0.003 |
| 80% | 80% | Intragenomic mobility | 12.0 | 0.001 |
| 80% | 85% | Intragenomic mobility | 12.0 | 0.002 |
| 80% | 90% | Intragenomic mobility | 12.0 | 0.002 |
| 80% | 95% | Intragenomic mobility | 10.8 | 0.003 |
| 90% | 80% | Intragenomic mobility | 12.0 | 0.001 |
| 90% | 85% | Intragenomic mobility | 12.0 | 0.002 |
| 90% | 90% | Intragenomic mobility | 12.0 | 0.002 |
| 90% | 95% | Intragenomic mobility | 10.8 | 0.003 |
| 50% | 80% | Intra- and intergenomic mobility | 16.0 | 0.001 |
| 50% | 85% | Intra- and intergenomic mobility | 14.1 | 0.001 |
| 50% | 90% | Intra- and intergenomic mobility | 12.1 | 0.001 |
| 50% | 95% | Intra- and intergenomic mobility | 12.1 | 0.001 |
| 70% | 80% | Intra- and intergenomic mobility | 16.0 | 0.001 |
| 70% | 85% | Intra- and intergenomic mobility | 14.1 | 0.001 |
| 70% | 90% | Intra- and intergenomic mobility | 12.1 | 0.001 |
| 70% | 95% | Intra- and intergenomic mobility | 12.1 | 0.001 |
| 80% | 80% | Intra- and intergenomic mobility | 16.0 | 0.001 |
| 80% | 85% | Intra- and intergenomic mobility | 14.1 | 0.001 |
| 80% | 90% | Intra- and intergenomic mobility | 12.1 | 0.001 |
| 80% | 95% | Intra- and intergenomic mobility | 12.1 | 0.001 |
| 90% | 80% | Intra- and intergenomic mobility | 16.0 | 0.001 |
| 90% | 85% | Intra- and intergenomic mobility | 14.1 | 0.001 |
| 90% | 90% | Intra- and intergenomic mobility | 12.1 | 0.001 |
| 90% | 95% | Intra- and intergenomic mobility | 12.1 | 0.001 |

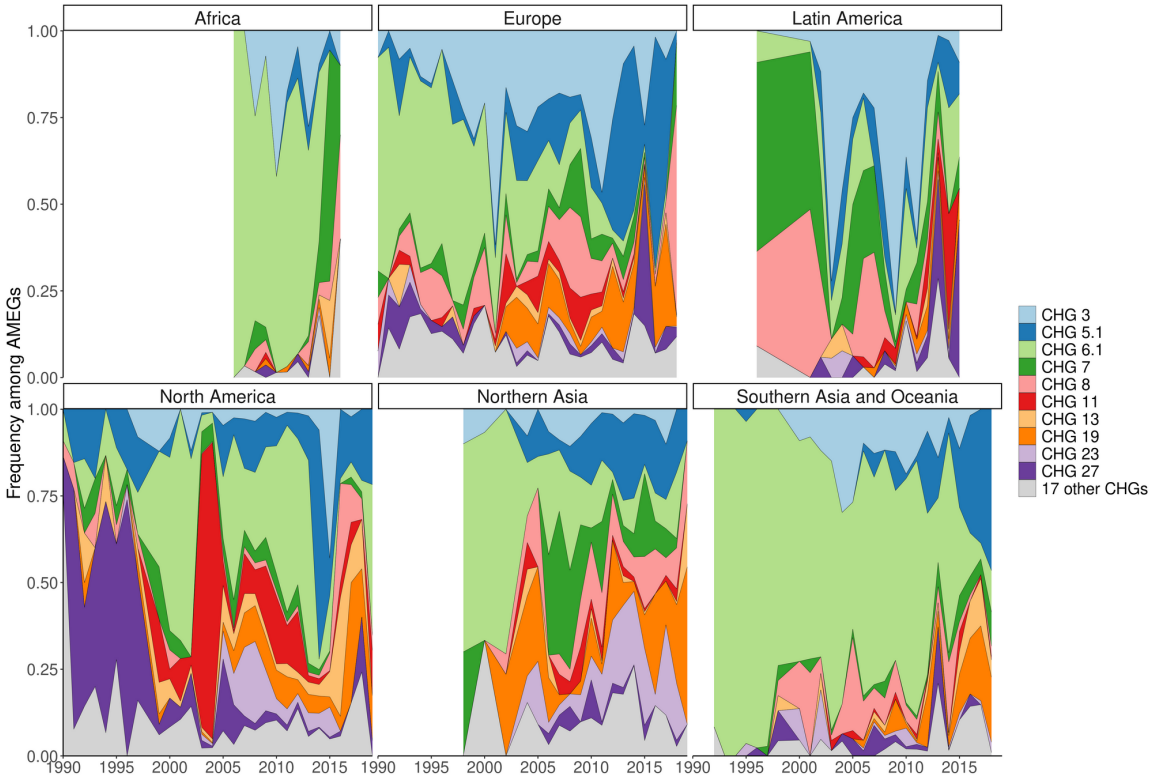


**Figure S1: Evolution of continental CHG frequencies among sampled AMEGs between 1990 and 2019.** IMAGE24 regions were gathered to higher continental ensembles in order to more representative frequencies. Western Europe, Central Europe, and "Ukraine+" were gathered to Europe; Canada and the USA to North America; the rest of Americas to Latin America; all African landmasses to Africa; Oceania, Southeastern Asia, "Indonesia+", and "India+" to Southern Asia and Oceania; and Korea, Japan, Middle East, Turkey, "China+", "Asia-Stan", and "Russia+" to Northern Asia. CHG frequencies were only computed when at least 10 ACBs had been sampled in a continent in a year.

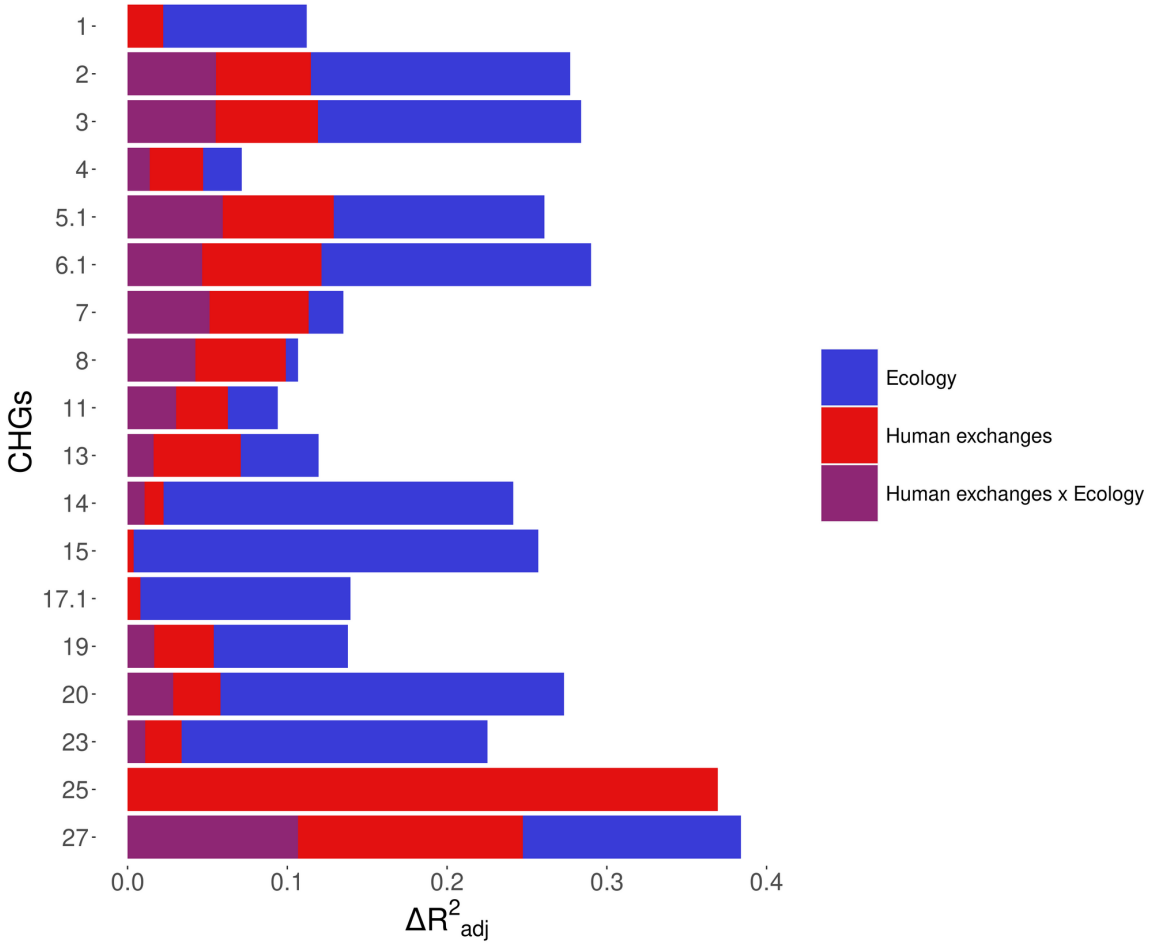


**Figure S2: Relative importance of several factors to explain the distribution of aminoglycoside-resistant bacteria in the world between 1997 and 2018.** Logistic regressions with a spatial Matérn correlation structure were computed to explain the frequency of 18 CHGs in samples. This figure represents the contribution of each variable in each selected model, as the fraction of adjusted McFadden's pseudo-R2 explained by adding this variable.

# 10 Annexe 2 : Informations supplémentaires de la section V

Les informations supplémentaires de la section V sont exposées dans la section suivante : Figure S1, et Tables S1–S27.

**Table S1: Classifications of pathogenicity for species included in HGT networks**

| Species | WHO | NIAID | PHIBASE | WARDEH |
|---|---|---|---|---|
| *Achromobacter denitrificans* | No | No | No | Yes |
| *Achromobacter ruhlandii* | No | No | No | No |
| *Achromobacter sp001299435* | No | No | No | No |
| *Achromobacter xylosoxidans* | No | No | No | Yes |
| *Acinetobacter baumannii* | Yes | No | Yes | Yes |
| *Acinetobacter nosocomialis* | No | No | No | No |
| *Acinetobacter pittii* | No | No | No | No |
| *Acinetobacter sp001696605* | No | No | No | No |
| *Aeromonas hydrophila* | No | No | No | Yes |
| *Aeromonas salmonicida* | No | No | No | Yes |
| *Alcaligenes faecalis* | No | No | No | No |
| *Bosea sp900112185* | No | No | No | No |
| *Brucella melitensis* | No | Yes | Yes | Yes |
| *Burkholderia cenocepacia* | No | No | No | Yes |
| *Burkholderia cepacia* | No | No | Yes | Yes |
| *Citrobacter braakii* | Yes | No | No | Yes |
| *Citrobacter freundii* | Yes | No | No | Yes |
| *Citrobacter portucalensis* | Yes | No | No | No |
| *Citrobacter werkmanii* | Yes | No | No | No |
| *Corynebacterium sp001815985* | No | No | No | No |
| *Enterobacter cloacae B* | Yes | No | No | Yes |
| *Enterobacter cloacae J* | Yes | No | No | Yes |
| *Enterobacter cloacae M* | Yes | No | No | Yes |
| *Enterobacter himalayensis* | Yes | No | No | No |
| *Enterobacter kobei* | Yes | No | No | Yes |
| *Enterococcus faecalis* | No | No | Yes | Yes |
| *Escherichia albertii* | Yes | No | No | No |
| *Escherichia coli* | Yes | Yes | Yes | Yes |
| *Escherichia coli C* | Yes | Yes | Yes | Yes |
| *Escherichia coli D* | Yes | Yes | Yes | Yes |
| *Escherichia dysenteriae* | Yes | Yes | No | No |
| *Escherichia fergusonii* | Yes | No | No | No |
| *Escherichia flexneri* | Yes | Yes | No | No |
| *Gallibacterium anatis* | No | No | No | No |
| *Klebsiella A oxytoca* | Yes | No | No | No |
| *Klebsiella pneumoniae* | Yes | No | Yes | Yes |
| *Klebsiella quasipneumoniae* | Yes | No | No | No |
| *Klebsiella quasivariicola* | Yes | No | No | No |
| *Klebsiella variicola* | Yes | No | No | No |
| *Leclercia adecarboxylata* | Yes | No | No | Yes |
| *Leclercia sp002902985* | Yes | No | No | No |
| *Mannheimia haemolytica* | No | No | No | Yes |
| *Metakosakonia intermedia* | Yes | No | No | No |
| *Moraxella A osloensis* | No | No | No | No |
| *Morganella morganii* | Yes | No | No | Yes |
| *Ochrobactrum intermedium* | No | No | No | No |
| *Pantoea B calida* | Yes | No | No | No |
| *Pantoea sp002920175* | Yes | No | No | No |
| *Pantoea stewartii* | Yes | No | No | No |
| *Pasteurella multocida* | Yes | No | No | Yes |
| *Proteus mirabilis* | Yes | No | No | Yes |
| *Providencia rettgeri D* | Yes | No | No | Yes |

| | | | | |
|---|---|---|---|---|
| *Pseudomonas A stutzeri* | No | No | No | No |
| *Pseudomonas E guariconensis* | No | No | No | No |
| *Pseudomonas E monteilli B* | No | No | No | No |
| *Pseudomonas E protegens* | No | No | No | No |
| *Pseudomonas E psychrophila A* | No | No | No | No |
| *Pseudomonas E sp002282595* | No | No | No | No |
| *Pseudomonas aeruginosa* | Yes | No | Yes | Yes |
| *Pseudomonas aeruginosa A* | Yes | No | Yes | Yes |
| *Salmonella enterica* | Yes | Yes | Yes | Yes |
| *Salmonella enterica E* | Yes | Yes | Yes | Yes |
| *Serratia marcescens I* | Yes | No | Yes | No |
| *Shewanella xiamenensis* | No | No | No | No |
| *Sphingobium yanoikuyae* | No | No | No | No |
| *Staphylococcus aureus* | Yes | Yes | Yes | Yes |
| *Staphylococcus epidermidis* | No | No | No | Yes |
| *Staphylococcus haemolyticus* | No | No | No | Yes |
| *Staphylococcus hominis* | No | No | No | Yes |
| *Stenotrophomonas maltophilia B* | No | No | No | Yes |
| *Stenotrophomonas maltophilia F* | No | No | No | Yes |
| *Streptomyces albidoflavus* | No | No | No | No |
| *Streptomyces bacilliaris* | No | No | No | No |
| *Trueperella pyogenes* | No | No | No | No |
| *Wohlfahrtiimonas chitiniclastica* | No | No | No | No |

**Table S2: Binary regression on the likelihood for a gene to have been horizontally transmitted.**

| **Variable** | **Estimate** | **Std. deviation** | **Z** | **p** |
|---|---|---|---|---|
| **Intercept** | -5.833 | 0.708 | -8.24 | **<2.2e-16** |
| **Intragenomic mobility MGEs** | 3.553 | 0.110 | 32.42 | **<2.2e-16** |
| **Intergenomic mobility MGEs** | 5.485 | 0.123 | 44.44 | **<2.2e-16** |
| **Nested MGEs** | 6.716 | 0.298 | 22.51 | **<2.2e-16** |
| **Human pathogens** | 3.229 | 0.075 | 42.85 | **<2.2e-16** |

**Table S3: Proportion of AMEGs included in each directed HGT network**

| CHG | Level | Putative HT genes | After the $d_2^*{}_{max}$ filter | After the $\Delta d_2^*$ filter |
|---|---|---|---|---|
| CHG 1 | Species | 580 | 499 | 405 |
| | Genera | | | 405 |
| CHG 2 | Species | 544 | 346 | 113 |
| | Genera | | | 113 |
| CHG 13 | Species | 969 | 955 | 507 |
| | Genera | | | 518 |
| CHG 14 | Species | 671 | 668 | 657 |
| | Genera | | | 656 |
| CHG 20 | Species | 862 | 723 | 464 |
| | Genera | | | 608 |
| CHG 25 | Species | 55 | 47 | 25 |
| | Genera | | | 32 |

**Table S4: Comparison of HGT networks at the level of genera with resampled networks.** T- and p-values displayed in the table are the statistics from a Welch's 1-sample t-test comparing the metric (diameter or degree assortativity) measured on each network to the distribution of 300 network metrics after genomes were randomly removed from the dataset. P-values in bold are lower than 0.05.

| CHG | Topological metric | 5% of genomes removed | | 10% of genomes removed | | 30% of genomes removed | |
|---|---|---|---|---|---|---|---|
| | | t | p | t | p | t | p |
| CHG 1 | Diameter | -3.5e-3 | 0.99 | 0.68 | 0.50 | -3.23 | **1.7e-3** |
| | Degree assortativity | -0.43 | -0.41 | -1.49 | 0.14 | -0.58 | 0.56 |
| CHG 2 | Diameter | 0.13 | 0.90 | 6.3e-2 | 0.95 | 0.20 | 0.84 |
| | Degree assortativity | -2.36 | **2.0e-2** | -8.0e-2 | 0.94 | -0.11 | 0.91 |
| CHG 13 | Diameter | -3.89 | **1.8e-4** | -3.59 | **5.2e-4** | -8.40 | **3.3e-13** |
| | Degree assortativity | 1.70 | 9.3e-2 | 6.0e-2 | 0.95 | -0.83 | 0.41 |
| CHG 14 | Diameter | 1.55 | 0.12 | -0.75 | 0.46 | 0.38 | 0.71 |
| | Degree assortativity | -0.69 | 0.49 | -2.06 | **4.2e-2** | 0.44 | 0.66 |
| CHG 20 | Diameter | -1.88 | 6.6e-2 | -3.17 | **2.0e-3** | -3.72 | **3.3e-4** |
| | Degree assortativity | -1.36 | 0.18 | -8.9e-2 | 0.93 | 1.70 | 9.2e-2 |
| CHG 25 | Diameter | -0.86 | 0.39 | -2.28 | **2.5e-2** | -0.62 | 0.54 |
| | Degree assortativity | -0.87 | 0.39 | -0.24 | 0.81 | 0.87 | 0.39 |

**Table S5: Comparison of HGT networks at the level of genera with randomly rewired networks.** T- and p-values displayed in the table are the statistics from a Welch's 1-sample t-test comparing the metric (diameter or degree assortativity) measured on each network to the distribution of 300 network metrics after partial random rewiring. P-values in bold are lower than 0.05.

| **CHG** | **Topological metric** | **10% of edges rewired** | | **30% of edges rewired** | | **50% of edges rewired** | |
|---|---|---|---|---|---|---|---|
| | | **t** | **p** | **t** | **p** | **t** | **p** |
| CHG 1 | Diameter | -1.57 | 0.12 | 6.29 | **8.7e-9** | 11.1 | **4.3e-19** |
| | Degree assortativity | 1.71 | 9.0e-2 | 3.06 | **2.8e-3** | 4.57 | **1.4e-5** |
| CHG 2 | Diameter | 2.54 | **1.3e-2** | 7.33 | **6.3e-11** | 9.84 | **2.4e-16** |
| | Degree assortativity | 1.34 | 0.18 | 3.93 | **1.6e-4** | 5.23 | **9.8e-7** |
| CHG 13 | Diameter | 1.99 | **4.9e-2** | 8.42 | **3.0e-13** | 6.32 | **7.6e-9** |
| | Degree assortativity | 1.03 | 0.30 | 2.55 | **1.2e-2** | 1.81 | 7.3e-2 |
| CHG 14 | Diameter | 5.49 | **3.0e-7** | 11.5 | **6.7e-20** | 14.3 | **6.7e-26** |
| | Degree assortativity | 1.21 | 0.23 | 3.09 | **2.6e-3** | 2.53 | **1.3e-2** |
| CHG 20 | Diameter | 1.68 | 0.10 | 5.25 | **8.6e-7** | 7.23 | **1.0e-10** |
| | Degree assortativity | 0.37 | 0.71 | 4.19 | **6.0e-5** | 5.30 | **7.1e-7** |
| CHG 25 | Diameter | 2.55 | **1.2e-2** | 4.42 | **2.5e-5** | 3.40 | **9.7e-4** |
| | Degree assortativity | -1.36 | 0.18 | 0.77 | 0.44 | -9.0e-2 | 0.93 |

**Table S6: Poisson PGLMM on the out-degree of species**.

| **Variable** | **Estimate** | **Std. deviation** | **Z** | **p** |
|---|---|---|---|---|
| **Intercept** | 0.788 | 0.328 | 2.404 | **1.6e-2** |
| **Other pathogens** | -0.561 | 0.543 | -1.034 | 0.30 |
| **Highly-problematic pathogens** | -1.537 | 0.405 | -3.792 | **1.5e-4** |

**Table S7: Binomial regression on the likelihood of horizontal gene transfer events in the 6 HGT networks built at the level of genera, without COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -8.64 | 0.819 | -10.5 | 0 *** |
| Phylogenetic distance | -0.853 | 0.332 | -2.57 | $1.02 \times 10^{-2}$ * |
| Intergenomic mobility MGEs | 5.88 | 0.536 | 11 | 0 *** |
| Intragenomic mobility MGEs | 0.171 | 0.533 | 0.321 | 0.748 |
| Nested MGEs | -3.81 | 0.677 | -5.63 | $1.8 \times 10^{-8}$ *** |
| Ecological distance | -0.612 | 0.424 | -1.44 | 0.149 |
| Phylogenetic distance × Intergenomic mobility MGEs | 0.178 | 0.472 | 0.377 | 0.706 |
| Phylogenetic distance × Intragenomic mobility MGEs | 1.23 | 0.41 | 3.01 | $2.63 \times 10^{-3}$ ** |
| Phylogenetic distance × Nested MGEs | 3.95 | 0.449 | 8.79 | 0 *** |

**Table S8: Binomial regression on the likelihood of horizontal gene transfer events in the 6 HGT networks built at the level of genera, including COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -3.51 | 0.927 | -3.79 | $1.49 \times 10^{-4}$ *** |
| Phylogenetic distance | -1.74 | 0.327 | -5.32 | $1.01 \times 10^{-7}$ *** |
| COUSIN18 | 0.476 | 0.143 | 3.33 | $8.63 \times 10^{-4}$ *** |
| Intergenomic mobility MGEs | -0.985 | 0.531 | -1.85 | $6.38 \times 10^{-2}$ * |
| Intragenomic mobility MGEs | -1.7 | 0.545 | -3.12 | $1.81 \times 10^{-3}$ ** |
| Nested MGEs | -9.27 | 0.695 | -13.3 | 0 *** |
| Ecological distance | -1.27 | 0.414 | -3.08 | $2.1 \times 10^{-3}$ ** |
| Phylogenetic distance × Intergenomic mobility MGEs | 0.221 | 0.523 | 0.423 | 0.672 |
| Phylogenetic distance × Intragenomic mobility MGEs | 1.52 | 0.408 | 3.73 | $1.95 \times 10^{-4}$ *** |
| Phylogenetic distance × Nested MGEs | 3.31 | 0.457 | 7.24 | $4.41 \times 10^{-13}$ *** |
| COUSIN18 × Intergenomic mobility MGEs | -0.644 | 0.401 | -1.61 | 0.108 |
| COUSIN18 × Intragenomic mobility MGEs | $7.93 \times 10^{-3}$ | 0.226 | $3.5 \times 10^{-2}$ | 0.972 |
| COUSIN18 × Nested MGEs | -1.37 | 0.205 | -6.68 | $2.42 \times 10^{-11}$ *** |

**Table S9: Binomial regression on the likelihood of horizontal gene transfer events in the 6 HGT networks built at the level of species, without COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -7.33 | 0.901 | -8.14 | $4.44 \times 10^{-16}$ *** |
| Phylogenetic distance | -1.21 | 0.566 | -2.15 | $3.17 \times 10^{-2}$ * |
| Intergenomic mobility MGEs | 5.98 | 0.715 | 8.37 | $1.11 \times 10^{-16}$ *** |
| Intragenomic mobility MGEs | 0.761 | 0.914 | 0.833 | 0.405 |
| Nested MGEs | -0.262 | 0.981 | -0.267 | 0.789 |
| Ecological distance | -2 | 0.301 | -6.62 | $3.53 \times 10^{-11}$ *** |
| Phylogenetic distance × Intergenomic mobility MGEs | 1.32 | 0.595 | 2.22 | $2.66 \times 10^{-2}$ * |
| Phylogenetic distance × Intragenomic mobility MGEs | 0.347 | 0.638 | 0.544 | 0.587 |
| Phylogenetic distance × Nested MGEs | 2.43 | 0.651 | 3.74 | $1.87 \times 10^{-4}$ *** |

**Table S10: Binomial regression on the likelihood of horizontal gene transfer events in the 6 HGT networks built at the level of species, including COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -3.36 | 0.909 | -3.7 | $2.2 \times 10^{-4}$ *** |
| Phylogenetic distance | -1.52 | 0.376 | -4.03 | $5.59 \times 10^{-5}$ *** |
| COUSIN18 | $7.96 \times 10^{-2}$ | 0.16 | 0.497 | 0.619 |
| Intergenomic mobility MGEs | -0.232 | 0.607 | -0.383 | 0.702 |
| Intragenomic mobility MGEs | 0.521 | 0.887 | 0.587 | 0.557 |
| Nested MGEs | -6.86 | 0.908 | -7.55 | $4.42 \times 10^{-14}$ *** |
| Ecological distance | -1.6 | 0.416 | -3.84 | $1.22 \times 10^{-4}$ *** |
| Phylogenetic distance × Intergenomic mobility MGEs | 0.439 | 0.484 | 0.907 | 0.365 |
| Phylogenetic distance × Intragenomic mobility MGEs | -0.222 | 0.471 | -0.471 | 0.638 |
| Phylogenetic distance × Nested MGEs | 2.69 | 0.549 | 4.91 | $9.17 \times 10^{-7}$ *** |
| COUSIN18 × Intergenomic mobility MGEs | 0.296 | 0.366 | 0.809 | 0.418 |
| COUSIN18 × Intragenomic mobility MGEs | -0.201 | 0.225 | -0.892 | 0.372 |
| COUSIN18 × Nested MGEs | -0.75 | 0.23 | -3.26 | $1.1 \times 10^{-3}$ ** |

**Table S11: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 1 HGT network built at the level of genera, without COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -4.21 | 1.14 | -3.71 | $2.09 \times 10^{-4}$ *** |
| Phylogenetic distance | -2.82 | 0.488 | -5.78 | $7.36 \times 10^{-9}$ *** |
| Intergenomic mobility MGEs | $-1.2 \times 10^{2}$ | 47.3 | -2.54 | $1.1 \times 10^{-2}$ * |
| Intragenomic mobility MGEs | 5.48 | 1.65 | 3.31 | $9.29 \times 10^{-4}$ *** |
| Nested MGEs | -1.75 | 1.58 | -1.11 | 0.269 |
| Ecological distance | -1.77 | 0.664 | -2.66 | $7.81 \times 10^{-3}$ ** |
| Phylogenetic distance × Intergenomic mobility MGEs | 48.7 | 17.3 | 2.81 | $4.88 \times 10^{-3}$ ** |
| Phylogenetic distance × Intragenomic mobility MGEs | 2.08 | 0.77 | 2.7 | $6.9 \times 10^{-3}$ ** |
| Phylogenetic distance × Nested MGEs | 0.376 | 1.06 | 0.353 | 0.724 |

**Table S12: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 1 HGT network built at the level of genera, including COUSIN18 as an effect.**
*: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -20.8 | 6.2 | -3.35 | $8.1 \times 10^{-4}$ *** |
| Phylogenetic distance | -6.9 | 1.17 | -5.9 | $3.55 \times 10^{-9}$ *** |
| COUSIN18 | 4.35 | 0.887 | 4.9 | $9.56 \times 10^{-7}$ *** |
| Intergenomic mobility MGEs | $-1.25 \times 10^{2}$ | 52.5 | -2.38 | $1.71 \times 10^{-2}$ * |
| Intragenomic mobility MGEs | 25 | 8.02 | 3.11 | $1.85 \times 10^{-3}$ ** |
| Nested MGEs | 14.3 | 6.09 | 2.35 | $1.87 \times 10^{-2}$ * |
| Ecological distance | -0.567 | 0.821 | -0.691 | 0.49 |
| Phylogenetic distance × Intergenomic mobility MGEs | 67.9 | 23.9 | 2.84 | $4.56 \times 10^{-3}$ ** |
| Phylogenetic distance × Intragenomic mobility MGEs | 7.31 | 1.44 | 5.07 | $3.91 \times 10^{-7}$ *** |
| Phylogenetic distance × Nested MGEs | 2.62 | 1.56 | 1.68 | $9.2 \times 10^{-2}$ * |
| COUSIN18 × Intergenomic mobility MGEs | -11.1 | 5.57 | -2 | $4.56 \times 10^{-2}$ * |
| COUSIN18 × Intragenomic mobility MGEs | -3.49 | 1.37 | -2.55 | $1.08 \times 10^{-2}$ * |
| COUSIN18 × Nested MGEs | -3.88 | 1.23 | -3.15 | $1.65 \times 10^{-3}$ ** |

**Table S13: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 1 HGT network built at the level of species, without COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -10.4 | 4.7 | -2.22 | $2.63 \times 10^{-2}$ * |
| Phylogenetic distance | 1.33 | 1.93 | 0.69 | 0.49 |
| Intergenomic mobility MGEs | $-1.06 \times 10^{2}$ | 56.2 | -1.88 | $6.01 \times 10^{-2}$ * |
| Intragenomic mobility MGEs | 12.6 | 4.86 | 2.59 | $9.51 \times 10^{-3}$ ** |
| Nested MGEs | 4.53 | 4.9 | 0.925 | 0.355 |
| Ecological distance | -2.09 | 0.838 | -2.49 | $1.27 \times 10^{-2}$ * |
| Phylogenetic distance × Intergenomic mobility MGEs | 40.1 | 20.8 | 1.93 | $5.34 \times 10^{-2}$ * |
| Phylogenetic distance × Intragenomic mobility MGEs | -3.4 | 1.99 | -1.71 | $8.71 \times 10^{-2}$ * |
| Phylogenetic distance × Nested MGEs | -3.68 | 2.16 | -1.7 | $8.85 \times 10^{-2}$ * |

**Table S14: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 1 HGT network built at the level of species, including COUSIN18 as an effect.**
*: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -8.37 | 5.17 | -1.62 | 0.106 |
| Phylogenetic distance | -2.08 | 3.37 | -0.616 | 0.538 |
| COUSIN18 | -1.2 | 0.81 | -1.48 | 0.139 |
| Intergenomic mobility MGEs | $-1 \times 10^{2}$ | 52.1 | -1.92 | $5.5 \times 10^{-2}$ * |
| Intragenomic mobility MGEs | 10.3 | 5.4 | 1.91 | $5.57 \times 10^{-2}$ * |
| Nested MGEs | 3.17 | 5.42 | 0.586 | 0.558 |
| Ecological distance | -1.4 | 0.894 | -1.57 | 0.118 |
| Phylogenetic distance × Intergenomic mobility MGEs | 39.4 | 20.5 | 1.92 | $5.42 \times 10^{-2}$ * |
| Phylogenetic distance × Intragenomic mobility MGEs | $3.46 \times 10^{-2}$ | 3.41 | $1.01 \times 10^{-2}$ | 0.992 |
| Phylogenetic distance × Nested MGEs | 0.104 | 3.56 | $2.93 \times 10^{-2}$ | 0.977 |
| COUSIN18 × Intergenomic mobility MGEs | 18.5 | 23.5 | 0.789 | 0.43 |
| COUSIN18 × Intragenomic mobility MGEs | 1.13 | 0.887 | 1.28 | 0.201 |
| COUSIN18 × Nested MGEs | 2.11 | 1.6 | 1.32 | 0.187 |

**Table S15: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 2 HGT network built at the level of genera, without COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -2.94 | 1.37 | -2.14 | $3.2 \times 10^{-2}$ * |
| Phylogenetic distance | -2.83 | 0.665 | -4.25 | $2.12 \times 10^{-5}$ *** |
| Intergenomic mobility MGEs | 1.57 | 1.65 | 0.948 | 0.343 |
| Intragenomic mobility MGEs | 39.6 | 24.6 | 1.61 | 0.107 |
| Nested MGEs | -12.2 | 5.25 | -2.33 | $1.98 \times 10^{-2}$ * |
| Ecological distance | -4.17 | 1.46 | -2.87 | $4.14 \times 10^{-3}$ ** |
| Phylogenetic distance × Intergenomic mobility MGEs | -5.35 | 2.33 | -2.3 | $2.15 \times 10^{-2}$ * |
| Phylogenetic distance × Intragenomic mobility MGEs | $-6.44 \times 10^{2}$ | $2.82 \times 10^{2}$ | -2.28 | $2.25 \times 10^{-2}$ * |
| Phylogenetic distance × Nested MGEs | 27 | 4.3 | 6.29 | $3.21 \times 10^{-10}$ *** |

**Table S16: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 2 HGT network built at the level of genera, including COUSIN18 as an effect.**
*: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | $1.99 \times 10^{15}$ | $2.98 \times 10^{6}$ | $6.67 \times 10^{8}$ | 0 *** |
| Phylogenetic distance | $-1.28 \times 10^{15}$ | $1.09 \times 10^{6}$ | $-1.18 \times 10^{9}$ | 0 *** |
| COUSIN18 | $1.29 \times 10^{14}$ | $3.32 \times 10^{5}$ | $3.89 \times 10^{8}$ | 0 *** |
| Intergenomic mobility MGEs | $-3.11 \times 10^{15}$ | $4.94 \times 10^{6}$ | $-6.3 \times 10^{8}$ | 0 *** |
| Intragenomic mobility MGEs | $-1.02 \times 10^{16}$ | $1.21 \times 10^{7}$ | $-8.4 \times 10^{8}$ | 0 *** |
| Nested MGEs | $-6.05 \times 10^{15}$ | $5.12 \times 10^{6}$ | $-1.18 \times 10^{9}$ | 0 *** |
| Ecological distance | $-3.04 \times 10^{15}$ | $2.54 \times 10^{6}$ | $-1.2 \times 10^{9}$ | 0 *** |
| Phylogenetic distance × Intergenomic mobility MGEs | $2 \times 10^{15}$ | $2.36 \times 10^{6}$ | $8.5 \times 10^{8}$ | 0 *** |
| Phylogenetic distance × Intragenomic mobility MGEs | $1.75 \times 10^{15}$ | $7.01 \times 10^{6}$ | $2.5 \times 10^{8}$ | 0 *** |
| Phylogenetic distance × Nested MGEs | $3.28 \times 10^{15}$ | $3.03 \times 10^{6}$ | $1.08 \times 10^{9}$ | 0 *** |
| COUSIN18 × Intergenomic mobility MGEs | $-1.99 \times 10^{14}$ | $1.3 \times 10^{6}$ | $-1.53 \times 10^{8}$ | 0 *** |
| COUSIN18 × Intragenomic mobility MGEs | $-1.85 \times 10^{15}$ | $4.07 \times 10^{6}$ | $-4.54 \times 10^{8}$ | 0 *** |
| COUSIN18 × Nested MGEs | $3.58 \times 10^{14}$ | $2.23 \times 10^{6}$ | $1.6 \times 10^{8}$ | 0 *** |

**Table S17: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 2 HGT network built at the level of species, without COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -3.17 | 1.36 | -2.34 | $1.92 \times 10^{-2}$ * |
| Phylogenetic distance | -3.66 | 1.18 | -3.11 | $1.86 \times 10^{-3}$ ** |
| Intergenomic mobility MGEs | 2.99 | 1.62 | 1.85 | $6.39 \times 10^{-2}$ * |
| Intragenomic mobility MGEs | $1.32 \times 10^{2}$ | $3 \times 10^{3}$ | $4.4 \times 10^{-2}$ | 0.965 |
| Nested MGEs | -48.5 | 11.4 | -4.24 | $2.28 \times 10^{-5}$ *** |
| Ecological distance | -2.25 | 1.07 | -2.1 | $3.6 \times 10^{-2}$ * |
| Phylogenetic distance × Intergenomic mobility MGEs | -3.29 | 1.46 | -2.26 | $2.38 \times 10^{-2}$ * |
| Phylogenetic distance × Intragenomic mobility MGEs | $-4.96 \times 10^{2}$ | $1.11 \times 10^{4}$ | $-4.48 \times 10^{-2}$ | 0.964 |
| Phylogenetic distance × Nested MGEs | 22.1 | 5.06 | 4.36 | $1.32 \times 10^{-5}$ *** |

**Table S18: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 2 HGT network built at the level of species, including COUSIN18 as an effect.**
*: p<0.05, **: p<0.01, ***: p<0.001.

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -11.5 | 3.13 | -3.66 | $2.49 \times 10^{-4}$ *** |
| Phylogenetic distance | 1.54 | 1.17 | 1.32 | 0.188 |
| COUSIN18 | $5.15 \times 10^{-2}$ | 0.779 | $6.61 \times 10^{-2}$ | 0.947 |
| Intergenomic mobility MGEs | 13.1 | 3.43 | 3.82 | $1.35 \times 10^{-4}$ *** |
| Intragenomic mobility MGEs | 23.7 | $1.06 \times 10^{2}$ | 0.224 | 0.823 |
| Nested MGEs | -2.53 | 7.84 | -0.323 | 0.747 |
| Ecological distance | -4.34 | 1.61 | -2.69 | $7.15 \times 10^{-3}$ ** |
| Phylogenetic distance × Intergenomic mobility MGEs | -10.3 | 1.99 | -5.17 | $2.34 \times 10^{-7}$ *** |
| Phylogenetic distance × Intragenomic mobility MGEs | -71.7 | $1.98 \times 10^{2}$ | -0.362 | 0.717 |
| Phylogenetic distance × Nested MGEs | 2.24 | 3.77 | 0.593 | 0.553 |
| COUSIN18 × Intergenomic mobility MGEs | -0.997 | 1.29 | -0.774 | 0.439 |
| COUSIN18 × Intragenomic mobility MGEs | 1.53 | $1.36 \times 10^{2}$ | $1.12 \times 10^{-2}$ | 0.991 |
| COUSIN18 × Nested MGEs | 0.964 | 3.21 | 0.301 | 0.764 |

**Table S19: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 13 HGT network built at the level of genera, without COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -4.17 | 2.1 | -1.99 | $4.71 \times 10^{-2}$ * |
| Phylogenetic distance | -3.12 | 1.28 | -2.44 | $1.47 \times 10^{-2}$ * |
| Intergenomic mobility MGEs | -2.32 | 2.9 | -0.801 | 0.423 |
| Intragenomic mobility MGEs | -0.365 | 2.24 | -0.163 | 0.871 |
| Nested MGEs | -3.2 | 2.21 | -1.45 | 0.147 |
| Ecological distance | 0.449 | 0.404 | 1.11 | 0.267 |
| Phylogenetic distance × Intergenomic mobility MGEs | 1.65 | 2.33 | 0.706 | 0.48 |
| Phylogenetic distance × Intragenomic mobility MGEs | 2.32 | 1.35 | 1.73 | $8.43 \times 10^{-2}$ * |
| Phylogenetic distance × Nested MGEs | 4.01 | 1.31 | 3.06 | $2.2 \times 10^{-3}$ ** |

**Table S20: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 13 HGT network built at the level of genera, including COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | 1.93 | 4.66 | 0.415 | 0.678 |
| Phylogenetic distance | -7.14 | 3.59 | -1.99 | $4.68 \times 10^{-2}$ * |
| COUSIN18 | 0.495 | 0.355 | 1.39 | 0.164 |
| Intergenomic mobility MGEs | -12.1 | 6.96 | -1.74 | $8.11 \times 10^{-2}$ * |
| Intragenomic mobility MGEs | -6.36 | 4.82 | -1.32 | 0.187 |
| Nested MGEs | -10.1 | 4.79 | -2.11 | $3.48 \times 10^{-2}$ * |
| Ecological distance | 0.246 | 0.473 | 0.52 | 0.603 |
| Phylogenetic distance × Intergenomic mobility MGEs | 6.48 | 5.33 | 1.22 | 0.224 |
| Phylogenetic distance × Intragenomic mobility MGEs | 6.85 | 3.67 | 1.86 | $6.22 \times 10^{-2}$ * |
| Phylogenetic distance × Nested MGEs | 8.4 | 3.59 | 2.34 | $1.94 \times 10^{-2}$ * |
| COUSIN18 × Intergenomic mobility MGEs | $-9.28 \times 10^{-2}$ | 1.06 | $-8.78 \times 10^{-2}$ | 0.93 |
| COUSIN18 × Intragenomic mobility MGEs | 0.199 | 0.395 | 0.503 | 0.615 |
| COUSIN18 × Nested MGEs | -1.09 | 0.377 | -2.89 | $3.86 \times 10^{-3}$ ** |

**Table S21: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 13 HGT network built at the level of species, without COUSIN18 as an effect.**
*: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | 11.2 | 4.32 | 2.59 | $9.61 \times 10^{-3}$ ** |
| Phylogenetic distance | -12 | 3.06 | -3.91 | $9.27 \times 10^{-5}$ *** |
| Intergenomic mobility MGEs | -18.8 | 5 | -3.76 | $1.72 \times 10^{-4}$ *** |
| Intragenomic mobility MGEs | -15.7 | 4.35 | -3.61 | $3.02 \times 10^{-4}$ *** |
| Nested MGEs | -20.6 | 4.47 | -4.62 | $3.86 \times 10^{-6}$ *** |
| Ecological distance | -0.867 | 0.6 | -1.44 | 0.149 |
| Phylogenetic distance × Intergenomic mobility MGEs | 10.6 | 3.93 | 2.71 | $6.75 \times 10^{-3}$ ** |
| Phylogenetic distance × Intragenomic mobility MGEs | 10.4 | 3.09 | 3.36 | $7.66 \times 10^{-4}$ *** |
| Phylogenetic distance × Nested MGEs | 13.3 | 3.12 | 4.26 | $2.02 \times 10^{-5}$ *** |

**Table S22: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 13 HGT network built at the level of species, including COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | 10.6 | 4.83 | 2.2 | $2.77 \times 10^{-2}$ * |
| Phylogenetic distance | -11.4 | 3.42 | -3.32 | $8.88 \times 10^{-4}$ *** |
| COUSIN18 | 0.353 | 0.418 | 0.846 | 0.398 |
| Intergenomic mobility MGEs | -19.3 | 6.6 | -2.93 | $3.4 \times 10^{-3}$ ** |
| Intragenomic mobility MGEs | -14.2 | 4.86 | -2.92 | $3.48 \times 10^{-3}$ ** |
| Nested MGEs | -21 | 5.01 | -4.19 | $2.78 \times 10^{-5}$ *** |
| Ecological distance | -1.52 | 0.681 | -2.24 | $2.52 \times 10^{-2}$ * |
| Phylogenetic distance × Intergenomic mobility MGEs | 5.55 | 5.6 | 0.99 | 0.322 |
| Phylogenetic distance × Intragenomic mobility MGEs | 9.76 | 3.45 | 2.83 | $4.72 \times 10^{-3}$ ** |
| Phylogenetic distance × Nested MGEs | 13.7 | 3.5 | 3.92 | $8.68 \times 10^{-5}$ *** |
| COUSIN18 × Intergenomic mobility MGEs | 0.155 | 1.16 | 0.134 | 0.893 |
| COUSIN18 × Intragenomic mobility MGEs | 0.247 | 0.438 | 0.565 | 0.572 |
| COUSIN18 × Nested MGEs | -0.778 | 0.456 | -1.71 | $8.78 \times 10^{-2}$ * |

**Table S23: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 20 HGT network built at the level of genera, without COUSIN18 as an effect.**
*: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | $-7.94 \times 10^{14}$ | $2.96 \times 10^{6}$ | $-2.69 \times 10^{8}$ | 0 *** |
| Phylogenetic distance | $1.18 \times 10^{15}$ | $1.23 \times 10^{6}$ | $9.61 \times 10^{8}$ | 0 *** |
| Intergenomic mobility MGEs | $-7.59 \times 10^{14}$ | $5.27 \times 10^{6}$ | $-1.44 \times 10^{8}$ | 0 *** |
| Intragenomic mobility MGEs | $-1.84 \times 10^{15}$ | $3.79 \times 10^{6}$ | $-4.85 \times 10^{8}$ | 0 *** |
| Nested MGEs | $2.1 \times 10^{15}$ | $7.55 \times 10^{6}$ | $2.78 \times 10^{8}$ | 0 *** |
| Ecological distance | $-4.42 \times 10^{15}$ | $2.92 \times 10^{6}$ | $-1.51 \times 10^{9}$ | 0 *** |
| Phylogenetic distance × Intergenomic mobility MGEs | $-2.43 \times 10^{15}$ | $3.67 \times 10^{6}$ | $-6.61 \times 10^{8}$ | 0 *** |
| Phylogenetic distance × Intragenomic mobility MGEs | $-1.04 \times 10^{15}$ | $3.12 \times 10^{6}$ | $-3.34 \times 10^{8}$ | 0 *** |
| Phylogenetic distance × Nested MGEs | $7.04 \times 10^{14}$ | $6.09 \times 10^{6}$ | $1.16 \times 10^{8}$ | 0 *** |

**Table S24: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 20 HGT network built at the level of genera, including COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | $-6.52 \times 10^{15}$ | $6.88 \times 10^{6}$ | $-9.48 \times 10^{8}$ | 0 *** |
| Phylogenetic distance | $1.86 \times 10^{15}$ | $3.59 \times 10^{6}$ | $5.17 \times 10^{8}$ | 0 *** |
| COUSIN18 | $-4.34 \times 10^{14}$ | $1.34 \times 10^{6}$ | $-3.23 \times 10^{8}$ | 0 *** |
| Intergenomic mobility MGEs | $9.52 \times 10^{15}$ | $1.76 \times 10^{7}$ | $5.42 \times 10^{8}$ | 0 *** |
| Intragenomic mobility MGEs | $1.11 \times 10^{16}$ | $2.83 \times 10^{7}$ | $3.94 \times 10^{8}$ | 0 *** |
| Nested MGEs | $-1.65 \times 10^{15}$ | $2.35 \times 10^{7}$ | $-7 \times 10^{7}$ | 0 *** |
| Ecological distance | $-9.04 \times 10^{14}$ | $8.13 \times 10^{6}$ | $-1.11 \times 10^{8}$ | 0 *** |
| Phylogenetic distance × Intergenomic mobility MGEs | $-4.39 \times 10^{15}$ | $7.55 \times 10^{6}$ | $-5.81 \times 10^{8}$ | 0 *** |
| Phylogenetic distance × Intragenomic mobility MGEs | $-8.51 \times 10^{15}$ | $1.47 \times 10^{7}$ | $-5.79 \times 10^{8}$ | 0 *** |
| Phylogenetic distance × Nested MGEs | $3.29 \times 10^{15}$ | $1.04 \times 10^{7}$ | $3.16 \times 10^{8}$ | 0 *** |
| COUSIN18 × Intergenomic mobility MGEs | $2.22 \times 10^{15}$ | $3.17 \times 10^{7}$ | $6.99 \times 10^{7}$ | 0 *** |
| COUSIN18 × Intragenomic mobility MGEs | $1.24 \times 10^{16}$ | $4.11 \times 10^{7}$ | $3.03 \times 10^{8}$ | 0 *** |
| COUSIN18 × Nested MGEs | $-5.4 \times 10^{15}$ | $3.75 \times 10^{7}$ | $-1.44 \times 10^{8}$ | 0 *** |

**Table S25: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 20 HGT network built at the level of species, without COUSIN18 as an effect.**
*: p<0.05, **: p<0.01, ***: p<0.001.

| **Variable** | **Estimate** | **Conditional standard error** | **t** | **p** |
|---|---|---|---|---|
| Intercept | -9.43 | 2.17 | -4.35 | $1.38 \times 10^{-5}$ *** |
| Phylogenetic distance | 1.45 | 1.04 | 1.39 | 0.164 |
| Intergenomic mobility MGEs | 4.74 | 2.58 | 1.84 | $6.57 \times 10^{-2}$ * |
| Intragenomic mobility MGEs | 1.91 | 2.99 | 0.638 | 0.523 |
| Nested MGEs | 1.56 | 2.58 | 0.602 | 0.547 |
| Ecological distance | 0.831 | 0.858 | 0.968 | 0.333 |
| Phylogenetic distance × Intergenomic mobility MGEs | -1.75 | 1.72 | -1.02 | 0.308 |
| Phylogenetic distance × Intragenomic mobility MGEs | -0.605 | 2.22 | -0.272 | 0.786 |
| Phylogenetic distance × Nested MGEs | -3.52 | 4.11 | -0.858 | 0.391 |

**Table S26: Binomial regression on the likelihood of horizontal gene transfer events in the CHG 20 HGT network built at the level of species, including COUSIN18 as an effect.** *: p<0.05, **: p<0.01, ***: p<0.001.

| Variable | Estimate | Conditional standard error | t | p |
|---|---|---|---|---|
| Intercept | -9.38 | 7.7 | -1.22 | 0.223 |
| Phylogenetic distance | -2.68 | 15.9 | -0.169 | 0.866 |
| COUSIN18 | -1.6 | 8.87 | -0.18 | 0.857 |
| Intergenomic mobility MGEs | 6.51 | 8.35 | 0.779 | 0.436 |
| Intragenomic mobility MGEs | -19.4 | 28.2 | -0.687 | 0.492 |
| Nested MGEs | 3.79 | 8 | 0.474 | 0.635 |
| Ecological distance | 0.761 | 1.17 | 0.65 | 0.516 |
| Phylogenetic distance × Intergenomic mobility MGEs | 1.08 | 16.9 | $6.38 \times 10^{-2}$ | 0.949 |
| Phylogenetic distance × Intragenomic mobility MGEs | 5.11 | 65.4 | $7.83 \times 10^{-2}$ | 0.938 |
| Phylogenetic distance × Nested MGEs | 0.409 | 17.4 | $2.35 \times 10^{-2}$ | 0.981 |
| COUSIN18 × Intergenomic mobility MGEs | 2.31 | 9.82 | 0.235 | 0.814 |
| COUSIN18 × Intragenomic mobility MGEs | -16.5 | 52.4 | -0.316 | 0.752 |
| COUSIN18 × Nested MGEs | 1.95 | 9.16 | 0.213 | 0.832 |

**Table S27: Reciprocate transfers in directed HGT networks**

| CHG | Proportion of edges | Species 1 | Species 2 |
|---|---|---|---|
| CHG 1 | 8/31 | *Brucella melitensis* | *Acinetobacter baumannii* |
| | | *Pseudomonas aeruginosa* | *Acinetobacter baumannii* |
| | | *Pseudomonas aeruginosa* | *Serratia marcescens I* |
| | | *Pseudomonas aeruginosa* | *Shewanella xiamenensis* |
| CHG 13 | 8/116 | *Pseudomonas aeruginosa* | *Pseudomonas A stutzeri* |
| | | *Pseudomonas aeruginosa* | *Pseudomonas E monteilii B* |
| | | *Pseudomonas E psychrophila* | *Trueperella pyogenes,* |
| | | *Pseudomonas aeruginosa* | *Salmonella enterica* |
| CHG 20 | 4/40 | *Klebsiella pneumoniae* | *Enterobacter himalayensis* |
| | | *Klebsiella pneumoniae* | *Escherichia flexneri* |

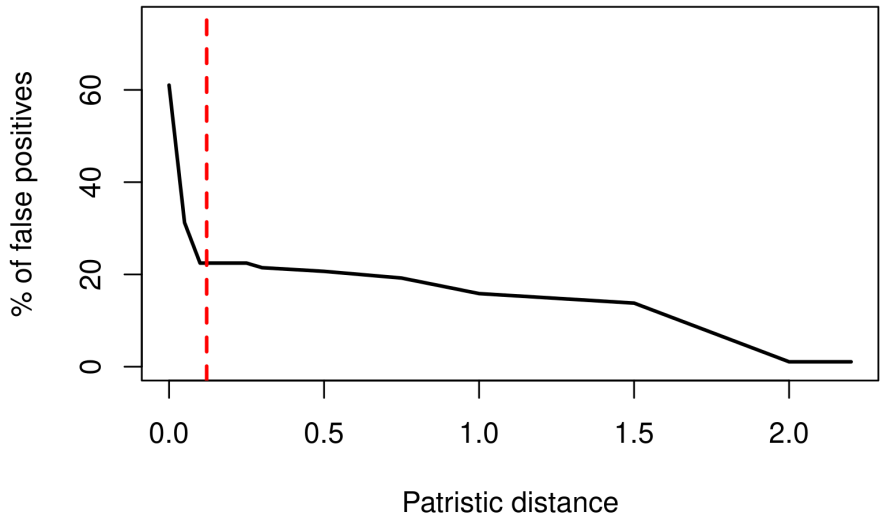


**Figure S1: False positives for implicit phylogenetic HGT detection in simulated sequence alignments for CHG 11.** The set of species is the one in which CHG 11 was sampled, and sequence length is median protein length for CHG 11. The red line corresponds to minimum patristic distance between two species of different genera.

# 11 Annexe 3 : Sources de sélection sur le biais d'usage de codons

L'annexe 3 présente un article de perspective discutant des différents types de sélection pouvant s'exercer sur le biais d'usage de codon. Ma participation à ce projet est la rédaction de la *Text box 1* sur le potentiel rôle des éléments transposable à la mise en place et au renforcement du biais d'usage de codon. En effet, l'intégration des éléments transposables dépend de la composition nucléotidique des séquences : soit du fait de motifs préférés, soit du fait de propriétés mécaniques de l'ADN associées. Ainsi, suivant l'usage de codon local, la probabilité d'intégration d'un élément transposable pourrait être plus ou moins élevée. Or, l'intégration d'un élément transposable dans une séquence codante peut empêcher l'expression de ce gène. La disruption de l'expression peut sélectionner, selon les contextes, pour un enrichissement en certains motifs, ou au contraire leur évitement. Cette sélection modulerait la composition nucléotidique locale, et donc l'usage de codons.

Durant ma dernière année de thèse j'ai encadré un stage de Master 1 se proposant de tester cette hypothèse. L'objectif de ce stage bioinformatique était de tester si différentes catégories de gènes étaient appauvris ou enrichis en motifs d'insertion d'un élément transposable très répandu (IS10) dans les génomes de 19 espèces bactériennes, et si cette caractéristique variait en fonction de l'écologie de ces espèces. Ce stage s'est principalement focalisé sur les gènes dont la disruption produit des phénotypes hypermutateurs. En effet, l'évolvabilité chez les bactéries peut être très liée à la régulation du taux de mutation, et il se pourrait donc que la sélection pour un enrichissement en motif d'insertion dans les gènes hypermutateurs dépende avant tout de l'écologie. Notamment, la prédiction est que les gènes à potentiel hypermutateur sont plus enrichis en motifs d'intégration chez les espèces vivant dans les environnements les plus variables. Par la suite, ce projet a également tenté de généraliser ce patron observé à d'autres types de gènes.

# Read between the Lines: Diversity of Nontranslational Selection Pressures on Local Codon Usage

Martijn Callens, Léa Pradier, Michael Finnegan, Caroline Rose, and Stéphanie Bedhomme*

Centre d'Ecologie Fonctionnelle et Evolutive, CNRS, Université de Montpellier, Université Paul Valéry Montpellier 3, Ecole Pratique des Hautes Etudes, Institut de Recherche pour le Développement, Montpellier, France

*Corresponding author: E-mail: stephanie.bedhomme@cefe.cnrs.fr.

**Accepted:** 28 April 2021

Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

## Abstract

Protein coding genes can contain specific motifs within their nucleotide sequence that function as a signal for various biological pathways. The presence of such sequence motifs within a gene can have beneficial or detrimental effects on the phenotype and fitness of an organism, and this can lead to the enrichment or avoidance of this sequence motif. The degeneracy of the genetic code allows for the existence of alternative synonymous sequences that exclude or include these motifs, while keeping the encoded amino acid sequence intact. This implies that locally, there can be a selective pressure for preferentially using a codon over its synonymous alternative in order to avoid or enrich a specific sequence motif. This selective pressure could—in addition to mutation, drift and selection for translation efficiency and accuracy—contribute to shape the codon usage bias. In this review, we discuss patterns of avoidance of (or enrichment for) the various biological signals contained in specific nucleotide sequence motifs: transcription and translation initiation and termination signals, mRNA maturation signals, and antiviral immune system targets. Experimental data on the phenotypic or fitness effects of synonymous mutations in these sequence motifs confirm that they can be targets of local selection pressures on codon usage. We also formulate the hypothesis that transposable elements could have a similar impact on codon usage through their preferred integration sequences. Overall, selection on codon usage appears to be a combination of a global selection pressure imposed by the translation machinery, and a patchwork of local selection pressures related to biological signals contained in specific sequence motifs.

**Key words:** codon usage, synonymous mutations, gene expression regulation, sequence targeting antiviral immune systems, transposable elements.

## Significance statement

The frequency of use of synonymous codons varies between species and is known to be under selection for translation speed and accuracy. In this review, we argue that an additional local selection pressure on codon usage is generated by sequence motifs conveying different biological signals such as transcription and translation initiation, mRNA maturation, antiviral immune system targets or preferred transposable elements insertion sequences. Alternative synonymous sequences can be favored or disfavored because they contain these motif sequences. We review experimental and bioinformatic evidence for these local selection pressures.

## Introduction

The redundancy of the genetic code is a consequence of the existence of synonymous codons, which differ by their nucleotide triplets but code for the same amino acid. The different codons within a synonymous codon family are not used at equal frequencies; this codon usage bias (CUB) can vary between species and between genes within a species (Grantham et al. 1981; Ikemura 1985). CUB is shaped by mutation, selection, and drift (Bulmer 1991; Hershberg and Petrov 2008; Plotkin and Kudla 2011; Shah and Gilchrist 2011). Selection on CUB is generally assumed to be driven by its effects on translation efficiency (Tuller, Waldman, et al.

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

2010) and accuracy (Kurland 1992; Stoletzki and Eyre-Walker 2007), mediated by the coevolution of translation machinery and CUB: an association between the frequency of use of a codon and the availability of the corresponding decoding tRNA has been established for various genomes (e.g., Duret 2000; Rocha 2004). Codon usage has been shown to modulate the rate and efficiency of translation, with examples ranging from decreases in viral capsid protein production leading to virus attenuation (Coleman et al. 2008) to 58% translation elongation rate increases in human cell lines (Yan et al. 2016).

Selection on codon usage does not always act in the direction of higher translation efficiency, and this direction can vary across the genome and within genes. For example, in many prokaryotic and eukaryotic species the first 30–50 bp of genes often present an accumulation of codons which are at low frequency in the rest of the genome. This has been associated with a localized slow translation, preventing ribosomal collisions downstream (Tuller, Carmi, et al. 2010). In bacteria, it has been established that the corresponding part of the mRNA presents a reduced folding energy compared to the rest of the mRNA, which is assumed to favor translation initiation. An analysis of over 400 bacteria genomes confirmed that codons overrepresented at the beginning of the genes are those that reduce mRNA folding around the translation start, regardless of whether these codons are frequent or rare (Bentele et al. 2013).

Ribosome profiling and other technical advances have led to an in-depth understanding of the complex relationship between codon usage, translation efficiency regulation, and proteome composition. They enabled, for example, descriptions of the effect of codon usage on mRNA secondary structure (Katz 2003) and accessibility to ribosomes (Kudla et al. 2009) as well as the measure of the rate of ribosomal drop-off at low-frequency codons producing truncated proteins (Yang et al. 2019). The kinetic coupling of translational speed and protein folding has been described in detail (Pechmann and Frydman 2013; Yu et al. 2015; Chaney et al. 2017; Zhao et al. 2017). Finally, the modulatory role of codon usage in mRNA decay and stability has been documented in bacteria (Boël et al. 2016), single celled eukaryotic yeast (Radhakrishnan et al. 2016), and between different tissues in humans (Burow et al. 2018). In particular, in human cells, codon usage is a key determinant of the routing of mRNA towards P-bodies which are cytoplasmic organelles involved in mRNA storage and decay (Courel et al. 2019). These phenomena have been reviewed by Brule and Grayhack (2017) and are not the focus of the present review.

The existence of alternative synonymous sequences suggests that protein coding genes could potentially contain or exclude sequence motifs with biologically meaningful signals in addition to simply coding for an amino acid sequence. These biological signals can take the form of motifs in the actual nucleotide sequence, or in the biophysical properties of this sequence (secondary structure, hairpins, stiffness, etc.). The presence of these "other codes" is particularly recognized for biological signals involved in gene expression (e.g., Bergman and Tuller 2020), and it has been suggested that the genetic code is better suited for encoding this additional information than the vast majority of the potential alternative genetic codes (Itzkovitz and Alon 2007). We argue here that the potential for genes to contain information beyond the code of the amino acid sequence implies that specific nucleotide sequences can be favored or disfavored, because of the biological signal they carry. This can result in selection on local codon usage for reasons other than its consequences on translation accuracy and efficiency. In this review, we compile the different biological signals that can be contained in nucleotide sequences. We further discuss patterns of avoidance or enrichment of these sequence motifs and, when available, we present experimental evidence of the phenotypic effects of synonymous mutations in relation to these biological signals. Figure 1 provides a summary of the elements discussed in this review.

## Sequence motifs involved in gene expression regulation

### Promoter, Near-Promoter and Antisense Promoter Sequences

Promoters in bacteria are characterized by two consensus sequences, TATAAT and TTGACA, respectively located 10 and 35 bp upstream of the transcriptional start site (Browning and and Busby 2004). Active promoter sequences are not necessarily an exact consensus sequence but usually contain only three or four of the six nucleotides (Kinney et al. 2010). Promoter sequences, or sequences within a short mutational distance from a promoter sequence, are likely to occur within DNA sequences because they are short and moderately conserved. Indeed, 10% of 100 bp random sequences exhibit promoter activity in *Escherichia coli*, and within 250 generations 60% of random sequences evolved functional promoter activity due to a single mutation (Yona et al. 2018). The potential of a given sequence to evolve a functional promoter can be beneficial in terms of plasticity and evolvability of the transcription network. It can even be beneficial when occurring in a coding sequence: for example, in bacteria, synonymous mutations at the end of the coding sequence of a gene have been shown to be beneficial because they create a promoter from which the next gene in the operon is transcribed, and this overexpression is advantageous in specific environmental conditions (Ando et al. 2014; Kershner et al. 2016). However, the appearance of a new promoter within a coding sequence can also lead to an overproduction of RNA transcripts, sequestration of RNA polymerase, and an overall reduction in gene expression (Lamberte et al. 2017). Hahn (2003) found that coding sequences across



Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

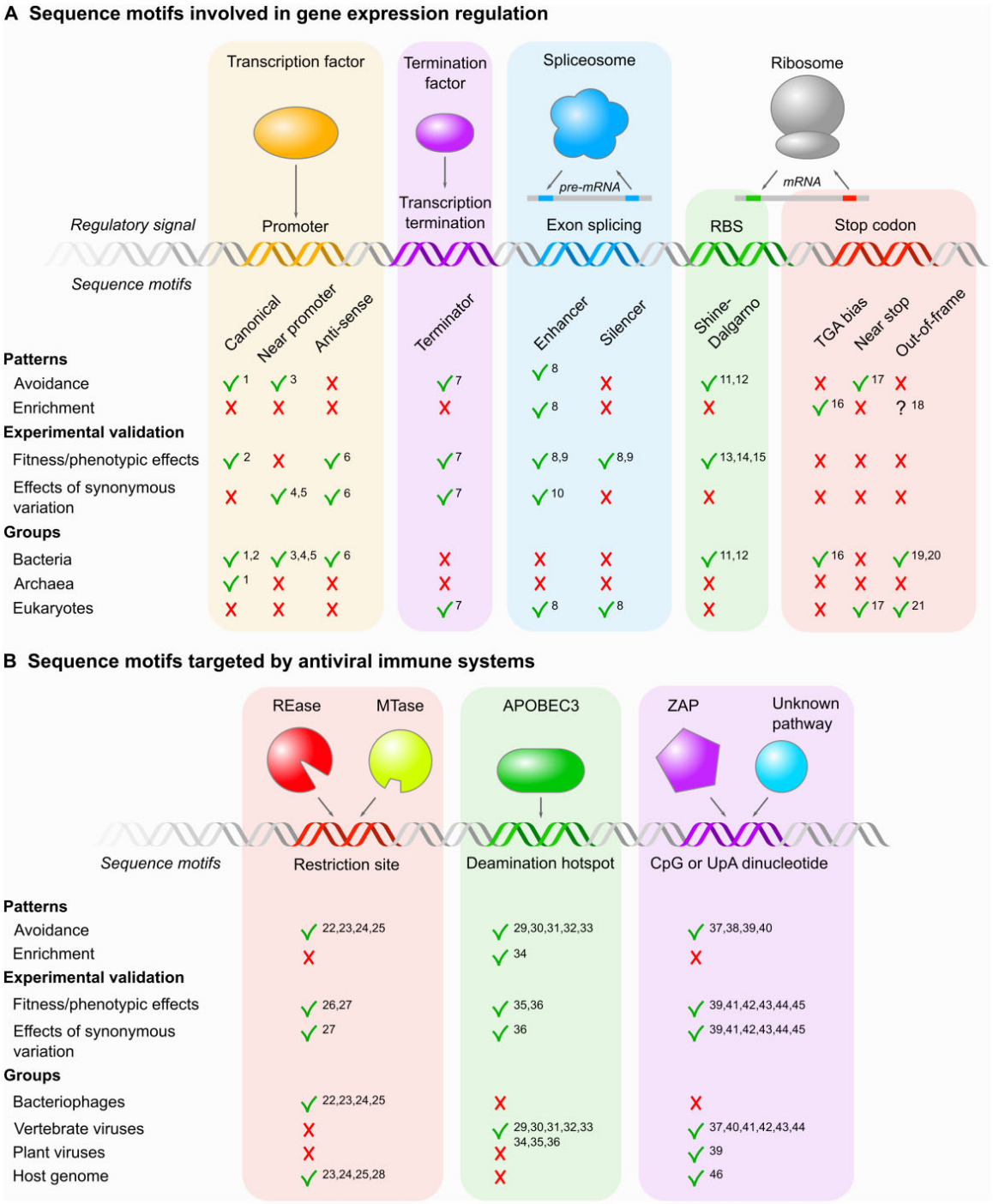


FIG. 1.—(A) Observed avoidance or enrichment of sequence motifs involved in gene expression regulation and potential phenotypic effects. Different processes depend on particular sequence motifs in the DNA or mRNA for their regulation (colored boxes from left to right: transcription initiation, transcription termination, gene splicing, translation initiation, translation termination). Green checks indicate if there is evidence in the literature for avoidance or enrichment of particular sequence motifs, if the presence or absence of these sequence motifs has observable phenotypic effects and if these phenotypic effects can be modified through synonymous variation. An "?" indicates this issue is debated. The bottom rows indicate in which domains of life these





Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

Eubacteria and Archaea are under selection to avoid canonical promoter sequences, and Yona et al. (2018) computationally showed that the *E. coli* coding genome is depleted in sequences close to promoter sequences. Furthermore, this avoidance pattern is even stronger for essential genes, for which perturbation is extremely costly. This suggests that specific intragenic combinations of codons corresponding to promoter or near-promoter sequences are generally disadvantageous but can also be beneficial in specific genomic and environmental situations.

Intragenic promoters are, however, present on the antisense strand in a diversity of bacterial species (Cohen et al. 2016). Transcription from antisense promoters produces RNA fragments that are strictly complementary to the mRNAs produced from the sense strand and can hybridize with them. Antisense transcripts often lead to some repression of translation because the presence of RNA duplexes along mRNA can inhibit translation and target mRNA for degradation (Brantl 2007; Brophy and Voigt 2016). It is unclear when and to what degree the presence of these antisense promoters is spurious or favored by selection because of their role in translational regulation (Gophna 2018). Urtecho et al. (2020) showed experimentally that *E. coli* genes containing antisense promoter sequences had lower transcript levels. This study also revealed that the portions of the sense strand complementary to the antisense promoters contain many codons present at low frequency in the rest of the genome. These sequences thus seem to be constrained both by their role in amino acid coding and as antisense promoters with a regulatory function. In this context, synonymous mutations could have a phenotypic impact by affecting the functionality of antisense promoters and consequently the transcript levels of the genes containing them.

### Ribosome Binding Sequences

Translation of mRNA is initiated by the binding of a ribosome to the ribosomal binding site (RBS). Across all bacterial species, the consensus RBS consists of a 6–7 bp motif found 5–10 bp upstream of the start codon and complementary to the 3′ tail of the 16S ribosomal RNA (Shine and Dalgarno 1974). RBSs are relatively short and sequences that are one or two mutations away from the consensus Shine–Dalgarno sequence can be a functional RBS (Omotajo et al. 2015). Intragenic RBSs may promote spurious internal translation initiation leading to the production of frame-shifted or truncated protein (Whitaker et al. 2015), which is expected to have negative fitness effects (Drummond and Wilke 2009). Intragenic RBSs are also known to increase the rate of ribosomal frameshifting during translation elongation. In some cases, this has been shown to be "programmed frameshifting" allowing the production of two different functional proteins from the same coding sequence (Devaraj and Fredrick 2010; Chen et al. 2014). However, cases of spurious ribosomal frameshifting during translation elongation are likely to have negative consequences. In various bacterial species, internal RBSs have also been shown to induce translational pauses by directly binding to the ribosome and thereby reducing the local translation elongation rate (Li et al. 2012; Schrader et al. 2014), leading to a reduction in the quantity of protein produced (Osterman et al. 2020). This slow local translation can have a positive effect on fitness by allowing correct protein folding or down-regulating protein translation (Fluman et al. 2014; Frumkin et al. 2017), or a negative effect if this down-regulation is maladaptive. Like promoter sequences, RBSs also have a high probability of occurring by chance in coding sequences, given their small size. It is difficult to predict whether these motifs will be favored or disfavored by selection because of the diversity of mechanistic and fitness consequences intragenic RBSs can have. The vast majority of prokaryotic protein coding sequences are depleted of internal RBSs (Itzkovitz et al. 2010; Diwan and Agashe 2016). Using a comparative approach, Hockenberry et al. (2018) showed that strong intragenic RBSs detected in *E. coli* present a low level of conservation across *Enterobacteriales* and that sequences downstream of internal RBSs are strongly depleted of ATG start codons. Both observations suggest a negative effect of the presence of these sequences. The general

observations have been made. References: [1]Hahn (2003), [2]Lamberte et al. (2017), [3]Yona et al. (2018), [4]Ando et al. (2014), [5]Kershner et al. (2016), [6]Urtecho et al. (2020), [7]Zhou et al. (2018), [8]Savisaar and Hurst (2017), [9]Sterne-Weiler et al. (2011), [10]Mueller et al. (2015), [11]Itzkovitz et al. (2010), [12]Diwan and Agashe (2016), [13]Schrader et al. (2014), [14]Li et al. (2012), [15]Osterman et al. (2020), [16]Eyre-Walker (1996), [17]Johnson et al. (2011), [18]Morgens et al. (2013), [19]Tse et al. (2010), [20]Abrahams and Hurst (2018), and [21]Bertrand et al. (2015). (B) Observed avoidance or enrichment of sequence motifs targeted by antiviral immune systems and potential phenotypic effects. Different types of antiviral immune systems are considered (colored boxes from left to right: bacterial R-M systems [Rease-MTase]; mammalian apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3 [APOBEC3] mediated innate immunity; eukaryotic antiviral pathways targeting CpG or UpA dinucleotides of which the zinc-finger antiviral protein [ZAP] is known to act in vertebrates but for plants the molecular pathways are yet to be elucidated). Green checks indicate if there is evidence in the literature for avoidance or enrichment of particular sequence motifs, if the presence or absence of these sequence motifs has observable phenotypic effects and if these phenotypic effects can be modified through synonymous variation. The bottom rows indicate in which host groups observations have been made in their infecting viruses or in the host genome itself. References : [22]Sharp (1986), [23]Karlin et al. (1992), [24]Rocha (2001), [25]Rusinov et al. (2018), [26]Pleška et al. (2016), [27]Pleška and Guet (2017), [28]Gelfand and Koonin (1997), [29]Warren et al. (2015), [30]Poulain et al. (2020), [31]Martinez et al. (2019), [32]Chen and MacCarthy (2017), [33]Verhalen et al. (2016), [34]Monajemi et al. (2014), [35]Armitage et al. (2012), [36]Sato et al. (2014), [37]Chen et al. (2014), [38]Simmonds et al. (2013), [39]Ibrahim et al. (2019), [40]Xia (2020), [41]Burns et al. (2009), [42]Gaunt et al. (2016), [43]Takata et al. (2017), [44]Fros et al. (2017), [45]Trus et al. (2020), and [46]Burge et al. (1992).



Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

pattern emerging from these data is a pattern of selection against intragenic RBSs although they may be favored by local selection when their regulatory effect on protein elongation is beneficial. Regardless of the direction of selection on intragenic RBSs, these selective pressures have the potential to impact local codon usage (Li et al. 2012).

### Overlapping and Near-Overlapping Genes

Overlapping genes are widespread in bacterial genomes because of their high gene density: a study analyzing 699 bacterial species revealed more than 90% have at least one overlapping gene pair (OGP), while some genomes harbor up to 3,000 OGPs (Ahnert et al. 2008). Additionally, a high proportion of codirectional gene pairs are "near-OGPs" with less than 40 bps between the two genes (Pallejà et al. 2009). As a consequence, the upstream gene sequence provides both the code for its own amino acid sequence and the promoter and RBS of the downstream gene. For OGPs, the 3′ end of the upstream gene also codes for the amino acid sequence of the downstream gene (Huvet and Stumpf 2014). The double role of these regions constrains the codon usage and partially explains why CUB on the end of bacterial genes is often different from the rest of the genome (Eyre-Walker 1996).

### Stop, Near-Stop and Out-of-Frame Stop Codons

Stop-codon usage is under similar global selection pressure as other codons. In particular, a correlation has been established between stop codon use and availability of the corresponding release factor (Korkmaz et al. 2014). Stop-codon usage is additionally under specific selection pressure in many upstream genes of OGPs in prokaryotes; which often share 1 or 4 bp with the downstream gene, resulting in the overlap of the upstream gene stop codon with the downstream gene ATG start codon. This overlap restricts the choice for stop codons and favors the use of TGA (Eyre-Walker 1996).

Some amino-acid coding codons, called *near-stop codons*, have only one nucleotide difference from stop codons. Near-stop codons can lead to processivity errors when mutations or transcription/translation errors occur (Freistroffer et al. 2000). As processivity errors lead to the production of truncated proteins, they are costly, particularly if they occur late in translation. Selection is predicted to disfavor near-stop codons within coding regions, with a gradual increase in selection pressure along the coding sequence. To our knowledge, only one study has attempted to test this prediction (Johnson et al. 2011), which found evidence for the predicted pattern in coding regions of yeast and humans. Additionally, this selection pressure against near-stops seems to be released in the 30–50 codons upstream of the stop codon. However, certain amino-acids are coded only by near-stop codons, while other amino-acids can be coded by both near-stop and nonnear-stop codons. This result should therefore be regarded with some caution because no correction was made for amino-acid usage. If the hypothesis were verified across species, this would indicate that avoidance of near-stop codons partially shapes the CUB for the four amino-acids coded both by near-stop and non near-stop codons (Leucine, Serine, Arginine, and Glycine).

Finally, the ambush hypothesis proposes that selection might favor out-of-frame stop codons in coding regions, allowing translation to be rapidly aborted when ribosomal frame-shifts occur, thereby reducing the cost of producing a long nonfunctional polypeptide (Seligmann and Pollock 2004). Various studies (Singh and Pardasani 2009; Tse et al. 2010; Bertrand et al. 2015; Abrahams and Hurst 2018) have tried to test the ambush hypothesis, but disagree on the interpretation of the analysis performed and no general conclusion has been reached for now. Indeed, a vast majority of the studies detected an enrichment of out-of-frame stop codons in coding sequences but this enrichment is not significantly more pronounced than the enrichment in other out of frame codons (Morgens et al. 2013). If out-of-frame stop codons are indeed enriched in coding regions, this will have an impact on the specific in-frame codons used.

### Transcription Termination Sequences

Transcription termination signals may play an important role in shaping CUB in eukaryotes. Endonucleolytic cleavage of nascent eukaryotic mRNAs is followed by synthesis of the polyadenosine (poly(A)) tail at specific *cis*-acting polyadenylation sites. These sites, called poly(A) signals, are generally highly conserved AU-rich motifs, mutations in which lead to defects in mRNA processing (Tian and Manley 2017). Using the eukaryotic model organism *Neurospora crassa*, Zhou et al. (2018) demonstrated experimentally that rare codons led to premature transcription termination by creating putative poly(A) sequences. This is because there is a strong preference for C/G nucleotides at the wobble positions of *N. crassa* codons, so genes with rare codons contain higher A/U frequencies and are more likely to lead to the formation of poly(A) signal motifs. Zhou et al (2018) also showed, using a bioinformatics approach, a similar consequence of rare codon usage in mice. The authors suggest that preferences in codon usage may have coevolved with transcription termination machinery to avoid costly premature termination of transcription in GC-rich eukaryotes.

### Exon Silencing and Exon Enhancer Sequences

In eukaryotic gene expression, transcription is followed by splicing—a process through which nonprotein coding introns are removed from the pre-mRNA, and protein-coding exons are joined to produce a mature mRNA. Splicing is catalyzed by a large RNA–protein complex that recognizes specific sequence motifs in the pre-mRNA, both within introns and

Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

exons (Abramowicz and Gos 2018). Exons contain Exonic Splice Enhancers (ESE) and Exonic Splice Silencers (ESS), which enhance integration into the mature mRNA or silence it, respectively. Disruption of ESE sites can cause the skipping of exons, leading to the production of dysfunctional proteins. Conversely, the creation of new ESS sites can lead to a similar outcome, by skipping previously included exons. Many ESE sites are involved in interactions with RNA-binding proteins (RBPs) and a selective pressure to conserve or avoid RBP motifs has been shown in primates and rodents (Savisaar and Hurst 2017). Interestingly, the strength of selection to conserve ESEs has been linked to effective population size. Wu and Hurst (2015) showed, in a study across 30 different species that mean intron size predicts ESE density, with mean intron size negatively correlating with effective population size. This argument also holds within species, with higher ESE density at genes with larger and more numerous introns.

Perturbation of exon encoded regulatory information has been associated with numerous human pathologies, including cystic fibrosis, Lynch syndrome, breast cancer, muscular dystrophy and haemophilia B (Sterne-Weiler et al. 2011; Savisaar and Hurst 2017). A comparative study (Fairbrother et al. 2004) showed that exon ends, where ESE are located, contain fewer single nucleotide polymorphisms than the central region of exons, and linked this pattern to the highly conserved splicing regulatory information encoded at exon extremities. Additionally, an experimental approach determined that 23% of synonymous mutations across exon 7 of the human *SMN1* gene decrease exon integration into mRNA (Mueller et al. 2015). This suggests that for some genes, splicing signals are encoded over the whole length of the exon. Thus, avoidance and maintenance of splice signals and other nonsplicing-associated RBP motifs could influence codon usage over extensive portions of the coding genome.

## Sequence Motifs Targeted by Antiviral Immune Systems

Viral reproduction depends on their host's cellular machinery because viruses release their genetic material directly into the cytoplasm of host cells where replication, transcription, and translation occur. The genetic material of viruses is thus a direct target for intracellular antiviral immune systems that recognize foreign nucleic acids based on specific sequence motifs, subsequently degrade the viral genetic material, and thus impede viral replication. In response, viruses have evolved sophisticated mechanisms to evade host immune responses such as DNA modification, the production of proteins that inhibit the action of certain restriction systems, the use of unusual bases in their genetic material and virus-encoded methylation (Tock and Dryden 2005; Harris and Dudley 2015). However, to evade immune systems that rely on the recognition of specific sequence motifs, the simplest strategy is to avoid these sequence motifs in their genetic material. Viruses have been shown to effectively evade host immune responses through synonymous mutations that remove target sequence motifs from their genome—while keeping the integrity of their coding sequences (Pleška and Guet 2017; Takata et al. 2017). This mechanism appears to be widespread, and the following sections provide an overview of the avoidance of sequence motifs in viral genomes that can be recognized by different antiviral immune systems.

### Recognition Sites for Restriction–Modification Systems

Bacterial restriction–modification (R-M) systems target recognition sites on double stranded DNA molecules that are generally composed of a 4–8 bp palindromic sequence. R-M systems consist of two enzymes: a restriction endonuclease (REase) and a methyltransferase (MTase). The REase cleaves the DNA at the recognition site, creating a double strand break. During bacterial DNA replication, the MTase methylates cytosine and adenine bases at the same recognition site, protecting it from cleavage by the REase. Through the combined action of the MTase and the REase, R-M systems can discriminate between host and foreign DNA containing recognition sites, and consequently cleave only the foreign DNA (Tock and Dryden 2005).

The biological consequences of recognition sites have been widely studied in bacteriophages, because they are the primary target of REases. The increasing availability of phage genomes from the 1980s onward has allowed testing for the avoidance of recognition sites that could be cleaved by the REases of their hosts (e.g., Sharp 1986; Karlin et al. 1992; Blaisdell et al. 1996; Rocha 2001; Rusinov et al. 2018). Indeed, in many phages, there seems to be selection for eliminating recognition sites that could be targeted by their host, resulting in a significant avoidance of these motifs (Sharp 1986). However, this strategy of avoiding host immune defences does not seem to be universal among phages, and three general factors have been identified that influence the occurrence of recognition site avoidance. First, recognition site avoidance is strongly dependent on the genetic material of the phage: dsDNA and ssDNA phages often avoid recognition site motifs, while RNA phages do not (Rocha 2001; Rusinov et al. 2018). This pattern is expected, as RNA phages are not targeted by REases, which only act on double stranded DNA. Although ssDNA phages are also resistant to restriction during their infective stage, they go through a double stranded stage during replication within the host, providing a window for REase attack and thus for selection to act against recognition site motifs. Second, the occurrence of restriction site avoidance depends on the type of R-M system: avoidance is often observed for recognition sites targeted by orthodox Type II R-M systems, but usually not for recognition sites of Type I and Type III R-M systems (Sharp 1986; Rusinov et al. 2018). There





Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

are several explanations for this observation. In Type II systems, the REase and the MTase are independent enzymes with separate DNA recognition domains, while Type I and Type III systems function as hetero-oligomeric complexes with a single sequence recognition domain (Tock and Dryden 2005). Sharing of recognition domains between R and M factors makes it easier to change the specificity of Type I and Type III systems than that of Type II systems. This instigates a phage-bacteria arms-race with rapid changes in the specificity of host defence, rendering recognition site avoidance a less efficient strategy for long-term avoidance of host immune defence using Type I or Type III R-M systems (Rusinov et al. 2018). Several phages are also known to produce universal antirestriction proteins that can inhibit the action of Type I or Type III R-M systems, and are thus protected against restriction even when recognition sites are present in their genome (e.g., SAMase in phage T3, Karlin et al. 1992). Due to the high diversity in Type II R-M systems, such a universal defence could be more difficult to establish (Rusinov et al. 2018). Type I and Type III systems also often require two recognition sites to be present on opposing strands, so avoidance can additionally be achieved by removing a recognition site from only one strand (Tock and Dryden 2005). Third, bacteriophage lifestyle also seems to be a determining factor for the strength of selection against recognition sites, with lytic phages showing a higher degree of recognition site avoidance than temperate phages (Sharp 1986; Karlin et al. 1992; Rocha 2001; Rusinov et al. 2018), probably because temperate phages integrate into the genome of the host where their DNA will be methylated and thereby escape restriction.

Pleška and Guet (2017) provided direct experimental support for the phenotypic effect of synonymous mutation through recognition site changes in bacteriophage $\lambda$ cI857, a conditionally lytic phage of *E. coli*. This phage contains five EcoRI restriction sites, into which synonymous mutations were introduced. They observed that all individual synonymous point mutations increased the likelihood of phage escape, although at a variable rate. The combination of five synonymous mutations, one in each restriction site, provided full escape from restriction by EcoRI. These experimental data represent direct evidence for strong phenotypic effects of synonymous mutations located in a restriction site.

Although the genomes of bacteria encoding R-M systems are assumed to be protected from self-restriction through methylation of recognition sites, several studies have found that many bacterial genomes also show significant recognition site avoidance (Karlin et al. 1992; Gelfand and Koonin 1997; Rocha 2001; Rusinov et al. 2018). This indicates that there is a substantial selective pressure on bacterial genomes to avoid recognition sites and prevent self-restriction. For example, the EcoRI recognition site is avoided in the *E. coli* genome (Gelfand and Koonin 1997). Pleška et al. (2016) experimentally demonstrated that the genomic DNA of *E. coli* is frequently cleaved by EcoRI, and this might be caused by differences in expression levels of the REase and MTase. By comparing the probability of escaping restriction and levels of selfrestriction by two restriction enzymes, Pleška et al. (2016) suggested a trade-off between the efficiency of defence against phages and selfrestriction, which can be mitigated by restriction site avoidance in the host genome.

### APOBEC3 Hotspots

APOBEC3 (apolipoprotein B mRNA-editing enzyme, catalytic subunit 3 or A3) enzymes belong to a family of mutagenic cytidine deaminases that transform cytidine to uracil in DNA or RNA. A3s participate in mammalian innate immunity against retrotransposons, exogenous viruses and endogenous viruses, in which they induce mutations that restrict their replication (Harris and Dudley 2015). A3s have a specific preferred deamination context, called a deamination "hotspot." For example, the 5′TC motif is a hotspot for A3B, while 5′CCC is a hotspot for A3G. Preferred motifs of a particular APOBEC can be changed through a small number of amino-acid changes in the hotspot recognition loop (Kohli et al. 2009), and the expanded A3 gene repertoire in mammals is assumed to be the result of gene duplication and diversification of preferred motifs in response to selective pressures from various viral infections (Münk et al. 2012).

The antiviral action of A3s has been found to exert a mutational and selective pressure on many viral genomes. Recent studies indicated an elevated C to U mutation rate in SARS-CoV2, which can be attributed to the action of A3 (Di Giorgio et al. 2020; Ratcliff and Simmonds 2021; Rice et al. 2021). Viral genomes also often exhibit a depletion of A3 hotspots (Warren et al. 2015; Chen and MacCarthy 2017; Martinez et al. 2019; Poulain et al. 2020). Such a depletion has been recorded in as many as 22% of all human viruses, and is most striking for 5′TC motifs that occupy the second and third position in a codon, where a deamination of the third codon position is always synonymous (Poulain et al. 2020). Furthermore, a high genomic GC content also provides protection against A3s because it tends to minimize the presence of hotspots (Chen and MacCarthy 2017). However, a complete avoidance of A3 hotspots is generally difficult to obtain, because it often requires multiple nonsynonymous mutations that would be detrimental to the virus (Martinez et al. 2019).

Depletion of A3 hotspots is only apparent in certain viral families, with members of the papillomaviruses, polyomaviruses, coronaviruses, and autonomous parvoviruses showing the strongest depletion (Verhalen et al. 2016; Warren et al. 2015; Poulain et al. 2020). This pattern could be caused by a higher A3 pressure on these viral families, either because they infect cell types with higher A3 expression levels, because they induce A3 expression in their host, or because they lack

Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

proteins that inhibit A3 activity (Warren et al. 2015; Verhalen et al. 2016). HIV, for example, is highly susceptible to A3G, but can effectively avoid deamination by the production of the vif protein that neutralizes A3G, reducing the need for A3G motif avoidance (Harris and Dudley 2015).

Although the action of A3-induced hypermutation is expected to have predominantly inactivating effects on HIV-1 (Armitage et al. 2012), some studies found evidence that during early infection HIV-1 can sometimes benefit from A3-induced hypermutation (Wood et al. 2009; Monajemi et al. 2014; Sato et al. 2014). This benefit is caused by accelerated evolution and diversification of positions targeted by the adaptive immune system, allowing for a quick evasion from the initial immune response. There are indications for positive selection on several codon sites within A3 hotspots of the envelope gene of HIV-1 that diversify during the early stages of infection (Wood et al. 2009). Sato et al. (2014) furthermore experimentally showed that in HIV-1 vif mutants, the action of A3D/F can promote in vivo viral diversification leading to a conversion of coreceptor usage. It has been hypothesized that this could explain an observed enrichment of A3 hotspots in cytotoxic T-cell epitope encoding portions of the HIV genome (Monajemi et al. 2014), but it remains unclear how selection for deaminated hotspots during early infection is counteracted by selection for unmodified hotspots during viral transmission.

### CpG and UpA Dinucleotides

Frequencies of CpG and UpA dinucleotides are often significantly depleted in both vertebrate and plant RNA viruses (Cheng et al. 2013; Simmonds et al. 2013; Ibrahim et al. 2019; Xia 2020). This depletion can be partially caused by the viral genome mirroring the nucleotide composition of the host mRNA, which avoids CpG and UpA for reasons other than interactions with antiviral immune systems (Beutler et al. 1989). However, experimental evidence suggests that plant- and vertebrate RNA viruses are additionally subjected to a selective pressure for CpG and UpA avoidance imposed by the host's antiviral immunity. Artificially increasing CpG and UpA dinucleotides, through synonymous mutations in protein coding genes or mutations in noncoding regions, was shown to strongly decrease replication in a large variety of viruses such as poliovirus (Burns et al. 2009), Influenza A (Gaunt et al. 2016), HIV-1 (Takata et al. 2017), the human enteric echovirus 7 (Fros et al. 2017), the potato virus Y (Ibrahim et al. 2019), and Zika virus (Trus et al. 2020). Fros et al. (2017) furthermore inferred that this effect was not caused by a lower translation efficiency due to changes in codon usage, thus suggesting the action of an intrinsic defence pathway present in the host cells acting on CpG and UpA dinucleotides. Takata et al. (2017) partially confirmed this by showing that the zinc-finger antiviral protein (ZAP) is involved in inhibiting virion production through targeting CpG dinucleotides in the RNA of HIV-1. Based on these findings, Xia (2020) proposed that the extreme CpG deficiency in SARS-CoV-2 could contribute to its high virulence in humans by allowing it to successfully avoid ZAP-mediated antiviral immunity. The immune pathways targeting CpG and UpA dinucleotides of plant viruses have not been elucidated, but analogous processes to those in vertebrates might also operate in plants (Ibrahim et al. 2019).

## Conclusions and Perspectives

We have reviewed a number of biological mechanisms that are likely to exert selection pressure on local codon usage for reasons other than selection for translation accuracy and efficiency. In the light of these different elements, selection on codon usage appears to be a combination of a global selection pressure imposed by the translation machinery, and a patchwork of local selection pressures linked to the enrichment or avoidance of specific nucleotide sequences that contain biological signals. However, contrary to the translational selection, the local, nontranslational selection pressures do not apply to all genomes, as some are specific to viruses or to prokaryotes (see fig. 1 for an overview). It is also important to realize that some sequence patterns could be subject to multiple selection pressures. For example, a palindromic sequence could be under selection both because it is the preferred insertion site for certain Transposable Elements (TEs) (see Box 1) and also because it is a restriction site. Specific selection pressures can therefore not be simply deduced by finding that a specific pattern is avoided or enriched in a genome, or a part of the genome. Knowledge of the evolutionary history of the species is generally necessary to make inferences about selective pressures (e.g., associations with specific TEs, specific restriction enzymes encoded and levels of selfrestriction). Additionally, for most mechanisms reviewed (except R-M motifs and CpG/UpA motifs), there are reports of both avoidance and enrichment of the same motif or of positive and negative effects on fitness of the addition or removal of these motifs. In these cases, the direction of selection is determined by factors that range from environmental conditions to surrounding sequences. Testing for avoidance or enrichment at a scale at which both might occur can lead to negative results or to errors in the estimation of the strength of the selection pressure. Finally, for all motifs, avoidance or enrichment patterns can be obtained through both synonymous and nonsynonymous mutations, but synonymous mutations are generally expected to have lower direct fitness effects and for this reason represent a priori a preferred way of avoiding or enriching specific patterns. Yet, when an avoidance or enrichment is observed, it cannot be excluded that nonsynonymous mutations contributed to this pattern.

Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

From a methodological point of view, the detection of over- or under-representation of a particular sequence motif in a genome is often not a trivial task, and is an important issue in computational biology. This detection requires an appropriate model of the genome that assumes the absence of a selective pressure on the sequence motif to which observed frequencies can be compared. A wide range of methods have been developed for this task, including simple estimations using the product of nucleotide or k-mer frequencies and approaches using Markov models (see e.g., Rusinov et al. 2018 for a comparison of methods). Given these methodological difficulties, several authors have noted that some observations of sequence motif avoidance or enrichment are inconclusive and can be artifacts of an erroneous methodology (Sharp 1986; Morgens et al. 2013). It is also a well-known problem that the inference of selection on codon usage by comparative sequence analysis can be confounded by mutational bias, as both processes can produce similar motif enrichment/avoidance and codon usage patterns (Laurin-Lemay et al. 2018). Mutation biases can affect codon usage on both a genome-wide and a local scale (Duret 2002). Disentangling the effects of selection and mutational bias on codon usage is thus not an easy task, and is still a subject of much debate (Galtier et al. 2018; Laurin-Lemay et al. 2018). Along the same lines, inference of selection on codon usage can be erroneous because factors such as amino acid usage bias or gene expression are not considered. For example, it was assumed that translational inefficient codons are selected at the 5′ end of bacterial signal peptides because they can facilitate protein secretion (Power et al. 2004). However, Cope et al. (2018) refuted this hypothesis by showing that the 5′ end of bacterial signal peptides show no differences in CUB compared to cytoplasmic proteins after correcting for amino acid usage and gene expression. In the studies cited in the present review, selection is often inferred based on deviations from genome-wide nucleotide or k-mer frequencies. However, these generally do not account for context-dependent mutational biases or amino acid usage (although see e.g., Wood et al. 2009 accounting for mutational hotspots). The usage of more elaborate models accounting for multiple confounding factors could thus nuance the assumption of selection when observing avoidance or enrichment of a particular sequence motif. Ideally, the fitness effects of synonymous mutations are empirically determined to provide unequivocal evidence for selective pressures on these synonymous positions (Pleška and Guet 2017).

Patterns of avoidance or enrichment in specific motifs or codons are thus not necessarily the product of selection. Conversely, the existence of selection for or against a motif does not necessarily result in the enrichment or avoidance of this motif because it depends on the selection coefficients and the effective population size. For translational selection, selection coefficients on synonymous mutations are generally assumed to be weak (Sharp and Li 1986) and translational selection is only expected to shape codon usage when the effective population size is large enough so that selection can overcome drift, as stated by the nearly neutral theory (Ohta and Gillespie 1996). Consequently, translational selection is assumed to shape the codon usage of species with large effective population sizes, such as many microorganisms and some invertebrate animals, but not (or less) in species with a small effective population size such as larger mammals (Galtier et al. 2018). For nontranslational selection on codon usage, selection coefficients are generally unknown, but they probably vary widely between selective pressures and synonymous sites (e.g., selection against near-stop codons might be weak while selection on avoiding sequence motifs targeted by antiviral immune systems might be stronger). To estimate the potential impact of nontranslational selective pressures on the codon usage of a particular species, both the selection coefficient acting on synonymous variation and the effective population size of the species will need to be considered. However, sometimes extrapolation might not be so straightforward as selection coefficients on synonymous variation might be indirectly affected by the effective population size (Wu and Hurst 2015). Future studies investigating the importance of nontranslational selective pressures for shaping codon usage in a wide variety of organisms will be of particular interest to address this issue.

Selection on codon usage thus appears as a complex phenomenon composed of a mix of global and local pressures. The local pressures are both diverse and specific to certain genome groups, the level of evidence of their existence also varies and it is very likely that some "other codes" of DNA have yet to be uncovered. For example, all the elements for selection against or for the presence of preferred target sequences for TEs are present (see Box 1), but to our knowledge, these patterns and the potential effects on selection and evolution of local codon usage have not yet been investigated. To get a complete and accurate picture of the patchwork of local selective pressures on codon usage and its evolution, more work is required to rigorously identify their molecular signature, to experimentally measure the fitness effects of synonymous mutations in the identified patterns, and to test new hypotheses.

Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

## Text box 1.

### Is Local Codon Usage Influenced by Transposable Elements?

Transposable elements (TEs) are DNA sequences that have the ability to change their position (i.e., to transpose) within or between genomes. TEs are widely spread across all eukaryotic and prokaryotic genomes, and their effects on genome structure and organism fitness are manifold (see Bourque et al. 2018 for a review): 1) TEs increase genome size by accumulating in genomes (Naville et al. 2019). 2) They create new recombination sites and thereby induce chromosome rearrangements (Lönnig and Saedler 2002). 3) They enhance the expression of genes, for example, by introducing new *cis*-regulatory elements in their neighborhood (Salces-Ortiz et al. 2020). (iv) They are a source of novel mutations: either by disrupting the expression of the genes they integrate into, or by introducing new genes (Jangam et al. 2017). Thus, the phenotypic changes induced by TEs range from adaptive (Salces-Ortiz et al. 2020) to lethal (Tsugeki et al. 1996). The sign and amplitude of the fitness effect depends mainly on the TE content and on its insertion site.

Many TE families show strong preferences for their insertion sites (Levin and Moran 2011), but some have dispersed integration patterns, and exhibit low or no preference, for example, ~500,000 copies of the L1 retroelement can be found throughout the human genome. For TEs showing an integration site preference, a precise nucleotide pattern is often required, for example the conserved 60 bp attnTn7 sequence required for the integration of Tn7 in bacterial chromosomes (Kuduvalli 2001; Parks and Peters 2007). The preferred integration site can also be a shorter, less conserved palindromic sequence, as for example the 6 bp motif where Tn10 preferentially inserts (Halling and Kleckner 1982). Other TE families show preferences for certain parts of the genome: some integrate in gene-rich regions but avoid coding regions, for example, *Drosophila* P element often integrates 500 bp upwards of transcription start sites (Bellen et al. 2011) and others integrate specifically in heterochromatin and other weakly expressed regions, for example, in *Saccharomyces cerevisiae*, 90% of Ty5 integration events occur in heterochromatin at telomeres (Zou and Voytas 1997). In many cases, the likelihood of transposition to a site mostly depends on DNA mechanical properties: namely DNA deformability, curvature, and melting (see Arinkin et al. 2019 for a review). Unwinding and bending of DNA allows precise cleavage of the target site, and renders integration irreversible (Morris et al. 2016; Ru et al. 2018). DNA melting allows the conjugative transposons to easily recombine with many insertion sites regardless of homology (Rubio-Cosials et al. 2018). Even when recognition by the transposase requires a few precise invariant base pairs (e.g., several DDE transposases require invariant T/A nucleotides in the sequence in order to integrate), DNA helix flexibility may be necessary to allow recognition and integration through base-flipping and formation of a base-specific contact zone with the transposase (Morris et al. 2016). Structural properties of DNA directly depend on sequence composition. GC content decreases thermostability and bendability but increases DNA curvature (Vinogradov 2003). The deformability of TE integration sites is suggested to be linked to their palindromicity, to their enrichment in T/A pairs (Arinkin et al. 2019) and in pyrimidine-purine base steps (Maskell et al. 2015; Morris et al. 2016).

The codon usage of transposable elements and the evolutionary forces shaping it have been investigated and debated (Lerat et al. 2002; Jia and Xue 2009; Southworth et al. 2019). It is also well established that the observed distribution of TEs in genomes is the result of both TE integration preferences and selection against the integration of TEs at certain loci (Sultana et al. 2017). However, to our knowledge, selection pressure on DNA motifs preferred for TE insertion, the resulting avoidance or enrichment and the potential impact on local codon usage has not been studied. However, by combining knowledge on TE insertion fitness effects and on the nature of preferred insertion sites, predictions can be derived. Local codon usage is likely to be a determinant of the local abundance of TE integration sites, either because synonymous versions of local sequences differ in their content of sequence-specific integration sites or palindromes, or because nucleotide sequence determines DNA mechanical properties (Olson et al. 1998) which favor or disfavor TE integration. Synonymous polymorphisms that increase the likelihood of TE integration will be less fit and purged from the population. This would give rise to a local codon usage preference that reduces the number of insertion motifs in coding regions. This evolutionary scenario should be most prevalent when fitness is highly correlated with gene expression, that is, in organisms with few redundant genes and/or a fast life cycle, and this selection for avoidance of integration sites should also be stronger for essential genes.

TE insertions can also have positive fitness effects, as adaptation to novel environments can be achieved by loss-of-function mutations, particularly in bacteria (reviewed in Hottes et al. 2013). In fluctuating environments, it might be advantageous to have the capacity to remobilize previously lost gene functions. In this context, we could imagine that gene expression could switch between "off" and "on" states through the integration/excision of nonreplicative TEs





Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

(e.g., via cut-and-paste transposition mechanism). Local codon usage preference could thus be under selection to increase the likelihood of transposon integration in these genes. Both predictions for enrichment and avoidance of TE integration sites can be tested by comparing the frequency of TE integration sites in different gene categories. Predictions for enrichment can additionally be tested by analyzing whole genome sequencing data from experimental evolution studies involving stressful conditionsfluctuating over an extended period.

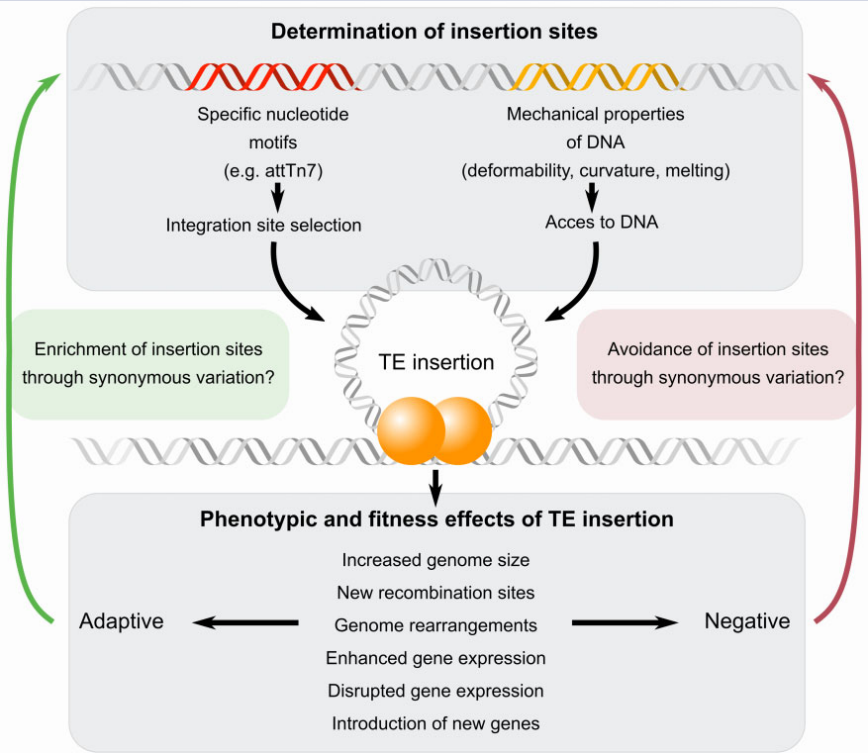


FIG. —How could transposable elements exert local selection pressures on codon usage?

## Acknowledgments

This work was supported by an ERC grant (HGTCODONUSE grant number 682819) to S.B.

## Data Availability

This paper does not include new data.

## Literature Cited

Abrahams L, Hurst LD. 2018. Refining the ambush hypothesis: evidence that GC- and AT-rich bacteria employ different frameshift defence strategies. Genome Biol Evol. 10:1153–1173.

Abramowicz A, Gos M. 2018. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet. 59:253–268.

Ahnert SE, Fink TMA, Zinovyev A. 2008. How much non-coding DNA do eukaryotes require? J Theor Biol. 252:587–592.

Ando H, Miyoshi-Akiyama T, Watanabe S, Kirikae T. 2014. A silent mutation in *mabA* confers isoniazid resistance on *Mycobacterium tuberculosis*: *mabA* mutation confers INH resistance on *Mtb*. Mol Microbiol. 91:538–547.

Arinkin V, Smyshlyaev G, Barabas O. 2019. Jump ahead with a twist: DNA acrobatics drive transposition forward. Curr Opin Struct Biol. 59:168–177.

Armitage AE, et al. 2012. APOBEC3G-induced hypermutation of human immunodeficiency virus type-1 is typically a discrete "all or nothing" phenomenon. PLoS Genet. 8:e1002550.

Bellen HJ, et al. 2011. The *Drosophila* gene disruption project: progress using transposons with distinctive site specificities. Genetics 188:731–743.

Bentele K, Saffert P, Rauscher R, Ignatova Z, Blüthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol. 9:675.



Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

Bergman S, Tuller T. 2020. Widespread non-modular overlapping codes in the coding regions. Phys Biol. 17:031002.
Bertrand RL, Abdel-Hameed M, Sorensen JL. 2015. Limitations of the 'ambush hypothesis' at the single-gene scale: what codon biases are to blame? Mol Genet Genomics. 290(2):493–504.
Beutler E, Gelbart T, Han JH, Koziol JA, Beutler B. 1989. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc Natl Acad Sci U S A. 86:192–196.
Blaisdell BE, Campbell AM, Karlin S. 1996. Similarities and dissimilarities of phage genomes. Proc Natl Acad Sci U S A. 93:5854–5859.
Boël G, et al. 2016. Codon influence on protein expression in *E. coli* correlates with mRNA levels. Nature 529:358–363.
Bourque G, et al. 2018. Ten things you should know about transposable elements. Genome Biol. 19(1):199.
Brantl S. 2007. Regulatory mechanisms employed by cis-encoded antisense RNAs. Curr Opin Microbiol. 10:102–109.
Brophy JAN, Voigt CA. 2016. Antisense transcription as a tool to tune gene expression. Mol Syst Biol. 12:854.
Browning DF, Busby SJW. 2004. The regulation of bacterial transcription initiation. Nat Rev Microbiol. 2:57–65.
Brule CE, Grayhack EJ. 2017. Synonymous codons: choose wisely for expression. Trends Genet. 33:283–297.
Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897–907.
Burge C, Campbell AM, Karlin S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci U S A. 89(4):1358–1362.
Burns CC, et al. 2009. Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. J Virol. 83:9957–9969.
Burow DA, et al. 2018. Attenuated codon optimality contributes to neural-specific mRNA decay in Drosophila. Cell Rep. 24:1704–1712.
Chaney JL, et al. 2017. Widespread position-specific conservation of synonymous rare codons within coding sequences. PLoS Comput Biol. 13:e1005531.
Chen J, et al. 2014. Dynamic pathways of −1 translational frameshifting. Nature 512:328–332.
Chen J, MacCarthy T. 2017. The preferred nucleotide contexts of the AID/APOBEC cytidine deaminases have differential effects when mutating retrotransposon and virus sequences compared to host genes. PLoS Comput. Biol. 13:e1005471.
Cheng X, et al. 2013. CpG usage in RNA viruses: data and hypotheses. PLoS ONE 8:e74109.
Cohen O, et al. 2016. Comparative transcriptomics across the prokaryotic tree of life. Nucleic Acids Res. 44:W46–53.
Coleman JR, et al. 2008. Virus attenuation by genome-scale changes in codon pair bias. Science 320:1784–1787.
Cope AL, Hettich RL, Gilchrist MA. 2018. Quantifying codon usage in signal peptides: gene expression and amino acid usage explain apparent selection for inefficient codons. Biochim Biophys Acta BBA - Biomembr. 1860:2479–2485.
Courel M, et al. 2019. GC content shapes mRNA storage and decay in human cells. eLife 8:e49708.
Devaraj A, Fredrick K. 2010. Short spacing between the Shine-Dalgarno sequence and P codon destabilizes codon-anticodon pairing in the P site to promote +1 programmed frameshifting: ribosomal frameshifting. Mol Microbiol. 78:1500–1509.
Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv. 6:eabb5813.
Diwan GD, Agashe D. 2016. The frequency of internal Shine–Dalgarno-like motifs in prokaryotes. Genome Biol Evol. 8:1722–1733.
Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 10:715–724.
Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 12:640–649.
Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. 16(7):287–289.
Eyre-Walker A. 1996. The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use. J Mol Evol. 42:73–78.
Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. 2:e268.
Fluman N, Navon S, Bibi E, Pilpel Y. 2014. mRNA-programmed translation pauses in the targeting of *E. coli* membrane proteins. eLife 3:e03440.
Freistroffer DV, Kwiatkowski M, Buckingham RH, Ehrenberg M. 2000. The accuracy of codon recognition by polypeptide release factors. Proc Natl Acad Sci U S A. 97:2046–2051.
Fros JJ, et al. 2017. CpG and UpA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. eLife 6:e29112.
Frumkin I, et al. 2017. Gene architectures that minimize cost of gene expression. Mol Cell. 65:142–153.
Galtier N, et al. 2018. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Mol Biol Evol. 35:1092–1103.
Gaunt E, et al. 2016. Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. eLife 5:e12735.
Gelfand MS, Koonin EV. 1997. Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. Nucleic Acids Res. 25:2430–2439.
Gophna U. 2018. The unbearable ease of expression—how avoidance of spurious transcription can shape G+C content in bacterial genomes. FEMS Microbiol Lett. 365:fny26. doi: 10.1093/femsle/fny267.
Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9(1):213–213.
Hahn MW. 2003. The effects of selection against spurious transcription factor binding sites. Mol Biol Evol. 20:901–906.
Halling SM, Kleckner N. 1982. A symmetrical six-base-pair target site sequence determines Tn10 insertion specificity. Cell 28:155–163.
Harris RS, Dudley JP. 2015. APOBECs and virus restriction. Virology 479-480:131–145
Hershberg R, Petrov DA. 2008. Selection on codon bias. Annu Rev Genet. 42:287–299.
Hockenberry AJ, Jewett MC, Amaral LAN, Wilke CO. 2018. Within-gene Shine–Dalgarno sequences are not selected for function. Mol Biol Evol. 35:2487–2498.
Hottes AK, et al. 2013. Bacterial adaptation through loss of function. PLoS Genet. 9:e1003617.
Huvet M, Stumpf MP. 2014. Overlapping genes: a window on gene evolvability. BMC Genomics. 15:721.
Ibrahim A, et al. 2019. A functional investigation of the suppression of CpG and UpA dinucleotide frequencies in plant RNA virus genomes. Sci Rep. 9:18359.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2:13–34.
Itzkovitz S, Alon U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res. 17:405–412.
Itzkovitz S, Hodis E, Segal E. 2010. Overlapping codes within protein-coding sequences. Genome Res. 20:1582–1589.



Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

Jangam D, Feschotte C, Betrán E. 2017. Transposable element domestication as an adaptation to evolutionary conflicts. Trends Genet. 33:817–831.
Jia J, Xue Q. 2009. Codon usage biases of transposable elements and host nuclear genes in *Arabidopsis thaliana* and *Oryza sativa*. Genomics Proteomics Bioinformatics. 7(4):175–184.
Johnson LJ, et al. 2011. Stops making sense: translational trade-offs and stop codon reassignment. BMC Evol Biol. 11:227.
Karlin S, Burge C, Campbell AM. 1992. Statistical analyses of counts and distributions of restriction sites in DNA sequences. Nucleic Acids Res. 20:1363–1370.
Katz L. 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. 13:2042–2051.
Kershner JP, et al. 2016. A synonymous mutation upstream of the gene encoding a weak-link enzyme causes an ultrasensitive response in growth rate. J Bacteriol. 198:2853–2863.
Kinney JB, Murugan A, Callan CG, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci U S A. 107:9158–9163.
Kohli RM, et al. 2009. A portable hot spot recognition loop transfers sequence preferences from APOBEC family members to activation-induced cytidine deaminase. J Biol Chem. 284:22898–22904.
Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. J Biol Chem. 289:30334–30342.
Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. Science 324:255–258.
Kuduvalli PN. 2001. Target DNA structure plays a critical role in Tn7 transposition. EMBO J. 20:924–932.
Kurland CG. 1992. Translational accuracy and the fitness of bacteria. Annu Rev Genet. 26:29–50.
Lamberte LE, et al. 2017. Horizontally acquired AT-rich genes in *Escherichia coli* cause toxicity by sequestering RNA polymerase. Nat Microbiol. 2:16249.
Laurin-Lemay S, Philippe H, Rodrigue N. 2018. Multiple factors confounding phylogenetic detection of selection on codon usage. Mol. Biol. Evol. 35:1463–1472.
Lerat E, Capy P, Biémont C. 2002. Codon usage by transposable elements and their host genes in five species. J Mol Evol. 54:625–637.
Levin HL, Moran JV. 2011. Dynamic interactions between transposable elements and their hosts. Nat Rev Genet. 12:615–627.
Li G-W, Oh E, Weissman JS. 2012. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484:538–541.
Lönnig W-E, Saedler H. 2002. Chromosome rearrangements and transposable elements. Annu Rev Genet. 36:389–410.
Martinez T, Shapiro M, Bhaduri-McIntosh S, MacCarthy T. 2019. Evolutionary effects of the AID/APOBEC family of mutagenic enzymes on human gamma-herpesviruses. Virus Evol. 5:vey040. doi: 10.1093/ve/vey040.
Maskell DP, et al. 2015. Structural basis for retroviral integration into nucleosomes. Nature 523:366–369.
Monajemi M, et al. 2014. Positioning of APOBEC3G/F mutational hotspots in the human immunodeficiency virus genome favors reduced recognition by CD8+ T cells. PLoS One 9:e93428.
Morgens DW, Chang CH, Cavalcanti AR. 2013. Ambushing the ambush hypothesis: predicting and evaluating off-frame codon frequencies in prokaryotic genomes. BMC Genomics. 14:418.
Morris ER, Grey H, McKenzie G, Jones AC, Richardson JM. 2016. A bend, flip and trap mechanism for transposon integration. eLife 5:e15537.
Mueller WF, Larsen LSZ, Garibaldi A, Hatfield GW, Hertel KJ. 2015. The silent sway of splicing by synonymous substitutions. J Biol Chem. 290:27700–27711.
Münk C, Willemsen A, Bravo IG. 2012. An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. BMC Evol Biol. 12:71.
Naville M, et al. 2019. Massive changes of genome size driven by expansions of non-autonomous transposable elements. Curr Biol. 29:1161–1168.e6.
Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. Theor Popul Biol. 49:128–142.
Olson WK, Gorin AA, Lu X-J, Hock LM, Zhurkin VB. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A. 95:11163–11168.
Omotajo D, Tate T, Cho H, Choudhary M. 2015. Distribution and diversity of ribosome binding sites in prokaryotic genomes. BMC Genomics. 16:604.
Osterman IA, et al. 2020. Translation at first sight: the influence of leading codons. Nucleic Acids Res. 48:6931–6942.
Pallejà A, García-Vallvé S, Romeu A. 2009. Adaptation of the short intergenic spacers between co-directional genes to the Shine-Dalgarno motif among prokaryote genomes. BMC Genomics. 10:537.
Parks AR, Peters JE. 2007. Transposon Tn7 is widespread in diverse bacteria and forms genomic islands. J. Bacteriol. 189:2170–2173.
Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol. 20:237–243.
Pleška M, et al. 2016. Bacterial autoimmunity due to a restriction-modification system. Curr Biol 26:404–409.
Pleška M, Guet CC. 2017. Effects of mutations in phage restriction sites during escape from restriction–modification. Biol Lett. 13(12):20170646.
Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 12:32–42.
Poulain F, Lejeune N, Willemart K, Gillet NA. 2020. Footprint of the host restriction factors APOBEC3 on the genome of human viruses. PLoS Pathog. 16:e1008718
Power PM, Jones RA, Beacham IR, Bucholtz C, Jennings MP. 2004. Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of *Escherichia coli*. Biochem Biophys Res Commun. 322:1038–1044.
Radhakrishnan A, et al. 2016. The DEAD-box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. Cell 67:122–132.e9.
Ratcliff J, Simmonds P. 2021. Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. Virology 556:62–72.
Rice AM, et al. 2021. Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design. Mol Biol Evol. 38:67–83.
Rocha EPC. 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 14:2279–2286.
Rocha EPC. 2001. Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. Genome Res. 11:946–958.
Ru H, et al. 2018. DNA melting initiates the RAG catalytic pathway. Nat Struct Mol Biol. 25:732–742.
Rubio-Cosials A, et al. 2018. Transposase-DNA complex structures reveal mechanisms for conjugative transposition of antibiotic resistance. Cell 173:208–220.e20.
Rusinov IS, Ershova AS, Karyagina AS, Spirin SA, Alexeevski AV. 2018. Avoidance of recognition sites of restriction-modification systems is

Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

a widespread but not universal anti-restriction strategy of prokaryotic viruses. BMC Genomics. 19:885.
Salces-Ortiz J, Vargas-Chavez C, Guio L, Rech GE, González J. 2020. Transposable elements contribute to the genomic response to insecticides in *Drosophila melanogaster*. Philos. Trans. R. Soc. B Biol. Sci. 375:20190341.
Sato K, et al. 2014. APOBEC3D and APOBEC3F potently promote HIV-1 diversification and evolution in humanized mouse model. PLoS Pathog. 10:e1004453.
Savisaar R, Hurst LD. 2017. Both maintenance and avoidance of RNA-binding protein interactions constrain coding sequence evolution. Mol Biol Evol. 34:1110–1126.
Schrader JM, et al. 2014. The coding and noncoding architecture of the *Caulobacter crescentus* genome. PLoS Genet. 10:e1004463.
Seligmann H, Pollock DD. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. DNA Cell Biol. 23:701–705.
Shah P, Gilchrist MA. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proc Natl Acad Sci U S A. 108:10231–10236.
Sharp P. 1986. Molecular evolution of bacteriophages: evidence of selection against the recognition sites of host restriction enzymes. Mol Biol Evol. 3:75–83.
Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. 24:28–38.
Shine J, Dalgarno L. 1974. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc Natl Acad Sci U S A. 71:1342–1346.
Simmonds P, Xia W, Baillie J, McKinnon K. 2013. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla –selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics. 14:610.
Singh TR, Pardasani KR. 2009. Ambush hypothesis revisited: evidences for phylogenetic trends. Comput Biol Chem. 33:239–244.
Southworth J, Grace CA, Marron AO, Fatima N, Carr M. 2019. A genomic survey of transposable elements in the choanoflagellate *Salpingoeca rosetta* reveals selection on codon usage. Mob DNA. 10:44.
Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR. 2011. Loss of exon identity is a common mechanism of human inherited disease. Genome Res. 21:1563–1571.
Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol Biol Evol. 24:374–381.
Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet. 18:292–308.
Takata MA, et al. 2017. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature 550:124–127.
Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol. 18:18–30.
Tock MR, Dryden DT. 2005. The biology of restriction and anti-restriction. Curr Opin Microbiol. 8:466–472.
Trus I, et al. 2020. CpG-recoding in zika virus genome causes host-age-dependent attenuation of infection with protection against lethal heterologous challenge in mice. Front Immunol. 10:3077.
Tse H, Cai JJ, Tsoi H-W, Lam EP, Yuen K-Y. 2010. Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. BMC Genomics. 11:491.
Tsugeki R, Kochieva EZ, Fedoroff NV. 1996. A transposon insertion in the Arabidopsis SSR16 gene causes an embryo-defective lethal mutation. Plant J. 10:479–489.
Tuller T, Carmi A, et al. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141:344–354.
Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. 107:3645–3650.
Urtecho G, et al. 2020. Genome-wide functional characterization of *Escherichia coli* promoters and regulatory elements responsible for their function. bioRxiv. doi: 10.1101/2020.01.04.894907.
Verhalen B, Starrett GJ, Harris RS, Jiang M. 2016. Functional upregulation of the DNA cytosine deaminase APOBEC3B by polyomaviruses. J Virol. 90:6379–6386.
Vinogradov AE. 2003. DNA helix: the importance of being GC-rich. Nucleic Acids Res. 31:1838–1844.
Warren CJ, Van Doorslaer K, Pandey A, Espinosa JM, Pyeon D. 2015. Role of the host restriction factor APOBEC3 on papillomavirus evolution. Virus Evol. 1(1):vev015.
Whitaker WR, Lee H, Arkin AP, Dueber JE. 2015. Avoidance of truncated proteins from unintended ribosome binding sites within heterologous protein coding sequences. ACS Synth Biol. 4:249–257.
Wood N, et al. 2009. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. PLoS Pathog. 5:e1000414.
Wu X, Hurst LD. 2015. Why selection might be stronger when populations are small: intron size and density predict within and between-species usage of exonic splice associated *cis*-motifs. Mol Biol Evol. 32:1847–1861.
Xia X. 2020. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. Mol Biol Evol. 37:2699–2705
Yan X, Hoek TA, Vale RD, Tanenbaum ME. 2016. Dynamics of translation of single mRNA molecules in vivo. Cell 165:976–989.
Yang Q, et al. 2019. eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. Nucleic Acids Res. 47:9243–9258.
Yona AH, Alm EJ, Gore J. 2018. Random sequences rapidly evolve into de novo promoters. Nat Commun. 9:1530.
Yu C-H, et al. 2015. Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Mol Cell. 59:744–754.
Zhao F, Yu C-H, Liu Y. 2017. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. Nucleic Acids Res. 45:8484–8492.
Zhou Z, Dang Y, Zhou M, Yuan H, Liu Y. 2018. Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation. eLife 7:e33569. doi: 10.7554/eLife.33569.
Zou S, Voytas DF. 1997. Silent chromatin determines target preference of the Saccharomyces retrotransposon Ty5. Proc Natl Acad Sci U S A. 94:7412–7416.

**Associate editor:** Paul Sharp

Downloaded from https://academic.oup.com/gbe/article/13/9/evab097/6263832 by guest on 15 February 2022

## Abstract :

In front of the antibiotic pressure, a wide range of resistance mechanisms exist. They represent an increasing public health issue, which does and will have major consequences on human mortality. Current strategies —mostly based on antibiotic consumption reductions— globally fail to contain this persistent issue. Moreover, even though it is undeniable that the extensive antibiotic use during the past decades has been a driver of antibiotic resistance evolution, its impact on the emergence and the propagation of antibiotic resistance is often overestimated compared to other factors. First, selection on antibiotic resistance cannot be restricted to the medical and veterinary uses of antibiotics. Second, the evolution (i.e. *the emergence, maintenance* and *spread*) of a trait in populations also relies on other evolutionary forces and ecological processes. Finally, the intertwining of these factors induces an intricate evolution of resistant phenotypes, relying on several scales ranging from genetic determinism to environmental dynamics. The aim of this thesis is therefore, through an eco-evolutionary approach, to improve the understanding of forces and mechanisms shaping directions and movements of antibiotic resistance genes. To do so, this thesis focuses on aminoglycoside resistance by modifying enzymes. The results of this work highlight the importance of ecological and genomic factors in the circulation of resistance genes, and thus conclude that integrative studies of resistance are thus required. To begin with, this thesis reveals that the distribution of aminoglycoside resistance is loosely linked to variation in aminoglycoside consumption across Europe, but strongly structured by globalization (trade, migration) and even more by ecological connectivity. Subsequent analysis underscores the underestimated role of non-pathogenic bacteria in the horizontal transfer of antibiotic resistance genes at large phylogenetic scales. In the light of these results, this thesis opens up a course of reflection on the study and management of antibiotic resistance. **Keywords :** *Antibiotic resistance, Ecology, Evolution, Horizontal gene transfer*

---

## Résumé :

Face à la pression des antibiotiques de nombreux mécanismes de résistances existent. Ces derniers représentent un problème de santée publique qui a et va avoir des conséquences importantes sur la mortalité humaine. Malgré diverses stratégies — principalement basées sur la réduction de la consommation d'antibiotiques — ce problème persistant peine à être contenu. De plus, même s'il est indéniable que l'utilisation massive d'antibiotiques au cours des dernières décennies a été un moteur de l'évolution de la résistance aux antibiotiques, son impact sur l'émergence et la propagation de la résistance aux antibiotiques est souvent surestimé par rapport à d'autres facteurs. Premièrement, la sélection sur la résistance aux antibiotiques ne peut pas être limitée aux antibiotiques à usage humain. Deuxièmement, l'évolution (c'est-à-dire *l'émergence, le maintien* et *la diffusion*) d'un trait dans une population repose également sur d'autres forces évolutives et processus écologiques. Enfin, l'entrelacement de ces facteurs résulte en une évolution complexe du phénotype résistant, dépendante de plusieurs échelles allant du déterminisme génétique aux dynamiques environnementales. L'objectif de cette thèse est donc, en s'appuyant sur une approche éco-évolutive, d'améliorer la compréhension des forces et des mécanismes qui façonnent les directions et le mouvement des gènes de résistance aux antibiotiques. Pour ce faire, cette thèse se concentre sur la résistance aux aminoglycosides par le biais d'enzymes d'altération. Les résultats de ces travaux mettent en évidence l'importance des facteurs écologiques et génomiques dans la circulation des gènes de résistance, et concluent ainsi à la nécessité d'études intégratives de la résistance. Dans un premier temps, cette thèse révèle que la distribution de la résistance aux aminoglycosides n'est que très peu liée aux variations de consommation d'aminoglycosides en Europe, mais fortement structurée par la mondialisation (commerce, migration) et plus encore par la connectivité écologique. Une analyse ultérieure souligne le rôle sous-estimé des bactéries non pathogènes dans le transfert horizontal des gènes de résistance aux antibiotiques à grande échelle phylogénétique. A la lumière de ces résultats, cette thèse propose des pistes de réflexion sur l'étude et la gestion des résistances aux antibiotiques.

**Mots-clés :** *Résistance antibiotique, Écologie, Évolution, Transfert horizontal de gène*

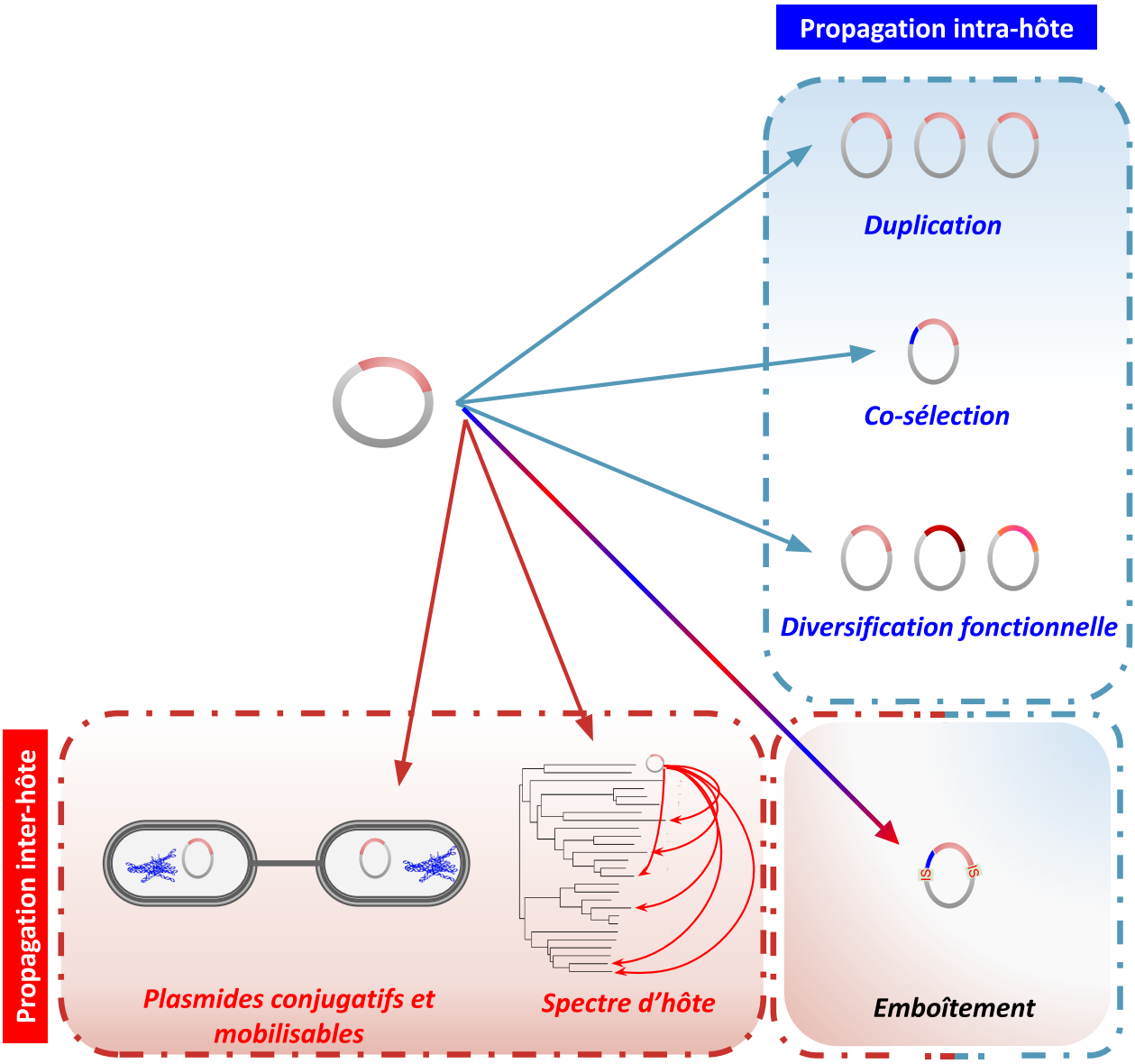


**FIGURE 8 –** Le rôle des éléments génétiques mobiles dans la propagation des gènes de résistance aux antibiotiques.

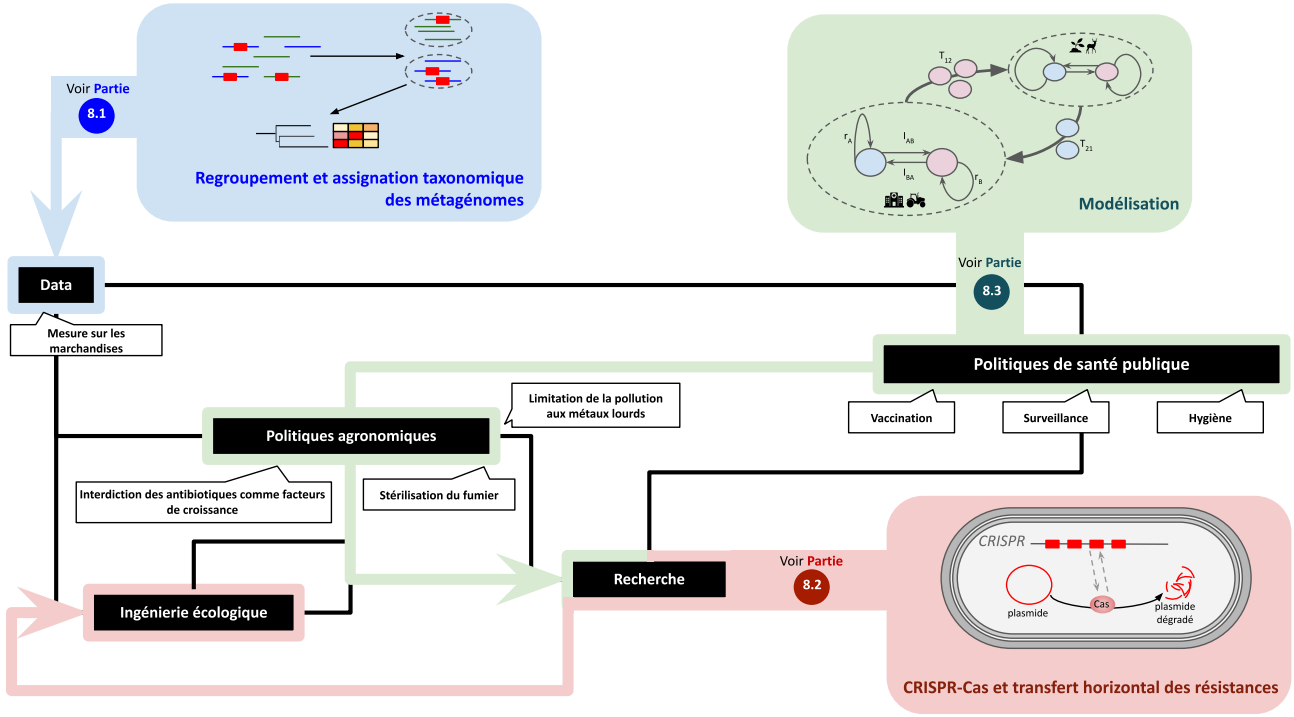


**FIGURE 9 –** Stratégies et perspectives dans la surveillance et la lutte contre les résistances antibiotiques. Les perspectives évoquées dans cette figure sont détaillées dans la section 8.