How capsules protect the cell, shape genetic exchanges and drive bacterial evolution

Matthieu Haudiquet

Université Paris Cité
Institut PASTEUR

École doctorale FIRE 474

Laboratoire Génomique évolutive des microbes, Institut Pasteur

How capsules protect the cell, shape genetic exchanges and drive bacterial evolution

Par Matthieu HAUDIQUET

Thèse de doctorat de Génétique, omiques, bioinformatique et biologie des systèmes

Dirigée par Eduardo ROCHA

Soutenue publiquement le 16 Septembre 2022, devant un jury composé de :

Marianne De Paepe	Rapporteuse	INRAE, MICALIS, Jouy-en-Josas
Vincent Daubin	Rapporteur	CNRS, LBBE, Université Lyon 1
Kathryn Holt	Examinatrice	LSHTM AMR Centre, London
Olivier Dussurget	Examinateur	Université Paris Cité, Paris
Olaya Rendueles	Membre invitée	CNRS & Institut Pasteur, Paris
Eduardo Rocha	Directeur de thèse	CNRS & Institut Pasteur, Paris

A mon père Franck, ma mère Virginie et ma sœur Mathilde Haudiquet.

Le chat

Les amoureux fervents et les savants austères
Aiment également, dans leur mûre saison,
Les chats puissants et doux, orgueil de la maison,
Qui comme eux sont frileux et comme eux sédentaires.

Amis de la science et de la volupté,
Ils cherchent le silence et l'horreur des ténèbres ;
L'Erèbe les eût pris pour ses coursiers funèbres,
S'ils pouvaient au servage incliner leur fierté.

Ils prennent en songeant les nobles attitudes
Des grands sphinx allongés au fond des solitudes,
Qui semblent s'endormir dans un rêve sans fin ;

Leurs reins féconds sont pleins d'étincelles magiques,
Et des parcelles d'or, ainsi qu'un sable fin,
Etoilent vaguement leurs prunelles mystiques.

Charles Baudelaire, *Les Fleurs du mal*

La vie antérieure

J'ai longtemps habité sous de vastes portiques
Que les soleils marins teignaient de mille feux,
Et que leurs grands piliers, droits et majestueux,
Rendaient pareils, le soir, aux grottes basaltiques.

Les houles, en roulant les images des cieux,
Mêlaient d'une façon solennelle et mystique
Les tout-puissants accords de leur riche musique
Aux couleurs du couchant reflété par mes yeux.

C'est là que j'ai vécu dans les voluptés calmes,
Au milieu de l'azur, des vagues, des splendeurs
Et des esclaves nus, tout imprégnés d'odeurs,

Qui me rafraîchissaient le front avec des palmes,
Et dont l'unique soin était d'approfondir
Le secret douloureux qui me faisait languir.

Charles Baudelaire, *Les Fleurs du mal*

Titre : Comment la capsule protège la cellule, façonne les échanges génétiques et influence l'évolution bactérienne.

Résumé :
Les échanges génétiques avec d'autres individus sont à l'origine de l'adaptation de nombreuses bactéries. Ils comprennent l'acquisition de nouveaux gènes et la propagation d'allèles favorables par transfert horizontal de gènes et recombinaison homologue. Ces échanges sont médiés par des éléments génétiques mobiles tels que les plasmides et les phages, et peuvent conduire à l'acquisition de nouvelles fonctionnalités comme la résistance aux antibiotiques ou à l'acquisition de facteurs de virulence. Un élément particulier présent dans à la surface de nombreuses cellules bactériennes - la capsule - affecte les taux d'échange génétique et évolue rapidement à travers eux. La capsule est présente dans tous les pathogènes nosocomiaux multirésistants de haute priorité ESKAPEs, et représente la première structure à traverser pour les virions et les systèmes conjugatifs. Dans cette thèse, j'ai cherché à caractériser l'interaction entre le transfert horizontal de gènes et la capsule bactérienne. J'ai utilisé *Klebsiella pneumoniae* comme modèle, et j'ai combiné plusieurs approches, notamment la génomique comparative et la phylogénie, ainsi que des experience de transfert de gènes et des mutants exprimant different sérotypes. Cette thèse met en lumière l'impact de la composition de l'enveloppe cellulaire sur l'évolution bactérienne, et montre que les capsules agissent comme des gardiens du transfer horizontal de gène.

Mots clefs : Microbiologie, *Klebsiella pneumoniae*, Capsule, Serotype, Génomique, Génomique comparative, Phages, Conjugaison

Title : How capsules protect the cell, shape genetic exchanges and drive bacterial evolution

Abstract :
Genetic exchanges with other individuals drive the adaptation of many bacteria. They include the acquisition of novel genes and the spread of favorable alleles by horizontal gene transfer and homologous recombination. These exchanges are mediated by mobile genetic elements such as plasmids and phages, and can lead to the gain of new functionalities like antibiotic resistance or to the acquisition of virulence factors. One particular element present in many bacterial cells - the capsule - is thought to both affect the rates of genetic exchange and to rapidly evolve through them. The capsule is present in all high-priority multidrug-resistant nosocomial pathogens ESKAPEs, and represents the first structure to cross for virions and conjugative systems. In this thesis, I aimed at characterizing the interplay between horizontal gene transfer and the bacterial capsule. I used Klebsiella pneumoniae as a model, and combined several approaches including comparative genomics and phylogeny, as well as gene transfer assays and serotype swap mutants. Overall, this thesis sheds light on the impact of the cell envelope composition on bacterial evolution, and show that capsules are gatekeepers for mobile genetic elements.

Keywords : Microbiology, Klebsiella pneumoniae, Capsule, Serotype, Genomics, Comparative genomics, Phages, Conjugation

Résumé substantiel en français :

Cette thèse est le résultat de plusieurs années de travail entre 2018 et 2022 à l'Institut Pasteur, dans le laboratoire de génomique évolutive des microbes dirigé par Eduardo Rocha, sous la supervision d'Eduardo et Olaya Rendueles. Il est composé d'une Introduction générale, où je dresse l'état de l'art

des sujets étroitement liés à mes questions, suivie d'une section Contribution avec des introductions spécifiques aux méthodes que j'ai utilisées et aux articles que j'ai (co-)écrits directement liés à mon projet de doctorat. Mes contributions au domaine sont réparties entre deux points de vue, la génomique comparative *in silico* et les expériences *in vitro*. La conclusion présente un aperçu intégré de mes contributions. Enfin, j'ai participé à plusieurs autres articles indirectement liés à mon sujet, mais toujours en rapport avec *Klebsiella* et les échanges génétiques qui sont présentés dans les Annexes.

Les échanges génétiques sont à l'origine de l'adaptation de nombreuses bactéries. Chez la plupart des espèces bactériennes, les échanges génétiques sont assurés par des éléments génétiques mobiles qui peuvent transférer leur propre matériel génétique, ou celui de leur hôte, aux cellules receveuses. Par conséquent, les gènes circulent dans les population, ce qui favorise l'évolution des bactéries, notamment des pathogènes. Deux processus sont associés avec les échanges génétiques : le remplacement de gènes préexistants par des allèles différents par recombinaison homologue, et l'acquisition de nouveaux gènes. Ces échanges peuvent conduire à l'acquisition de nouvelles fonctionnalités, notamment la résistance aux antibiotiques ou la virulence. Les éléments génétiques mobiles comme les phages et les plasmides conjugatifs codent des structures physiques, telles que les virions et les pili conjugatifs, capables de traverser l'enveloppe cellulaire des cellules réceptrices. Tout déterminant restreignant ou donnant accès aux éléments génétiques mobiles peut donc avoir un impact sur le flux génétique au sein d'une population. Une caractéristique essentielle de l'enveloppe de nombreuses espèces bactériennes, la capsule, constitue souvent le premier point de contact des éléments génétiques mobiles avec la cellule.

Les capsules bactériennes sont des couches protectrices contre les stress abiotiques et biotiques. Elles forment une barrière hautement hydratée qui ralentit la dessiccation et exclut les molécules hydrophobes toxiques telles que les détergents. Les capsules ont également été décrites comme protégeant contre la prédation par les protozoaires, et l'infection par les bactériophages en masquant les récepteurs de surface. Enfin, les capsules peuvent améliorer la capacité des bactéries à infecter leurs hôtes et sont donc considérées comme des facteurs de virulence. Le rôle des capsules dans la virulence est multiple, car elles protègent les cellules contre le système immunitaire de diverses manières. Premièrement, elles masquent les antigènes immunogènes présents à la surface de la cellule. Ensuite, l'épaisseur et la charge négative de la capsule préviennent la lyse et la phagocytose, et protègent contre les protéines antimicrobiennes. Enfin, les capsules des agents pathogènes humains sont généralement composées de motifs polysaccharidiques imitant ceux de l'hôte, qui ne peuvent pas être ciblés efficacement par les anticorps. Le rôle protecteur global des capsules contribue à expliquer pourquoi

les espèces bactériennes codant pour des capsules sont associées à une plus grande distribution environnementale, c'est-à-dire qu'elles peuvent être trouvées dans plus d'environnements que les autres espèces. Les capsules évoluent rapidement et présentent une énorme diversité chimique et génétique. La diversification de la composition chimique des capsules se produit par le biais d'un changement génétique dans le locus de la capsule, parfois par mutation, mais souvent par recombinaison avec de l'ADN transféré horizontalement.

L'hypothèse principale motivant cette thése est que les capsules jouent un rôle majeur dans l'adaptation bactérienne. En effet, elles facilitent la survie en protégeant contre les stress biotiques et abiotiques. De plus, elles pourraient moduler les taux d'échange génétique car elles représentent une barrière physique à franchir pour les éléments génétiques mobiles. Cette hypothèse est soutenue par des découvertes récentes selon lesquelles (i) les capsules sont plus fréquentes chez les bactéries environnementales que chez les pathogènes, (ii) les bactéries avec des capsules occupent des niches plus diverses, (iii) les bactéries capsulées ont des pan-génomes plus grands, et (iv) accumulent plus d'éléments génétiques mobiles. Il y a donc une énigme concernant l'évolution des capsules : la capsule a besoin d'éléments génétiques mobiles pour varier par transfert horizontal de gènes, mais peut bloquer l'acquisition de ces mêmes éléments génétiques mobiles. Comment donc l'interaction entre capsules et éléments génétiques mobiles a-t-elle un impact sur l'évolution bactérienne ?

Modèle

Dans cette thèse, j'ai utilisé *Klebsiella pneumoniae* comme modèle pour étudier l'interaction entre les éléments génétiques mobiles et la capsule. *Klebsiella pneumoniae* est une espèce capsulée appartenant à la famille des Enterobacteriaceae, tout comme *Salmonella* et *Escherichia*. C'est une espèce ubiquitaire, et un commensal du tube digestif de nombreux eucaryotes. On peut la trouver dans le sol, l'eau douce, associée aux racines et aux feuilles des plantes, aux insectes et aux animaux, y compris les humains. Associée aux plantes, *K. pneumoniae* est une diazotrophe capable de fixer l'azote atmosphérique et est souvent attachée aux poils absorbants des racines des plantes qui bénéficient de cet azote. Associée aux mammifères comme l'homme, *K. pneumoniae* est un commensal du nez, de la bouche, des poumons et de l'intestin qui fait partie de la flore bactérienne saine. Il s'agit toutefois d'un pathogène opportuniste qui entraîne un large panel de maladies chez les patients immunodéprimés, et fait partie du groupe de pathogènes ESKAPE. Les ESKAPEs sont un groupe d'espèces bactériennes virulentes et multi-résistantes identifié par l'OMS, qui comprend *Enterococcus faecium,*

Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa et *Enterobacter spp*.

Klebsiella pneumoniae est une espèce particulièrement adaptée pour étudier la capsule et les élément génétiques mobiles. En effet, cette espèce acquiert à un rythme rapide des élément génétiques mobiles codant notamment pour des gènes de résistance aux antibiotiques, et des facteurs de virulence. De plus, *Klebsiella pneumoniae* est recouverte d'une capsule polysaccharidique qui présente une large diversité génétique et chimique, comptant certainement plus de 160 sérotypes différents.

Objectifs

L'objectif de cette thèse est de caractériser l'interaction entre la capsule bactérienne et le transfert horizontal de gènes. Plus précisément, je voulais faire la lumière sur l'impact de la capsule sur le flux de gènes entre les populations bactériennes, et comment ce processus peut façonner l'évolution de la capsule. Klebsiella pneumoniae représente un modèle de choix, car elle exprime de manière ubiquitaire une grande capsule, code souvent des MGEs, et son évolution représente une menace pour notre système de santé global. Les objectifs de ma thèse étaient les suivants :

- Déduire les échanges génétiques passés entre isolats séquencés de Klebsiella pneumoniae, en relation avec leur capsule.
- Mesurer les taux d'échanges génétiques chez Klebsiella pneumoniae en relation avec leur capsule.
- Comprendre l'évolution de la capsule à la lumière des interactions avec les éléments génétiques mobiles.

Contributions

Mes travaux de thèse ont donné lieu à la publication d'un article de premier auteur et d'un autre article qui sera bientôt soumis. Ces études reposent sur deux piliers méthodologiques : la génomique comparative et la biologie expérimentale. Dans la première étude, j'ai exploité des outils du domaine de la génomique comparative pour analyser des milliers de génomes de *Klebsiella pneumoniae*, déduire leurs échanges génétiques passés et proposer un modèle d'évolution de la capsule. J'ai complété mes résultats par un modèle expérimental simple qui soutenait les conclusions de l'analyse génomique. Dans la deuxième étude, j'ai utilisé une approche miroir partant de l'ingénierie expérimentale des génomes et de transfert de gènes *in vitro*, complétée par une analyse génomique des plasmides de *Klebsiella pneumoniae*. Par conséquent, mes contributions sur l'interaction entre les capsules bactériennes et le transfert horizontal de gènes seront présentées d'abord sous l'angle de la génomique,

puis sous celui de la biologie expérimentale. J'ai également participé à plusieurs autres études, qui sont présentées dans les annexes.

L'étude du flux de gènes chez *K. pneumoniae* a révélé que le transfert de gènes se produit plus fréquemment entre les souches du même sérotype, ce qui entraîne un biais de flux de gènes intra-sérotypes. Cela peut s'expliquer par des effets écologiques, car les populations ayant des sérotypes similaires ont tendance à occuper des niches similaires et ont plus de possibilités d'échanges génétiques. Ce biais peut également s'expliquer par la parenté génétique, puisque les souches étroitement apparentées ont tendance à échanger du matériel génétique plus fréquemment. On pourrait s'attendre à ce que les effets écologiques et les effets de parenté aient un impact similaire sur les échanges de matériel génétique médiés par la conjugaison et par les phages, l'analyse des flux génétiques liés aux plasmides a révélé qu'elle était en fait biaisée vers les échanges inter-sérotypes. À l'inverse, le flux génétique des phages tempérées sont fortement orientés vers les échanges intra-sérotypes. Dans l'ensemble, le biais intra-sérotype semble être largement déterminé par les phages, éclipsant le biais opposé observé dans les plasmides.

Pourquoi le flux génétique médié par les phages est-il plus élevé entre les souches d'un même sérotype ? Les phages virulents naturels d'espèces capsulées, telles que *K. pneumoniae* ou *Acinetobacter baumannii*, sont généralement spécifiques à un, ou parfois plusieurs, sérotypes. La coévolution de longue date des phages avec leurs hôtes capsulés a donné lieu à une relation écologique particulière : chez les espèces typiquement non-capsulées, leurs phages sont bloqués par l'expression d'une capsule, mais chez les espèces généralement capsulées, leurs phages sont dépendants de la capsule pour infecter leurs hôtes. C’est cette spécificité envers le sérotype capsulaire qui provoque le biais d’échanges génétiques intra-serotype.

Pourquoi le flux génétique médié par la conjugaison est-il plus élevé entre les souches de sérotypes différents ? Au moins deux observations peuvent expliquer ce phénomène : i) l'inactivation de la capsule est fréquente, et ces cellules non capsulées ont des fréquences de conjugaison jusqu'à 100 fois plus élevées et ii) les sérotypes de capsule ont un impact différent sur l'efficacité de la conjugaison. En identifiant d'abord les loci capsulaires inactivés, et en analysant le flux génétique de ces isolats, nous avons identifié les cellules non capsulées comme des hub de conjugaison. Nous avons également fourni des preuves expérimentales que les cellules non encapsulées ont des taux de réception et de don plus élevés, et que ces hubs de conjugaison fournissent des voies de transfert entre sérotypes distincts. En construisant des échanges de sérotypes isogéniques avec une nouvelle méthode, nous avons montré

que l'efficacité de la conjugaison est déterminée par le sérotype de la capsule du donneur et du receveur. Par exemple, les sérotypes K1 et K2 sont associés à des efficacités de conjugaison relativement plus faibles que les sérotypes K24 et K3. De plus, les cellules hébergeant le même sérotype de capsule n'ont pas une fréquence de conjugaison plus élevée, et il semble y avoir peu ou pas d'interaction entre la capsule du donneur et celle du receveur pendant la conjugaison. En conséquence, la conjugaison est généralement plus efficace entre les cellules K1 et K3 qu'entre les cellules K1. Ainsi, les différences d'efficacité de la conjugaison peuvent entraîner un transfert préférentiel entre les cellules de sérotypes différents. Les raisons mécanistiques pour lesquelles les sérotypes sont associés à des efficacités de conjugaison différentes sont encore inconnues. Pour répondre à cette question, les échanges de sérotypes isogéniques seront essentiels pour délimiter les différences précises entre sérotypes, indépendamment du fond génétique. Notre hypothèse principale est que l'épaisseur inhérente au sérotype est en corrélation négative avec l'efficacité de conjugaison, car elle diminue la proximité physique entre les cellules. Cette hypothèse est conforme à l'observation selon laquelle les clones hypervirulents de *K. pneumoniae* ont tendance à présenter une diversité génétique relativement plus faible, car leurs capsules épaisses peuvent permettre au système immunitaire de mieux s'échapper mais réduire les taux de conjugaison.

TABLE OF CONTENT

Remerciements

Pendant l'écriture de cette thèse, j'ai souvent ravivé ma motivation en pensant au moment rêvé où, libéré, je me poserai enfin pour écrire ces remerciements. Quel plaisir de pouvoir les écrire de mon bureau au 6eme étage du bâtiment François Jacob, un mois après ma soutenance.

Merci aux membres de mon jury pour leur travail de relecture, leur intérêt, la pertinence de leurs commentaires et questions. Merci à vous Vincent Daubin, Marianne De Paepe, Kathryn Holt, et Olivier Dussurget. Merci aussi aux membres de mon comite de thèse, Nienke Buddelmeiser et Olivier Tenaillon, pour votre accompagnement, votre guidance et votre bienveillance.

Merci Eduardo pour ton accueil et ta confiance tout au long de ces années. Ce fut un réel plaisir de t'avoir comme directeur de thèse. Merci de m'avoir formé, avec bienveillance et patience, à la Recherche. Merci pour ton soutien, tes conseils et ta justesse. Cela va me manquer de ne plus passer te voir à ton bureau pour te parler de mes résultats, de mes problèmes, et pour te dire « Ça n'a pas marché, mais j'y suis presque ! ». Je te suis reconnaissant pour tout le temps que tu as passé à m'aider, que ce soit sur les manuscrits, les slides, et autres. Et je suis fier de dire que j'ai préparé ma thèse dans ton laboratoire.
Merci Olaya de m'avoir accompagné tout au long de cette aventure, dans les hauts et les bas. Merci pour tes conseils à la paillasse, pour ton aide et tes commentaires pendant l'écriture, et de m'avoir aidé à ne pas (trop) m'éparpiller. Longue vie à nos papiers !

Merci Amandine pour tous ces bons moments de travail mais aussi de rigolade dans notre petit P2 du 5ème étage. Le travail c'est tout de suite plus agréable avec une collègue et une amie sur laquelle on peut compter, et ça aide à relativiser les tracas des manips qui ne marchent pas. Je te souhaite bon courage pour la suite, car sans moi, ce sera forcément moins fun nan ? J'ai envie de résumer ces années à des phrases qu'on se disait à tour de rôle : « Ça marche pas » ; « Un résultat négatif, c'est un résultat en soit, non ? » ; « QUI UTILISE LA CENTRI ?! » ; « C'est de la recherche de toute façon c'est pas sensé marcher ! » ; « Hm, t'utilises la hôte … ? ».
Merci Marie, pour tes centaines de bons conseils sur la génomique, pour ces milliers de cafés partagés sur la terrasse, parfois juste pour se raconter nos vies, ou bien râler sur les IS dans les génomes. Merci pour tes figures scientifiques dignes d'une illustratrice professionnelle, et ta bienveillance. Tu es une source d'inspiration pour moi, et j'espère avoir la chance de retravailler avec toi un jour. Et bien sûr

merci parce que grâce à toi, on a toujours bien rigolé au 6ème étage. Cette thèse n'existerait pas sans ton soutien.

Obrigado à Jorge pour nos discussions scientifiques passionnantes, pour les pastéis de nata, pour ta bienveillance et tes conseils (et pour le wGRR.py !!). Danke Eugen, for all the food (!!), all the laugh, all the phage discussions, and your incredible kindness. Merci mon Charles, un vrai crack de la recherche, mon MOASS buddy, pour tous tes conseils, tes histoires de fou, les barres de rire et tout le reste (les brunchs !!!). Gracias Manuel for the science, the beers, the nights out in bad Parisian bars with the Spanish mafia. Sorry I forgot to thank you during the defense, but you know I was overwelmed at the thought of thanking you. Merci mon Rémi pour ton aide au début de ma thèse quand je ne connaissais rien à la génomique, tu as été le premier à me montrer la voie de SLURM et à me convertir au IQTree-isme. Et merci pour tous ces bons moments. Gracias Neris, for being the most talkative and fun *Klebsiella* enthusiast/hater of all the other students that shared an office with me. Jorge, Eugen, Charles, Manuel, Rémi et Neris cette thèse aurait été tellement moins intéressante et fun sans vous, merci pour tout. Merci aux autres membres du GEM lab qui ont fait ce bout de chemin avec moi, Camille, Amandine P., Fanny W., Martin, Elif, Julien, Bertrand, ce fut un plaisir !
Finalement, merci aux Padawans du labo : Julie, Eloi et Fanny. Merci de m'avoir toujours traité avec le respect qu'il se doit à un ancien, et aussi, finalement, *de rien*, car j'ai été comme un exemple pour vous n'est-ce pas ? Je vous souhaite de réussir et de plus tard, me recruter dans votre labo car vous êtes simplement trop forts.

Merci, merci, merci à mes amis.

Ceux de toujours, OP. Pierre, Thomas R., Gauthier, Thomas V., Antoine, Thibaut. Quelle chance de vous avoir, c'est impossible de vous remercier à votre juste valeur ici. Merci pour les vacances de fou, les Discord, les visites aux quatre coins du monde. Votre amitié est une constante qui me permet d'avancer. Sans vous, pas de Matthieu, pas de thèse. Soyez fier car grâce à vous, le monde sait enfin que la capsule impacte les échanges génétiques entre bactéries.
Ceux de Paris, qui sont comme une seconde famille. Maxime, même si tu as oublié de me remercier à ma juste valeur dans ta thèse, merci pour tout mon frère. Merci pour les Zombies, pour les Mercredidi, les voyages. Merci de me montrer qu'un fou furieux peut devenir chercheur. Grégoire, mon grand Gregory, le plus beau Scout d'Orléans, le babos des Gobelins, le craqueur de cou. Merci de me montrer qu'un joueur de carte Magic qui peint des Warhammers peut devenir chercheur. Merci Nji TBB (artiste, productrice, chercheuse), merci Tom.G (le D du 15), merci Charly (guitariste-chercheur),

merci Sylvain (Dark Sylvain des montagnes), merci Eli (la littéraire du lot), merci Alice (Cat mom) et Robin (Cat dad), merci Egill (surfeur-chercheur), merci Louis (Jazzman-biophysicien), merci Vincent-Alexis (GOB-man extraordinaire).

Merci à Bart et Nilou de m'avoir toujours rappelé que la thèse, c'est beaucoup moins important que plein d'autres trucs comme : sortir dans le jardin à 5h du matin, manger des croquettes, manger de la pâtée, manger des friandises, dormir, se battre ou encore rien faire du tout.

Merci à ma famille. Maman, Papa et Mathilde merci d'avoir toujours cru en moi. Papa, merci pour tous les documentaires et les *C'est pas sorcier* qu'on a regardé ensemble, c'est ce qui m'a mis sur cette voie. Maman, merci de m'avoir toujours écouté avec attention même quand je parlais pendant des heures des cours, du lab, de science. Mathilde, « ah t'es là toi », merci de toujours me faire marrer (et buguer aussi parfois) et pour tes œuvres qui apparaissent dans cette thèse. Je vous dédie cette thèse à tous les trois car sans votre soutien, rien de tout cela n'aurait été possible. Maintenant vous n'avez plus qu'à la lire !

Et merci à Léone, la *bad cop* de mon *good cop*, pour tout le bonheur que tu m'apportes. Tu es celle qui a payé le prix le plus cher pendant cette thèse : avoir à m'écouter parler de bactéries depuis l'encadrement de la porte à minuit quand tout ce que tu voulais c'était dormir. Merci pour ton coaching sur Rocket League. Merci pour toutes les fois où tu m'as ranimé malgré les Zombies. Merci pour les GOB partagées, les soirées à refaire le monde, les escapes, les voyages, les cats, les repet' de talk, … la vie quoi. Merci future docteure.

Preamble

This thesis is the results of several years of work between 2018 and 2022 at the Institut Pasteur, in the Microbial Evolutionary Genomics lab led by Eduardo Rocha, under the supervision of Eduardo and Olaya Rendueles. It is composed of a General Introduction, where I draw the state-of-the-art of topics closely related to my questions, followed by a Contribution section with specific introductions to the methods I used and the articles I have (co-)authored directly related to my PhD project. My contributions to the field are split between two views, *in silico* comparative genomics and *in vitro* experiments. The Conclusion presents an integrated overview of my contributions. Finally, I have participated in several other articles indirectly related to my subject, but always related to *Klebsiella* and genetic exchanges that are presented in the Annexes.

Genetic exchanges drive the adaptation of many bacteria. They include the replacement of pre-exiting genes by different alleles via homologous recombination, and the acquisition of new genes. These exchanges are mediated by **mobile genetic elements** such as **plasmids** and **bacteriophages**, and can lead to the gain of new functionalities including antibiotics resistance or virulence. One particular element present in many bacterial cells – the capsule - is thought to both affect the rates of genetic exchange and to rapidly evolve through them.

The capsule forms the outermost layer of the cell in more than half of the sequenced bacterial species, and as such, it is the first point of contact between the cell and its surroundings. Capsules are membrane-bound polysaccharides synthesized by a multi-step pathway encoded in a capsule locus. They diversify at high rates and frequently **evolve** via horizontal gene transfer, suggesting that capsules undergo strong selective pressures. Proposed sources of selection for capsule evolution and diversification are related abiotic stress tolerance, optimal nutrient utilization, host's immune system escape, and protection from predation by protozoa and bacteriophages. Capsule evolution is intense, but depends on horizontal gene transfer mediated by mobile genetic elements, for which they represent a physical barrier to cross. How then does this interaction between capsules and MGEs impact bacterial evolution?

This thesis focuses on ***Klebsiella pneumoniae***, a capsulated diderm bacterium belonging to the Enterobacteriaceae family. It has a large tropism, for instance, it can be found in the soil, associated to plants, but it is best known as an opportunistic pathogen colonizing a wide variety of hosts, including

human, where it causes pneumonia, recurrent urinary infections and acute abscesses in the liver. *Klebsiella pneumoniae* clonal complexes can be associated with hypervirulence, multidrug resistance or, recently, both. It is the 'K' from the ESKAPE pathogens, the six most significant multidrug resistant nosocomial pathogens.

Given the importance of genetic transfer in the evolution and adaptation of microorganisms, and notably in the acquisition of antimicrobial resistance genes, the aim of this thesis is to **understand the interplay between horizontal gene transfer and the capsule during the evolution of *Klebsiella pneumoniae***. This is key to understand the importance of the cell envelope on bacterial evolution, how the evolution of MGE and bacteria are impacted by their physical interaction, and the emergence of multi-drug resistant and/or virulent clones by accumulation of genes through HGT in capsulated species.

General Introduction

Bacteria

Bacteria are single-celled organisms forming one of the three Domains of Life. In this section, I will first briefly describe the bacterial cell, and provide a primer on bacterial genome composition and organization.

Bacterial cell

The bacterial cell is composed of DNA, RNA, proteins, lipids and other molecules involved in the cell's metabolism, *i.e.* all the chemical and biological transformations supporting life (Figure 1). The inner medium is called the cytoplasm, whereas the cell is delimited by the envelope.

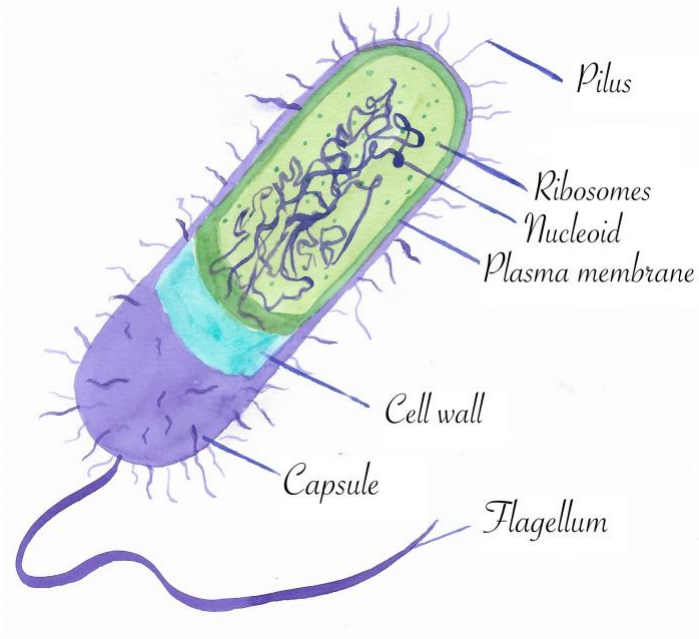


Figure 1 - The bacterial cell. An additional membrane above the cell wall is present in diderm species. (Credit: Mathilde Haudiquet)

Bacteria typically have one chromosome made of a circular stretch of DNA, compacted by specialized histone-like proteins within the **nucleoid**. The nucleoid is often associated with the cytoplasmic membrane of the cell, but is largely viewed as free-floating within the cytoplasm. Chromosome size

varies from a few hundreds of kilobases to more than 14 megabases, depending on the species. Many cells also carry extra-chromosomal DNA elements called plasmids.

Upon transcription of the DNA into messenger RNA, protein synthesis is carried out by **ribosomes** which are molecular machines made of proteins and ribozymes (catalytic RNA) accounting for most of the cell's mass. Proteins localize either in the nucleoid, in the cytoplasm or anchored in the cell's envelope.

The cell envelope regroups the structures surrounding the cytoplasm and delimiting the borders of the cell, including the cytoplasmic membrane, the cell wall, the outer membrane, and additional layers like the S-layer or the capsule. **Cytoplasmic membranes**, sometimes called plasma membranes, are found in all three domains of life and are composed of phospholipids organized in a polar bilayer. They surround the cytoplasm and provide a protection against diffusion of ions and molecules. Numerous proteins are present within the membrane, carrying out various functions such as water diffusion, molecules import and export, or attachment to surfaces. Phospholipid bilayers associated with proteins form selectively permeable membranes. On top of this membrane, the **cell wall** provides structural integrity for the cell and protection from the internal turgor pressure caused by the much higher concentration of molecules inside the cell than outside the cell. The bacterial cell wall is distinguished from that of Archaea and Eukarya by the presence of peptidoglycan composed of poly-N-acetylglucosamine and N-acetylmuramic acid. Peptidoglycan is located immediately outside the cytoplasmic membrane and is responsible for the rigidity of the envelope and shape of the cell, while being porous enough not to impede molecule's diffusion. While almost all bacterial cell walls contain peptidoglycan, not all cell walls have the same organization. This is reflected in particular by the monoderm and diderm classification.

Monoderms have one membrane composed of a phospholipid bilayer, covered in a thick peptidoglycan layer. Notorious members of the monoderms include *Staphylococcus, Enterococcus* and *Streptococcus* species. **Diderms** are composed of a thinner peptidoglycan layer but are surrounded with an additional membrane called the **outer membrane**. This outer membrane is anchored to the peptidoglycan by proteins, and is covered with the lipopolysaccharide (LPS). The LPS is a large molecule composed of the lipid A, core oligosaccharide, and the O-antigen. The O-antigen is the most distal part of the molecule and is composed of various repeats of glycans. It is covalently attached to the core, which is less variable than the O-antigen, itself bonded to the lipid A anchored in the outer membrane. Together, these three components participate to the structural stability of the diderms cell envelope.

Finally, monoderm and diderm bacterial cells are often surrounded by an additional, outermost protective layer called the **capsule**. This structure of the cell envelope is described in details in its own section: The bacterial capsule.

Bacterial genomes

The bacterial genome is encoded on a large, typically circular, DNA molecule called a chromosome. Additionally, some species sometimes carry secondary chromosomes, and often smaller DNA molecules such as plasmids and extra-chromosomal prophages can be present in the cytoplasm. In this section, I overview their **composition** and **organization**, and present some hypotheses explaining these observations.

Genome regions can be classed between **protein-coding regions** and **non-coding regions**. Protein-coding, and often non-coding regions, encode the information for a function, for example an enzyme catalyzing a chemical reaction, or a promoter regulating the expression of an adjacent region.
Bacterial genomes are very **dense in protein-coding regions**, which account for 85-90% (McCutcheon and Moran 2012) of the DNA. Moreover, most functional regions are organized in operons, which form a condensed array of co-transcribed and co-regulated protein-coding regions, leading to high genetic linkage among them. Proteins are translated from the messenger RNA which is transcribed and read DNA triplets called codons and corresponding to specific amino acids. A start and a stop codon enclose the coding regions in open reading frames (ORFs). Several codons correspond to the same amino acid, but synonymous codons are not randomly distributed within bacterial genomes, which is known as the **codon usage bias** [1]. This is due to selection on translation optimization [1,2], co-evolving with host tRNA abundance [3]. It is also the result of mutational bias and base composition, including the impact of the GC-content [4], transcription-coupled repair [5], and leading-/lagging strand bias [6].
The 10-15% remaining DNA is composed of non-coding regions, and encompass functional and non-functional sequences. Functional non-coding DNA is usually involved in gene regulation, in the form of promoter sequences, enhancers, ribosome-binding sites, origin of replication, replication terminus, or functional RNAs. Non-functional non-coding DNA includes decayed protein-coding regions called pseudogenes, and fragments of transposons and other integrated mobile genetic elements.

Chromosome replication often starts from one origin, called *oriC*, at which two replication forks form and move in opposite direction, forming a theta-like structure. The two forks terminate at the terminus region *ter*, which is diametrically opposed to the *oriC*. Forks move from the origin to the terminus and are composed of a leading and a lagging strand. Leading and lagging strands asymmetry and gradient ploidy between origin-proximal and distal regions are two important factors driving the organization and composition of bacterial genomes [7] (Figure 2).

Leading and lagging strands have asymmetric nucleotide compositions [8], and this is hypothesized to be due to asymmetric mutation biases [9,10], particularly the excess of cytosine deamination (leading to C/T transition) of single-stranded DNA. Additionally, **essential genes** are mostly found in the same direction as the **leading strand**, and the selection for this co-orientation stems from the deleterious impact of DNA polymerase / RNA polymerase collisions that may disrupt replication, but most importantly produce truncated transcripts [11].

In fast-growing cells, several initiations can start successively before complete replication, which accelerates replication and leads to a higher copy-number for *oriC*-proximal genes. As a result, genes necessary for fast-growth and under **strong dosage effects** like transcription and translation are over-represented close to the origin [12]. Hence, replication-associated gene dosage is an important determinant of chromosome organization and dynamics, especially among fast-growing bacteria.

Coding-regions farther from the *oriC* also tend have higher mutation rates resulting in elevated evolutionary rates [13,14]. This may be due to the preferential use of specific repair systems close to the terminus region [13,14]. Generally, the DNA **G+C content** is lower close to the terminus [13]. It has been proposed that GC alleles are more frequently under positive selection [15], but this does not explain why G+C3, the third codon position GC content, is lower close to the terminus in many species [13]. Different hypothesis may explain this observation, including replication-associated phenomena counteracting the G/C to T/A transitions bias [16], that may be driven by cytosine deamination. Overall, replication appears to be an important mechanism responsible for the GC gradient between the origin and terminus of bacterial chromosomes.

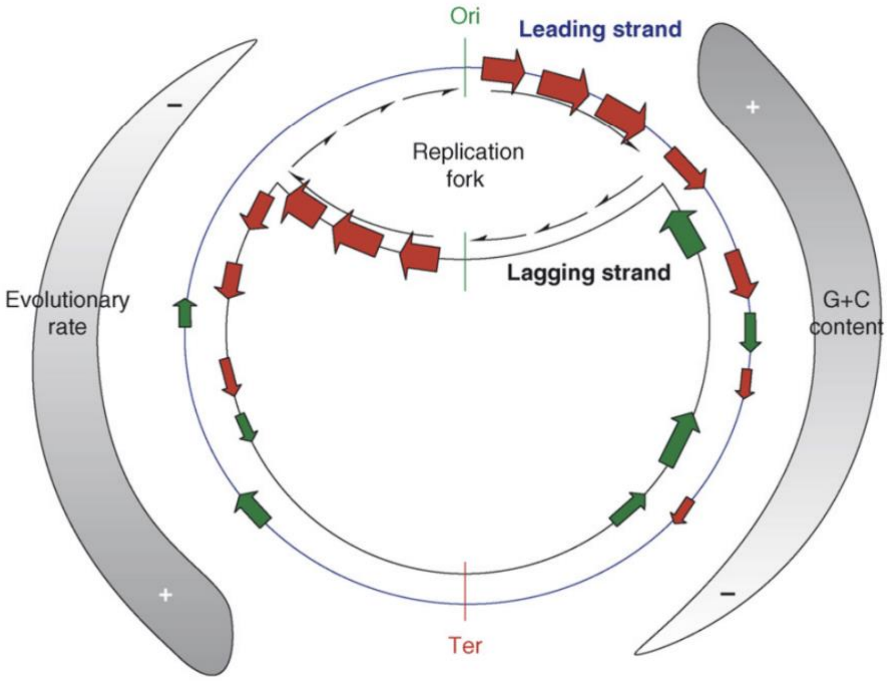


Figure 2 – Replication constraints on bacterial chromosomal organization. Two kinds of biases are detected along the bacterial chromosome: asymmetries owing to the existence of a leading and a lagging strand, and biases related to the proximity of the origin and terminus of replication (Ori and Ter). Essential genes are represented by red arrows and non-essential genes are shown in green. The thickness of an arrow is proportional to the expression rate of the gene it represents. Essential genes are preferentially located on the leading strand and highly expressed genes, especially those related to transcription and translation, tend to be closer to the origin of replication in fast-growing bacteria. The evolutionary rate and the G + C content (gray gradients) are respectively increasing and decreasing with distance to the origin. Figure and legend adapted from [17].

Bacterial genomes are populated with (semi-)autonomous genetic elements called mobile genetic elements [18]. The term mobile relates to their capacity to move across genomic regions either within the cell (intracellular mobility) or between different cell genomes (intercellular mobility). Intracellular mobile elements like transposons can piggy-back intercellular mobile elements like conjugative plasmids, and they are both responsible for horizontal DNA transfer between bacterial genome, which is the focus of the next section.

Horizontal DNA Transfer

Horizontal DNA transfer refers to the movement of genetic information between organisms, as opposed to the vertical DNA transfer that occurs during cell duplication [19–21]. DNA transfer can result in different fates for incoming foreign DNA (Figure 3). **Integrated** foreign DNA may replace native genes, or replace homologous genes with other alleles (or deletions) by a process called **allelic recombination** [22–24]. If the foreign DNA contains non-homologous genes, and is bordered by homologous regions, homologous recombination can lead to the incorporation of novel genes by **horizontal gene transfer** (HGT) [22,23]. Additionally, non-homologous recombination can lead to the integration of novel genes by a process called site-specific recombination, leading to HGT [25,26]. Finally, foreign DNA may also stay **unintegrated** in the cell in the form of a plasmid [27,28].

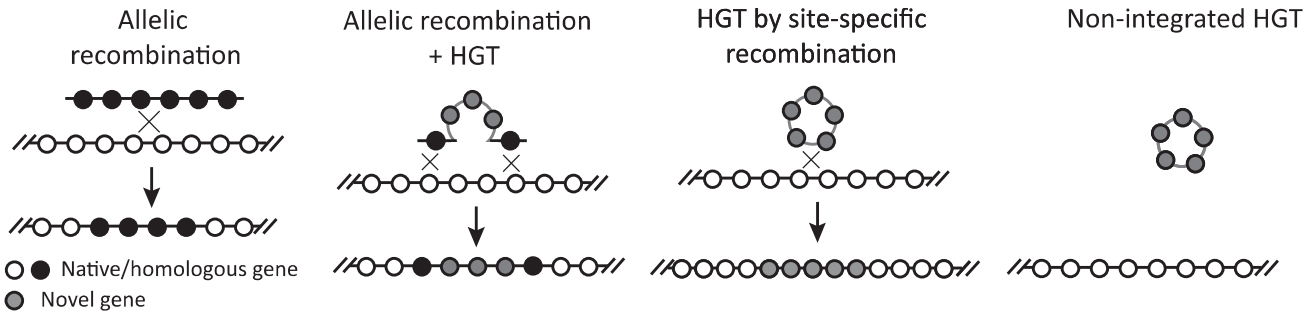


Figure 3 - Fate of horizontally transferred DNA. Modified from [23]

Historically, horizontal transfer of a trait was first demonstrated by Frederick Griffith in 1928 in a pioneering study on the virulence of *Streptococcus pneumoniae* [29]. He showed that mice co-injected with live avirulent strains and heat-inactivated virulent strains died as a result of bacterial infection. His result indicated that a thermostable factor was able to modify the heredity of a trait, turning the avirulent strain into a virulent one. This factor was later found to be DNA [30], transferred from the dead bacteria to the live bacteria through a process called natural transformation. Importantly for this thesis, the avirulent strains used by Griffith were phenotypically “rough”, lacking a polysaccharide capsule. The gene transferred in Griffith experiment was the functional copy of a gene involved in capsule synthesis which was deficient in the avirulent rough strains.

Allelic recombination and HGT drives the evolution of most bacterial species [21,31–33]. These exchanges are mediated by the physical structure of mobile genetic elements (MGE) such as

conjugative systems and virions, and can lead to the gain of new functionalities like antibiotic resistance or virulence factors [34–37]. MGEs transfer from one bacterial host to another via various mechanisms, and their spread depends on host-encoded, and MGE-encoded, factors. In the following sections, I will introduce the different mechanisms supporting horizontal DNA transfer in bacteria, the determinants impacting those transfers and establishment of foreign DNA in recipients cells. I will then discuss the consequences of genetic exchanges on bacterial genomes and bacterial ecology, with a focus on the acquisition of human-relevant traits.

Mechanisms of Horizontal DNA Transfer

Natural transformation

Natural transformation corresponds to the active import of environmental DNA within cells. Around 80 bacterial species including monoderms and diderms are known to engage in transformation [38], but many more species are suspected to do so. This uncertainty stems from the observation that many species encode the necessary genes involved in competence [39], but conditions inducing the expression of this state are unknown for the majority of species [40]. Hence, the relative importance of natural transformation compared with other mechanisms of HGT is not known. Moreover, the idea that natural transformation evolved as a result of natural selection for new trait acquisition by HGT is still debated. Indeed alternative hypothesis have been proposed, such as chromosomal curing from parasitic MGE [41], selection for nutrient uptake [42] or DNA repair [43].

The main steps involved in the natural transformation (Figure 4) of DNA leading to its introduction in the recipient bacteria have been studied in model organisms such as *Bacillus subtilis* [38], *Streptococcus pneumoniae* [44] (monoderms) and *Neisseria gonorrhoeae* [45], *Haemophilus influenzae* (diderms) [38]. First, double stranded DNA (dsDNA) is released in the environment, generally via cell lysis. This implicates that the DNA substrate usually occurs in adverse conditions where many bacteria die. Then, specific type IV pili coated with adhesins with a high affinity toward DNA bind dsDNA and direct it to the cell membrane(s) [38,46]. In diderms, the type IV pilus retracts within the outer-membrane secretin, which is not necessary for monoderms since they lack the outer membrane. Then, one strand is translocated into the cytoplasm by a translocase, namely ComA/ComEC proteins, whereas the other strand is degraded. The single-stranded DNA is then able to engage in RecA-mediated homologous recombination. [38,46].

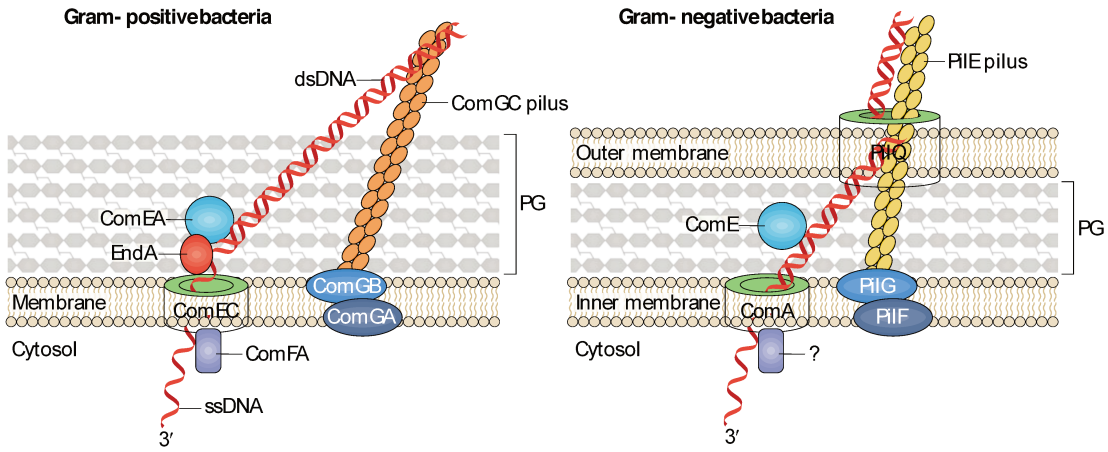


Figure 4 – DNA import mechanisms in Gram-positive and Gram-negative bacteria. Modified from [38].

Membrane vesicle-mediated transfer

Bacteria can release **membrane vesicles** containing an array of molecules, including RNA, linear DNA and plasmids [47]. The function of membrane vesicles is not fully understood, yet they appear to be involved in stress response [47]. HGT by membrane vesicle has been observed in laboratory conditions in a few species of monoderm and diderm [48–50], however no defined genetic pathway have been described to regulate their formation and **loading with DNA**. The proposed mechanism involves membrane vesicle formation and cargo-loading, through budding for example, and latter **fusion with another cell**. Outstanding questions remain regarding membrane vesicle-mediated like:

- Can it transfer DNA outside the laboratory ?
- Is the process regulated from the donor, is DNA actively loaded ?
- Is the process regulated from the recipient, is there a “competence” state ?

Another mechanism through which membrane vesicle formation and fusion may indirectly lead to HGT is the acquisition of phage sensitivity ASEN through receptor transfer [51]. Sensitive hosts carrying a phage receptor may transmit membrane vesicles coated with this receptor, latter fusing with resistant cells. Phage-mediated genetic transfer is described in the next section.

Transfer via vectors

DNA transfer via **vector** refers to the exchange of DNA between cells via the physical structure of **mobile genetic elements** encoding their own inter-cellular transfer systems. **Phages** form particles

called virions to spread, while **conjugative elements** like plasmids and Integrative and Conjugative Elements (ICE) rely on conjugation systems. In the following sections I will describe how DNA transfer takes place between bacteria via phage-mediated transfer and conjugation-mediated transfer.

Phage-mediated genetic transfer

Bacteriophages, or phages, are viruses parasitizing bacteria. Their existence was postulated by Frederick Twort in 1915 after identifying transmissible bacterial lysis zones on his cultures [52], and formally described by Félix d'Hérelle in 1917 [53], who coined the term bacteriophage, from the Greek "eater of bacteria". They were rapidly used to treat bacterial infections, but quickly abandoned by the West in favor of broad-spectrum antibiotics after their success in treating wounded soldiers during WWI. The study of bacteriophages regained momentum in 1940s as pioneers like Max Delbrück, Salvador Luria and Alfred Hershey studying bacteriophages-bacteria interactions revolutionized molecular biology. They discovered that mutations conferring phage-resistance arise spontaneously and randomly before adaptation [54], and that phages can also mutate and infect resistant hosts [55]. Later in the 1950s, researchers started uncovering the ability of phages to transfer traits. Victor Freeman showed that phages could convert avirulent strains of *Corynebacterium diphtheriae* into virulent pathogens, which he called lysogenic conversion [56]. Norton Zinder and Joshua Lederberg discovered that phages can transfer traits between bacteria, by transporting fragments of their genetic materials, which they called transduction [57]. Altogether, the study of phages revealed early on that they are able to efficiently **lyse bacteria**, but also **support genetic transfer** between their hosts.

Phage genetic material is composed of DNA or RNA, either double-stranded or single-stranded, packaged into a capsid. Capsid architecture, nucleic acid composition and the presence of specific structures on the virion are diverse and can used to class phages into different families [58], such as *Myoviridae* (e.g. T4, P1)*, Siphoviridae* (e.g. Lambda), *Podoviridae* (e.g. T7, P22), *Inoviridae* (*e.g.* M13) or *Microviridae* (*e.g.* phiX174). More recently, the principal phage taxonomy which is developed by The International Committee on Taxonomy of Viruses has adopted a genomics approach, relying on gene content and sequence homology and similarity to classify phages.

Phages rely on different modes of infection to spread and persist in bacterial communities. **Virulent phages**, such as T4, infect bacteria and directly engage in the **lytic cycle**: they hijack the cell machinery to replicate, assemble virions and lyse the cell to release their newly formed virions. **Temperate**

phages, such as Lambda, can engage in an additional cycle called lysogeny. During the **lysogenic cycle**, phages stay dormant inside their host in the form of a prophage, usually via chromosomal integration. Upon induction, temperate phages can also engage in the lytic cycle, leading to cell lysis and the release of virions. Beyond the lytic-lysogenic categorization, there exists an array of lifestyles, summarized on Table 1, following the updated nomenclature proposed by [59]. Notably, filamentous phages establish **chronic infections**, either productive, *i.e.* virions are released [60], or non-productive, *i.e.* virions are produced but not released [61]. Some phages, including virulent ones, can also engage in a stalled phage development state, in which the unintegrated phage genome is asymmetrically passed on to daughter cells upon cell division. This phenomenon called **pseudolysogeny** has mainly been observed in starved cells [62], but also in growing cultures [63]. Infection modes of phages are diverse, and support a multitude of horizontal gene transfer mechanisms, either of their own genetic material or of their host which are described below.

	Lytic infection	*Pseudo-lysogeny*	*Productive chronic infection*	*Non-productive chronic infection*	*Lysogeny integrated*	*Lysogeny non-integrated*
production of viral particles	+	–	+	+	–	–
progeny release by cell lysis	+	–	–	–	–	–
progeny release by budding or extrusion	–	–	+	–	–	–
no progeny release	–	+	–	+	+	+
episome	–	+	±	+	–	+
genome integration	–	–	±	–	+	–
inducible	–	–	±	?	+	+
asymmetric division of the episomes	–	+	–	–	–	–

Table 1 – Phage infection modes according to several properties. Virulent phages typically correspond to the first column “Lytic infection” and temperate phages typically correspond to “Lysogeny, integrated”. Adapted from [59]

Temperate phages engaged in the **lysogeny** cycle are part of the cell’s genome and are vertically transmitted when cells divide. While some temperate phages called phage-plasmids are able to maintain themselves as extrachromosomal elements [64], most temperate phages integrate directly in the chromosome of their host via site-specific recombinases called phage integrases [65]. In phage

Lambda, the integrase catalyzes the integration of the phage DNA between its self-encoded *attP* site and a host-encoded site *attB* [65,66]. Prophage integration can result in gene disruption and hence the loss of certain functions [67]. Temperate phages can also encode cargo genes, or "morons", which are not involved in the virus life cycle but can provide new traits to the host bacterium [68]. The expression of cargo genes leading to phenotypic changes for the bacteria is called **lysogenic conversion** [69]. Hence, lysogeny is a form of HGT where the transferred DNA is the phage. Lysogeny also favors genetic exchanges by capture of host's gene within phages. For example, phage Lambda can capture other prophages or bacterial genes by homologous recombination, at frequencies ranging between 10^{-4} and 10^{-6} depending on the extent of homology between the DNA segments [70]. This phenomenon is mainly catalyzed by phage-encoded Rad52-like recombinases during the lytic cycle [70].

Lysogeny is very common in bacteria, with nearly half of complete genomes encoding at least one prophage, and some up to fifteen [71]. Capture of chromosomal genes is also frequent, for example in *E. coli*, an antibiotics resistance gene inserted in the defective prophage Rac was captured by Lambda and found in lysogens at a frequency of 5.10^{-8} in mice gut microbiota in the absence of selection [72].

Phages package their own genetic material inside the capsid of the virion. However, phage DNA packaging processes are not error-free, and can lead to the subsequent packaging of their host's DNA instead of their own. Those phage particles can infect other hosts but will inject either a mix of phage/bacterial DNA, or solely bacterial DNA. This process is called transduction and is subdivided in several types that are specialized transduction, generalized transduction, lateral transduction. Moreover, phages can be hijacked by parasites called phage satellites, and those satellites may also perform transduction. The mechanisms are represented on Figure 5.

Specialized transduction is a special case of lysogenization with imprecise prophage excision, first discovered in phage Lambda [73]. In specialized transduction, the excision of a prophage occurs with a site adjacent to the *att* site of the prophage, leading to the capture of neighboring chromosome genes. The improperly excised prophage then undergoes replication and packaging, with an upper-limit imposed by the capsid size (50kb for Lambda [74]). As a result, only regions close to the integration site of temperate phages can be transduced. The production rate for transducing particle has been estimated to 1:10,000 for phage Lambda [74], and the frequency of successful transduction 100-fold lower at around 1:1,000,000 successful transduction per virion [73].

Generalized transduction refers to the aberrant packaging of the bacterial DNA into the capsid [57,75]. During the lytic cycle, phages degrade the host's DNA to fuel their own replication, and can mistakenly package pieces of the bacterial DNA instead of their own. This mechanism is typically restricted to phages with a headful packaging, which rely on a *pac* site to start the capsid packaging but cut DNA non-specifically when the capsid is full. Generalized transduction is not completely random, as revealed by virion sequencing and mapping on the host chromosome sequence [74,76]. This may be due to phage-encoded factors, like the specificity of the *pac* site recognition, and host-encoded factor, like the presence of endogenous *pac* sites in the chromosome [76,77]. Transducing particles are produced at various frequencies depending on the phage, and has been estimated between 4.5-6% for the model *E. coli* phage P1 [74]. The lysis of a bacterial culture containing 10^6 bacteria releasing a hundred virion each containing 50kb of DNA would mean that the bacterial chromosome is encoded 45-60 times among 10^6 transducing particles.

Lateral transduction is a process by which very long fragments of bacterial DNA are transferred to another cell [78]. In lateral transduction, the prophage begins its replication before its excision from the chromosome, leading to the replication of adjacent chromosomal DNA. Successive encapsidation after initiation from the phage genome leads to the headful encapsidation of large and consecutive stretches of bacterial DNA into the virion.

Finally, phages can themselves be parasitized by genetic elements called **phage satellites**. Satellites cannot produce virions on their own, instead, they hijack their "helper" phage's machinery to spread [79]. Most studied satellites come in the form of phage-inducible chromosomal islands (PICI) [79], including the well-characterized *Staphylococcus aureus* pathogenicity islands (SAPI), PICI-like elements (PLE) and the P4 satellite of the P2 *E. coli* temperate phage [80]. Similarly to temperate phages, satellites remain in the bacterial host genome and may bring new traits like virulence factors [81], or anti-phage defense systems [82]. Satellites are ubiquitous in bacteria [83], for example, P4-like elements are present in 44% of sequenced *E. coli* genomes [84].

Some bacterial and archaeal species produce virus-like particles containing DNA called **gene transfer agents** (GTA). GTAs have been described in a few species such as *Rhodobacter capsulatus* [85] or *Bartonella spp.* [86]. They seem to have emerged from independent domestication events of prophages by their host [87], as they usually share many homologous genes with phages but lack the ability to form infectious particles. Like prophages, genes regulating the production of GTA are expressed upon DNA-damage stress, environmental cues and quorum-sensing, and are released via cell lysis [88]. In

contrast to generalized transduction, GTAs seem to have no sequence-specificity determinants, and package random host DNA sequences [87].

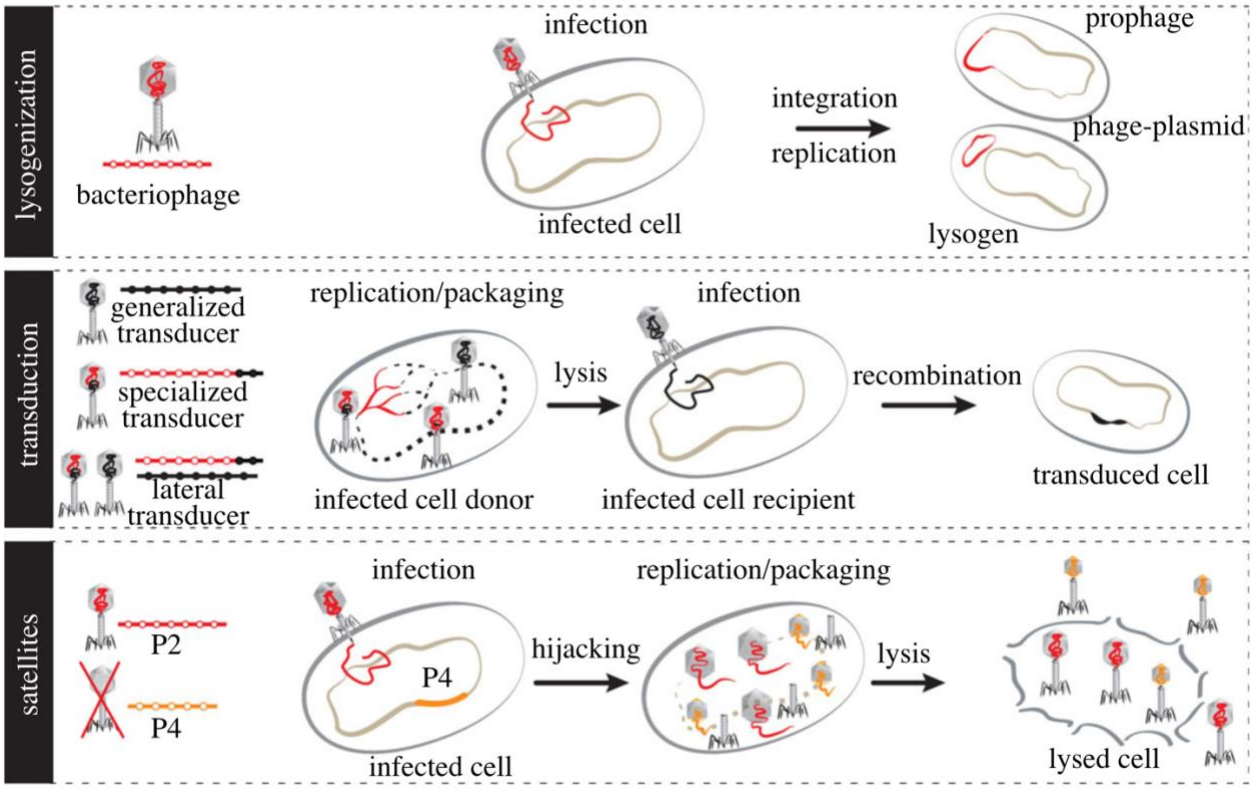


Figure 5 – Main mechanisms of phage-mediated HGT. Adapted from [89]

Conjugation-mediated genetic transfer

Bacterial conjugation refers to the transfer of DNA from one cell to another by direct contact. Conjugation was discovered by Joshua Lederberg and Edward Tatum in 1946 [90]. They identified a contact-dependent transfer of genetic markers conferring antibiotics resistance in closely related strains of *Escherichia coli.* Esther Lederberg, who was trying to map the position of phage Lambda in the genome, identified the fertility factor F as the genetic element responsible for the transfer of DNA [91]. This factor was later identified to be carried on a plasmid, which was shown to be able to integrate and excise out of the chromosome of *E. coli.* Whereas the excised circular form of F transfers only itself, the integrated form of F is able to transfer along with the entire bacterial chromosome into a recipient cell. F's capacity to transfer very large area of the bacterial chromosome from its insertion

point proved very useful in the 1960's to map the bacterial chromosome by timely interruption of conjugation to identify genetic markers locations [92].

Mobile genetic elements encoding the genes necessary to spread via conjugation are called **conjugative elements**. They fall into two main categories: **conjugative plasmids** and **Integrative and Conjugative Elements** (ICEs) [93,94]. Plasmid are extrachromosomal elements usually encoding replication and partition systems for their maintenance, while ICEs integrate in the bacterial chromosome via site-specific recombinases, or sometimes via relaxase-mediated integration [25,95,96], and are maintained by chromosomal replication. Those two elements harbor similar genetic repertoire [94] with key differences relating to their lifestyle (e.g. integrases, partition systems). Integrated conjugative plasmids and episomal ICEs have been described, and interconversion between the two is frequent [97]. Conjugative plasmids and ICEs have similar median sizes (46kb vs. 52kb) but conjugative plasmids have a higher genetic plasticity and exhibit a size variation coefficient twice higher than ICEs [94]. The latter, however, transfer more frequently between distantly related taxa, and thus seem to have a broader host ranger [94].

The conjugation machinery encoded on ICEs and plasmids rely on a set of conserved proteins [98]. Eight distinct mating-pair formation (MPF) types have been characterized based on genetic similarity, including four frequently found in Proteobacteria: MPF_F (based on plasmid F), MPF_I (based on IncI plasmids like RP4), MPF_T (based on *Agrobacterium tumefaciens* Ti plasmid) and MPF_G (based on ICE HIN1056) [28,99]. The mechanism of conjugation appears to be conserved across MPF types. ICEs first excise out of the chromosome and circularize upon entering the conjugative cycle [95,100]. Conjugation starts with the action of the relaxase that nicks the dsDNA of plasmids and ICEs at the *oriT*. *OriT* nicking initiates the rolling-circle replication of plasmid in the donor. The relaxase forms a nucleoprotein filament with the ssDNA, which is actively transferred into the recipient cell through a type IV secretion system [101]. Direct contact and stabilization of the cell-cell interaction is enabled by the mating pair formation system [98], which forms a membrane-spanning protein complex along with a sex pilus (Figure 6).

Conjugative MGEs lead to the transfer of new functionalities encoded on the element, but can also mediate the transfer of chromosomal DNA or other MGE.
Analogous to specialized transduction, ICEs may **excise imprecisely** from the chromosome, bringing flanking genes along with themselves, especially those with low sequence specificity [100]. ICEs also

have a propensity to **accretion**, forming tandem arrays with other integrative elements with identical or similar recognition sites [102], and may be capable of moving together.

Integrated conjugative plasmids such as F, and some ICEs can start **conjugation from the chromosome** itself in a process called *Hfr*-like conjugation [103–105]. In this scenario, ICEs and integrated plasmids starts their conjugation cycle before excision from the *oriT* and can transfer up to the entire chromosome, with the direction depending on the orientation of the inserted conjugative element. The frequency of this mechanism is, however, currently unknown.

The conjugation system of plasmids and ICEs can be exploited by genetic elements that do not encode a complete system of their own. These elements are said mobilizable, because they can engage in conjugation but only when a functional conjugation system is present in the host [28]. The minimum requirement for a mobilizable element to transfer is to carry an origin of transfer (*oriT*) compatible with the relaxase of the conjugative element they parasite. Both mobilizable plasmids and integrative and mobilizable elements [106] have been described, but mainly mobilizable plasmids carrying their own *oriT* and relaxase have been experimentally investigated. A large-scale study on plasmids has estimated that mobilizable plasmids encoding a relaxase are as frequent as conjugative plasmids in proteobacteria [28], around 23% of all plasmids. However, considering that plasmids may only contain an *oriT* to be mobilized, the number of mobilizable plasmids is likely higher.

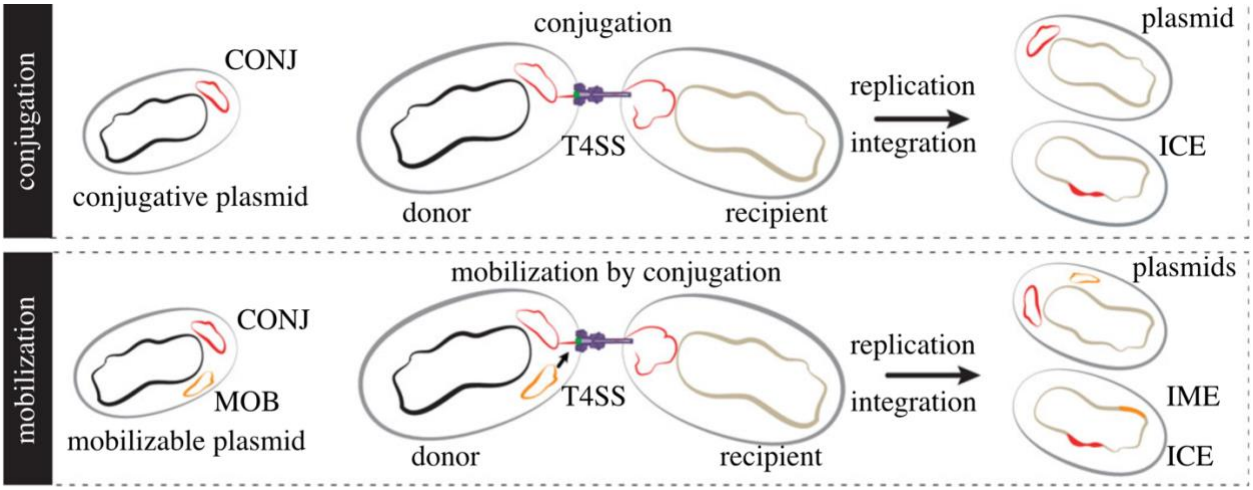


Figure 6 – Conjugation-mediated HGT. Adapted from [89]

Determinants of vector-mediated HDT

Phage and conjugative systems infection success depends on multiple host determinants that shape their **host-range**. Those determinants can either be **extracellular** and mediate the interaction between the physical structures of MGE and the cellular envelope, or **intracellular** and mediate the establishment of DNA into the host genome.

Cell envelope determinants

Phage virions interact with the cell envelope to recognize and infect their host. Infection is a multi-step process involving adsorption at the surface and injection of phage DNA across the cell membrane(s). Adsorption is mediated by interactions between proteins protruding from the virion, called **receptor-binding proteins (RBP)**, and ligands accessible from the cell surface, called receptors [107,108]. RBPs are located outwards the virion, at the tip of the tail, on tail-spikes, on the baseplate, or all around the capsid, depending on their morphology. RBP-mediated adsorption on the cell surface typically happens in two steps, first by reversible binding between the RBP and a receptor, sometimes followed with a "walk" at the surface, and then irreversible binding between the same or different receptor/RBP pairs (Figure 7) [109]. Irreversible binding leads to conformational changes in the virion that trigger the injection of its genetic material [110]. RBPs are specific to their ligand, which can be a protein, glycoprotein, polysaccharide or carbohydrate moieties [107]. In diderms, one of the most studied model is the T4 coliphage which binds reversibly to the porin OmpC or the LPS, depending on the strain, and irreversibly bind the outer core region of the LPS [111,112].

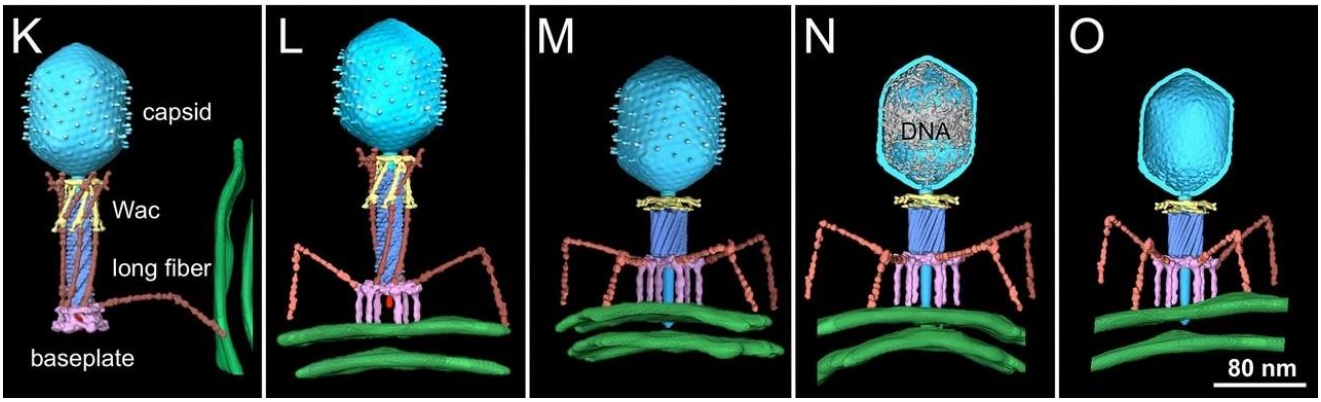


Figure 7 - Phage T4 adsorption mechanism. Pictured are rendered 3D tomograms, shown as central slices, of individual virions after 30 s (K), 1 min (L), 3 min (N), 5 min (N), and 10 min (O) of infection.

Long-fiber tips RBP first reversibly binds to OmpC or the LPS, and engage in a walk from one receptor to another. Baseplate-bound tail RBPs then irreversibly bind the outer core LPS, leading to needle-like injection of the phage DNA (Capsid is full of DNA in N, and empty in O). Figure adapted from [111].

Phage host-range (or host sensitivity) is driven by the occurrence of specific RBP/receptor pairs, resulting in narrow inteactions. Phages can have one, or several different RBPs, and are generally thought to be **strain-specific**, even though broad host-range phage abundance may be underestimated due to isolation bias [113]. Phage specificity can contract [114], or expand [115], indicating that host specificity is an evolvable trait. Host-range contraction evolves rapidly by loss of RBPs, or mutation in RBPs [114], while expansion also involves more complex processes such as RBP recombination including additional RBP acquisition [116]. In turn, phage-sensitive bacteria may develop resistance by mutation in the receptor, and phage-predation selects for diversification of cell envelope features [117–119]. This process is further discussed in the capsule section of the introduction, in Interaction with HGT.

Finally, phages can protect the cell they have infected from further infections, especially from their own virus, which is called **superinfection exclusion**. This phenomenon has been described in temperate phages [120] and virulent phages [121], mostly via repression of phage genes expression. However, superinfection exclusion can take place by cell surface modifications, for example, by synthesis of membrane-bound proteins directly preventing RBP adsorption [120], or by modifying the structure of the receptor via a phage-encoded enzyme [122,123].

Conjugation systems interact with the cell envelope of the donor and the recipient to form a mating-pair [98]. On the **recipient end**, the conjugation pilus must bind to its surface to establish the mating-pair, and inject DNA through the envelope. On the **donor end**, the MPF system is a complex structure that needs to be synthetized and assembled correctly. Finally, **mating-pair stabilization** can also be impacted by more complex recipient/donor and conjugative element/cell surface interactions.

There are two different conditions in which mating-pair can form, on a **solid surface** such as within a biofilm, or **in liquid** between free-floating cells. Solid surface mating is much more efficient than liquid mating, because i) cells are already in close, stable proximity and ii) some conjugation systems are upregulated during biofilm growth [124,125]. In turn, conjugation systems themselves can favor biofilm formation [125,126].

Given that some plasmids are capable of trans-Domain conjugation on surfaces, for example between bacterial and plant [127] or yeast cells [128], and that those plasmids can also conjugate into bacterial cells, conjugation seem to be a receptor-less process un-avoidable by recipients. Some conjugative elements, like the R64 MPF_I plasmid, encode additional type IV pili with high affinity for cell receptors like the LPS [129], but those only increase conjugation efficiency in liquid, because they increase cell-cell interactions and facilitate mating-pair formation. Accordingly, there appear to be no essential *Bacillus subtilis* [130] and *E. coli* gene for conjugation reception, and only some LPS mutations decrease the reception efficiency by *c.a.* 30% in liquid [131]. There are evidence that some surface molecules may decrease conjugation reception [132]. For example in the ICE St3 of *Streptococcus thermophilus*, reception efficiency on solid surface is increased in mutants lacking (lipo-)teichoic acid, exopolysaccharides, or with lower lipoproteins levels [133]. Finally, early observations on the conjugation of *Haemophilus influenzae*, a capsulated diderm, suggested that strains harboring the same capsular polysaccharide may engage more efficiently in conjugation on solid surfaces [134], but the hypothesis was not formally tested.

The **impact of the donor** cell on conjugation, including the impact of the donor's cell envelope, may also play a role in the efficiency of conjugation. In ICE St3 conjugation on surface, mutations decreasing cell surface molecules that increased reception efficiency actually lowered the donation efficiency [133]. These results suggest that cell envelope determinants may be important for the proper positioning, assembly or activity of the conjugation machinery in the donor cell.

Finally, conjugation may be impacted by the **interaction between recipients, donors and conjugative elements**. In the case of mutations increasing the reception efficiency in ICE St3, their effect may be due to their higher cell-cell interactions and elevated biofilm formation [133]. Additionally, recipients may be protected from conjugation by pre-established conjugative elements, which often encode conjugation exclusion systems against themselves and related elements. Two distinct exclusion systems have been described, entry exclusion and surface exclusion. In F-like plasmids, **entry exclusion** is mediated by TraS in the inner-membrane of the recipient cell, and TraG which may translocate from the donor outer-membrane to the recipient [135], leading to abortion of conjugation after mating-pair formation. **Surface exclusion** is mediated by TraT [136], an abundant outer-membrane protein that destabilizes mating-pair formation when the same allele is present in both donor and recipient [137].

Overall, cell surface features may negatively impact conjugation efficiency by impeding mating pair stabilization especially in liquid, while surface mating appear to be unavoidable by the recipient,

except via exclusion systems. Hence, pre-existing conjugative elements and cell-cell interactions mediated by the envelopes appear to be the most important factors for the formation of the mating pair.

Intracellular determinants

Having crossed the cell envelope, foreign DNA belonging to a MGE or another bacterium must establish in the genome to be vertically inherited. Establishment of foreign DNA is restricted, or facilitated, by an array of intracellular determinants encoded by the host. Those include defense mechanisms, homologous recombination and the replication machinery. In turn, MGE often encode their own factors to integrate and persist in bacterial genome.

Defense systems

Bacteria can protect themselves from foreign DNA and MGEs infection via intracellular factors known as defense systems, analogous to an immune system (Figure 8). The diversity of defense systems is large and is still actively being discovered and characterized [138]. The **homologous recombination** pathway can act as a defense system, especially via the action of the RecBCD complex which is an exonuclease degrading linear dsDNA as carried by numerous phages [139]. **Restriction-modification systems** are genetic elements able to digest incoming foreign DNA, composed of a DNA methylase (antidote) preventing the action of its cognate restriction enzyme (poison). Recently acquired DNA is not methylated and hence quickly degraded by the restriction enzyme, except if it came from a previous host with compatible restriction-modification system [140,141]. **Abortive infection** systems can detect phage infection and trigger cell growth arrest or death, preventing the virus from completing the lytic cycle [142]. **Adaptive immunity** systems, like the CRISPR/Cas system, can acquire fragments from foreign DNA, store it as a spacer in the spacer array, and later use it as a probe to recognize and degrade it [143]. CRISPR/Cas systems specifically target DNA sequences from MGEs [144], and hence protect against infection while allowing non-MGE DNA to undergo homologous recombination [145]. In the past years, the identification and characterization of novel defense systems has greatly accelerated, revealing an unexpected diversity of mechanisms with large consequences on our understanding of bacteria-MGE interactions and ecology [138,146], and potential applications in biotechnology.

Defense systems do protect bacteria against MGEs, the vectors of HGT, but are frequently found within the MGE themselves [82,147]. Additionally, MGEs often encode anti-defense mechanisms,

anti-anti-defense, and so on [148]. Hence, defense systems cannot be resumed as bacterial immune systems, as they are also involved in MGE-MGE interactions [147].

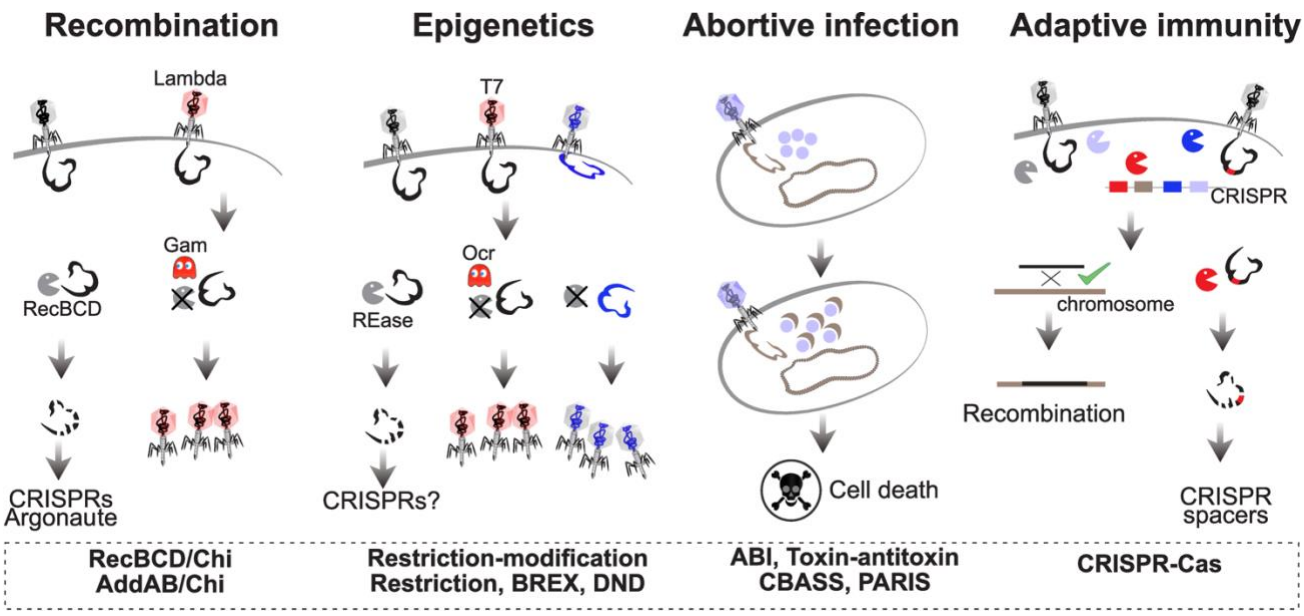


Figure 8 - Overview of defense mechanisms against foreign DNA. Figure from [147].

After passing the defense-system frontlines, foreign DNA can establish either by **integration** in the recipient genome and subsequent vertical inheritance or by **extra-chromosomal maintenance** as an extrachromosomal element called a plasmid, replicated and segregated upon cell division.

Establishment by integration

Integration occurs via genetic recombination which can happen via different mechanisms. **Homologous recombination** is a type of genetic recombination in which sequences are exchanged between two similar molecules of DNA [24]. It is a conserved mechanism across the three Domains of life, and is involved in homologous recombination repair of harmful DNA double-strand breaks. It is also involved in the incorporation of new sequences in the genome, and depend on the extant and degree of homology. Two cross-overs are needed to replace a sequence by another, and such events can lead to allelic replacement, integration of new genes, or deletions [22,23]. The strand exchange process is catalyzed by the RecA protein which is able to bind ssDNA and form a nucleoprotein filament. The nucleoprotein filament engages in homology search, scanning through the cell dsDNA for homologous sequences [24]. The unwound dsDNA bound to the nucleoprotein filament forms an heteroduplex that can either be homologous, leading to no change or single nucleotide polymorphism

upon incorporation. If foreign DNA contains heterologous sequence flanked by homologous regions, the recombination intermediates contains a loop of non-homologous ssDNA which must be filed into a dsDNA after strand exchange takes place.
Homologous recombination promotes HGT and allelic replacement between bacteria [22], but this process is countered by the mismatch repair pathway [149]. Indeed, mismatch repair suppresses homologous recombination when the heteroduplex DNA contains excessive mismatched nucleotides [149]. This process increases the accuracy of HR repair, but also contributes to limit inter-species recombination and can thus act as a gene flow barrier [150].

Other types of genetic recombination exist in bacteria, which also lead to DNA shuffling between or within DNA molecules. These mechanisms are called **specialized recombination**, and rely on molecular machineries leading to transposition and site-specific recombination [25]. Many MGEs, including transposons and ICEs, integrate bacterial genomes via those mechanisms [100,151].
Transposition is the process by which genetic elements move between different locations of the genome via specialized enzymes called transposases [25]. Most transposases are able to recognize a specific small DNA sequence as a target for integration, but can also have non-specific activities especially when their recognition site is absent from the genome [151]. Transposases often travel with cargo genes carrying various functions together forming a transposon, *i.e.* a group of functions able to copy-paste or cut-and-paste by itself [152]. In the case of a transposon, both the transposase and cargo genes are surrounded by specific DNA repeats that serve as borders for the element and recognition sites for the transposase [153,154]. In bacterial genomes, lone transposases form of class of elements referred to as Insertion Sequences [154].
Site-specific recombination is a reaction in which DNA strands are broken and exchanged at precise positions of two target DNA loci, leading to non-homologous DNA rearrangement [25]. This process is catalyzed by site-specific recombinases, which recognize specific target sequences and are able to break, exchange and seal DNA fragments in a reciprocal manner [26].

Establishment by extra-chromosomal maintenance

Finally, MGEs can remain unintegrated in the bacterial genome and persist vertically by (semi-)autonomous replication in the form of **plasmids**, typically circular but sometimes linear [155]. Intracellular factors such as the host's **replication** machinery, MGE's replication and **partitioning** systems can interfere with plasmids persistence, giving rise to plasmid host-range [27,156,157].

Accordingly, plasmids can be classified into narrow and broad host ranges, with broad host range defined as the ability to maintain among bacteria belonging to different phylogenetic subgroups, *e.g.* different species [157].

The interaction between the **plasmid replication** machinery and host factors dictates if a plasmid can successfully replicate. There are three main types of plasmid replication mechanisms, which may rely on the host replication initiation system, called theta, rolling-circle and strand-displacement replication [27].
The theta replication mechanism is analogous to chromosomal replication, with two symmetrical forks starting from a single origin of replication [158]. Theta replication rely on several plasmid-encoded elements which are the Rep, iterons, DnaA boxes and an AT-rich region [159]. For example, plasmid pSC101 encodes a RepA protein which can recruit *E. coli* DnaA at the origin to initiate replication [27]. RepA expression is controlled by a negative feedback loop, enabling control over the plasmid copy-number [160]. Hence, Rep proteins interact with the host's initiation factors such as DnaA, and this interaction may not be strong enough to initiate replication, defining a replicative host-range.
Plasmids can also encode their own replication initiation systems, and be dependent on the host's replication machinery solely for elongation and termination. Both strand-displacement [161] and rolling-circle replicative plasmids [162] have generally broader host-range than theta replicating plasmids because they encode various sets of primases, helicases and Rep initiator proteins [158].

Plasmids also often encode mechanisms to ensure their distribution in both daughter cells upon cell division which are called **partition systems**. Partition systems are typically composed of three elements, the centromere-like site, centromere-binding protein(s) and a motor protein. The partition complex involves plasmid pairing at the centromere-like site via centromere-binding proteins that are then pulled apart by the motor protein in opposite directions like the two bacterial poles [156]. Partition systems are subdivided in classes I-IV based primarily on the type of the motor protein [156].

Plasmids present in the same cell can interfere with one another, leading to the inability of one or both plasmid to stably coexist over a few generations. This phenomenon is called **plasmid incompatibility** and stems from the two processes resumed above, namely replication and partition. Plasmids with the same replication initiation mechanism [163], or the same partition system [164], compete for replication and repartition in daughter cells, usually leading to the loss of one plasmid especially in the absence selection. Incompatible plasmids are said to belong to the same incompatibility group (Inc),

but except for well-studied replicon and partition system, this classification is limited by the lack of experimental validation [165].

Consequences of HGT

In the following sections I will detail the different consequences of HGT on bacterial genomes, bacterial ecology, and a focus on the spread of clinically-relevant traits in pathogenic strains.

On bacterial (pan-)genomes

Bacterial evolution is shaped by horizontal DNA transfer, a process reflected by the high density of horizontally transferred DNA in bacterial genomes [20]. Expansions of protein families arise more frequently by HGT (xenologs) than gene duplication (paralogs) and xenologs tend to persist longer than paralogs [33]. **Genome size** increases with genetic acquisitions, but strains of the same species tend to have a stable genome size, even those with very high rates of HGT. This may be due to a strong deletional bias toward neutral or slightly disadvantageous sequences in Bacteria, which counteracts genetic acquisitions and duplications [166]. In any case, genome size variations are **mechanistically driven by the genetic gain/loss ratio.**

Integration of foreign DNA can occur virtually at any chromosomal location via homologous recombination or specialized recombination. However, horizontally transferred genes are not uniformly distributed along the chromosome [167,168]. In fact, bacterial chromosomes are composed of hotspots of HGT*, i.e.* regions flanked by the same core genes but displaying very high genetic plasticity and MGE concentration. An analysis of 80 species estimated that, on average, less than 2% of the largest hotspots accumulate more than half of the horizontally acquired genes [167].
MGE integrases often target specific regions like tRNA and tmRNA without disrupting their sequence and function [169]. This can lead to the accretion of integrative MGE in the same spots one after another [170,171]. As a result, tRNA and tmRNA are frequently are found in the vicinity of 29% HGT hotspots [167]. Conversely, IS harboring DEE recombinases, which have low sequence specificity for their insertion, are not frequently found in chromosomal hotspots [167]. The location of HGT hotspot also depends on their composition in MGEs. Indeed, phages tend to integrate more frequently closer to the replication terminus [172], suggesting that their integration is more costly in highly expressed regions close to the *ori*, while ICEs integration position does not correlate with the *ori* → *ter* axis [167]. Hence, site-specificity of MGE integrases can lead to the formation and diversification of hotspots, and their location depend on the type of MGE presumably because they may incur different location-related fitness effects.

Even though most integrases are found in hotspots, more than half of the hotspots do not contain an integrase or identifiable MGE [167]. This may be due to the fact that hotspots can also diversify via double homologous recombination at the flanking core genes, an hypothesis supported by the observations that flanking core genes have high recombination rates and transformable species have relatively more hotspots [167]. For example, the capsule locus of several species including *Streptococcus pneumoniae* and *Klebsiella pneumoniae* is a chromosomal hotspot with tremendous genetic diversity and high flanking gene recombination rates [173–175].

Frequent influx and deletion of genes in bacteria have a tremendous impact on the gene repertoire of bacterial species. Indeed, closely related strains often harbor a common set of conserved and often essential, homologous genes, as well as many accessory genes recently acquired [176]. The accessory genome is mainly comprised of MGE, like prophages and plasmids [177,178]. In groups of related strains, like a species, the intersection of homologues between all strains is called the **core-genome**, which can be used to infer phylogenies [179]. In genomics, homologs groups are approximated by clustering of gene sequences into gene families. Core genes are not exempt of allelic exchange but are usually consistent with a single tree [180], however the impact of recombination on phylogenies is debated [181]. The core genome concept is usually approximated by the persistent genome, which is the set of gene families present in 90-99% of the strains, and is particularly helpful as bacterial genomes assemblies can be incomplete (see Genome assemblies), or too divergent (see Pan- and core-genome). The union of all distinct gene families found in related strains is called the **pangenome**, which typically increases with the number of sampled isolates [178]. The pan-genome rarefaction curve may eventually reach a plateau, meaning that the pangenome is “closed” and the genetic diversity is almost completely sampled. However, most bacterial species pangenomes are “open”, suggesting that their genetic diversity is not bound by the definition of their species [168,182].

Species with large effective population size (N_e, the number of individuals giving rise to an offspring in the next generation) tend to have relatively larger genomes and open pangenomes [178]. This observation seemed puzzling at first, since natural selection is more effective at high N_e [183,184]. and should result in rapid removal or fixation of new genetic variants, limiting the accumulation of mildly deleterious (or neutral) genes. Additionally, many bacterial genes only provide a fitness advantage under very specific conditions [185]. Thus one of the explanations for the correlation between pangenome openness and N_e is that populations with large effective sizes inhabit diverse environments and often migrate to other niches [178]. Adaptation toward each environment is accompanied with acquisition of new variants and genes, so sampling generalist species from diverse environments will

be reflected by a large pangenome. Hence large pangenomes have been proposed to be the result of adaptive, not neutral, evolution [178]. This reasoning also indicates that generalist species have high rates of genetic acquisition, in part because they have more opportunities to interact with other cells. For example, *Klebsiella pneumoniae*, an ubiquitous species, has an estimated core genome of less than 20% its pangenome, and each new isolate adds on average 134 new pan-genes (https://pangenome.org/Klebsiella_pneumoniae). Accordingly, species with low N_e and isolated niches like endosymbionts and obligatory intracellular pathogens have small genome size and closed pangenomes [186,187], because they have less distinct environments to adapt to, less opportunity for HGT, and are more subject to drift (Figure 9).

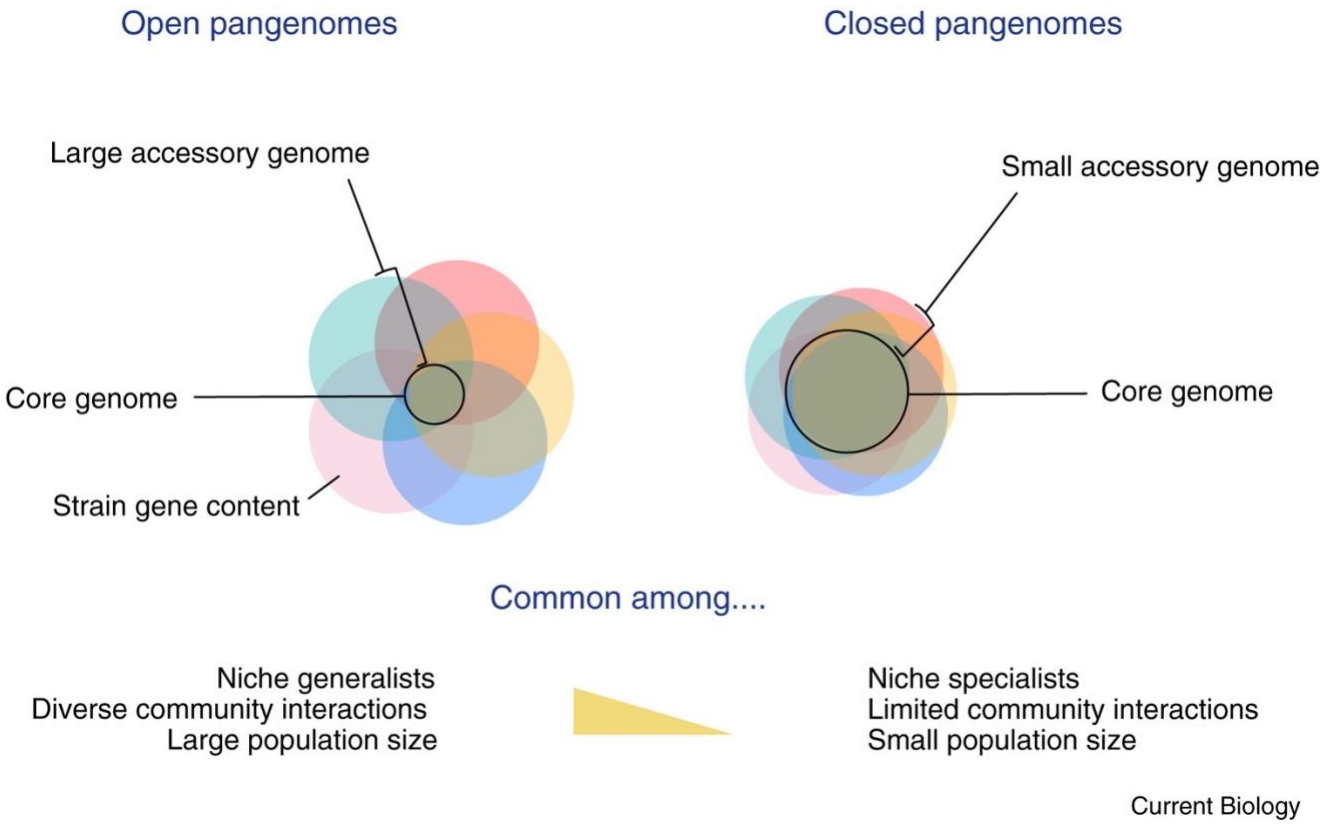


Figure 9 – The bacterial pangenome concept. Species pangenomes can either be open (left) or closed (right). Pangenome openness correlates with core genome proportion and ecological characteristics including niche diversity, community interactions and population size. Figure from [182]

On bacterial ecology

As discussed above, HGT is associated with the colonization of novel niches. Indeed, HGT can lead to the incorporation of a whole set of new functions and promote the evolutionary leaps that allow bacteria to rapidly adapt to new conditions [188,189]. In turn, bacterial interactions may be favored by HGT, because cooperative traits can be transferred horizontally [190,191], especially between geographically close cells. Moreover, MGE infection may enforce cooperation by spreading cooperative traits to surrounding cheaters [35,190,192].

The number of different ecological niches on Earth is tremendous, as reflected by the one trillion estimated microbial species inhabiting the planet [193]. Horizontal transfer of homologous genes leading to allelic recombination can introduce new alleles and remove deleterious mutations, but is unlikely to create radically novel traits [31]. Horizontal gene transfer, on the other hand, can incorporate whole new enzymes and operons, enabling bacteria to acquire new traits like metabolic pathways, virulence factors or resistance mechanisms. For example, methylotrophic bacteria acquired the ability to utilize methanol as a carbon source from methanogenic Archaea [194], enabling them to colonize methanol-rich plant leaves surface. In a study of 53 *Escherichia coli* strains, it was found that almost 2000 metabolic innovations arose from the horizontal acquisition of a single DNA region less than 30kb long [189]. In *Listeria monocytogenes*, a facultative intracellular pathogen, virulence is caused by the ancestral acquisition of *Listeria* pathogenicity islands, encoding an array of virulence factors [195,196] and sometimes transferred by HGT to other members of the genus like *Listeria innocua* [197]. Perhaps the most infamous case of HGT leading to virulence is the lysogenic conversion of *Vibrio cholerae* by phage CTXΦ which encodes the cholera toxin, leading to increased survival in the host and deadly infections [69].

Finally, antibiotic resistance genes (ARG) transfer is the main mechanism supporting the evolution of antibiotic resistance in bacteria[37,198,199]. ARG are mostly encoded in transposons and integrons carried by plasmids [200,201] and ICEs [198], and hence ARG are mainly spread by conjugation [37]. Such plasmids, called resistance plasmids, can be as large as several hundred kilobases and contain multiple ARG conferring resistance to many different antibiotics. The metabolic cost of large resistance plasmids can be quickly compensated by mutations in the plasmid or in the chromosome [202]. Generally, alleviation of fitness cost participates in the maintenance of plasmids and can promote the spread of ARG in the absence of selective pressure [202].

Bacteria seldom live alone, be it as single free-floating cells, or within clonal populations. Bacterial species have overlapping niches and this leads to the formation of communities that interact both within and between species. Among bacterial communities, cooperative behaviors are frequent and typically related to foraging, building, reproducing, dispersing and communicating [203]. The most common medium for cooperation is by the production of "public goods" whose individual cost of production is compensated by the benefits they procure to individual and the group [204]. However, evolutionary theory predicts that public goods cooperation is at risk of invasion by selfish cells, called cheaters, who do not cooperate (*e.g.* do not produce public goods) but benefit from the cooperators (*e.g.* benefit from public goods) [205]. In any case, numerous studies have shown the existence of cooperative traits. For example, the production of siderophores, iron-scavenging molecules, in the pathogen *Pseudomonas aeruginosa* is a typical case of cooperation via public good, where the emergence of cheaters is documented *in vivo* [206]. In general, genes involved in secretion are over-represented among MGE, especially plasmids [35]. In fact, MGE can theoretically enforce cooperation by i) spreading cooperative traits and ii) re-establishing cooperative traits in cheater mutants [204]. This view is debated, and it has been proposed that while HGT can help cooperative genes initially invade a population, it has less influence on the longer-term maintenance of cooperation [207].

Overall, HGT has a deep impact on bacterial evolution by allowing the rapid spread of numerous functions leading to the colonization of novel niches, and favoring cooperation among bacteria.

The bacterial capsule

The bacterial capsule is the outermost structure of the cell and is present at the cell envelope of monoderm and diderm species. Capsules are generally synthesized through complex, multi-step biosynthesis pathways which require the expression of many genes regrouped in capsule loci [208]. They are very frequent cell envelope structures, as 52% of bacterial species with more than four complete genomes were found to frequently encode one or more capsule loci [209]. Moreover, the nosocomial multi-resistant group of ESKAPE bacteria, are all capsulated [210]. Capsules are typically made of polysaccharides, but occasionally of proteins, anchored to the cell surface. They are highly diverse in term of assembly machineries, which are classed according to four main groups. Some species harbor diverse capsular polysaccharides, which are classed into serotypes. The term capsule serotypes stems from the historical use of serological methods to different bacterial strains [211].

Roles

Bacterial capsules are **protective layers** against abiotic and biotic stresses. They forms a highly hydrated barrier that slows down **desiccation** and excludes toxic hydrophobic molecules such as **detergents** [212]. Capsules have also been described to protect against protozoa **grazing** [213], and bacteriophages infection by masking surface receptors [214,215]. To note, the interaction between phages and capsules will be detailed in the section Interaction with HGT. Finally, capsules can enhance the ability of bacteria to infect their hosts and are thus considered virulence factors [34].
Capsule's role in virulence is diverse, as they protect cells against the immune system in various ways. First, they **mask the immunogenic antigens** present at the cell surface [216]. Secondly, the thickness and negative-charge of the capsule **impedes complement-mediated killing and phagocytosis** [217], and **protect against antimicrobial proteins** [218]. Finally, capsules of human pathogens are generally composed of polysaccharide motifs **mimicking the host's *self***, which can't be efficiently targeted by antibodies [219].

The overall protective role of capsules contributes to explains why bacterial species encoding capsules are associated with a broader environmental breadth, *i.e.* they can be found in more environments than other species [209].

Synthesis of polysaccharide capsules

Polysaccharide capsule synthesis pathways are multi-step enzymatic reactions taking place in the cytosol and across the cell membrane(s). These pathways are diverse and fall into distinct groups based on their assembly machineries. The exact steps underlying capsule synthesis have been characterized in model species like *Escherichia coli*, *Streptococcus pneumoniae* and *Salmonella enterica*.

Group I capsules, also called **Wzx/Wzy-dependent** capsules, have been most studied in *Escherichia coli* (Serotype K30) [220], *Streptococcus pneumoniae* [221] and *Klebsiella pneumoniae* [222]. Synthesis starts in the cytosol by the loading of the first sugar (e.g. glucose-1-P) on the lipid carrier undecaprenyl phosphate by an initiating glycosyltransferase. Specific glycosyltransferase then assemble individual repeat units onto the carrier lipid, which are then flipped across the membrane by Wzx. Repeat units are then polymerized by Wzy into full-length polymers. In diderm, the chain length is determined by Wzc phosphorylation state, regulated by the Wzc auto kinase activity and its cognate phosphatase Wzb. Capsule polymers are exported by Wzc through an outer-membrane porin called Wza and subsequently bound non-covalently by the outer-membrane protein Wzi [223]. In monoderm, the Wzd/Wze complex is responsible for translocation, and possibly attachment, of capsular polysaccharide to the cell membrane [224].

Group II and III capsules, also called **ABC-dependent** capsules, have been most studied in *Escherichia coli*, where serotype K5 is the model of group II capsules [225,226], and serotype K10 the model of group III [227]. The main difference with group I capsule is the transport of the repeat units bound to undecaprenyl phosphate by a transmembrane protein with an ATP-binding cassette called an ABC transporter. Both Wzx/Wzy and ABC dependent capsules rely on homologous proteins such as Wza for export through the outer membrane proteins [208].

Group IV capsules, like *Escherichia coli* K40 serotype [228,229], are similar to Group I capsules but their overall translocation mechanism is currently unknown. They rely on different initiating glycosyltransferase than group I capsules, but use an homologue of Wzy for repeat-unit polymerization.

Synthase-dependent capsules, like the hyaluronic capsule of *Streptococcus spp.*, are produced via a single protein, the synthase, responsible for the initiation, polymerization and translocation of the capsule [221]. The synthase-dependent capsule corresponding to serotype 3 of *Streptococcus pneumoniae* is not bound to the peptidoglycan or outer-membrane [230], and may thus be considered an exopolysaccharide.

Genetics of Wzx/Wzy-dependent capsules

Genes involved in capsule synthesis are typically encoded in multiple adjacent operons within genetic clusters called **capsule loci**, which are among the fastest evolving loci in bacterial genomes [118,231]. In this section I will describe the genetic organization of capsule loci. I will especially focus on Wzx/Wzy-dependent capsules, since they are the most abundant, account for *ca.* 70% of identified loci in sequenced genomes [209], and are the main focus of this thesis.

Capsule loci are most often found in single chromosomal location [174,224,232], sometimes in tandem controlled by phase variation [233], or on mobile genetic elements [209]. Capsule loci are composed of capsule group-specific genes and serotype-specific genes, with a conserved organization (Figure 10) [222]. Group-specific genes encode conserved functions involved in the assembly machinery and are encoded at the 5' and 3' ends of capsule loci. Serotype-specific genes are highly variable genes that determine the chemical composition of the repeat-unit and are encoded in the central region of capsule loci [220,224,234].

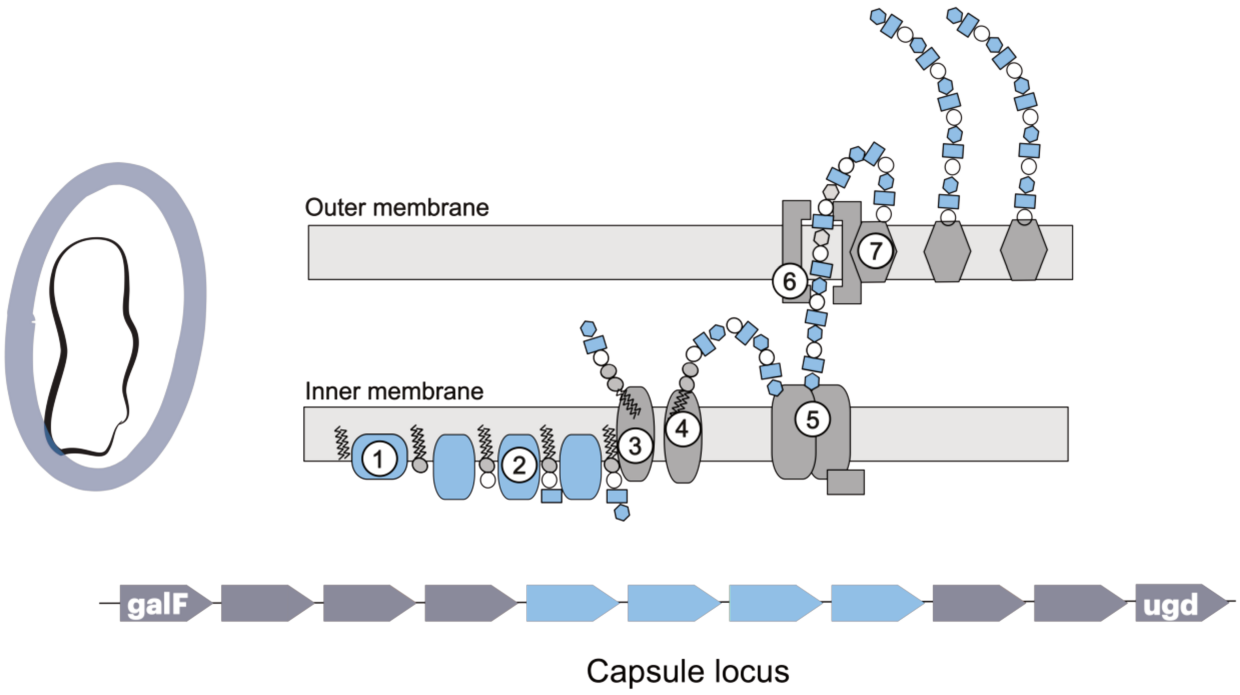


Figure 10 – The prototypical genetic organization and biosynthetic pathway of the group I capsule of *Klebsiella pneumoniae*. On the left, a bacterium with a capsule represented in blue to denote the serotype. On the right, a schematic assembly of the capsule, with serotype-specific repeat-units in blue. Serotype-specific sugar processing enzymes in blue, correspond to the variable, middle region, of the

capsule locus. Group-specific, export proteins are represented in grey, corresponding to the bordering regions of the capsule locus.

The genetic content of capsule loci has been shown to be a very good predictor of capsule serotype in species like *Streptococcus pneumoniae* [224,235], *Escherichia coli* [236], *Acinetobacter baumannii* [237] and *Klebsiella pneumoniae* [232,234]. Advances in comparative genomics have led to the development of specialized tools to infer the capsule type from sequencing data of bacterial isolates. Many pathogenic species have historically been typed by sera raised against capsule polysaccharide [238], and sometimes capsule-specific bacteriophages [239], because they provided a rapid test to classify isolates into serotypes. For example, there are currently 97 *S. pneumoniae* serotype structure available [240], and 74 for *K. pneumoniae* [234]. The combination of strains serotyping and sequencing have led to the precise identification of genetic clusters corresponding almost perfectly to the serotype by comparative genomics [224,234]. Distinct capsule genetic clusters are called **capsule locus types**. Specialized softwares such as Kaptive can infer capsule locus types – an thus predict the serotype, with high accuracy based on curated reference databases, whole-locus and by-gene alignments [232,241,242], as resumed on Figure 11. Identification of capsule loci in bacterial genomes has accelerated the rate of discovery of new putative serotypes, for example, there are now more than 70 capsule locus types in *K. pneumoniae* defined solely by their genetic composition [241].

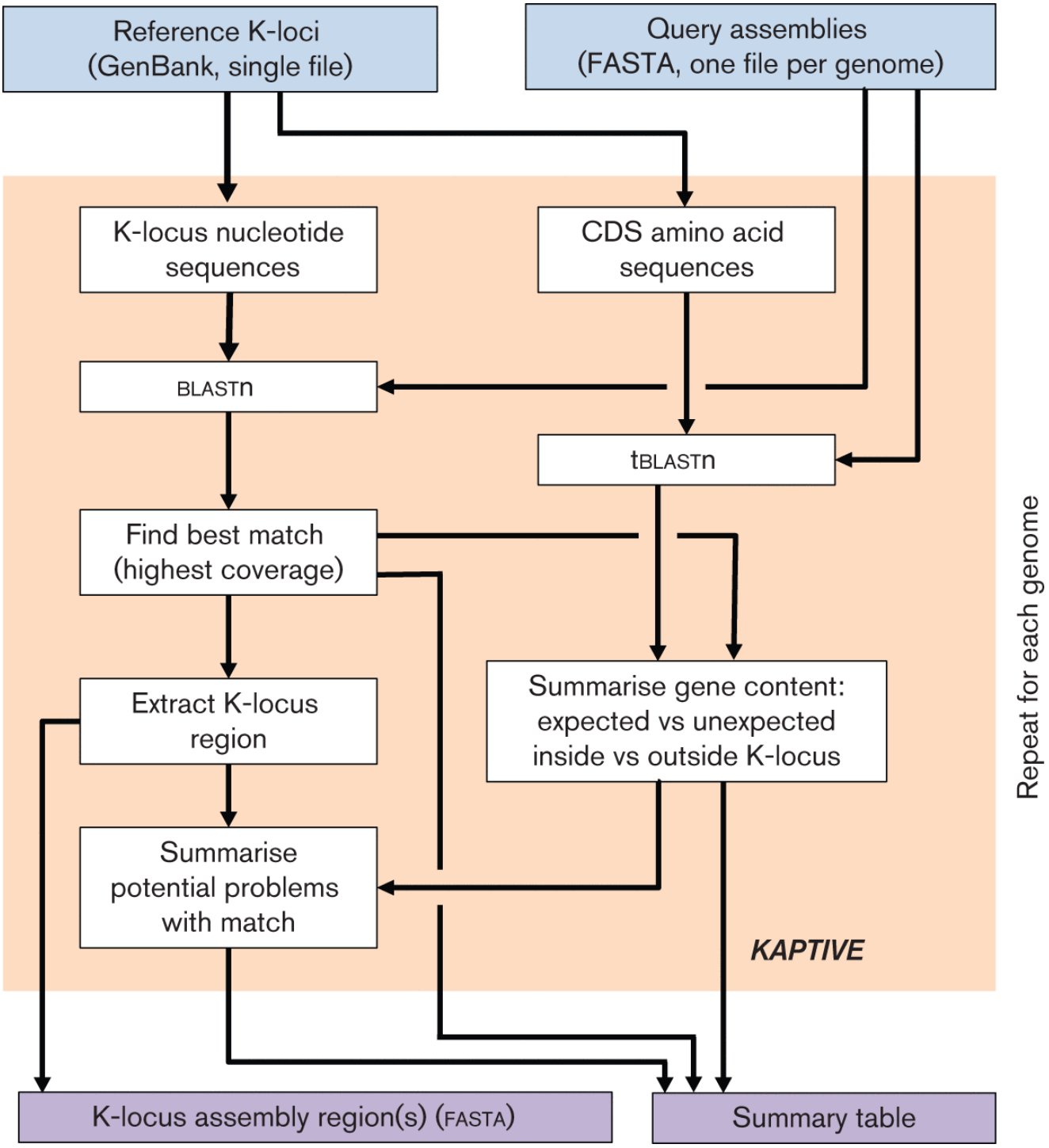


Figure 11 - Summary of the capsule serotype analysis procedure by Kaptive. Figure from [232].

Capsule repeat-unit characterization is now lagging behind the discovery of new capsule locus type. For those new putative serotypes, there is currently no method available to predict the chemical composition of the repeat-unit solely based on the DNA sequence [224,234]. This represents a

challenge in metabolism modeling, because it requires to know i) what is the chemical reaction catalyzed capsule locus encoded enzymes, ii) in which order sugar residues are processed and iii) if genes located outside the capsule locus intervene in the repeat-unit synthesis. Certain simple predictions can be made accurately (Figure 12), for example in *K. pneumoniae* the WbaP is responsible for the synthesis of the initial galactose residue in the repeat unit, while WcaJ is responsible for the initiation with glucose [234]. However, the presence of *manCB* genes, responsible for the synthesis of mannose, only correlates with the presence of mannose in the repeat-unit in the absence of *gmd* and *wcaG* that appear to convert mannose into fucose [234]. Hence, further work on metabolic pathway modeling is necessary to be able to predict the structure of the repeat-unit solely from the genetic composition of capsular loci.

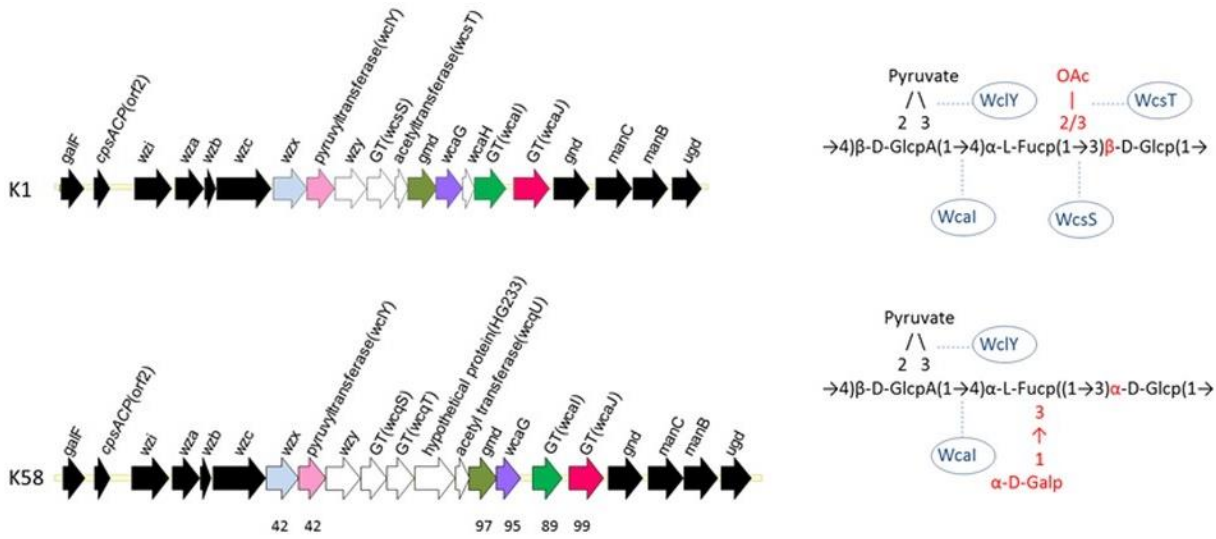


Figure 12 – Correspondence between the genetic composition of capsule loci and chemical composition of the capsule repeat-unit. Figure from [234]

Evolution

Capsules loci are fast-evolving genetic elements showing a tremendous chemical and genetic diversity. Change in chemical composition can occur via genetic change in the capsule locus, by mutation and recombination with horizontally transferred DNA.

Mutations in the capsule locus have been shown to either completely inhibit capsule synthesis by gene inactivation, leading to non-capsulated cells [243–246], change the chemical composition [234,247], or the expression [246].

Recombination has been shown to shuffle parts of the locus with other sugar-processing enzymes [174], or completely replace the whole locus with one of another type [248–250]. Hence, ***de novo* serotype birth** may come from mutation, gene shuffling facilitated by Insertion Sequences (IS) [175] or both, while **serotype replacement** occurs with a pre-existing serotype. Such replacements by recombination have been called serotype switch [248], or swaps [251]. However, a serotype switch can also refer to a switch between capsulated/non-capsulated, or mucoid/non-mucoid phenotypes [252], while a **serotype swap** precisely refers to serotype replacement with the complete capsule locus of another bacterium.

Why and how serotype swaps occur in natural populations is not fully understood. As for the **why**, different and sequential selective pressures may explain the frequency of serotype swaps in pathogens, such as interaction with host immune system, cell-cell interactions favoring colonization, or capsule-specific phage predation (Figure 13) [118]. Regarding the **how**, the genetic organization of capsular loci in conserved syntenic blocks has been proposed to favor serotype swaps [222], since the most conserved and homologous recombination-prone genes surround the serotype-specific genes, and the capsule locus is typically encoded in a defined chromosomal spot [174]. Additionally, transposable elements like Insertion Sequences (IS) are frequently found within [175], or around [222] capsular loci, which may facilitate their translocation from the donor's chromosome to MGEs and/or the recipient's chromosome.

Serotype swap has mainly been studied in *Streptococcus pneumoniae*, a naturally competent species with high recombination rates, for which vaccines based on capsule antigens are available. Prevnar started commercializing a highly protective vaccine, PCV**7,** targeting the seven most prevalent

serotypes involved in children infections in 2000 in the USA [253]. However, infections with strains expressing non-targeted serotypes rapidly increased, leading to the release of PCV**13** (13 targeted serotypes) in 2010 and PCV**20** (20 targeted serotypes) in 2021 [254]. Serotype replacement in the pathogen population was shown to occur via clonal expansion of strains carrying non-targeted serotypes, but also via serotype swap [255]. However, the study of hundreds of *Streptococcus* genomes, including non-pathogenic strains, revealed that swaps occurred frequently even before large-scale vaccination programs [256], suggesting that immunity-induced pressures are not the sole driver of serotype swaps.

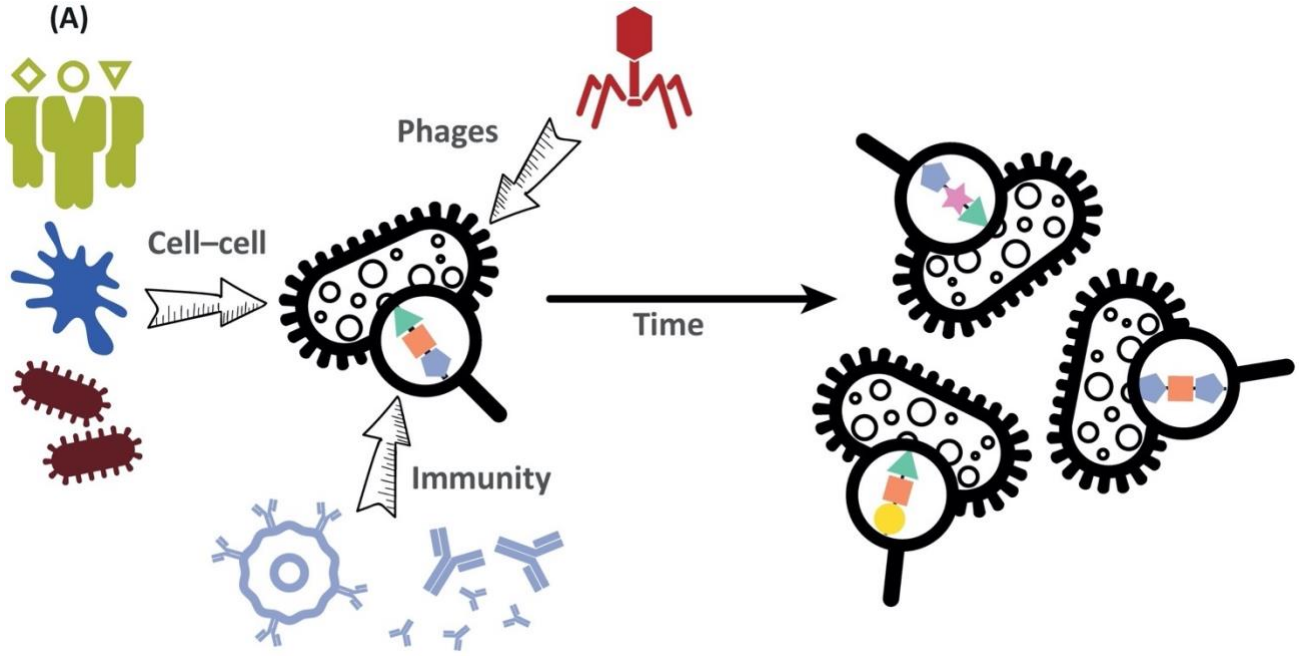


Figure 13 – Candidate selective pressures for the diversification of capsule serotype. Those selective pressures may lead to capsule inactivation, *de novo* serotype birth, or serotype swap. Figure from [118]

Interactions between capsule and MGEs

In capsulated bacteria, the capsule is the outermost layer of the cell envelope making it the first point of contact between virions and conjugative systems. Hence, this physical barrier may also represent a genetic transfer barrier, because the physical structures of MGE (virion, conjugative pili) interact with the cell surface to infect cells (See Cell envelope determinants).

In *Escherichia coli*, which can encode up to four different capsules [209], upregulation of the colanic acid capsule (related to group I capsule) is one of the most general mechanisms of phage resistance [257,258]. This is thought to be due to the thickness-dependent coverage of phage receptors, which disrupt the activity of phage receptor-binding proteins [258]. Other examples include the group II capsule K5 of *E. coli* which provides protection against T4 phage infection [259], and the group I capsule K30 that provide protection against phage O9a by masking the LPS [260]. In *Streptococcus pneumoniae*, which encode a group I capsule with various serotypes, but is frequently found non-encapsulated [243,244], there are fewer studies investigating phage-capsule interaction. Three phages isolated from a non-capsulated strain were able to infect other non-capsulated strains, but none of 41 capsulated cells [261]. In the gut commensal *Bacteroidetes thetaiotaomicron,* which encodes different capsular polysaccharide modulated by phase-variation, the expression of non-permissive capsules is selected under phage predation, enabling survival [262]. Hence, capsules can protect the cell from phage infections by masking surface receptors. Since phages are not only predators but also vectors of HGT, capsules may impact the gene flow.

Capsules are not an insurmountable barrier, and can be overcome by some phages, as evidenced by the large array of isolated phages that infect capsulated strains. For example, *E. coli* expressing the K5 capsule are indeed resistant to T4, but are sensitive to so-called K5-specific phages [263]. For K5-specific phages, the capsule is not a resistance factor, but a sensitivity factor, and non-capsulated mutants are fully phage-resistant [259,263]. In *Klebsiella pneumoniae*, which is ubiquitously capsulated, almost all identified phages depend on capsule production [264–268]. This indicates that phages able to overcome the capsule barrier may be dependent on the capsule for successful infection.

The dependency of phages toward the capsule of their host comes from the characteristics of phage virions and their receptor-binding proteins (RBP) (See Cell envelope determinants). The RBP of capsule-infecting phages are endowed with **capsule depolymerases** that not only bind, but also

specifically cleave capsular polymers [269,270]. They decorate virions in the form of tail-spike proteins connected to the baseplate or the tail, directly via conserved N-terminal domains [271], or indirectly via an adapter protein [272] (Figure 14). Biochemically, capsule depolymerases fall into two enzymatic classes, lyases and hydrolases, but lyases appear to be the most common [269]. Both hydrolases and lyases have strong substrate specificities for di- or tri-saccharides, and given the diversity of capsular polysaccharide they are usually specific toward one serotype [265]. Virions can encode one depolymerase, present in several copies on the virion, or several different depolymerases, present each in one or several copies on the virion [273], so phage host-range tends to broaden with the number of different depolymerases [274]. As of today, capsule depolymerases architecture has been modeled for many phage families of *Klebsiella pneumoniae* [273], their presence can be predicted via sequence homology [270], but their exact substrate specificity still requires chemical characterization [275].

Depolymerases have a modular structure with an N-terminal region involved in virion binding, a middle-region encoding the catalytic site, and a C-terminal involved in di- or trimerization. This modularity support a model where phages can rapidly shift host-range by recombination between catalytic sites, and there are evidence that this process happens in nature [276]. This modularity could be leveraged to engineer depolymerases for therapy [277].

Overall, capsule depolymerases are an important determinant of phage host-range, lead to serotype specificity, and may favor genetic transfer between strains of the same serotype by phage-mediated HGT. Alternatively, phages may encode up to eleven different depolymerases [274], and actually drive inter-serotype genetic exchanges.

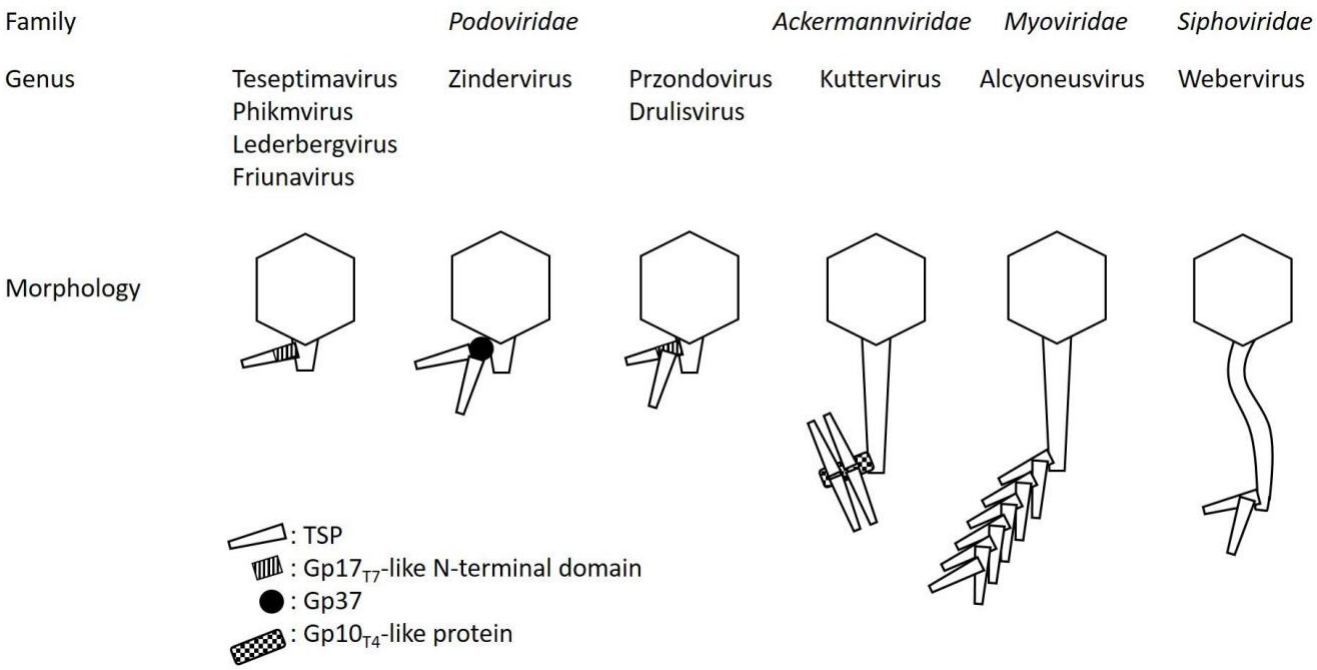


Figure 14 - Capsule depolymerases (tail-spike proteins, TSP) architecture in different phage families. Different mode of attachment to the virion are illustrated, with or without an adapter protein (Gp37, $Gp10_{T4}$-like protein). Figure from [269].

The impact of the capsule and its composition on conjugation has seldom been studied. While theoretically, the capsule represents a physical barrier that may prevent mating-pair stabilization and decrease conjugation rates, capsule-encoding species tend to carry more mobile genetic elements, including plasmids, than non-capsulated species [278]. Moreover, it has been hypothesized that strains of the same capsule serotype may engage in conjugation more efficiently [134], but this remains to be formally tested. Some conjugative systems are associated with type IV pili with lectin-like domains that bind to the oligosaccharide of the LPS [279], suggesting that capsule polysaccharide could be targeted in a similar way. Additionally, conjugative pili usually encode an hydrolase (VirB1) that could play a role in capsule degradation, but this protein appears to be necessary for pilus assembly across the peptidoglycan rather than crossing the peptidoglycan of the donor [280]. Hence, the impact of the capsule on conjugation is poorly understood.

Klebsiella pneumoniae: a sugar-coated playground for MGEs

Klebsiella pneumoniae is a capsulated diderm species belonging to the Enterobacteriaceae family, like the well-known *Salmonella* and *Escherichia* genera. It is an ubiquitous species, and a commensal of the digestive tract of many eukaryotes. It can be found in soil, fresh water, associated with plant roots and leaves, insects and animals including humans [281,282]. Associated with plants, *K. pneumoniae* is a diazotroph capable of fixing atmospheric nitrogen and are often attached to the absorptive hairs of plant roots which benefit from their nitrogen [283]. Associated with mammals like humans, *K. pneumoniae* is a nose, mouth, lung and gut commensal which is part of the healthy bacterial flora [281]. It is, however, an opportunistic pathogen leading to a wide array of diseases in immune-compromised patients [281], and is part of the virulent and multi-drug resistant ESKAPE pathogen group from the WHO, which comprises *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter spp.* [210].

Pathotypes

Four main pathotypes are associated with *K. pneumoniae* strains [246,284]. Classical (**cKP**) strains are commensal strains but also opportunistic pathogens that can be efficiently treated with antibiotics [246]. Multi-resistant, notably extended-spectrum β-lactamases and carbapenemases-producing (**CR-Kp**) strains are nosocomial opportunistic pathogens acquired and spread in clinical settings [199,285]. CR-KP infections are notoriously difficult to treat since they are usually resistant to all antibiotics classes. Hypervirulent (**hvKP**) strains, are community-acquired pathogens causing, among others, liver-abscesses, septicemia and meningitis [284]. Infection by hvKP strains can be efficiently treated with antibiotics, but progress very rapidly and have a high mortality rate [284]. Finally, recently isolated strains show a convergent phenotype called **CR-hvKP**, causing community-acquired deadly infections in healthy patients that are untreatable with antibiotic treatment [286,287]. Such convergent strains seem to have emerged independently during the last few years and represent a serious threat for human health.

Capsule

One the most prominent feature of *K. pneumoniae* is the production of a nearly ubiquitous group I (Wzx/Wzy-dependent) capsule that gives a mucoid aspect to bacterial colonies cultivated *in vitro*

(Figure 15). There is one single capsule locus in *K. pneumoniae*, located between the *galF* and *ugd* core genes, which has increased rate of recombination and HGT compared to the rest of the genome [231,234]. This locus contains conserved genes involved in capsule assembly and export, flanking a highly variable region encoding enzymes that determine the oligosaccharide combination, linkage, and modification (and thus the serotype) [208,232]. To date, the chemical structure of 79 distinct capsule serotype have been determined, and genomic association studies have revealed that the genetic content of the capsule locus is a good predictor of the serotype [232,234]. Such predictions are called capsular locus types (CLTs) to distinguish them from experimentally determined serotypes. Analysis of the capsule locus of thousands of isolates have revealed that there are more than 160 genetically distinct CLTs, which can be easily identified with specialized tools like Kaptive [232,241,242].

K. pneumoniae capsules can extend well beyond the outer membrane, up to 420 nm, which is 140 times the average size of the peptidoglycan layer[288]. They enhance cellular survival to the immune response, including macrophages [216], and antimicrobial peptides [218], being a major virulence factor of the species. Intriguingly, multidrug-resistant lineages exhibit higher capsular diversity than the hypervirulent ones [282], which are almost exclusively of the serotype K1 and K2 [289,290].

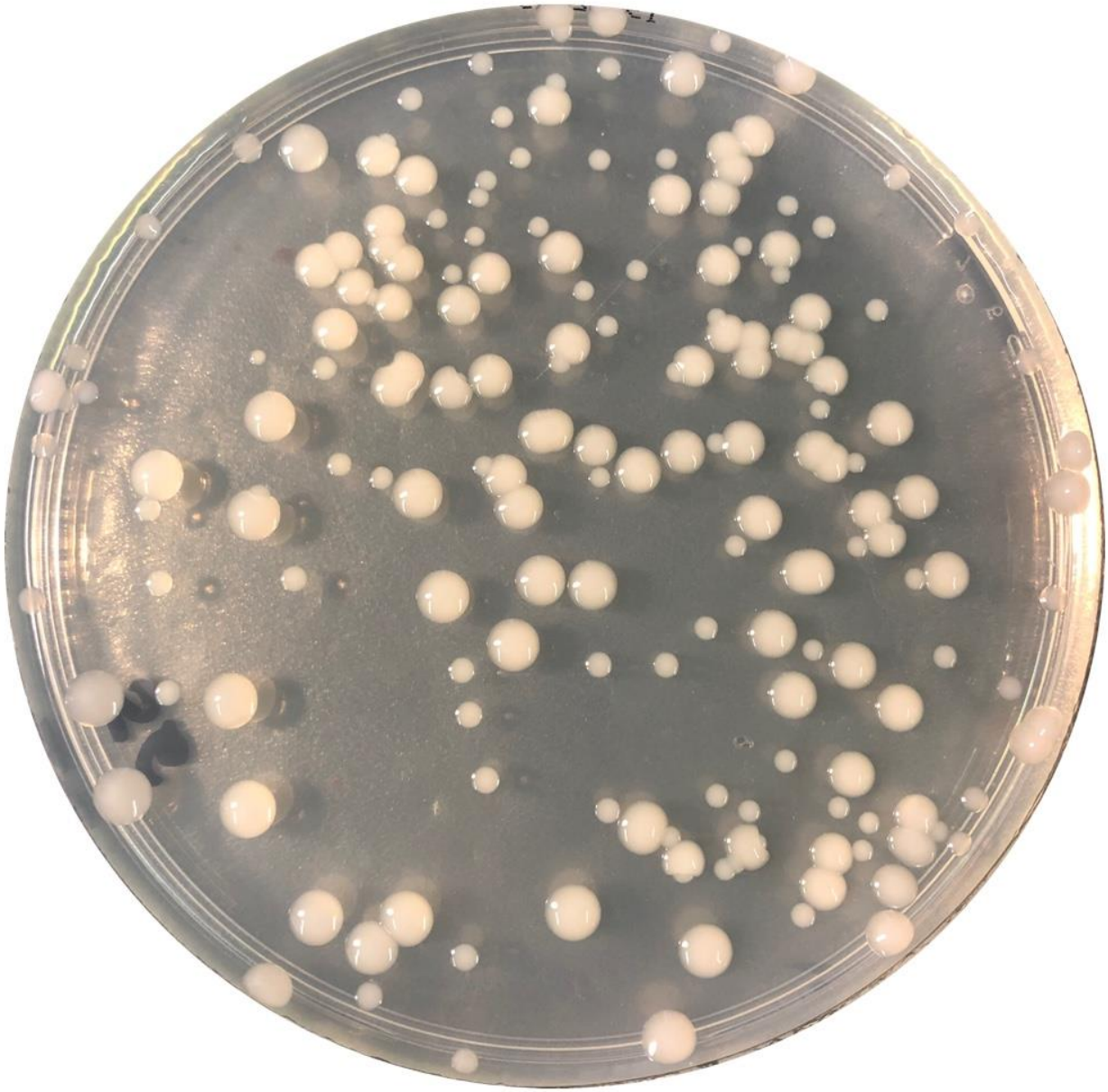


Figure 15 – Bacterial colonies of *Klebsiella pneumoniae* NTUH-K2044 (pathotype HvKP, serotype K1) isolated from a Taiwanese patient with liver abscess and meningitis. Colonies were grown on an LB agar plate at 37c overnight. Larger, mucoid colonies correspond to the wildtype capsulated clones. Smaller, translucent colonies correspond to non-capsulated clones that emerge during *in vitro* cultivation.

Genomics

K. pneumoniae genomes are between 5-6Mpb in size, and encode between 5000-6000 genes. The *K. pneumoniae* species tree, based on 1000-2000 core genes, comprises hundreds of deep-branching

lineages with *c.a.* 0.5% nucleotide divergence [282]. These lineages correspond to independent clones, or clonal groups, which had historically been defined with multi-locus sequence typing schemes and are associated with various traits and pathotypes [173,291]. In addition to the 1000-2000 genes core genes, *K. pneumoniae* harbors a massive genetic diversity in its accessory genome, and its pan-genome is estimated in the order of 100,000 distinct gene families. This genetic diversity stems from a seemingly large propensity for the species to acquire and distribute MGEs, and hence engage in horizontal gene transfer. This is evidenced by the fact that on average, *K. pneumoniae* isolates carry between four to six distinct plasmids, which is more than other Enterobacteriaceae [292], as well as between four to six distinct prophages [293] and often encode ICEs [294]. Chromosomal recombination is also frequently observed within [175], but also between [295], clonal groups and frequently involves the capsule locus [175]. As an extreme example, one clinical isolate was found to be a genetic chimera with 2,9Mpb of its chromosome belonging to clonal group 23, and 2.4Mpb belonging to another unknown clonal group [295].

Evolution of resistance and virulence

The evolution of *K. pneumoniae* into dangerous pathotypes like CR-KP, hvKP and CR-hvKP is the result of the acquisition of antibiotic resistance genes (ARG) and virulence factors [282]. However, evolution of antibiotics resistance appears to be distinct from the evolution of hypervirulence [290], leading to a structuration of the population between classical, resistant and virulent clones. Convergence of resistance and virulence within the same clones may be predictable due to preferential evolutionary paths [286,290]. Numerous novel ARG were first discovered in *K. pneumoniae*, especially ones conferring carbapenem-resistance such as KPC, OXA-48 or NMD-1. Those genes were later identified in other species across the world, such as in the other members of the ESKAPE group. Hence, the wide ecological distribution and high HGT rates of *K. pneumoniae* places it as a key trafficker of ARG [292]. Many different clonal groups are associated with antibiotics resistance, which suggest that the genetic background plays a minor role in the evolution from cKP to CR-KP [282]. Accordingly, multi-resistant isolates harbor diverse capsule serotypes. ARG are mainly encoded within transposons and integrons [296] carried on conjugative plasmids [36], which facilitate their spread via diverse routes, represented on Figure 16.

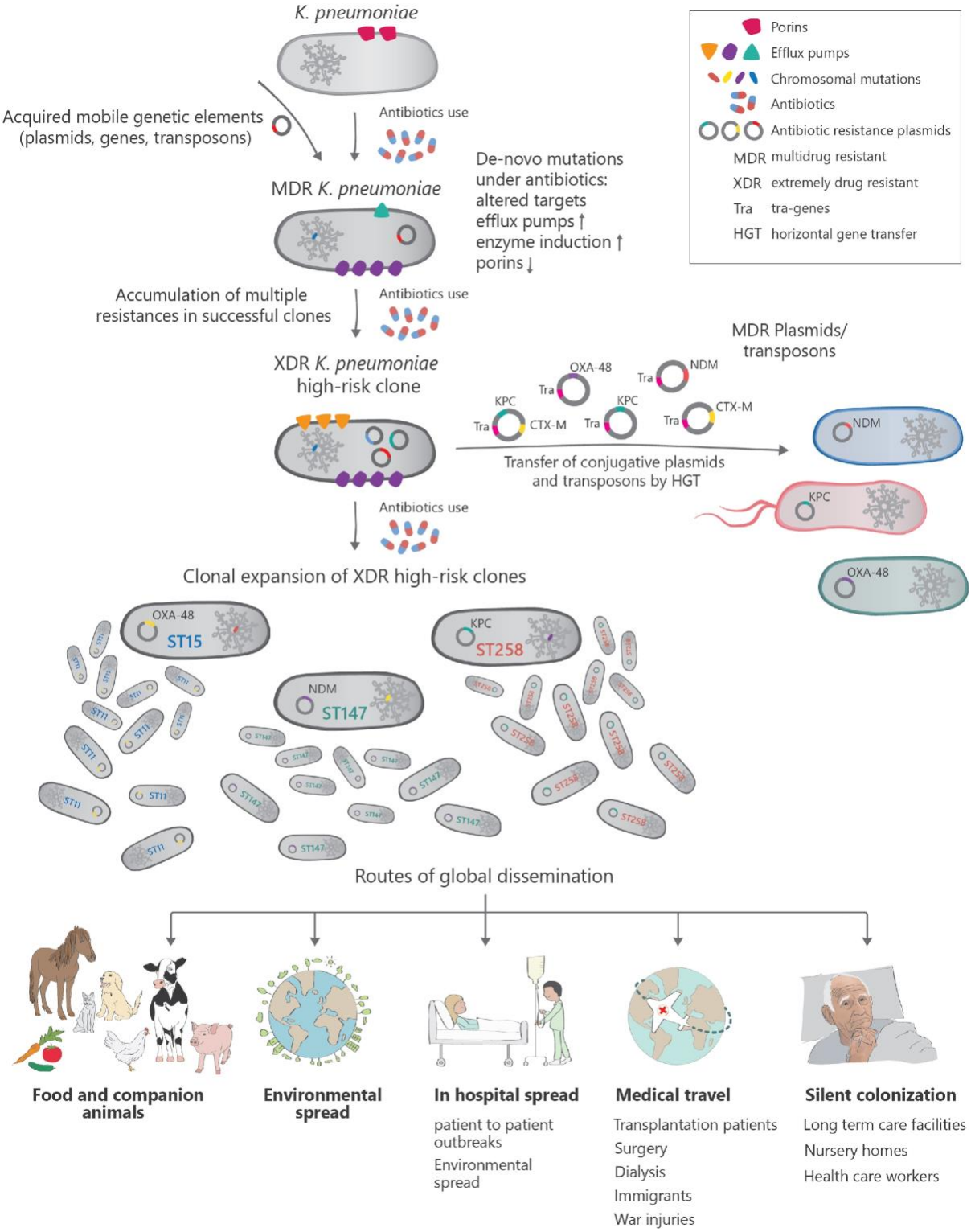


Figure 16 - Routes of global dissemination of antibiotics resistance by *Klebsiella pneumoniae*. Figure from [199].

K. pneumoniae is generally an opportunistic pathogen, but some hypervirulent hvKP clones have the ability to provoke community-acquired infections. Those strains are characterized by several features: capsule serotype, clonal group and an array of virulence factors. There are two **capsule serotypes**

generally associated with HvKP clones: K1 and K2. Those two serotypes have been shown to promote *K. pneumoniae* infection in the host [289,295,297]. While **K1 hvKP** are monophyletic, belonging to the clonal group CG23 [295], **K2 hvKP** are more genetically diverse and belong to five distinct clonal groups [290]. Both K1 and K2 hvKP strains harbor multiple virulence factors, including the colibactin toxin, the aerobactin, salmochelin and yersiniabactin siderophores, and the RmpA/A2 regulators of mucoid phenotype [173]. Those virulence factors can be encoded on so-called virulence plasmids, like the 219kb pLVPKP [298], which are not conjugative but may be mobilized by other conjugative elements [299]. Virulence factors are also encoded on integrative and conjugative elements, like ICEKp1 which carries a similar virulence region than pLVPKP but is able to self-mobilize and spread by conjugation [300].

While there was little overlap between multidrug resistant *vs*. hypervirulent lineages before the 2010s, recent studies have reported the emergence of carbapenem-resistant hypervirulent strains leading to high mortality rates [301]. Mechanisms for the emergence of CR-hvKP can be summarized in three patterns: (i) CR-KP acquiring a hypervirulent phenotype; (ii) hvKP acquiring a carbapenem-resistant phenotype; and (iii) cKP acquiring both a carbapenem resistance and hypervirulence hybrid plasmid. Those three patterns depend on MGE transfer and highlight the importance of understanding HGT in bacterial pathogens populations.

Aim of this thesis

Hypothesis

The main hypothesis driving my work is that capsules play a major role in bacterial adaptation. Indeed, they facilitate survival by protecting against biotic and abiotic stresses. Moreover, they might modulate the rates of genetic exchange because they represent a physical barrier to cross for mobile genetic elements. This hypothesis is supported by recent findings that (i) capsules are more frequent in environmental bacteria than in pathogens (ii) bacteria with capsules occupy more diverse niches, (iii) capsulated bacteria have larger pan-genomes, and (iv) accumulate more mobile genetic elements [209,278].

There is thus a conundrum regarding capsule evolution: the capsule needs mobile genetic elements to vary by horizontal gene transfer, but may block the acquisition of the very same mobile genetic elements. Hence, capsule's impact on gene flow may also have consequences regarding capsule evolution.

Objectives

The aim of this thesis is to characterize the interplay between the bacterial capsule and horizontal gene transfer. More precisely, I wanted to shed light on the impact of the capsule on the gene flow between bacterial populations, and how this process may shape capsule evolution. *Klebsiella pneumoniae* represents a model of choice, because it ubiquitously expresses a large capsule, often encodes MGEs, and its evolution represents a threat for our global health system. The objectives of my thesis were the followings:

- **Infer the past genetic exchanges** between *Klebsiella pneumoniae* sequenced isolates, in relation to their capsule.
- **Measure the rates of genetic exchanges** in *Klebsiella pneumoniae* in relation to their capsule.
- **Understand capsule evolution** in the light of interactions with mobile genetic elements.

Contributions

My PhD work has led to the publication of one published and one soon-to-be submitted first-author articles, which rely on two methodological backbones: comparative genomics and experimental biology. In the first study, I leveraged tools from the comparative genomics field to analyze thousands of *Klebsiella pneumoniae* genome assemblies, infer their past genetic exchanges and propose a model for the evolution of the capsule. I completed my findings with a simple experimental model that supported the conclusions of the genomics analysis. In the second study, I used a mirror approach where I started from experimental genome engineering and *in vitro* gene transfer assays and then completed my findings with a genomic analysis of *Klebsiella pneumoniae* plasmids. Hence, my contributions on the **interplay between bacterial capsules and horizontal gene transfer** will be first through **the lens of genomics**, and then through the **lens of experimental biology**. I have also participated in several other studies, which are presented in the annexes.

The interplay between bacterial capsules and horizontal gene transfer

Through the lens of genomics

Introduction

Genomics is a broad, and relatively recent, field of biology consisting in the study of the genome sequences. It relies on experimental data, mainly DNA sequences from organisms generated via sequencing, and their analysis with computational tools. Comparative genomics is a subfield of genomics consisting in the comparison of related organisms' genomes, the inference of their evolutionary relationships and the study of their evolution. At the beginning of my thesis, more than 5,000 draft assemblies of *Klebsiella pneumoniae* were publicly available in the RefSeq database of the NCBI. Using a recent software called Kaptive [232,241,242], able to precisely infer the capsule serotype from genome assemblies of *K. pneumoniae,* I first investigated the interplay between capsules and HGT using comparative genomics, with the following aims:

- Characterize the impact of the capsule and its serotype on past genetic exchanges in *K. pneumoniae* sequenced genomes.
- Understand the dynamics of evolution of the capsule, notably the mechanisms at play leading to serotype swaps.

In the following section, I will introduce genomics concepts, along with the tools and methods I used to investigate my questions.

Genome assemblies

Most sequenced bacterial genomes are generated from the DNA extracted of a clonal culture, via *de novo* assembly, which is the process of reconstructing the original DNA sequences using only the sequencing reads from this isolate. Sequencing reads are fragments of sequenced DNA varying in size depending on the sequencing method. Illumina sequencing methods, also called short-read sequencing, typically produce 100-250bp reads which can be paired, meaning that DNA fragments are sequenced from each extremity, providing information on sequence contiguity. PacBio and Nanopore sequencing, also called long-read sequencing, produce reads from 5,000 – 30,000bp on average. Assembly is typically performed with the following steps (Figure 17): finding overlaps between reads, building a

graph with all read connections, simplifying the graph and traversing the graph to trace a path corresponding to the consensus sequence. Assembly graphs are sensitive to sequencing errors, which can introduce false edges and nodes, to sequence heterogeneity in the case of contamination or polymorphisms. They are also sensitive to DNA repeats, which form highly connected nodes in the graph. The impact of DNA repeats is critically determined by the length of the reads and of the repeats. This may be known as the law of repeats, as stated by Torsten Seemann:

"It is impossible to resolve repeats of length S,
unless you have reads longer than S.
It is impossible to resolve repeats of length S,
unless you have reads longer than S."

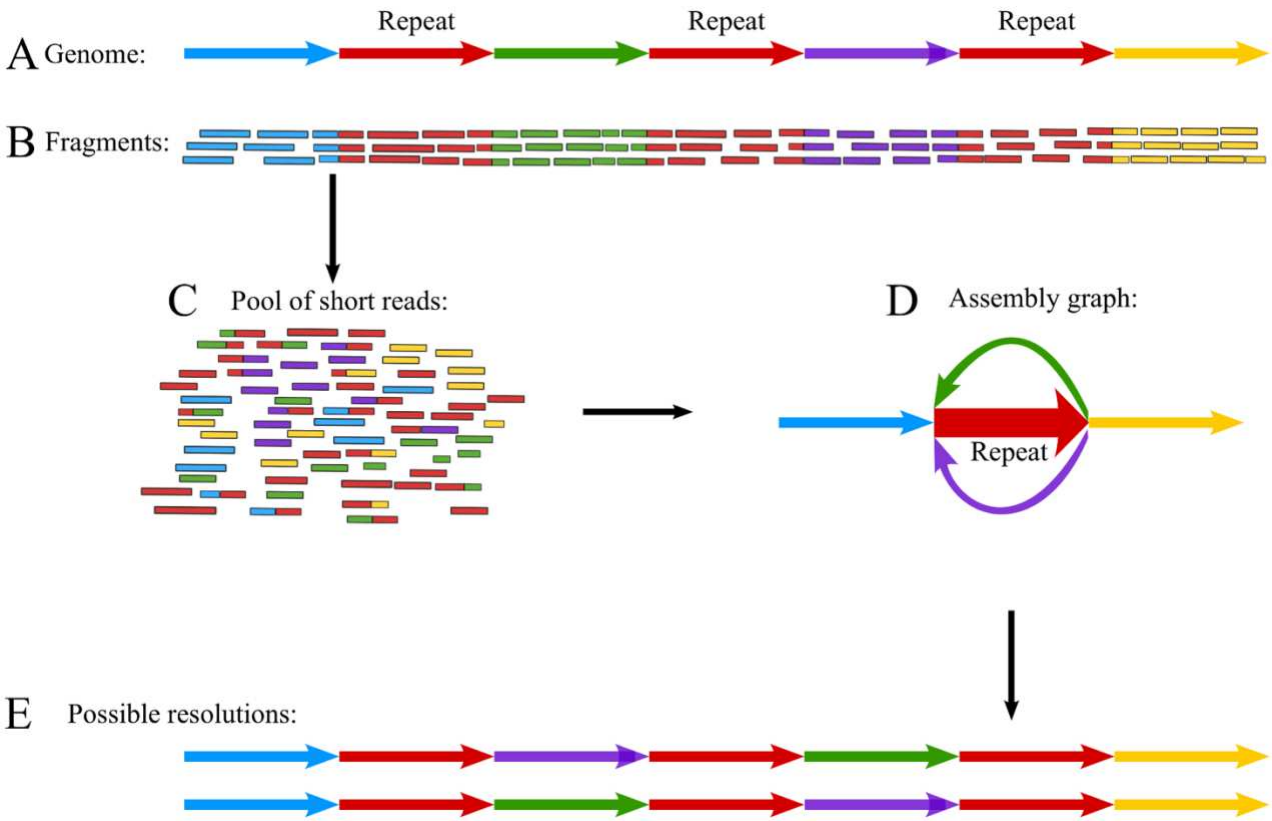


Figure 17 – Genome assembly at repeats larger than sequencing reads. A. The true genome representation. B. Fragmentation of the genome prior to sequencing. C. Sequencing reads. D. An example of an assembly graph constructed by read overlapping, arrows represent the bubble formed

by repeats in assembly graphs. E. The graph cannot be resolved by a single resolution, breaking the assembly. Figure adapted from [302]

The end-result of sequencing and assembly is a text file containing the sequence coded in ATCG, in continuous sequences called contigs, and can be classed in two categories. If the assembly is not fully resolved and the DNA molecule(s) are split between several contigs, it is called a **draft assembly** (gapped assembly). If the assembly consist of one contig per DNA molecule, it is called a **complete assembly** (un-gapped assembly). In bacteria, a complete assembly is typically composed of a single contig corresponding to the chromosome, along with a single contig for each plasmid. The quality of a draft assembly can be measured via different metrics, which may require prior knowledge or not. Completeness can be gauged using a set of genes that are universally distributed as orthologs across particular clades of species [303], for example ribosomal proteins and rRNA genes. Assembly quality can also be measured via the number of total contigs, overall genome length, biggest contig size, or more complex metrics such as Nx, equals to the largest contig length, L, such that using contigs of length longer than L accounts for at least x% of the bases in the assembly [304].

Genome assemblies are generally deposited in public databases, such as the RefSeq [305] database from NCBI which provides high quality, curated assemblies. Criteria to be included in RefSeq are quite stringent to ensure the overall quality of the dataset, and are listed here:
https://www.ncbi.nlm.nih.gov/assembly/help/anomnotrefseq/
This list includes, among others:

- **Normal gene to sequence ratio** - the ratio of the number of predicted genes to the length of the genome divided by 1000 is far outside the usual range for a Complete Genome assembly. The NCBI Prokaryotic Genome Annotation Pipeline typically expects to find an average of one gene for every 1,000 nucleotides in a genome assembly. The typical range is 0.8 to1.2; anything outside the range 0.5 to 1.5 is considered abnormal.
- **Assembly fragmentation** - a prokaryotic assembly with contig L50 above 500, contig N50 below 5000, or more than 2,000 contigs.
- **Presence of key genes** – including ribosomal proteins, rRNA and tRNA genes.

In 2022, around 250,000 bacterial genomes assemblies had been included in RefSeq, spanning 12,500 different species. The number of deposited sequences in public databases is increasing rapidly, and is currently outpacing our analysis capabilities. At the beginning of this thesis in September 2018 there

were about 5,500 *Klebsiella pneumoniae* genome assemblies on RefSeq. By the end of this thesis in 2022, there were more than 13,000 assemblies.

Genome annotation

Genome assemblies are simple text files with the sequence organized in contigs and sometimes scaffolds, which represent contig links. For instance, a bacterial chromosome should be one contig, but can be split into several contigs if the assembly is incomplete. While transcription factors and tRNA can read the code *in vivo*, researchers have developed informatics tools to decode genomes *in silico*. This process is called genome annotation, and starts with several general steps:

- Identifying protein-coding regions (Open reading frames, ORFs)
- Identifying non-coding, functional regions (*e.g.* regulatory elements)
- Associating biological information to these elements.

Differentiating protein coding from non-coding regions relies on **Open Reading Frame prediction**. Several tools have been developed to call the ORFs of bacterial genomes. During my thesis, I used a software called **Prodigal** [306], which starts by identifying all putative ORFs and then uses multiple rounds of dynamic programming to refine the selection and learn their upstream motifs from the input sequence. Prodigal presents the advantage of having a clear set of biologically relevant rules, is highly accurate, fast and consistent across assemblies. It also translates each ORF into its corresponding amino-acid sequence using the adequate translation table. There are known pitfalls in prokaryotic ORF identification. First, there can be several canonical, and sometimes non-canonical, start sites for a given protein-coding region, and prediction tools are biased toward canonical start sites too far upstream [307]. Secondly, short protein-coding regions are frequently missed because short ORFs (<150-300bp) lack the strong statistical signals allowing successful identification and may overlap with larger ORFs [307,308].

Predicting the function of a protein is typically performed by finding homologous sequences with experimentally characterized functions. **Sequence homology** refers to sequence similarity due to a common evolutionary origin. Homology-based function prediction is typically performed by sequence similarity search with specialized tools like Blast [309], HMMER [310], MASH [311], or MMSEQS2 [312].

Blast can be used to search nucleic acid (Blastn) and protein (Blastp) sequences against a database sequence and uses short words local alignments that are extended and scored according the match similarity [309]. Blast can perform sequence search with nucleotides and proteins sequences in all directions (Blastx, tBlastn, tBlastx). However, remote homologs with very low sequence similarity are hard to find with Blast.

HMMER is a tool that relies on the conversion of the query in a profile Hidden Markov Model (HMM), which are probabilistic models of single sequence or an alignment of sequences [310]. Profile HMM capture position-specific sequence conservation, changes, gaps and insertions. They are well suited to detect divergent homologous sequences because they are more sensitive than simpler local alignment algorithms such as Blast. Profile HMM can be made from protein or nucleic acid sequence, and used to search protein or nucleic acid databases.

The distance between two sequences can also be computed without alignment, typically by reducing the sequence(s) to compressed sketch representations, as in the **MASH** software [311]. Sketches are made from small words called k-mer, which are hashed, *i.e.* linked to an identifier. Alignment-free methods such as MASH are typically several orders of magnitude faster than alignment-based methods to detect similar sequences, at the cost of accuracy, which is particularly useful for the clustering of very large datasets.

Alignment-free methods can be leveraged to speed-up the computational load associated with sequence alignment. In this optic, **MMSEQS2** [312] searching is composed of three stages: a k-mer sketching and matching step, vectorized un-gapped alignment, and gapped alignment. The first stage is crucial for the improved performance, because it filters out most of the comparisons before the computationally intensive steps of sequence alignment.

Genome annotation extends further than ORF calling and function prediction. Proteins often function in **macromolecular systems**, which are involved in key cellular processes. Those macromolecular systems can be nanomachines (*e.g.* ribosomes, flagellum) or molecular pathways (*e.g.* capsule biosynthesis). Genomes also contain mobile genetic elements, which often encode macromolecular systems such as conjugative pili, or viral particles. The systematic detection of such systems is not an easy task, because of a number of reasons [313]:

- Systems are made of different component with different dispensability (some are essential, some are not)
- Key components may have homologs in other systems
- Systems components may evolve at very different rates (some have very conserved sequences, some have very variable sequences)

These difficulties can be partly circumvented by searching for the whole set of components of the system because the detection it should lead to more accurate inference. This is especially relevant if the genes encoding these components are organized in highly conserved ways. In bacterial genomes, genes are often organized in organized operon [314], and macromolecular systems are often encoded in one or several contiguous operon [315,316] ensuring tight regulation, correct function, and potentially co-transfer [317]. Overall, complex systems encoded by bacterial genomes are generally in linked **multi-genic systems** which can be detected by searching for its individual components with specialized tools such as **MacSyFinder** [313,318]. Example of multi-genic systems encoded in one locus include capsule loci and conjugative systems. In the context of detecting multiple contiguous genes spanning several Kb of DNA, draft genomes present an additional challenge to the identification of complex systems because they may split linked genes into different contigs, abrogating the positional information associated with the genes.

Mobile genetic elements such as plasmids and prophages can also be seen as contiguous multi-genic systems. In draft genomes, they are more difficult to detect, because they are frequently split into several contigs.

Plasmid detection is generally straightforward in complete genomes, because plasmids form circular contigs. However, plasmids often encode insertion sequences and are thus rich in DNA repeats [152], and are frequently split between contigs in draft assemblies from short-read sequencing. Several approaches have been developed to classify contigs between plasmids and chromosome. Specific plasmid sequences, such as replication and segregation systems, can be used as **plasmid-specific "probes"** [319] but such methods may only detect one out of several contigs corresponding to a given plasmid. Another method relies on the nucleotide composition, like **k-mer frequencies**, to predict if the contig resembles more to the chromosome, or plasmids. These methods usually require the training of machine learning models with complete genomes for which chromosomal and plasmid contigs are known [320].

Prophage detection consists in the identification and delimitation of inserted phage genomes. Several tools have been developed for this task, such as VirSorter [321] and PHASTER [322]. PHASTER relies on two distinct databases, one for phage-specific genes clustered from known temperate phages, and one for bacteria-specific genes, from persistent genomes. First, it identifies putative prophages by searching for phage gene clusters genes, then searches for an integrase and if it finds one, scan the region for small repeats indicative of *att* sites. If putative *att* sites are found, they are considered as the border of the element. If an integrase or *att* sites are not found, the region is delimited via a bi-

directional sliding window that stops when no phage genes are found. The region is then given a score, based on the number of key phage features such as integrase, tail or baseplate proteins. Given the large diversity of phages, this method may only be accurate to identify regions similar to known phages. The delimitation may also be inaccurate, especially if no *att* sites are found, or if multiple prophages are in close distance. Finally, prophages split into two or more contigs may not be detected, or may be given a low confidence score.

Overall, systematic genome annotation is a complex task requiring a multitude of specialized tools, and is negatively impacted by genome assembly fragmentation.

Pan- and core-genome

The the pangenome is the union of all gene families present in a set of genomes, while the core genome is the intersection. Computationally, pangenomes are typically built by clustering the predicted proteins of all the genomic dataset based on their sequence similarity (Figure 18). Further refinement based on ORF synteny may help differentiating orthologues from paralogues, but such analysis is difficult in draft genomes and species with high rates of genome re-organizations. In this thesis, I used the PanACoTa tool [323] to download complete and draft assemblies of *K. pneumoniae* from RefSeq, annotate the ORFs, and build the pan and core genomes.

PanACoTa [323] relies on the clustering algorithm of MMSEQS2 to build pan-genomes. Estimating gene families requires several input parameters, such as the minimum coverage and sequence identity threshold. While low values of coverage and identity lead to the clustering of non-orthologous genes, higher values may lead to overestimation of the pan-genome size by separating groups of orthologous genes [323,324]. Hence, the most adapted thresholds may be very different when comparing diverse or closely related genomes, but also when attempting to cluster gene families with different evolutionary rates. For a single species pangenome, it is generally accepted that >80% bidirectional sequence coverage and identity is appropriate to group the majority of orthologs [168].

Another contentious point for the appropriate clustering of gene families is the impact of pseudogenes. Diverse mutations (In/Del, SNP, and generally frameshift mutation) lead to the presence of early stop codons in ORFs, resulting in truncated proteins which may not pass the coverage threshold set for the

clustering method. In that case, most truncated ORFs may not cluster within their correct gene family, leading to the creation of artifactual, often singleton, families. This problem may be solved by a protein-to-DNA sequence search [325], whereby gene family clusters are corrected when the DNA sequence reach high coverage and identity, while the protein sequences only reach the identity threshold.

The definition of the core genome is sometimes too strict to identify the intersection of gene families across a set of genomes. For example, any assembly error leading to the absence of one family in one particular genome will exclude the family from the core. Hence the core genome is typically approximated by the persistent genome, defined as gene families present in more than X% genome assemblies. This threshold is usually set between 90-99% of genomes, depending on the size and quality of the dataset, because of potential sequencing errors or missing gene in draft assemblies.

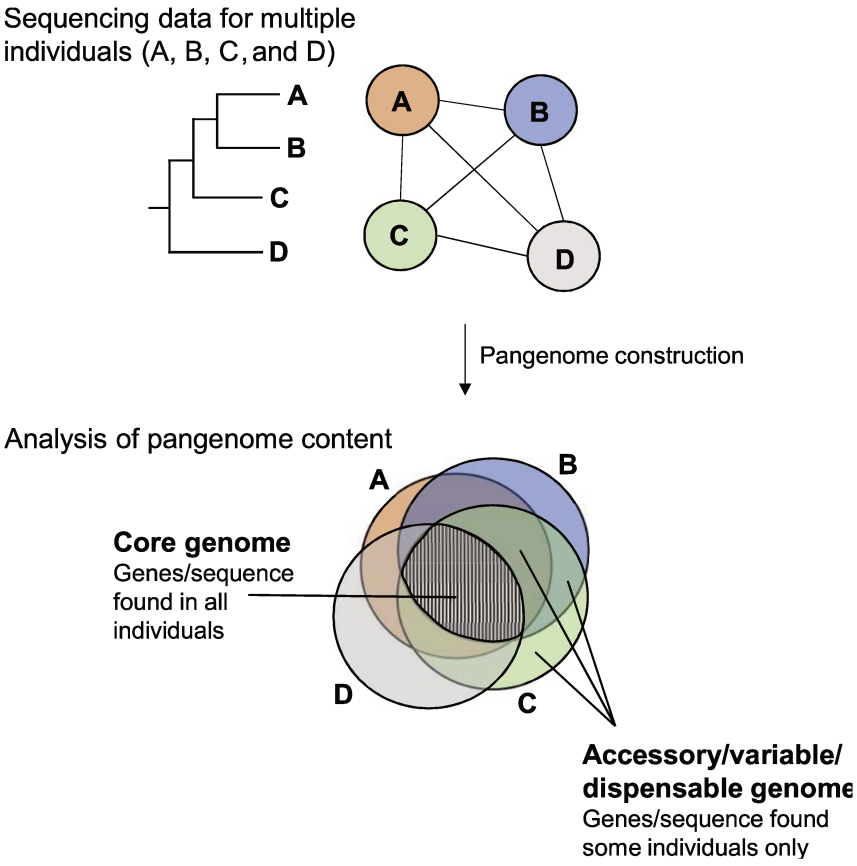


Figure 18 – Pan-genome inference from genomic datasets. The graph represents an orthologous gene present in all four genomes. [326]

Open-source script: GALOPA: Gain, Loss, Persistence and Absence pangenome mapping across phylogenetic trees.

1 – Considerations

Bacterial genomes are very plastic and often acquire and lose genes with passing generations. Gene family acquisitions are particularly interesting because they arise from horizontal gene transfer events. Since I wanted to **quantify the gene flow** between strains of *K. pneumoniae* in relation with their capsule state and serotype, I built an automated pipeline called GaLoPA to map all the pangenome families on a phylogenetic tree (with PanACoTa), reconstruct their presence/absence at each branch of the tree (with Pastml) and infer the gene **GA**in, **LO**ss, **P**ersistence, **A**bsence state for all gene families for all branches. Gene flow can then be represented by co-gains, *i.e.* pairs of genomes having acquired the same gene-family, and co-gains can be classed according to other states in the tree like for example the capsule serotype.

2 – Method

A – Data needed

The method relies on a dataset of genomes, either complete or draft, from the same species. This dataset must first be processed through an annotation pipeline to identify the ORFs and generate the pan- and core-genome, for example with the PanACoTa framework. The required outputs from PanACoTa are:

- The pan-genome table
- A core(/persistent) gene concatenate alignment for phylogenetic inference

The core gene concatenate alignment can be used to generate a phylogenetic tree. To do so, I used IQ-Tree with the automated ModelFinder algorithm, and the ultra-fast bootstrap to assess the robustness of the branches. The tree must be rooted, which can be done by adding a small group of genomes belonging to another species (Ideally, a closely related species, for example one of *Klebsiella pneumoniae*'s most related species is *Klebsiella quasipneumoniae* subsp. *quasipneumoniae*).

B – Softwares needed

- PanACoTa [323] to download, annotate, build the pan/core genomes, and concatenated alignment
- IQ-Tree [327] for phylogenetic inference
- PastML [328] for ancestral reconstruction of gene families presence/absence
- The GaLoPA script with the following dependencies:
    - Phangorn [329]
    - Tidytree (https://github.com/YuLab-SMU/tidytree)
    - Dplyr [330]

C – Ancestral State Reconstruction

This method relies on PastML to infer ancestral characters on a rooted phylogenetic tree with annotated tips, using maximum likelihood. Maximum likelihood approaches are based on probabilistic models of character evolution along tree branches. From a theoretical standpoint, these methods have some optimality guaranty, at least in the absence of model violations [328].

PastML presents several advantages compared with other tools. It is particularly adapted for the ancestral reconstruction of binary traits in large trees, because it is orders of magnitudes faster, but as accurate, as previous tools. It also provides a novel method, called marginal posterior probabilities approximation (MPPA) which does not rely on a predefined threshold on the probabilities. Indeed, MPPA chooses a subset of likely states that minimizes the prediction error for every node. It may keep multiple state predictions per node but only when they have similar and high probabilities. Hence, nodes with both “presence” and “absence” states are treated as “unknown” events in GaLoPA.

The GaLoPA script can perform:

- Adequate naming of the tree nodes (keeping the bootstrap support values)
- Tree rooting via the *midpoint* function from R package Phangorn (if the tree was not rooted beforehand)
- Removal of singleton gene families (gene families present in only one genome)
- A fast function to generate the presence/absence of each gene families for each genome in separate tables compatible with PastML
- Automated parallelization of PastML on a computing cluster

- Concatenation of the results into a complete presence/absence/unknown table for each gene family in each node of the tree, including the singletons.

Note: This step is computationally intensive. PastML can infer the ancestral state of binary traits in large trees (1000-10,000 tips) in approximately 5 minutes with the MPPA method. For a 4,000 genomes dataset of *K. pneumoniae* with approximately 80,000 distinct gene families, it would take more than 6 months on a standard computer. By launching each gene family in parallel on a computing cluster, this step can be as fast as one hour.

D – Gain, Loss, Persistence, Absence mapping

While ancestral states are inferred per node, evolutionary events such as gene gains must be inferred per branches. The script is designed to compare each node with its parental node. The rules defining events are simple:

Parental node →	Offspring node:	**Event**
Absence	→ Presence	**Gain**
Presence	→ Absence	**Loss**
Presence	→ Presence	**Persistence**
Absence	→ Absence	**Absence**

E – Comparison with Count

To check the reliability of the GaLoPA method, we compared it with Count [331]. Count is a software package for the evolutionary analysis of gene family sizes (phylogenetic profiles), or other numerical census-type characters along a phylogeny. We used a dataset of 4,000 *K. pneumoniae* genomes. We split the species tree (cuttree function in R, package stats) in 50 smaller groups and, for the groups that took less than a month of computing time with Count (2,500 genomes), we compared the results of Count to those of PastML. The 2 methods were highly correlated in term of number of inferred gains per branch (Spearman correlation test, Rho = 0.88, p-value < 0.0001).

F – Limitations

There are a number of limitations in this method:

- This approach is dependent on the accuracy of the species tree and the clustering of the gene families.
- The evolutionary model I used is the Felsenstein, 1981 model, where the rate of changes from i to j ($i \neq j$) is proportional to the equilibrium frequency of j. This is the default, recommended model for PastML, but I did not explore the other models.
- It considers that a gene family is either present or absent. However, a genome can contain several genes corresponding to the same gene family, obscuring evolutionary events. To account for this limitation, the copy number of gene families in the tips of the tree is present in the output, so those families can be filtered out.

3 – Summary

The approach is schematized below in Figure 19, for a phylogenetic tree and a single gene family ***g***. The initial presence/absence of ***g*** is only known for terminal branches, presence of ***g*** is represented with a blue circle. The ancestral reconstruction performed by Pastml can predict which internal node of the tree harbored ***g***, as represented by internal blue circles. A tree-wide comparison between parent/offspring nodes can predict if the branch incurred a gain or a loss of ***g***, or if ***g*** persisted or was absent, as represented by stars on the branches. In this example, the most likely scenario is that ***g*** was acquired twice, and hence co-acquired once. The annotated output of GaLoPA is represented on Figure 20.

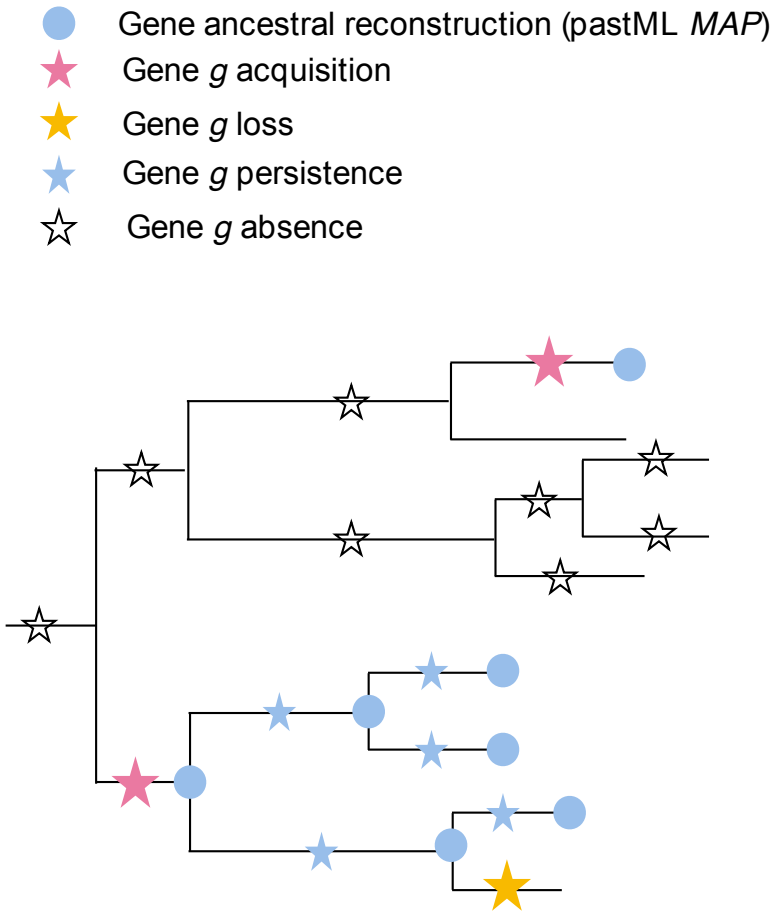


Figure 19 – GaLoPA approach example scheme.

label	name	child_state	label_parent	parent_state	event	branch.length	ufb	copy_number	is_terminal
ACHA001.0321.00003	fam_2011	1	n2	0	gain	0.01374629		1	TRUE
ACHA001.0321.00003	fam_2225	0	n2	1	loss	0.01374629		0	TRUE
ACHA001.0321.00003	fam_1969	1	n2	1	persistence	0.01374629		1	TRUE
ACHA001.0321.00003	fam_2228	0	n2	0	absence	0.01374629		0	TRUE
ACHA001.0321.00003	fam_3263	1	n2	0	gain	0.01374629		2	TRUE
n3	fam_1969	1	n2	1	persistence	0.003032434	100		FALSE
n3	fam_224	1	n2	0	gain	0.003032434	100		FALSE

Label	Branch label
Name	Pangenome family ID
Child_state	Presence/absence in the branch (label)
label_parent	parental node name
parent_state	Presence/absence in the parental branch
Event	gain / loss / persistence / absence
branch.length	branch length (label)
Ufb	UFB value of the label (None if root/terminal)
copy_number	Copy number of the gene family in this branch (only in terminal branches)
is_terminal	TRUE if the label correspond to a terminal branch (i.e. a genome)

Figure 20 – The annotated output of the GaLoPA pipeline. Each line corresponds to a gene family ("name"), and a branch, with the "label" column corresponding to the offspring node, and the "label_parent" to the parental node.

3 – Availability

The scripts are available at:

https://github.com/matthieu-haudiquet/galopa

Research article: Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen Klebsiella pneumoniae.

Matthieu Haudiquet*, Amandine Buffet, Olaya Rendueles, Eduardo P. C. Rocha

Published in PLOS BIOLOGY.

This article aims at understanding the impact of the bacterial capsule and of the capsule serotype on the gene flow in *Klebsiella pneumoniae* populations. It also provides a new model for the evolution of the capsule by highlighting the role of non-capsulated variants as a missing link during serotype swap.

PLOS BIOLOGY

RESEARCH ARTICLE

Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen *Klebsiella pneumoniae*

Matthieu Haudiquet[1,2]*, **Amandine Buffet**[1], **Olaya Rendueles**[1☯], **Eduardo P. C. Rocha**[1☯]

1 Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris, France, **2** Ecole Doctoral FIRE–Programme Bettencourt, CRI, Paris, France

☯ These authors contributed equally to this work.
* matthieu.haudiquet@pasteur.fr

Citation: Haudiquet M, Buffet A, Rendueles O, Rocha EPC (2021) Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen *Klebsiella pneumoniae*. PLoS Biol 19(7): e3001276. https://doi.org/10.1371/journal.pbio.3001276

Academic Editor: J. Arjan G. M. de Visser, Wageningen University, NETHERLANDS

Received: January 13, 2021

Accepted: May 7, 2021

Published: July 6, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pbio.3001276

Data Availability Statement: All individual quantitative observations (S1 to S18 Dataset) that underlie the data summarized in the figures and results of the manuscript are available at: https://

Abstract

Mobile genetic elements (MGEs) drive genetic transfers between bacteria using mechanisms that require a physical interaction with the cellular envelope. In the high-priority multidrug-resistant nosocomial pathogens (ESKAPE), the first point of contact between the cell and virions or conjugative pili is the capsule. While the capsule can be a barrier to MGEs, it also evolves rapidly by horizontal gene transfer (HGT). Here, we aim at understanding this apparent contradiction by studying the covariation between the repertoire of capsule genes and MGEs in approximately 4,000 genomes of *Klebsiella pneumoniae* (Kpn). We show that capsules drive phage-mediated gene flow between closely related serotypes. Such serotype-specific phage predation also explains the frequent inactivation of capsule genes, observed in more than 3% of the genomes. Inactivation is strongly epistatic, recapitulating the capsule biosynthetic pathway. We show that conjugative plasmids are acquired at higher rates in natural isolates lacking a functional capsular locus and confirmed experimentally this result in capsule mutants. This suggests that capsule inactivation by phage pressure facilitates its subsequent reacquisition by conjugation. Accordingly, capsule reacquisition leaves long recombination tracts around the capsular locus. The loss and regain process rewires gene flow toward other lineages whenever it leads to serotype swaps. Such changes happen preferentially between chemically related serotypes, hinting that the fitness of serotype-swapped strains depends on the host genetic background. These results enlighten the bases of trade-offs between the evolution of virulence and multidrug resistance and caution that some alternatives to antibiotics by selecting for capsule inactivation may facilitate the acquisition of antibiotic resistance genes (ARGs).

figshare.com/projects/Supplementary_Datasets/114459 The 3980 genomes assemblies analyzed in this study are publicly available from the RefSeq database (accession numbers in S1 Dataset) All the genomes of strains used for experimental evolution and conjugation assays are publicly available from ENA (accession numbers in S1 Table).

Funding: This work was supported by an ANR JCJC (Agence national de recherche) grant [ANR 18 CE12 0001 01 ENCAPSULATION] awarded to O. R. The laboratory is funded by a Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' grant [ANR-10-LABX-62-IBEID], the INCEPTION program[PIA/ANR-16-CONV-0005], and the FRM [EQU201903007835]. M.H. has received funding from the FIRE Doctoral School (Centre de Recherche Interdisciplinaire, programme Bettencourt) to attend conferences. The funders had no role in the study design, data collection and interpretation, or the decision to submit the work for publication.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: ARG, antibiotic resistance gene; BBH, bidirectional best hits; CFU, colony-forming unit; CLT, capsular locus type; DAP, diaminopimelic acid; GT, glycosyl transferases; GTA, gene transfer agent; HGT, horizontal gene transfer; ICE, integrative conjugative element; Kpn, *Klebsiella pneumoniae*; LB, Luria–Bertani; MGE, mobile genetic element; MPF, mating pair formation; ST, sequence type; WGA, whole-genome alignment; wGRR, gene repertoire relatedness weighted by sequence identity; WT, wild type.

Introduction

Mobile genetic elements (MGE) drive horizontal gene transfer (HGT) between bacteria, which may result in the acquisition of virulence factors and antibiotic resistance genes (ARGs) [1,2]. DNA can be exchanged between cells via virions or conjugative systems [3,4]. Virions attach to specific cell receptors to inject their DNA into the cell, which restricts their host range [5]. When replicating, bacteriophages (henceforth phages) may package bacterial DNA and transfer it across cells (transduction). Additionally, temperate phages may integrate into the bacterial genome as prophages, eventually changing the host phenotype [4]. In contrast, DNA transfer by conjugation involves mating pair formation (MPF) between a donor and a recipient cell [6]. Even if phages and conjugative elements use very different mechanisms of DNA transport, both depend crucially on interactions with the cell envelope of the recipient bacterium. Hence, changes in the bacterial cell envelope may affect their rates of transfer.

Klebsiella pneumoniae (Kpn) is a gut commensal that has become a major threat to public health [7,8] because it is acquiring MGEs encoding ARGs and virulence factors at a fast pace [2,9]. This propensity is much higher in epidemic nosocomial multidrug-resistant lineages than in hypervirulent strains producing infections in the community [10]. Kpn is a particularly interesting model system to study the interplay between HGT and the cell envelope because it is covered by a nearly ubiquitous Group I (or Wzx/Wzy dependent) polysaccharide capsular structure [11,12], which is the first point of contact with incoming MGEs. Similar capsule loci are present across the bacterial phylogeny [13]. There is one single capsule locus in Kpn [14], located between *galF* and *ugd*, which has increased rate of recombination and HGT compared to the rest of the genome [11,15,16]. This locus contains conserved genes encoding the proteins necessary for the assembly and export of the capsule, which is a multistep biosynthesis pathway. These conserved genes flank a highly variable region encoding enzymes that determine the oligosaccharide combination, linkage, and modification (and thus the serotype) [17]. The biochemical determination of the serotype has not been done for the bacteria corresponding to the most recently sequenced genomes. But the genetic content of the capsule locus has been shown to be a very good predictor of the capsule serotype. Such predictions are called capsular locus types (CLTs) to distinguish them from experimentally determined serotypes [17]. Here, we will use CLT to refer to the genomic predictions and serotype to mention the capsule type. There are more than 140 genetically distinct CLTs, of which 76 have well-characterized chemical structures and are referred to as serotypes [17]. Kpn capsules can extend well beyond the outer membrane, up to 420 nm, which is 140 times the average size of the peptidoglycan layer [18]. They enhance cellular survival to bacteriocins, immune response, and antibiotics [19–21], being a major virulence factor of the species. Intriguingly, the multidrug-resistant lineages of Kpn exhibit higher capsular diversity than the virulent ones, which are almost exclusively of the serotype K1 and K2 [10].

By its size, the capsule can hide phage receptors and block phage infection [22]. Since most Kpn are capsulated, many of its virulent phages evolved to overcome the capsule barrier by encoding serotype-specific depolymerases in their tail proteins [23,24]. For the same reason, phages have evolved to use the capsule for initial adherence before attaching to the primary cell receptor. Hence, instead of being hampered by the capsule, many Kpn phages have become dependent on it [25,26]. This means that the capsule may affect the rates of HGT positively or negatively depending on how it enables or blocks phage infection. Furthermore, intense phage predation may select for capsule swap or inactivation, because this renders bacteria resistant to serotype-specific phages. Serotype swaps may allow cells to escape phages to which they were previously sensitive, but they may also expose them to new infectious phages. In contrast, capsule inactivation can confer pan-resistance to capsule-dependent phages [25].

Regarding the effect of capsules on conjugation, very little is known, except that it is less efficient between a few different serotypes of *Haemophilus influenzae* [27]. The interplay between MGEs (phages and conjugative elements) and the capsule has the potential to strongly impact Kpn evolution in terms of both virulence and antibiotic resistance because of the latter's association with specific serotypes and MGEs.

HGT in Kpn is thought to take place by conjugation or in virions, since it is not part of the known naturally transformable bacteria [28]. Hence, the capsule needs MGEs to vary by HGT, but may block the acquisition of the very same MGEs. Moreover, capsulated species are associated with higher rates of HGT [29]. There is thus the need to understand the capsule's precise impact on gene flow and how the latter affects capsule evolution. Here, we leverage a very large number of genomes of Kpn to investigate these questions using computational analyses that are complemented with experimental data. As a result, we propose a model of capsule evolution involving loss and regain of function. This model explains how the interplay of the capsules with different MGEs can either lower, increase, or rewire gene flow depending on the way capsules affect their mechanisms of transfer.

Results

Gene flow is higher within than between serotype groups

We reasoned that if MGEs are specifically adapted to serotypes, then genetic exchanges should be more frequent between bacteria of similar serotypes. We used Kaptive [17] to predict the CLT in 3,980 genomes of Kpn. Around 92% of the isolates could be classed with good confidence level. They include 108 of the 140 previously described CLTs of *Klebsiella* spp. The pangenome of the species includes 82,730 gene families, which is 16 times the average genome. It contains 1,431 single copy gene families present in more than 99% of the genomes that were used to infer a robust rooted phylogenetic tree of the species (average ultra-fast bootstrap of 98%, Fig 1A). Rarefaction curves suggest that we have extensively sampled the genetic diversity of Kpn genomes, its CLTs, plasmids, and prophages (Fig 1B). We then inferred the gains and losses of each gene family of the pangenome using PastML and focused on gene gains in the terminal branches of the species tree predicted to have maintained the same CLT from the node to the tip (91% of branches). This means that we can associate each of these terminal branches with one single serotype. We found significantly more genes acquired (co-gained) in parallel by different isolates having the same CLT than expected by simulations assuming random distribution in the phylogeny (1.95×, Z-test $p < 0.0001$, Fig 1C). This suggests that Kpn exhibits more frequent within-serotype than between-serotype genetic exchanges.

Given the tropism of Kpn phages to specific serotypes, we wished to clarify if phages contribute to the excess of intra-CLTs genetic exchanges. Since transduction events cannot be identified unambiguously from the genome sequences, we searched for prophage acquisition events, i.e., for the transfer of temperate phages from one bacterial genome to another. We found that 97% of the strains were lysogens, with 86% being poly-lysogens, in line with our previous results in a much smaller dataset [25]. In total, 9,886 prophages were identified in the genomes. Their genes account for 16,319 families (19.5%) of the species pangenome (Fig 1B). We then measured the gene repertoire relatedness weighted by sequence identity (wGRR) between all pairs of prophages. The wGRR is a measure of genetic similarity that amounts to 0 if there are no homologs between two genomes, and one if all genes of the smaller genome have a homolog with 100% identity in the other genome. This matrix was clustered, resulting in 2,995 prophage families whose history of vertical and horizontal transmissions was inferred using the species phylogenetic tree (see "Prophage detection"). We found 3,269 independent infection events and kept one prophage for each of them. We found that pairs of independently

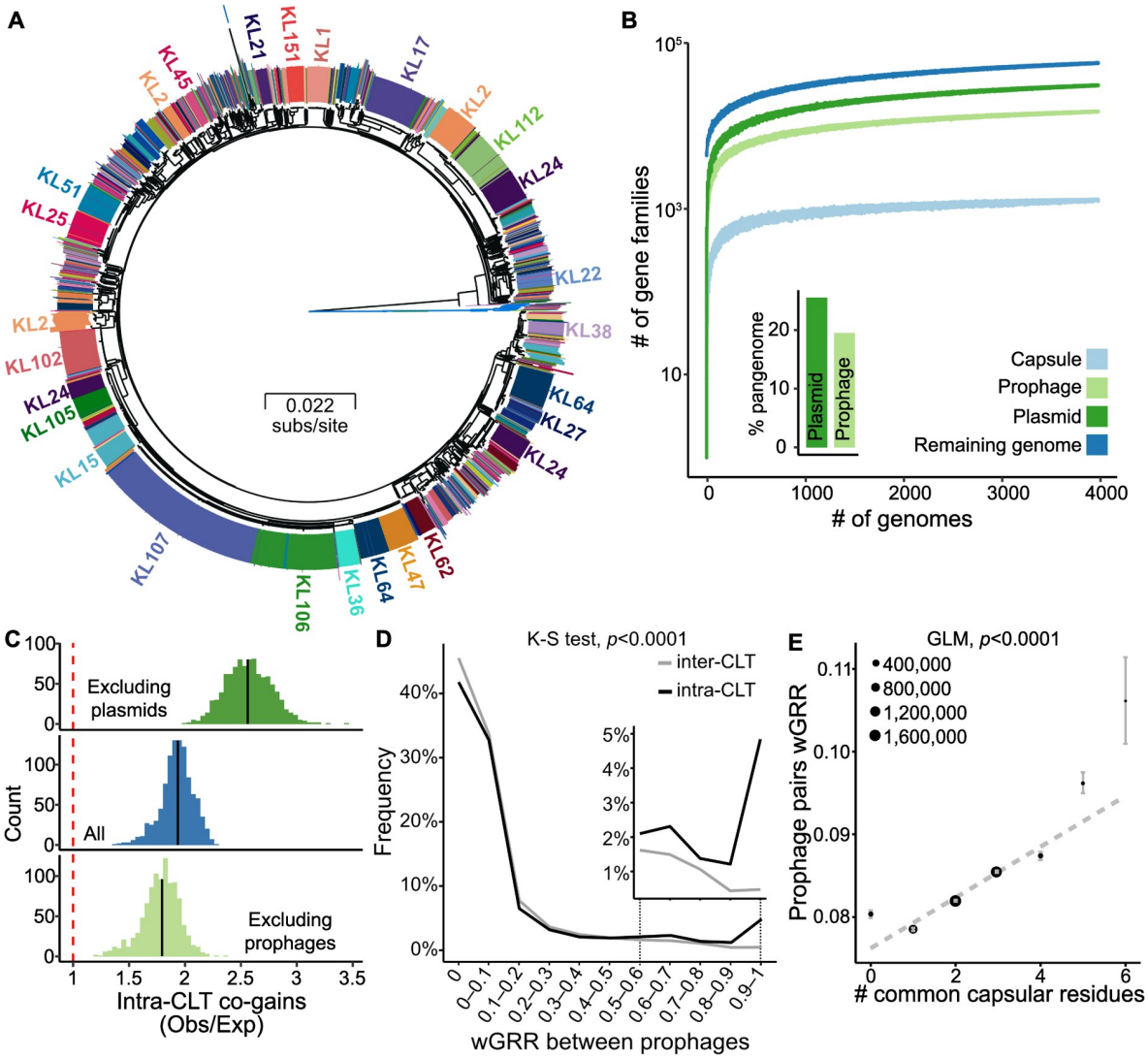


Fig 1. Gene flow is higher between strains of the same serotype. (A) Core genome phylogenetic tree with the 22 *Klebsiella quasipneumoniae* subsp. *similipneumoniae* (misannotated as Kpn in RefSeq) strains as an outgroup (blue branches). The annotation circle represents the 108 CLTs predicted by Kaptive. The largest clusters of CLTs (>20 isolates) are annotated (full list in https://doi.org/10.6084/m9.figshare.14673156). **(B)** Rarefaction curves of the pangenome of prophages, plasmids, capsule genes, and all remaining genes (Genome). The points represent 50 random samples for each bin (bins increasing by 10 genomes). The inset bar plot represents the percentage of gene families of the Kpn pangenome including genes of plasmids or prophages (https://doi.org/10.6084/m9.figshare.14673141). **(C)** Histograms of the excess of intra-CLT co-gains in relation to those observed inter-CLT (Observed/Expected ratio obtained by 1,000 simulations). The analysis includes all genes (center), excludes prophages (bottom), or excludes plasmids (top) (https://doi.org/10.6084/m9.figshare.14673147). **(D)** Gene repertoire relatedness between independently acquired prophages in bacteria of different (inter-CLT, gray) or identical CLT (intra-CLT, black). The insert is a zoom of the distribution for the highest values of wGRR (https://doi.org/10.6084/m9.figshare.14673144). **(E)** Linear regression of the wGRR between pairs of prophages and the number of capsular residues in common between their hosts. The points represent the mean for each category, with their size corresponding to the number of pairs per category. Error bars represent the standard error of the mean. The regression was performed on the original raw data, but only the averages are represented for clarity (https://doi.org/10.6084/m9.figshare.14673165). CLT, capsular locus type; GLM, generalized linear model; Kpn, *Klebsiella pneumoniae*; wGRR, gene repertoire relatedness weighted by sequence identity.

https://doi.org/10.1371/journal.pbio.3001276.g001

infecting prophages are 1.7 times more similar when in bacteria with identical rather than different CLTs (Fig 1D; two-sample Kolmogorov–Smirnov test, $p < 0.0001$). To confirm that phage-mediated HGT is favored between strains of the same CLT, we repeated the analysis of gene co-gains after removing the prophages from the pangenome. As expected, the preference

toward same-CLT exchanges decreased from 1.95× to 1.73× (Fig 1C). This suggests that HGT tends to occur more frequently between strains of identical serotypes than between strains of different serotypes, a trend that is amplified by the transfer of temperate phages.

Most of the depolymerases that allow phages to overcome the capsule barrier act on specific disaccharides or trisaccharides, independently of the remaining monomers [30–32]. This raises the possibility that phage-mediated gene flow could be higher between strains whose capsules have common oligosaccharide residues. To test this hypothesis, we compiled the information on the 76 capsular chemical structures previously described [33]. The genomes with these CLTs, 59% of the total, show a weak but significant proportionality between prophage similarity, and the number of similar residues in their host capsules (Fig 1E), i.e., prophages, are more similar between bacteria with more biochemically similar capsules.

Recombination swaps biochemically related capsules

To understand the genetic differences between serotypes and how these could facilitate swaps, we compared the gene repertoires of the different capsular loci (between *galF* and *ugd*, S1 Fig). As expected from previous studies [11,17], this analysis revealed a clear discontinuity between intra-CLT comparisons that had mostly homologous genes and the other comparisons, where many genes (average = 10) lacked homologs across serotypes (Wilcoxon test, $p < 0.0001$, Fig 2A). As a result, the capsule pangenome contains 325 gene families that are specific to a CLT (out of 547, see "Pan- and persistent genomes"). This implies that serotype swaps require the acquisition of multiple novel genes by horizontal transfer. To quantify and identify these CLT swaps, we inferred the ancestral CLT in the phylogenetic tree and found a rate of 0.282 swaps per branch (see "Serotype swaps identification"). We then identified 103 highly confident swaps, some of which occurred more than once (Fig 2B). We used the chemical characterization of the capsules described above to test if swaps were more likely between serotypes with more similar chemical composition. Indeed, swaps occurred between capsules with an average of 2.42 common sugars (mean Jaccard similarity 0.54), more than the average value across all other possible CLT pairs (1.98, mean Jaccard similarity 0.38, Wilcoxon test, $p < 0.0001$, S2A Fig). Interestingly, the wGRR of the swapped loci is only 3% higher than the rest of pairwise comparisons (S2B Fig). This suggests that successful swaps are poorly determined by the differences in gene repertoires. Instead, they are more frequent between capsules that have more similar chemical composition.

The existence of a single capsule locus in Kpn genomes suggests that swaps occur by homologous recombination at flanking conserved sequences [11]. We used Gubbins [34] to detect recombination events in the 25 strains with terminal branch serotype swaps and closely related completely assembled genomes (see "Detection of recombination tracts"). We found long recombination tracts encompassing the capsule locus in 24 of these 25 genomes, with a median length of 100.3 kb (Fig 2C). At least one border of the recombination tract was less than 3-kb away from the capsule locus in 11 cases (46%). Using sequence similarity to identify the origin of the transfer, we found that most recombination events occurred between distant strains and no specific clade (Fig 2D). We conclude that serotype swaps occur by recombination at the flanking genes with DNA from genetically distant isolates but chemically related capsules.

Capsule inactivation follows specific paths, might be driven by phage predation, and spurs HGT

We sought to investigate whether the aforementioned swaps occur by an intermediary step where cells have inactivated capsule loci. To do so, we first established the frequency of inactivated capsular loci. We used the Kaptive software to detect missing genes, expected to be

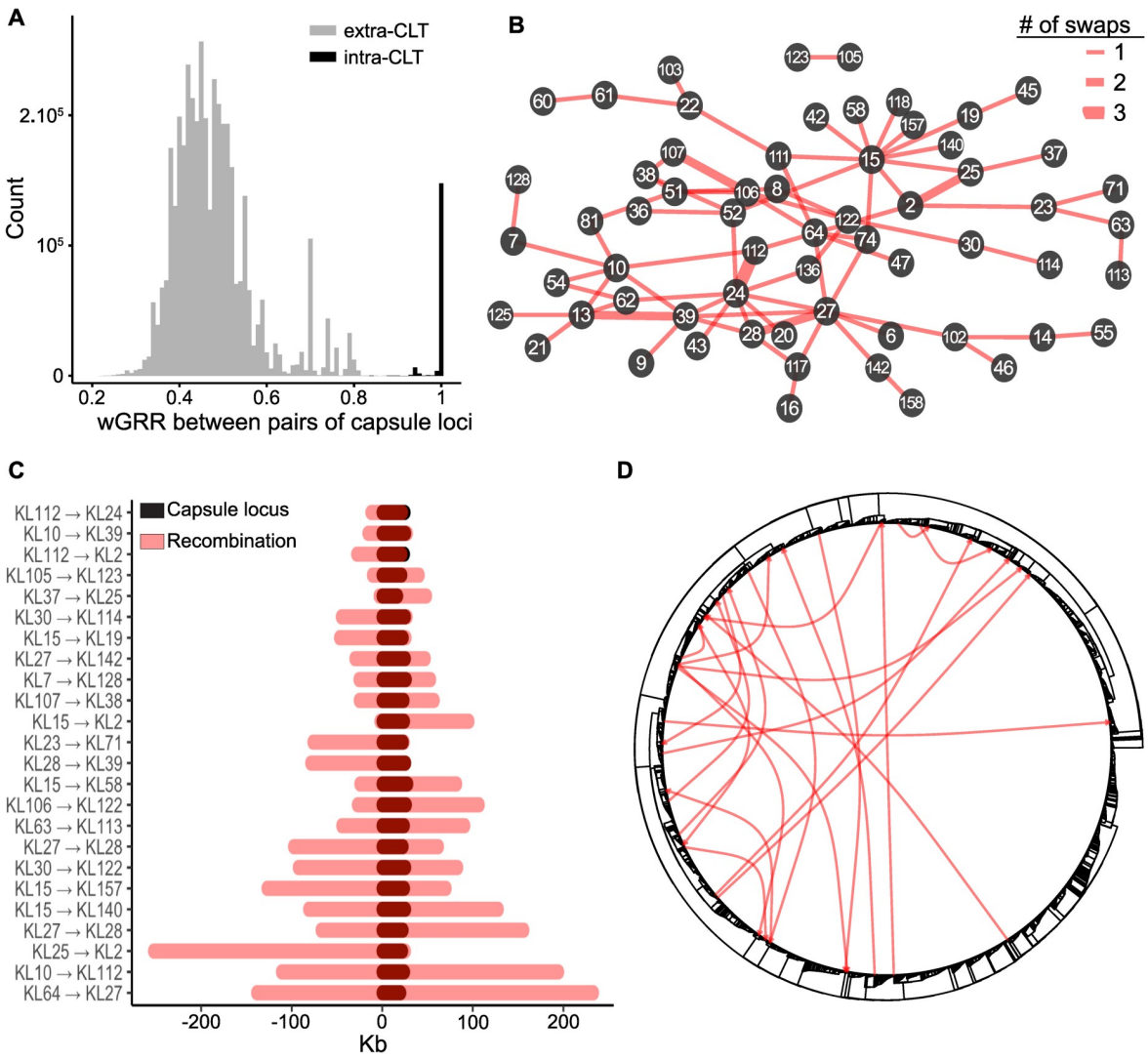


Fig 2. Homologous recombination events lead to frequent CLT swaps. (A) Histogram of the comparisons of gene repertoire relatedness (wGRR) between capsular loci of the same (intra-) or different (inter-) CLT (https://doi.org/10.6084/m9.figshare.14673171). **(B)** Network of CLT swaps identified by ancestral state reconstruction, with edge thickness corresponding to the number of swaps, and numbers **x** within nodes corresponding to the CLT (KL**x**). **(C)** Recombination encompassing the capsule locus detected with Gubbins. The positions of the tracts are represented in the same scale, where the first base of the *galF* gene was set as 0 (https://doi.org/10.6084/m9.figshare.14673150). **(D)** Putative donor–recipient pairs involved in the CLT swaps of panel C indicated in the Kpn tree. CLT, capsular locus type; Kpn, *Klebsiella pneumoniae*; wGRR, gene repertoire relatedness weighted by sequence identity.

https://doi.org/10.1371/journal.pbio.3001276.g002

encoded in capsule loci found on a single contig. We also used the Kaptive database of capsular proteins to detect pseudogenes using protein–DNA alignments in all genomes. We found 55 missing genes and 447 pseudogenes, among 9% of the loci. The frequency of pseudogenes was not correlated with the quality of the genome assembly (see "Identification of capsule pseudogenes and inactive capsule loci"), and all genomes had at least a part of the capsule locus. We cannot exclude that some of these mutations fixated during passage prior to sequencing if these conditions strongly select for capsule loss. However, many isolates harbor several pseudogenes, which suggests that capsule inactivation is not due to very strong selection of one inactivating mutation during passage. We classed 11 protein families as essential for capsule production (Table A in S1 Text). At least one of these essential genes was missing in 3.5% of

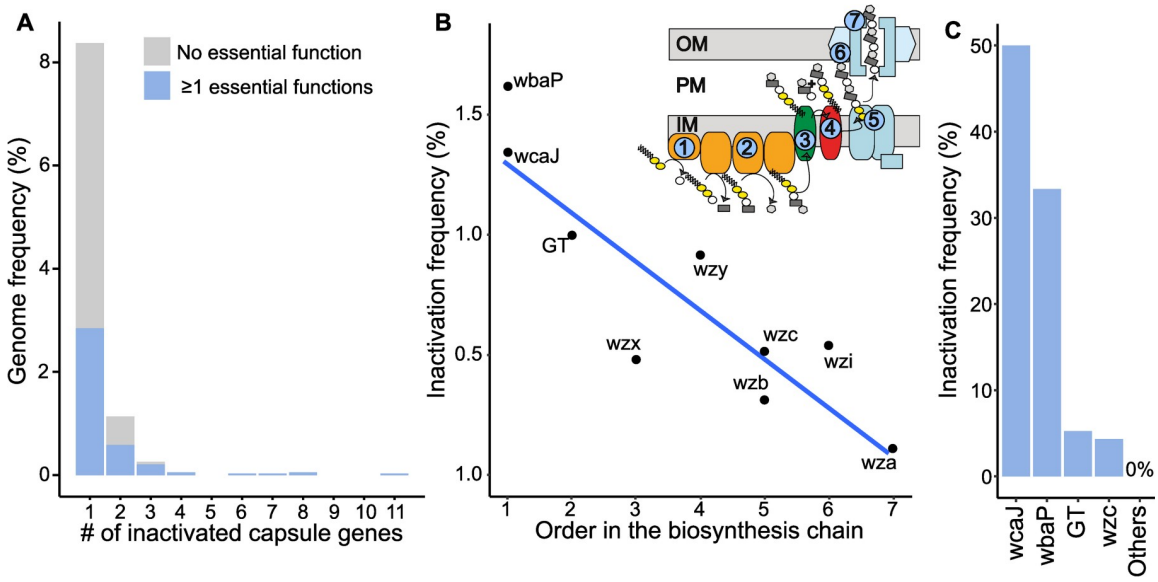


Fig 3. Loss of function in the capsule locus. (A) Distribution of the number of inactivated capsule genes per genome, split in 2 categories: loci lacking a functional essential capsule gene (blue, non-capsulated strains), and other loci only lacking nonessential capsule genes (white, not categorized as non-capsulated strains) (https://doi.org/10.6084/m9.figshare.14673153). **(B)** Linear regression between the inactivation frequency and the rank of each gene in the biosynthesis pathway ($p = 0.001$, $R^2 = 0.78$). The numbers in the scheme of the capsule assembly correspond to the order in the biosynthesis pathway (https://doi.org/10.6084/m9.figshare.14673153). **(C)** Frequency of inactivated capsule genes arising in the non-capsulated clones isolated in 8 different strains after approximately 20 generations in LB growth medium. Genomes containing several missing genes and pseudogenes are not included (https://doi.org/10.6084/m9.figshare.14673177). GT, glycosyl transferases; LB, Luria–Bertani.

https://doi.org/10.1371/journal.pbio.3001276.g003

the loci, which means these strains are likely non-capsulated (Fig 3A). These variants are scattered in the phylogenetic tree with no particular clade accounting for the majority of these variants (S3 Fig), e.g., there are non-capsulated strains in 61 of the 617 sequence types (STs) identified by Kleborate. These results suggest that capsule inactivation has little phylogenetic inertia, i.e., it is a trait that changes very quickly, either because the variants are counterselected or because capsules are quickly reacquired. Accordingly, we could not detect significant phylogenetic inertia with Pagel Lambda test [35] ($p > 0.05$). Hence, capsule inactivation is frequent, but non-capsulated lineages do not persist for long periods of time.

We further investigated the genetic pathways leading to capsule inactivation. Interestingly, we found that the pseudogenization frequency follows the order of biosynthesis of the capsule (linear regression, $p = 0.001$, $R^2 = 0.78$), with the first (*wbaP* or *wcaJ*) and second steps (glycosyl transferases, GT) being the most commonly inactivated when a single essential gene is a pseudogene (Fig 3B). The overall frequency of gene inactivation drops by 14% per rank in the biosynthesis chain. To confirm these results, we made several controls. To show that this is not simply an effect of gene length, we normalized the inactivation frequency by gene length and found the same relationship (S4A Fig, $p = 0.005$, $R^2 = 0.7$). To check that this was not a mutation hotspot induced by simple sequence repeats, which are prone to polymerase-slippage–induced mutations and sequencing errors, we searched for mono- and di-nucleotide tracts. We found that their frequency in the commonly inactivated pair *wcaJ*/*wbaP* was smaller than in the rarely mutated group of genes *wzc*/*wzx*/*wzy* (S4B Fig). Finally, we verified that our analysis was not impacted by a higher allelic diversity in these genes in the reference database. We found that the most genetically diverse genes were not the most inactivated (S4C Fig). Together, these results support the idea that selection plays a key role in the fixation of inactivating mutations in the early genes of the capsule biosynthetic pathway.

To test if similar results are found when capsules are counterselected in the laboratory, we analyzed a subset of populations stemming from a short evolution experiment in which populations of 8 different strains of *Klebsiella* spp. were diluted daily during 3 days (approximately 20 generations) under agitation in Luria–Bertani (LB), a medium known to select for capsule inactivation [36]. Each strain belongs to a different phylogenetic group (ST), encompassing 6 different serotypes and isolation sources (Table B in S1 Text). We have previously shown that under such conditions, phage pressure accelerates capsule loss [25]. After 3 days, non-capsulated clones emerged in 22 out of 24 populations from 8 different ancestral strains. We isolated one non-capsulated clone for Illumina sequencing from each population and searched for the inactivating mutations with the same pipeline as for the genomic dataset. We also compared our method with two popular tools used to detect new mutations in evolved isolates, *breseq* [37] and *snippy*, which rely on read mapping onto a reference genome. All 3 approaches yielded comparable conclusions, although some mutations were not found by all 3. Overall, 13 out of the 16 inactivated or deleted genes found by our method were also identified by either *breseq*, *snippy*, or both. Read mapping approaches detected other types of mutations, like intergenic mutations, and few mutations in the rest of the genome, which are not detectable by our targeted approach (https://doi.org/10.6084/m9.figshare.14673177). We found that most of these were localized in *wcaJ* and *wbaP* (Fig 3C). In accordance with our comparative genomics analysis, we found fewer loss-of-function mutations in GTs and *wzc* and none in the latter steps of the biosynthetic pathway, except for one large deletion event encompassing almost all the capsular locus (https://doi.org/10.6084/m9.figshare.14673177). Studies focusing on the mutations conferring phage resistance in Kpn have also reported an abundance of loss-of-function mutations in capsule genes leading to a non-capsulated phenotype [25,38], especially *wcaJ* [39,40]. These results strongly suggest that mutations leading to the loss of capsule production impose a fitness cost determined by the position of the inactivated gene in the biosynthesis pathway.

Once a capsule locus is inactivated, the function can be reacquired by (1) reversion mutations fixing the broken allele; (2) restoration of the inactivated function by acquisition of a gene from another bacterium, eventually leading to a chimeric locus; and (3) replacement of the entire locus leading to a CLT swap. Our analyses of pseudogenes provide some clues on the relevance of the 3 scenarios. We found 111 events involving nonsense point mutations. These could eventually be reversible (scenario 1) if the reversible mutation arises before other inactivating changes accumulate. We also observed 269 deletions (100 of more than 2 nucleotides) in the inactivated loci. These changes are usually irreversible in the absence of HGT. We then searched for chimeric loci (scenario 2), i.e., CLTs containing at least 1 gene from another CLT. We found 35 such loci, accounting for approximately 0.9% of the dataset (for example, a *wzc_KL1* allele in an otherwise KL2 loci), with only one occurrence of a *wcaJ* allele belonging to another CLT, and none for *wbaP*. Finally, the analysis of recombination tracts detailed above reveals frequent replacement of the entire locus between *galF* and *ugd* by recombination (S1 Fig, scenario 3).

Since reacquisition of the capsule function might often require HGT, we enquired if capsule inactivation was associated with higher rates of HGT. Indeed, the number of genes gained by HGT per branch of the phylogenetic tree is higher in branches where the capsule was inactivated than in the others (2-sample Wilcoxon test, $p < 0.0001$, Fig 4A), even if these branches have similar sizes (S6 Fig). We compared the number of phages and conjugative systems acquired in the branches where capsules were inactivated against the other branches. This revealed significantly more frequent (3.6 times more) acquisition of conjugative systems (Fisher exact test, $p < 0.0001$, Fig 4B) upon capsule inactivation. This was also the case, to a lesser extent, for prophages. Intriguingly, we observed even higher relative rates of acquisition

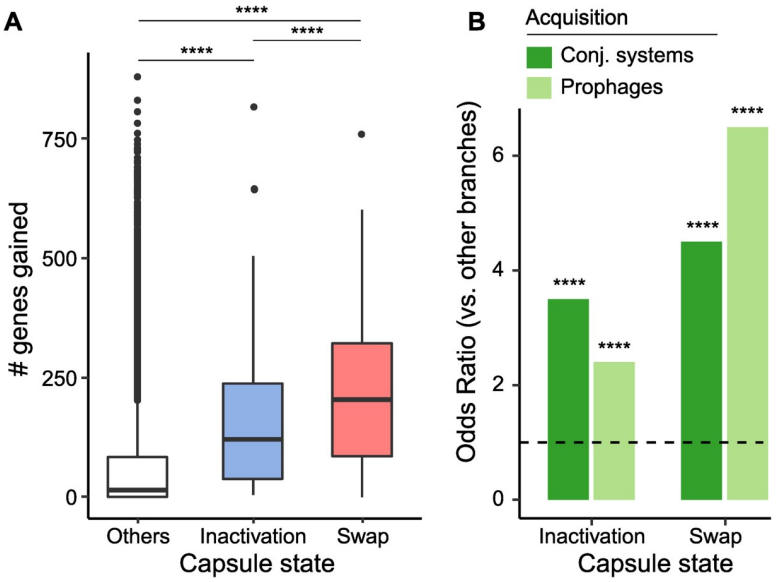


Fig 4. Changes in capsule state impact HGT. (A) Number of genes gained in terminal branches of the phylogenetic tree where the capsule was inactivated, swapped, and in the others (2-samples Wilcoxon tests). **(B)** Increase in the frequency of acquisition of prophages and conjugative systems on branches where the capsule was either inactivated or swapped, relative to the other branches, as represented by odds ratio (Fisher exact test). ****: p-value < 0.0001 (https://doi.org/10.6084/m9.figshare.14673159). HGT, horizontal gene transfer.

https://doi.org/10.1371/journal.pbio.3001276.g004

of genes and MGEs in branches where the serotype was swapped (Fig 4A and 4B). The latter branches are 2.7 times longer than the others (S6 Fig), which may be due to the recombination tracts associated with the swap. The difference in branch lengths between the swapped and inactivated categories are on the order of magnitude of the difference in the number of genes gained, suggesting approximately similar rates of gene gain in both categories. In contrast, the difference in terms of gained genes between branches with capsule swaps and branches where capsules remained unchanged is much larger than 2.7. Hence, branches where capsules were swapped, like those where they were inactivated, represent periods of more frequent acquisition of conjugative elements and phages relative to the periods where the capsule remained unchanged. The acquisition rate is particularly larger for conjugative systems in branches where the capsule was inactivated and larger for prophages in branches that swapped (6.5 versus 4.5 times more). Overall, the excess of HGT in periods of capsule inactivation facilitates the reacquisition of a capsule and the novel acquisition of other potentially adaptive traits.

Conjugative systems are frequently transferred across serotypes

The large size of the Kpn capsule locus is difficult to accommodate in the phage genome, and the tendency of phages to be serotype specific makes them unlikely vectors of novel capsular loci. Also, the recombination tracts observed in Fig 2C are too large to be transduced by most temperate phages of Kpn, whose prophages average 46 kb [25]. Since Kpn is not naturally transformable, we hypothesized that conjugation is the major driver of capsule acquisition. Around 80% of the strains encode a conjugative system, and 94.4% have at least one. Plasmids alone make 25.5% of the pangenome (Fig 1B). To these numbers, one should add the genes present in integrative conjugative elements (ICEs). Unfortunately, there is currently no method to identify ICEs accurately in draft genomes. By subtracting the total number of conjugative elements from those identified in plasmids, we estimate that 41% of the conjugative systems in Kpn are not in plasmids but in ICEs. Since ICEs and conjugative plasmids have

approximately similar sizes [41], the joint contribution of ICEs and plasmids in the species pangenome is very large.

We identified independent events of infection by conjugative systems as we did for prophages (see above). The 5,144 conjugative systems fell into 252 families with 1,547 infection events. On average, pairs of conjugative systems acquired within the same CLT were only 3% more similar than those in different CLTs. This suggests that phage- and conjugation-driven HGT have very different patterns, since the former tend to be serotype specific, whereas the latter are very frequently transferred across serotypes. This opposition is consistent with the analysis of co-gains (Fig 1C), which were much more serotype dependent when plasmids were excluded from the analysis and less serotype dependent when prophages were excluded. To further test our hypothesis, we calculated the number of CLTs where one could find each family of conjugative systems or prophages and then compared these numbers with the expectation if they were distributed randomly across the species. The results show that phage families are present in much fewer CLTs than expected, whereas there is no bias for conjugative systems (Fig 5). We conclude that conjugation spreads plasmids across the species regardless of

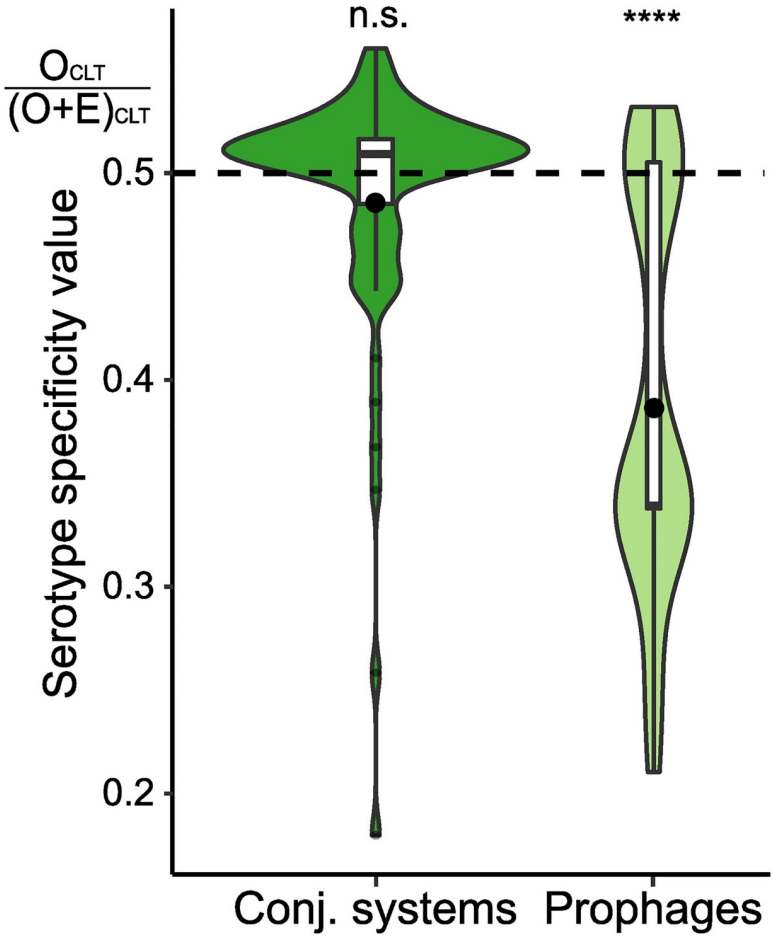


Fig 5. Serotype specificity of prophages and conjugative systems. O_{CLT}: observed number of serotypes infected per family of homologous element. E_{CLT}: expected number of serotypes infected per family of homologous element generated by 1,000 simulations (see "CLT specificity"). This measure aims to capture the number of different CLT that each prophage and conjugative system were able to infect, compared to what was expected by chance given the phylogeny. When the elements distribute randomly across CLTs, the value is 0.5 (dashed line). Very low values indicate high serotype specificity. One-sample Wilcoxon test. ****: p-value < 0.0001 (https://doi.org/10.6084/m9.figshare.14673162). CLT, capsular locus type.

https://doi.org/10.1371/journal.pbio.3001276.g005

serotype. Together, these results reinforce the hypothesis that conjugation drives genetic exchanges between strains of different serotypes, decreasing the overall bias toward same-serotype exchanges driven by phages.

Capsule inactivation results in increased conjugation frequency

Together, these elements led us to hypothesize that capsule inactivation results in higher rates of acquisition of conjugative elements. This is consistent with the observation that terminal branches associated with inactive capsules have a higher influx of conjugative systems than prophages (Fig 4B). In the absence of published data on the impact of the capsule on conjugation frequency, we tested experimentally our hypothesis on a diverse set of *Klebsiella* isolates composed of four strains from different STs: three *Klebsiella pneumoniae sensu stricto* (serotypes K1 and K2) and one *Klebsiella variicola* (serotype K30, also found in Kpn). To test the role of the capsule in plasmid acquisition, we analyzed the conjugation frequency of these strains and their non-capsulated counterparts, deprived of *wcaJ*, the most frequent pseudogene in the locus both in the genome data and in our experimental evolution (see "Conjugation assay"). For this, we built a plasmid that is mobilized in *trans*, i.e., once acquired by the new host strain, it cannot further conjugate, due to their lack of a compatible conjugative system. This allows to measure precisely the frequency of conjugation between the donor and the recipient strain. In agreement with the results of the computational analysis, we found that the frequency of conjugation is systematically and significantly higher in the mutant than in the associated wild type (WT) for all four strains (paired Wilcoxon test, p-value = 0.002, Fig 6). On average, non-capsulated strains conjugated 8.06 times more than capsulated strains. In strain CG43, this difference was 20-fold.

Interestingly, the difference in conjugation rates between the mutants and their WT is inversely proportional to the conjugation frequency in the WT, possibly because some WT strains already conjugate at very high rates. Cell densities were normalized to the same optical density (OD_{600} = 0.9) before mating, and the donor culture was the same for all recipient strains per biological replicate. Cells were allowed to conjugate for one hour. Still, there were slightly fewer colony-forming unit (CFU; 3.2× less, paired Wilcoxon test, $p < 0.001$) on the membranes with non-capsulated mutants than on those with the WT. This may lead to a slight underestimation of the conjugation frequency in non-capsulated mutants. As a consequence, we may have underestimated the differences in conjugation frequency. Overall, these experiments show that the ability to receive a conjugative element is increased in the absence of a functional capsule. Hence, non-capsulated variants acquire more genes by conjugation than the others. Interestingly, if the capsule is transferred by conjugation, this implies that capsule inactivation favors the very mechanism leading to its subsequent reacquisition.

Discussion

The specificity of many Kpn phages to one or a few chemically-related serotypes is presumably caused by their reliance on capsules to adsorb to the cell surface and results from the long-standing coevolution of phages with their Kpn hosts. One might invoke environmental effects to explain these results, since populations with identical or closely related serotypes might often co-occur and thus potentiate more frequent cross infections. However, the same ecological bias would be expected to increase the rates of conjugation between identical or closely related serotypes. This could not be detected, suggesting that bias in intra-serotype gene flow is mediated by phages. It is well known that some phages carry capsule depolymerases acting on disaccharides or trisaccharides [30–32] that may be similar across serotypes and thus favor transfer of phages between these cells. This fits our previous studies on the infection networks

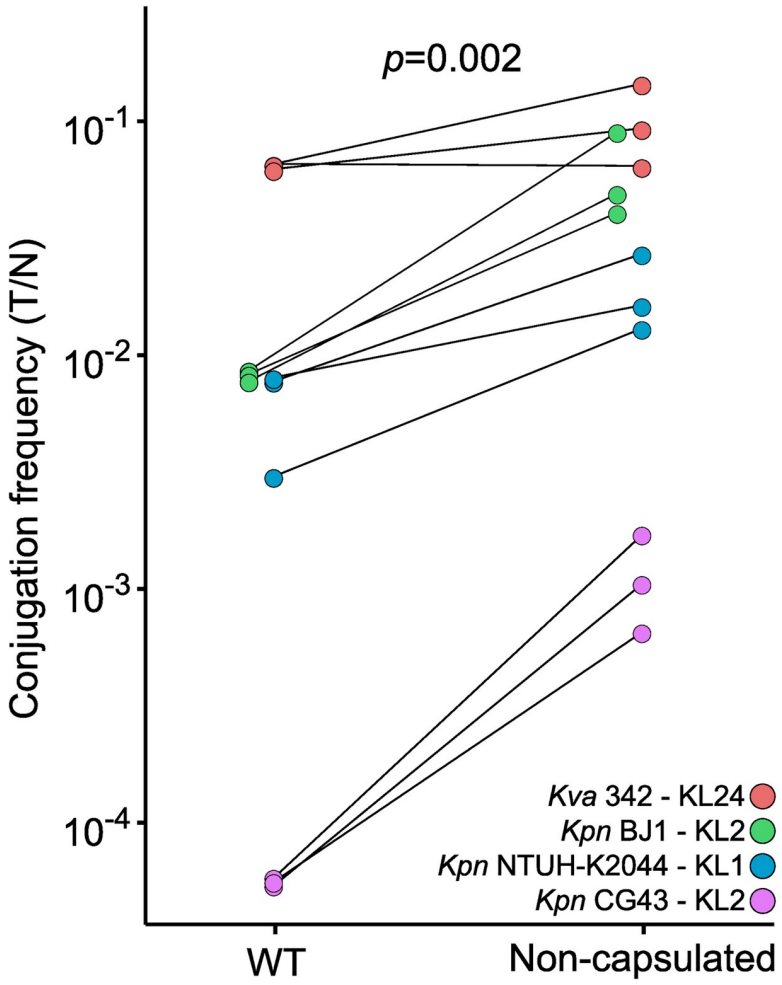


Fig 6. Capsules negatively impact conjugation. Conjugation frequency of WT and their associated non-capsulated (*ΔwcaJ*) mutants. The conjugation efficiency (T/N, see "Conjugation assay") is represented on a log scale. Each pair of points represents a biological replicate. Strains and their capsule locus type are listed in the order in which they appear from the top to the bottom of the plot. The *p*-value for the paired Wilcoxon test is displayed (https://doi.org/10.6084/m9.figshare.14673168). Kpn, *Klebsiella pneumoniae*; Kva, Klebsiella variicola; CG, clonal group; WT, wild type.

https://doi.org/10.1371/journal.pbio.3001276.g006

of Kpn prophages [25] and suggests that a population of cells encoding and expressing a given serotype has more frequent phage-mediated genetic exchanges with bacteria of identical or chemically similar serotypes (Fig 7A). In this context, phages carrying multiple capsule depolymerases have broader host range and may have a key role in phage-mediated gene flow between very distinct serotypes. For example, one broad host range virulent phage has been found to infect 10 distinct serotypes because it encodes an array of at least 9 depolymerases [24].

The interplay between the capsule and conjugative elements has been much less studied. Our comparative genomics analyses reveal that conjugation occurs across strains of identical or different CLTs at similar rates. Furthermore, our experimental data shows that non-capsulated mutants are up to 20 times more receptive to plasmid conjugation than the capsulated WT bacteria, an effect that seems more important for the WT that are poor recipients. These results are likely to be relevant not only for non-capsulated strains, but also for those not expressing the capsule at a given moment. If so, repression of the expression of the capsule may allow bacteria to escape phages and endure extensive acquisition of conjugative elements.

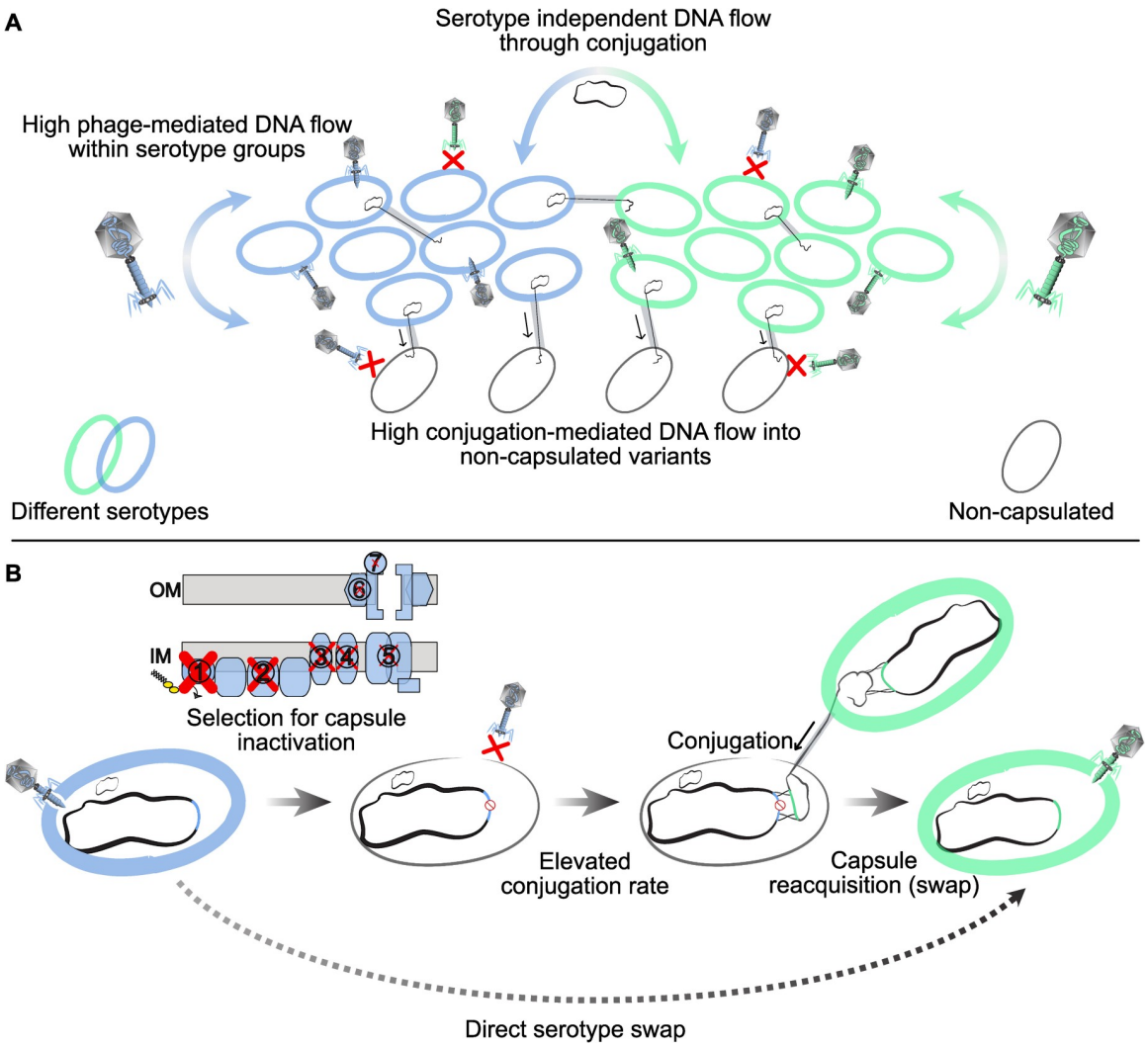


Fig 7. A model for the interplay between serotypes and mobile elements. (A) The capsule impacts Kpn gene flow. A bacterial population expressing a given serotype (blue or green) preferentially exchanges DNA by phage-mediated processes with bacteria of identical or chemically similar serotypes. Such flow may be rare toward non-capsulated bacteria because they are often resistant to Kpn phages. In contrast, conjugation occurs across serotypes and is more frequent to non-capsulated bacteria. **(B)** Proposed model for serotype swaps in Kpn. The capsule locus is colored according to its type. Capsule inactivation is occasionally adaptive. The pseudogenization process usually starts by the inactivation of the genes involved in the early stages of the capsule biosynthesis, as represented by the size of the red cross on the capsule assembly scheme. Non-capsulated strains are often protected from Kpn phage infections while acquiring more genes by conjugation. This increases the likelihood of capsule reacquisition. Such reacquisition can bring a new serotype, often one that is chemically similar to the previous one, and might be driven by conjugation because of its high frequency in non-capsulated strains. Serotype swaps rewire phage-mediated genetic transfers. IM, inner membrane; Kpn, *Klebsiella pneumoniae*; OM, outer membrane.

https://doi.org/10.1371/journal.pbio.3001276.g007

These results may also explain a long-standing conundrum in Kpn. The hypervirulent lineages of Kpn, which are almost exclusively of serotypes K1 and K2 [10,42], have reduced pangenome, plasmid, and capsule diversity. They also often carry additional factors like *rmpA* upregulating the expression of the capsule [42]. In contrast, they are very rarely multidrug resistant. Our data suggest that the protection provided by thick capsules hampers the acquisition of conjugative elements, which are the most frequent vectors of antibiotic resistance. Furthermore, the moments of capsule swap, repression of expression, or inactivation are expected to be particularly deleterious for hypervirulent clones, because the capsule is a virulence factor,

thus further hampering their ability to acquire the conjugative MGEs that carry antibiotic resistance. This may have favored a specialization of the clones into either hypervirulent or multidrug resistance. Unfortunately, the capsule is not an insurmountable barrier for conjugation, and recent reports have uncovered the emergence of multidrug-resistant hypervirulent clones [43,44]. The most worrisome pattern of evolution of such strains is the fusion of the resident virulence plasmids of hypervirulent clones with highly mobile multidrug resistance plasmids [45]. Strikingly, the recently isolated MS3802 clone, which belongs to the hypervirulent lineage ST23 and harbors a chimeric virulence/resistance plasmid, is string test negative and encodes a strongly degraded KL1 capsule locus [46]. Capsule inactivation may have facilitated the acquisition of a conjugative resistance plasmid that co-integrated with the resident virulence plasmid.

We observed that branches of isolates lacking a functional capsule have higher rates of acquisition of conjugative systems than prophages, whereas those where there was a capsule swap have the inverse pattern (Fig 4B). What could justify these differences in the interplay of the capsule with phages and conjugative elements? Phages must adsorb on the cell surface, whereas there are no critical positive determinants for incoming conjugation pilus [47]. As a result, serotype swaps may affect much more the flow of phages than that of conjugative elements. Capsule inactivation may have an opposite effect on phages and plasmids: it removes a point of cell attachment for phages, decreasing their infection rates, and removes a barrier to the conjugative pilus, increasing their ability to transfer DNA. Hence, when a bacterium has an inactivated capsule, e.g., because of phage predation, it becomes more permissive for conjugation. In contrast, when a bacterium acquires a novel serotype, it may become sensitive to novel phages resulting in rapid turnover of its prophage repertoire. These results implicate that conjugation should be much more efficient at spreading traits across the entire Kpn species than phage-mediated mechanisms, whose role would be important for HGT between strains of similar or closely related serotypes (Fig 7A).

The existence of serotype swaps has been extensively described in the literature for Kpn [16] and many other species [48,49]. Whether these swaps imply a direct serotype replacement, or an intermediate non-capsulated state, is not sufficiently known. Several processes are known to select for capsule inactivation in some bacteria, including growth in rich medium [36], phage pressure, and immune response [25,26,39,50] (Fig 7B). Because of the physiological effects of these inactivations, and their impact on the rates and types of HGT, it is important to quantify the frequency of inactivated (or silent) capsular loci and the mechanisms favoring it. Our study of pseudogenization of capsular genes revealed a few percent of putative non-capsulated strains scattered in the species tree, opening the possibility that non-capsulated strains are a frequent intermediate step of serotype swap. The process of capsule inactivation is shaped by the capsule biosynthesis pathway, the frequency of pseudogenization decreasing linearly with the rank of the gene in the capsule biosynthesis pathway. This suggests a major role for epistasis in the evolutionary pathway leading to non-capsulated strains. Notably, the early inactivation of later genes in the biosynthesis pathway, while the initial steps are still functional, can lead to the sequestration of key molecules at the cell envelope or the toxic accumulation of capsule intermediates (Fig 7B). Accordingly, Δ*wza* and Δ*wzy* mutants, but not Δ*wcaJ*, lead to defects in the cell envelope of the strain Kpn SGH10 [51]. In *Acinetobacter baumannii*, high-density transposon mutagenesis also recently revealed that inactivation of genes involved in the last steps of capsule biosynthesis is much more deleterious than those encoding the early steps [52].

Capsule reacquisition is more likely driven by conjugation than by phages, which are generally dependent on the capsule to infect their host. Hence, the increased rate of acquisition of conjugative elements by non-capsulated strains may favor the process of capsule reacquisition,

and, eventually, serotype swap. Cycles of gain and loss of capsular loci have been previously hypothesized in the naturally transformable species *Streptococcus pneumoniae*, because vaccination leads to counterselection of capsulated strains, and natural transformation seems to increase recombination in non-capsulated clades [53]. Accordingly, tracts with a median length 42.7 kb encompassing the capsule locus were found in serotype-swapped *S. pneumoniae* isolates [54]. In Kpn, such tracts were more than twice larger, which is consistent with the role of conjugation in capsule swap. Interestingly, these swaps are more frequent between CLTs encoding serotypes with common sugar residues, independently of their overall genetic relatedness. Understanding this result will require further work, but it suggests that genomic adaptation to the production of specific activated sugars can lead to genetic incompatibilities with other capsular genes.

Our results are relevant to understand the interplay between the capsule and other mobile elements in Kpn or other bacteria. We expect to observe more efficient conjugation when the recipient bacteria lack a capsule in other species. For example, higher conjugation rates of non-capsulated strains may help explain their higher recombination rates in *Streptococcus* [55]. The serotype specificity of phages also opens intriguing possibilities for them and for virion-derived elements. For instance, some *Escherichia coli* strains able to thrive in freshwater reservoirs have capsular loci acquired in a single-block horizontal transfer from Kpn [56]. This could facilitate interspecies phage infections (and phage-mediated HGT), since Kpn phages may now be adsorbed by these strains. Gene transfer agents (GTA) are co-options of virions for intraspecies HGT that are frequent among alpha-proteobacteria [57] (but not yet described in Kpn). They are likely to have equivalent serotype specificity, since they attach to the cell envelope using structures derived from phage tails. Indeed, the infection by the *Rhodobacter capsulatus* GTA model system depends on the host Wzy capsule [58], and non-capsulated variants of this species are phage resistant [59] and impaired in GTA-mediated transfer [60]. Our general prediction is that species where cells tend to be capsulated are going to coevolve with phages, or phage-derived tail structures, such that the latter will tend to become serotype specific. We speculate that future developments on the systematic detection of depolymerase genes will shed light on depolymerase swaps between phages, as a reciprocal phenomenon to capsule serotype swaps. Here, we could not make such an analysis since we identified very few depolymerases. Similar difficulties to find depolymerases were recently observed [25,61], suggesting that many such proteins remain to be identified.

These predictions have an impact in the evolution of virulence and antimicrobial therapy. Some alternatives to antibiotics—phage therapy, depolymerases associated with antibiotics, pyocins, and capsular polysaccharide vaccines—may select for the inactivation of the capsule [25,38,39]. Such non-capsulated variants have often been associated with better disease outcomes [62], lower antibiotic tolerance [21], and reduced virulence [20]. However, they can also be more successful colonizers of the urinary tract [63]. Our results suggest that these non-capsulated variants are at higher risk of acquiring resistance and virulence factors through conjugation, because ARGs and virulence factors are often found in conjugative elements in Kpn and in other nosocomial pathogens. Conjugation may also eventually lead to the reacquisition of functional capsules. At the end of the inactivation–reacquisition process, recapitulated on Fig 7, the strains may be capsulated, more virulent, and more antibiotic resistant.

Materials and methods

Genomes

We used the PanACoTa tool to generate the dataset of genomes [64]. We downloaded all the 5,805 genome assemblies labeled as *Klebsiella pneumoniae sensu stricto* (*Kpn*) from NCBI

RefSeq (accessed on October 10, 2018). We removed lower-quality assemblies by excluding genomes with L90>100. The pairwise genetic distances between all remaining genomes of the species was calculated by order of assembly quality (L90) using MASH [65]. Strains that were too divergent (MASH distance >6%) to the reference strain or too similar (<0.0001) to other strains were removed from further analysis. The latter tend to have identical capsule serotypes, and their exclusion does not eliminate serotype swap events. This resulted in a dataset of 3,980 strains which were re-annotated with *prokka* (v1.13.3) [66] to use consistent annotations in all genomes. Erroneous species annotations in the GenBank files were corrected using Kleborate (https://github.com/katholt/Kleborate). This step identified 22 *Klebsiella quasipneumoniae* subspecies *similipneumoniae* (Kqs) genomes that were used to root the species tree and excluded from further analyses. The accession number for each analyzed genome is presented in https://doi.org/10.6084/m9.figshare.14673156, along with all the annotations identified in this study.

Pan- and persistent genome

The pangenome is the full repertoire of homologous gene families in a species. We inferred the pangenome with the connected-component clustering algorithm of MMSeqs2 (*release 5*) [67] with pairwise bidirectional coverage >0.8 and sequence identity >0.8. The persistent genome was built from the pangenome, with a persistence threshold of 99%, meaning that a gene family must be present in single copy in at least 3,940 genomes to be considered persistent. Among the 82,730 gene families of the Kpn pangenome, there were 1,431 gene families present in 99% of the genomes, including the Kqs. We used mlplasmids to identify the "plasmid" contigs (default parameters, species "*Klebsiella pneumoniae*" [68]). To identify the pangenome of capsular loci present in the Kaptive database, we used the same method as above, but we lowered the sequence identity threshold to >0.4 to put together more remote homologs.

Phylogenetic tree

To compute the species phylogenetic tree, we aligned each of the 1,431 protein families of the persistent genome individually with mafft (v7.407) [69] using the option *FFT-NS-2*, back-translated the sequences to DNA (i.e., replaced the amino acids by their original codons) and concatenated the resulting alignments. We then made the phylogenetic inference using *IQ-TREE* (*v1.6.7.2*) [70] using ModelFinder (-m TEST) [71] and assessed the robustness of the phylogenetic inference by calculating 1,000 ultra-fast bootstraps (-bb 1,000) [72]. There were 220,912 parsimony-informative sites over a total alignment of 1,455,396 bp, and the best-fit model without gamma correction was a general time-reversible model with empirical base frequencies allowing for invariable sites (GTR+F+I). We did not use the gamma correction because of branch length scaling issues, which were 10 times longer than with simpler models, and is related to an optimization problem with big datasets in IQ-TREE. The tree is very well supported, since the average ultra-fast bootstrap support value was 97.6% and the median was 100%. We placed the Kpn species root according to the outgroup formed by the 22 *Kqs* strains. The tree, along with Kleborate annotations, can be visualized and manipulated in https://microreact.org/project/kk6mmVEDfa1o3pGQSCobdH/9f09a4c3.

Capsule locus typing

We used Kaptive [17] with default options and the "K locus primary reference" to identify the CLT of strains. The predicted CLT is assigned a confidence level, which relies on the overall alignment to the reference CLT, the allelic composition of the locus, and its fragmentation level. We assigned the CLT to "unknown" when the confidence level of Kaptive was indicated

as "none" or "low," as suggested by the authors of the software. This only represented 7.9% of the genomes. After this filtering, we simply considered that 2 CLT are the same if they are both annotated with the same KLx name.

Identification of capsule pseudogenes and inactive capsule loci

We first compiled the list of missing expected genes from Kaptive, which is only computed by Kaptive for capsule loci encoded in a single contig. Then, we used the Kaptive reference database of Kpn capsule loci to retrieve capsule reference genes for all the identified serotypes. We searched for sequence similarity between the proteins of the reference dataset and the 3,980 genome assemblies using blastp and tblastn (v.2.9.0) [73]. We then searched for the following indications of pseudogenization: stop codons resulting in protein truncation, frameshift mutations, insertions, and deletions (https://doi.org/10.6084/m9.figshare.14673153). Truncated and frameshifted coding sequences covering at least 80% of the original protein in the same reading frame were considered functional. Additionally, a pseudogene did not result in a classification of inactivated function if we could identify an intact homolog or analog. For example, if *wcaJ_KL1* has a frameshift, but *wcaJ_KL2* was found in the genome, the pseudogene was flagged and not used to define non-capsulated mutants. Complete gene deletions were identified by Kaptive among capsular loci encoded on a single contig. We built a dictionary of genes that are essential for capsule production by gathering a list of genes (annotated as the gene name in the Kaptive database) present across all CLTs and which are essential for capsule production according to experimental evidence (Table A in S1 Text). The absence of a functional copy of these essential genes resulted in the classification "non-capsulated" (except *wcaJ* and *wbaP*, which are mutually exclusive). To correlate the pseudogenization frequency with the order in the capsular biosynthesis process, we first sought to identify all glycosyl transferases from the different CLTs and grouped them in one category. To do so, we retrieved the GO molecular functions listed on UniProtKB of the genes within the Kaptive reference database. For the genes that could be ordered in the biosynthesis chain (Table A in S1 Text), we computed their frequency of inactivation by dividing the count of inactivated genes by the total number of times it is present in the dataset.

To test that sequencing errors and contig breaks were not leading to an excess of pseudogenes in certain genomes, we correlated the number of pseudogenes (up to 11) and missing genes with 2 indexes of sequence quality, namely, the sequence length of the shortest contig at 50% of the total genome length (N50) and the smallest number of contigs whose length sum makes up 90% of genome size (L90). We found no significant correlation in both cases (Spearman correlation test, p-values > 0.05), suggesting that our results are not strongly affected by sequencing artifacts and assembly fragmentation.

Genetic similarity

We searched for sequence similarity between all proteins of all prophages or conjugative systems using MMSeqs2 with the sensitivity parameter set to 7.5. The hits were filtered (e-value $< 10^{-5}$, $\geq$35% identity, coverage >50% of the proteins) and used to compute the set of bidirectional best hits (BBH) between each genome pair. BBH were used to compute the gene repertoire relatedness between pairs of genomes (weighted by sequence identity):

$$\mathrm{wGRR}_{A,B} = \sum_i \frac{id(A_i, B_i)}{\min(\#A, \#B)},$$

as previously described [74], where A_i and B_i are the pair i of homologous proteins present in A and B (containing respectively $\#A$ and $\#B$ proteins), $id(A_i, B_i)$ is their sequence identity, and

min(#*A*, #*B*) is the number of proteins of the element encoding the fewest proteins (#*A* or #*B*). wGRR varies between 0 and 1. It amounts to 0 if there are no homologous proteins between the genomes, and one if all genes of the smaller genome have a homolog in the other genome. Hence, the wGRR accounts for both frequency of homology and degree of similarity among homologs.

Inference of genes ancestral states

We inferred the ancestral state of each pangenome family with PastML (v1.9.23) [75] using the maximum-likelihood algorithm MPPA and the F81 model. We also tried to run Count [76] with the ML method to infer gene gains and losses from the pangenome, but this took a prohibitive amount of computing time. To check that PastML was producing reliable results, we split our species tree (*cuttree* function in R, package stats) in 50 groups and for the groups that took less than a month of computing time with Count (2,500 genomes), we compared the results of Count to those of PastML. The 2 methods were highly correlated in term of number of inferred gains per branch (Spearman correlation test, Rho = 0.88, p-value < 0.0001). We used the results of PastML, since it was much faster and could handle the whole tree in a single analysis. Since the MPPA algorithm can keep several ancestral states per node if they have similar and high probabilities, we only counted gene gains when both ancestral and descendant nodes had one single distinct state (absent ➔ present).

Analysis of conjugative systems

To detect conjugative systems, Type IV secretion systems, relaxases, and infer their MPF types, we used TXSScan with default options [77]. We then extracted the protein sequence of the conjugation systems and used these sequences to build clusters of systems by sequence similarity. We computed the wGRR (see "Genetic similarity") between all pairs of systems and clustered them in wGRR families by transitivity when the wGRR was higher than 0.99. This means that some members of the same family can have a wGRR <0.99. This threshold was defined based on the analysis of the shape of the distribution of the wGRR (S5A Fig). We used a reconstruction of the presence of members of each gene family in the species phylogenetic tree to infer the history of acquisition of conjugative elements (see "Inference of genes ancestral state"). To account for the presence of orthologous families, i.e., those coming from the same acquisition event, we kept only 1 member of a wGRR family per acquisition event. For example, if a conjugative system of the same family is present in 4 strains, but there were 2 acquisition events, we randomly picked 1 representative system for each acquisition event (in this case, 2 elements, 1 per event). Elements that resulted from the same ancestral acquisition event are referred as orthologous systems. We combined the predictions of mlplasmids and TXSScan to separate conjugative plasmids from ICEs. The distribution of conjugative system's MPF type in the chromosomes and plasmids is shown in S7 Fig.

Prophage detection

We used PHASTER [78] to identify prophages in the genomes (accessed in December 2018). The category of the prophage is given by a confidence score that corresponds to "intact," "questionable," or "incomplete." We kept only the "intact" prophages because other categories often lack essential phage functions. We further removed prophage sequences containing more than 3 transposases after annotation with ISFinder [79] because we noticed that some loci predicted by PHASTER were composed of arrays of insertion sequences. We built clusters of nearly identical prophages with the same method used for conjugative systems. The wGRR threshold for clustering was also defined using the shape of the distribution (S5B Fig). The

definition of orthologous prophages follows the same principle than that of conjugative systems, they are elements that are predicted to result from one single past event of infection.

Serotype swaps identification

We inferred the ancestral state of the capsular CLT with PastML using the maximum-likelihood algorithm MPPA, with the recommended F81 model [75]. In the reconstruction procedure, the low confidence CLTs were treated as missing data. This analysis revealed that serotype swaps happen at a rate of 0.282 swaps per branch, which are, on average, 0.000218 substitutions/site long in our tree. CLT swaps were defined as the branches where the descendant node state was not present in the ancestral node state. In 92% of the swaps identified by MPPA, there was only 1 state predicted for both ancestor and descendant node, and we could thus precisely identify the CLT swaps. These swaps were used to generate the network in Fig 3A.

Detection of recombination tracts

We detected recombination tracts with Gubbins v2.4.0 [34]. Our dataset is too large to build one meaningful whole-genome alignment (WGA). Gubbins is designed to work with closely related strains, so we split the dataset into smaller groups defined by a single ST. We then aligned the genomes of each ST with snippy v4.3.8 (https://github.com/tseemann/snippy), as recommended by the authors in the documentation. The reference genome was picked randomly among the complete assemblies of each ST. We analyzed the 25 groups in which a CLT swap happened (see above) and for which a complete genome was available as a reference. We launched Gubbins independently for each WGA, using default parameters. We focused on the terminal branches to identify the recombination tracts resulting in CLT swap. We enquired on the origin of the recombined DNA using a sequence similarity approach. We used blastn [73] (-task megablast) to find the closest match of each recombination tract by querying the full tract against our dataset of genome assemblies and mapped the closest match based on the bit score onto the species tree.

Identification of co-gains

We used the ancestral state reconstruction of the pangenome families to infer gene acquisitions at the terminal branches. We then quantified how many times an acquisition of the same gene family of the pangenome (i.e., co-gains) occurred independently in genomes of the same CLT. This number was compared to the expected number given by a null model where the CLT does not impact the gene flow. The distribution of the expectation of the null model was made by simulation in R, taking into account the phylogeny and the distribution of CLTs. In each simulation, we used the species tree to randomly redistribute the CLT trait on the terminal branches (keeping the frequencies of CLTs equal to those of the original data). We ran 1,000 simulations and compared them with the observed values with a 1-sample Z-test [80]:

$$Acquisition\ specificity\ score = \sum_g \frac{I_g \times (I_g - 1)}{T_g \times (T_g - 1)},$$

where the numerator is the number of pairs with gains in a CLT, and the denominator is the number of all possible pairs. With each gene family of the pangenome g, the number of gene gains in strains of the same CLT I, and the total number of gene gains T. This corresponds to the sum of total number of co-gains within a CLT, normalized by the total number of co-gains for each gene. This score captures the amount of gene acquisitions that happened within

strains of the same CLT. If the observed score is significantly different than the simulations assuming random distribution, it means that there was more genetic exchange within CLT groups than expected by chance.

CLT specificity

We used the ancestral reconstruction of the acquisition of prophages and conjugative systems to count the number of distinct CLT in which such an acquisition occurred. For example, one prophage family can be composed of 10 members, coming from 5 distinct infection events in the tree: 2 in KL1 bacteria and 3 in KL2 bacteria. Therefore, we count 5 acquisitions in 2 CLT ($CLT_{obs} = 2$). The null model is that of no CLT specificity. The distribution of the expected number of CLT infected following the null model was generated by simulation ($n = 1{,}000$), as described above (see "Identification of co-gains"), and we plotted the specificity score as follows:

$$\text{Specificity score} = \frac{CLT_{obs}}{(CLT_{obs} + \overline{CLT_{exp}})},$$

where CLT_{obs} is the observed number of CLT infected, and $\overline{CLT_{exp}}$ is the mean number of CLT infected in the simulations. Thus, the expected value under nonspecificity is 0.5.

Under our example of 5 acquisitions in 2 CLT ($CLT_{obs} = 2$), a $\overline{CLT_{exp}}$ of 2 would mean that there is no difference between the observed and expected distributions across CLTs, and the specificity score is 0.5. Values lower than 0.5 indicate a bias toward regrouping of elements in a smaller than expected number of CLTs, whereas values higher than 0.5 indicate over-dispersion across CLTs. The statistics computed on Fig 5 are the comparison of all the specificity scores for all the prophage and conjugative systems families to the null model (score = 0.5) with a 1-sample Wilcoxon signed rank test.

Handling of draft assemblies

Since more than 90% of our genome dataset is composed of draft assemblies, i.e., genomes composed of several chromosomal contigs, we detail here the steps undertaken to reduce the impact of such fragmentation on our analysis. We only included prophages and conjugative systems that are localized on the same contig (see "Prophage detection" and "Analysis of conjugative systems"). Kaptive is able to handle draft assemblies and adjust the confidence score accordingly when the capsule locus is fragmented, so we relied on the Kaptive confidence score to annotate the CLT, which was treated as missing data in all the analysis when the score was below "Good" (see "Capsule locus typing"). For the detection of missing capsular genes, performed by Kaptive, we verified that only non-fragmented capsular loci are included (see "Identification of capsule pseudogenes and inactive capsule loci"). For the detection of capsule pseudogenes, we included all assemblies and flagged pseudogenes that were localized on the border of a contig (last gene on the contig). Out of the 502 inactivated/missing genes, 47 were localized at the border of a contig. We repeated the analysis presented on Fig 3B after removing these pseudogenes and found an even better fit for the linear model at $R^2 = 0.77$ and $p = 0.004$. Of note, such contig breaks are likely due to IS insertions, forming repeated regions that are hard to assemble, so we kept them in the main analysis.

Analyses of lab-evolved non-capsulated clones

To pinpoint the mechanisms by which a diverse set of strains became non-capsulated, we took advantage of an experiment performed in our lab and described previously in [36]. Briefly, 3

independent overnight cultures of 8 strains (Table B in S1 Text) were diluted 1:100 into 5 mL of fresh LB and incubated at 37˚C under agitation. Each independent population was diluted 1:100 into fresh LB every 24 hours for 3 days (approximately 20 generations). We then plated serial dilutions of each population. A single non-capsulated clone per replicate population was isolated based on their translucent colony morphology, except in 2 replicate populations where all colonies plated were capsulated. We performed DNA extraction with the guanidium thiocyanate method [81], with modifications. DNA was extracted from pelleted cells grown overnight in LB supplemented with 0.7 mM EDTA. Additionally, RNAse A treatment (37˚C, 30 minutes) was performed before DNA precipitation. Each clone ($n = 22$) was sequenced by Illumina with 150pb paired-end reads, yielding approximately 1 GB of data per clone. The reads were assembled with Unicycler v0.4.4 [82], and the assemblies were checked for pseudogenes (see "Identification of capsule pseudogenes and inactive capsule loci"). We also used *breseq* [37] and *snippy* (https://github.com/tseemann/snippy) to verify that there were no further undetected mutations in the evolved sequenced clones (https://doi.org/10.6084/m9.figshare.14673177).

Generation of capsule mutants

Isogenic capsule mutants were constructed by an in-frame deletion of *wcaJ* by allelic exchange as reported previously [36]. Deletion mutants were first verified by Sanger, and Illumina sequencing revealed that there were no off-target mutations.

Conjugation assay

Construction of pMEG-Mob plasmid. A mobilizable plasmid named pMEG-Mob was built by assembling the region containing the origin of transfer of the pKNG101 plasmid [83] and the region containing the origin of replication, kanamycin resistance cassette, and IPTG-inducible *cfp* from the pZE12:CFP plasmid [84] (see Table C in S1 Text, and plasmid map, S8 Fig). We amplified both fragments of interest by PCR with the Q5 high fidelity DNA polymerase from New England Biolabs (NEB), with primers adapted for Gibson assembly designed with Snapgene, and used the NEB HiFi Builder mix following the manufacturer's instructions to assemble the 2 fragments. The assembly product was electroporated into electro-competent *E. coli* DH5α strain. KmR colonies were isolated, and correct assembly was checked by PCR. Cloned pMEG-Mob plasmid was extracted using the QIAprep Spin Miniprep Kit, and electroporation into the donor strain *E. coli* MFD λ-pir strain [85]. The primers used to generate pMEG-Mob are listed in Table D in S1 Text.

Conjugation assay. Recipient strains of *Klebsiella* spp. were diluted at 1:100 from a LB overnight into fresh LB in a final volume of 3 mL. Donor strain *E. coli* MFD λ-pir strain (diaminopimelic acid (DAP) auxotroph), which is carrying the pMEG-Mob plasmid, exhibited slower growth than *Klebsiella* strains and was diluted at 1:50 from an overnight into fresh LB + DAP (0.3 mM) + Kanamycin (50 μg/ml). Cells were allowed to grow at 37˚C until late exponential growth phase (Optical density; OD of 0.9 to 1) and adjusted to an OD of 0.9. The cultures were then washed twice in LB and mixed at a 1:1 donor–recipient ratio. Donor–recipient mixes were then centrifuged for 5 minutes at 13,000 rpm, resuspended in 25-μL LB+DAP, and deposited on a MF-Millipore Membrane Filter (0.45 μm pore size) on nonselective LB+DAP plates. The mixes were allowed to dry for 5 minutes with the lid open, and then incubated at 37˚C. After 1 hour, membranes were resuspended in 1-mL phosphate-buffered saline (PBS) and thoroughly vortexed. Serial dilutions were plated on selective (LB+Km) and non-selective (LB+DAP) plates to quantify the number of transconjugants (T) and the total

number of cells (N). The conjugation efficiency was computed with the following:

$$\text{Conjugation efficiency} = \frac{T}{N}$$

This simple method is relevant in our experimental setup because the plasmid can only be transferred from the donor strain to the recipient strain, and the duration of the experiment only allowed for minimal growth [86]. The lack of the conjugative machinery of MPF type I (the MPF type of RK2) within the plasmid and in the recipient strains prevents the transfer across recipient strains (see "Analysis of conjugative systems").

Data analysis

All the data analyses were performed with R version 3.6 and Rstudio version 1.2. We used the packages ape v5.3 [87], phangorn v2.5.5 [88], and treeio v1.10 [89] for the phylogenetic analyses. Statistical tests were performed with the base package stats. For data frame manipulations and simulations, we also used dplyr v0.8.3 along with the tidyverse packages [90] and data.table v1.12.8.

Supporting information

S1 Fig. Comparison of 2 capsular loci. Two CLTs (KL112 and KL24) involved in CLT swap, with the essential genes for capsule expression colored in blue. Gray tracks correspond to the sequence identity (computed using blastn) above 90% (see scale) to indicate highly similar homologs (liable to recombine). CLT, capsular locus type.
(TIFF)

S2 Fig. Similarity between swapped CLT and other CLT. (A) Comparison of sugar composition similarity (Jaccard similarity) between swapped vs. others CLTs. **(B)** Comparison of genetic similarity (wGRR) between swapped vs. others CLTs. The *p*-value displayed is for the 2-sample Wilcoxon test (https://doi.org/10.6084/m9.figshare.14673180). CLT, capsular locus type; wGRR, gene repertoire relatedness weighted by sequence identity.
(TIFF)

S3 Fig. Phylogenetic distribution of the inactivated capsular loci. The blue dots represent the putative inactivated capsules, which have at least 1 essential gene for capsule production pseudogenized or deleted (https://doi.org/10.6084/m9.figshare.14673156).
(TIFF)

S4 Fig. Controls for the inactivated capsule gene analysis. (A) Linear regression between the inactivation frequency normalized by average gene length and the rank of each gene in the biosynthesis pathway ($p = 0.005$, $R^2 = 0.7$) (https://doi.org/10.6084/m9.figshare.14673183). **(B)** Number of SSR in the core capsule genes in the Kaptive reference database (https://doi.org/10.6084/m9.figshare.14673174). **(C)** Genetic diversity of core capsule genes within the Kaptive reference database, represented by the percent of identity of all pairwise alignments of the proteins from different reference capsule loci (https://doi.org/10.6084/m9.figshare.14673192). SSR, simple sequence repeats.
(TIFF)

S5 Fig. Distribution of the similarity measured by wGRR between pairs of conjugative systems (A) and between pairs of prophages (B) for wGRR >0. The arrows represent the threshold (wGRR >0.99) set for clustering into families of highly similar elements. Since we performed transitive clustering to build the families, some elements belonging to the same

families have wGRR <0.99. We annotated the distribution of conjugation systems belonging to the same MPF type, which shows that systems of the same MPF are very similar but are below the selected threshold for clustering (https://doi.org/10.6084/m9.figshare.14673144 and https://doi.org/10.6084/m9.figshare.14673186). MPF, mating pair formation; wGRR, gene repertoire relatedness weighted by sequence identity.
(TIFF)

S6 Fig. Changes in capsule state and branch length. The capsule state changes among branches of the species tree are represented on the *x* axis, and the branch length is represented on the *y* axis in substitution per site. Individual points represent the mean for each group, and the bars represent the standard error. The *p*-values for *the t* test are represented on top of each comparisons (https://doi.org/10.6084/m9.figshare.14673159). We also performed a 2-sample Wilcoxon test to compare the medians ("Others" vs. "inactivation": $p < 0.0001$; "Others" vs. "Swap": $p < 0.0001$).
(TIFF)

S7 Fig. Distribution of conjugation system MPF types. Conjugation systems are classified in 2 categories according to their genomic location, which was predicted with the mlplasmids classifier. The MPF was predicted with the CONJscan module of MacSyfinder. Absolute number of systems are displayed for each category (https://doi.org/10.6084/m9.figshare.14673189). MPF, mating pair formation.
(TIFF)

S8 Fig. pMEG-Mob plasmid genetic map. pMEG-Mob was constructed by Gibson assembly from plasmids pKNG101 and pZE12. It encodes a colE1/pUC origin of replication (high copy number), a selectable marker (Kanamycin resistance cassette, green), the mobilizable region of pKNG101 which is composed of the origin of transfer of RK2 and 2 genes involved in conjugation (*traJ* and *traK*), a counter selectable marker (*sacB*), and an inducible CFP gene (IPTG induction). pMEG-Mob can only be mobilized in *trans* and thus can only be transferred from a strain expressing the RK2 conjugative machinery, which is absent from the panel of strains we used as recipients.
(TIFF)

S1 Text. Supporting tables with references. Supporting information containing detailed list of essential capsule genes, strains, plasmids, and primers used in this study.
(PDF)

Acknowledgments

We thank Rafał Mostowy, Jorge Moura de Sousa, Nienke Buddlelmeijer, Olivier Tenaillon, Marie Touchon, and other lab members for fruitful discussions. We thank Sylvain Brisse for providing us with *Klebsiella* strains. We thank Christiane Forestier and Damien Balestrino for providing the pKNG101 plasmid and Jean-Marc Ghigo and Christophe Beloin for the gift of pZE12::CFP used to construct pMEG-Mob and *E. coli* MFD λ-pir.

Author Contributions

Conceptualization: Matthieu Haudiquet, Olaya Rendueles, Eduardo P. C. Rocha.

Data curation: Matthieu Haudiquet, Olaya Rendueles, Eduardo P. C. Rocha.

Formal analysis: Matthieu Haudiquet.

Funding acquisition: Olaya Rendueles, Eduardo P. C. Rocha.

Investigation: Matthieu Haudiquet, Olaya Rendueles, Eduardo P. C. Rocha.

Methodology: Matthieu Haudiquet.

Project administration: Olaya Rendueles, Eduardo P. C. Rocha.

Resources: Matthieu Haudiquet, Amandine Buffet, Olaya Rendueles, Eduardo P. C. Rocha.

Software: Matthieu Haudiquet, Eduardo P. C. Rocha.

Supervision: Olaya Rendueles, Eduardo P. C. Rocha.

Validation: Matthieu Haudiquet, Olaya Rendueles, Eduardo P. C. Rocha.

Visualization: Matthieu Haudiquet.

Writing – original draft: Matthieu Haudiquet, Olaya Rendueles, Eduardo P. C. Rocha.

Writing – review & editing: Matthieu Haudiquet, Olaya Rendueles, Eduardo P. C. Rocha.

References

1. Diard M, Hardt W-D. Evolution of bacterial virulence. FEMS Microbiol Rev. 2017; 41:679–697. https://doi.org/10.1093/femsre/fux023 PMID: 28531298
2. Navon-Venezia S, Kondratyeva K, Carattoli A. *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. FEMS Microbiol Rev. 2017; 41:252–275. https://doi.org/10.1093/femsre/fux013 PMID: 28521338
3. Cabezón E, Ripoll-Rozada J, Peña A, de la Cruz F, Arechaga I. Towards an integrated model of bacterial conjugation. FEMS Microbiol Rev. 2015; 39:81–95. https://doi.org/10.1111/1574-6976.12085 PMID: 25154632
4. Touchon M, Moura de Sousa JA, Rocha EP. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. Curr Opin Microbiol. 2017; 38:66–73. https://doi.org/10.1016/j.mib.2017.04.010 PMID: 28527384
5. Bertozzi Silva J, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. FEMS Microbiol Lett. 2016; 363:fnw002. https://doi.org/10.1093/femsle/fnw002 PMID: 26755501
6. de la Cruz F, Frost LS, Meyer RJ, Zechner EL. Conjugative DNA metabolism in Gram-negative bacteria. FEMS Microbiol Rev. 2010; 34:18–40. https://doi.org/10.1111/j.1574-6976.2009.00195.x PMID: 19919603
7. 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; 37:186–192. https://doi.org/10.1038/s41587-018-0009-7 PMID: 30718869
8. Rice LB. Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE. J Infect Dis. 2008; 197:1079–1081. https://doi.org/10.1086/533452 PMID: 18419525
9. Yang X, Wai-Chi Chan E, Zhang R, Chen S. A conjugative plasmid that augments virulence in Klebsiella pneumoniae. Nat Microbiol 2019; 4:2039–2043. https://doi.org/10.1038/s41564-019-0566-7 PMID: 31570866
10. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. PLoS Genet. 2019;15. https://doi.org/10.1371/journal.pgen.1008114 PMID: 30986243
11. Pan Y-J, Lin T-L, Chen C-T, Chen Y-Y, Hsieh P-F, Hsu C-R, et al. Genetic analysis of capsular polysaccharide synthesis gene clusters in 79 capsular types of Klebsiella spp. Sci Rep. 2015; 5:15573. https://doi.org/10.1038/srep15573 PMID: 26493302
12. Follador R, Heinz E, Wyres KL, Ellington MJ, Kowarik M, Holt KE, et al. The diversity of Klebsiella pneumoniae surface polysaccharides. Microb Genom. 2016; 2:e000073. https://doi.org/10.1099/mgen.0.000073 PMID: 28348868
13. Rendueles O, Garcia-Garcerà M, Néron B, Touchon M, Rocha EPC. Abundance and co-occurrence of extracellular capsules increase environmental breadth: Implications for the emergence of pathogens. PLoS Pathog. 2017; 13:e1006525. https://doi.org/10.1371/journal.ppat.1006525 PMID: 28742161

14. Wyres KL, Gorrie C, Edwards DJ, Wertheim HFL, Hsu LY, Van Kinh N, et al. Extensive Capsule Locus Variation and Large-Scale Genomic Recombination within the *Klebsiella pneumoniae* Clonal Group 258. Genome Biol Evol. 2015; 7:1267–1279. https://doi.org/10.1093/gbe/evv062 PMID: 25861820

15. Mostowy RJ, Holt KE. Diversity-Generating Machines: Genetics of Bacterial Sugar-Coating. Trends Microbiol. 2018; 26:1008–1021. https://doi.org/10.1016/j.tim.2018.06.006 PMID: 30037568

16. Holt KE, Lassalle F, Wyres KL, Wick R, Mostowy RJ. Diversity and evolution of surface polysaccharide synthesis loci in Enterobacteriales. ISME J. 2020; 14:1713–1730. https://doi.org/10.1038/s41396-020-0628-0 PMID: 32249276

17. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. Identification of *Klebsiella* capsule synthesis loci from whole genome data. Microb Genomics. 2016; 2:e000102. https://doi.org/10.1099/mgen.0.000102 PMID: 28348840

18. Wang H, Wilksch JJ, Lithgow T, Strugnell RA, Gee ML. Nanomechanics measurements of live bacteria reveal a mechanism for bacterial cell protection: the polysaccharide capsule in *Klebsiella* is a responsive polymer hydrogel that adapts to osmotic stress. Soft Matter. 2013; 9:7560–7567. https://doi.org/10.1039/C3SM51325D

19. Campos MA, Vargas MA, Regueiro V, Llompart CM, Albertí S, Bengoechea JA. Capsule polysaccharide mediates bacterial resistance to antimicrobial peptides. Infect Immun. 2004; 72:7107–7114. https://doi.org/10.1128/IAI.72.12.7107-7114.2004 PMID: 15557634

20. Cortés G, Borrell N, de Astorza B, Gómez C, Sauleda J, Albertí S. Molecular Analysis of the Contribution of the Capsular Polysaccharide and the Lipopolysaccharide O Side Chain to the Virulence of Klebsiella pneumoniae in a Murine Model of Pneumonia. Infect Immun. 2002; 70:2583. https://doi.org/10.1128/IAI.70.5.2583-2590.2002 PMID: 11953399

21. Fernebro J, Andersson I, Sublett J, Morfeldt E, Novak R, Tuomanen E, et al. Capsular Expression in Streptococcus pneumoniae Negatively Affects Spontaneous and Antibiotic-Induced Lysis and Contributes to Antibiotic Tolerance. J Infect Dis. 2004; 189:328–338. https://doi.org/10.1086/380564 PMID: 14722899

22. Soundararajan M, von Bünau R, Oelschlaeger TA. K5 Capsule and Lipopolysaccharide Are Important in Resistance to T4 Phage Attack in Probiotic E. coli Strain Nissle 1917. Front Microbiol. 2019; 10:2783. https://doi.org/10.3389/fmicb.2019.02783 PMID: 31849915

23. Latka A, Maciejewska B, Majkowska-Skrobek G, Briers Y, Drulis-Kawa Z. Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during the infection process. Appl Microbiol Biotechnol. 2017; 101:3103–3119. https://doi.org/10.1007/s00253-017-8224-6 PMID: 28337580

24. Pan Y-J, Lin T-L, Chen C-C, Tsai Y-T, Cheng Y-H, Chen Y-Y, et al. *Klebsiella* Phage ΦK64-1 Encodes Multiple Depolymerases for Multiple Host Capsular Types. J Virol. 2017; 91:e02457–16. https://doi.org/10.1128/JVI.02457-16 PMID: 28077636

25. de Sousa JAM, Buffet A, Haudiquet M, Rocha EPC, Rendueles O. Modular prophage interactions driven by capsule serotype select for capsule loss under phage predation. ISME J. 2020; 14:2980–2996. https://doi.org/10.1038/s41396-020-0726-z PMID: 32732904

26. Hsieh P-F, Lin H-H, Lin T-L, Chen Y-Y, Wang J-T. Two T7-like Bacteriophages, K5-2 and K5-4, Each Encodes Two Capsule Depolymerases: Isolation and Functional Characterization. Sci Rep. 2017; 7:4624. https://doi.org/10.1038/s41598-017-04644-2 PMID: 28676686

27. Stuy JH. Plasmid transfer in *Haemophilus influenzae*. J Bacteriol. 1979; 139:520–529. https://doi.org/10.1128/jb.139.2.520-529.1979 PMID: 313393

28. Johnston C, Martin B, Fichant G, Polard P, Claverys J-P. Bacterial transformation: distribution, shared mechanisms and divergent control. Nat Rev Microbiol. 2014; 12:181–196. https://doi.org/10.1038/nrmicro3199 PMID: 24509783

29. Rendueles O, de Sousa JAM, Bernheim A, Touchon M, Rocha EPC. Genetic exchanges are more frequent in bacteria encoding capsules. PLoS Genet. 2018; 14:e1007862. https://doi.org/10.1371/journal.pgen.1007862 PMID: 30576310

30. Buerret M, Joseleau J-P. Depolymerization of the capsular polysaccharide from *Klebsiella* K19 by the glycanase associated with particles of *Klebsiella* bacteriophage φ19. Carbohydr Res. 1986; 157:27–51. https://doi.org/10.1016/0008-6215(86)85058-3 PMID: 3815416

31. Rieger-Hug D, Stirm S. Comparative study of host capsule depolymerases associated with *Klebsiella* bacteriophages. Virology. 1981; 113:363–378. https://doi.org/10.1016/0042-6822(81)90162-8 PMID: 7269247

32. Thurow H, Niemann H, Stirm S. Bacteriophage-borne enzymes in carbohydrate chemistry: Part I. On the glycanase activity associated with particles of *Klebsiella* bacteriophage No. 11. Carbohydr Res. 1975; 41:257–271. https://doi.org/10.1016/s0008-6215(00)87024-x PMID: 236830

33. Patro LPP, Rathinavelan T. Targeting the Sugary Armor of Klebsiella Species. Front Cell Infect Microbiol. 2019; 9:367. https://doi.org/10.3389/fcimb.2019.00367 PMID: 31781512

34. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015; 43:e15. https://doi.org/10.1093/nar/gku1196 PMID: 25414349

35. Pagel M. Inferring the historical patterns of biological evolution. Nature. 1999; 401:877–884. https://doi.org/10.1038/44766 PMID: 10553904

36. Buffet A, Rocha EPC, Rendueles O. Nutrient conditions are primary drivers of bacterial capsule maintenance in Klebsiella. Proc Biol Sci. 2021; 288:20202876. https://doi.org/10.1098/rspb.2020.2876 PMID: 33653142

37. Deatherage DE, Barrick JE. Identification of mutations in laboratory evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol. 2014; 1151:165–188. https://doi.org/10.1007/978-1-4939-0554-6_12 PMID: 24838886

38. Hesse S, Rajaure M, Wall E, Johnson J, Bliskovsky V, Gottesman S, et al. Phage Resistance in Multidrug-Resistant Klebsiella pneumoniae ST258 Evolves via Diverse Mutations That Culminate in Impaired Adsorption. MBio. 2020; 11:e02530–19. https://doi.org/10.1128/mBio.02530-19 PMID: 31992617

39. Tan D, Zhang Y, Qin J, Le S, Gu J, Chen L, et al. A Frameshift Mutation in wcaJ Associated with Phage Resistance in Klebsiella pneumoniae. Microorganisms. 2020; 8:378. https://doi.org/10.3390/microorganisms8030378 PMID: 32156053

40. Cai R, Wang G, Le S, Wu M, Cheng M, Guo Z, et al. Three Capsular Polysaccharide Synthesis-Related Glucosyltransferases, GT-1, GT-2 and WcaJ, Are Associated With Virulence and Phage Sensitivity of Klebsiella pneumoniae. Front Microbiol. 2019;10. https://doi.org/10.3389/fmicb.2019.00010 PMID: 30728810

41. Cury J, Oliveira PH, de la Cruz F, Rocha EPC. Host Range and Genetic Plasticity Explain the Coexistence of Integrative and Extrachromosomal Mobile Genetic Elements. Mol Biol Evol. 2018; 35:2230–2239. https://doi.org/10.1093/molbev/msy123 PMID: 29905872

42. Wyres KL, Lam MMC, Holt KE. Population genomics of *Klebsiella pneumoniae*. Nat Rev Microbiol. 2020; 18:344–359. https://doi.org/10.1038/s41579-019-0315-1 PMID: 32055025

43. Chen Y, Marimuthu K, Teo J, Venkatachalam I, Cherng BPZ, De Wang L, et al. Acquisition of Plasmid with Carbapenem-Resistance Gene blaKPC2 in Hypervirulent *Klebsiella pneumoniae*, Singapore. Emerg Infect Dis. 2020; 26:549–559. https://doi.org/10.3201/eid2603.191230 PMID: 32091354

44. Lam MMC, Wyres KL, Wick RR, Judd LM, Fostervold A, Holt KE, et al. Convergence of virulence and MDR in a single plasmid vector in MDR *Klebsiella pneumoniae* ST15. J Antimicrob Chemother. 2019; 74:1218–1222. https://doi.org/10.1093/jac/dkz028 PMID: 30770708

45. Lan P, Jiang Y, Zhou J, Yu Y. A global perspective on the convergence of hypervirulence and carbapenem resistance in Klebsiella pneumoniae. J Glob Antimicrob Resist. 2021; 25:26–34. https://doi.org/10.1016/j.jgar.2021.02.020 PMID: 33667703

46. Hernández M, López-Urrutia L, Abad D, De Frutos Serna M, Ocampo-Sosa AA, Eiros JM. First Report of an Extensively Drug-Resistant ST23 Klebsiella pneumoniae of Capsular Serotype K1 Co-Producing CTX-M-15, OXA-48 and ArmA in Spain. Antibiotics (Basel). 2021;10. https://doi.org/10.3390/antibiotics10020157 PMID: 33557209

47. Pérez-Mendoza D, de la Cruz F. *Escherichia coli* genes affecting recipient ability in plasmid conjugation: Are there any? BMC Genomics. 2009; 10:71. https://doi.org/10.1186/1471-2164-10-71 PMID: 19203375

48. Chang B, Nariai A, Sekizuka T, Akeda Y, Kuroda M, Oishi K, et al. Capsule Switching and Antimicrobial Resistance Acquired during Repeated *Streptococcus pneumoniae* Pneumonia Episodes. J Clin Microbiol. 2015; 53:3318–3324. https://doi.org/10.1128/JCM.01222-15 PMID: 26269621

49. Swartley JS, Marfin AA, Edupuganti S, Liu LJ, Cieslak P, Perkins B, et al. Capsule switching of *Neisseria meningitidis*. Proc Natl Acad Sci U S A. 1997; 94:271–276. https://doi.org/10.1073/pnas.94.1.271 PMID: 8990198

50. Verma V, Harjai K, Chhibber S. Restricting ciprofloxacin-induced resistant variant formation in biofilm of *Klebsiella pneumoniae* B5055 by complementary bacteriophage treatment. J Antimicrob Chemother. 2009; 64:1212–1218. https://doi.org/10.1093/jac/dkp360 PMID: 19808232

51. Tan YH, Chen Y, Chu WHW, Sham L-T, Gan Y-H. Cell envelope defects of different capsule-null mutants in K1 hypervirulent *Klebsiella pneumoniae* can affect bacterial pathogenesis. Mol Microbiol. 2020; 113:889–905. https://doi.org/10.1111/mmi.14447 PMID: 31912541

52. Bai J, Dai Y, Farinha A, Tang AY, Syal S, Vargas-Cuebas G, et al. Essential gene analysis in Acinetobacter baumannii by high-density transposon mutagenesis and CRISPR interference. J Bacteriol. 2021. https://doi.org/10.1128/JB.00565-20 PMID: 33782056

53. Andam CP, Hanage WP. Mechanisms of genome evolution of *Streptococcus*. Infect Genet Evol. 2015; 33:334–342. https://doi.org/10.1016/j.meegid.2014.11.007 PMID: 25461843

54. Croucher NJ, Kagedan L, Thompson CM, Parkhill J, Bentley SD, Finkelstein JA, et al. Selective and Genetic Constraints on Pneumococcal Serotype Switching. PLoS Genet. 2015;11. https://doi.org/10.1371/journal.pgen.1005095 PMID: 25826208

55. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014; 46:305–309. https://doi.org/10.1038/ng.2895 PMID: 24509479

56. Nanayakkara BS, O'Brien CL, Gordon DM. Diversity and distribution of *Klebsiella* capsules in *Escherichia coli*. Environ Microbiol Rep. 2019; 11:107–117. https://doi.org/10.1111/1758-2229.12710 PMID: 30411512

57. Lang AS, Zhaxybayeva O, Beatty JT. Gene transfer agents: phage-like elements of genetic exchange. Nat Rev Microbiol. 2012; 10:472–482. https://doi.org/10.1038/nrmicro2802 PMID: 22683880

58. Westbye AB, Kuchinski K, Yip CK, Beatty JT. The Gene Transfer Agent RcGTA Contains Head Spikes Needed for Binding to the *Rhodobacter capsulatus* Polysaccharide Cell Capsule. J Mol Biol. 2016; 428:477–491. https://doi.org/10.1016/j.jmb.2015.12.010 PMID: 26711507

59. Flammann HT, Weckesser J. Composition of the cell wall of the phage resistant mutant Rhodopseudomonas capsulata St. Louis RC1-. Arch Microbiol. 1984; 139:33–37. https://doi.org/10.1007/BF00692708

60. Brimacombe CA, Stevens A, Jun D, Mercer R, Lang AS, Beatty JT. Quorum-sensing regulation of a capsular polysaccharide receptor for the Rhodobacter capsulatus gene transfer agent (RcGTA). Mol Microbiol. 2013; 87:802–817. https://doi.org/10.1111/mmi.12132 PMID: 23279213

61. Townsend EM, Kelly L, Gannon L, Muscatt G, Dunstan R, Michniewski S, et al. Isolation and Characterization of Klebsiella Phages for Phage Therapy. Phage (New Rochelle). 2021; 2:26–42. https://doi.org/10.1089/phage.2020.0046 PMID: 33796863

62. Kostina E, Ofek I, Crouch E, Friedman R, Sirota L, Klinger G, et al. Noncapsulated *Klebsiella pneumoniae* bearing mannose-containing O antigens is rapidly eradicated from mouse lung and triggers cytokine production by macrophages following opsonization with surfactant protein D. Infect Immun. 2005; 73:8282–8290. https://doi.org/10.1128/IAI.73.12.8282-8290.2005 PMID: 16299325

63. Ernst CM, Braxton JR, Rodriguez-Osorio CA, Zagieboylo AP, Li L, Pironti A, et al. Adaptive evolution of virulence and persistence in carbapenem-resistant *Klebsiella pneumoniae*. Nat Med. 2020; 26:705–711. https://doi.org/10.1038/s41591-020-0825-4 PMID: 32284589

64. Perrin A, Rocha EPC. PanACoTA: A modular tool for massive microbial comparative genomics. bioRxiv. 2020 [cited 5 Oct 2020]. https://doi.org/10.1101/2020.09.11.293472

65. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016; 17:132. https://doi.org/10.1186/s13059-016-0997-x PMID: 27323842

66. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153 PMID: 24642063

67. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017; 35:1026–1028. https://doi.org/10.1038/nbt.3988 PMID: 29035372

68. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, Corander J, et al. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. Microb Genomics. 2018; 4:e000224. https://doi.org/10.1099/mgen.0.000224 PMID: 30383524

69. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30:772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

70. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32:268–274. https://doi.org/10.1093/molbev/msu300 PMID: 25371430

71. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017; 14:587–589. https://doi.org/10.1038/nmeth.4285 PMID: 28481363

72. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018; 35:518–522. https://doi.org/10.1093/molbev/msx281 PMID: 29077904

73. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. https://doi.org/10.1186/1471-2105-10-421 PMID: 20003500

74. Bobay L-M, Rocha EPC, Touchon M. The Adaptation of Temperate Bacteriophages to Their Host Genomes. Mol Biol Evol. 2013; 30:737–751. https://doi.org/10.1093/molbev/mss279 PMID: 23243039

75. Ishikawa SA, Zhukova A, Iwasaki W, Gascuel O. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. Mol Biol Evol. 2019; 36:2069–2085. https://doi.org/10.1093/molbev/msz131 PMID: 31127303

76. Csűös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics. 2010; 26:1910–1912. https://doi.org/10.1093/bioinformatics/btq315 PMID: 20551134

77. Abby SS, Rocha EPC. Identification of Protein Secretion Systems in Bacterial Genomes Using MacSyFinder. Methods Mol Biol. 2017; 1615:1–21. https://doi.org/10.1007/978-1-4939-7033-9_1 PMID: 28667599

78. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016; 44:W16–21. https://doi.org/10.1093/nar/gkw387 PMID: 27141966

79. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006; 34:D32–36. https://doi.org/10.1093/nar/gkj014 PMID: 16381877

80. Touchon M, Perrin A, de Sousa JAM, Vangchhia B, Burn S, O'Brien CL, et al. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. PLoS Genet. 2020; 16:e1008866. https://doi.org/10.1371/journal.pgen.1008866 PMID: 32530914

81. Pitcher DG, Saunders NA, Owen RJ. Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. Lett Appl Microbiol. 1989; 8:151–156. https://doi.org/10.1111/j.1472-765X.1989.tb00262.x

82. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017; 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595 PMID: 28594827

83. Kaniga K, Delor I, Cornelis GR. A wide-host-range suicide vector for improving reverse genetics in gram-negative bacteria: inactivation of the blaA gene of *Yersinia enterocolitica*. Gene. 1991; 109:137–141. https://doi.org/10.1016/0378-1119(91)90599-7 PMID: 1756974

84. Lutz R, Bujard H. Independent and Tight Regulation of Transcriptional Units in *Escherichia coli* Via the LacR/O, the TetR/O and AraC/I1-I2 Regulatory Elements. Nucleic Acids Res. 1997; 25:1203–1210. https://doi.org/10.1093/nar/25.6.1203 PMID: 9092630

85. Ferrières L, Hémery G, Nham T, Guérout A-M, Mazel D, Beloin C, et al. Silent mischief: bacteriophage Mu insertions contaminate products of *Escherichia coli* random mutagenesis performed using suicidal transposon delivery plasmids mobilized by broad-host-range RP4 conjugative machinery. J Bacteriol. 2010; 192:6418–6427. https://doi.org/10.1128/JB.00621-10 PMID: 20935093

86. Zhong X, Droesch J, Fox R, Top EM, Krone SM. On the meaning and estimation of plasmid transfer rates for surface-associated and well-mixed bacterial populations. J Theor Biol. 2012; 294:144–152. https://doi.org/10.1016/j.jtbi.2011.10.034 PMID: 22085738

87. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019; 35:526–528. https://doi.org/10.1093/bioinformatics/bty633 PMID: 30016406

88. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011; 27:592–593. https://doi.org/10.1093/bioinformatics/btq706 PMID: 21169378

89. Wang L-G, Lam TT-Y, Xu S, Dai Z, Zhou L, Feng T, et al. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Mol Biol Evol. 2020; 37:599–603. https://doi.org/10.1093/molbev/msz240 PMID: 31633786

90. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019; 4:1686. https://doi.org/10.21105/joss.01686

Through the lens of experimental biology

Introduction

Experimental biology is a set of approaches that consist in the conduction of experiments to investigate biological phenomena. My research using comparative genomics revealed that the gene flow between *Klebsiella pneumoniae* strains had been shaped by the capsule by **1) favoring intra-serotype exchanges via phage-mediated HGT** and **2) favoring inter-serotype exchanges via conjugation-mediated HGT**. It also highlighted that, **3) non-capsulated variants were frequent and had higher plasmid acquisition rate**.

While finding **#1** could be explained by the fact that phages are serotype-specific, and fit with the theory that closely related strains have high rates of HGT, finding **#2** suggested that conjugation happened less frequently than expected between closely related cells. This result hinted that conjugation was affected by the capsule. Finding **#3** substantiated this hypothesis, showing that capsules could be barriers to the reception of conjugative elements. We thus wanted to understand how the capsule precisely impact conjugation. Since a capsule can cover both donor and recipients, and be of different serotypes, we investigated the donation and reception frequencies of conjugative plasmid in function of the serotype.

To precisely test the impact of the serotype on conjugation efficiency, I needed to **construct isogenic mutants with different serotypes.** This represented a challenge, since the capsule locus is approximatelu 30kb long and *K. pneumoniae* is not naturally competent. I also needed to **collect natural conjugative plasmids** with selectable markers and **devise a method to measure the conjugation frequency** from natural isolates of *K. pneumoniae.* In the following section, I present the protocol I have developed to construct serotype swap isogenic mutants. I also present an early version of the article based on this work.

Experimental protocol: Isogenic serotype swaps in Klebsiella pneumoniae via chromosomal engineering.

I have developed a new protocol to perform genetic exchange of capsular loci between natural isolates of *K. pneumoniae*. This protocol is the result of 8 months of trial and failure of different approaches, which have resulted in a reliable method based on Lambda Red recombineering and gap-repair mediated cloning of chromosomal loci.

I first attempted to swap capsule loci with a standard two-step allelic exchange method. For this, we relied on the highly similar bordering genes, *galF* and *ugd*, as the homology arms surrounding the loci, and cloned the K1 and K2 capsule locus via Gibson assembly into DH5alpha *E. coli* cells. Gibson assembly of the 30kb capsule locus on a vector backbone was challenging, because it required a lot of PCR primers and cycling optimization. I then conjugated the vectors into *K. pneumoniae* via the Mu-Free Donor *E. coli* strain [332], and selected for the first and second cross-over leading to capsule locus replacement. However, all the colonies with a positive PCR specific of the new capsule locus appeared non-capsulated on LB plates. Upon whole genome sequencing (WGS) of one BJ1::K1 and one NTUH-K2044::K2 clone, I validated that the new capsule locus replaced original one in its native locus and no other mutations were found outside the capsule locus. However, we identified the same insertion sequence in both swapped capsule loci, inserted in different places, and thus resulting in non-functional capsules. This IS is not found in the parental Kpn strains, suggesting that capsule inactivation happened prior to introduction in Kpn. Indeed, we traced back this IS from the genome of DH5alpha, and hypothesized that whole capsule loci of *K. pneumoniae* were rapidly inactivated when expressed into *E. coli* laboratory strains. Hence, I sought to develop a protocol to clone and swap capsule loci without the need for intermediate laboratory strains.

The first part of the protocol aims to generate a strain encoding its own capsule locus on a low-copy vector. A Δcapsule mutant of the focus *K. pneumoniae* strain is obtained by electroporation of a novel deletion cassette encoding the kanamycin resistance gene bordered by two FRT sites (KmFRT) and a restriction enzyme I-SceI cut site. The KmFRT cassette is then excised by FLP recombinase, leaving a genomic scar in place of the capsule locus containing the I-SceI cut site.
In parallel, a gap-repair mediated cloning approach is used to clone the capsule locus of the focus strain. This is done by electroporation of a linear vector bordered with the capsule locus 5' and 3' regions (facing inward) in *K. pneumoniae* expressing the Lambda Red recombinase. Colonies which

have circularized the linear vector by recombining the chromosomal locus are KanR. This vector is called pKAPTURE and also contains an I-SceI cut site.
Once a Δcapsule and pKAPTURE are generated, a serotype swap can be performed by electroporating the pKAPTURE in the Δcapsule and recovering capsulated colonies. Scarless integration is then performed by inducing the Lambda Red recombinase and the I-SceI restriction enzyme, which linearize the pKAPTURE vector in the cell, and induce a chromosomal double strand break between the two regions bordering the capsule locus. Surviving cells exhibiting a capsulated phenotype without selection on kanamycin can be checked to confirm successful integration. The integration is scarless because recombination happens at the borders of the capsule locus, which are >99% identical on average.

This new protocol has already been used successfully by other members of the lab and is being used in other capsule-related research project.

Scarless Serotype Swap protocol in *Klebsiella pneumoniae*

1 – Considerations

The aim of this protocol is to generate isogenic mutants encoding and expressing different capsule serotypes. The capsule locus of *Klebsiella pneumoniae* is defined as the minimal sequence leading to the expression of a functional and typeable capsule. It is typically defined as the genomic region located within the two core genes: 5'-*galF* and *ugd-3'*. On the 3' side of *ugd* is located a *gmd/wbgU* gene (transcribed in the opposite direction), and the O-antigen locus, which typically starts with the *tagG* gene (Figure 21). Any other gene present after *ugd* (especially if they are in the same transcription direction) could be involved in the capsule synthesis. This is referred to as the "the border dilemma" (Figure 22).

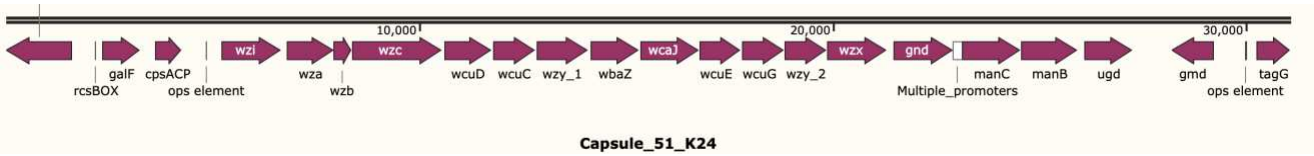


Figure 21 – Capsule locus of strain #51 of serotype K24. The capsule locus starts just before the rcsBOX (RcsAB binding site) and ends right after ugd. This typical organization is easy to exchange by homologous recombination: the borders are clearly galF and ugd.

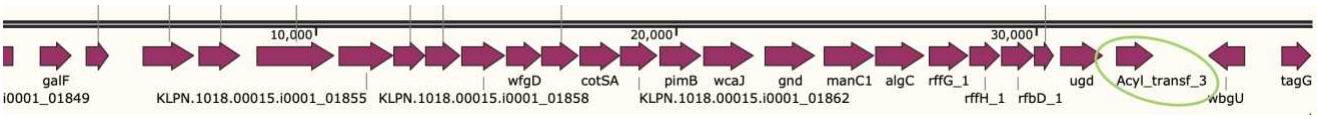


Figure 22 – The border dilemma of the K64 capsule. What is this Acyltransferase gene (green circle) and is it involved in capsule production and determining the K64 serotype?

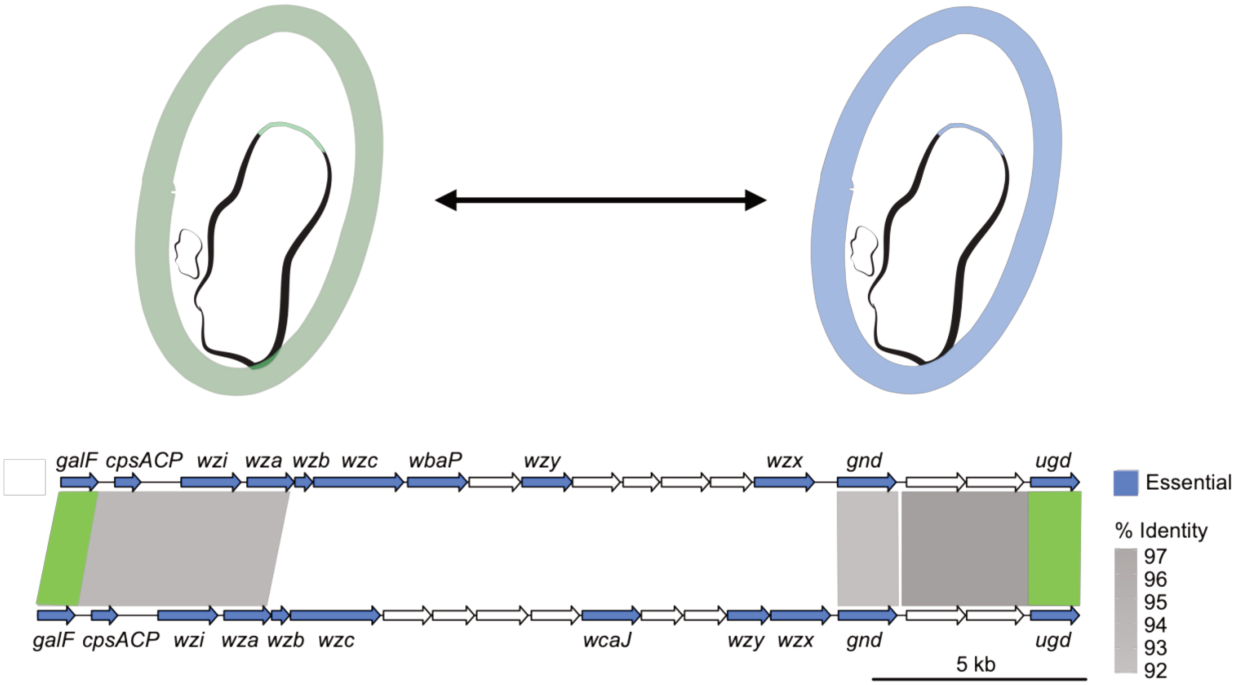


Figure 23 – Overview of a swap involving two strains with two different serotypes. Capsules are thick colored edges around cells. Two capsule loci are displayed on the bottom, with regions of high genetic similarity (homologous recombination-prone). The two genes highlighted with green tracks will undergo the cross-overs leading to the swap (>95% identity).

2 – Protocol

A - Material needed

Reagents:

- Hot start Q5 polymerase (NEB # M0493S)
- Electroporation cuvettes

Strains:

- Strain #751: *K. pneumoniae* BJ1 ΔK2::KmFRT_I-SceI
  source of capsule deletion cassette (KanR)
- Strain #745: *K. pneumoniae* BJ1 ΔK2:: I-SceI pKAPTURE
  source of pKAPTURE cassette (KanR)
- Strain #79: *E. coli* JM109 pKOBEG199 [333]
  source of pKOBEG199 (TcR)

- Strain #672: *E. coli* DH5alpha pTKRED [334]
  source of pTKRED (SpecR, Thermosensitive must be cultured at 30deg)
- Strain #543: *E. coli* DH5alpha pMPIII
  source of pMPIII (SpecR, Thermosensitive must be cultured at 30deg)

NB: It is very useful to generate large stocks of pKOBEG199/pTKRED/pMPIII. For pKOBEG199 and pTKRED, adding glucose to the culture can get higher yields and lower any counterselection from leaky expression of their enzymes.

Primers needed:

- Deletion cassette:
  #585 TGCCGGATATCATCCTTGACGG
  #590 AGGTTGTCGTACAGCGCAC
  Expected size 2.5kb – 68deg annealing, 1min30 elongation
- pKAPTURE cassette:
  #555 CTTCGAGGAGTGCGTCACC
  #560 CGTATTGTCATCGGTGAGCG Expected size 4kb – 68deg annealing, 2min elongation
- Verification deletion:
  #88 CCACAAAGGCAATTCCAAAG
  #489 TCTTCCGCCATACGGTT
  Expected size 3.5kb– 62deg annealing, 2min elongation
- Verification excision:
  #88 CCACAAAGGCAATTCCAAAG
  #489 TCTTCCGCCATACGGTT
  Expected size 2kb– 62deg annealing, 2min elongation

B – Protocol overview

In a serotype swap, the functionality of the cloned capsule locus is essential since it seems to be under high levels of purifying selection. An important control to test this, is to perform a complementation assay. Here is one way:

If:

- strain **A** encodes capsule locus KL**A** of serotype K**A** displays capsulated phenotype.
- strain ΔCapsule **A** encodes no capsule locus (genetic deletion) displays a non-capsulated phenotype.
- pKAPTURE_**A** encodes capsule locus KL**A** from strain **A**

then:

The introduction (and integration) of pKAPTURE_A in strain ΔCapsule **A** should lead to a capsulated phenotype. This is illustrated on figure 24. The **pK** strain (*trans* complemented) is the ideal stock strain to store a functional pKAPTURE since it is already in an adapted background and it is easy to pick capsulated colonies containing functional pKAPTURE vectors.

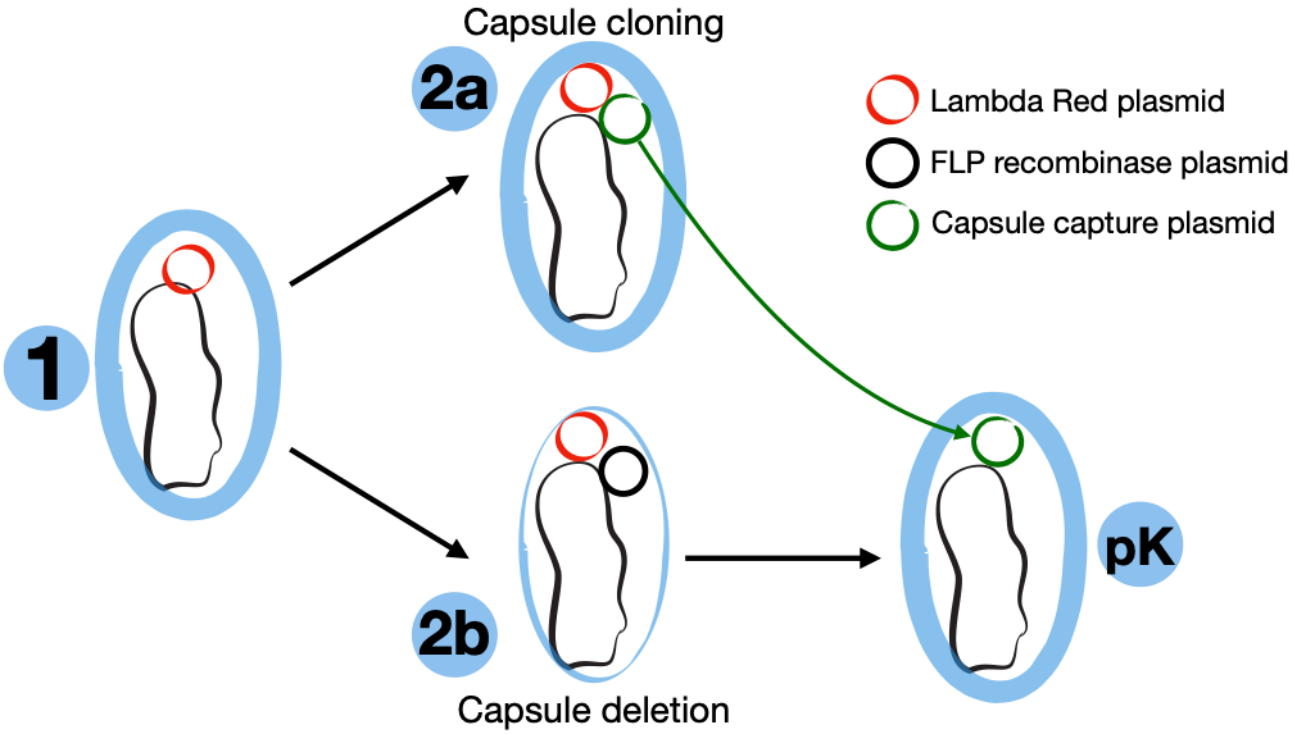


Figure 24 – Capsule complementation with pKAPTURE. Thick blue edges represent the capsule, thin blue edges represent the absence of capsule.

A serotype swap involving two strains can directly follow the *trans* complementation experiment. In this context, another strain with another capsule locus will be subject to capsule deletion and pKAPTURE integration. This is summarized on figure 25.

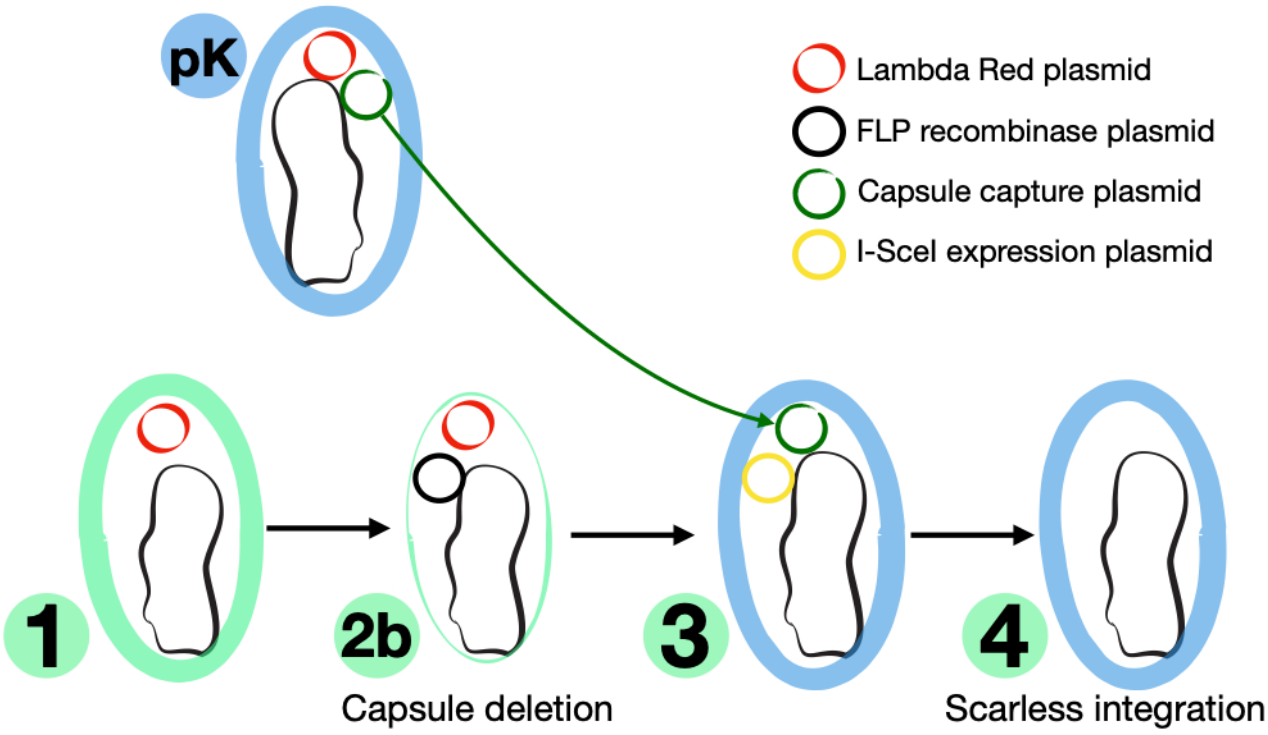


Figure 25 – Overview of a serotype swap between a donor (blue) and recipient (originally green)

I - Generating capsule deletion mutants

1- Electroporate pKOBEG199 into your focus strain

(See preparation of electrocompetent cells protocol)

(See also *https://international.neb.com/protocols/2012/06/21/making-your-own-electrocompetent-cells*)

- It is not necessary to prepare the cells on ice.
- Add 0.2% Glucose to the recovery medium after electroporation
- pKOBEG199 harboring cells must be plated on LB-Tetracycline(15ug/mL)-Glucose(0.2%) plates to avoid leaky expression of Lambda Red enzymes

2- Electroporate the deletion cassette into your pKOBEG199-harboring strain

a. Prepare some deletion cassette by PCR (See Primers/Strains needed)
b. Run a gel to verify if the expected product was amplified
c. Use a PCR purification kit or drop-dialysis to remove salts from the PCR reaction

d. Prepare electrocompetent cells after inducing the lambda red enzymes present on pKOBEG199

1. Overnight your strain in LB+Tetracycline(15ug/mL)+Glucose(0.2%) (37°c)
2. Overday your strain in LB+Tetracycline(15ug/mL)+EDTA(7uM) until it reaches an OD of 0.5 (37°c)
3. At OD=0.5 add 0.2% L-Arabinose to induce Lambda Red **for 30min** (37°c)
4. Prepare the electrocompetent cells **on ice**
5. Electroporate 1uL of salt-free deletion cassette
6. Recover in 1mL LB at 37deg with shaking in culture tube for 1h
7. Plate 100uL of the outgrowth on LB+Kana (Add glucose if you are worried about leaky Lambda red expression)
8. Identify non-capsulated colonies, re-streak on LB+Kana to purify and and perform the verification of deletion PCR.

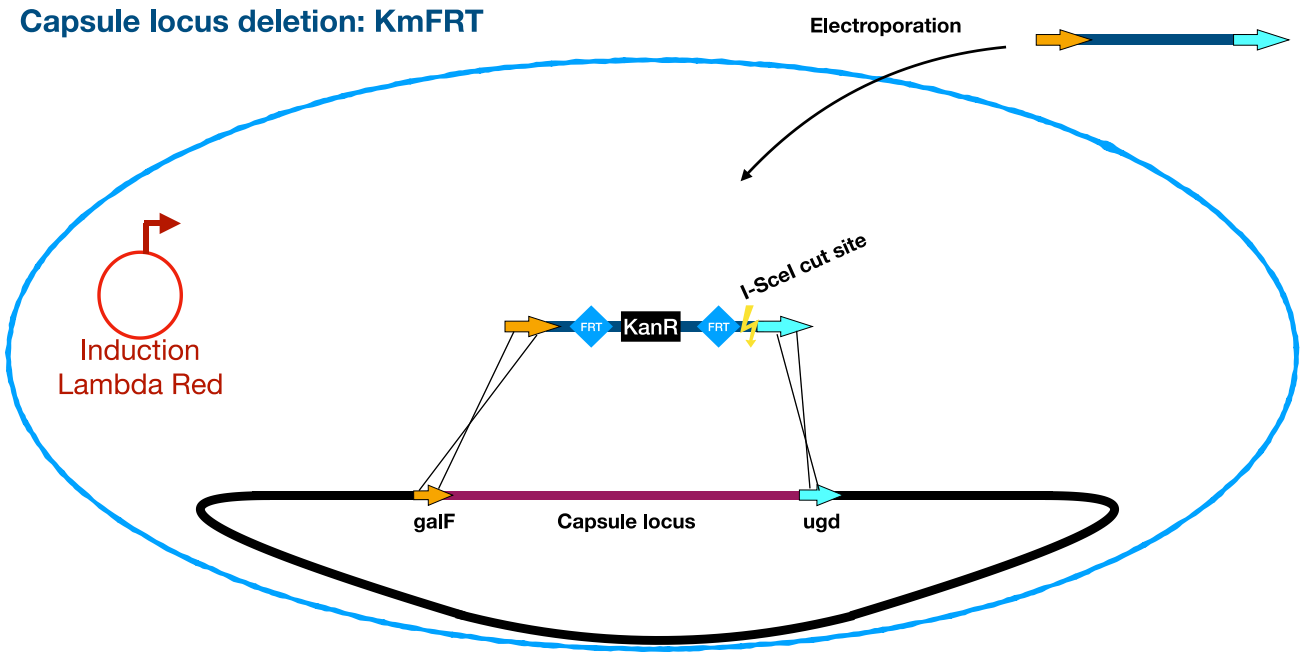


Figure 26 – Deletion of the capsule locus

3- Excise the KanMX resistance gene by expressing FLP:

1. Prepare electrocompetent cells of your non-capsulated cells (No need to do it on ice)
2. Electroporate 1uL of pMPIII miniprep
3. Recover in LB at 30°c for 1h30
4. Plate on LB+Spectinomycine (50ug/mL)
5. Incubate at 30°c
6. Pick 3 colonies and culture them overnight at 42°c to cure the pMPIII plasmid
7. Spread 100uL of 10^{-5} and 10^{-6} serial dilution and incubate at 37°c
8. Pick ~10 colonies and streak them in parallel on LB / LB+Kanamycin / LB+Tetracyclin / LB + Spectinomycin.
9. Identify a clone that only grows on LB and perform the verification of excision PCR. If it's good, this is your ΔCapsule mutant, congrats!

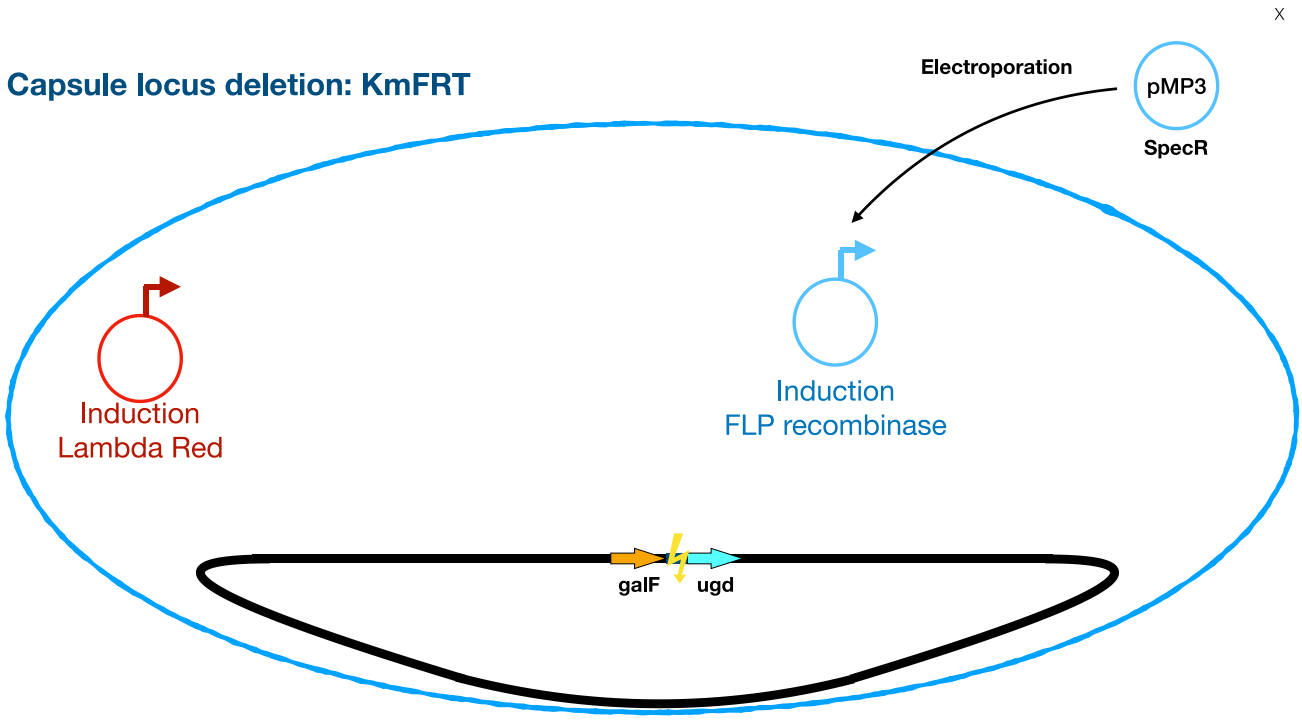


Figure 27 – Excision of the kanMX-FRT marker

II - Generating pKAPTURE vectors

1- Electroporate pKOBEG199 into your strain carrying the serotype of interest. (See step I - 1)

2- Electroporate the pKAPTURE into your pKOBEG199 strain:
    a. Prepare some pKAPTURE cassette by PCR (See Primers/Strains needed)
    b. Run a gel to verify if the expected product was amplified
    c. Use a PCR purification kit or drop-dialysis to remove salts from the PCR reaction
    d. Prepare electrocompetent cells after inducing the lambda red enzymes present on pKOBEG199:

1- Overnight your strain in LB+Tetracycline(15ug/mL)+Glucose(0.2%) (37°c)
2- Overday your strain in LB+Tetracycline(15ug/mL)+EDTA(7uM) until it reaches an OD of 0.2 (37°c)
3- At OD=0.2 add 0.2% L-Arabinose to induce Lambda Red **for 2 hours** (37°c)
4- Prepare the electrocompetent cells **on ice**
5- Electroporate 1uL of salt-free pKAPTURE cassette
6- Recover in 1mL LB at 37deg with shaking, in culture tube for 1h
7- Plate 100uL of the outgrowth on LB+Kana
8- Pick a few capsulated colonies, re-streak on LB+Kana and in parallel start independent cultures with each in LB+Kana+EDTA (7uM) overnight at 30°c to avoid capsule overproduction. Keep track of who is who.
9- Perform a miniprep for each culture. Elute in ddH2O and drop-dialysis all 30uL of the miniprep.
10- Electroporate this miniprep into your ΔCapsule mutant (Same strain as the pKATURE strain to do the complementation)

Add all 30uL of miniprep by using it to dilute the cells at the last step of electrocompetent cells preparation.

11- Recover in 1mL LB (Shaking, 37°c) and plate 100uL of the outgrowth.
12- Plate on LB+Kana.
13- Identify capsulated colonies and re-streak in parallel on LB / LB+Kana.
14- Isolate one colony that is: non-capsulated on LB and capsulated on LB+Kana. This is now your source of pKAPTURE, stock it. Well done!

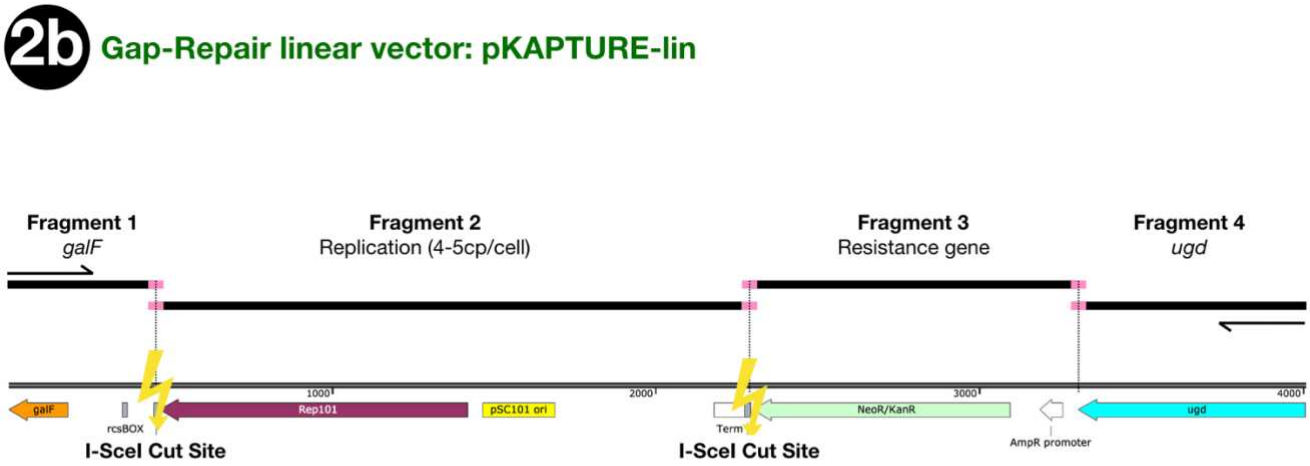
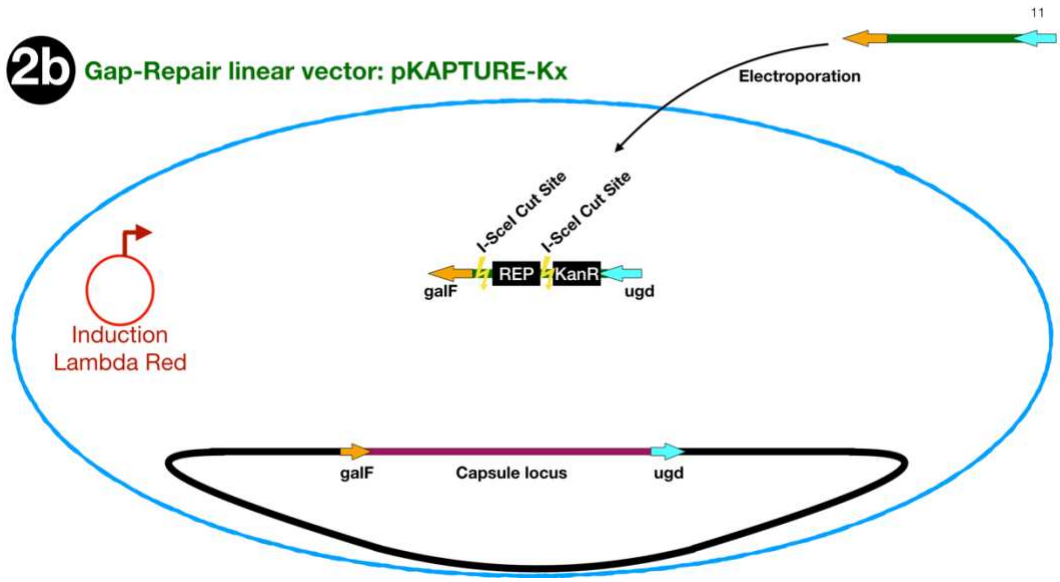


Figure 28 – Capture cassette construction and organization.



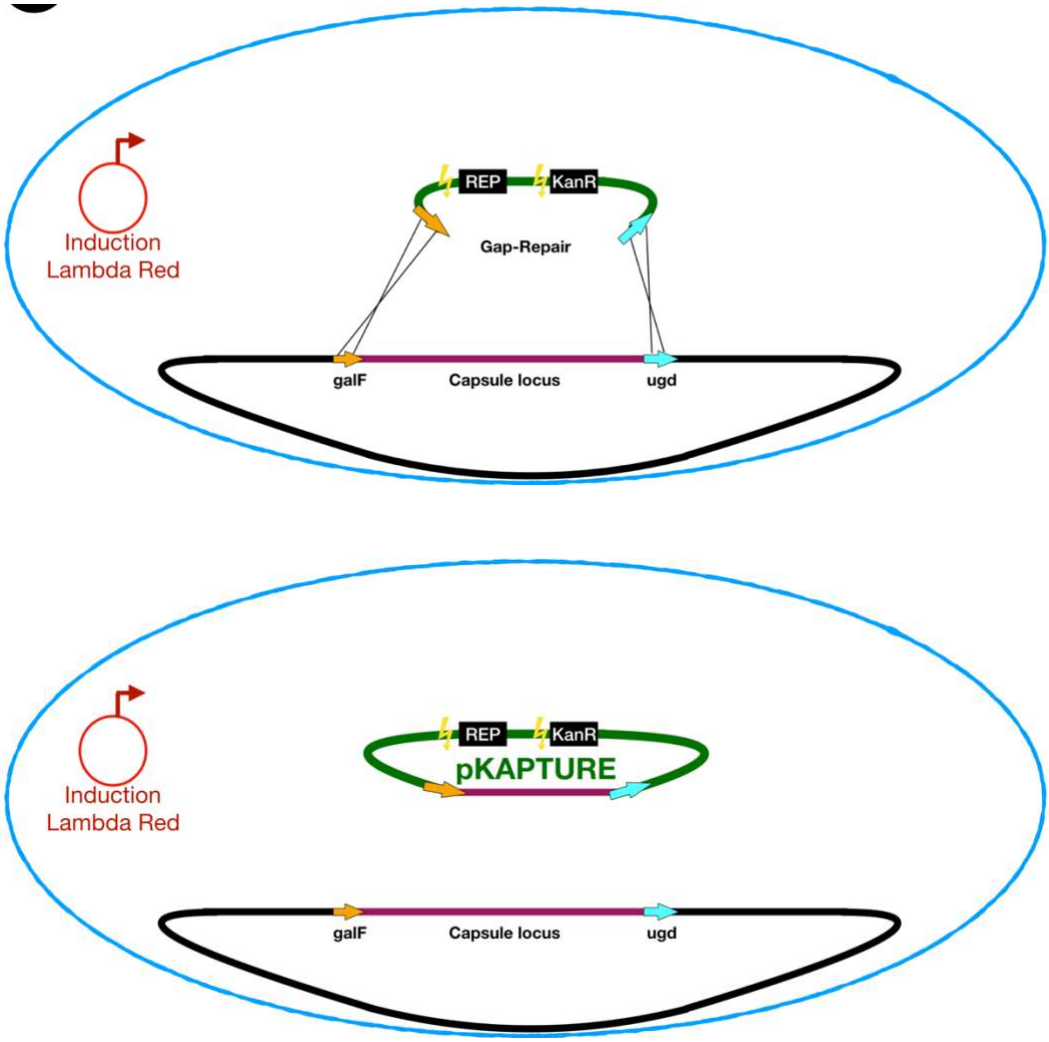


Figure 29 – Capsule locus cloning on pKAPTURE via Lambda Red induction and gap-repair recombination.

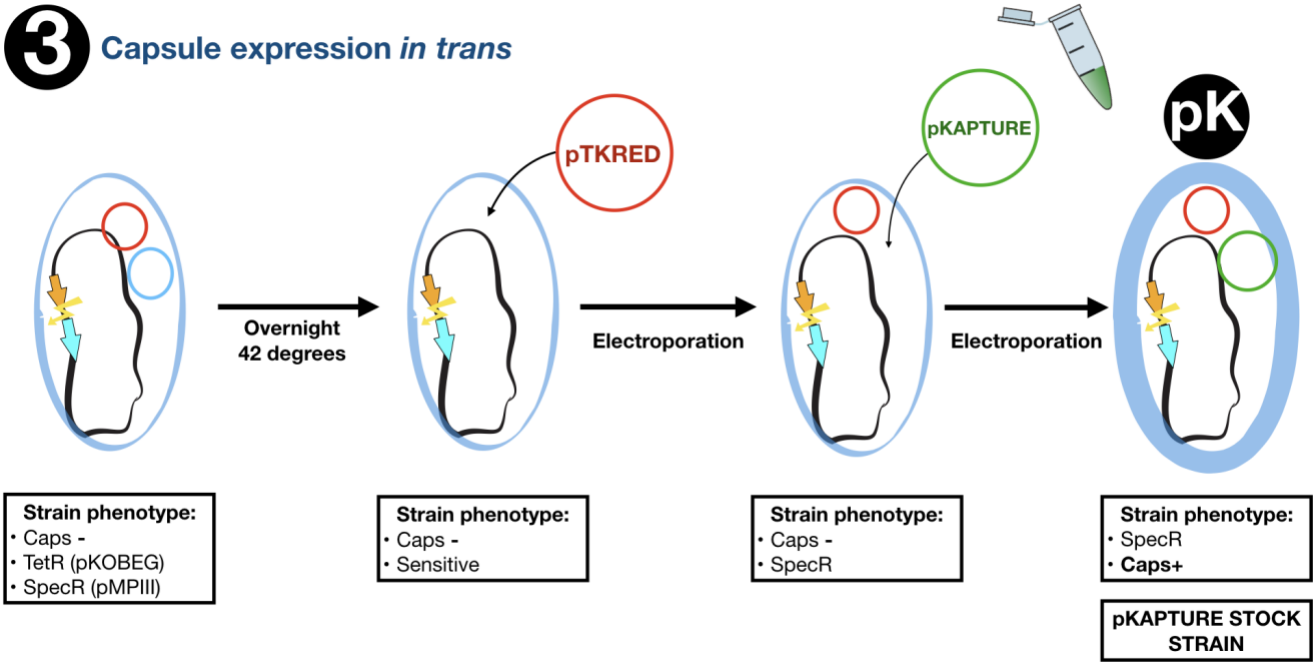


Figure 30 – Overview of capsule deletion and expression *in trans*.

III - Integrating capsule locus from pKAPTURE into Δcapsule mutants

1- Prepare a miniprep of your pKAPTURE from your stock strain (Step D.13)

Elute in ddH2O and drop-dialysis the resulting 30uL

2- Prepare electrocompetent cells of your Δcapsule mutants (No need to do it on ice)
3- Electroporate 30uL of pKAPTURE miniprep by diluting the competent cells in it
4- Recover in 1mL LB at 37°c for about 1h (Shaking, culture tube)
5- Plate on LB+Kana and grow overnight at 37°c
6- Pick a capsulated colony, parallel streak on LB / LB+Kana
(Must be non-caps on LB and capsulated on LB+Kana)
7- Inoculate a colony from LB+Kana into LB+Kana+EDTA (15mL) in the morning
8- Prepare electrocompetent cells once the culture reach OD~0.7 (no need to do it on ice)
9- Electroporate pTKRED
10- Recover at 30°c in 1mL LB+Kana+0.2%Glucose for 1h30 (Shaking, culture tube)
11- Plate on LB+Kana+Spectinomycine(50ug/mL)+Glucose(0.2%) and incubate at RT or 30°c
12- Pick several colonies and resuspend in 5mL M63b1 +Spectinomycine(100ug/mL) +0.2%L-Arabinose +0.2%Glyrerol
13- Cultivate at 30degrees (Shaking)

You can plate at the end of the day and the next day on LB (+Glucose0.2%) at 42°c to cure pTKRED. (Adapt the dilution factor before plating according to the turbidity of the culture)

14- Identify capsulated colonies and parallel streak on LB / LB+Spec / LB+Kana

15- Identify a clone that only grows on LB

16- Use primers specific to your new capsule locus to verify the strain.

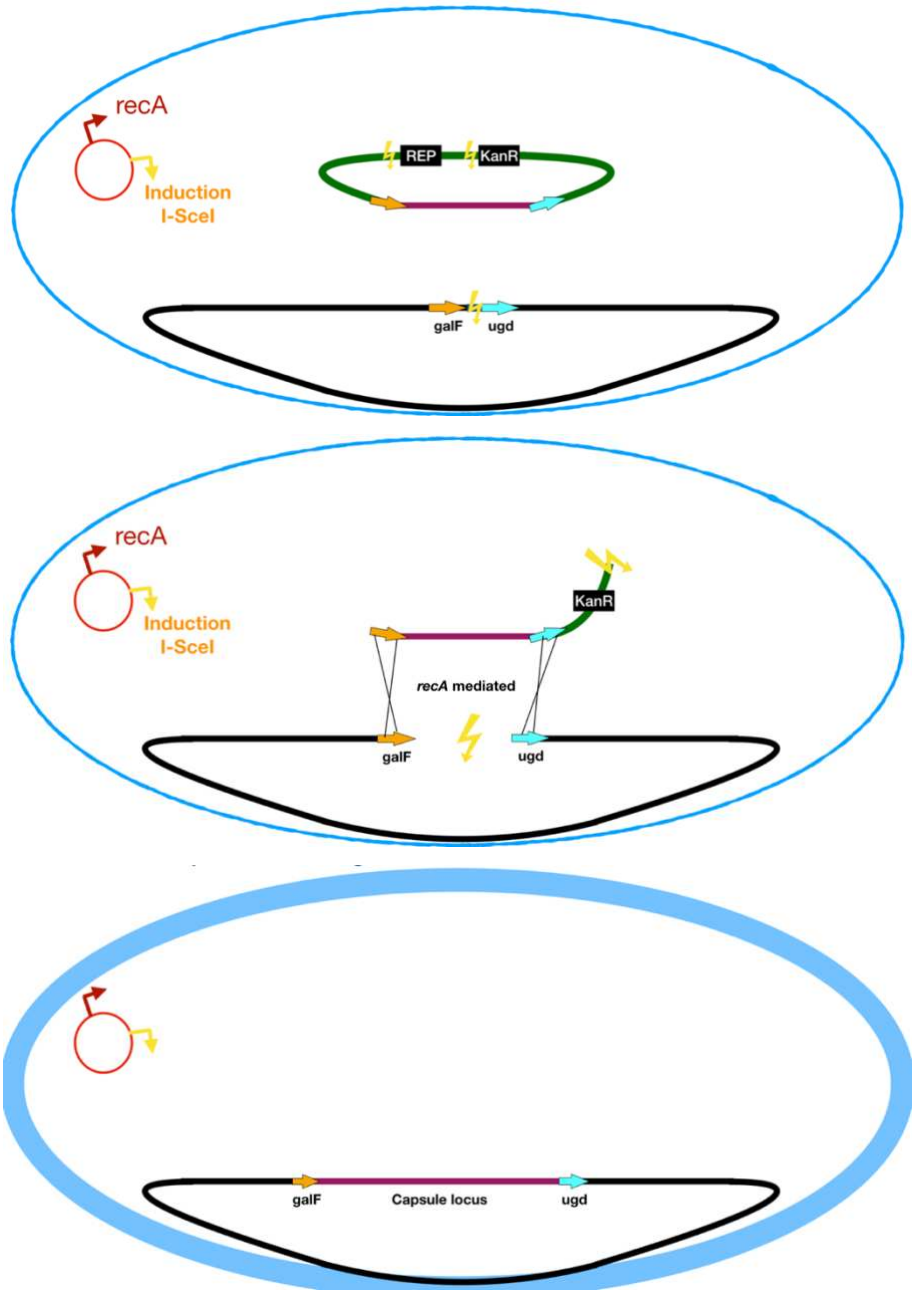


Figure 31 – Scarless integration via RecA-mediated homologous recombination following pKAPTURE linearization and capsule locus double strand break.

Research article: Capsule serotypes result in distinct phage infection patterns and frequency of plasmid conjugation.

Matthieu Haudiquet*, Olaya Rendueles, Eduardo Rocha

In preparation for submission.

This is an early version of the manuscript.

Title:

Capsule serotypes of *Klebsiella pneumoniae* determine phage sensitivity and plasmid conjugation efficiency.

Authors

Matthieu Haudiquet, Olaya Rendueles, Eduardo Rocha

[To be defined]

Preliminary abstract:

Klebsiella pneumoniae is a hub for mobile genetic elements carrying antibiotic resistance across enterobacteria. This propensity is modulated by many intra- and extracellular factors, such as the bacterial capsule, a ubiquitous feature of ESKAPE pathogens, which is both a virulence factor and a gatekeeper for horizontal gene transfer (HGT). Yet, the diversity of genetic backgrounds, mobile elements, and defense systems have hampered a rigorous test of the impact of capsules in phage and plasmid transfer. To characterize the interaction between mobile genetic elements and capsule composition, we combined comparative genomics with *in vitro* gene transfer experiments between serotype swap mutants. We confirm that capsule serotype shapes phage host range, and thus virion-mediated gene transfer. The exact switch of host range following capsule swaps, suggests that many phages may only require a capsule swap to lose or gain the ability to infect. On the other hand, conjugation efficiency is quantitatively affected by the capsule and its serotypes, which act as a general defense mechanism against conjugative plasmids. These results suggest that cell envelope composition and spatial hindrance have a major impact on intercellular interactions leading to HGT. Overall, capsules' interaction with mobile genetic elements results in a complex interplay between virulence, phage sensitivity and conjugation rates influencing the evolution of *Klebsiella pneumoniae*.

Introduction

Bacterial cells possess a dynamic cellular envelope armored against both biotic and abiotic stressors [1,2]. The envelope is often covered by a capsule composed of thick, membrane-bound polysaccharide polymers that forms the outermost layer of the cell[3]. Capsules protect cells against desiccation [4], bacteriophage predation[5,6] and protozoan grazing[7], as well as factors associated with host-pathogen interactions like antimicrobial peptides[8] and macrophages[9,10]. Hence, capsules are important colonization factors and sometimes also virulence factors. The most frequent capsules are of Group I capsules, also known as Wzx/Wzy-dependent capsules [3], and share common assembly pathways but highly diverse repeat units, called K antigen [11]. Polymerization of the K antigen and subsequent attachment on the cell surface leads to the expression of distinct capsule serotypes. These capsules are encoded in capsule loci with a genetic diversity corresponding to their chemical diversity[12,13]. It is not happenstance that capsules are so diverse, it is rather the result of strong selective pressure(s) driving the diversification and fixation dynamics of capsule loci[14]. Capsule genes are exchanged via homologous recombination with foreign DNA brought by horizontal gene transfer (HGT). Serotype swaps are frequent and can occur either by intra-locus recombination [15], or whole-loci exchanges by HGT [16,17].

If the capsule varies by HGT, it also affects the rates of HGT [16,18], which a key driver of bacterial evolution. MGEs are the most frequent, and often only, vectors of DNA transfer among bacteria. They can be found in almost all bacterial species[19]. DNA transfer by MGEs can occurs by two distinct mechanisms. Conjugation involves the transfer of ssDNA from one cell to another through a mating pair formation (MPF) apparatus with a type 4 secretion system that has been classed in 8 different families based on their phylogeny and gene composition [20]. Conjugation requires direct contact between the cells and can transfer very large amounts of genetic material in one single event [21]. Phages can transfer DNA under two major types of mechanism. Temperate phages can integrate the host genome and lead to lysogenic conversion. Alternatively, some phages can transduce bacterial DNA by a variety of mechanisms between cells [22]. Both processes require the access of the outer membrane to work, but the biochemical mechanisms are very different and may be affected differently by the existence of a capsule. Phages require cell receptors for adsorption and subsequent infection. There is extensive evidence that phages of capsulated bacteria are either blocked by the capsule [5], or dependent on the presence of a specific serotype for infection[23,24]. The latter encode serotype-specific capsule depolymerase enzymes to pass through the capsule barrier [25–27]. As a result, phages tend to drive genetic exchanges between strains of the same serotype[16]. On the other hand, MPF systems are specialized type IV secretion systems which fall into eight distinct types, based on their gene content and evolutive history, called Mpf types[20]. The three most prevalent Mpf types of Proteobacteria plasmids are the Mpf_F named from the F-plasmid, the Mpf_T from the Ti plasmid and the Mpf_I from the IncI R64 plasmid[28]. Little is known on the impact of the molecular differences between conjugation systems on plasmid transfer. To date, no specific receptor has been described for Mpf systems[29], but specific type 4 pili adhesins encoded in Mpf_I plasmids have been associated to increased conjugation efficiency toward certain recipients[30]. It was proposed, but not tested, that conjugation might be more frequent between bacteria with similar capsule serotypes [31]. We have recently shown that *Klebsiella pneumoniae* acquires one engineered plasmid from *E. coli* at higher frequency when lacking the capsule (REF). Recently, we have shown that the bacterial capsule can decrease the reception efficiency from an engineered donor-plasmid system in single gene mutants of *Klebsiella pneumoniae,* and that plasmids seem to drive genetic exchanges between different serotypes[16]. Hence, the interaction between virions, Mpf systems and the capsule impact the evolvability and evolution of capsules, which diversify via HGT.

HGT also supports the rapid evolution of pathogens, and most notably the acquisition of antibiotic resistance genes (ARG), endangering our healthcare systems[32]. Among those pathogens, the alarming multi-resistant nosocomial group ESKAPE, comprised of *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter spp.* represents models for the study of horizontal gene transfer[33]. Moreover, all the ESKAPEs encode capsular loci, and their capsule diversity can reach more than a hundred different serotypes[14]. Among those, *K. pneumoniae* encodes a thick capsule with diverse chemical composition, encompassing more than 130 predicted serotypes that follows the population structure but are frequently exchanged via serotype swap[13,34,35]. Indeed, *K. pneumoniae* population is structured between clones often associated with either hypervirulence or multi-drug resistance[36], and those traits are associated with specific serotypes. For example, hypervirulent clones of *K. pneumoniae* harbour almost exclusively the K1 or K2 serotypes[37], but rarely encode ARG[38]. Understanding the impact of capsule serotype on specific traits, however, requires to build complex isogenic mutants, *i.e.* precisely exchanging ~30kb

capsule loci between strains. So-called swap mutants have been generated in *Streptococcus*, a naturally competent genus with high transformation rates, and used to show that capsule serotypes are associated with different virulence levels[39–41]. In *K. pneumoniae*, which is not naturally competent, most studies on the capsule focus on single gene mutants, which have been generated to investigate the link between the capsule and virulence[42]. Moreover, the construction of mosaic strains harbouring, *inter alia*, different capsule loci showed that some hybrids expressed a new capsule serotype and had altered macrophage sensitivity[43].

Many mysteries remain regarding the factors modulating MGE infection in bacterial populations, especially regarding the involvement of the cell envelope, and thus of the capsule. A growing body of evidence places the capsule as an important gatekeeper of horizontal gene transfer[5,16,23,44], which may have shaped the negative correlation between virulence and resistance in K. pneumoniae. Here, we wanted to know to what extent the capsule serotype impacts horizontal gene transfer, especially of antibiotics resistance genes via conjugation, and how the interaction between capsules and MGEs can lead to distinct evolutionary paths. To do so, we built a set of isogenic capsule serotype swaps in K. pneumoniae, encoding and expressing different serotypes by leveraging the recombineering toolbox. Using previously identified phages, we first assessed whether phage sensitivity of the swaps was altered. We then measured their conjugation propensity with an array of clinical plasmids, estimating the conjugation efficiency of 468 Donor-Plasmid-Recipient (DpR) groups. We show that this impact correlates with the Mpf type of plasmids, which could explain why MpfF plasmids have been so successful in invading K. pneumoniae populations. Our results highlight slow and fast lanes of ARG spread, and contribute to explain the distinct evolutionary paths of hypervirulent strains of K. pneumoniae.

Results

Capsule loci exchange leads to functional capsules

To test the impact of the serotype on conjugation, we first sought to produce capsule swaps in three diverse strains to control for the capsule's host genetic background. We selected three isolates with relevant features to our study, namely they harboured a chemically characterised capsule serotype, and harboured different types and quantity of native MGEs.

We first attempted to swap capsule loci with a standard two-step allelic exchange method. For this, we relied on the highly similar bordering genes, *galF* and *ugd*, as the homology arms surrounding the loci, and cloned the K1 and K2 capsule locus via Gibson assembly into DH5alpha *E. coli* cells. We then conjugated the vectors into *K. pneumoniae* via the MFD *E. coli* strain, and selected for the first and second cross-over leading to capsule locus replacement. However, all the colonies with a positive PCR specific of the new capsule locus appeared non-capsulated on LB plates. Upon whole genome sequencing (WGS) of one BJ1::K1 and one NTUH-K2044::K2 clone, we validated that the new capsule locus replaced original one in its native locus and no other mutations were found outside the capsule locus. We identified identical insertion sequences in the two swaps, in different places, that resulted in non-functional capsules (Fig. 1A). This IS was acquired from the genome of DH5alpha, suggesting rapid inactivation of the capsule loci when expressed in the *E. coli* strain.

Hence, we sought to develop a protocol to clone and swap capsule loci without the need for intermediate laboratory strains. We used a Lambda Red gap-repair mediated cloning method to capture the different capsule loci directly within *K. pneumoniae* with a newly designed capture cassette including an I-SceI cut site outside the Km-FRT marker (See methods). We successfully cloned four different capsule loci corresponding to K1, K2, K3 and K24, which we then inserted in place of the previous capsule locus of our three selected strains. To do so, we generated full capsule mutants (Δcps) but leaving the two border genes *galF* and *ugd* intact. We then electroporated the cloned capsule vectors in the three strains and inserted the new capsule loci in its native position via a double-strand break and repair insertion protocol (Fig. 1B). We replaced the capsule loci of our three Δcps mutants with four new serotypes, including their own as complemented controls, resulting in a total of 12 swaps. WGS of capsule swaps showed that the different capsules had been integrated as expected, and revealed only one secondary mutation in NTUH-K2044::K24. We confirmed visually the presence of a capsule in all strains by optical microscopy (Fig 1C). We concluded that our method of capsule swap worked robustly, and that the replacement of whole capsule loci leads to the expression of a functional capsule.

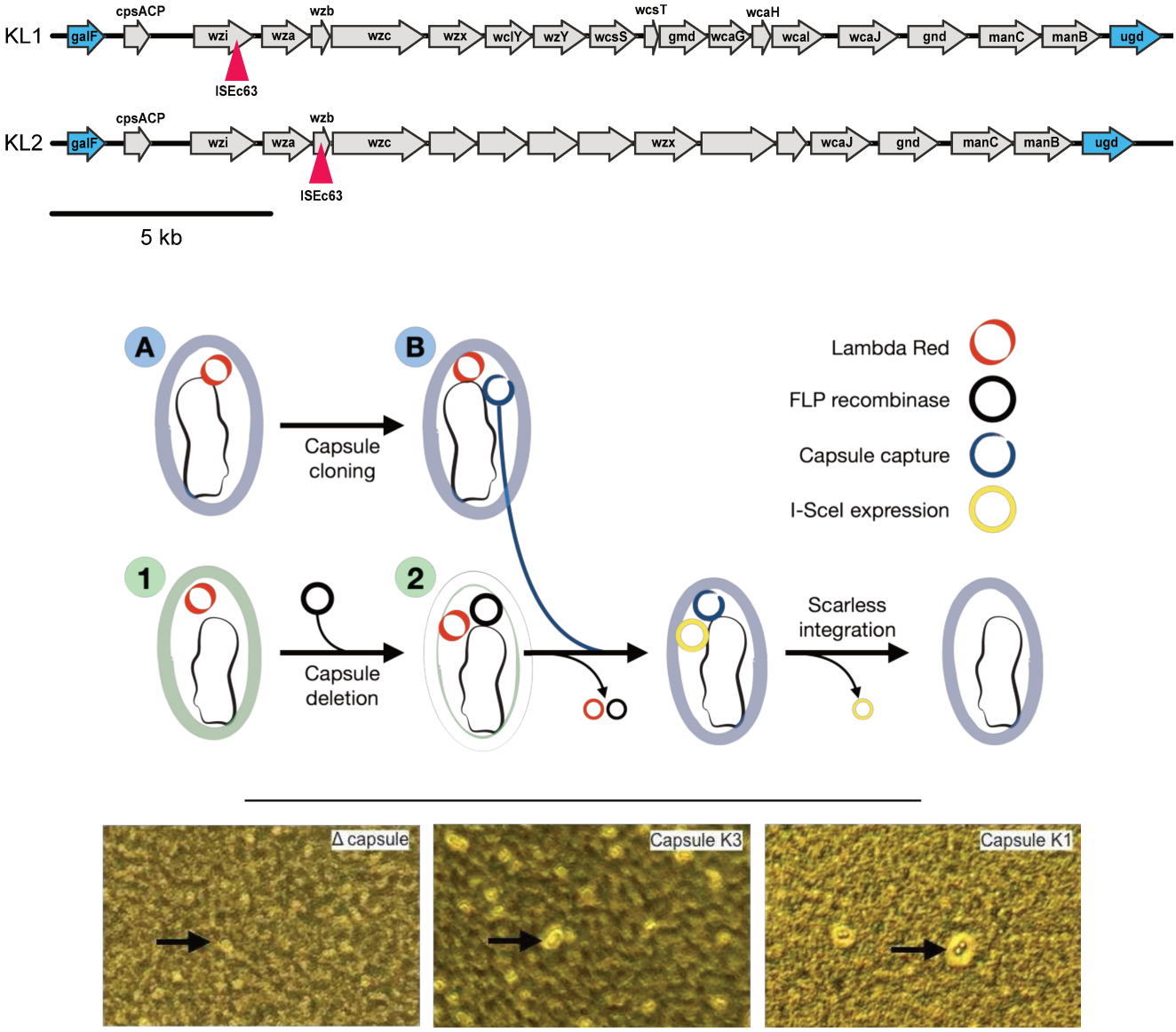


Figure 1 – Scarless Artificial Serotype Swap. A. Capsule loci corresponding to K1 and K2. Red triangles correspond to IS insertions found in the swaps generated from capsule loci sub cloned into DH5alpha and MFD. **B.** Protocol overview. The colours of the cells' envelope correspond to different capsule serotypes. Black circular line: chromosome. Smaller circles: vectors encoding various functions (see legend). **C.** Capsule visualization of swapped mutants

New serotype acquisition changes phage infection profiles [Not finished]

Capsule serotype is one of the drivers of phage host range in Kpn, leading to elevated within-serotype gene flow. However, it is unclear if capsule-infecting phages rely on secondary receptors, and if the expression of a given capsule locus leads to a given phage sensitivity.

We thus wanted to know if our swapped mutants had altered sensitivity to phages. We prepared lysates of temperate and virulent phages previously described, and tested their ability to lyse our swapped mutants (Figure 2).

[TO BE COMPLETED]

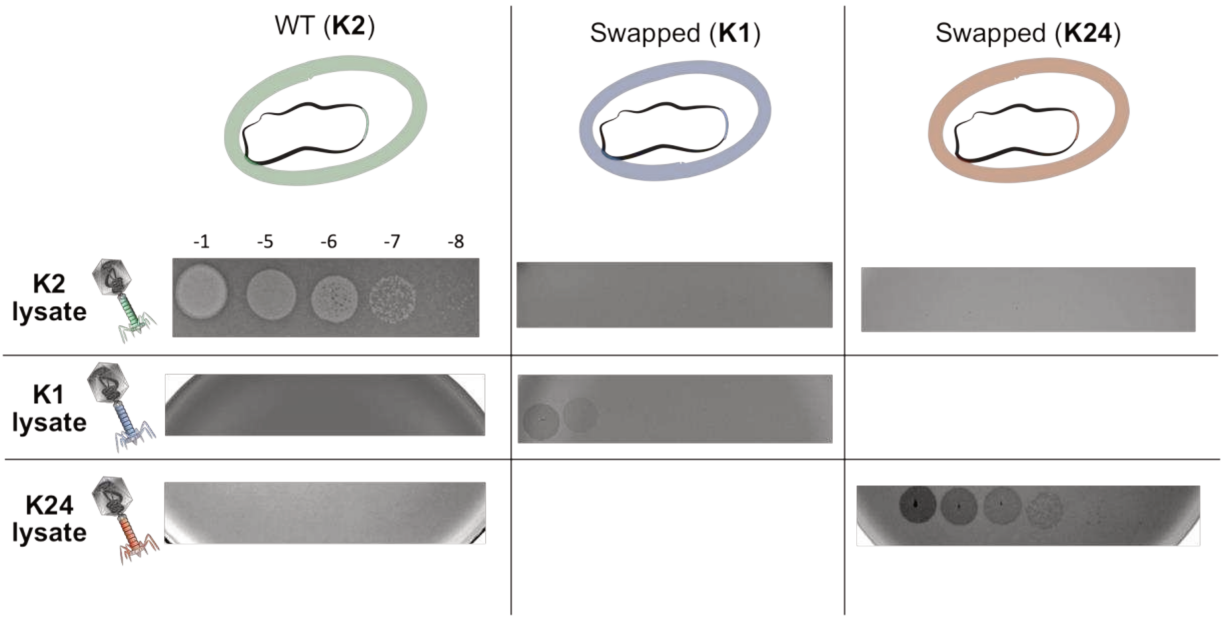


Figure 2 - Phage sensitivity changes upon serotype swap. Pictures of plaque assays with temperate phage lysates against capsule swaps. The dilution factor (log10) is indicated.

These results show that serotype swap sensitizes bacteria to new predator, and hint that the capsule may be the sole and only receptor needed for phage adsorption and entry in the cell. They also validate that the new capsule loci in the swapped mutants lead to the expression of a capsule with a chemical composition similar, if not identical, to their original host, regardless of their new genetic background.

Serotype swaps change conjugation efficiency

We have shown that serotype swap drastically changes the pattern of infection of phages, and thus of phage-mediated HGT. We thus wanted to investigate how capsule serotype impact the efficiency of conjugation-mediated HGT, the main driver of antibiotic resistance gene spread within nosocomial pathogens.

To precisely assess the role of the capsule in conjugation, we collected nine conjugative plasmids (Fig. 3A) with MPF types F, T and I and made conjugation assays from *E. coli* to the 162 *K. pneumoniae* combination of 3 mutants, 9 plasmids, and 6 capsule states (WT, Δcps, K1, K2, K3, K24), all done in three independent replicates (486 essays). We analysed the results in the light of each of these parameters: recipient strain, plasmid Mpf type and capsule serotype.

To **test if the reception efficiency was different between our three diverse strains**, we compared the distribution of transconjugant frequency (Fig 3B) between BJ1, NTUH-K2044, and ST45. We found that all three strains had different reception rates of conjugative plasmid (Paired Wilcoxon, $p<0.001$). Of note, three MPF_T plasmids, $p168C10_T$, $p486_T$ and $p580_{T\&F}$ could not be conjugated into strain ST45 efficiently enough to meet our detection threshold. We hypothesised that this was due to the low reception efficiency of the strain. We concluded that different genetic background are associated with broadly different rates of reception of conjugative plasmids.

We then enquired on the **differences in transconjugant frequency (TF) between plasmids encoding different** types of Mpf systems. Among our set of plasmids, we observed that the two Mpf_F plasmids had lower transfer efficiencies than Mpf_T, which themselves transferred less efficiently than Mpf_I plasmids (Paired Wilcoxon, $p<0.001$). These results suggest that types of Mpf are associated with different plasmid transfer rates.

We then tested **the role of the serotype on the reception efficiency** in *K. pneumoniae*. As expected from our previous study on *ΔwcaJ* single mutants, we found that the median TF of Δcps was 3.5 times higher than WT cells (Paired Wilcoxon test, $p<0.001$). We used $TF_{\Delta cps}$ to normalize the TF of the swaps, to enquire on the differences in reception efficiency between the four serotypes while entirely controlling for the genetic background (Fig 3D). We found that the expression of K1, K2, and K24 significantly decreased the reception efficiency (Paired Wilcoxon test, $p<0.001$), but not K3 ($p=X$). K1 and K2 ranked first in term of reception reduction, then K24 (K1/K2 vs. K24, $p<0.001$) and then K3, which was significantly different from the other three serotypes (all comparisons, $p<0.001$). We concluded that capsule serotypes are associated with different conjugation propensity on the recipient end. However, we observed that the p23 plasmid, belonging to MPF_F, behaved oppositely to other plasmids, including the other MPF_F plasmid pR1-*drd19,* and seemed to be better transferred into capsulated cells. These differences could be explained by the divergence of conjugation operons between the two plasmids, as they only share 51% total nucleic identity. Further, pR1 was not isolated in *K. pneumoniae*. This suggests that the p23 plasmid may be adapted to the presence of the capsule.

Since the recipient serotype and the plasmid Mpf type had an impact on the reception efficiency, we wanted to know if **their interaction** impacted the reception of conjugative plasmids. We analysed the normalized TF between serotypes according to the MPF type of the plasmids. We found that the two F-type plasmids were the least affected by the presence of a capsule independently of the serotype, whereas conjugation by plasmids with MPF_I and MPF_T were strongly inhibited by the capsule and this inhibition was dependant on the serotype (Fig 3E).

Hence, the capsule is most often a barrier to the acquisition of conjugative plasmids but this propensity is dependent on the serotype and the plasmid.

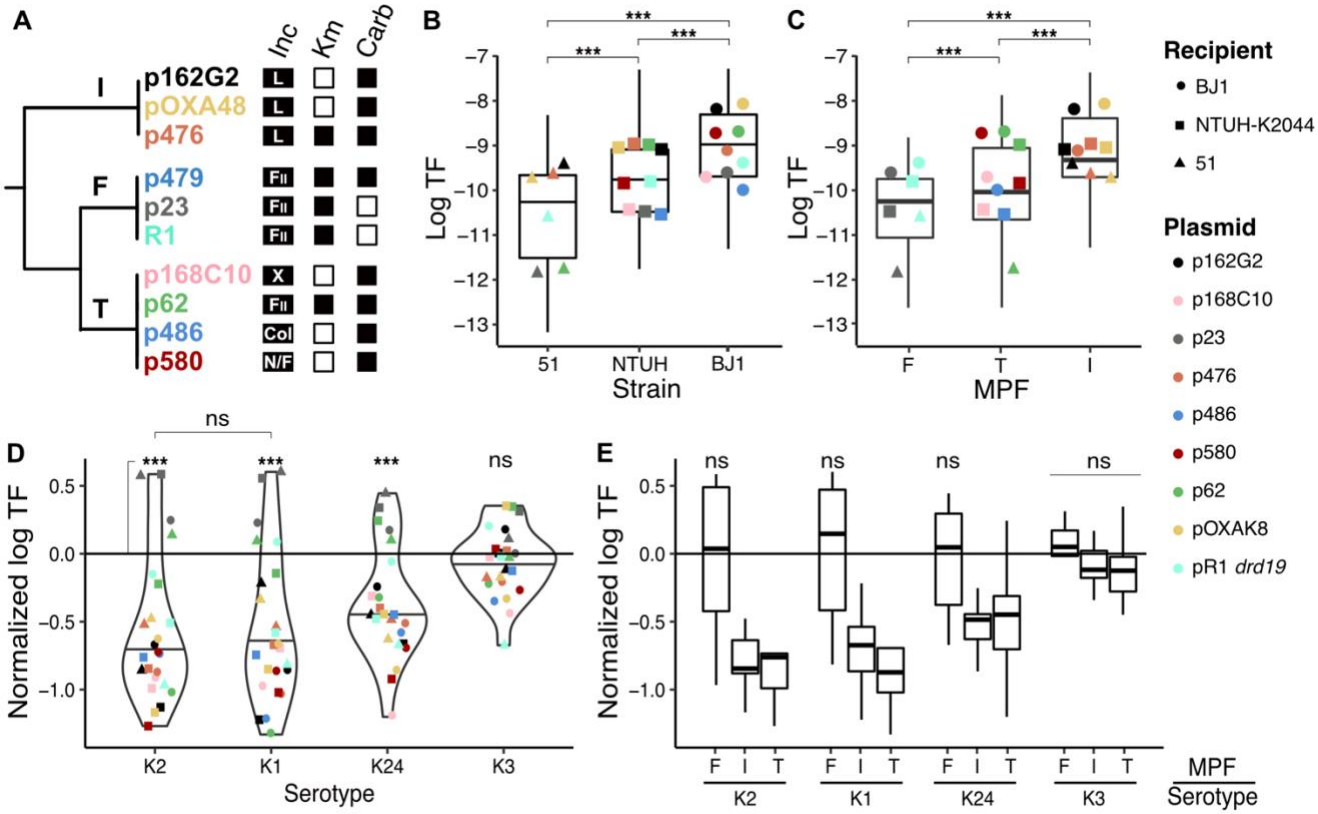


Figure 3 – Transconjugant frequency (TF) from E. coli to K. pneumoniae. TF was assessed for nine plasmids from E. coli DH10B donors into three different K. pneumoniae recipients, each with six capsule states (WT, Δcps, K1, K2, K3, K24) in independent triplicate.A. Plasmids characteristics. Cladogram based on the virb4 protein tree differentiating I, T and F Mpf, from Guglielmini et al., 2013. Plasmids names are coloured according to the legend. The predicted incompatibility group is displayed in the first column, the resistance phenotype for kanamycin (Km) and the carbapenem ertapenem (Carb) in the other columns. Black boxes indicate that the plasmid confers resistance to the corresponding antibiotic. B and C. Transconjugant frequency of the three K. pneumoniae strains (B) and of the three MPF (C) types. The y-axis represents the log10 Transconjugant Frequency (TF). Each dot represents the average TF for each plasmid, whereas the boxplots are drawn from all replicates and all plasmids. The x-axis is split according to the strains. D and E. Reception efficiency according to the capsule serotype (D) and interaction between plasmid Mpf type and capsule serotype (E). The y-axis represents the normalized log10 TF for each Plasmid-Recipient (PR) group, which was normalized by substracting the corresponding log10 TF of the Δcps. Below the solid line at 0, plasmid-recipient (PR) groups have a lower TF than the Δcps. Individual points represent the average of 3 biological replicates, and the horizontal bar represent the median for all PR groups.

Capsule expression and serotype lead to slow and fast lanes of conjugation

We previously showed that the reception of conjugative plasmid is quantitatively impacted by the capsule expression and serotype. However, conjugation is a two-faced coin comprising two processes, donation and reception. Additionally, it has been hypothesized that *Haemophilus influenzae* strains of the same serotype engage more efficiently in conjugation [31] , but our recent comparative genomics analyses of *K. pneumoniae* suggested that there was no preferential gene exchange via conjugation within strains with the same serotype [16]. To shed light on the impact of the capsule on donation and

reception of conjugative plasmids, we set up 972 distinct Donor-Plasmid-Recipient (D_PR) conjugation assays corresponding to three Donor-Recipient pairs, three plasmids ($p23_F$, $p62_T$, $p476_I$), six capsule states in both directions in independent triplicates. (See Experimental design of conjugation assays).

We first tested if plasmid transfer was impacted by the presence of a capsule in the recipient strain when the transfer took place between *K. pneumoniaet*. As expected, we found that non-capsulated recipients (Δcps) had generally higher transconjugant frequencies (Wilcoxon test, $p<0.001$), in a similar manner than what we observed for our *E. coli* to *K. pneumoniae* experiment. Thus, we used the non-capsulated mutant pairs to normalize each D_PR group (Fig 4A), and found that the capsule of the recipient and its serotype significantly impacted the conjugation efficiency.

We then performed the same analysis considering this time the serotype of the donor (Fig 4B). We found that the capsule of the donor also affected the conjugation efficiency, with the same ranks as recipient capsules, from least to most permissive capsules: K1/K2, K24, K3. Further, as observed previously, p23F did not behave the same as p476I and p62T, *i.e.* it did not seem to be negatively affected by the capsule. The capsule and its serotype did not seem to impact the transfer rate of p23F, both from the donor and recipient.

Next we wanted to investigate if capsule serotype played a role in the interaction between donor and recipient cells, and if cells of the same capsule serotype would better exchange plasmids than others. We first compared the TF of DpR couples in which both the donor and recipient had the same serotype against couples in which donor and recipients had different capsule (Fig 4C). We found that none of our nine DpR groups had a significantly different TF between same and different serotypes pairs (Multiple Wilcoxon test, all $p>0.05$). This shows that there is no preferential transfer by conjugation between cells expressing the same capsule (relative to the others).

We further enquired **on the interaction between donors, recipients, and their capsules** by leveraging the TF of Δcps mutants. We build a simple geometric model of interaction t(See methods) using $TF_{\Delta_p \to K}$ and $TF_{K_p \to \Delta}$ to compute an expected $TF_{K_p \to K}$ value (Fig. 4D). The observed to expected TF ratio distribution was not significantly different from 1 (Wilcoxon test, $p>0.05$), indicating that the model fitted well to the observed data, except for a few outliers that seemingly conjugated 10 times more (ratio>10) or 10 times less (ratio<0.1) than expected. This suggest that the impact of the donor and recipient capsules were generally independent. Our results also suggested the existence of preferential routes of conjugation (fig 4E). For example, there were intra-serotypes differences, e.g. K1-K1 transfer is less efficient than K3-K3 transfer (Wilcoxon test, $p=0.0007$). But also inter-serotypes, e.g. the K2-K3 transfer route is more efficient than the K1-K2 route Wilcoxon test, $p=0.004$). Finally, it highlights how Δcps mutants are highly efficient dealers of resistance conjugative plasmids in general (Pairwise wilcoxon test, $p<0.001$).

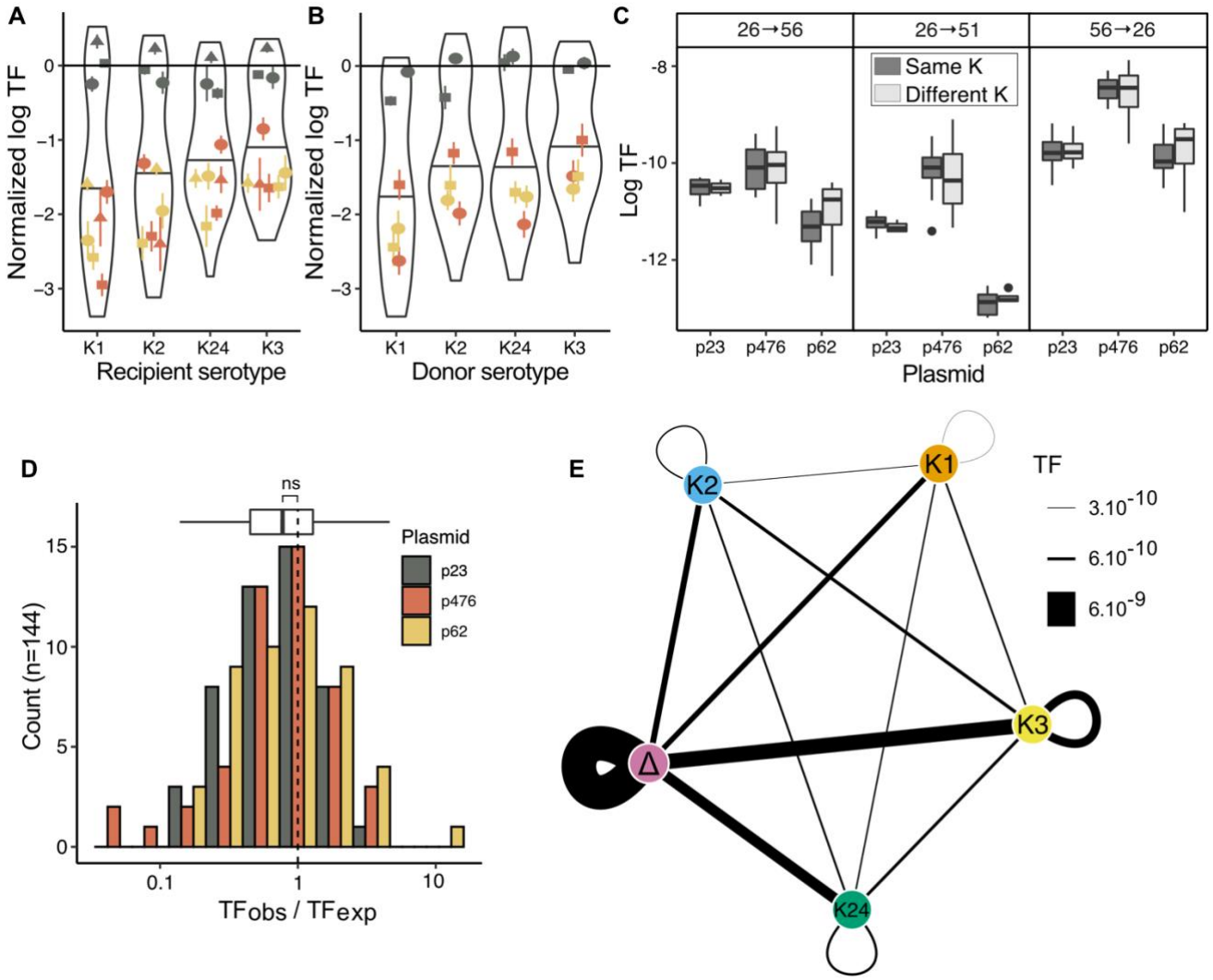


Figure 4 – Transconjugant frequency (TF) of **K. pneumoniae to K. pneumoniae.** TF was measured for three plasmids, three donor-recipient pairs, with each six capsule states (WT, Δcps, K1, K2, K3, K24).

A and **B. Impact of the capsule on plasmid reception (A) and donation (B).** The y-axis represents the log10 Transconjugant Frequency (TF). Each dot represents the average TF for each plasmid, whereas the violin plots are drawn from all replicates and all plasmids. **C.** Log_{10} Transconjugant Frequency (TF) of same or different donor/recipient capsule serotype. The x-axis is split between plasmids, and boxplots are colored according to two conditions: donor and recipients belong to the same serotype ("same K") and donor and recipient belong to different serotypes ("different K"). None of the differences between same K and different K are significant (Wilcoxon test, $p>0.05$). **D.** Histogram showing the distribution of all TFobs/TFexp for pairs of capsulated donor-recipient. A TFobs/TFexp of 1 suggest that there is no interaction between the serotype of the donor and the recipient, *i.e.* the TF of pairs of capsulated cells can be expected from the TF of non-capsulated with capsulated cells (See methods). The distribution is not significantly different from 1 (Wilcoxon test, $p>0.05$).**E**. Conjugation efficiency network between capsule states. Each node represents a capsule state. The width and transparence of edges is scaled to the average bidirectional TF for each pair (n=972 independent conjugation assays, 3 plasmids, 3 strains, 3 donor-recipient groups, 6 capsule states). Note that the scale is capped and thus not linear for extreme values.

We showed that the expression and serotype of capsules play a role in the donation and reception of conjugative plasmids, and that plasmids of different MPF types interacted differently with the capsule. This strongly suggested **the existence of preferential routes of transfer** for each plasmid. To test this and highlight the interaction of the capsule serotype with different plasmids, we first drew conjugation networks normalized by $TF_{\Delta\to\Delta}$ of each group (Fig 5A). We also used a Z-Score including the Δcps, indicative of the direction of the effect (relative to the average), and the strength of statistical difference (Fig 5B). Hence, the normalized TF captures the differences between serotypes relative to the non-capsulated, and the Z-score captures all differences against the group average. Our analyses show that $p476_I$ and $p62_T$ have symmetrical TF, implying that the impact of the capsule is independent of the direction of transfer. For example, for $p476_I$, $Z\text{-score}_{K3\to K1}$ is -0.61 and $Z\text{-score}_{K1\to K3}$ is -0.69. However, we observed that $p23_F$ has an asymmetrical TF, implying that the impact of the capsule is dependent on the direction of transfer. Specifically, K1 is a better than average recipient but a lower than average donor of $p23_F$, as shown by the asymmetry of the Z-score heatmap and of the networks edges. Taken together, these results highlight the importance of the serotype in shaping the conjugation efficiency between strains, and reveal that this effect is dependent on the plasmid.

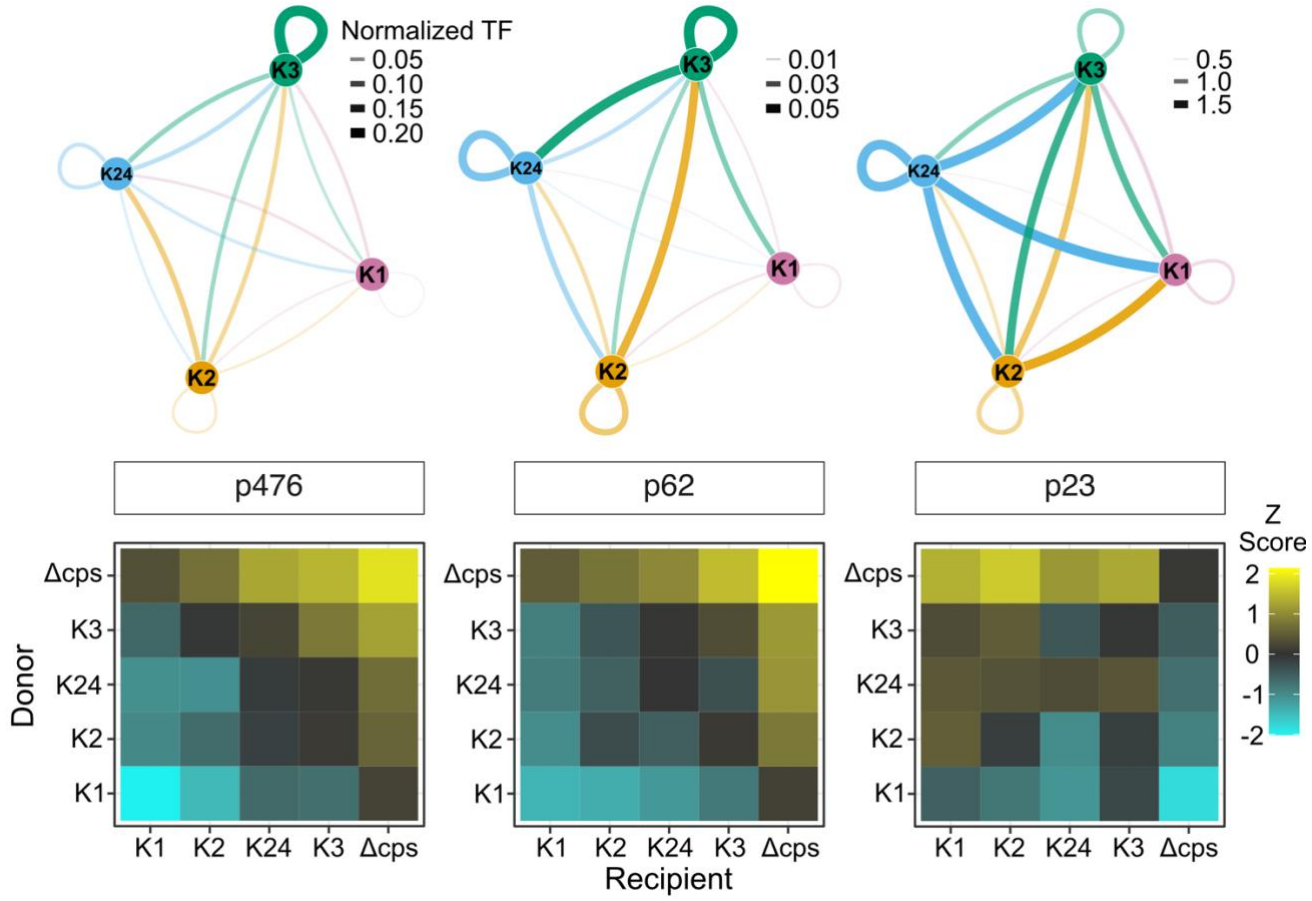


Figure 5 – Conjugation network between Kpn of different serotypes.

A. Plasmid specific conjugation networks. The nodes in the networks correspond to the serotypes, and the edges are colored according to the donor of the conjugative plasmid. Edges thickness and

transparency are scaled to the normalized TF for each pair, highlighting the differences between capsule serotype. The TF is normalized by dividing the average of each TF_{DpR} with the corresponding $TF_{\Delta\rightarrow\Delta}$, enabling visual comparisons between plasmids.

B. Plasmid specific conjugation matrix. Each tile of the matrix is colored according to the average Z-score.

The three focal plasmids, which rely on different types of Mpf for conjugative spread, are associated with different conjugation efficiencies and interactions with the capsule of *K. pneumoniae.* This suggests that different Mpf types are preferentially associated to certain serotypes. To test this, we analysed the distribution of Mpf types among publicly available genomes. We analysed 623 genomes of *K. pneumoniae,* built a phylogenetic tree, and determined the capsule serotype and plasmid content of each genome. We then identified conjugative plasmids with MacSyFinder and annotated clinically relevant genes such as virulence factors and antibiotics resistance genes with Kleborate. Then we categorized them between virulence plasmids, resistance plasmids, virulence and resistance plasmids (Vir & Res) and others (Fig. 6A). We found that 21% of all plasmids carried all the necessary genes for conjugationt, and that 56% of them carried at least one ARG. Among conjugative plasmids, 23% were of type I, 27% of type T and 50% of type F. Type F plasmids were the only conjugative plasmids to carry virulence factors (8%), among which half also encoded at least one resistance gene. We wanted to know if these plasmids were the fruit of recent transfer event. For this, we traced each plasmid's history of acquisition on the species tree (Fig. 6B). We identified plasmids acquired in the terminal branches of the tree, *i.e.* corresponding to recent acquisitions, and found that 68% of the conjugative plasmids had been recently in terminal branches. Hence, most plasmids are recent and they were likely acquired when the strain already had the current serotype.

We then tested if serotypes groups were associated with different amounts of recently acquired conjugative plasmids. We found that there are significant differences in terms of the frequency of acquisition of conjugative plasmid in strains with different serotypes (ANOVA, $p<0.001$), however this analysis only factors in the phylogenetic inertia of plasmids, and not of host strains. We then reanalysed the data focusing on our focal serotypes, namely K1, K2, K3, and K24. In agreement with our conjugative assays, ranking the serotypes in term of conjugation efficiency K1,K2 < K24 < K3, K1 and K2 had a median of 0 acquired conjugative plasmid, whereas K24 had a median of 1, and K3 of 2. Given the underlying phylogeny and the few K3 genomes included, we could only observe, but not test, that the observed trend were in accordance with our conjugation assays (Fig. 4E). Hence, type F conjugation systems are the most frequent among conjugative plasmids of *K. pneumoniae*, and the only ones to harbour virulence factors. Moreover, these results suggest that capsule serotype quantitatively impacts the acquisition of antibiotic resistance genes and virulence factors by conjugation in the population of *K. pneumoniae*.

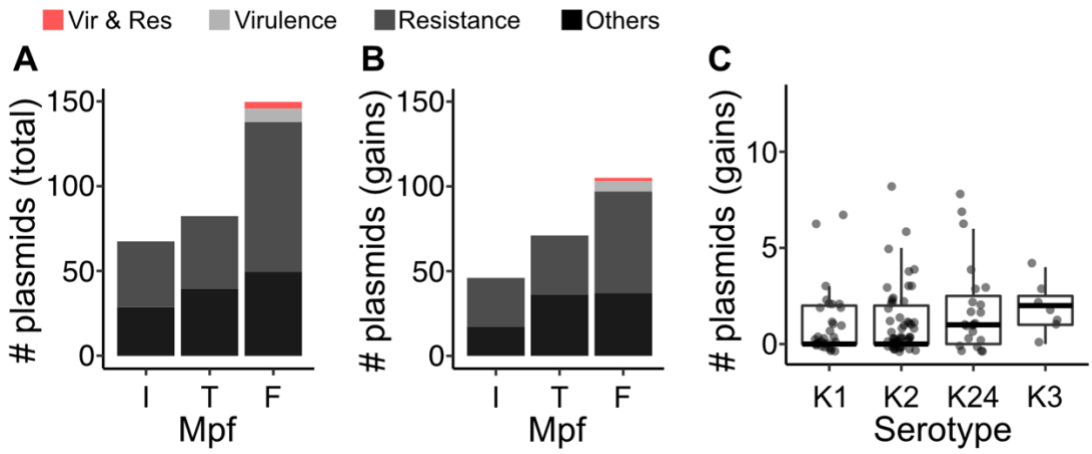


Figure 6 – Distribution and recent acquisition of MPFs in 623 complete K. pneumoniae genomes.

A. Distribution of conjugative plasmids in the dataset. The barplot colors correspond to the different categories of plasmids.

B. Distribution of acquired conjugative plasmids in the dataset. The barplot colors correspond to the different categories of plasmids. The data correspond to conjugative plasmids acquired in terminal branches.

C. Number of acquired plasmids between the serotypes used in this study. The data correspond to all plasmids acquired in terminal branches.

Discussion

Capsule loci are frequently exchanged in natural populations of capsulated bacteria, including *K. pneumoniae*, and their serotype correlates with different traits such as resistance and virulence[37,45]. However, the genetic background of natural isolates with different capsule serotypes precludes direct comparison between strains to enquire on serotype-specific traits. Here, we constructed isogenic swap mutants of natural isolates of *K. pneumoniae* to shed light on the interaction between MGEs, notably conjugative systems, and the capsule serotype. Isogenic swaps expressed visible and functional capsules in three different genetic backgrounds, suggesting that *K. pneumoniae* can readily exchange whole capsule loci and express the expected capsule serotype without prior adaptation. Within *K. pneumoniae*, capsule loci are ready-to-transfer genetic information.

Our attempts at cloning different capsule loci in laboratory *E. coli* strains such as DH5alpha and MFD resulted in the transfer of mutated, non-functional capsule loci to *K. pneumoniae*. Previous studies investigating bioconjugate vaccines showed that K1 and K2 polysaccharides could be produced and conjugated on the core oligosaccharide, using a specialized engineered *E. coli* strain, but only upon expression of the *rmpA* gene from *K. pneumoniae*[46]. Moreover, some natural isolates of *E. coli* have acquired whole capsule loci from *K. pneumoniae,* but these strains are restricted to a few phylotypes which co-acquired the O-antigen gene region as well, and do not encode a colanic acid capsule anymore[47]. Taken together, these results show that capsule of *K. pneumoniae* can be expressed in

other species but not in all genetic backgrounds, certainly due to genetic interference with endogenous capsule loci and O-antigen genes.

The large array of capsule serotypes in capsulated species indicates that strong and persistent selective pressures drive capsule diversification[14]. Such chemical diversification suggests that selection acts on the specific composition of capsule, and not simply on its expression. Here, the isogenic capsule mutants demonstrates that serotype swaps make *K. pneumoniae* resistant to previously infecting phages and sensitive to novel ones. This dual role shows that the capsule is the main receptor for phage infection and suggests that no secondary receptor is generally needed for infection. Indeed, non-capsulated *K. pneumoniae* are notoriously phage-resistant[48–50], and all tested phages could infect our three strains expressing a sensitive serotype. Hence, capsule-phage coevolution is a driver of capsule diversification, and capsule qualitatively shapes phage host-range and phage-mediated HGT (Fig. 7A).

Oppositely to phages, Mpf systems do not seem to rely on specific receptors to engage in conjugation. Indeed, the host range of conjugative plasmids depends primarily on the plasmids replication systems and their compatibility with the host cell. However, we had recently identified that capsule expression can be a barrier to conjugation of an engineered mobilizable plasmid. Additionally, some surface molecules have been shown to decrease conjugation efficiency of ICESt3 in recipient cells[51]. We hypothesized that, if capsule expression can decrease conjugation efficiency, capsule serotypes may have different impacts on conjugation. Using an array of natural conjugative plasmids and isogenic capsule mutants, we found that the acquisition of a new capsule serotype can either lead to a decrease or an increase in the acquisition and donation of conjugative plasmids. This effect was serotype dependent, as the rank from lowest to highest conjugation efficiency was the same in all three strains. Hence, the capsule expression and serotype is a determinant of the strain propensity to exchange conjugative plasmids (Fig 7B).

Here, we could test, and falsify, the hypothesis that bacteria with similar capsules have maximal conjugation rates. It is unlikely that cells of the same serotype stick more together, because this should result in stabilizing the mating-pair and increase conjugation efficiency. Even though such stabilization may not be an important factor of conjugation efficiency in surface mating, these results are compatible with our previous observation that plasmid drive inter-serotype genetic exchanges in *K. pneumoniae*[16]. This is further supported by our results showing, for example, that K1-K3 conjugation is more efficient than K1-K1 transfer. Conjugation-prone serotypes result in highways of plasmid transfer, and this may explain the association between some serotypes and antibiotics resistance, as they are often carried by conjugative plasmids. For example, K1 and K2 have historically had low amounts of resistance genes, and are the serotypes lowering the most conjugation efficiency. Plasmid-encoded virulence factors of *K. pneumoniae*, on the other hand, often rely on the presence of a capsule to confer an advantage, *i.e.* capsule itself is a virulence factor, and positive regulators of the capsule such as *rmpABC* further increase their immune system escape[38,52]. Hence, evolution toward increased resistance might happen at the cost of virulence, and vice versa, explaining the genetic stratification[36] of resistant and virulent populations of *K. pneumoniae*.

An exception to this model is the $p23_F$ plasmid, an Mpf_F plasmid isolated in a low virulence nosocomial strain. Although this plasmid did not seem impacted by the presence or serotype of the capsule, transfer

between *K. pneumoniae* cells revealed that this plasmid was more efficiently donated by non-capsulated cells than capsulated ones, like the other plasmids, but more efficiently received by capsulated cells. We postulate that the Mpf system of this plasmid may have an affinity towards the capsule, however not serotype-specific, but that this affinity is lowered when the donor strain itself is covered by a capsule, which still represents an additional structure to grow over for the type IV secretion pilus. Moreover, since capsule tends to be quickly inactivated when they do not procure a fitness advantage[53], this behaviour would still favour the surrounding capsulated cells. Paired with the observation that only plasmids of the same type (type F) carry virulence factors, and that some virulence factors only provide an advantage in capsulated cells, it is not unexpected that plasmids favouring capsulated recipient would carry virulence factors. Hence, further studies specifically addressing the conjugation efficiency of virulence plasmids in capsulated bacteria are needed to shed light on this scenario.

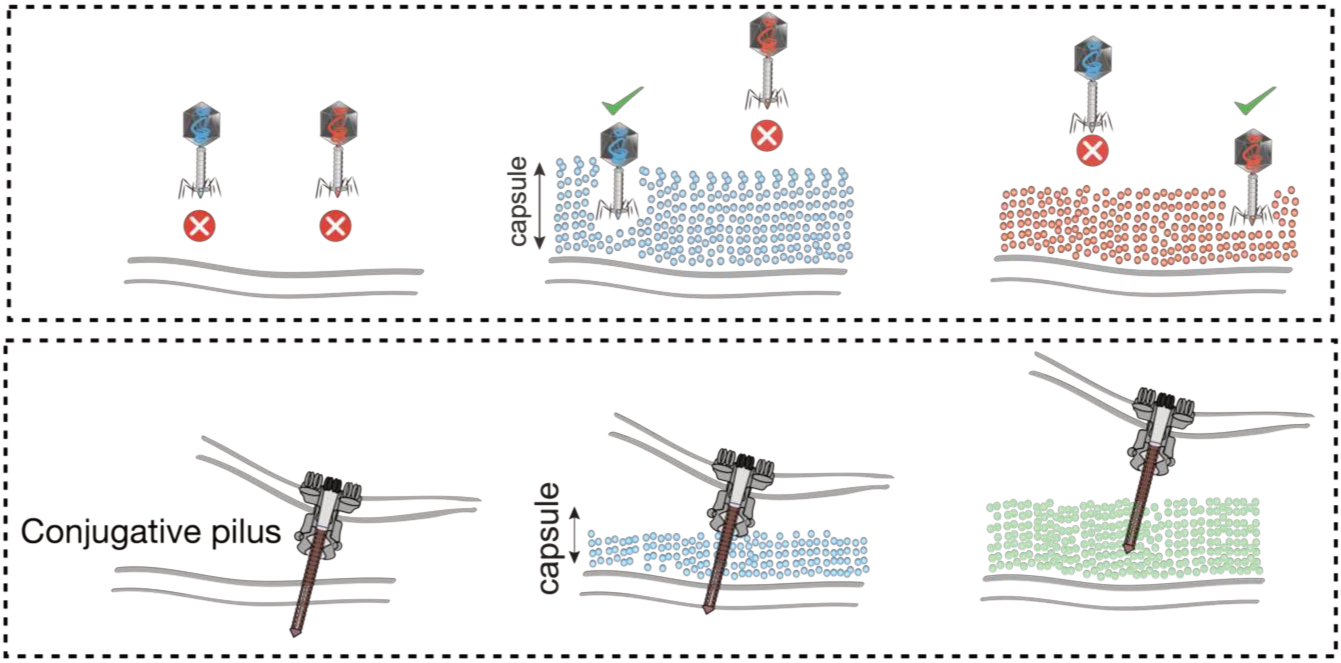


Figure 7 - Impact of the capsule on MGE-mediated horizontal gene transfer.

Material and methods

Bacterial strains

Strain	ST / K	Virulence score	ICEs	plasmids	prophage	Defense systems	Origin
BJ1	ST380 K2	2	T	2	0	AbiEii	Human liver abcess, France
NTUH-K2044	ST23 K1	4	T + G	1	0	Gabija, DRT, DISARM, CRISPR-Cas, Pif, TypeII-RM	Human liver abcess, Taiwan
#51	ST45 K24	1	T	2	3	AbiEii, Thoeris, typeI-RM	Human carriage, feces, Netherlands

Plasmid content:

Strain	Plasmid	Inc	Mpf
BJ1	P1	IncFIB	Type F
BJ1	P2	IncFII	Mob
NTUH-K2044	P1	IncHI1B	
#51	P1	IncFIB/IncFII	Type F
#51	P2	Col440I	

Conjugative plasmids

Plasmid	Source	MPF type	Comment	Ref
p162G2	Hospital	Type I	Almost the same as pOXA48-K8	This work
p476	Hospital	Type I		This work
pOXA-K8	Mazel Lab	Type I		
p168C10	Hospital	Type T		This work
p486	Hospital	Type T		This work
p62	Own collection	Type T		This work
p580	Hospital	Type T & F	Two independent conjugation loci	This work
p23	Own collection	Type F		This work
pR1-drd19	Ghigo Lab	Type F	Dead-regulator mutant	

Scarless artificial serotype swaps

The detailed protocol for scarless artificial serotype swaps (SASS) is available as Sup Text X. Briefly, we first constructed complete capsule loci deletion (Δcps mutants) in our target strains via a Lambda Red knockout with a a modified KmFRT cassette including an I-SceI cut site outside the FRT sites, and 500 bp of homology upstream *galF* and downstream *ugd*. Deletion of the capsule locus was confirmed phenotypically and through amplification of the deleted region (Δcps). The KanMX marker was then excised via expression of the FLP recombinase, leaving an 80bp scar between *galF* and *ugd* (which were left intact) containing the I-SceI cut site.

In parallel, we cloned the whole capsule loci (~30kb) of several different strains with different capsule serotypes via a Lambda Red gap-repair cloning approach. We built a cloning cassette encoding a KanMX resistance marker, an I-SceI cut site, and low copy pSC101 origin of replication. We further added ~500bp homology arms to encompass the first promoter of the capsule locus located 3' of *galF* and the terminator located 5' of *ugd.* This cassette is designed to circularize around the capsule locus via recombination and capture the whole locus to form a pKAPTURE vector[54].

We then purified the pKAPTURE vectors carrying the different capsule loci by miniprep and electroporated it in our Δcps mutants, effectively complementing the mutants with their own, or other, capsule loci.

The Δcps + pKAPTURE strains were electroporated with the pTKRED[55] plasmid, carrying an inducible I-SceI restriction enzyme, inducible Lambda Red system, and a functional copy of *recA*. We performed scarless integration by solely inducing the I-SceI enzyme overnight with selection to maintain pTKRED. The I-SceI enzyme linearizes the pKAPTURE plasmid, providing recombination proficient linear ends, and introduce a chromosomal double-strand break within the capsule deletion, which is lethal if unrepaired with pKAPTURE, resulting in the insertion of the capsule operon . When the repair occurs with the linearized plasmid, the I-SceI cut site is destroyed since the capsule locus recombines outside of the deletion. We identified capsulated colonies on LB plates without selection,

and sequenced the mutants to validate the proper scarless replacement of the capsule locus. All the 12 strains we sequenced carried the expected capsule swap and only one strain carried one off-target mutation (#56-K24) outside the capsule locus (see X).

Phage infection assay

Phage lysates were prepared as described elsewhere (REF). Briefly, overnight cultures were diluted 1:500 in fresh LB and allowed to grow until OD = 0.2. Mitomycin C was added to a final concentration of 5 µg/mL. After 4h hours at 37°C, cultures were centrifuged at 4000 rpm and the supernatant was filtered through 0.22um. Filtered supernatants were mixed with chilled PEG-NaCl 5X (PEG 8000 20% and 2.5M of NaCl) and mixed through inversion. Phages were allowed to precipitate for 15 min and pelleted by centrifugation 10 min at 13000 rpm at 4°C. The pellets were dissolved in TBS (Tris Buffer Saline, 50 mM Tris-HCl, pH 7.5, 150 mM NaCl). To test the susceptibility of wild type and capsule swapped strains to phages, overnight cultures in LB of strains were diluted 1:100 and allowed to grow until OD = 0.8. 250 µL of bacterial cultures were mixed with 3 mL of top agar (0.7% agar) and poured intro prewarmed LB plates to generate t bacteria overlay. Plates were allowed to dry before spotting serial dilutions of induced PEG-precipitated phages. Plates were incubated at 37° degrees for 4 hours and pictures were taken.

Isolation of conjugative plasmids

We screened the genomes of our own collection and the Klebsiella isolates of the National Reference Centre laboratory for Carbapenemase-producing Enterobacteriaceae at the Bicêtre Hospital to identify contigs resembling conjugative plasmids. We used plasmidfinder (ref) to retrieve plasmid contigs, MacSyFinder with TXSScan models (ref) to identify conjugation operons and annotate their mating-pair formation type, and ResFinder to annotate antibiotics resistance genes (ref). We gathered a list of contigs containing the following features: a plasmid replicase identified and typed by PlasmidFinder, a complete conjugation system, and at least once selectable antibiotic resistance (carbapenem or kanamycin resistance, absent in our swapped strains). Additionally, we included two extensively studied conjugative plasmids, pOXA48-K8 (Mpf_I) and the dead regulator version of the R1 plasmid, R1-drd19 (Mpf_F). The clinical isolates carried many other MGEs, so we purified our plasmids of interest through by conjugation into *E. coli* DH10B cells. We sequenced one transconjugant per plasmid, and validated that they only contained a single conjugative system.

Conjugation assay

We used *Escherichia coli* DH10B as donor for our *E. coli* to *Kpn* conjugation experiments because of several necessary traits: streptomycin resistance (selectable), leucine auxotrophy (counter-selectable), efficient plasmid maintenance (*recA1*, *endA1*, *relA1*), lack of restriction-modification systems and of other conjugative elements (ref DH10B genome paper).

For Eco to Kpn experiments, and owing to a slower *E. coli* growth rate compared to that Kpn, an overnight culture in LB with appropriate antibiotics (Atb) at 37°C was used to inoculate 3mL LB + Atb (diluted 1:100). In parallel, a culture was started from a single colony of the Kpn recipient strain, to

avoid the emergence of non-capsulated cells that can appear rapidly under laboratory conditions (Ref Buffet et al).

For Kpn to Kpn experiments the transconjugants of the Eco to Kpn experiments were used as donors for the Kpn to Kpn experiments. A single colony was inoculated in LB with appropriate antibiotics for the donor strains (Kpn + Plasmid), and without antibiotics for the recipient strains.

For both sets of experiments, cells reached an OD600 ≈ 1 after 4h of over-day growth, at which point donor cultures were centrifuged and resuspended in antibiotic-free LB. They were then mixed 1:1 vol/vol and a 15uL drop of the mixture was adsorbed onto 1mL LB-agar pads in a 24-well microtiter plate. The droplets were allowed to dry under the hood with laminar flow (5-10min). After one hour at 37°C, 1mL of Phosphate-buffer saline was added in each well, the plates sealed with a hydrophobic adhesive film and shaken at 120rpm for 10min to resuspend the lawn. The contents of each well was then transferred to a 96-well plate for serial dilutions, which were spotted on plates selecting for either donor cells, recipient cells, or transconjugants. The next day, colonies were counted at the appropriate dilution (between 3 and 30 colonies per spot).

To estimate the conjugation efficiency, we computed the transconjugant frequency (TF) at 1h with the following method:

$$Transconjugant\ frequency = \frac{T}{D.R} \cdot \Delta t$$

Where T is the transconjugants concentration (CFU/mL), D the donors concentration and R the recipients concentration. This method performs accurately to estimate the efficiency of conjugation under short conjugation time, and consistent D:R ratio when compared with far more complex population-based methods [56,57] (refs). TF is thus expressed in $mL.CFU^{-1}.hours^{-1}$, and represents the transfer rate constant. Importantly, we followed the recommendation presented in [57] allowing the T/DR quantity be highly accurate.

We compared this quantity with the widely used and simpler formula T/(D+R), and found a Pearson correlation coefficient of 0.97 ($p<0.001$) when we compared the log(TF) with log(T/(D+R)). Hence, these two values are interchangeable and our findings are robust to several methods. Since the TF value has been shown to be less susceptible to several parameters such as D:R ratio, we decided to use the TF in our study.

We used each $Donor_{plasmid}$ to Recipient ($D_p \rightarrow R$) group average ($\overline{TF_{D_p \rightarrow R}}$) and standard deviation (sd) to compute a Z-score for each observed estimate abbreviated TF_{obs} corresponding to the average transconjugant frequency of independent triplicates for a given Donor, Donor serotype, Plasmid, Recipient, Recipient serotype:

$$Zscore_{obs} = \frac{TF_{obs} - \overline{TF_{D_p \rightarrow R}}}{sd(TF_{D_p \rightarrow R})}$$

We also used the $TF_{\Delta_p \to \Delta}$ to normalize the TF for each $D_p \to R$ groups:

$$Normalized\ TF_{obs} = \frac{TF_{obs}}{TF_{\Delta_p \to \Delta}}$$

We leveraged the natural metabolism and resistance of the strains to distinguish donors, recipients, and transconjugants. We tested several antibiotics and carbon sources identified in (Blin ref) to prepare selective media corresponding to each of our three Kpn strain.

The strategy of selective plating for each donor-recipient-plasmid group is detailed in Sup X.

Capsule microscopy

We mixed 100uL of Kpn overnight culture cultivated in shaken LB at 37°C with 20uL of India Ink and smeared over a microscope slide and air-dried. We visualized the slides with 100x oil-immersed phase-contrast objective in a XXXX Leica/Olympus XXYY microscope. Pictures were taken with a XYZ camera.

Genomics analysis

[to do]

References

1. Mitchell AM, Silhavy TJ. Envelope stress responses: balancing damage repair and toxicity. Nat Rev Microbiol. 2019;17: 417–428. doi:10.1038/s41579-019-0199-0

2. Delhaye A, Collet J-F, Laloux G. A Fly on the Wall: How Stress Response Systems Can Sense and Respond to Damage to Peptidoglycan. Front Cell Infect Microbiol. 2019;9: 380. doi:10.3389/fcimb.2019.00380

3. Rendueles O, Garcia-Garcerà M, Néron B, Touchon M, Rocha EPC. Abundance and co-occurrence of extracellular capsules increase environmental breadth: Implications for the emergence of pathogens. PLOS Pathog. 2017;13: e1006525. doi:10.1371/journal.ppat.1006525

4. Tipton KA, Chin C-Y, Farokhyfar M, Weiss DS, Rather PN. Role of Capsule in Resistance to Disinfectants, Host Antimicrobials, and Desiccation in Acinetobacter baumannii. Antimicrob Agents Chemother. 2018;62. doi:10.1128/AAC.01188-18

5. Scholl D, Adhya S, Merril C. Escherichia coli K1's capsule is a barrier to bacteriophage T7. Appl Environ Microbiol. 2005;71: 4872–4874. doi:10.1128/AEM.71.8.4872-4874.2005

6. Soundararajan M, von Bünau R, Oelschlaeger TA. K5 Capsule and Lipopolysaccharide Are Important in Resistance to T4 Phage Attack in Probiotic E. coli Strain Nissle 1917. Front Microbiol. 2019;10: 2783. doi:10.3389/fmicb.2019.02783

7. Jung S-Y, Matin A, Kim KS, Khan NA. The capsule plays an important role in Escherichia coli K1 interactions with Acanthamoeba. Int J Parasitol. 2007;37: 417–423. doi:10.1016/j.ijpara.2006.10.012

8. Campos MA, Vargas MA, Regueiro V, Llompart CM, Albertí S, Bengoechea JA. Capsule polysaccharide mediates bacterial resistance to antimicrobial peptides. Infect Immun. 2004;72: 7107–7114. doi:10.1128/IAI.72.12.7107-7114.2004

9. Hyams C, Camberlein E, Cohen JM, Bax K, Brown JS. The Streptococcus pneumoniae Capsule Inhibits Complement Activity and Neutrophil Phagocytosis by Multiple Mechanisms. Infect Immun. 2010;78: 704–715. doi:10.1128/IAI.00881-09

10. Kostina E, Ofek I, Crouch E, Friedman R, Sirota L, Klinger G, et al. Noncapsulated Klebsiella pneumoniae bearing mannose-containing O antigens is rapidly eradicated from mouse lung and triggers cytokine production by macrophages following opsonization with surfactant protein D. Infect Immun. 2005;73: 8282–8290. doi:10.1128/IAI.73.12.8282-8290.2005

11. Whitfield C. Biosynthesis and assembly of capsular polysaccharides in Escherichia coli. Annu Rev Biochem. 2006;75: 39–68. doi:10.1146/annurev.biochem.75.103004.142545

12. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, et al. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet. 2006;2: e31. doi:10.1371/journal.pgen.0020031

13. Pan Y-J, Lin T-L, Chen C-T, Chen Y-Y, Hsieh P-F, Hsu C-R, et al. Genetic analysis of capsular polysaccharide synthesis gene clusters in 79 capsular types of Klebsiella spp. Sci Rep. 2015;5: 15573. doi:10.1038/srep15573

14. Mostowy RJ, Holt KE. Diversity-Generating Machines: Genetics of Bacterial Sugar-Coating. Trends Microbiol. 2018;26: 1008–1021. doi:10.1016/j.tim.2018.06.006

15. Mostowy RJ, Croucher NJ, De Maio N, Chewapreecha C, Salter SJ, Turner P, et al. Pneumococcal Capsule Synthesis Locus cps as Evolutionary Hotspot with Potential to Generate Novel Serotypes by Recombination. Mol Biol Evol. 2017;34: 2537–2554. doi:10.1093/molbev/msx173

16. Haudiquet M, Buffet A, Rendueles O, Rocha EPC. Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen Klebsiella pneumoniae. PLoS Biol. 2021;19: e3001276. doi:10.1371/journal.pbio.3001276

17. Croucher NJ, Kagedan L, Thompson CM, Parkhill J, Bentley SD, Finkelstein JA, et al. Selective and Genetic Constraints on Pneumococcal Serotype Switching. PLoS Genet. 2015;11. doi:10.1371/journal.pgen.1005095

18. Rendueles O, Sousa JAM de, Bernheim A, Touchon M, Rocha EPC. Genetic exchanges are more frequent in bacteria encoding capsules. PLOS Genet. 2018;14: e1007862. doi:10.1371/journal.pgen.1007862

19. García-Aljaro C, Ballesté E, Muniesa M. Beyond the canonical strategies of horizontal gene transfer in prokaryotes. Curr Opin Microbiol. 2017;38: 95–105. doi:10.1016/j.mib.2017.04.011

20. Guglielmini J, Néron B, Abby SS, Garcillán-Barcia MP, de la Cruz F, Rocha EPC. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. Nucleic Acids Res. 2014;42: 5715–5727. doi:10.1093/nar/gku194

21. de la Cruz F, Frost LS, Meyer RJ, Zechner EL. Conjugative DNA metabolism in Gram-negative bacteria. FEMS Microbiol Rev. 2010;34: 18–40. doi:10.1111/j.1574-6976.2009.00195.x

22. Touchon M, Moura de Sousa JA, Rocha EP. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. Curr Opin Microbiol. 2017;38: 66–73. doi:10.1016/j.mib.2017.04.010

23. de Sousa JAM, Buffet A, Haudiquet M, Rocha EPC, Rendueles O. Modular prophage interactions driven by capsule serotype select for capsule loss under phage predation. ISME J. 2020;14: 2980–2996. doi:10.1038/s41396-020-0726-z

24. Latka A, Leiman PG, Drulis-Kawa Z, Briers Y. Modeling the Architecture of Depolymerase-Containing Receptor Binding Proteins in Klebsiella Phages. Front Microbiol. 2019;10: 2649. doi:10.3389/fmicb.2019.02649

25. Rieger-Hug D, Stirm S. Comparative study of host capsule depolymerases associated with Klebsiella bacteriophages. Virology. 1981;113: 363–378. doi:10.1016/0042-6822(81)90162-8

26. Bertozzi Silva J, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. FEMS Microbiol Lett. 2016;363: fnw002. doi:10.1093/femsle/fnw002

27. Pires DP, Oliveira H, Melo LDR, Sillankorva S, Azeredo J. Bacteriophage-encoded depolymerases: their diversity and biotechnological applications. Appl Microbiol Biotechnol. 2016;100: 2141–2151. doi:10.1007/s00253-015-7247-0

28. Guglielmini J, de la Cruz F, Rocha EPC. Evolution of conjugation and type IV secretion systems. Mol Biol Evol. 2013;30: 315–331. doi:10.1093/molbev/mss221

29. Pérez-Mendoza D, de la Cruz F. Escherichia coli genes affecting recipient ability in plasmid conjugation: Are there any? BMC Genomics. 2009;10: 71. doi:10.1186/1471-2164-10-71

30. Ishiwa A, Komano T. PilV Adhesins of Plasmid R64 Thin Pili Specifically Bind to the Lipopolysaccharides of Recipient Cells. J Mol Biol. 2004;343: 615–625. doi:10.1016/j.jmb.2004.08.059

31. Stuy JH. Plasmid transfer in Haemophilus influenzae. J Bacteriol. 1979;139: 520–529. doi:10.1128/JB.139.2.520-529.1979

32. von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S, van Alphen LB, et al. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. Front Microbiol. 2016;7: 173. doi:10.3389/fmicb.2016.00173

33. Rice LB. Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE. J Infect Dis. 2008;197: 1079–1081. doi:10.1086/533452

34. Heinz E, Follador R, Thomson NR, Holt KE, Kowarik M, Wyres KL, et al. The diversity of Klebsiella pneumoniae surface polysaccharides. Microb Genomics. 2016;2. doi:10.1099/mgen.0.000073

35. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. Identification of Klebsiella capsule synthesis loci from whole genome data. Microb Genomics. 2016;2. doi:10.1099/mgen.0.000102

36. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae. PLoS Genet. 2019;15. doi:10.1371/journal.pgen.1008114

37. Yu W-L, Ko W-C, Cheng K-C, Lee C-C, Lai C-C, Chuang Y-C. Comparison of prevalence of virulence factors for Klebsiella pneumoniae liver abscesses between isolates with capsular K1/K2 and non-K1/K2 serotypes. Diagn Microbiol Infect Dis. 2008;62. doi:10.1016/j.diagmicrobio.2008.04.007

38. Walker KA, Miller VL. The intersection of capsule gene expression, hypermucoviscosity and hypervirulence in Klebsiella pneumoniae. Curr Opin Microbiol. 2020;54: 95–102. doi:10.1016/j.mib.2020.01.006

39. Hyams C, Yuste J, Bax K, Camberlein E, Weiser JN, Brown JS. Streptococcus pneumoniae resistance to complement-mediated immunity is dependent on the capsular serotype. Infect Immun. 2010;78: 716–725. doi:10.1128/IAI.01056-09

40. Okura M, Auger J-P, Shibahara T, Goyette-Desjardins G, Van Calsteren M-R, Maruyama F, et al. Capsular polysaccharide switching in Streptococcus suis modulates host cell interactions and virulence. Sci Rep. 2021;11: 6513. doi:10.1038/s41598-021-85882-3

41. Paton JC, Trappetti C. Streptococcus pneumoniae Capsular Polysaccharide. Microbiol Spectr. 2019;7. doi:10.1128/microbiolspec.GPP3-0019-2018

42. Tan YH, Chen Y, Chu WHW, Sham L-T, Gan Y-H. Cell envelope defects of different capsule-null mutants in K1 hypervirulent Klebsiella pneumoniae can affect bacterial pathogenesis. Mol Microbiol. 2020;113: 889–905. doi:10.1111/mmi.14447

43. Ofek I, Kabha K, Athamna A, Frankel G, Wozniak DJ, Hasty DL, et al. Genetic exchange of determinants for capsular polysaccharide biosynthesis between Klebsiella pneumoniae strains expressing serotypes K2 and K21a. Infect Immun. 1993;61: 4208–4216.

44. Bradshaw JL, Rafiqullah IM, Robinson DA, McDaniel LS. Transformation of nonencapsulated Streptococcus pneumoniae during systemic infection. Sci Rep. 2020;10. doi:10.1038/s41598-020-75988-5

45. Russo TA, Marr CM. Hypervirulent Klebsiella pneumoniae. Clin Microbiol Rev. 2019;32: e00001-19. doi:10.1128/CMR.00001-19

46. Feldman MF, Mayer Bridwell AE, Scott NE, Vinogradov E, McKee SR, Chavez SM, et al. A promising bioconjugate vaccine against hypervirulent Klebsiella pneumoniae. Proc Natl Acad Sci U S A. 2019;116: 18655–18663. doi:10.1073/pnas.1907833116

47. Nanayakkara BS, O'Brien CL, Gordon DM. Diversity and distribution of Klebsiella capsules in Escherichia coli. Environ Microbiol Rep. 2019;11: 107–117. doi:10.1111/1758-2229.12710

48. Hesse S, Rajaure M, Wall E, Johnson J, Bliskovsky V, Gottesman S, et al. Phage Resistance in Multidrug-Resistant Klebsiella pneumoniae ST258 Evolves via Diverse Mutations That Culminate in Impaired Adsorption. mBio. 2020;11: e02530-19. doi:10.1128/mBio.02530-19

49. Tan D, Zhang Y, Qin J, Le S, Gu J, Chen L, et al. A Frameshift Mutation in wcaJ Associated with Phage Resistance in Klebsiella pneumoniae. Microorganisms. 2020;8: 378. doi:10.3390/microorganisms8030378

50. Verma V, Harjai K, Chhibber S. Restricting ciprofloxacin-induced resistant variant formation in biofilm of Klebsiella pneumoniae B5055 by complementary bacteriophage treatment. J Antimicrob Chemother. 2009;64: 1212–1218. doi:10.1093/jac/dkp360

51. Dahmane N, Robert E, Deschamps J, Meylheuc T, Delorme C, Briandet R, et al. Impact of Cell Surface Molecules on Conjugative Transfer of the Integrative and Conjugative Element ICESt3 of Streptococcus thermophilus. Appl Environ Microbiol. 2018;84. doi:10.1128/AEM.02109-17

52. Hsu C-R, Lin T-L, Chen Y-C, Chou H-C, Wang J-T. The role of Klebsiella pneumoniae rmpA in capsular polysaccharide synthesis and virulence revisited. Microbiol Read Engl. 2011;157: 3446–3457. doi:10.1099/mic.0.050336-0

53. Buffet A, Rocha EPC, Rendueles O. Nutrient conditions are primary drivers of bacterial capsule maintenance in Klebsiella. Proc Biol Sci. 2021;288: 20202876. doi:10.1098/rspb.2020.2876

54. Thomason L, Court DL, Bubunenko M, Costantino N, Wilson H, Datta S, et al. Recombineering: genetic engineering in bacteria using homologous recombination. Curr Protoc Mol Biol. 2007;Chapter 1: Unit 1.16. doi:10.1002/0471142727.mb0116s78

55. Kuhlman TE, Cox EC. Site-specific chromosomal integration of large synthetic constructs. Nucleic Acids Res. 2010;38: e92. doi:10.1093/nar/gkp1193

56. Huisman JS, Benz F, Duxbury SJN, de Visser JAGM, Hall AR, Fischer EAJ, et al. Estimating plasmid conjugation rates: A new computational tool and a critical comparison of methods. Plasmid. 2022;121: 102627. doi:10.1016/j.plasmid.2022.102627

57. Zhong X, Droesch J, Fox R, Top EM, Krone SM. On the meaning and estimation of plasmid transfer rates for surface-associated and well-mixed bacterial populations. J Theor Biol. 2012;294: 144–152. doi:10.1016/j.jtbi.2011.10.034

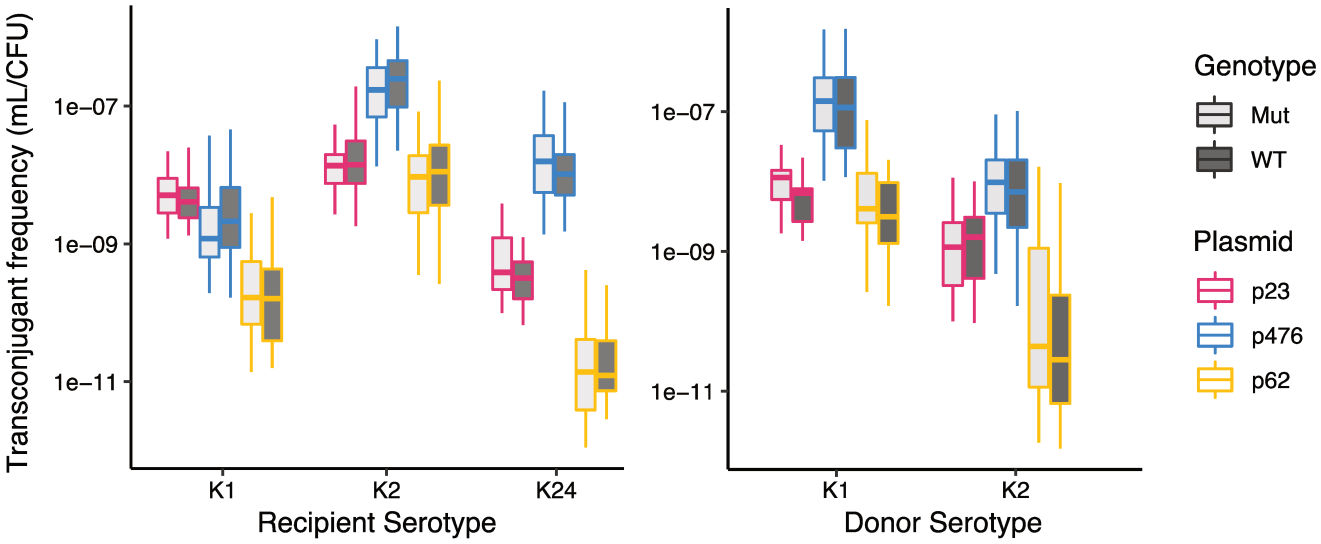


Figure S1 - WT vs. same-serotype swap (Mut) controls. None of WT vs Mut comparisons are significantly different (Wilcoxon test, p>0.05).

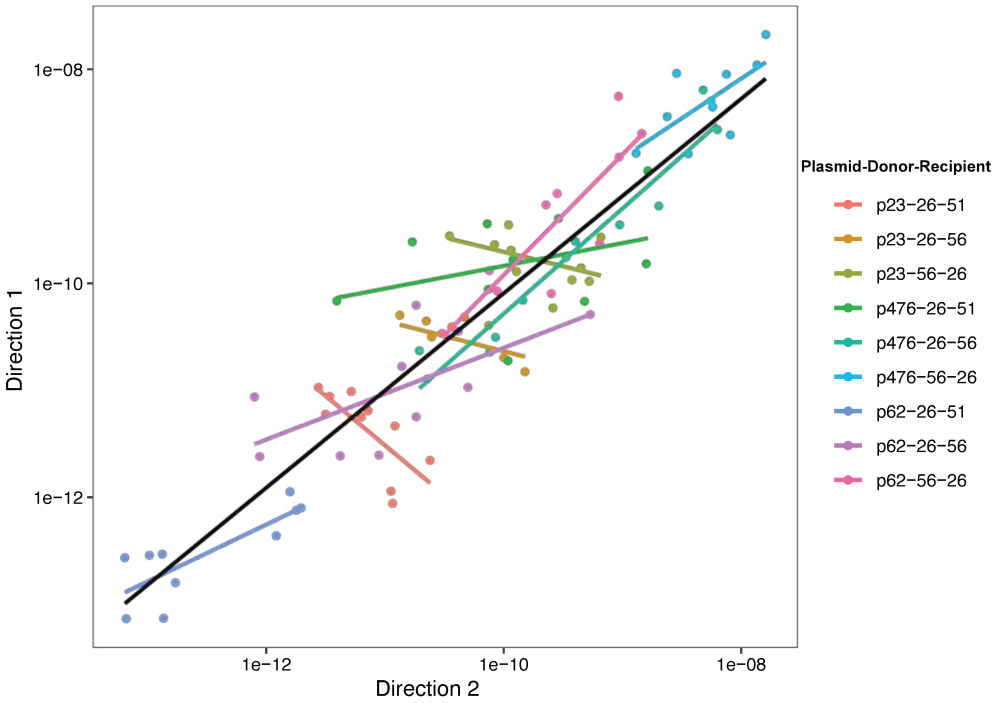


Figure S2 – Regression between the TF of $K_A \rightarrow K_B$ vs $K_B \rightarrow K_A$. The black line is the regression for all points, whereas colored lines and point correspond to each donor-recipient-plasmid groups. The overall regression is positive with a coefficient of 0.99 (R^2= 0.8, p<0.001). Note the inverse correlation for p23 groups.

Conclusions and perspectives

In most bacterial species, horizontal gene transfer (HGT) is mediated by mobile genetic elements (MGE) which can transfer their own, or their host's, genetic material to recipient cells [18,32,89,94]. As a result, genes flow through populations and drive the evolution of bacteria [21], notably pathogens [34,37,81,198]. MGEs encode physical structures, such as virions and conjugative pili, able to cross the cellular envelope of recipient cells [98,101,111]. Any determinant restraining, or giving access to MGEs may thus impact the gene flow within a population [214,215]. One key feature of the envelope of many bacterial species is the polysaccharide capsule [209], which is composed of membrane-bound polysaccharide chains [208] and constitutes the first point of contact of MGEs with the cell. Capsules harbor diverse chemical compositions called serotypes [224,231,234], which diversify by HGT [174,175,250]. Hence, capsules may represent a physical barrier against MGEs [259], but evolve via HGT. Moreover, capsulated species tend to have relatively higher rates of genetic exchange [278]. To solve this apparent conundrum, this thesis focused on **the interplay between capsules and HGT in *K. pneumoniae*, an MGE-rich ubiquitous species with a large serotype diversity.**

Studying the gene flow in *K. pneumoniae* revealed that gene transfer happened more frequently between strains of the same serotype, resulting in an **intra-serotype gene flow bias** [335]. This may be explained by ecological effects, because populations with similar serotypes might occupy similar niches and have more opportunities for genetic exchanges [336]. It may also be explained by genetic relatedness, since closely related strains tend to exchange genetic material more frequently [19,31,337,338]. One would expect that conjugation-mediated and phage-mediated HGT would be similarly impacted by ecological and relatedness effects, but focusing the gene flow analysis on plasmids revealed that it was actually biased toward **inter**-serotype exchanges. Conversely, temperate phage gene flow was highly biased toward **intra**-serotype exchanges. Overall, the intra-serotype bias appears to be largely driven by phages, overshadowing the opposite bias observed in plasmids.

Why is phage-mediated gene flow higher between strains of the same serotype? Naturally occurring virulent phages of capsulated species, such as *K. pneumoniae* or *Acinetobacter baumannii*, are generally specific to one, or sometimes several, serotypes [265,274,339]. The long-standing coevolution of phages with their capsulated hosts has resulted in an ecological love-hate relationship: in species typically non-capsulated, their phages are blocked by the expression of a capsule, but in species ubiquitously capsulated, their phages are dependent on the capsule for successful infection

(Figure 32A). And while *in vitro* phage isolation may not reflect the relevance of capsule dependency in nature, because they were isolated in conditions where the capsule is expressed (*e.g.* rich culture media at 37c), temperate phages having infected their host before isolation show extensive serotype specificity and dependency as well [293]. This concurs with the idea that **capsule serotype dependency is widespread in phages of capsulated species**, resulting in short-term selection for capsule inactivation [266,293,340], long-term selection for capsule diversification [118], and intra-serotype gene flow bias [335] (Figure 32A). While prophages represent phage-mediated HGT, gene flow mediated by other mechanisms like generalized and lateral transduction was not quantified during this work. This is due to the fact that such events leave no indication of the vector that brought foreign DNA into the cell.

Why is conjugation-mediated gene flow higher between strains of different serotype? At least two observations can explain this phenomenon: i) capsule inactivation is frequent [335], and those non-capsulated cells have up to 100-fold larger conjugation frequencies [335] and ii) capsule serotypes differently impact conjugation efficiency [341]. By first identifying pseudogenized capsule loci and analyzing the gene flow of those isolates, we identified **non-capsulated cells as conjugation hubs**. We further provided experimental evidence that non-capsulated cells have larger reception and donation rates [335,341], and that such conjugation hubs provide highways of transfer between distinct serotypes. By constructing **isogenic serotype swaps** with a novel method, we showed that conjugation efficiency is determined by the capsule serotype of both the donor and recipient. For example, serotypes K1 and K2 are associated with relatively lower conjugation efficiencies than K24 and K3 serotypes. Additionally, cells harboring the same capsule serotype do not have higher conjugation frequency, and there appear to be little to no interaction between the donor and recipient capsule during conjugation. Accordingly, conjugation is generally more efficient between K1 and K3 cells than between K1 cells. **Thus, differences in conjugation efficiency can result in preferential transfer between cells of different serotype.** The mechanistic reasons why serotypes are associated with different conjugation efficiencies are still unknown. To answer this question, isogenic serotype swaps will be key to delineate the precise differences between serotype, regardless of the genetic background. Our lead hypothesis is that the inherent thickness of serotype negatively correlates with conjugation efficiency, because it decreases physical proximity between cells (Figure 32A). This hypothesis is in line with the observation that hypervirulent clones of *K. pneumoniae* tend to have relatively lower genetic diversity [290,295], as their thick capsules may increase immune system escape but lower conjugation rates.

Conjugative systems belong to the type IV secretion system family and are divided into mating-pair formation (MPF) types based on the genetic content of the conjugation locus. In Enterobacteriaceae such as *K. pneumoniae*, conjugative plasmids often belong to the MPF_F, MPF_T or MPF_I types. MPF_T and MPF_I plasmids included in our study both displayed a symmetrical impact of the capsule on the donor and recipient (*i.e.* donation and reception efficiency are positively correlated). However, p23F, a MPF_F plasmid had an asymmetrical donor/recipient behavior displaying: i) preferential donation from non-capsulated cells, **ii) preferential reception into capsulated cells.** Recently, Low et al. [342] found that MPF_F plasmids often encode an outer-membrane protein (OMP) called TraN, which cooperates with distinct OMP in recipients to mediate mating pair stabilization and efficient conjugation. These TraN-OM receptor pairings reflected the distribution of resistance plasmids within Enterobacteriaceae: *e.g.* plasmids with the TraNβ allele were primarily found in *K. pneumoniae* which harbors its OmpK36 receptor. Given that i) OmpK36 is involved in antibiotic resistance, ii) OmpK36 and the capsule are co-regulated [343], and iii) the capsule may cover TraN (donor), and OmpK36 (recipient), one can postulate that a tripartite interaction between OMP, capsule and conjugation system could explain our results. Hypothetically, **capsule production could interfere with TraN in the donor, but increases the expression of OmpK36 in the recipient, explaining why non-capsulated cells conjugate more efficiently into capsulated cells.**

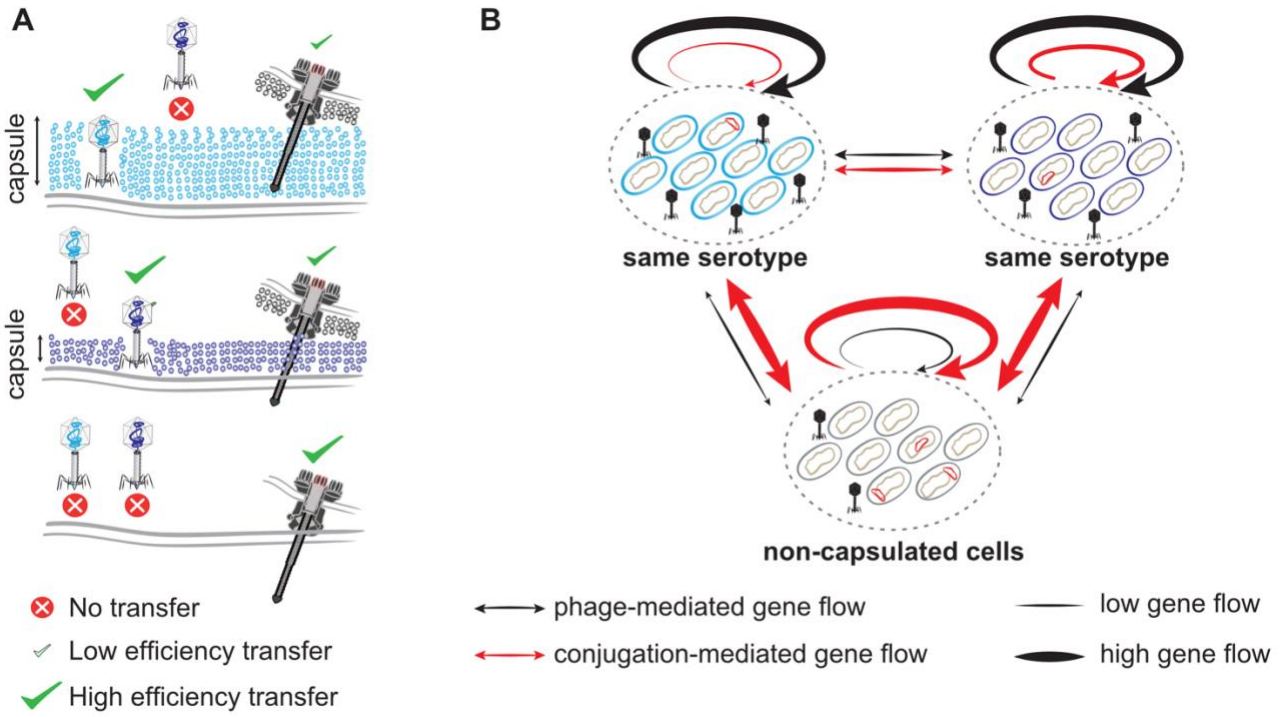


Figure 32 - Interaction between mobile genetic elements and capsules result in biased gene flow. **A.** Three cell envelopes are represented with distinct capsule serotype (Blue on top, purple in the middle) or no capsule (bottom). Phages are serotype-specific, while conjugation is differently impacted by the capsule presence/absence and serotype, possibly because of differences in thickness. Non-capsulated cells are resistant to capsule-dependent phages, and have higher conjugation (donation and reception) efficiencies. **B.** Three populations, corresponding to the three envelopes of panel A, are represented. Arrows represent the relative gene flow according to the vectors of HGT. Figure adapted from [89].

Shedding light on the interaction between MGEs and the capsule led to the formulation of **a model for capsule swap**, the process by which a capsule locus is replaced by another via horizontal gene transfer. Serotype-specific phage predation might be the strongest selective pressure on capsule composition, and rapidly select for capsule inactivation [266,293,344], however capsule are found in numerous species and associated with increased niche colonization [209], and are thus generally favorable. This thesis provides a new model focusing on **how** MGEs can drive serotype swap, presented in figure 33. This model is limited by our limited knowledge of broad-host range phages, which could theoretically transfer capsule loci from and to capsulated cells

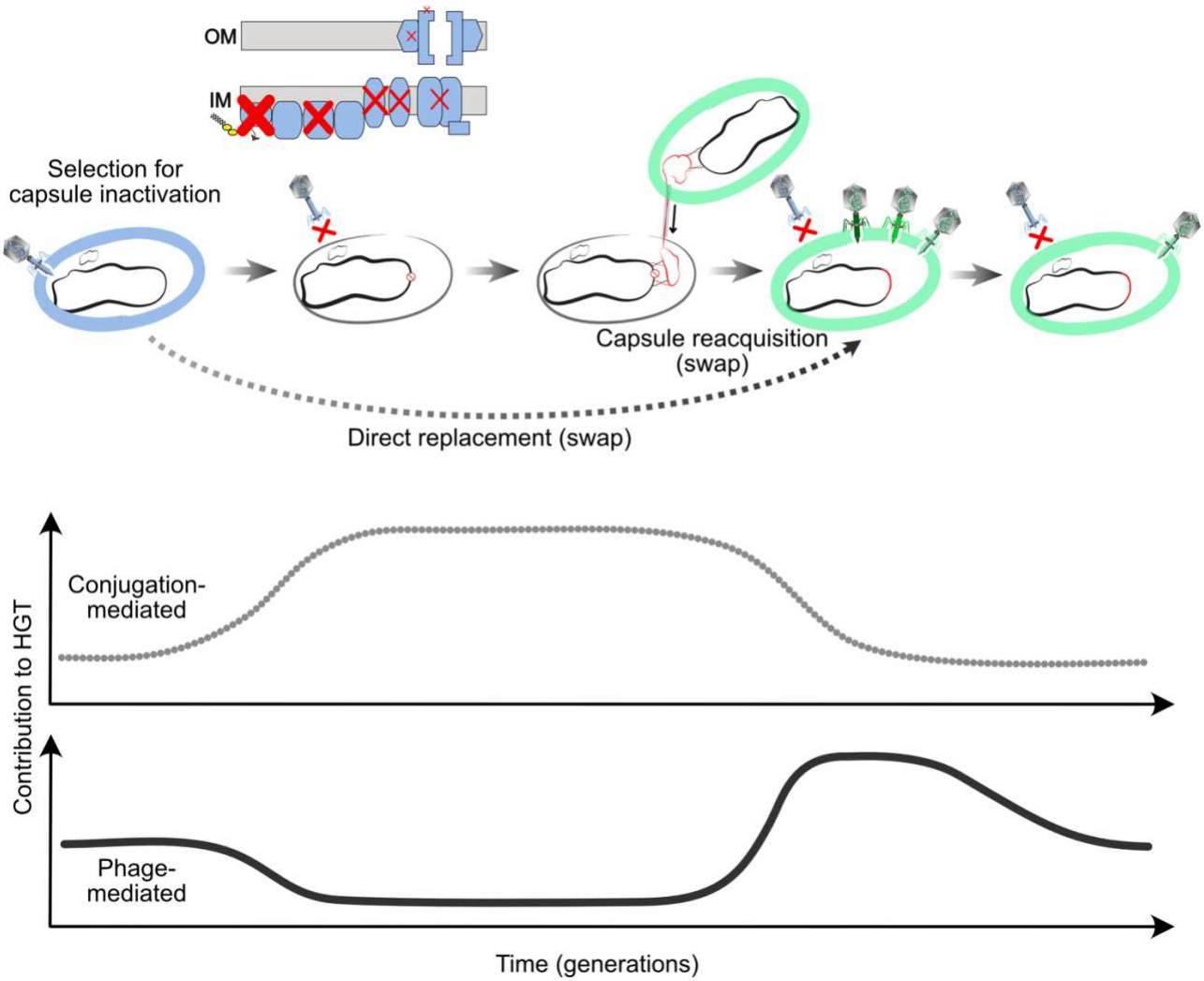


Figure 33 - Proposed model for serotype swaps in *K. pneumoniae* and its relationship with MGE-mediated DNA acquisition. The capsule locus is colored according to its type. Capsule inactivation is occasionally adaptive, *e.g.* in the context of phage predation. The pseudogenization process usually starts by the inactivation of the genes involved in the early stages of the capsule biosynthesis, as represented by the size of the red cross on the capsule assembly scheme. Non-capsulated strains are often protected from *K. pneumoniae* phage infections while acquiring more genes by conjugation. This increases the likelihood of capsule reacquisition. Such reacquisition can bring a new serotype, often one that is chemically similar to the previous one, and might be driven by conjugation because of its high frequency in non-capsulated strains. Recently swapped strains are associated with an increase in prophage acquisition. Finally, serotype swaps rewire phage-mediated genetic transfers.

These findings have broad implications for the evolution and treatment of capsulated pathogens, including the six members of the ESKAPEs [210]e. They highlight a trade-off between the evolution of virulence, correlated with non-immunogenic capsules like K1 and K2 [284,289], and the acquisition of antibiotic resistance genes by conjugation, more efficient between non-capsulated cells [335] and non-hypervirulent serotypes [341]. Because of the current antibiotics resistance crisis, new therapies are envisioned to treat the ESKAPEs, including phage therapy and capsule-based vaccines. Several studies have already documented the rapid emergence of non-capsulated variants during phage therapy against *K. pneumoniae* [266,268,340,345], and vaccine-induced immunity has been shown to select for novel serotypes and non-capsulated clones in *Streptococcus pneumoniae* [243]. As non-capsulated clones are hubs of conjugative transfer, they pose an increased risk of spreading antibiotics resistance genes in bacterial populations. Moreover, they may themselves re-acquire functional, phage-resistant or nonvaccine serotypes. In *S. pneumoniae*, non-capsulated variants display higher rates of HGT presumably because the capsule decreases transformation rates [346]. Hence, it was proposed that vaccines should also target non-capsulated pneumococci, as a way to lower the overall recombination rate of *S. pneumococcus* populations [347]. In the race toward a capsule-based *K. pneumoniae* vaccine [348], targeting ubiquitous surface antigens typically masked by the capsule could achieve the same goal.

While there was little overlap between multidrug resistant *vs*. hypervirulent lineages before the 2010s, recent studies have reported the emergence of carbapenem-resistant (CR) hypervirulent (Hv) strains of *K. pneumoniae* (KP) leading to high mortality rates [301]. Mechanisms for the emergence of CR-hvKP can be summarized in three patterns: (i) CR-KP acquiring a hypervirulent phenotype; (ii) hvKP acquiring a carbapenem-resistant phenotype; and (iii) classical KP acquiring both a carbapenem resistance and hypervirulence hybrid plasmid [286]. Genomics-based prediction have postulated that multi-drug resistant clones are more prone to new trait acquisitions, including virulence factors [290]. This is in line with our result that capsules of hypervirulent clones (K1 and K2) act as barriers to the donation and reception of conjugative plasmids [341]. However, hypervirulence seems to evolve in defined genetic backgrounds, and may not be transferable to every multidrug resistant clones, which are more diverse [282,290,295]. Overall, the convergence of hypervirulence and multi-drug resistance is the result of MGE transfer, and further studies should be conducted to address this evolutionary process.

Finally, MGEs themselves can modulate the expression and composition of the capsule, which may modify their own transfer rate. The transcriptional activators *rmpA, rmpB, rmpC* and *rmpD* (for

regulator of **m**ucoid **p**henotype) are frequently found on *K. pneumoniae* plasmids and ICEs, and modulate capsule gene expression, leading to increased capsule production or hyper-muco-viscosity [349–352]. Those transcriptional activators are considered virulence factors, and are associated with only a subset of serotypes, including K1 and K2 [295,353]. Virulence plasmids are typically only mobilizable in *K. pneumoniae* [286,295,299], however some self-conjugative plasmids carrying *rmpA* have been described. These plasmids harbor MPF_F conjugative systems, and this is in line with the observation that MPF_F plasmids seem to be less impacted by the capsule than other types [341]. While the *rmp* genes modify the expression of the capsule locus and are found on plasmids and ICES, one prophage of *Acinetobacter baumannii* has recently been shown to modify the structure of the capsule via an alternative, phage-encoded Wzy polymerase [354]. To note, increased capsule production is associated with phage resistance and probably lower conjugation rates, while capsule composition modification is associated with a shift in phage sensitivity and can also result in lower conjugation rates. In fact, one may speculate that MGE can leverage the capsule for their own benefit, by defending their host against other interfering MGEs. Further studies may address the hypothesis that capsule manipulation could represent an additional layer in the interplay between MGEs and their hosts.

References

1. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 1980;8: r49–r62.
2. Sørensen MA, Kurland CG, Pedersen S. Codon usage determines translation rate in *Escherichia coli*. J Mol Biol. 1989;207: 365–377. doi:10.1016/0022-2836(89)90260-x
3. Rocha EPC. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 2004;14: 2279–2286. doi:10.1101/gr.2896904
4. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci U S A. 2004;101: 3480–3485. doi:10.1073/pnas.0307827100
5. Francino MP, Ochman H. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. Mol Biol Evol. 2001;18: 1147–1150. doi:10.1093/oxfordjournals.molbev.a003888
6. Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol. 1996;13: 660–665. doi:10.1093/oxfordjournals.molbev.a025626
7. Rocha EPC. The replication-related organization of bacterial genomes. Microbiol Read Engl. 2004;150: 1609–1627. doi:10.1099/mic.0.26974-0
8. Wu CI. DNA strand asymmetry. Nature. 1991;352: 114. doi:10.1038/352114b0
9. Lobry JR, Sueoka N. Asymmetric directional mutation pressures in bacteria. Genome Biol. 2002;3: RESEARCH0058. doi:10.1186/gb-2002-3-10-research0058
10. Wu CI, Maeda N. Inequality in mutation rates of the two strands of DNA. Nature. 1987;327: 169–170. doi:10.1038/327169a0
11. Rocha EPC, Danchin A. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. 2003;31: 6570–6577. doi:10.1093/nar/gkg859
12. Couturier E, Rocha EPC. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol. 2006;59: 1506–1518. doi:10.1111/j.1365-2958.2006.05046.x
13. Daubin V, Perrière G. G+C3 structuring along the genome: a common feature in prokaryotes. Mol Biol Evol. 2003;20: 471–483. doi:10.1093/molbev/msg022
14. Sharp PM, Shields DC, Wolfe KH, Li WH. Chromosomal location and evolutionary rate variation in enterobacterial genes. Science. 1989;246: 808–810. doi:10.1126/science.2683084
15. Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 2010;6: e1001107. doi:10.1371/journal.pgen.1001107
16. Hershberg R, Petrov DA. Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genet. 2010;6: e1001115. doi:10.1371/journal.pgen.1001115
17. Abby S, Daubin V. Comparative genomics and the evolution of prokaryotes. Trends Microbiol. 2007;15: 135–141. doi:10.1016/j.tim.2007.01.007
18. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005;3: 722–732. doi:10.1038/nrmicro1235
19. Ge F, Wang L-S, Kim J. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 2005;3: e316. doi:10.1371/journal.pbio.0030316

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

21. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. Nat Rev Genet. 2015;16: 472–482. doi:10.1038/nrg3962

22. Lawrence JG, Retchless AC. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. Methods Mol Biol Clifton NJ. 2009;532: 29–53. doi:10.1007/978-1-60327-853-9_3

23. Rocha EPC. Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. Mol Biol Evol. 2018;35: 1338–1347. doi:10.1093/molbev/msy078

24. Smith GR. Homologous recombination in procaryotes. Microbiol Rev. 1988;52: 1–28. doi:10.1128/mr.52.1.1-28.1988

25. Hallet B, Sherratt DJ. Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. FEMS Microbiol Rev. 1997;21: 157–178. doi:10.1111/j.1574-6976.1997.tb00349.x

26. Stark WM, Boocock MR, Sherratt DJ. Catalysis by site-specific recombinases. Trends Genet TIG. 1992;8: 432–439.

27. del Solar G, Giraldo R, Ruiz-Echevarría MJ, Espinosa M, Díaz-Orejas R. Replication and Control of Circular Bacterial Plasmids. Microbiol Mol Biol Rev. 1998;62: 434–464.

28. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F. Mobility of plasmids. Microbiol Mol Biol Rev MMBR. 2010;74: 434–452. doi:10.1128/MMBR.00020-10

29. Griffith F. The Significance of Pneumococcal Types. J Hyg (Lond). 1928;27: 113–159. doi:10.1017/s0022172400031879

30. Avery OT, Macleod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. J Exp Med. 1944;79: 137–158. doi:10.1084/jem.79.2.137

31. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000;405: 299–304. doi:10.1038/35012500

32. Touchon M, Moura de Sousa JA, Rocha EP. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. Curr Opin Microbiol. 2017;38: 66–73. doi:10.1016/j.mib.2017.04.010

33. Treangen TJ, Rocha EPC. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLOS Genet. 2011;7: e1001284. doi:10.1371/journal.pgen.1001284

34. Diard M, Hardt W-D. Evolution of bacterial virulence. FEMS Microbiol Rev. 2017;41: 679–697. doi:10.1093/femsre/fux023

35. Nogueira T, Rankin DJ, Touchon M, Taddei F, Brown SP, Rocha EPC. Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. Curr Biol CB. 2009;19: 1683–1691. doi:10.1016/j.cub.2009.08.056

36. Ramirez MS, Traglia GM, Lin DL, Tran T, Tolmasky ME. Plasmid-Mediated Antibiotic Resistance and Virulence in Gram-Negatives: the *Klebsiella pneumoniae* Paradigm. Microbiol Spectr. 2014;2. doi:10.1128/microbiolspec.PLAS-0016-2013

37. von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S, van Alphen LB, et al. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. Front Microbiol. 2016;7: 173. doi:10.3389/fmicb.2016.00173

38. Johnston C, Martin B, Fichant G, Polard P, Claverys J-P. Bacterial transformation: distribution, shared mechanisms and divergent control. Nat Rev Microbiol. 2014;12: 181–196. doi:10.1038/nrmicro3199

39. Mell JC, Redfield RJ. Natural Competence and the Evolution of DNA Uptake Specificity. J Bacteriol. 2014;196: 1471–1483. doi:10.1128/JB.01293-13

40. Seitz P, Blokesch M. Cues and regulatory pathways involved in natural competence and transformation in pathogenic and environmental Gram-negative bacteria. FEMS Microbiol Rev. 2013;37: 336–363. doi:10.1111/j.1574-6976.2012.00353.x

41. Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD, Fraser C. Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. PLOS Biol. 2016;14: e1002394. doi:10.1371/journal.pbio.1002394

42. Redfield RJ. Do bacteria have sex? Nat Rev Genet. 2001;2: 634–639. doi:10.1038/35084593

43. Bernstein H, Byerly HC, Hopf FA, Michod RE. Genetic damage, mutation, and the evolution of sex. Science. 1985;229: 1277–1281. doi:10.1126/science.3898363

44. Salvadori G, Junges R, Morrison DA, Petersen FC. Competence in *Streptococcus pneumoniae* and Close Commensal Relatives: Mechanisms and Implications. Front Cell Infect Microbiol. 2019;9: 94. doi:10.3389/fcimb.2019.00094

45. Obergfell KP, Seifert HS. Mobile DNA in the Pathogenic Neisseria. Microbiol Spectr. 2015;3: MDNA3-0015–2014. doi:10.1128/microbiolspec.MDNA3-0015-2014

46. Chen I, Dubnau D. DNA uptake during bacterial transformation. Nat Rev Microbiol. 2004;2: 241–249. doi:10.1038/nrmicro844

47. Schwechheimer C, Kuehn MJ. Outer-membrane vesicles from Gram-negative bacteria: biogenesis and functions. Nat Rev Microbiol. 2015;13: 605–619. doi:10.1038/nrmicro3525

48. Dorward DW, Garon CF, Judd RC. Export and intercellular transfer of DNA via membrane blebs of Neisseria gonorrhoeae. J Bacteriol. 1989;171: 2499–2505. doi:10.1128/jb.171.5.2499-2505.1989

49. Fulsundar S, Harms K, Flaten GE, Johnsen PJ, Chopade BA, Nielsen KM. Gene Transfer Potential of Outer Membrane Vesicles of Acinetobacter baylyi and Effects of Stress on Vesiculation. Appl Environ Microbiol. 2014;80: 3469–3483. doi:10.1128/AEM.04248-13

50. Yaron S, Kolling GL, Simon L, Matthews KR. Vesicle-mediated transfer of virulence genes from *Escherichia coli* O157:H7 to other enteric bacteria. Appl Environ Microbiol. 2000;66: 4414–4420. doi:10.1128/AEM.66.10.4414-4420.2000

51. Tzipilevich E, Habusha M, Ben-Yehuda S. Acquisition of Phage Sensitivity by Bacteria through Exchange of Phage Receptors. Cell. 2017;168: 186-199.e12. doi:10.1016/j.cell.2016.12.003

52. Twort FW. An investigation on the nature of ultra-microscopic viruses. Acta Kravsi. 1961.

53. d'Herelle M. Sur un microbe invisible antagoniste des bacilles dysentériques. Acta Kravsi. 1917.

54. Luria SE, Delbrück M. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics. 1943;28: 491–511.

55. Luria SE. Mutations of Bacterial Viruses Affecting Their Host Range. Genetics. 1945;30: 84–99.

56. Freeman VJ. Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. J Bacteriol. 1951;61: 675–688. doi:10.1128/jb.61.6.675-688.1951

57. Zinder ND, Lederberg J. Genetic exchange in Salmonella. J Bacteriol. 1952;64: 679–699. doi:10.1128/jb.64.5.679-699.1952

58. Dion MB, Oechslin F, Moineau S. Phage diversity, genomics and phylogeny. Nat Rev Microbiol. 2020;18: 125–138. doi:10.1038/s41579-019-0311-5

59. Mäntynen S, Laanto E, Oksanen HM, Poranen MM, Díaz-Muñoz SL. Black box of phage-bacterium interactions: exploring alternative phage infection strategies. Open Biol. 2021;11: 210188. doi:10.1098/rsob.210188

60. Hay ID, Lithgow T. Filamentous phages: masters of a microbial sharing economy. EMBO Rep. 2019;20: e47427. doi:10.15252/embr.201847427

61. Onodera S, Olkkonen VM, Gottlieb P, Strassman J, Qiao XY, Bamford DH, et al. Construction of a transducing virus from double-stranded RNA bacteriophage phi6: establishment of carrier states in host cells. J Virol. 1992;66: 190–196. doi:10.1128/JVI.66.1.190-196.1992

62. Los M, Wegrzyn G, Neubauer P. A role for bacteriophage T4 rI gene function in the control of phage development during pseudolysogeny and in slowly growing host cells. Res Microbiol. 2003;154: 547–552. doi:10.1016/S0923-2508(03)00151-7

63. Cenens W, Makumi A, Mebrhatu MT, Lavigne R, Aertsen A. Phage-host interactions during pseudolysogeny: Lessons from the Pid/dgo interaction. Bacteriophage. 2013;3: e25029. doi:10.4161/bact.25029

64. Pfeifer E, Moura de Sousa JA, Touchon M, Rocha EPC. Bacteria have numerous distinctive groups of phage-plasmids with conserved phage and variable plasmid gene repertoires. Nucleic Acids Res. 2021;49: 2655–2673. doi:10.1093/nar/gkab064

65. Groth AC, Calos MP. Phage integrases: biology and applications. J Mol Biol. 2004;335: 667–678. doi:10.1016/j.jmb.2003.09.082

66. Campbell A. Comparative molecular biology of lambdoid phages. Annu Rev Microbiol. 1994;48: 193–222. doi:10.1146/annurev.mi.48.100194.001205

67. Canchaya C, Fournous G, Brüssow H. The impact of prophages on bacterial chromosomes. Mol Microbiol. 2004;53: 9–18. doi:10.1111/j.1365-2958.2004.04113.x

68. Taylor VL, Fitzpatrick AD, Islam Z, Maxwell KL. The Diverse Impacts of Phage Morons on Bacterial Fitness and Virulence. Adv Virus Res. 2019;103: 1–31. doi:10.1016/bs.aivir.2018.08.001

69. Waldor MK, Mekalanos JJ. Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. Science. 1996;272: 1910–1914. doi:10.1126/science.272.5270.1910

70. De Paepe M, Hutinet G, Son O, Amarir-Bouhram J, Schbath S, Petit M-A. Temperate Phages Acquire DNA from Defective Prophages by Relaxed Homologous Recombination: The Role of Rad52-Like Recombinases. PLoS Genet. 2014;10: e1004181. doi:10.1371/journal.pgen.1004181

71. Touchon M, Bernheim A, Rocha EP. Genetic and life-history traits associated with the distribution of prophages in bacteria. ISME J. 2016;10: 2744–2754. doi:10.1038/ismej.2016.47

72. De Paepe M, Tournier L, Moncaut E, Son O, Langella P, Petit M-A. Carriage of λ Latent Virus Is Costly for Its Bacterial Host due to Frequent Reactivation in Monoxenic Mouse Intestine. PLoS Genet. 2016;12: e1005861. doi:10.1371/journal.pgen.1005861

73. Morse ML, Lederberg EM, Lederberg J. Transduction in *Escherichia coli* K-12. Genetics. 1956;41: 142–156. doi:10.1093/genetics/41.1.142

74. Kleiner M, Bushnell B, Sanderson KE, Hooper LV, Duerkop BA. Transductomics: sequencing-based detection and analysis of transduced DNA in pure cultures and microbial communities. Microbiome. 2020;8: 158. doi:10.1186/s40168-020-00935-5

75. Lennox ES. Transduction of linked genetic characters of the host by bacteriophage P1. Virology. 1955;1: 190–206. doi:10.1016/0042-6822(55)90016-7

76. Garneau JR, Legrand V, Marbouty M, Press MO, Vik DR, Fortier L-C, et al. High-throughput identification of viral termini and packaging mechanisms in virome datasets using PhageTermVirome. Sci Rep. 2021;11: 18319. doi:10.1038/s41598-021-97867-3

77. Huang H, Masters M. Bacteriophage P1 pac sites inserted into the chromosome greatly increase packaging and transduction of *Escherichia coli* genomic DNA. Virology. 2014;468–470: 274–282. doi:10.1016/j.virol.2014.07.029

78. Chen J, Quiles-Puchalt N, Chiang YN, Bacigalupe R, Fillol-Salom A, Chee MSJ, et al. Genome hypermobility by lateral transduction. Science. 2018;362: 207–212. doi:10.1126/science.aat5867

79. Penadés JR, Christie GE. The Phage-Inducible Chromosomal Islands: A Family of Highly Evolved Molecular Parasites. Annu Rev Virol. 2015;2: 181–201. doi:10.1146/annurev-virology-031413-085446

80. Lindqvist BH, Dehò G, Calendar R. Mechanisms of genome propagation and helper exploitation by satellite phage P4. Microbiol Rev. 1993;57: 683–702. doi:10.1128/mr.57.3.683-702.1993

81. Novick RP, Ram G. Staphylococcal pathogenicity islands-movers and shakers in the genomic firmament. Curr Opin Microbiol. 2017;38: 197–204. doi:10.1016/j.mib.2017.08.001

82. Rousset F, Dowding J, Bernheim A, Rocha EPC, Bikard D. Prophage-encoded hotspots of bacterial immune systems. bioRxiv; 2021. p. 2021.01.21.427644. doi:10.1101/2021.01.21.427644

83. Fillol-Salom A, Martínez-Rubio R, Abdulrahman RF, Chen J, Davies R, Penadés JR. Phage-inducible chromosomal islands are ubiquitous within the bacterial universe. ISME J. 2018;12: 2114–2128. doi:10.1038/s41396-018-0156-3

84. Moura de Sousa JA, Rocha EPC. To catch a hijacker: abundance, evolution and genetic diversity of P4-like bacteriophage satellites. Philos Trans R Soc Lond B Biol Sci. 2022;377: 20200475. doi:10.1098/rstb.2020.0475

85. Marrs B. Genetic recombination in Rhodopseudomonas capsulata. Proc Natl Acad Sci U S A. 1974;71: 971–973. doi:10.1073/pnas.71.3.971

86. Barbian KD, Minnick MF. A bacteriophage-like particle from Bartonella bacilliformis. Microbiol Read Engl. 2000;146 (Pt 3): 599–609. doi:10.1099/00221287-146-3-599

87. Lang AS, Zhaxybayeva O, Beatty JT. Gene transfer agents: phage-like elements of genetic exchange. Nat Rev Microbiol. 2012;10: 472–482. doi:10.1038/nrmicro2802

88. Brimacombe CA, Stevens A, Jun D, Mercer R, Lang AS, Beatty JT. Quorum-sensing regulation of a capsular polysaccharide receptor for the Rhodobacter capsulatus gene transfer agent (RcGTA). Mol Microbiol. 2013;87: 802–817. doi:10.1111/mmi.12132

89. Haudiquet M, Sousa JM de, Touchon M, Rocha E. Selfish, promiscuous, and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. EcoEvoRxiv; 2021. doi:10.32942/osf.io/7t2jh

90. Lederberg J, Tatum EL. Gene recombination in *Escherichia coli*. Nature. 1946;158: 558. doi:10.1038/158558a0

91. Lederberg J, Cavalli LL, Lederberg EM. Sex Compatibility in *Escherichia coli*. Genetics. 1952;37: 720–730.

92. Taylor AL, Thoman MS. The Genetic Map of *Escherichia coli* K-12. Genetics. 1964;50: 659–677.

93. Burrus V, Marrero J, Waldor MK. The current ICE age: biology and evolution of SXT-related integrating conjugative elements. Plasmid. 2006;55: 173–183. doi:10.1016/j.plasmid.2006.01.001

94. Cury J, Oliveira PH, de la Cruz F, Rocha EPC. Host Range and Genetic Plasticity Explain the Coexistence of Integrative and Extrachromosomal Mobile Genetic Elements. Mol Biol Evol. 2018;35: 2230–2239. doi:10.1093/molbev/msy123

95. Carraro N, Burrus V. Biology of Three ICE Families: SXT/R391, ICEBs1, and ICESt1/ICESt3. Microbiol Spectr. 2014;2. doi:10.1128/microbiolspec.MDNA3-0008-2014

96. Draper O, César CE, Machón C, de la Cruz F, Llosa M. Site-specific recombinase and integrase activities of a conjugative relaxase in recipient cells. Proc Natl Acad Sci U S A. 2005;102: 16385–16390. doi:10.1073/pnas.0506081102

97. Guglielmini J, Quintais L, Garcillán-Barcia MP, de la Cruz F, Rocha EPC. The Repertoire of ICE in Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation. PLoS Genet. 2011;7: e1002222. doi:10.1371/journal.pgen.1002222

98. Cabezón E, Ripoll-Rozada J, Peña A, de la Cruz F, Arechaga I. Towards an integrated model of bacterial conjugation. FEMS Microbiol Rev. 2015;39: 81–95. doi:10.1111/1574-6976.12085

99. Guglielmini J, Néron B, Abby SS, Garcillán-Barcia MP, de la Cruz F, Rocha EPC. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. Nucleic Acids Res. 2014;42: 5715–5727. doi:10.1093/nar/gku194

100. Johnson CM, Grossman AD. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. Annu Rev Genet. 2015;49: 577–601. doi:10.1146/annurev-genet-112414-055018

101. Lawley TD, Klimke WA, Gubbins MJ, Frost LS. F factor conjugation is a true type IV secretion system. FEMS Microbiol Lett. 2003;224: 1–15. doi:10.1016/S0378-1097(03)00430-0

102. Hochhut B, Beaber JW, Woodgate R, Waldor MK. Formation of Chromosomal Tandem Arrays of the SXT Element and R391, Two Conjugative Chromosomally Integrating Elements That Share an Attachment Site. J Bacteriol. 2001;183: 1124–1132. doi:10.1128/JB.183.4.1124-1132.2001

103. Adelberg EA, Pittard J. CHROMOSOME TRANSFER IN BACTERIAL CONJUGATION. Bacteriol Rev. 1965;29: 161–172. doi:10.1128/br.29.2.161-172.1965

104. Haas D, Watson J, Krieg R, Leisinger T. Isolation of an Hfr donor of Pseudomonas aeruginosa PAO by insertion of the plasmid RP1 into the tryptophan synthase gene. Mol Gen Genet MGG. 1981;182: 240–244. doi:10.1007/BF00269664

105. Hochhut B, Marrero J, Waldor MK. Mobilization of plasmids and chromosomal DNA mediated by the SXT element, a constin found in Vibrio cholerae O139. J Bacteriol. 2000;182: 2043–2047. doi:10.1128/JB.182.7.2043-2047.2000

106. Coluzzi C, Guédon G, Devignes M-D, Ambroset C, Loux V, Lacroix T, et al. A Glimpse into the World of Integrative and Mobilizable Elements in Streptococci Reveals an Unexpected Diversity and Novel Families of Mobilization Proteins. Front Microbiol. 2017;8: 443. doi:10.3389/fmicb.2017.00443

107. Bertozzi Silva J, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. FEMS Microbiol Lett. 2016;363: fnw002. doi:10.1093/femsle/fnw002

108. Dowah ASA, Clokie MRJ. Review of the nature, diversity and structure of bacteriophage receptor binding proteins that target Gram-positive bacteria. Biophys Rev. 2018;10: 535–542. doi:10.1007/s12551-017-0382-3

109. Casjens SR, Molineux IJ. Short noncontractile tail machines: adsorption and DNA delivery by podoviruses. Adv Exp Med Biol. 2012;726: 143–179. doi:10.1007/978-1-4614-0980-9_7

110. González-García VA, Pulido-Cid M, Garcia-Doval C, Bocanegra R, van Raaij MJ, Martín-Benito J, et al. Conformational changes leading to T7 DNA delivery upon interaction with the bacterial receptor. J Biol Chem. 2015;290: 10038–10044. doi:10.1074/jbc.M114.614222

111. Hu B, Margolin W, Molineux IJ, Liu J. Structural remodeling of bacteriophage T4 and host membranes during infection initiation. Proc Natl Acad Sci U S A. 2015;112: E4919–E4928. doi:10.1073/pnas.1501064112

112. Washizaki A, Yonesaki T, Otsuka Y. Characterization of the interactions between *Escherichia coli* receptors, LPS and OmpC, and bacteriophage T4 long tail fibers. MicrobiologyOpen. 2016;5: 1003–1015. doi:10.1002/mbo3.384

113. Ross A, Ward S, Hyman P. More Is Better: Selecting for Broad Host Range Bacteriophages. Front Microbiol. 2016;7: 1352. doi:10.3389/fmicb.2016.01352

114. Holtzman T, Globus R, Molshanski-Mor S, Ben-Shem A, Yosef I, Qimron U. A continuous evolution system for contracting the host range of bacteriophage T7. Sci Rep. 2020;10: 307. doi:10.1038/s41598-019-57221-0

115. Yosef I, Goren MG, Globus R, Molshanski-Mor S, Qimron U. Extending the Host Range of Bacteriophage Particles for DNA Transduction. Mol Cell. 2017;66: 721-728.e3. doi:10.1016/j.molcel.2017.04.025

116. Haggård-Ljungquist E, Halling C, Calendar R. DNA sequences of the tail fiber genes of bacteriophage P2: evidence for horizontal transfer of tail fiber genes among unrelated bacteriophages. J Bacteriol. 1992;174: 1462–1477.

117. Lenski RE, Levin BR. Constraints on the Coevolution of Bacteria and Virulent Phage: A Model, Some Experiments, and Predictions for Natural Communities. Am Nat. 1985;125: 585–602.

118. Mostowy RJ, Holt KE. Diversity-Generating Machines: Genetics of Bacterial Sugar-Coating. Trends Microbiol. 2018;26: 1008–1021. doi:10.1016/j.tim.2018.06.006

119. Weinbauer MG, Rassoulzadegan F. Are viruses driving microbial diversification and diversity? Environ Microbiol. 2004;6: 1–11. doi:10.1046/j.1462-2920.2003.00539.x

120. Bondy-Denomy J, Qian J, Westra ER, Buckling A, Guttman DS, Davidson AR, et al. Prophages mediate defense against phage infection through diverse mechanisms. ISME J. 2016;10: 2854–2866. doi:10.1038/ismej.2016.79

121. Shi K, Oakland JT, Kurniawan F, Moeller NH, Banerjee S, Aihara H. Structural basis of superinfection exclusion by bacteriophage T4 Spackle. Commun Biol. 2020;3: 691. doi:10.1038/s42003-020-01412-3

122. Taylor VL, Udaskin ML, Islam ST, Lam JS. The D3 bacteriophage α-polymerase inhibitor (Iap) peptide disrupts O-antigen biosynthesis through mimicry of the chain length regulator Wzz in Pseudomonas aeruginosa. J Bacteriol. 2013;195: 4735–4741. doi:10.1128/JB.00903-13

123. Uetake H, Luria SE, Burrous JW. Conversion of somatic antigens in Salmonella by phage infection leading to lysis or lysogeny. Virology. 1958;5: 68–91. doi:10.1016/0042-6822(58)90006-0

124. Cook L, Chatterjee A, Barnes A, Yarwood J, Hu W-S, Dunny G. Biofilm growth alters regulation of conjugation by a bacterial pheromone. Mol Microbiol. 2011;81: 1499–1510. doi:10.1111/j.1365-2958.2011.07786.x

125. Molin S, Tolker-Nielsen T. Gene transfer occurs with enhanced efficiency in biofilms and induces enhanced stabilisation of the biofilm structure. Curr Opin Biotechnol. 2003;14: 255–261. doi:10.1016/s0958-1669(03)00036-3

126. Ghigo JM. Natural conjugative plasmids induce bacterial biofilm development. Nature. 2001;412: 442–445. doi:10.1038/35086581

127. Gordon JE, Christie PJ. The Agrobacterium Ti Plasmids. Microbiol Spectr. 2014;2. doi:10.1128/microbiolspec.PLAS-0010-2013

128. Soltysiak MPM, Meaney RS, Hamadache S, Janakirama P, Edgell DR, Karas BJ. Trans-Kingdom Conjugation within Solid Media from *Escherichia coli* to Saccharomyces cerevisiae. Int J Mol Sci. 2019;20: 5212. doi:10.3390/ijms20205212

129. Ishiwa A, Komano T. Thin pilus PilV adhesins of plasmid R64 recognize specific structures of the lipopolysaccharide molecules of recipient cells. J Bacteriol. 2003;185: 5192–5199. doi:10.1128/JB.185.17.5192-5199.2003

130. Johnson CM, Grossman AD. Identification of host genes that affect acquisition of an integrative and conjugative element in Bacillus subtilis. Mol Microbiol. 2014;93: 1284–1301. doi:10.1111/mmi.12736

131. Pérez-Mendoza D, de la Cruz F. *Escherichia coli* genes affecting recipient ability in plasmid conjugation: Are there any? BMC Genomics. 2009;10: 71. doi:10.1186/1471-2164-10-71

132. Johnson CM, Grossman AD. The Composition of the Cell Envelope Affects Conjugation in Bacillus subtilis. J Bacteriol. 2016;198: 1241–1249. doi:10.1128/JB.01044-15

133. Dahmane N, Robert E, Deschamps J, Meylheuc T, Delorme C, Briandet R, et al. Impact of Cell Surface Molecules on Conjugative Transfer of the Integrative and Conjugative Element ICESt3 of Streptococcus thermophilus. Appl Environ Microbiol. 2018;84. doi:10.1128/AEM.02109-17

134. Stuy JH. Plasmid transfer in Haemophilus influenzae. J Bacteriol. 1979;139: 520–529. doi:10.1128/JB.139.2.520-529.1979

135. Audette GF, Manchak J, Beatty P, Klimke WA, Frost LS. Entry exclusion in F-like plasmids requires intact TraG in the donor that recognizes its cognate TraS in the recipient. Microbiol Read Engl. 2007;153: 442–451. doi:10.1099/mic.0.2006/001917-0

136. Jalajakumari MB, Guidolin A, Buhk HJ, Manning PA, Ham LM, Hodgson AL, et al. Surface exclusion genes traS and traT of the F sex factor of *Escherichia coli* K-12. Determination of the nucleotide sequence and promoter and terminator activities. J Mol Biol. 1987;198: 1–11. doi:10.1016/0022-2836(87)90452-9

137. Harrison JL, Taylor IM, Platt K, O'Connor CD. Surface exclusion specificity of the TraT lipoprotein is determined by single alterations in a five-amino-acid region of the protein. Mol Microbiol. 1992;6: 2825–2832. doi:10.1111/j.1365-2958.1992.tb01462.x

138. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science. 2018;359: eaar4120. doi:10.1126/science.aar4120

139. Murphy KC. Lambda Gam protein inhibits the helicase and chi-stimulated recombination activities of *Escherichia coli* RecBCD enzyme. J Bacteriol. 1991;173: 5808–5821. doi:10.1128/jb.173.18.5808-5821.1991

140. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010;8: 317–327. doi:10.1038/nrmicro2315

141. Oliveira PH, Touchon M, Rocha EPC. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res. 2014;42: 10618–10631. doi:10.1093/nar/gku734

142. Lopatina A, Tal N, Sorek R. Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy. Annu Rev Virol. 2020;7: 371–384. doi:10.1146/annurev-virology-011620-040628

143. Marraffini LA. CRISPR-Cas immunity in prokaryotes. Nature. 2015;526: 55–61. doi:10.1038/nature15386

144. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science. 2008;322: 1843–1845. doi:10.1126/science.1165771

145. Watson BNJ, Staals RHJ, Fineran PC. CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene Transfer by Transduction. mBio. 2018;9: e02406-17. doi:10.1128/mBio.02406-17

146. Bernheim A, Sorek R. The pan-immune system of bacteria: antiviral defence as a community resource. Nat Rev Microbiol. 2020;18: 113–119. doi:10.1038/s41579-019-0278-2

147. Rocha EPC, Bikard D. Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? PLoS Biol. 2022;20: e3001514. doi:10.1371/journal.pbio.3001514

148. Penner M, Morad I, Snyder L, Kaufmann G. Phage T4-coded Stp: double-edged effector of coupled DNA and tRNA-restriction systems. J Mol Biol. 1995;249: 857–868. doi:10.1006/jmbi.1995.0343

149. Spies M, Fishel R. Mismatch Repair during Homologous and Homeologous Recombination. Cold Spring Harb Perspect Biol. 2015;7: a022657. doi:10.1101/cshperspect.a022657

150. Matic I, Rayssiguier C, Radman M. Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. Cell. 1995;80: 507–515. doi:10.1016/0092-8674(95)90501-4

151. Curcio MJ, Derbyshire KM. The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol. 2003;4: 865–877. doi:10.1038/nrm1241

152. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. 2014;38: 865–891. doi:10.1111/1574-6976.12067

153. Rice PA, Baker TA. Comparative architecture of transposase and integrase complexes. Nat Struct Biol. 2001;8: 302–307. doi:10.1038/86166

154. Siguier P, Filée J, Chandler M. Insertion sequences in prokaryotic genomes. Curr Opin Microbiol. 2006;9: 526–531. doi:10.1016/j.mib.2006.08.005

155. Hawkey J, Cottingham H, Tokolyi A, Wick RR, Judd LM, Cerdeira L, et al. Linear plasmids in Klebsiella and other Enterobacteriaceae. Microb Genomics. 2022;8. doi:10.1099/mgen.0.000807

156. Baxter JC, Funnell BE. Plasmid Partition Mechanisms. Microbiol Spectr. 2014;2. doi:10.1128/microbiolspec.PLAS-0023-2014

157. Jain A, Srivastava P. Broad host range plasmids. FEMS Microbiol Lett. 2013;348: 87–96. doi:10.1111/1574-6968.12241

158. Lilly J, Camps M. Mechanisms of Theta Plasmid Replication. Microbiol Spectr. 2015;3: PLAS-0029-2014.

159. Konieczny I, Bury K, Wawrzycka A, Wegrzyn K. Iteron Plasmids. Microbiol Spectr. 2014;2. doi:10.1128/microbiolspec.PLAS-0026-2014

160. Vocke C, Bastia D. Primary structure of the essential replicon of the plasmid pSC101. Proc Natl Acad Sci U S A. 1983;80: 6557–6561.

161. Rawlings DE, Tietze E. Comparative Biology of IncQ and IncQ-Like Plasmids. Microbiol Mol Biol Rev. 2001;65: 481–496. doi:10.1128/MMBR.65.4.481-496.2001

162. del Solar G, Moscoso M, Espinosa M. Rolling circle-replicating plasmids from gram-positive and gram-negative bacteria: a wall falls. Mol Microbiol. 1993;8: 789–796. doi:10.1111/j.1365-2958.1993.tb01625.x

163. Novick RP. Plasmid incompatibility. Microbiol Rev. 1987;51: 381–395. doi:10.1128/mr.51.4.381-395.1987

164. Austin S, Nordström K. Partition-mediated incompatibility of bacterial plasmids. Cell. 1990;60: 351–354. doi:10.1016/0092-8674(90)90584-2

165. Thomas CM. Plasmid Incompatibility. In: Bell E, editor. Molecular Life Sciences: An Encyclopedic Reference. New York, NY: Springer; 2021. pp. 1–3. doi:10.1007/978-1-4614-6436-5_565-2

166. Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. Trends Genet TIG. 2001;17: 589–596. doi:10.1016/s0168-9525(01)02447-7

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

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

169. Williams KP. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res. 2002;30: 866–875. doi:10.1093/nar/30.4.866

170. Bellanger X, Morel C, Gonot F, Puymege A, Decaris B, Guédon G. Site-specific accretion of an integrative conjugative element together with a related genomic island leads to cis mobilization and gene capture. Mol Microbiol. 2011;81: 912–925. doi:10.1111/j.1365-2958.2011.07737.x

171. Pavlovic G, Burrus V, Gintz B, Decaris B, Guédon G. Evolution of genomic islands by deletion and tandem accretion by site-specific recombination: ICESt1-related elements from Streptococcus thermophilus. Microbiol Read Engl. 2004;150: 759–774. doi:10.1099/mic.0.26883-0

172. Bobay L-M, Rocha EPC, Touchon M. The Adaptation of Temperate Bacteriophages to Their Host Genomes. Mol Biol Evol. 2013;30: 737–751. doi:10.1093/molbev/mss279

173. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. Proc Natl Acad Sci U S A. 2015;112: E3574-3581. doi:10.1073/pnas.1501049112

174. Mostowy RJ, Croucher NJ, De Maio N, Chewapreecha C, Salter SJ, Turner P, et al. Pneumococcal Capsule Synthesis Locus cps as Evolutionary Hotspot with Potential to Generate Novel Serotypes by Recombination. Mol Biol Evol. 2017;34: 2537–2554. doi:10.1093/molbev/msx173

175. Wyres KL, Gorrie C, Edwards DJ, Wertheim HFL, Hsu LY, Van Kinh N, et al. Extensive Capsule Locus Variation and Large-Scale Genomic Recombination within the *Klebsiella pneumoniae* Clonal Group 258. Genome Biol Evol. 2015;7: 1267–1279. doi:10.1093/gbe/evv062

176. Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, et al. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A. 2002;99: 17020–17024. doi:10.1073/pnas.252529799

177. Jackson RW, Vinatzer B, Arnold DL, Dorus S, Murillo J. The influence of the accessory genome on bacterial pathogen evolution. Mob Genet Elem. 2011;1: 55–65. doi:10.4161/mge.1.1.16432

178. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017;2: 17040. doi:10.1038/nmicrobiol.2017.40

179. Lerat E, Daubin V, Moran NA. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. PLoS Biol. 2003;1: E19. doi:10.1371/journal.pbio.0000019

180. Daubin V, Moran NA, Ochman H. Phylogenetics and the cohesion of bacterial genomes. Science. 2003;301: 829–832. doi:10.1126/science.1086568

181. Stott CM, Bobay L-M. Impact of homologous recombination on core genome phylogenies. BMC Genomics. 2020;21: 829. doi:10.1186/s12864-020-07262-x

182. Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. The Ecology and Evolution of Pangenomes. Curr Biol CB. 2019;29: R1094–R1103. doi:10.1016/j.cub.2019.08.012

183. Kimura M. The neutral theory of molecular evolution and the world view of the neutralists. Genome. 1989;31: 24–31. doi:10.1139/g89-009

184. Ohta T. The Nearly Neutral Theory of Molecular Evolution. Annu Rev Ecol Syst. 1992;23: 263–286.

185. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature. 2018;557: 503–509. doi:10.1038/s41586-018-0124-0

186. Gil R, Sabater-Muñoz B, Latorre A, Silva FJ, Moya A. Extreme genome reduction in Buchnera spp.: toward the minimal genome needed for symbiotic life. Proc Natl Acad Sci U S A. 2002;99: 4454–4458. doi:10.1073/pnas.062067299

187. Veyrier FJ, Dufort A, Behr MA. The rise and fall of the Mycobacterium tuberculosis genome. Trends Microbiol. 2011;19: 156–161. doi:10.1016/j.tim.2010.12.008

188. Groisman EA, Ochman H. Pathogenicity islands: bacterial evolution in quantum leaps. Cell. 1996;87: 791–794. doi:10.1016/s0092-8674(00)81985-6

189. Pang TY, Lercher MJ. Each of 3,323 metabolic innovations in the evolution of *E. coli* arose through the horizontal transfer of a single DNA segment. Proc Natl Acad Sci U S A. 2019;116: 187–192. doi:10.1073/pnas.1718997115

190. Dimitriu T, Lotton C, Bénard-Capelle J, Misevic D, Brown SP, Lindner AB, et al. Genetic information transfer promotes cooperation in bacteria. Proc Natl Acad Sci U S A. 2014;111: 11103–11108. doi:10.1073/pnas.1406840111

191. Rankin DJ, Rocha EPC, Brown SP. What traits are carried on mobile genetic elements, and why? Heredity. 2011;106: 1–10. doi:10.1038/hdy.2010.24

192. Mc Ginty SE, Rankin DJ, Brown SP. Horizontal gene transfer and the evolution of bacterial cooperation. Evol Int J Org Evol. 2011;65: 21–32. doi:10.1111/j.1558-5646.2010.01121.x

193. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proc Natl Acad Sci U S A. 2016;113: 5970–5975. doi:10.1073/pnas.1521291113

194. Chistoserdova L, Vorholt JA, Thauer RK, Lidstrom ME. C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic Archaea. Science. 1998;281: 99–102. doi:10.1126/science.281.5373.99

195. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, et al. Comparative genomics of Listeria species. Science. 2001;294: 849–852. doi:10.1126/science.1063447

196. Prokop A, Gouin E, Villiers V, Nahori M-A, Vincentelli R, Duval M, et al. OrfX, a Nucleomodulin Required for Listeria monocytogenes Virulence. mBio. 2017;8: e01550-17. doi:10.1128/mBio.01550-17

197. Johnson J, Jinneman K, Stelma G, Smith BG, Lye D, Messer J, et al. Natural Atypical Listeria innocua Strains with Listeria monocytogenes Pathogenicity Island 1 Genes. Appl Environ Microbiol. 2004;70: 4256–4266. doi:10.1128/AEM.70.7.4256-4266.2004

198. Botelho J, Schulenburg H. The Role of Integrative and Conjugative Elements in Antibiotic Resistance Evolution. Trends Microbiol. 2021;29: 8–18. doi:10.1016/j.tim.2020.05.011

199. Navon-Venezia S, Kondratyeva K, Carattoli A. *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. FEMS Microbiol Rev. 2017;41: 252–275. doi:10.1093/femsre/fux013

200. Grohmann E, Muth G, Espinosa M. Conjugative plasmid transfer in gram-positive bacteria. Microbiol Mol Biol Rev MMBR. 2003;67: 277–301, table of contents. doi:10.1128/MMBR.67.2.277-301.2003

201. Waters VL. Conjugative transfer in the dissemination of beta-lactam and aminoglycoside resistance. Front Biosci J Virtual Libr. 1999;4: D433-456. doi:10.2741/waters

202. Wein T, Hülter NF, Mizrahi I, Dagan T. Emergence of plasmid stability under non-selective conditions maintains antibiotic resistance. Nat Commun. 2019;10: 2595. doi:10.1038/s41467-019-10600-7

203. Crespi BJ. The evolution of social behavior in microorganisms. Trends Ecol Evol. 2001;16: 178–183. doi:10.1016/s0169-5347(01)02115-2

204. Lee IPA, Eldakar OT, Gogarten JP, Andam CP. Bacterial cooperation through horizontal gene transfer. Trends Ecol Evol. 2022;37: 223–232. doi:10.1016/j.tree.2021.11.006

205. Hamilton WD. The Evolution of Altruistic Behavior. Am Nat. 1963;97: 354–356. doi:10.1086/497114

206. Griffin AS, West SA, Buckling A. Cooperation and competition in pathogenic bacteria. Nature. 2004;430: 1024–1027. doi:10.1038/nature02744

207. Dewar AE, Thomas JL, Scott TW, Wild G, Griffin AS, West SA, et al. Plasmids do not consistently stabilize cooperation across bacteria but may promote broad pathogen host-range. Nat Ecol Evol. 2021;5: 1624–1636. doi:10.1038/s41559-021-01573-2

208. Whitfield C, Wear SS, Sande C. Assembly of Bacterial Capsular Polysaccharides and Exopolysaccharides. Annu Rev Microbiol. 2020;74: 521–543. doi:10.1146/annurev-micro-011420-075607

209. Rendueles O, Garcia-Garcerà M, Néron B, Touchon M, Rocha EPC. Abundance and co-occurrence of extracellular capsules increase environmental breadth: Implications for the emergence of pathogens. PLOS Pathog. 2017;13: e1006525. doi:10.1371/journal.ppat.1006525

210. Rice LB. Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE. J Infect Dis. 2008;197: 1079–1081. doi:10.1086/533452

211. Ørskov I. Serological investigations in the Klebsiella group. I. New capsule types. Acta Pathol Microbiol Scand. 1955;36: 449–453. doi:10.1111/j.1699-0463.1955.tb04640.x

212. Tipton KA, Chin C-Y, Farokhyfar M, Weiss DS, Rather PN. Role of Capsule in Resistance to Disinfectants, Host Antimicrobials, and Desiccation in Acinetobacter baumannii. Antimicrob Agents Chemother. 2018;62. doi:10.1128/AAC.01188-18

213. Plante CJ, Shriver AG. Differential lysis of sedimentary bacteria by Arenicola marina L.: examination of cell wall structure and exopolymeric capsules as correlates. J Exp Mar Biol Ecol. 1998;229: 35–52. doi:10.1016/S0022-0981(98)00039-2

214. Scholl D, Adhya S, Merril C. *Escherichia coli* K1's capsule is a barrier to bacteriophage T7. Appl Environ Microbiol. 2005;71: 4872–4874. doi:10.1128/AEM.71.8.4872-4874.2005

215. Wilkinson BJ, Holmes KM. Staphylococcus aureus cell surface: capsule as a barrier to bacteriophage adsorption. Infect Immun. 1979;23: 549–552.

216. Kostina E, Ofek I, Crouch E, Friedman R, Sirota L, Klinger G, et al. Noncapsulated *Klebsiella pneumoniae* bearing mannose-containing O antigens is rapidly eradicated from mouse lung and triggers cytokine production by macrophages following opsonization with surfactant protein D. Infect Immun. 2005;73: 8282–8290. doi:10.1128/IAI.73.12.8282-8290.2005

217. Hyams C, Camberlein E, Cohen JM, Bax K, Brown JS. The Streptococcuspneumoniae Capsule Inhibits Complement Activity and Neutrophil Phagocytosis by Multiple Mechanisms. Infect Immun. 2010;78: 704–715. doi:10.1128/IAI.00881-09

218. Campos MA, Vargas MA, Regueiro V, Llompart CM, Albertí S, Bengoechea JA. Capsule polysaccharide mediates bacterial resistance to antimicrobial peptides. Infect Immun. 2004;72: 7107–7114. doi:10.1128/IAI.72.12.7107-7114.2004

219. Cress BF, Englaender JA, He W, Kasper D, Linhardt RJ, Koffas MAG. Masquerading microbial pathogens: capsular polysaccharides mimic host-tissue molecules. FEMS Microbiol Rev. 2014;38: 660–697. doi:10.1111/1574-6976.12056

220. Drummelsmith J, Whitfield C. Gene products required for surface expression of the capsular form of the group 1 K antigen in *Escherichia coli* (O9a:K30). Mol Microbiol. 1999;31: 1321–1332. doi:10.1046/j.1365-2958.1999.01277.x

221. Yother J. Capsules of *Streptococcus pneumoniae* and other bacteria: paradigms for polysaccharide biosynthesis and regulation. Annu Rev Microbiol. 2011;65: 563–581. doi:10.1146/annurev.micro.62.081307.162944

222. Rahn A, Drummelsmith J, Whitfield C. Conserved organization in the cps gene clusters for expression of *Escherichia coli* group 1 K antigens: relationship to the colanic acid biosynthesis locus and the cps genes from *Klebsiella pneumoniae*. J Bacteriol. 1999;181: 2307–2313. doi:10.1128/JB.181.7.2307-2313.1999

223. Bushell SR, Mainprize IL, Wear MA, Lou H, Whitfield C, Naismith JH. Wzi Is an Outer Membrane Lectin that Underpins Group 1 Capsule Assembly in *Escherichia coli*. Struct England1993. 2013;21: 844–853. doi:10.1016/j.str.2013.03.010

224. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, et al. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet. 2006;2: e31. doi:10.1371/journal.pgen.0020031

225. Pazzani C, Rosenow C, Boulnois GJ, Bronner D, Jann K, Roberts IS. Molecular analysis of region 1 of the *Escherichia coli* K5 antigen gene cluster: a region encoding proteins involved in cell surface expression of capsular polysaccharide. J Bacteriol. 1993;175: 5978–5983. doi:10.1128/jb.175.18.5978-5983.1993

226. Petit C, Rigg GP, Pazzani C, Smith A, Sieberth V, Stevens M, et al. Region 2 of the *Escherichia coli* K5 capsule gene cluster encoding proteins for the biosynthesis of the K5 polysaccharide. Mol Microbiol. 1995;17: 611–620. doi:10.1111/j.1365-2958.1995.mmi_17040611.x

227. Orskov I, Nyman K. Genetic mapping of the antigenic determinants of two polysaccharide K antigens, K10 and K54, in *Escherichia coli*. J Bacteriol. 1974;120: 43–51. doi:10.1128/jb.120.1.43-51.1974

228. Amor PA, Whitfield C. Molecular and functional analysis of genes required for expression of group IB K antigens in *Escherichia coli*: characterization of the his-region containing gene clusters for multiple

cell-surface polysaccharides. Mol Microbiol. 1997;26: 145–161. doi:10.1046/j.1365-2958.1997.5631930.x

229. Whitfield C. Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli*. Annu Rev Biochem. 2006;75: 39–68. doi:10.1146/annurev.biochem.75.103004.142545

230. Sørensen UB, Henrichsen J, Chen HC, Szu SC. Covalent linkage between the capsular polysaccharide and the cell wall peptidoglycan of *Streptococcus pneumoniae* revealed by immunochemical methods. Microb Pathog. 1990;8: 325–334. doi:10.1016/0882-4010(90)90091-4

231. Holt KE, Lassalle F, Wyres KL, Wick R, Mostowy RJ. Diversity and evolution of surface polysaccharide synthesis loci in Enterobacteriales. ISME J. 2020;14: 1713–1730. doi:10.1038/s41396-020-0628-0

232. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. Identification of Klebsiella capsule synthesis loci from whole genome data. Microb Genomics. 2016;2. doi:10.1099/mgen.0.000102

233. Porter NT, Canales P, Peterson DA, Martens EC. A Subset of Polysaccharide Capsules in the Human Symbiont Bacteroides thetaiotaomicron Promote Increased Competitive Fitness in the Mouse Gut. Cell Host Microbe. 2017;22: 494-506.e8. doi:10.1016/j.chom.2017.08.020

234. Pan Y-J, Lin T-L, Chen C-T, Chen Y-Y, Hsieh P-F, Hsu C-R, et al. Genetic analysis of capsular polysaccharide synthesis gene clusters in 79 capsular types of Klebsiella spp. Sci Rep. 2015;5: 15573. doi:10.1038/srep15573

235. Kapatai G, Sheppard CL, Al-Shahib A, Litt DJ, Underwood AP, Harrison TG, et al. Whole genome sequencing of *Streptococcus pneumoniae*: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. PeerJ. 2016;4: e2477. doi:10.7717/peerj.2477

236. Bessonov K, Laing C, Robertson J, Yong I, Ziebell K, Gannon VPJ, et al. ECTyper: in silico *Escherichia coli* serotype and species prediction from raw and assembled whole-genome sequence data. Microb Genomics. 2021;7. doi:10.1099/mgen.0.000728

237. Wyres KL, Cahill SM, Holt KE, Hall RM, Kenyon JJ. Identification of Acinetobacter baumannii loci for capsular polysaccharide (KL) and lipooligosaccharide outer core (OCL) synthesis in genome assemblies using curated reference databases compatible with Kaptive. Microb Genomics. 2020;6: e000339. doi:10.1099/mgen.0.000339

238. Orskov F, Orskov I. The serology of capsular antigens. Curr Top Microbiol Immunol. 1990;150: 43–63.

239. Dutton GG, Parolis H, Joseleau JP, Marais MF. The use of bacteriophage depolymerization in the structural investigation of the capsular polysaccharide from Klebsiella serotype K3. Carbohydr Res. 1986;149: 411–423. doi:10.1016/s0008-6215(00)90061-2

240. Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, et al. Pneumococcal Capsules and Their Types: Past, Present, and Future. Clin Microbiol Rev. 2015;28: 871–899. doi:10.1128/CMR.00024-15

241. Lam MMC, Wick RR, Judd LM, Holt KE, Wyres KL. Kaptive 2.0: updated capsule and lipopolysaccharide locus typing for the *Klebsiella pneumoniae* species complex. Microb Genomics. 2022;8. doi:10.1099/mgen.0.000800

242. Wick RR, Heinz E, Holt KE, Wyres KL. Kaptive Web: User-Friendly Capsule and Lipopolysaccharide Serotype Prediction for Klebsiella Genomes. J Clin Microbiol. 2018;56. doi:10.1128/JCM.00197-18

243. Andam CP, Hanage WP. Mechanisms of genome evolution of Streptococcus. Infect Genet Evol. 2015;33: 334–342. doi:10.1016/j.meegid.2014.11.007

244. Bradshaw JL, Rafiqullah IM, Robinson DA, McDaniel LS. Transformation of nonencapsulated *Streptococcus pneumoniae* during systemic infection. Sci Rep. 2020;10. doi:10.1038/s41598-020-75988-5

245. Dorman MJ, Feltwell T, Goulding DA, Parkhill J, Short FL. The Capsule Regulatory Network of *Klebsiella pneumoniae* Defined by density-TraDISort. Chang Y-F, editor. mBio. 2018;9: e01863-18. doi:10.1128/mBio.01863-18

246. Ernst CM, Braxton JR, Rodriguez-Osorio CA, Zagieboylo AP, Li L, Pironti A, et al. Adaptive evolution of virulence and persistence in carbapenem-resistant *Klebsiella pneumoniae*. Nat Med. 2020;26: 705–711. doi:10.1038/s41591-020-0825-4

247. Hsu C-R, Liao C-H, Lin T-L, Yang H-R, Yang F-L, Hsieh P-F, et al. Identification of a capsular variant and characterization of capsular acetylation in *Klebsiella pneumoniae* PLA-associated type K57. Sci Rep. 2016;6: 31946. doi:10.1038/srep31946

248. Croucher NJ, Kagedan L, Thompson CM, Parkhill J, Bentley SD, Finkelstein JA, et al. Selective and Genetic Constraints on Pneumococcal Serotype Switching. PLoS Genet. 2015;11. doi:10.1371/journal.pgen.1005095

249. Swartley JS, Marfin AA, Edupuganti S, Liu LJ, Cieslak P, Perkins B, et al. Capsule switching of Neisseria meningitidis. Proc Natl Acad Sci U S A. 1997;94: 271–276. doi:10.1073/pnas.94.1.271

250. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, Liñares J, et al. Pneumococcal Capsular Switching: A Historical Perspective. J Infect Dis. 2013;207: 439–449. doi:10.1093/infdis/jis703

251. Liu B, Park S, Thompson CD, Li X, Lee JC. Antibodies to Staphylococcus aureus capsular polysaccharides 5 and 8 perform similarly in vitro but are functionally distinct in vivo. Virulence. 2017;8: 859–874. doi:10.1080/21505594.2016.1270494

252. Chiarelli A, Cabanel N, Rosinski-Chupin I, Zongo PD, Naas T, Bonnin RA, et al. Diversity of mucoid to non-mucoid switch among carbapenemase-producing *Klebsiella pneumoniae*. BMC Microbiol. 2020;20: 325. doi:10.1186/s12866-020-02007-y

253. Pai VB, Heyneman CA, Erramouspe J. Conjugated heptavalent pneumococcal vaccine. Ann Pharmacother. 2002;36: 1403–1413. doi:10.1345/aph.1A048

254. Hurley D, Griffin C, Young M, Scott DA, Pride MW, Scully IL, et al. Safety, Tolerability, and Immunogenicity of a 20-Valent Pneumococcal Conjugate Vaccine (PCV20) in Adults 60 to 64 Years of Age. Clin Infect Dis Off Publ Infect Dis Soc Am. 2021;73: e1489–e1497. doi:10.1093/cid/ciaa1045

255. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013;45: 656–663. doi:10.1038/ng.2625

256. Chaguza C, Cornick JE, Andam CP, Gladstone RA, Alaerts M, Musicha P, et al. Population genetic structure, antibiotic resistance, capsule switching and evolution of invasive pneumococci before conjugate vaccination in Malawi. Vaccine. 2017;35: 4594–4602. doi:10.1016/j.vaccine.2017.07.009

257. Mutalik VK, Adler BA, Rishi HS, Piya D, Zhong C, Koskella B, et al. High-throughput mapping of the phage resistance landscape in *E. coli*. PLoS Biol. 2020;18: e3000877. doi:10.1371/journal.pbio.3000877

258. Rousset F, Cui L, Siouve E, Becavin C, Depardieu F, Bikard D. Genome-wide CRISPR-dCas9 screens in *E. coli* identify essential genes and phage host factors. PLOS Genet. 2018;14: e1007749. doi:10.1371/journal.pgen.1007749

259. Soundararajan M, von Bünau R, Oelschlaeger TA. K5 Capsule and Lipopolysaccharide Are Important in Resistance to T4 Phage Attack in Probiotic *E. coli* Strain Nissle 1917. Front Microbiol. 2019;10: 2783. doi:10.3389/fmicb.2019.02783

260. Nesper J, Hill CMD, Paiment A, Harauz G, Beis K, Naismith JH, et al. Translocation of group 1 capsular polysaccharide in *Escherichia coli* serotype K30. Structural and functional analysis of the outer membrane lipoprotein Wza. J Biol Chem. 2003;278: 49763–49772. doi:10.1074/jbc.M308775200

261. Bernheimer HP, Tiraby JG. Inhibition of phage infection by pneumococcus capsule. Virology. 1976;73: 308–309. doi:10.1016/0042-6822(76)90085-4

262. Porter NT, Hryckowian AJ, Merrill BD, Fuentes JJ, Gardner JO, Glowacki RWP, et al. Phase-variable capsular polysaccharides and lipoproteins modify bacteriophage susceptibility in Bacteroides thetaiotaomicron. Nat Microbiol. 2020;5: 1170–1181. doi:10.1038/s41564-020-0746-5

263. Gupta DS, Jann B, Schmidt G, Golecki JR, Ørskov I, Ørskov F, et al. Coliphage K5, specific for *E. coli* exhibiting the capsular K5 antigen. FEMS Microbiol Lett. 1982;14: 75–78. doi:10.1111/j.1574-6968.1982.tb08638.x

264. Pieroni P, Rennie RP, Ziola B, Deneer HG. The use of bacteriophages to differentiate serologically cross-reactive isolates of *Klebsiella pneumoniae*. J Med Microbiol. 1994;41: 423–429. doi:10.1099/00222615-41-6-423

265. Rieger-Hug D, Stirm S. Comparative study of host capsule depolymerases associated with Klebsiella bacteriophages. Virology. 1981;113: 363–378. doi:10.1016/0042-6822(81)90162-8

266. Tan D, Zhang Y, Qin J, Le S, Gu J, Chen L, et al. A Frameshift Mutation in wcaJ Associated with Phage Resistance in *Klebsiella pneumoniae*. Microorganisms. 2020;8: 378. doi:10.3390/microorganisms8030378

267. Thurow H, Niemann H, Stirm S. Bacteriophage-borne enzymes in carbohydrate chemistry: Part I. On the glycanase activity associated with particles of Klebsiella bacteriophage No. 11. Carbohydr Res. 1975;41: 257–271. doi:10.1016/S0008-6215(00)87024-X

268. Verma V, Harjai K, Chhibber S. Restricting ciprofloxacin-induced resistant variant formation in biofilm of *Klebsiella pneumoniae* B5055 by complementary bacteriophage treatment. J Antimicrob Chemother. 2009;64: 1212–1218. doi:10.1093/jac/dkp360

269. Knecht LE, Veljkovic M, Fieseler L. Diversity and Function of Phage Encoded Depolymerases. Front Microbiol. 2019;10: 2949. doi:10.3389/fmicb.2019.02949

270. Pires DP, Oliveira H, Melo LDR, Sillankorva S, Azeredo J. Bacteriophage-encoded depolymerases: their diversity and biotechnological applications. Appl Microbiol Biotechnol. 2016;100: 2141–2151. doi:10.1007/s00253-015-7247-0

271. Cuervo A, Pulido-Cid M, Chagoyen M, Arranz R, González-García VA, Garcia-Doval C, et al. Structural characterization of the bacteriophage T7 tail machinery. J Biol Chem. 2013;288: 26290–26299. doi:10.1074/jbc.M113.491209

272. Leiman PG, Battisti AJ, Bowman VD, Stummeyer K, Mühlenhoff M, Gerardy-Schahn R, et al. The structures of bacteriophages K1E and K1-5 explain processive degradation of polysaccharide capsules and evolution of new host specificities. J Mol Biol. 2007;371: 836–849. doi:10.1016/j.jmb.2007.05.083

273. Latka A, Leiman PG, Drulis-Kawa Z, Briers Y. Modeling the Architecture of Depolymerase-Containing Receptor Binding Proteins in Klebsiella Phages. Front Microbiol. 2019;10: 2649. doi:10.3389/fmicb.2019.02649

274. Pan Y-J, Lin T-L, Chen C-C, Tsai Y-T, Cheng Y-H, Chen Y-Y, et al. Klebsiella Phage ΦK64-1 Encodes Multiple Depolymerases for Multiple Host Capsular Types. J Virol. 2017;91: e02457-16. doi:10.1128/JVI.02457-16

275. Hsieh P-F, Lin H-H, Lin T-L, Chen Y-Y, Wang J-T. Two T7-like Bacteriophages, K5-2 and K5-4, Each Encodes Two Capsule Depolymerases: Isolation and Functional Characterization. Sci Rep. 2017;7: 4624. doi:10.1038/s41598-017-04644-2

276. Scholl D, Adhya S, Merril CR. Bacteriophage SP6 is closely related to phages K1-5, K5, and K1E but encodes a tail protein very similar to that of the distantly related P22. J Bacteriol. 2002;184: 2833–2836. doi:10.1128/JB.184.10.2833-2836.2002

277. Latka A, Lemire S, Grimon D, Dams D, Maciejewska B, Lu T, et al. Engineering the Modular Receptor-Binding Proteins of Klebsiella Phages Switches Their Capsule Serotype Specificity. mBio. 2021;12: e00455-21. doi:10.1128/mBio.00455-21

278. Rendueles O, Sousa JAM de, Bernheim A, Touchon M, Rocha EPC. Genetic exchanges are more frequent in bacteria encoding capsules. PLOS Genet. 2018;14: e1007862. doi:10.1371/journal.pgen.1007862

279. Ishiwa A, Komano T. PilV Adhesins of Plasmid R64 Thin Pili Specifically Bind to the Lipopolysaccharides of Recipient Cells. J Mol Biol. 2004;343: 615–625. doi:10.1016/j.jmb.2004.08.059

280. Christie PJ, Whitaker N, González-Rivera C. Mechanism and structure of the bacterial type IV secretion systems. Biochim Biophys Acta. 2014;1843: 1578–1591. doi:10.1016/j.bbamcr.2013.12.019

281. Paczosa MK, Mecsas J. *Klebsiella pneumoniae*: Going on the Offense with a Strong Defense. Microbiol Mol Biol Rev MMBR. 2016;80: 629–661. doi:10.1128/MMBR.00078-15

282. Wyres KL, Lam MMC, Holt KE. Population genomics of *Klebsiella pneumoniae*. Nat Rev Microbiol. 2020;18: 344–359. doi:10.1038/s41579-019-0315-1

283. Grabbe R, Klopprogge K, Schmitz RA. Fnr Is required for NifL-dependent oxygen control of nif gene expression in *Klebsiella pneumoniae*. J Bacteriol. 2001;183: 1385–1393. doi:10.1128/JB.183.4.1385-1393.2001

284. Russo TA, Marr CM. Hypervirulent *Klebsiella pneumoniae*. Clin Microbiol Rev. 2019;32: e00001-19. doi:10.1128/CMR.00001-19

285. Wyres KL, Hawkey J, Hetland MAK, Fostervold A, Wick RR, Judd LM, et al. Emergence and rapid global dissemination of CTX-M-15-associated *Klebsiella pneumoniae* strain ST307. J Antimicrob Chemother. 2019;74: 577–581. doi:10.1093/jac/dky492

286. Lan P, Jiang Y, Zhou J, Yu Y. A global perspective on the convergence of hypervirulence and carbapenem resistance in *Klebsiella pneumoniae*. J Glob Antimicrob Resist. 2021;25: 26–34. doi:10.1016/j.jgar.2021.02.020

287. Li Y, Hu D, Ma X, Li D, Tian D, Gong Y, et al. Convergence of carbapenem resistance and hypervirulence leads to high mortality in patients with postoperative *Klebsiella pneumoniae* meningitis. J Glob Antimicrob Resist. 2021;27: 95–100. doi:10.1016/j.jgar.2021.02.035

288. Wang H, Wilksch JJ, Lithgow T, Strugnell RA, Gee ML. Nanomechanics measurements of live bacteria reveal a mechanism for bacterial cell protection: the polysaccharide capsule in Klebsiella is a responsive polymer hydrogel that adapts to osmotic stress. Soft Matter. 2013;9: 7560–7567. doi:10.1039/C3SM51325D

289. Yu W-L, Ko W-C, Cheng K-C, Lee C-C, Lai C-C, Chuang Y-C. Comparison of prevalence of virulence factors for *Klebsiella pneumoniae* liver abscesses between isolates with capsular K1/K2 and non-K1/K2 serotypes. Diagn Microbiol Infect Dis. 2008;62. doi:10.1016/j.diagmicrobio.2008.04.007

290. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. PLoS Genet. 2019;15. doi:10.1371/journal.pgen.1008114

291. Diancourt L, Passet V, Verhoef J, Grimont PAD, Brisse S. Multilocus Sequence Typing of *Klebsiella pneumoniae* Nosocomial Isolates. J Clin Microbiol. 2005;43: 4178–4182. doi:10.1128/JCM.43.8.4178-4182.2005

292. Wyres KL, Holt KE. *Klebsiella pneumoniae* as a key trafficker of drug resistance genes from environmental to clinically important bacteria. Curr Opin Microbiol. 2018;45: 131–139. doi:10.1016/j.mib.2018.04.004

293. de Sousa JAM, Buffet A, Haudiquet M, Rocha EPC, Rendueles O. Modular prophage interactions driven by capsule serotype select for capsule loss under phage predation. ISME J. 2020;14: 2980–2996.

294. Lam MMC, Wick RR, Wyres KL, Gorrie CL, Judd LM, Jenney AWJ, et al. Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in *Klebsiella pneumoniae* populations. Microb Genomics. 2018;4: e000196. doi:10.1099/mgen.0.000196

295. Struve C, Roe CC, Stegger M, Stahlhut SG, Hansen DS, Engelthaler DM, et al. Mapping the Evolution of Hypervirulent *Klebsiella pneumoniae*. mBio. 2015;6: e00630. doi:10.1128/mBio.00630-15

296. Néron B, Littner E, Haudiquet M, Perrin A, Cury J, Rocha EPC. IntegronFinder 2.0: Identification and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in Klebsiella. Microorganisms. 2022;10: 700. doi:10.3390/microorganisms10040700

297. Tan YH, Chen Y, Chu WHW, Sham L-T, Gan Y-H. Cell envelope defects of different capsule-null mutants in K1 hypervirulent *Klebsiella pneumoniae* can affect bacterial pathogenesis. Mol Microbiol. 2020;113: 889–905. doi:10.1111/mmi.14447

298. Chen Y-T, Chang H-Y, Lai Y-C, Pan C-C, Tsai S-F, Peng H-L. Sequencing and analysis of the large virulence plasmid pLVPK of *Klebsiella pneumoniae* CG43. Gene. 2004;337: 189–198. doi:10.1016/j.gene.2004.05.008

299. Xu Y, Zhang J, Wang M, Liu M, Liu G, Qu H, et al. Mobilization of the nonconjugative virulence plasmid from hypervirulent *Klebsiella pneumoniae*. Genome Med. 2021;13: 119. doi:10.1186/s13073-021-00936-5

300. Lin T-L, Lee C-Z, Hsieh P-F, Tsai S-F, Wang J-T. Characterization of Integrative and Conjugative Element ICEKp1-Associated Genomic Heterogeneity in a *Klebsiella pneumoniae* Strain Isolated from a Primary Liver Abscess. J Bacteriol. 2008;190: 515–526. doi:10.1128/JB.01219-07

301. Lam MMC, Wyres KL, Wick RR, Judd LM, Fostervold A, Holt KE, et al. Convergence of virulence and MDR in a single plasmid vector in MDR *Klebsiella pneumoniae* ST15. J Antimicrob Chemother. 2019;74: 1218–1222. doi:10.1093/jac/dkz028

302. Sipos B, Massingham T, Stütz AM, Goldman N. An improved protocol for sequencing of repetitive genomic regions and structural variations using mutagenesis and next generation sequencing. PloS One. 2012;7: e43359. doi:10.1371/journal.pone.0043359

303. Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: Assessing Genomic Data Quality and Beyond. Curr Protoc. 2021;1: e323. doi:10.1002/cpz1.323

304. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29: 1072–1075. doi:10.1093/bioinformatics/btt086

305. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, et al. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res. 2021;49: D1020–D1028. doi:10.1093/nar/gkaa1105

306. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119. doi:10.1186/1471-2105-11-119

307. Korandla DR, Wozniak JM, Campeau A, Gonzalez DJ, Wright ES. AssessORF: combining evolutionary conservation and proteomics to assess prokaryotic gene predictions. Bioinformatics. 2019;36: 1022–1029. doi:10.1093/bioinformatics/btz714

308. Storz G, Wolf YI, Ramamurthi KS. Small proteins can no longer be ignored. Annu Rev Biochem. 2014;83: 753–777. doi:10.1146/annurev-biochem-070611-102400

309. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421

310. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7: e1002195. doi:10.1371/journal.pcbi.1002195

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

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

313. Abby SS, Néron B, Ménager H, Touchon M, Rocha EPC. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PloS One. 2014;9: e110726. doi:10.1371/journal.pone.0110726

314. Zaslaver A, Mayo A, Ronen M, Alon U. Optimal gene partition into operons correlates with gene functional order. Phys Biol. 2006;3: 183–189. doi:10.1088/1478-3975/3/3/003

315. Lathe WC, Snel B, Bork P. Gene context conservation of a higher order than operons. Trends Biochem Sci. 2000;25: 474–479. doi:10.1016/s0968-0004(00)01663-7

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

317. Lawrence JG, Roth JR. Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters. Genetics. 1996;143: 1843–1860.

318. Abby SS, Rocha EPC. Identification of Protein Secretion Systems in Bacterial Genomes Using MacSyFinder. Methods Mol Biol. 2017;1615: 1–21. doi:10.1007/978-1-4939-7033-9_1

319. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, 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: 3895–3903. doi:10.1128/AAC.02412-14

320. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, Corander J, et al. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. Microb Genomics. 2018;4: e000224. doi:10.1099/mgen.0.000224

321. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3: e985. doi:10.7717/peerj.985

322. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016;44: W16-21. doi:10.1093/nar/gkw387

323. Perrin A, Rocha EPC. PanACoTA: a modular tool for massive microbial comparative genomics. NAR Genomics Bioinforma. 2021;3: lqaa106. doi:10.1093/nargab/lqaa106

324. Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. Nucleic Acids Res. 2018;46: e5. doi:10.1093/nar/gkx977

325. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. GigaScience. 2019;8: giz119. doi:10.1093/gigascience/giz119

326. Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. Trends Genet TIG. 2020;36: 132–145. doi:10.1016/j.tig.2019.11.006

327. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37: 1530–1534. doi:10.1093/molbev/msaa015

328. Ishikawa SA, Zhukova A, Iwasaki W, Gascuel O. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. Mol Biol Evol. 2019;36: 2069–2085. doi:10.1093/molbev/msz131

329. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27: 592–593. doi:10.1093/bioinformatics/btq706

330. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4: 1686. doi:10.21105/joss.01686

331. Csűös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics. 2010;26: 1910–1912. doi:10.1093/bioinformatics/btq315

332. Ferrières L, Hémery G, Nham T, Guérout A-M, Mazel D, Beloin C, et al. Silent mischief: bacteriophage Mu insertions contaminate products of *Escherichia coli* random mutagenesis performed using suicidal

transposon delivery plasmids mobilized by broad-host-range RP4 conjugative machinery. J Bacteriol. 2010;192: 6418–6427. doi:10.1128/JB.00621-10

333. Balestrino D, Haagensen JAJ, Rich C, Forestier C. Characterization of type 2 quorum sensing in *Klebsiella pneumoniae* and relationship with biofilm formation. J Bacteriol. 2005;187: 2870–2880. doi:10.1128/JB.187.8.2870-2880.2005

334. Kuhlman TE, Cox EC. Site-specific chromosomal integration of large synthetic constructs. Nucleic Acids Res. 2010;38: e92. doi:10.1093/nar/gkp1193

335. Haudiquet M, Buffet A, Rendueles O, Rocha EPC. Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen *Klebsiella pneumoniae*. PLoS Biol. 2021;19: e3001276. doi:10.1371/journal.pbio.3001276

336. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. Nature. 2011;480: 241–244. doi:10.1038/nature10571

337. Choi I-G, Kim S-H. Global extent of horizontal gene transfer. Proc Natl Acad Sci U S A. 2007;104: 4489–4494. doi:10.1073/pnas.0611557104

338. Rayssiguier C, Thaler DS, Radman M. The barrier to recombination between *Escherichia coli* and Salmonella typhimurium is disrupted in mismatch-repair mutants. Nature. 1989;342: 396–401. doi:10.1038/342396a0

339. Beamud B, García-González N, Gómez-Ortega M, González-Candelas F, Domingo-Calap P, Sanjuan R. Genetic determinants of host tropism in Klebsiella phages. bioRxiv; 2022. p. 2022.06.01.494021. doi:10.1101/2022.06.01.494021

340. Hesse S, Rajaure M, Wall E, Johnson J, Bliskovsky V, Gottesman S, et al. Phage Resistance in Multidrug-Resistant *Klebsiella pneumoniae* ST258 Evolves via Diverse Mutations That Culminate in Impaired Adsorption. mBio. 2020;11: e02530-19. doi:10.1128/mBio.02530-19

341. Haudiquet M, Rendueles O, Rocha EPC. Capsule serotypes result in distinct phage infection patterns and frequency of plasmid conjugation. In preparation. 2022.

342. Low WW, Wong JLC, Beltran LC, Seddon C, David S, Kwong H-S, et al. Mating pair stabilization mediates bacterial conjugation species specificity. Nat Microbiol. 2022;7: 1016–1027. doi:10.1038/s41564-022-01146-4

343. Xu L, Wang M, Yuan J, Wang H, Li M, Zhang F, et al. The KbvR Regulator Contributes to Capsule Production, Outer Membrane Protein Biosynthesis, Antiphagocytosis, and Virulence in *Klebsiella pneumoniae*. Infect Immun. 2021;89: e00016-21. doi:10.1128/IAI.00016-21

344. Cai R, Wang G, Le S, Wu M, Cheng M, Guo Z, et al. Three Capsular Polysaccharide Synthesis-Related Glucosyltransferases, GT-1, GT-2 and WcaJ, Are Associated With Virulence and Phage Sensitivity of *Klebsiella pneumoniae*. Front Microbiol. 2019;10. doi:10.3389/fmicb.2019.01189

345. Hung C-H, Kuo C-F, Wang C-H, Wu C-M, Tsao N. Experimental Phage Therapy in Treating *Klebsiella pneumoniae*-Mediated Liver Abscesses and Bacteremia in Mice. Antimicrob Agents Chemother. 2011;55: 1358–1365. doi:10.1128/AAC.01123-10

346. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46: 305–309. doi:10.1038/ng.2895

347. Langereis JD, de Jonge MI. Non-encapsulated *Streptococcus pneumoniae*, vaccination as a measure to interfere with horizontal gene transfer. Virulence. 2017;8: 637–639. doi:10.1080/21505594.2017.1309492

348. Feldman MF, Mayer Bridwell AE, Scott NE, Vinogradov E, McKee SR, Chavez SM, et al. A promising bioconjugate vaccine against hypervirulent *Klebsiella pneumoniae*. Proc Natl Acad Sci U S A. 2019;116: 18655–18663. doi:10.1073/pnas.1907833116

349. Cheng HY, Chen YS, Wu CY, Chang HY, Lai YC, Peng HL. RmpA regulation of capsular polysaccharide biosynthesis in *Klebsiella pneumoniae* CG43. J Bacteriol. 2010;192: 3144–3158. doi:10.1128/JB.00031-10

350. Lai Y-C, Peng H-L, Chang H-Y. RmpA2, an activator of capsule biosynthesis in *Klebsiella pneumoniae* CG43, regulates K2 cps gene expression at the transcriptional level. J Bacteriol. 2003;185: 788–800. doi:10.1128/JB.185.3.788-800.2003

351. Walker KA, Miner TA, Palacios M, Trzilova D, Frederick DR, Broberg CA, et al. A *Klebsiella pneumoniae* Regulatory Mutant Has Reduced Capsule Expression but Retains Hypermucoviscosity. mBio. 2019;10: e00089-19. doi:10.1128/mBio.00089-19

352. Walker KA, Treat LP, Sepúlveda VE, Miller VL. The Small Protein RmpD Drives Hypermucoviscosity in *Klebsiella pneumoniae*. mBio. 2020;11: e01750-20. doi:10.1128/mBio.01750-20

353. Zhang Y, Zhao C, Wang Q, Wang X, Chen H, Li H, et al. High Prevalence of Hypervirulent *Klebsiella pneumoniae* Infection in China: Geographic Distribution, Clinical Characteristics, and Antimicrobial Resistance. Antimicrob Agents Chemother. 2016;60: 6115–6120. doi:10.1128/AAC.01127-16

354. Arbatsky NP, Kasimova AA, Shashkov AS, Shneider MM, Popova AV, Shagin DA, et al. Involvement of a Phage-Encoded Wzy Protein in the Polymerization of K127 Units To Form the Capsular Polysaccharide of Acinetobacter baumannii Isolate 36-1454. Microbiol Spectr. 2022; e0150321. doi:10.1128/spectrum.01503-21

List of figures

Annexes

Annex #1

Research article: Modular prophage interactions driven by capsule serotype select for capsule loss under phage predation.

Jorge A. M. de Sousa*, Amandine Buffet, Matthieu Haudiquet, Eduardo P. C. Rocha, Olaya Rendueles

Published in The ISME Journal.

For this work, I provided the *Klebsiella* phylogenetic tree represented on Figure 1, as well as in-depth phage genome annotation and genome comparisons represented on Figure 2.

The ISME Journal (2020) 14:2980–2996
https://doi.org/10.1038/s41396-020-0726-z

ARTICLE

Modular prophage interactions driven by capsule serotype select for capsule loss under phage predation

Jorge A. M. de Sousa[1] · Amandine Buffet[1] · Matthieu Haudiquet[1,2] · Eduardo P. C. Rocha[1] · Olaya Rendueles[1]

Received: 7 January 2020 / Revised: 15 July 2020 / Accepted: 20 July 2020 / Published online: 30 July 2020

Abstract

Klebsiella species are able to colonize a wide range of environments and include worrisome nosocomial pathogens. Here, we sought to determine the abundance and infectivity of prophages of *Klebsiella* to understand how the interactions between induced prophages and bacteria affect population dynamics and evolution. We identified many prophages in the species, placing these taxa among the top 5% of the most polylysogenic bacteria. We selected 35 representative strains of the *Klebsiella pneumoniae* species complex to establish a network of induced phage–bacteria interactions. This revealed that many prophages are able to enter the lytic cycle, and subsequently kill or lysogenize closely related *Klebsiella* strains. Although 60% of the tested strains could produce phages that infect at least one other strain, the interaction network of all pairwise cross-infections is very sparse and mostly organized in modules corresponding to the strains' capsule serotypes. Accordingly, capsule mutants remain uninfected showing that the capsule is a key factor for successful infections. Surprisingly, experiments in which bacteria are predated by their own prophages result in accelerated loss of the capsule. Our results show that phage infectiousness defines interaction modules between small subsets of phages and bacteria in function of capsule serotype. This limits the role of prophages as competitive weapons because they can infect very few strains of the species complex. This should also restrict phage-driven gene flow across the species. Finally, the accelerated loss of the capsule in bacteria being predated by their own phages, suggests that phages drive serotype switch in nature.

Introduction

Phages are one of the most abundant entities on Earth. They are found in multiple environments, typically along with their host bacteria, including in the human microbiome. Many recent studies focused on virulent phages, which follow exclusively a lytic cycle. In contrast, temperate phages, which can either follow a lytic cycle, or integrate into the host genome and produce a lysogen, have been comparatively less studied. Integrated phages, hereafter referred to as prophages, replicate vertically with the host and are typically able to protect them from infections by similar phages, the so-called resistance to superinfection [1–3]. Most prophage genes are silent and have little impact in bacterial fitness as long as there is no induction of the lytic cycle [1]. If the prophage remains in the genome for a very long period of time it may be inactivated by mutations. A few studies suggest that many prophages are inactive to some extent [4, 5]. Upon induction, some of them cannot excise (cryptic prophages), replicate, infect, or produce viable progeny. Prophage inactivation and further domestication may lead to the co-option of some phage functions by the bacterial host [6]. For instance, some bacteriocins result from the domestication of phage tails [7, 8]. In contrast, intact prophages can be induced (by either extrinsic or intrinsic factors) and resume a lytic cycle, producing viable viral particles.

Temperate phages affect the evolution of gene repertoires and bacterial population dynamics [9–11] by two key mechanisms. First, induction of prophages by a small subset of a population produces virions that can infect susceptible bacteria and thus facilitate colonization [9, 12, 13]. Second, they drive horizontal gene transfer between bacteria.

Supplementary information The online version of this article (https://doi.org/10.1038/s41396-020-0726-z) contains supplementary material, which is available to authorized users.

✉ Olaya Rendueles
olaya.rendueles-garcia@pasteur.fr

[1] Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris 75015, France

[2] Ecole Doctorale FIRE—Programme Bettencourt, CRI, Paris 75005, France

Around half of the sequenced genomes contain identifiable prophages [14, 15], e.g., a third of *Escherichia coli*'s pangenome is in prophages [16]. The frequency of prophages is higher in bacteria with larger genomes, in pathogens, and in fast-growing bacteria [15]. Lysogenization and the subsequent expression of some prophage genes may result in phenotypic changes in the host, e.g., many pathogens have virulence factors encoded in prophages [17]. Prophages may also facilitate horizontal transfer between bacteria by one of several mechanisms of transduction [1, 18, 19]. Interestingly, bacterial populations can acquire adaptive genes from their competitors by killing them with induced prophages and recovering their genes by generalized transduction [20]. While these mechanisms have been explored in many experimental and computational studies, the impact of temperate phages in the diversity of bacterial lysogens is still poorly understood.

Here, we assess the relevance of prophages in the biology of *Klebsiella* spp., a genus of bacteria capable of colonizing a large range of environments. The genus includes genetically diverse species of heterotrophic facultative aerobes that have been isolated from numerous environments, including the soil, sewage, water, plants, insects, birds and mammals [21]. *Klebsiella* spp. can cause various diseases such as urinary tract infections, acute liver abscesses, pneumonia, infectious wounds, and dental infections [22, 23]. They commonly cause severe hospital outbreaks associated with multidrug resistance, and *K. pneumoniae* is one of the six most worrisome antibiotic-resistant (ESKAPE) pathogens. The versatility of *Klebsiella* spp. is associated with a broad and diverse metabolism [24], partly acquired by horizontal gene transfer [23, 25, 26]. In addition, *Klebsiella* spp. code for an extracellular capsule that is highly variable within the species. This capsule is a high molecular weight polysaccharide made up of different repeat units of oligosaccharides. Combinations of different oligosaccharides are referred to as serotypes. In *K. pneumoniae* there are 77 serologically defined serotypes, and numbered from K1 to K77 [27] and more than 130 serotypes were identified computationally. The latter are noted from KL1 to KL130 [28, 29], and are usually referred as capsule locus types (or CLT). The capsule is considered a major virulence factor, required, for instance, in intestinal colonization [30]. It also provides resistance to the immune response and to antibiotics [31–33]. From an ecological point of view, the capsule is associated with bacteria able to colonize diverse environments [34, 35]. Its rapid diversification may thus be a major contributor to *Klebsiella*'s adaptive success, including in colonizing clinical settings.

We have recently shown that species of bacteria encoding capsular loci undergo more frequent genetic exchanges and accumulate more mobile genetic elements, including prophages [35]. This is surprising because capsules were proposed to decrease gene flow [36] and some phages are known to be blocked by the capsule [37–39]. However, several virulent phages of *Klebsiella* are known to have depolymerase activity in their tail fibers [40–44]. These depolymerases specifically digest oligosaccharidic bonds in certain capsules [45] and allow phages to access the outer membrane and infect bacteria [46]. Since depolymerases are specific to one or a few capsule types [44, 47], this implicates that some phages interact with capsules in a serotype-specific manner [42, 44, 48]. In addition, the capsule could facilitate cell infection because phages bind reversibly to it, prior to the irreversible binding to the specific cell receptor [49].

To date, the number and role of prophages in *Klebsiella*'s population biology is not well known. *Klebsiella* are interesting models to study the role of prophages, because of the interplay between the capsule, phage infections, and also the influence of the former in *Klebsiella's* colonization of very diverse ecological niches. In this work, we sought to characterize the abundance and distribution of *Klebsiella* temperate phages, and experimentally assess their ability to re-enter the lytic cycle and lysogenize different *Klebsiella* strains. By performing more than 1200 pairwise combinations of lysates and host strains, we aim to pinpoint the drivers of prophage distribution and elucidate some of the complex interactions that shape phage–bacteria interactions in *Klebsiella*.

Results

Prophages are very abundant in the genomes of *Klebsiella*

We used PHASTER [50] to analyse 254 genomes of eight *Klebsiella* species (and two subspecies). We detected 1674 prophages, of which 55% are classified as "intact" by PHASTER and are the most likely to be complete and functional. These "intact" prophages were present in 237 out of the 254 genomes (see Methods). The remaining prophages were classed as "questionable" (20%) and "incomplete" (25%, Figs. 1a, b and S1). The complete list of bacterial genomes and prophages is available in Dataset S1. Most of the genomes were polylysogenic, encoding more than one prophage (Fig. S1). However, the number of prophages varied markedly across genomes, ranging from 1 to 16, with a median of 6 per genome. In addition, the total number of prophages in *Klebsiella* spp. varied significantly across species (Kruskal–Wallis, $df = 6$, $P < 0.001$, Fig. 1b). More specifically, the average number of "intact" prophages varied between eight for *K. oxytoca* and less than one for *K. quasipneumoniae subsp. quasipneumoniae*. *K. pneumoniae*, the most represented species of our dataset (~77% of the

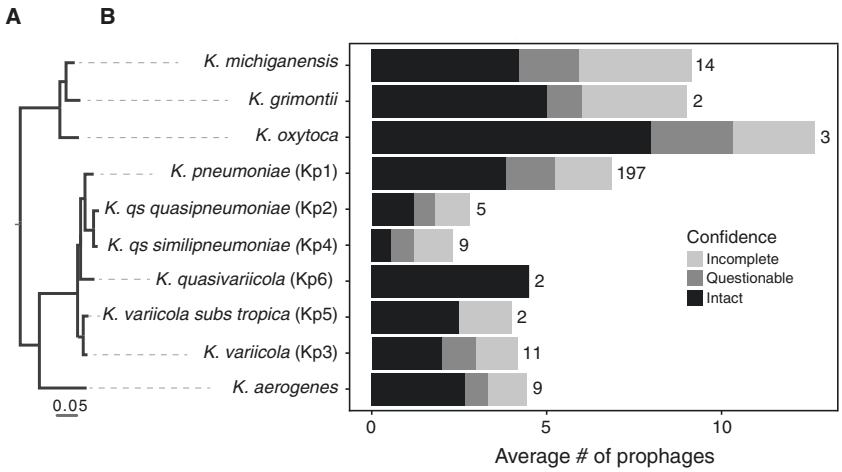


Fig. 1 Prophage distribution in *Klebsiella* genus. a Rooted phylogenetic tree of *Klebsiella* species used in this study based on the core genes. **b** Average number of prophages per genome. PHASTER prediction for completeness is indicated. Numbers represent the total number of genomes analysed of each species.

genomes), has an average of nine prophages per genome, of which four are classified as "intact" (Fig. 1b). As expected, both the number of prophages per genome and the average number of prophages per species are correlated positively with genome size (Spearman's rho = 0.49, $P < 0.001$, Spearman's rho = 0.76, $P = 0.01$, respectively) (Fig. S2A, B). When compared with the one hundred most sequenced bacterial species, the number of prophages in *Klebsiella* is very high. *K. pneumoniae* ranks within the 5th percentile of the most prophage-rich species, comparable with *E. coli* and *Yersinia enterocolitica* (Fig. S3). This shows that prophages are a sizeable fraction of the genomes of *Klebsiella* and may have an important impact in its biology.

Our experience is that prophages classed as "questionable" and "incomplete" often lack essential phage functions. Hence, all the remaining analyses were performed on "intact" prophages, unless indicated otherwise. These elements have 5% lower GC content than the rest of the chromosome (Wilcoxon test, $P < 0.001$, Fig. S2B), as typically observed for horizontally transferred genetic elements [51, 52]. Their length varies from 13 to 137 kb, for an average of 46 kb. Since temperate dsDNA phages of enterobacteria are usually larger than 30 kb [53], this suggests that a small fraction of the prophages might be incomplete (Fig. S2C). Among the "intact" prophages detected in the 35 strains analysed from our laboratory collection (Dataset S1 and Fig. S1) and isolated from different environments and representative of the genetic diversity of the *K. pneumoniae* species complex [24], we chose 11 to characterize in detail in terms of genetic architecture (Fig. 2). A manual and computational search for recombination sites (*att*) in these 11 prophages (See Methods, Fig. 2) showed that some might be larger than predicted by PHASTER. To verify the integrity of these phages, we searched for genes encoding structural functions —head, baseplate, and tail—and found them in the eleven prophages. Two of these prophages (#62 (2) and #54 (1)) have a protein of *ca.* 4200 aa (Fig. 2), homologous to the tail component (gp21) of the linear phage-plasmid phiKO2 of *K. oxytoca* [54]. The prophages had integrases and were flanked by recombination sites, suggesting that they could still be able to excise from the chromosome (Fig. 2).

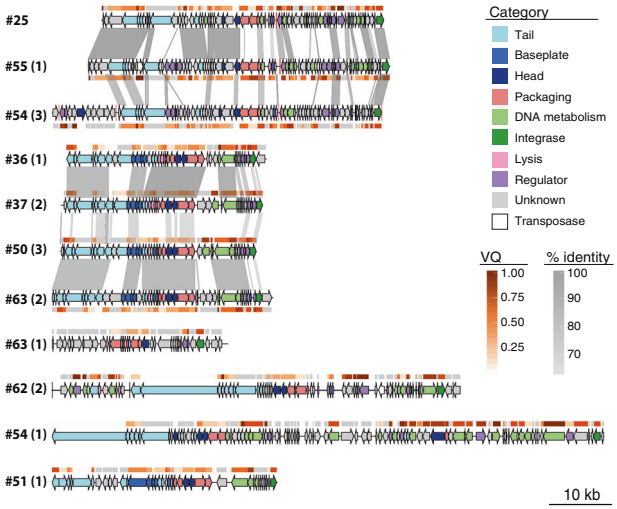


Fig. 2 Genomic organization of eleven of the prophages in this study. Numbers correspond to the host genomes, as displayed in Fig. 3. The number in parenthesis identifies the prophage in the genome. All prophages are classified as "intact" except #63 (1) ("questionable"). Genome boundaries correspond to attL/R sites. Arrows represent predicted ORFs and are oriented according to transcriptional direction. Colors indicate assigned functional categories, tRNAs are represented as red lines and the sequences are oriented based on the putative integrase localization. Local *blastn* alignments (option *dc-megablast*) are displayed between pairs of related prophages, colored according to the percentage of identity. The Viral Quotient (VQ) from pVOG is displayed below or on top of each ORF, with gray meaning that there was no match in the pVOG profiles database and thus no associated VQ value. Prophages #62 (2) and #54(1) have inserted the core gene *icd*, and the boundaries correspond to *icd* on the right and the most distal *att* site found, which is likely to be a remnant prophage border. The most proximal *att* site is also annotated (vertical black line). This figure was generated using the R package GenoPlotR v0.8.9 [103].

Nevertheless, some prophages encoded a small number of insertion sequences, which accumulate under lysogeny and degrade prophages [55].

Klebsiella spp. prophages can be released into the environment

To further characterize the prophages of *K. pneumoniae* species complex, we sought to experimentally assess their ability to excise and be released into the environment. The analysis described above identified 95 prophages in the 35 strains analyzed from our collection (Fig. 3). Among these, 40 were classed as "intact" (including the ten prophages whose genomic composition is characterized in Fig. 2). Eleven strains had no "intact" prophages, and four of these also lacked "questionable" prophages. Hence, our collection is representative of the distribution of prophages in the species, with some genomes containing one or multiple putatively functional prophages and others being prophage-free.

To test prophage induction, we grew these 35 strains and added mitomycin C (MMC) to exponentially growing cultures. Viable prophages are expected to excise and cause cell death as a result of phage replication and outburst. In agreement with PHASTER predictions, the addition of MMC to the strains lacking "intact" and "questionable" prophages showed no significant cell death (Figs. 3 and S4), with the single exception of mild cell death at high doses of MMC for strain NTUH K2044 (#56). Three of the seven strains lacking "intact" but having some "questionable" prophages showed very mild cell death upon induction, two others exhibited a dose-dependent response, and the remaining two showed rapid cell death (#44 and #35). This suggests that at least some of the "questionable" prophages are still inducible and able to kill the host. In addition, all the 24 strains with "intact" prophages showed signs of cell death *ca*. 1 h after exposure to MMC (Figs. 3 and S4). This occurred in a dose-dependent manner, consistent with prophage induction.

These results suggest that most strains have inducible prophages. To verify the release of prophage DNA to the environment, we tested the presence of 17 different phages from eleven different strains by PCR, both in induced and non-induced PEG-precipitated filtered supernatants

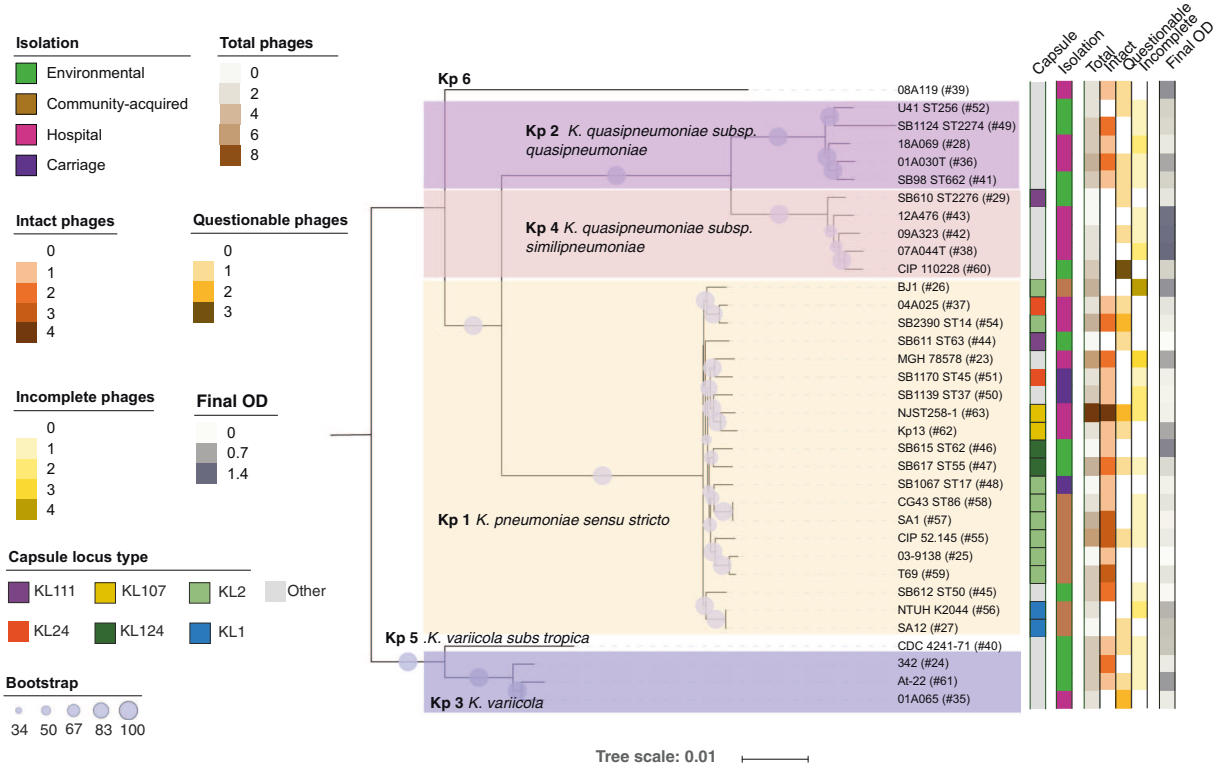


Fig. 3 Phylogenetic tree of the 35 Klebsiella strains. The tree was built using the protein sequences of 3009 families of the core genome of representative strains from the *Klebsiella pneumoniae* species complex. The first column determines the capsule locus type and the second column provides information about the environment from which it was isolated. The next columns indicate the total, "intact", "questionable", and "incomplete" prophages detected in the genomes by PHASTER. The last column shows the final absorbance of a culture after induction by mitomycin C. Background color indicates different *Klebsiella* spp. The size of the circles along the branches are proportional to bootstrap values ranging between 34 and 100. The gray color indicates other CLTs that are only present once in the dataset. The hash symbol represents the number of the strain in our collection and is used for simplicity throughout the text.

(Fig. S5A). Indeed, all 17 phages were amplified, indicating that viral genomes are released into the environment. As expected, amplification bands were weaker in the non-induced supernatant compared with the induced (Fig. S5A). We further verified the recircularization of these prophages in induced supernatants. We detected the recircularization of eight phages induced from six strains (#25, #36, #62, #63, #37, and #50), five of which were classified as "intact", and three as "questionable" by PHASTER (Fig. S5B). Despite using different primers and PCR settings, we did not obtain a clean PCR product for eight phages. This is most likely due to the aforementioned inaccuracies in the delimitation of the prophages, as these phages are detected in induced supernatants (Fig. S5A). Overall, our results suggest that most "intact" and some of the "questionable" prophages are able to induce the lytic cycle and lyse the cells.

The prophage–bacteria interaction network is sparse

Prophages may impact intraspecies competition [9]. In order to characterize this effect in the *K. pneumoniae* species complex, we produced three independent lysates of all 35 strains, by MMC induction and PEG-precipitation of the filtered supernatants (see Methods). We tested the ability of all lysates to produce a clearing on bacterial overlays of every strain. This resulted in 1225 possible lysate–bacteria combinations, with some lysates being potentially composed of multiple types of phage and different proportions across the three independent replicates. We first tested whether the number of different prophages in each strain's genome correlated with its ability to infect the other strains. To do so, we calculated the infection score of each strain, i.e., the average frequency at which its lysate infects the 35 strains of our panel. We observed a positive and significant association between the number of predicted "intact" phages and the infection score (rho = 0.16, $P < 0.001$). Overall, 21 strains out of the 35 (60%) produced a lysate that could infect another strain, but only 75 of the 1225 possible lysate–strain combinations showed clearances or plaques (Figs. 4a and S6). To confirm that plaques are caused by the activity of phages, we added lysates to growing cultures of sensitive bacteria, which resulted in bacterial death. Consistent with the action of phages, bacterial death was prevented upon the addition of citrate, a known inhibitor of phage adsorption

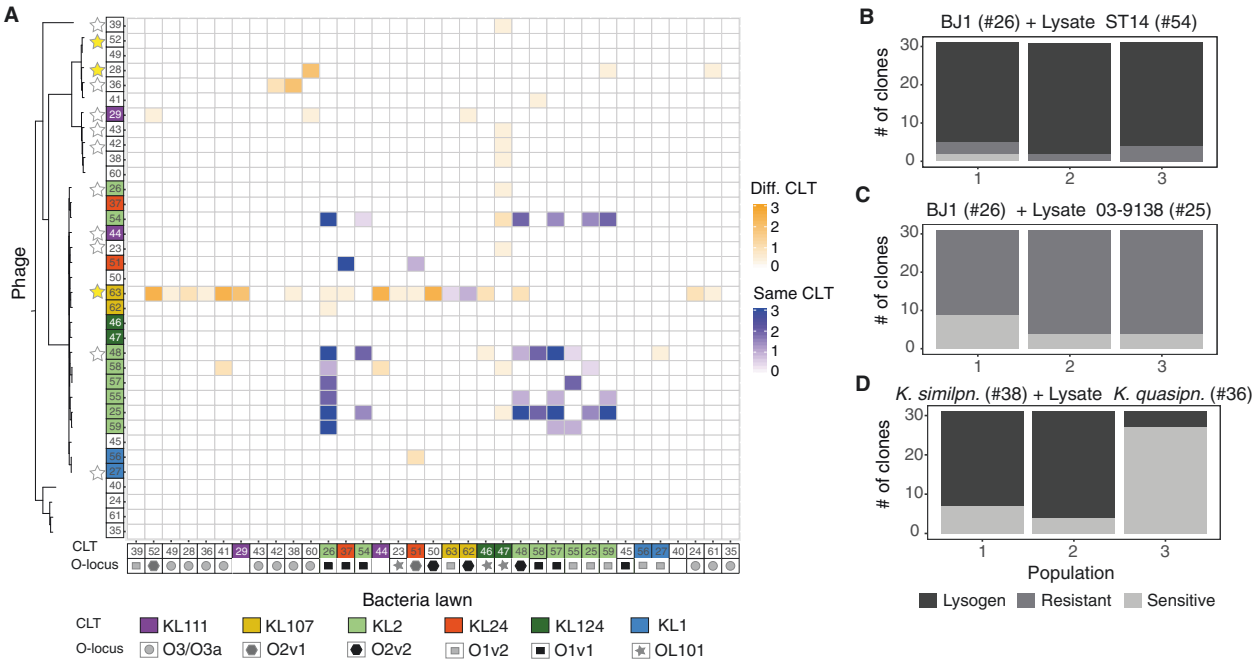


Fig. 4 Some *Klebsiella* encode viable prophages that can infect and lysogenize other strains. a Infection matrix indicating the ability of induced and PEG-precipitated supernatants of all strains to form inhibition halos on lawns of all strains. This was repeated three times with three independently produced lysates. The sum of the three experiments is shown. The strains are ordered by phylogeny, and the colors on top of numbers indicate capsule locus types (CLT). Geometric figures along the *x*-axis indicate the LPS serotype (O-locus). Stars along the *y*-axis indicate that the genome codes for putative genes involved in colicin production but lacking a lysis gene (white) or has a full colicin operon (yellow). Shades of blue indicate infections between lysates produced by bacteria from the same CLT as the target bacteria, whereas shades of orange indicate cross-CLT infections. **b–d** Barplots indicate the number of resistant, sensitive, or lysogenized clones from each independent replicate population exposed 24 h to lysates from other strains. Each bar represents the 31 survivor clones from each of three independent experiments. Gray scale indicates whether: (i) the clones are still sensitive to the phage (light gray), (ii) they became resistant by becoming a lysogen (dark gray) as determined by cell death upon addition of MMC and PCR targeting prophage sequences, or (iii) they became resistant by other mechanisms.

(Fig. S6A). Surprisingly, 5 out of the 21 strains could be reinfected by their own lysate. This suggests the presence of ineffective repressor mechanisms protecting from superinfection, and is consistent with the observed spontaneous induction of prophages in lysogens in some of these strains (Figs. S5 and S6C).

Some of the inhibition halos observed in the experiments could result from bacteriocins induced by MMC [56]. To test if this could be the case, we searched for putative bacteriocins in the genomes and found them in 13 of the 35 genomes (Table S1, Fig. 4). However, most of them had very low (<50%) protein sequence identity with known proteins and were in genomes lacking recognizable lysis proteins. This may explain why 3 of the 13 strains failed to produce an inhibition halo on all tested strains and five showed mild inhibition of a single target strain (#47). In three of the strains (#63, #48, and #36) we could confirm phage production by recircularization (Fig. S5), showing that phages are produced. Indeed, phages from strain #36 can lysogenize strain #38 (see below, Fig. 4d), and phages from strain #48 are able to produce plaques on overlays of strain BJ1(Fig. S6B). Strain #63 has a full bacteriocin operon, with high identity to colicin E7. It is thus possible that some of the inhibition we observed from the lysates of this strain could be caused by bacteriocins. Note that we also observed that at least two of the prophages of strain #63 are able to excise and circularize (Fig. S5 and see annotation Fig. 2). Thus, for this particular strain, we cannot precisely identify the cause of the inhibition halos. Finally, from the two remaining strains (#29 and #28, both *K. quasipneumoniae*), only #28 encodes a lysis protein, but they both mildly infect three strains each. Overall, our results suggest that for the majority of the cases shown here, inhibitions seem to be caused by phage activity and are not due to bacteriocins.

Our analysis shows that most lysates cannot infect other strains. For the 75 phage infections we do observe, only 17 (23%) occurred in all three independently produced lysates (Fig. S7), hinting to an underlying stochasticity in the induction and infection process. Interestingly, we observe that infections of target bacteria with lysates from bacteria with the same CLT seem both more reproducible and effective, compared with those lysates produced from bacteria with different CLT (Fig. S7), suggesting that the capsule may play a crucial role during phage infection.

Overall, our analysis shows that most lysates cannot infect other strains resulting in a sparse matrix of phage-mediated competitive interactions.

Induced *Klebsiella* prophages can lysogenize other strains

An induced prophage can, upon infection of a new host, generate new lysogens. To test the ability of the induced prophages to lysogenize other strains, we challenged *K. pneumoniae* BJ1 (#26) that lacks prophages with lysates from two strains (#54 and #25). These strains have the same CLT as BJ1 (KL2) and we have PCR evidence that they produce viral particles (see above) (Fig. S5). We grew BJ1 in contact with these two lysates, and from the surviving BJ1 cells, we isolated 93 clones from three independently challenged cultures. We reexposed these clones to a lysate from strain #54, to test if they had become resistant. Almost all clones (96%) of BJ1 could grow normally upon rechallenge (Fig. 4b), suggesting that they had acquired resistance either by lysogenization or some other mechanism. Most of these clones displayed significant cell death upon exposure to MMC (whereas the ancestral strain was insensitive), which is consistent with lysogenization. Analyses by PCR showed that 57 out of 82 clones from the three independent BJ1 cultures exposed to #54 lysate acquired at least one of the "intact" prophages and at least four of them acquired both (shown in Fig. 2). In contrast, no BJ1 clones became lysogens when challenged with the lysate of strain #25 (as tested by exposure to MMC and PCR verification, Fig. 4c), indicating that the surviving clones became resistant to phage infections by other mechanisms. Overall, this shows that lysogenization of BJ1 is dependent on the infecting phages, with some driving the emergence of alternative mechanisms of resistance to infection in the bacterial host.

To test the taxonomic range of infections caused by induced prophages, we also exposed a Kp 4 (*K. quasipneumoniae subsp similipneumoniae*) (#38) to a lysate from a Kp 2 (*K. quasipneumoniae subsp quasipneumoniae*) (#36) that consistently infects #38 even if it belongs to a different subspecies and has a different capsular CLT (Fig. 4a). Survivor clones of strain #38 were lysogenized (in 61 out of 93 surviving clones) exclusively by one of the phages present in strain #36, as confirmed by PCR of the phage genome (Fig. 4d). Interestingly, four of these lysogenized clones did not display cell death when exposed to MMC, suggesting that their prophages might not be fully functional. Taken together, our results show that some *Klebsiella* prophages can be transferred to and lysogenize other strains, including from other subspecies.

Resistance to superinfection and bacterial defenses do not explain the interaction matrix

To investigate the determinants of infections in the interaction matrix, we first hypothesized that resident prophages could hamper infections by lysates of other strains. Consistently, the most sensitive strain in our panel (#26) does not have any detectable prophages, rendering it sensitive to infection by numerous lysates. However, we found no negative association between the number of prophages in

the target strain and the number of times it was infected (rho = 0.04, $P = 0.166$). Repression of infecting phages is expected to be highest when there are similar resident prophages in the strain. Even if closer strains are more likely to carry the same prophages, the interaction matrix clearly shows that phages tend to infect the most closely related strains. To determine if the presence of similar prophages shapes the infection matrix, we calculated the genetic similarity between all "intact" prophages using weighted gene repertoire relatedness (wGRR) (see Methods). The frequency of infection was found to be independent from the similarity between prophages (determined as higher than 50% wGRR for phage pairs, Odds ratio = 0.55, P value = 0.055) when analysing only "intact" prophages or also including the "questionable" (Fig. S8A). Ultimately, resident phages may repress incoming phages if they have very similar repressors, but pairs of dissimilar phages (wGRR < 50%) with similar repressors (>80% sequence identity) are only ~0.3% of the total (see Methods, Fig. S8B). Thus, the analysis of both phage sequence similarity and similarity between their repressors suggests that the observed interaction matrix is not strongly influenced by the resistance to superinfection provided by resident prophages.

Bacterial defense systems block phage infections and are thus expected to influence the interaction matrix. To study their effect, we analysed systems of CRISPR-Cas and restriction-modification (R-M), since these are the most common, the best characterized to date, and those for which tools for their computational detection are available [57, 58]. We found CRISPR-Cas systems in 8 of the 35 strains and tested if the strains that were infected by the lysates of other bacteria were less likely to encode these systems. We observed no correlation between the number of unsuccessful infections and the presence of CRISPR-Cas systems in a genome ($P > 0.05$, Wilcoxon test). Accordingly, the majority of strains (77%) lack spacers against any of the prophages of all the other strains and we found no correlation between the presence of CRISPR spacers targeting incoming phages and the outcome of the interaction (i.e., infection or not) (Odds ratio = 1.28 Fisher's exact test, P value = 1, Dataset S1, Fig. S9A). Actually, only 5% of the pairs with null interactions (no infection) concerned a target strain with CRISPR spacers matching a prophage in the lysate-producing strain. This indicates that CRISPR-Cas systems do not drive the patterns observed in the infection matrix.

The analysis of R-M systems is more complex because their specificity is often harder to predict. We searched for these systems in the 35 genomes and identified the following systems: 37 type I, 158 type II, 4 type IIG, 7 type III, and 7 type IV. We then investigated whether the systems present in the target strains could protect from phages in the lysates. Almost all the strains studied here have systems that could target at least one prophage from the lysate-producing strains (Fig. S9B, see Methods). However, the joint distribution of R-M systems and targeted prophages does not explain the outcome observed in the infection matrix (Odds ratio = 0.78, Fisher's exact test on a contingency table, P value = 0.4, Fig. S9). We thus conclude that the distribution of either CRISPR-Cas or R-M bacterial defense systems in the bacterial strains targeted by the phages in the lysates does not explain the network of phage–bacteria interactions we observe.

The capsule plays a major role in shaping phage infections

The first bacterial structures that interact with phages are the capsule and the LPS. We thus tested whether their serotypes shape the infection network of the *Klebsiella* prophages. We used Kaptive to serotype the strains from the genome sequence and tested if the infections were more frequent when the target bacteria and the one producing the lysate were from the same serotype. We found no significant effect for the LPS serotypes (Fig. 4a, Fisher's exact test $P = 0.37$). In contrast, we observed 35 cross-strain infections between lysates and sensitive bacteria from the same CLT out of 105 possible combinations (33%), whereas only 3.6% of the possible 1120 inter-CLT infections were observed (40 infections in total, Fig. 4a, $P < 0.0001$, Fisher's exact test). For example, strain #37 was only infected by lysates produced by the other strain with the KL24 CLT (#51) (Fig. 2). Similarly, strains #57, #55, and #25 were only infected by lysates from strains of the same CLT (KL2) (Fig. 4a). These results are independent of the genetic relatedness, as we observed infections between strains with the same CLT that are phylogenetically distant. Intriguingly, the lysate of one strain alone (#63) produced an inhibition halo in lawns of fifteen strains from different CLTs, but (as mentioned above) this may be caused by a bacteriocin present in this strain's genome. To control for the putative effect of bacteriocins in the association between infections and the CLT, we restricted the analysis to strains lacking bacteriocin homologs ($N = 22$) and those with a CLT determined with high confidence by Kaptive ($N = 29$) (see Methods). We confirmed in both cases that CLT is positively associated with cross-strain infections ($P < 0.0001$, Fisher's exact test for both).

The specificity of induced prophages for strains with a CLT similar to the original cell could be explained by the presence of depolymerases in their tail fibers. Visual observation of plaques showed that the latter are small and are not surrounded by enlarged halos typically indicative of depolymerase activity (Fig. S6). We also searched the genomes of the prophages for published depolymerase domains using protein profiles and a database of known depolymerases (see Methods, Table S2). We only found hits to two protein

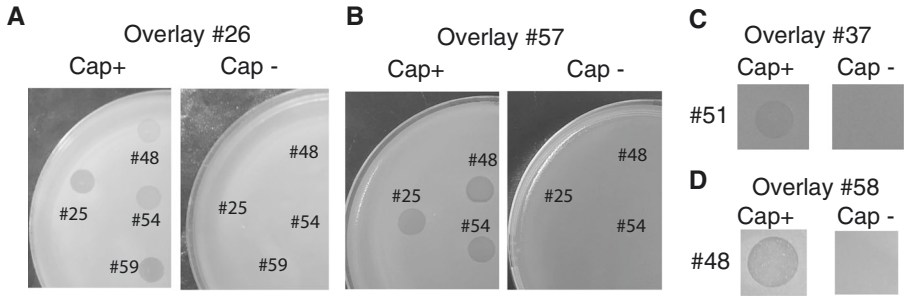


Fig. 5 The capsule is required for phage infection. Lawns of different strains, wild type (Cap+) and capsule null mutant (Cap-) in contact with PEG-precipitated supernatant of other strains. Isogenic capsule mutants of strains #26 and #37 were constructed by an in-frame deletion of *wza* gene, and a *wcaJ* deletion for strains #57 and #58 (see Methods).

profiles (Pectate_lyase_3 and Peptidase_S74) with poor *e*-value (>10^{-10}) and a low identity homolog (<50%) to one protein sequence (YP_009226010.1) (Tables S3 and S4). Moreover, these homologs of depolymerases were found in multiple prophages that are hosted in strains differing in CLTs, suggesting that they are either not functional, or that they do not target a specific CLT. It is also possible that the CLT has switched (i.e., changed by recombination [59]) since the phage originally infected the strain. This suggests that the depolymerases present in these prophages (or at least those that we could identify) do not explain the patterns observed in the interaction matrix, raising the possibility that novel depolymerases remain to be found.

It is often mentioned in the literature that capsules protect against phages [39]. However, the results above show that induced prophages tend to infect strains of similar CLT, and capsule mutants of the strain Kp36 were previously shown to become resistant to the virulent phage 117 [60]. Hence, we wondered if strains lacking a capsule would also be immune to temperate phages, and so we tested whether isogenic capsule mutants from different CLTs (KL2, KL30, and KL1) were sensitive to phages present in the lysates. We verified by microscopy and biochemical capsule quantification that these mutants lacked a capsule and then challenged them with the lysates of all 35 strains. Wild type strains #56 (K1) and #24 (K30) are resistant to all lysates, as are their capsule mutants. Capsule mutants from #26 (BJ1), #57 and #58 (all KL2), and #37 (KL24) were resistant to phages, whilst their respective wild-types are all sensitive to phage infections (Fig. 5). Hence, phages targeting these strains only infect when a capsule is present. These results show that the loss of the capsule does not make bacteria more sensitive to phages. On the contrary, the capsule seems required for infection.

Phage predation stimulates capsule loss

Given that non-capsulated mutants are resistant to phages (as shown above and also observed in other studies [60]), we hypothesized that phages that infect a particular strain could drive the loss of its capsule. To test this, we performed a short-term evolution experiment (ten days, ca 70 generations). We assessed the emergence of non-capsulated bacteria in three different strains: (i) a strain with no temperate phages (BJ1, #26); (ii) a strain that produces phages infecting many strains including itself (#54) even in the absence of MMC induction (Fig. S6C); and (iii) a strain that is resistant to its own phages (#36). Both strains #36 and #54 produce infectious phages which can lysogenize other strains (#38 and #26, respectively, Fig. 4b). Six independent populations of each strain were evolved in four different environments: (i) LB, (ii) LB supplemented with 0.2% citrate to inhibit phage adsorption, (iii) LB with MMC to increase the phage titer, and (iv) LB with MMC and 0.2% citrate to control for the effect of faster population turnover due to prophage induction and the subsequent cellular death. We expected that passages in rich medium might lead to non-capsulated mutants, even in the absence of phages, as previously observed [61–63]. This process should be accelerated if phage predation drives selection of the non-capsulated mutants (presence of phages, absence of citrate, in LB). It should be further accelerated if the intensity of phage predation increases (under MMC). As expected, all strains progressively lost their capsule albeit at different rates (Fig. 6a). To allow comparisons between treatments and strains, we calculated the area under the curve during the first five days, where most of the capsule loss took place (Fig. 6b). Strain BJ1 lacks prophages and shows no significant differences between the treatments. Strain #54 lost its capsule faster when prophages were induced (MMC) and citrate relieved this effect in accordance with the hypothesis that the speed of capsule loss depends on the efficiency of phage infection (MM+citrate). Similarly, in the treatment with citrate the capsule is lost at a slower rate than in the LB treatment. In the latter, the few events of spontaneous prophage induction generate a basal level of predation that is sufficient to increase the rate of loss of the capsule (albeit not to the levels of the MMC treatment). Note that the

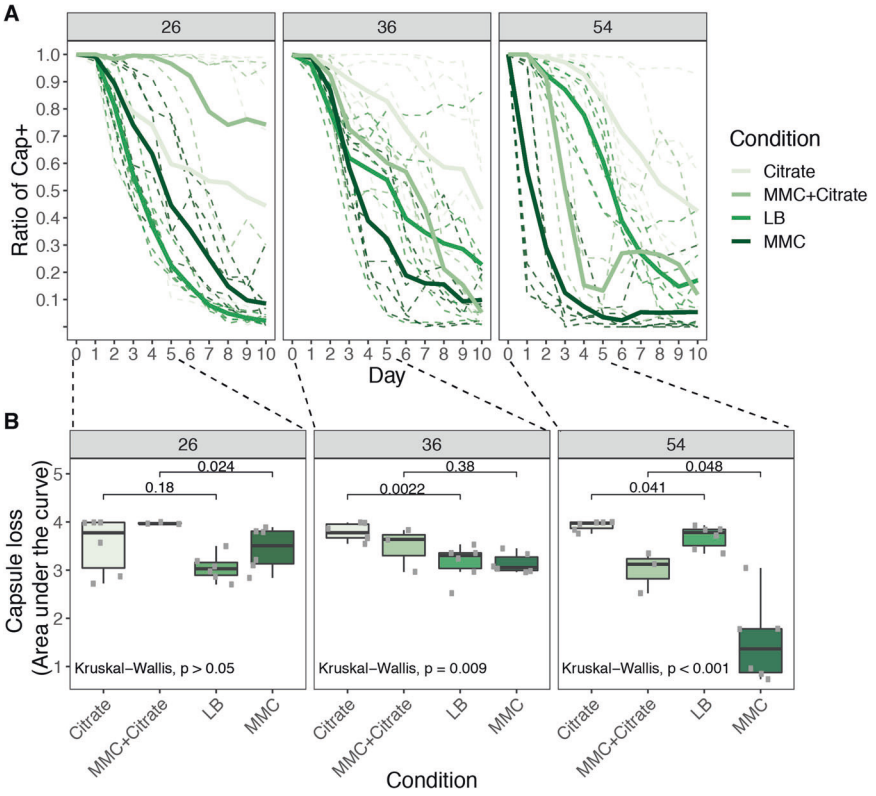


Fig. 6 Loss of capsule in three *Klebsiella* strains. a Ratio of capsulated clones throughout the ten days before daily passages of each culture. Shades of green represent the different environments in which evolution took place. MMC stands for mitomycin C. Full lines represent the average of the independent populations of the same strain and environment (shades of green). Dashed lines represent each of the independent populations. **b** Area under the curve during the first five days of the experiment.

combination of citrate and MMC did not significantly affect bacterial growth, and thus the effect we observe seems to be caused by phage predation (Figs. S6 and S10). Finally, strain #36 showed no significant difference between the experiments with MMC and LB. This suggests that the amount of phages in the environment does not affect the rate of capsule loss in this strain, consistent with it being insensitive to its own phage. Intriguingly, adding citrate lowered the rate of capsule loss in this strain, a result that suggests that even if phage infection is inefficient, there may be small deleterious impacts due to the presence of phages. Taken together, these results show that effective predation by induced prophages selects for the loss of the capsule in the lysogen.

Discussion

We provide here a first comprehensive analysis of the distribution of prophages in the *Klebsiella* genus, their genetic composition and their potential to excise, infect and lysogenize other strains. The number of prophages varies significantly across the species of the genus, but most genomes of *Klebsiella* encode for a phage or its remnants. *K. pneumoniae* is one of the species with more prophages among widely sequenced bacteria, suggesting that temperate phages are particularly important for its biology. This rapid turnover of prophages, already observed in other species, may contribute to the phenotypic differences between strains. Since many of these prophages seem to retain the ability to excise, form viable virions, and lysogenize other bacteria, they could spur adaptation when transferring adaptive traits (including antibiotic resistance in clinical settings). Further work will be needed to assess the phenotypic consequences of lysogeny and the frequency of transduction, but these results already show that the contribution of prophages to *Klebsiella* genomes is significant.

Phage–bacteria interactions shape a myriad of biological processes. Several recent studies have detailed infection networks to understand the ecological traits and the molecular mechanisms shaping them [64–66]. These analyses use isolated phages from laboratory stocks, or phages recovered from coevolution experiments [64]. In addition, these studies tend to focus on virulent phages rather than temperate, either because they envision some sort of phage therapy or because virulent phages give simpler phenotypes. Several recent computational studies have described the different phage families and the beneficial traits they may impart to their hosts, such as virulence factors [16, 65, 67, 68]. A few have explored the natural diversity of temperate phages of a species and experimentally assessed their ability to cross-infect other strains [65]. Here, we sought to generate a network of temperate phage-mediated bacterial interactions in naturally infected

Klebsiella spp, including competitive killing through phage induction and also the transfer of prophages between strains. These results are especially pertinent for *K. pneumoniae* in clinical settings because many antibiotics stimulate prophage induction [17, 69, 70] and facilitate phage infection [71, 72]. Also, phages in the mammalian gut, the most frequent habitat of *K. pneumoniae*, tend to be temperate and result from in situ prophage induction [65, 73].

We studied a large cohort of strains from the *K. pneumoniae* species complex, representing considerable diversity in terms of the number and types of prophages present in lysogenic strains. However, the extrapolation of some of the results presented here to the *Klebsiella* genus as a whole must be done with care. Further work, and a larger picture, will be possible once genomes from other clades of the genus are available. Moreover, since it would be unfeasible to generate mutants for all the strains, or perform evolution experiments with the entire cohort, some experimental assays performed here were limited to a reduced number of strains and environments. It is thus possible that some observations may not be consistently obtained with other genetic backgrounds, or ecological contexts (e.g., spatially structured environments). Nevertheless, our work considers a total of 1225 lysate–bacteria interactions, which already allows the characterization of a wide range of possible outcomes.

Our experimental setup was specifically designed to address the natural variation in the ability of lysates to infect other strains. In order to capture this variation, the network of infections was performed thrice with three independently generated lysates, which led to stochastic variation in the infection outcomes. This has been rarely characterized before for poly-lysogens. One possible explanation for the observed stochasticity in the outcomes of infections could be that the resident phages are unfit due to the accumulation of deleterious mutations during lysogeny. This could affect, for instance, the number of produced virions, and thus the infection efficiency. Alternatively, there could be a noteworthy degree of natural variation in the frequencies of each induced phage across the independently generated lysates. This means that the underrepresentation of certain phages could lead to unsuccessful infections. Having multiple prophages is expected to be costly (in terms of gene expression and cell death by induction), but extends the range of competitors that can be affected by prophage induction. Finally, infections of target bacteria with lysates from bacteria with the same CLT seem both more reproducible and effective, compared with those lysates produced from bacteria with different CLT (Fig. S7), suggesting that stochasticity in infections can also be a consequence of capsule–phage interactions, e.g., if the former is unevenly distributed throughout the cell envelope [74] or if there is stochasticity in its expression [75, 76]. Future experiments tackling diverse ecological scenarios (e.g., cocultures between three or more strains) will help understanding the causes and consequences of multiple infections producing poly-lysogens for competitive and evolutionary interactions between strains.

Overall, we found that 60% of the strains could produce at least one lysate that led to phage infection of at least one other strain. The number of strains able to produce phages is comparable with what is observed in other species like *Pseudomonas aeruginosa* (66%) [77] and *Salmonella enterica* (68%) [78]. However, we only observed 6% of all the possible cross-infections. This is much less than in *P. aeruginosa*, where a set of different lysogenic strains derived from PA14 was infected by *ca* 50% of the lysates tested [77]. Similarly, in *S. enterica*, ~35% of cross-infections were effective [78]. This suggests that the likelihood that a prophage from one strain is able to infect another *Klebsiella* strain is relatively small. Hence, when two different *Klebsiella* strains meet, they will be very often immune to the prophages of the other strains. This implicates that prophages would be less efficient in increasing the competitive ability of a *Klebsiella* strain than in other species. Finally, this also implies that phage-mediated HGT in *Klebsiella* may not be very efficient in spreading traits across the species, which means that, in some situations, the capsule could slow down evolution by phage-mediated horizontal gene transfer.

Interestingly, 5 out of 35 strains can be infected by their own induced prophages. It is commonly assumed that lysogens are always resistant to reinfection by the same phage [79], but the opposite is not unheard of. It has been previously reported that *E. coli* strains could be reinfected by the same phage twice, both by phage lambda [80] and by phage P1 [81]. At this time, we can only offer some speculations for this lack of superinfection immunity. For example, it has been estimated that there are only ten free dimmers of Lambda's cI repressor in a typical cell [82]. The infection of several phages at the same time may titer this limited amount of repressor, such that some incoming phages are able to induce resident prophages to enter a lytic cycle. This could be amplified when prophages accumulate deleterious mutations in the repressor or in the binding sites of the repressor [83–85], further decreasing its ability to silence the prophage or the incoming phage. This could also explain why we observe frequent prophage spontaneous induction (strain #54, Fig. S6C).

The small number of cross-infectivity events could not be attributed to bacterial defense systems (R-M or CRISPR-Cas) or LPS serotypes. The similarity between prophages or their repressor proteins, potentially facilitating superinfection immunity [77], also failed to explain the observed infection patterns. Instead, the relatively few cases of phage infection in the matrices are grouped in modules that seem determined by the capsule composition, since there is a vast overrepresentation of infections between bacteria of the same CLT. This seems independent of their genetic

relatedness, since we observed infections between strains with the same CLT that are phylogenetically distant. This CLT specificity is consistent with a recent study that focuses on a carbapenemase-producing *K. pneumoniae* CG258, in which very few phages could lyse bacteria with different CLT [86]. The CLT specificity of phage-encoded depolymerases [40–42, 44, 47, 48, 87, 88] could explain these results. Yet, in our set, few of such enzymes were detected, they exhibited very low identity to known depolymerases, and did not seem to correlate with the CLT. At this stage, it is difficult to know if depolymerases are rare in temperate phages or if they are just too different from known depolymerases, since these were mostly identified from virulent phages. If depolymerases are indeed rare in temperate phages, this raises the important question of the alternative mechanisms that underlie their capsule specificity.

Most of the literature on other species concurs that the capsule is a barrier to phages by limiting their adsorption and access to cell receptors [37–39]. In contrast, our results show that the capsule is often required for infection by *Klebsiella* induced prophages. This is in agreement with a recent study where inactivation of *wcaJ*, a gene essential for capsule synthesis, rendered the strain resistant to phage infection by a virulent phage [60]. Specific interactions between the phage and the capsule could be caused by the latter stabilizing viral adsorption or allowing more time for efficient DNA injection [49]. This may have resulted in phages selecting for the ability to recognize a given capsule serotype. The specificity of the interactions between temperate phages and bacteria, caused by capsule composition, has outstanding implications for the ecology of these phages because it severely limits their host range. Indeed, we report very few cases (*ca*. 3%) of phages infecting strains with different CLTs, and this could be an overestimation if we discard strain #63 because of its potential bacteriocin activity. The consequence of this evolutionary process is that phage pressure results in selection for the loss of capsule because non-capsulated bacteria are resistant to phage infection. Our experimental evolution captures the first steps of this coevolution dynamic, suggesting that phage predation selects very strongly for capsule loss in the infected strains. Since capsules are prevalent in *Klebsiella*, this suggests that it may be reacquired later on.

Our results might also provide insights regarding the possible use of phages to fight the increasing challenge of antibiotic resistance in *Klebsiella* infections. Although more work is needed to understand how to best use virulent phages to control *Klebsiella* infections, our results already hint that phage therapy may, at least in a first step, lead to capsule loss. While such treatments may be ineffective in fully clearing *Klebsiella* infections (due to the large diversity of existing serotypes), they can select for non-capsulated mutants. The latter are expected to be less virulent [89], because the capsule is a major virulence factor in *K. pneumoniae*. The increase in frequency of non-capsulated mutants may also increase the efficiency of traditional antimicrobial therapies, as the capsule is known to increase tolerance to chemical aggressions, including antibiotics and cationic antimicrobial peptides produced by the host [89].

Materials and methods

Strains and growth conditions

We used 35 *Klebsiella* strains that were selected based on MLST data, and representative of the phylogenetic and clonal diversity of the *K. pneumoniae* species complex [24]. Strains were grown in LB at 37° and under shaking conditions (250 rpm).

Genomes

254 genomes of *Klebsiella* species (of which 197 of *K. pneumoniae*) and one *Cedecea* sp (outgroup) were analysed in this study. This included all complete genomes belonging to *Klebsiella* species from NCBI, downloaded February 2018, and 29 of our own collection [24]. We corrected erroneous NCBI species annotations using Kleborate typing [28]. 253 genomes were correctly assigned to its species, with a "strong" level of confidence, as annotated by Kaptive. Only two genomes were classified with a "weak" level of confidence by the software. All information about these genomes is presented in Dataset S1.

Identification of prophages

To identify prophages, we used a freely available computational tool, PHASTER [50], and analysed the genomes on September 2018. The completeness or potential viability of identified prophages are identified by PHASTER as "intact", "questionable" or "incomplete" prophages. All results presented here were performed on the "intact" prophages, unless stated otherwise. Results for all prophages ("questionable", "incomplete") are presented in the supplemental material. Primers used for phage detection and recircularization are presented in Table S5.

Prophage characterization

(i) *Prophage delimitation*. Prophages are delimited by the *attL* and *attR* recombination sites used for phage recircularization and theta replication. In some instances, these sites were predicted by PHASTER. When none were found,

we manually searched for them either by looking for similar *att* sites in related prophages or by searching for interrupted core bacterial genes. (ii) *Functional annotation* was performed by combining multiple tools: *prokka* v1.14.0 [90], pVOG profiles [91] searched for using *HMMER* v3.2.1 [92], the PHASTER Prophage/Virus DB (version Aug 14, 2019), *BLAST 2.9.0+* [93], and the *ISFinder* webtool [94]. For each protein and annotation tool, all significant matches (*e*-value < 10^{-5}) were kept and categorized in dictionaries. If a protein was annotated as "tail" in the description of a matching pVOG profile or PHASTER DB protein, the gene was categorized as tail. Results were manually curated for discrepancies and ties. For proteins matching more than one pVOG profile, we attributed the Viral Quotient (VQ) associated to the best hit (lowest *e*-value). The VQ is a measure of how frequent a gene family is present in phages, and ranges from zero to one with higher values meaning that the family is mostly found in viruses. (iii) *Repressor identification*. Repressors in the prophages were detected using specific HMM protein profiles for the repressor, available in the pVOG database [91]. We selected those with a VQ higher than 0.8. These profiles were matched to the *Klebsiella* intact prophage sequences using HMMER v3.1b2, and we discarded the resulting matches whose best domain had an *e*-value of more than 0.0001 or a coverage of less than 60%. We further selected those with at least one of the following terms in the descriptions of all the proteins that compose the HMM: immunity, superinfection, repressor, exclusion. This resulted in a set of 28 profiles (Table S6). The similarity between repressors was inferred from their alignments (assessed with the *align.globalxx* function in from the *pairwise2* module in biopython, v1.74), by dividing the number of positions matched by the size of the smallest sequence (for each pair of sequences).

Core genome

(i) *Klebsiella* spp. ($N = 255$) (Figs. 1a and S1). The core genome was inferred as described in [95]. Briefly, we identified a preliminary list of orthologs between pairs of genomes as the list of reciprocal best hits using end-gap free global alignment, between the proteome of a pivot and each of the other strains proteome. Hits with less than 80% similarity in amino acid sequences or more than 20% difference in protein length were discarded. (ii) *Laboratory collection of Klebsiella genomes* ($N = 35$) (Fig. 4). The pangenome was built by clustering all protein sequences with Mmseqs2 (v1-c7a89) [96] with the following arguments: *-cluster-mode 1* (connected components algorithm), *-c 0.8 –cov-mode 0* (default, coverage of query and target >80%) and *–min-seq-id 0.8* (minimum identity between sequences of 80%). The core genome was taken from the pangenome by retrieving families that were present in all genomes in single copy.

Phylogenetic trees

To compute both phylogenetic trees, we used a concatenate of the multiple alignments of the core genes aligned with MAFFT v7.305b (using default settings). The tree was computed with IQ-Tree v1.4.2 under the GTR model and a gamma correction (GAMMA). We performed 1000 ultrafast bootstrap experiments (options *–bb* 1000 and *–wbtl*) on the concatenated alignments to assess the robustness of the tree topology. The vast majority of nodes were supported with bootstrap values higher than 90% (Fig. S1). The *Klebsiella* spp. ($N = 255$) tree had 1,106,022 sites, of which 263,225 parsimony-informative. The phylogenetic tree of the laboratory collection strains ($N = 35$) was built using 2,800,176 sites, of which 286,156 were parsimony-informative.

Capsule serotyping

We used Kaptive, integrated in Kleborate, with default options [28]. Serotypes were assigned with overall high confidence levels by the software (see Dataset S1): 13 were a perfect match to the sequence of reference, 144 had very high confidence, 33 high, 35 good, 11 low, and 19 none. From the strains used in the experiments, two were a perfect match to reference strains, 20 had very high confidence, 3 were high, 7 good and 3 for which the assignment had low confidence.

Bacteriocin detection

We checked for the presence of bacteriocins and other bacterial toxins using BAGEL4 [97]. The results are reported in Table S1.

Depolymerase detection

We checked for the presence of 14 different HMM profiles associated with bacteriophage-encoded depolymerases from multiple bacterial species [98], Table S2). The profiles were matched against the complete set of prophage proteins using HMMER v3.1b2, filtering by the *e*-value of the best domain (maximum 10^{-3}) and the coverage of the profile (minimum 30%). Five additional sequences of depolymerases, validated experimentally in lytic phages of *Klebsiella* (see references in Table S1), were also matched against our dataset using *blastp* (version 2.7.1+, default parameters, and filtering by the *e*-value (maximum 10^{-5}), identity (40%), and coverage (40%).

Identification of CRISPR arrays and R-M systems

(i) *CRISPR-Cas arrays*. We used CRISPRCasFinder [99] (v4.2.18, default parameters) to identify the CRISPR arrays

in all *Klebsiella* genomes used in this study. We excluded arrays with less than three spacers. We then matched each spacer sequence in each array with the complete prophage nucleotide sequences *(blastn* version 2.7.1+, with the -task blastn-short option). Only matches with a spacer coverage of at least 90%, a maximum *e*-value of 10^{-5} and a minimum nucleotide identity of 90% were retained. The resulting matches indicate prophages that are targeted by these spacers, and the full set of these results are presented in Dataset S1. (ii) *R-M systems* were identified using the highly specific and publicly available HMM profiles in https://gitlab.pasteur.fr/erocha/RMS_scripts. If a single protein matched multiple systems, the best hit for each protein-R-M pair was selected. To assess the likelihood that a strain can defend itself against infection by prophages induced from another strain using R-M systems, we calculated the similarity between these proteins (all versus all) using BLAST (*blastp* version 2.7.1+, filtering by identity >50%, *e*-value < 0.0001). We then considered that two R-M associated proteins can target similar recognition sites if their identity is either at above 50% (for Type II and IV REases), 55% (for Type IIG), 60% (for Type II MTases)), or 80% (for Type I and III MTases and Type I and III REases), according to [100]. We inferred the recognition sites targeted by these systems using REBASE (http://rebase.neb.com/rebase/rebase.html, default parameters). We selected the recognition site associated to the protein with the best score, and further chose those whose identity obeyed the thresholds enumerated above (for each type of RMS). The nucleotide motifs associated with these recognition sites were searched for in the intact prophage genomes from the 35 bacterial strains using the Fuzznuc program from the EMBOSS suite (version 6.6.0), with the default parameters, using the option to search the complementary strand when the motif was not a palindrome. Finally, for each pair of bacterial strains A–B (where strain A produces the lysate and strain B is used as a bacterial lawn), we looked for prophages in strain A whose genomes contain recognition sites targeted by R-M systems present in strain B and absent from strain A. If a recognition site was found in any prophage from strain A, we consider that prophage to be putatively targeted by (at least) one R-M defense system of strain B. Because some R-M systems require two (similar) recognition sites in order to effectively target incoming DNA [101], we also did a separate analysis where we require that each motif is found twice in each individual prophage genome. This analysis resulted in similar qualitative results.

Prophage experiments

(i) *Growth curves:* 200 µL of diluted overnight cultures of *Klebsiella* spp. (1:100 in fresh LB) were distributed in a 96-well plate. Cultures were allowed to reach $OD_{600} = 0.2$ and either MMC to 0, 1, or 3 µg/mL or PEG-precipitated induced and filtered supernatants at different PFU/ml was added. Growth was then monitored until late stationary phase. (ii) *PEG-precipitation of virions*. Overnight cultures were diluted 1:500 in fresh LB and allowed to grow until $OD_{600} = 0.2$. MMC was added to final 5 µg/mL. In parallel, non-induced cultures were grown. After 4 h at 37 °C, cultures were centrifuged at 4000 rpm and the supernatant was filtered through 0.22 um. Filtered supernatants were mixed with chilled PEG-NaCl 5X (PEG 8000 20% and 2.5 M of NaCl) and mixed through inversion. Virions were allowed to precipitate for 15 min and pelleted by centrifugation 10 min at 13,000 rpm at 4 °C. The pellets were dissolved in TBS (Tris Buffer Saline, 50 mM Tris-HCl, pH 7.5, 150 mM NaCl). (iii) *All-against-all infection*. Overnight cultures of all strains were diluted 1:100 and allowed to grow until $OD_{600} = 0.8$. 1 mL of bacterial cultures were mixed with 12 mL of top agar (0.7% agar), and 3 mL of the mixture was poured onto a prewarmed LB plate and allowed to dry. Ten microliters of freshly prepared and PEG-precipitated lysates were spotted on the top agar and allowed to grow for 4 h at 37° prior to assessing clearance of bacterial cultures. This was repeated in three independent temporal blocks. (iv) *Calculating plaque forming units (PFU)*. Overnight cultures of sensitive strains were diluted 1:100 and allowed to grow until $OD_{600} = 0.8$. 250 µL of bacterial cultures were mixed with 3 mL of top agar (0.7% agar) and poured intro prewarmed LB plates. Plates were allowed to dry before spotting serial dilutions of induced and non-induced PEG-precipitate virions. Plates were left overnight at room temperature and phage plaques were counted.

Evolution experiment

Three independent clones of each strain (#54, #26, and #36) were used to initiate each evolving of the three evolving populations in four different environments: (i) LB, (ii) LB supplemented with 0.2% citrate, (iii) LB with mytomycin C (MMC, 0.1 µg/mL), and (iv) LB with MMC (0.1 µg/mL) and supplemented with 0.2% citrate. Populations were allowed to grow for 24 h at 37°. Each day, populations were diluted 1:100 and plated on LB to count for capsulated and non-capsulated phenotypes. This experiment was performed in two different temporal blocks and its results combined.

wGRR calculations and network building

We searched for sequence similarity between all proteins of all phages using mmseqs2 [96] with the sensitivity parameter set at 7.5. The results were converted to the blast format for analysis and were filtered with the following parameters: *e*-value lower than 0.0001, at least 35% identity

between amino acids, and a coverage of at least 50% of the proteins. The filtered hits were used to compute the set of bi-directional best hits between each phage pair. This was then used to compute a score of gene repertoire relatedness for each pair of phage genomes, weighted by sequence identity, computed as following:

$$wGRR = \frac{\sum_i^p \mathrm{id}(\mathrm{A}_i, \mathrm{B}_i)}{\min(\#\mathrm{A}, \#\mathrm{B})},$$

where A_i and B_i is the pair *i* of homologous proteins present in *A* and *B* (containing respectively #A and #B proteins and *p* homologs), id(A_i,B_i) is the percent sequence identity of their alignment, and min(#*A*,#*B*) is the total number of proteins of the smallest prophage, the one encoding the smallest number of proteins (*A* or *B*). wGRR varies between zero and one. It amounts to zero if there are no orthologs between the elements, and one if all genes of the smaller phage have an ortholog 100% identical in the other phage. Hence, the wGRR accounts for both frequency of homology and degree of similarity among homologs. For instance, three homologous genes with 100% identity between two phages, where the one with the smallest genome is 100 proteins long, would result in a wGRR of 0.03. The same wGRR value would be obtained with six homologous genes with 50% identity. The phage network was built with these wGRR values, using the *networkx* and *graphviz* Phyton (v2.7) packages, and the *neato* algorithm.

Generation of capsule mutant

Isogenic capsule mutants were generated by an in-frame knock-out deletion of gene *wza* (strain #26, #24, #56, #37) or gene *wcaJ* (strain #57 and #58) by allelic exchange. Upstream and downstream sequences of the *wza* or *wcaJ* gene (>500 pb) were amplified using Phusion Master Mix then joined by overlap PCR. All primers used are listed in Table S5. The PCR product was purified using the QIAquick Gel Extraction Kit after electrophoresis in agarose gel 1% and then cloned with the Zero Blunt® TOPO® PCR Cloning Kit (Invitrogen) into competent *E. coli* DH5α strain. KmR colonies were isolated and checked by PCR. A positive Zero Blunt® TOPO® plasmid was extracted using the QIAprep Spin Miniprep Kit, digested for 2 h at 37 °C with ApaI and SpeI restriction enzymes and ligated with T4 DNA ligase overnight to digested pKNG101 plasmid. The ligation was transformed into competent *E. coli* DH5α pir strain, and again into *E. coli* MFD λ−pir strain [102], which was used as a donor strain for conjugation into *Klebsiella* spp. Conjugations were performed for 24 h at 37°. Single cross-over mutants (transconjugants) were selected on Streptomycin plates (200 µg/mL). Plasmid excision was performed after a 48 h culture incubated at 25 °C and double cross-over mutants were selected on LB without salt plus 5% sucrose at the room temperature. To confirm the loss of the plasmid, colonies were tested for their sensitivity to streptomycin and mutants were confirmed by PCR across the deleted region and further verified by Illumina sequencing.

Data availability

Raw data are available in Dryad: https://doi.org/10.5061/dryad.qfttdz0f3.

Acknowledgements We thank Pedro Oliveira for making available the profiles for the RMS systems, Marie Touchon for help in the analysis of CRISPR, and Amandine Perrin for help with building pan-genomes.

Funding MH is funded by an ANR JCJC (Agence national de recherche) grant [ANR 18 CE12 0001 01 ENCAPSULATION] awarded to OR. JAMS is supported by an ANR grant [ANR 16 CE12 0029 02 SALMOPROPHAGE] awarded to EPCR The laboratory is funded by a Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (grant ANR-10-LABX-62-IBEID). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

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

References

1. Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H. Prophage genomics. Microbiol Mol Biol Rev. 2003;67:238–76.
2. Lu MJ, Henning U. Superinfection exclusion by T-even-type coliphages. Trends Microbiol. 1994;2:137–9.
3. Susskind MA, Wright A, Botstein D. Superinfection exclusion by P22 prophage in lysogens of *Salmonella typhimurium*. IV. genetics and physiology of sieB exclusion. Virology. 1974;62:367–84.
4. Asadulghani M, Ogura Y, Ooka T, Itoh T, Sawaguchi A, Iguchi A, et al. The defective prophage pool of *Escherichia coli* O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. PLoS Pathog. 2009;5:e1000408.
5. Matos RC, Lapaque N, Rigottier-Gois L, Debarbieux L, Meylheuc T, Gonzalez-Zorn B, et al. *Enterococcus faecalis* prophage

dynamics and contributions to pathogenic traits. PLoS Genet. 2013;9:e1003539.
6. Touchon M, Bobay LM, Rocha EP. The chromosomal accommodation and domestication of mobile genetic elements. Curr Opin Microbiol. 2014a;22:22–9.
7. Nakayama K, Takashima K, Ishihara H, Shinomiya T, Kageyama M, Kanaya S, et al. The R-type pyocin of *Pseudomonas aeruginosa* is related to P2 phage, and the F-type is related to lambda phage. Mol Microbiol. 2000;38:213–31.
8. Winstanley C, Langille MG, Fothergill JL, Kukavica-Ibrulj I, Paradis-Bleau C, Sanschagrin F, et al. Newly introduced genomic prophage islands are critical determinants of *in vivo* competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. Genome Res. 2009;19:12–23.
9. Bossi L, Fuentes JA, Mora G, Figueroa-Bossi N. Prophage contribution to bacterial population dynamics. J Bacteriol. 2003; 185:6467–71.
10. Fortier LC, Sekulovic O. Importance of prophages to evolution and virulence of bacterial pathogens. Virulence. 2013;4:354–65.
11. Nanda AM, Thormann K, Frunzke J. Impact of spontaneous prophage induction on the fitness of bacterial populations and host-microbe interactions. J Bacteriol. 2015;197:410–9.
12. Brown SP, Le Chat L, De Paepe M, Taddei F. Ecology of microbial invasions: amplification allows virus carriers to invade more rapidly when rare. Curr Biol. 2006;16:2048–52.
13. Sousa JAM, Rocha EPC. Environmental structure drives resistance to phages and antibiotics during phage therapy and to invading lysogens during colonisation. Sci Rep. 2019;9:3149.
14. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife. 2015;4:e08490.
15. Touchon M, Bernheim A, Rocha EP. Genetic and life-history traits associated with the distribution of prophages in bacteria. ISME J. 2016;10:2744–54.
16. Bobay LM, Rocha EP, Touchon M. The adaptation of temperate bacteriophages to their host genomes. Mol Biol Evol. 2013; 30:737–51.
17. Wagner PL, Waldor MK. Bacteriophage control of bacterial virulence. Infect Immun. 2002;70:3985–93.
18. Chen J, Quiles-Puchalt N, Chiang YN, Bacigalupe R, Fillol-Salom A, Chee MSJ, et al. Genome hypermobility by lateral transduction. Science. 2018;362:207–12.
19. Touchon M, Moura de Sousa JA, Rocha EP. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. Curr Opin Microbiol. 2017;38:66–73.
20. Haaber J, Leisner JJ, Cohn MT, Catalan-Moreno A, Nielsen JB, Westh H, et al. Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. Nat Commun. 2016;7:13333.
21. Brisse S, Grimont F, Grimont PAD. The genus *Klebsiella*. The Prokaryotes. New York, USA: Springer; 2006. p. 159–96.
22. Lee CR, Lee JH, Park KS, Jeon JH, Kim YB, Cha CJ, et al. Antimicrobial resistance of hypervirulent *Klebsiella pneumoniae*: epidemiology, hypervirulence-associated determinants, and resistance mechanisms. Front Cell Infect Microbiol. 2017;7:483.
23. Navon-Venezia S, Kondratyeva K, Carattoli A. *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. FEMS Microbiol Rev. 2017;41:252–75.
24. Blin C, Passet V, Touchon M, Rocha EPC, Brisse S. Metabolic diversity of the emerging pathogenic lineages of *Klebsiella pneumoniae*. Environ Microbiol. 2017;19:1881–98.
25. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. PLoS Genet. 2019;15:e1008114.
26. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. Proc Natl Acad Sci USA. 2015;112:E3574–81.
27. Mori M, Ohta M, Agata N, Kido N, Arakawa Y, Ito H, et al. Identification of species and capsular types of *Klebsiella* clinical isolates, with special reference to *Klebsiella planticola*. Microbiol Immunol. 1989;33:887–95.
28. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. Identification of *Klebsiella* capsule synthesis loci from whole genome data. Micro Genom. 2016;2:e000102.
29. Pan YJ, Lin TL, Chen CT, Chen YY, Hsieh PF, Hsu CR, et al. Genetic analysis of capsular polysaccharide synthesis gene clusters in 79 capsular types of *Klebsiella* spp. Sci Rep. 2015;5:15573.
30. Favre-Bonte S, Licht TR, Forestier C, Krogfelt KA. *Klebsiella pneumoniae* capsule expression is necessary for colonization of large intestines of streptomycin-treated mice. Infect Immun. 1999;67:6152–6.
31. Alvarez D, Merino S, Tomas JM, Benedi VJ, Alberti S. Capsular polysaccharide is a major complement resistance factor in lipopolysaccharide O side chain-deficient *Klebsiella pneumoniae* clinical isolates. Infect Immun. 2000;68:953–5.
32. Campos MA, Vargas MA, Regueiro V, Llompart CM, Alberti S, Bengoechea JA. Capsule polysaccharide mediates bacterial resistance to antimicrobial peptides. Infect Immun. 2004;72:7107–14.
33. Doorduijn DJ, Rooijakkers SHM, van Schaik W, Bardoel BW. Complement resistance mechanisms of *Klebsiella pneumoniae*. Immunobiology. 2016;221:1102–9.
34. Rendueles O, Garcia-Garcera M, Neron B, Touchon M, Rocha EPC. Abundance and co-occurrence of extracellular capsules increase environmental breadth: Implications for the emergence of pathogens. PLoS Pathog. 2017;13:e1006525.
35. Rendueles O, de Sousa JAM, Bernheim A, Touchon M, Rocha EPC. Genetic exchanges are more frequent in bacteria encoding capsules. PLoS Genet. 2018;14:e1007862.
36. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46:305–9.
37. Moller AG, Lindsay JA, Read TD. Determinants of phage host range in *Staphylococcus* species. Appl Environ Microbiol. 2019;85:e00209–19.
38. Negus D, Burton J, Sweed A, Gryko R, Taylor PW. Poly-gamma-(D)-glutamic acid capsule interferes with lytic infection of *Bacillus anthracis* by *B. anthracis*-specific bacteriophages. Appl Environ Microbiol. 2013;79:714–7.
39. Scholl D, Adhya S, Merril C. *Escherichia coli* K1's capsule is a barrier to bacteriophage T7. Appl Environ Microbiol. 2005;71:4872–4.
40. Niemann H, Frank N, Stirm S. *Klebsiella* serotype-13 capsular polysaccharide: primary structure and depolymerization by a bacteriophage-borne glycanase. Carbohydr Res. 1977a;59:165–77.
41. Niemann H, Kwiatkowski B, Westphal U, Stirm S. *Klebsiella* serotype 25 capsular polysaccharide: primary structure and depolymerization by a bacteriophage-borne glycanase. J Bacteriol. 1977b;130:366–74.
42. Pan YJ, Lin TL, Chen CC, Tsai YT, Cheng YH, Chen YY, et al. *Klebsiella* phage PhiK64-1 encodes multiple depolymerases for multiple host capsular types. J Virol. 2017;91:e02457–02416.
43. Bessler W, Freund-Molbert E, Knufermann H, Rduolph C, Thurow H, Stirm S. A bacteriophage-induced depolymerase active on *Klebsiella* K11 capsular polysaccharide. Virology. 1973;56:134–51.
44. Thurow H, Niemann H, Rudolph C, Stirm S. Host capsule depolymerase activity of bacteriophage particles active on *Klebsiella* K20 and K24 strains. Virology. 1974;58:306–9.

45. Latka A, Maciejewska B, Majkowska-Skrobek G, Briers Y, Drulis-Kawa Z. Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during the infection process. Appl Microbiol Biotechnol. 2017;101:3103–19.
46. Scholl D, Rogers S, Adhya S, Merril CR. Bacteriophage K1-5 encodes two different tail fiber proteins, allowing it to infect and replicate on both K1 and K5 strains of *Escherichia coli*. J Virol. 2001;75:2509–15.
47. Lin TL, Hsieh PF, Huang YT, Lee WC, Tsai YT, Su PA, et al. Isolation of a bacteriophage and its depolymerase specific for K1 capsule of *Klebsiella pneumoniae*: implication in typing and treatment. J Infect Dis. 2014;210:1734–44.
48. Pan YJ, Lin TL, Chen YY, Lai PH, Tsai YT, Hsu CR, et al. Identification of three podoviruses infecting *Klebsiella* encoding capsule depolymerases that digest specific capsular types. Micro Biotechnol. 2019;12:472–86.
49. Bertozzi Silva J, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. FEMS Microbiol Lett. 2016;363:fnw002.
50. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016;44:W16–21.
51. Daubin V, Lerat E, Perriere G. The source of laterally transferred genes in bacterial genomes. Genome Biol. 2003;4:R57.
52. Rocha EP, Danchin A. Base composition bias might result from competition for metabolic resources. Trends Genet. 2002;18:291–4.
53. Bobay LM, Touchon M, Rocha EP. Pervasive domestication of defective prophages by bacteria. Proc Natl Acad Sci USA. 2014;111:12127–32.
54. Casjens SR, Gilcrease EB, Huang WM, Bunny KL, Pedulla ML, Ford ME, et al. The pKO2 linear plasmid prophage of *Klebsiella oxytoca*. J Bacteriol. 2004;186:1818–32.
55. Leclercq S, Cordaux R. Do phages efficiently shuttle transposable elements among prokaryotes? Evolution. 2011;65:3327–31.
56. Ghazaryan L, Tonoyan L, Ashhab AA, Soares MI, Gillor O. The role of stress in colicin regulation. Arch Microbiol. 2014;196:753–64.
57. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. Nat Rev Microbiol. 2010;8:317–27.
58. Samson JE, Magadan AH, Sabri M, Moineau S. Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol. 2013;11:675–87.
59. Wyres KL, Gorrie C, Edwards DJ, Wertheim HF, Hsu LY, Van Kinh N, et al. Extensive capsule locus variation and large-scale genomic recombination within the *Klebsiella pneumoniae* clonal group 258. Genome Biol Evol. 2015;7:1267–79.
60. Tan D, Zhang Y, Qin J, Le S, Gu J, Chen LK et al. A frameshift mutation in *wcaJ* associated with phage resistance in *Klebsiella pneumoniae*. Microorganisms. 2020;8:378.
61. Buffet A, Rocha EPC, Rendueles O. Selection for the bacterial capsule in the absence of biotic and abiotic aggressions depends on growth conditions. BioRxiv. 2020. https://doi.org/10.1101/2020.04.27.059774.
62. Julianelle LA. Bacterial variation in cultures of Friedlander's *Bacillus*. J Exp Med. 1928;47:889–902.
63. Randall WA. Colony and antigenic variation in *Klebsiella pneumoniae* types A, B and C. J Bacteriol. 1939;38:461–77.
64. Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS. Statistical structure of host-phage interactions. Proc Natl Acad Sci USA. 2011;108:E288–97.
65. Mathieu A, Dion M, Deng L, Tremblay D, Moncaut E, Shah SA, et al. Virulent coliphages in 1-year-old children fecal samples are fewer, but more infectious than temperate coliphages. Nat Commun. 2020;11:385.
66. Weitz JS, Poisot T, Meyer JR, Flores CO, Valverde S, Sullivan MB, et al. Phage-bacteria infection networks. Trends Microbiol. 2013;21:82–91.
67. Brueggemann AB, Harrold CL, Rezaei Javan R, van Tonder AJ, McDonnell AJ, Edwards BA. Pneumococcal prophages are diverse, but not without structure or history. Sci Rep. 2017;7:42976.
68. Castillo D, Kauffman K, Hussain F, Kalatzis P, Rorbo N, Polz MF, et al. Widespread distribution of prophage-encoded virulence factors in marine *Vibrio* communities. Sci Rep. 2018;8:9973.
69. Allen HK, Looft T, Bayles DO, Humphrey S, Levine UY, Alt D, et al. Antibiotics in feed induce prophages in swine fecal microbiomes. mBio. 2011;2:e00260–11.
70. Otsuji N, Sekiguchi M, Iijima T, Takagi Y. Induction of phage formation in the lysogenic *Escherichia coli* K-12 by mitomycin C. Nature. 1959;184:1079–80.
71. Comeau AM, Tetart F, Trojet SN, Prere MF, Krisch HM. Phage-antibiotic synergy (PAS): beta-lactam and quinolone antibiotics stimulate virulent phage growth. PLoS One. 2007;2:e799.
72. Kim M, Jo Y, Hwang YJ, Hong HW, Hong SS, Park K et al. Phage-antibiotic synergy via delayed lysis. Appl Environ Microbiol. 2018;84:e02085–18.
73. De Paepe M, Tournier L, Moncaut E, Son O, Langella P, Petit MA. Carriage of lambda latent virus is costly for its bacterial host due to frequent reactivation in monoxenic mouse intestine. PLoS Genet. 2016;12:e1005861.
74. Phanphak S, Georgiades P, Li R, King J, Roberts IS, Waigh TA. Super-resolution fluorescence microscopy study of the production of K1 capsules by *Escherichia coli*: evidence for the differential distribution of the capsule at the poles and the equator of the cell. Langmuir. 2019;35:5635–46.
75. Krinos CM, Coyne MJ, Weinacht KG, Tzianabos AO, Kasper DL, Comstock LE. Extensive surface diversity of a commensal microorganism by multiple DNA inversions. Nature. 2001;414:555–8.
76. Tzeng YL, Thomas J, Stephens DS. Regulation of capsule in *Neisseria meningitidis*. Crit Rev Microbiol. 2016;42:759–72.
77. Bondy-Denomy J, Qian J, Westra ER, Buckling A, Guttman DS, Davidson AR, et al. Prophages mediate defense against phage infection through diverse mechanisms. ISME J. 2016;10:2854–66.
78. Zhang YF, LeJeune JT. Transduction of bla(CMY-2), tet(A), and tet(B) from *Salmonella enterica* subspecies enterica serovar Heidelberg to S-Typhimurium. Vet Microbiol. 2008;129:418–25.
79. Brussow H, Kutter E. Phage ecology. Bacteriophages: biology and application. Boca Raton, Florida: CRC Press; 2005. p. 129–64.
80. Calef E, Marchelli C, Guerrini F. The formation of superinfection-double lysogens of phage lambda in *Escherichia coli* K-12. Virology. 1965;27:1–10.
81. Scott JR, West BW, Laping JL. Superinfection immunity and prophage repression in phage P1. IV. The c1 repressor bypass function and the role of c4 repressor in immunity. Virology. 1978;85:587–600.
82. Bakk A, Metzler R. Nonspecific binding of the OR repressors CI and Cro of bacteriophage lambda. J Theor Biol. 2004;231:525–33.
83. Casjens S. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol. 2003;49:277–300.
84. Hendrix RW, Smith MCM, Burns RN, Ford ME, Hatfull GF. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. Proc Natl Acad Sci USA. 1999;96:2192–7.
85. Ramisetty BCM, Sudhakari PA. Bacterial 'Grounded' Prophages: hotspots for genetic renovation and innovation. Front Genet. 2019;10:65.
86. Venturini C, Ben Zakour N, Bowring B, Morales S, Cole R, Kovach Z et al. *K. pneumoniae* ST258 genomic variability and bacteriophage susceptibility. bioRxiv. 2019. https://doi.org/10.1101/628339.
87. Hsieh PF, Lin HH, Lin TL, Chen YY, Wang JT. Two T7-like bacteriophages, K5-2 and K5-4, each encodes two capsule depolymerases: isolation and functional characterization. Sci Rep. 2017;7:4624.

88. Majkowska-Skrobek G, Latka A, Berisio R, Squeglia F, Maciejewska B, Briers Y, et al. Phage-borne depolymerases decrease *Klebsiella pneumoniae* resistance to innate defense mechanisms. Front Microbiol. 2018;9:00209–19.
89. Paczosa MK, Mecsas J. *Klebsiella pneumoniae*: going on the offense with a strong defense. Microbiol Mol Biol Rev. 2016;80: 629–61.
90. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
91. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017;45: D491–D498.
92. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195.
93. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinforma. 2009;10:421.
94. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34:D32–6.
95. Touchon M, Cury J, Yoon EJ, Krizova L, Cerqueira GC, Murphy C, et al. The genomic diversification of the whole Acinetobacter genus: origins, mechanisms, and consequences. Genome Biol Evol. 2014b;6:2866–82.
96. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–8.
97. van Heel AJ, de Jong A, Song C, Viel JH, Kok J, Kuipers OP. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. Nucleic Acids Res. 2018; 46:W278–81.
98. Pires DP, Oliveira H, Melo LD, Sillankorva S, Azeredo J. Bacteriophage-encoded depolymerases: their diversity and biotechnological applications. Appl Microbiol Biotechnol. 2016; 100:2141–51.
99. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Neron B, et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. 2018;46:W246–51.
100. Oliveira PH, Touchon M, Rocha EP. Regulation of genetic flux between bacteria by restriction-modification systems. Proc Natl Acad Sci USA. 2016;113:5658–63.
101. Mucke M, Kruger DH, Reuter M. Diversity of Type II restriction endonucleases that require two DNA recognition sites. Nucleic Acids Res. 2003;31:6079–84.
102. Ferrieres L, Hemery G, Nham T, Guerout AM, Mazel D, Beloin C, et al. Silent mischief: bacteriophage Mu insertions contaminate products of *Escherichia coli* random mutagenesis performed using suicidal transposon delivery plasmids mobilized by broad-host-range RP4 conjugative machinery. J Bacteriol. 2010;192:6418–27.
103. Guy L, Kultima JR, Andersson SG. genoPlotR: comparative gene and genome visualization in R. Bioinformatics. 2010;26: 2334–5.

Annex #2

Review article: Selfish, promiscuous, and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations.

Matthieu Haudiquet* , Jorge Moura de Sousa* , Marie Touchon* and Eduardo P. C. Rocha

Accepted for publication in Philosophical transactions of the royal society B.

I participated in the writing of the introduction and discussion, and the section “Mobile genetic elements-cell envelope interactions are key to successful transfer”.

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: Haudiquet M, de Sousa JM, Touchon M, Rocha EPC. 2022 Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations. *Phil. Trans. R. Soc. B* 20210234. https://doi.org/10.1098/rstb.2021.0234

Received: 29 October 2021
Accepted: 6 December 2021

One contribution of 11 to a discussion meeting issue 'Genomic population structures of microbial pathogens'.

Subject Areas:
evolution, genomics, microbiology

Keywords:
horizontal gene transfer, evolution, defence systems, bacteriophages, satellites, plasmids

Author for correspondence:
Eduardo P. C. Rocha
e-mail: erocha@pasteur.fr

†These authors contributed equally to this study.

THE ROYAL SOCIETY PUBLISHING

Selfish, promiscuous and sometimes useful: how mobile genetic elements drive horizontal gene transfer in microbial populations

Matthieu Haudiquet†, Jorge Moura de Sousa†, Marie Touchon† and Eduardo P. C. Rocha

Institut Pasteur, Université de Paris, CNRS UMR3525, Microbial Evolutionary Genomics, Paris 75015, France

JMdS, 0000-0003-3530-8550; EPCR, 0000-0001-7704-822X

Horizontal gene transfer (HGT) drives microbial adaptation but is often under the control of mobile genetic elements (MGEs) whose interests are not necessarily aligned with those of their hosts. In general, transfer is costly to the donor cell while potentially beneficial to the recipients. The diversity and plasticity of cell-MGEs interactions, and those among MGEs, results in complex evolutionary processes where the source, or even the existence of selection for maintaining a function in the genome is often unclear. For example, MGE-driven HGT depends on cell envelope structures and defense systems, but many of these are transferred by MGEs themselves. MGEs can spur periods of intense gene transfer by increasing their own rates of horizontal transmission upon communicating, eavesdropping, or sensing the environment and the host physiology. This may result on high-frequency transfer of host genes unrelated with the MGE. Here, we review how MGEs drive HGT and how their transfer mechanisms, selective pressures, and genomic traits affect gene flow, and therefore adaptation, in microbial populations. The encoding of many adaptive niche-defining microbial traits in MGEs means that intragenomic conflicts and alliances between cells and their MGEs are key to microbial functional diversification.

This article is part of a discussion meeting issue 'Genomic population structures of microbial pathogens'.

1. Introduction

The gene repertoires of microbial species change very fast and their pangenomes are often orders of magnitude larger than the average genome [1,2]. Most such genes are acquired by horizontal gene transfer (HGT) driven by mobile genetic elements (MGEs). Yet, MGEs are autonomous genetic agents that may proliferate even when they have a negative impact on host fitness. Gene flow is thus a rich provider of novel functions to microbial genomes but is largely out of the control of the recipient cells. On the one hand, this means that microbial adaptation depends heavily on the trade-off between gaining advantageous functions by MGE-driven HGT and the costs associated with these elements. On the other hand, as genomes contain many MGEs and these often interact antagonistically, gene flow is shaped by a complex interplay between the host and its many MGEs, as well as between the MGEs themselves. These interactions depend on the characteristics of the MGEs and on the host genetic background, notably its ability to control infections of deleterious MGEs and to integrate the novel genetic information. Ultimately, many rare genes in microbial populations may be effectively under selection because they are adaptive for the MGEs carrying them. Whether this affects cell fitness, and in which sense, it is most often unclear.

Here, we review how MGEs drive, but also constrain microbial evolution by HGT. While our text focuses on bacteria, where mechanisms are better known and examples more abundant, it is often also applicable to the interactions between Archaea and their MGEs. We start by a short summary of the mechanisms of transfer of MGEs, highlighting recent findings on their interactions.

2. Genomes as playgrounds of mobile genetic elements

MGEs drive transfer between bacteria either by transferring themselves between cells or by mediating the transfer of chromosomal DNA (figure 1). Some mechanisms of HGT do not depend on MGEs [3], most notably natural transformation [4], but their relevance across bacteria in the acquisition of novel genes remains to be understood (e.g. [5]). In this review, we focus on the role of MGEs as drivers of HGT and will not expand on these other processes. MGEs can be classified in terms of their mechanisms of autonomous horizontal (conjugation or viral particles) or vertical transmission (extrachromosomal or integrative). There is extensive genetic diversity within each type of MGE, which can complicate their identification and characterization. Furthermore, some MGEs are parasites or competitors of other MGEs, establishing complex ecological dynamics within populations.

The most frequent mechanism of conjugation involves a relaxase that nicks and attaches to a single strand of DNA. The nucleoprotein filament is then transferred between physically close cells by a type IV secretion system resulting in the replication of the element [6]. Conjugation can transfer vast amounts of DNA, up to entire chromosomes. Conjugative elements are called plasmids when extrachromosomal, and integrative conjugative elements (ICEs) when they integrate the chromosome. Despite a clear distinction made between these two types of elements in the literature, they encode similar conjugative machineries for horizontal transmission and are present across most bacterial clades [7]. Moreover, ICEs capable of autonomous replication and plasmids integrated in chromosomes have been described [8,9], suggesting only a few key differences, and also potentially frequent interconversions, between the two types of elements.

The ability of conjugative elements to transfer between cells can be exploited by mobilizable elements that are present in the same host. Interactions between mobilizable and conjugative elements have been studied more in detail in plasmids. Mobilizable plasmids are typically smaller than conjugative plasmids and do not encode the conjugative pilus required for autonomous HGT. Some encode a relaxase that interacts with pili encoded by conjugative elements present in the cell. Such mobilizable plasmids are at least as abundant as conjugative plasmids and tend to encode similar types of traits [10]. Many other plasmids lack even a relaxase and their mechanisms of transfer, as well as their interactions with other MGEs, are poorly known. Despite the exploitative interaction between these two types of MGEs, it is not known whether this systematically imparts a significant cost for the conjugative plasmid.

The contribution of temperate bacteriophages (phages) for HGT is complicated by their role as bacterial predators. Upon cell entry, temperate phages can opt between active reproduction and cell lysis (lytic cycle), or lysogeny, where they replicate synchronously with the host either integrated in the chromosome or as phage-plasmids. Half of the available bacterial genomes are recognizably lysogens [11], and some prophages encode traits adaptive to the host, like virulence factors and bacteriocins [12], but can also kill their hosts by induction of the lytic cycle [13]. The effect of temperate phages in bacterial fitness may thus depend on physiological and environmental conditions (see below). Phages can also transfer bacterial genes by generalized, specialized or lateral transduction [14,15]. Each mechanism differentially impacts the scope and efficiency of transfer of bacterial traits. For example, specialized transduction transfers only a few chromosomal genes in the neighbourhood of the prophage, whereas generalized transduction transfers genes from across the chromosome. Lateral transduction occurs when phage replication starts while the prophage is yet integrated in the chromosome, and can result in the transfer of extensive neighbouring chromosomal regions [16]. Of note, the amount of DNA packaged by phages is limited by the virion size, which in temperate phages tends to accommodate around 50 kb (with large variations across phages). As a result, a bacterial chromosome can only be transferred by transduction when fragmented across multiple virions. But since cells can liberate many phages, the extent of bacterial DNA transferred by transduction can be huge. A back-of-the-envelope calculation has estimated that a single lysate of phages that infect *Staphylococcus aureus* has the potential to encode up to 20 000 copies of an entire bacterial chromosome in transduction particles [17].

Despite being parasites of bacteria, phages have their own parasites. Phage satellites are small mobile elements (*ca* 7–18 kb) lacking components of the viral particle for autonomous transfer. Instead, they encode sophisticated mechanisms to hijack the particles of 'helper' phages to transfer between cells [18]. Three main types of phage satellites have been described: P4 in Enterobacterales [19], phage-inducible chromosomal islands in Enterobacterales and Firmicutes [20], and phage-inducible chromosomal island like-elements (PLEs) in *Vibrio* spp. [21]. Many other types of satellites may be still uncovered, and the ones that are known seem very abundant and diverse. For example, almost half of *E. coli* genomes have between one and three P4-like satellites [19]. Phage satellites can impact their bacterial hosts at different levels: by transducing chromosomal DNA [15], by encoding virulence factors [22], or by encoding anti-MGE defense systems [23]. Satellites are costly to phages because they hijack their particles, thereby decreasing phage burst size. However, there is significant variation in this cost, depending on the satellite-helper pair. Some PLEs completely abolish phage reproduction [24], whereas P4 has, under certain conditions, a much lower impact on phage reproductive fitness [25].

As satellites are mobilized by phages and mobilizable plasmids by conjugative elements, there are other MGEs that can be mobilized by these parasites of parasites [26,27]. This makes them parasites of parasites of parasites of bacteria (which may themselves be parasites of Eukaryotes). While the full scope of ecological interactions between all these MGEs is not very well known, it is clearly a multi-layered complex network that opens paths for both conflicts and alliances in the cell. As an example of such complex interactions, prophages interact not only with other prophages and satellites, e.g. by repressing or actively targeting them [28] (figure 2*a*), but also with other MGEs, particularly with conjugative or mobilizable elements, which can encode anti-phage defenses [29] or be mobilized by phages [30]. Further, and despite their potential costs for

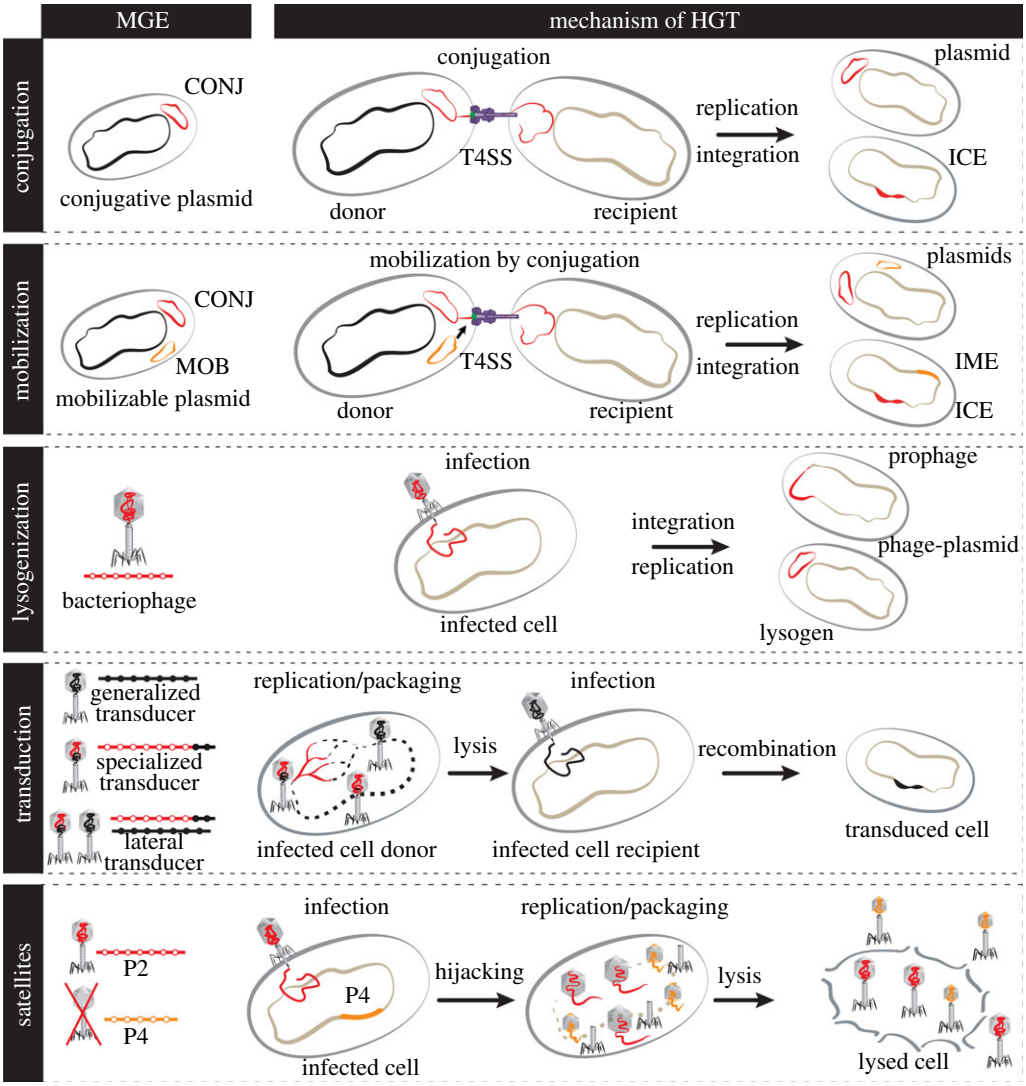


Figure 1. Major mechanisms of HGT driven by MGEs. CONJ, conjugative element; MOB, element mobilizable by conjugation. T4SS, type IV secretion system; ICE, Integrative conjugative element; IME, integrative mobilizable element.

bacterial reproduction, there are also synergies between MGEs and host cells: phage satellites encode defense systems against phages that they cannot parasitize, which favors the other MGEs in the genome, including prophages, and the host cell [23]. Finally, MGEs can exchange genetic material between them, and with their host, through transposable elements [31] or different recombination mechanisms [32]. For example, a chromosomal gene conferring resistance to carbapenem antibiotics in *Pseudomonas aeruginosa* originated from a conjugative plasmid, with the transfer from plasmid to bacterial chromosome likely being mediated by transposases [33].

The abundance and diversity of MGEs, and the myriad of their possible interactions, set up a scenario where bacteria are a playground for MGEs and their genomes are shaped by the associated eco-evolutionary wargames. The following sections of this review will thus address the different ways in which these interactions affect the networks of gene transfer that shape microbial evolution.

3. Interplay between ecology and mobile genetic elements shapes horizontal gene transfer

The transfer of an MGE requires that either cells meet for conjugation, or that viral particles diffuse far enough to find susceptible hosts. Therefore, the size and the diversity of the gene pool for a species depends on the composition of microbial communities. Metagenomics data have shown that transfers occur more frequently between isolates from similar environments [34,35]. Similar conclusions were obtained by searching for highly similar genes across different genomes [36,37]. These results have spurred proposals that the dynamic interplay among hosts, MGEs and environments shapes networks of genetic exchanges within communities [38]. Accordingly, the lineages that are most prevalent across different habitats within *Listeria* spp. have higher rates of HGT [39]. The frequency of genetic exchanges mediated by MGEs is expected to depend on the density of cell hosts in the community, which may explain why the densely populated human gut is a hotspot of genetic exchanges [34,40]. It also depends on the physical distances that can be covered by MGEs outside of the cell. These distances are extremely small for conjugative elements because they require direct cell–cell contact for transfer. Phages can survive for long periods of time in the environment [41], which allows their dispersion across large geographical distances, e.g. in aquatic environments. Hence, phage-driven HGT is more likely to result in direct transfers across segregated microbial communities than conjugation.

Structured environments, like biofilms, are thought to be the most frequent types of microbial environments in the

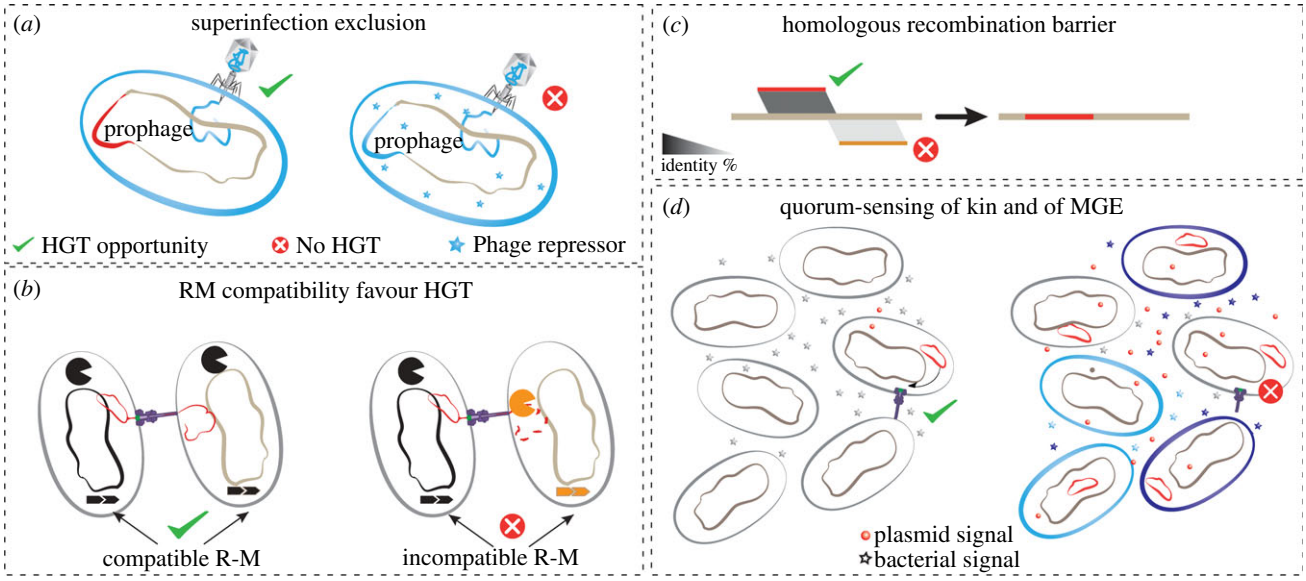


Figure 2. Recombination, defense and communication shape HGT. (*a*) Prophages protect from other phages by many mechanisms, including superinfection exclusion and repression of gene expression. (*b*) Plasmids can eavesdrop the quorum-sensing mechanisms of the host cell and use their own to promote their conjugation when there are many closely related hosts without plasmids in the neighbourhood of the host cell. (*c*) Homologous recombination (HR) requires high similarity between the exogenous DNA and the chromosome. (*d*) Bacteria with compatible restriction-modification (R-M) systems can exchange DNA at higher rates because the DNA is marked with the correct epigenetic markers and is not restricted by the recipient cell.

planet [42]. The structure of the environment is important because it shapes the physiological response of individual cells, the networks of interactions between microbes, and the transmission dynamics of their MGEs [43]. Conjugative systems mate more efficiently in solid surfaces [44,45] and conjugation can thus take place at very high rates in the outer layers of biofilms [46,47]. Plasmids that lack adaptive genes and are only maintained through high transfer frequencies are thus more likely to persist in biofilms [48]. Interestingly, conjugation itself spurs the formation of biofilms [49], thus driving conditions that effectively favour the transfer of conjugative elements. In contrast, limited diffusion of phage particles hinders phage amplification in structured environments, thereby decreasing the generation of phage genetic diversity and making phage-host antagonistic coevolution less predictable [50–52]. Habitat structure and composition are therefore key determinants of the rate and type of MGE-driven HGT.

4. Mobile genetic element manipulation of the timing of gene transfer

Several mechanisms increase the rates of genetic exchanges under conditions of maladaptation, i.e. when the acquisition of novel functions is more likely to have a positive impact on fitness. Expression of competence for natural transformation is usually under the control of conserved regulatory circuits of the recipient cell, even if several plasmids have been described to repress transformation [53,54]. In most other cases, the decision for transfer is under the control of MGEs, not of the host or recipient cells. In theory, investment in horizontal and vertical transmission are equally important for the success of the MGE at the evolutionary time scale [55]. Hence, very costly MGEs are expected to have lower rates of vertical transmission but can still prosper if their rates of HGT are high. The investment on the different types of transmission may vary. When the host viability is in risk, the investment in horizontal transmission is much more rewarding that the investment in vertical transmission. This results in an intense exodus of MGEs from the cell to increase their chances of survival, corresponding to a shift in investment from vertical to horizontal transmission in search for better hosts. The consequence for microbial populations is an increase in the rates of HGT.

MGEs can sense cues that indicate the cell is no longer a promising host for vertical transmission, and thus shift its investment from vertical towards horizontal transmission. For example, certain DNA lesions lead to the activation of the SOS response which favours the induction of prophages [56,57] and conjugative elements [58]. Because of their effect on cell physiology, including induction of SOS in some bacteria, antibiotics can spur the transfer of phages [59] and conjugative plasmids [60]. Inflammatory responses in the gut also increase conjugative transfer and prophage induction, fostering the spread of functions such as those associated with virulence and antibiotic resistance [61,62]. These processes are under the control of the MGE and can be costly, and sometimes lethal, to the donor cells. Occasionally, they result in the acquisition of adaptive genes by a recipient cell.

The timing and source of gene flow in populations may also be conditioned by social processes. Quorum-sensing allows bacteria to assess the abundance of closely related cells in a population. Similarly, MGEs have evolved to sense bacterial quorum-sensing signals to eavesdrop on bacterial communication and decide when to invest in horizontal transmission [63]. MGEs also encode their own quorum-sensing systems that further informs them on the presence of similar elements in neighbouring bacteria. Conjugative plasmids use it to transfer between cells when the environment is crowded with closely related bacteria that lack the plasmid [64,65] (figure 2*b*). Temperate phages use it to favour lysogeny when the density of similar phages in the environment is high [66] and to induce the lytic cycle when the concentration of susceptible hosts is high [67,68].

Although systems of molecular communication have only recently been uncovered in MGEs, it is possible that several other strategies of communication underlie their interactions with other MGEs and with their potential hosts [69].

5. Scope of horizontal gene transfer as the result of mobile genetic element-host interactions

Since a lot of HGT relies on the ability of MGEs to transfer horizontally between hosts, their host-range will determine the rate at which adaptive traits can be transferred across different species. In general terms, the efficiency of HGT decreases with the phylogenetic distance between donor and recipient cells [70]. The magnitude of this effect depends on the mechanism of transfer of MGEs. Conjugative elements, which do not require specific cell receptors, often have large host ranges and can transfer elements across genera or even phyla [71]. Phage host ranges are usually narrower and can be limited to a small number of strains having a specific cell receptor or serotype [72] (see below). The host range of the many MGEs that exploit other MGEs to transfer across cells is poorly known. Some mobilizable plasmids might have a very broad host range because they can hijack conjugative systems from different conjugative plasmids [27]. Similarly, the host range of phage satellites depends on their ability to hijack multiple phages.

Once an MGE has successfully passed the envelope and the cell defense barriers, it still endures functional constraints because the molecular mechanisms used by the MGE for horizontal transmission (e.g. production of viral particles or conjugative pili) may not work in the novel genetic background, thereby restricting the MGE effective host range. For example, conjugative pili are specialized in specific membrane structures and those functioning in cells with an outer membrane usually do not work in cells lacking it [73]. How functions related to vertical transmission work (or do not work) in the novel genetic background of recipient cells also contributes to explain differences in host range. Site-specific recombinases allow MGEs to integrate at highly conserved regions of the chromosome, like tRNA genes, without inactivating them [74]. These integrases function in very different genetic backgrounds, facilitating transfer of MGEs across distantly related taxa with little fitness impact for the host. The higher sensitivity of plasmid replicases to the genetic background relative to ICE integrases contributes to explain why the latter have even broader host ranges than the former [75]. The broad host range of conjugative elements and their high genetic plasticity may explain why these elements are the major vectors of the ongoing large-scale transfer of antibiotic resistance from soil bacteria to human pathogens [76].

DNA integrating into the genome by homologous recombination must have high sequence identity with the chromosome (figure 2*c*) [77]. This mechanism is important for allelic exchanges in core genes, which in many species results in rates of introduction of nucleotide changes higher than those caused by mutation rates [78]. In bacteria that are not naturally transformable, these allelic exchanges require MGE-driven HGT. Yet, core genes are systematically absent from MGEs. Conjugation or transduction are the most likely candidates to provide the chromosomal DNA required for allelic exchanges. Recent studies show that lateral transduction can drive the transfer of vast amounts of chromosomal DNA within species [79]. However, we still lack quantitative measures of the relative importance of these different processes in shaping patterns of recombination in natural populations. While recombination might allow the integration of exogenous DNA it may also favour the deletion of MGEs from the chromosome [80]. Unfortunately, most of these recombination processes leave very few, if any, traces of the vehicle of transfer of the exogenous DNA into the cell, which is also why the real-world impact of some types of HGT are still so difficult to quantify (e.g. generalized transduction). As a result, the mechanisms of acquisition of exogenous DNA allowing allelic exchanges in core genes by homologous recombination remain largely hypothetical and based on extrapolation from data of laboratory experiments.

6. Mobile genetic elements-cell envelope interactions are key to successful transfer

MGE-driven HGT requires an initial interaction between the recipient cell envelope and the structural component of the MGE that interfaces with it, be it the tip of the conjugative pilus or the tail of the phage. Viral particles interact with cells via phage-encoded receptor-binding proteins (RBPs), which enable their adsorption and stabilization at the cell surface before DNA is injected into the cell [81]. RBPs are very specific to their corresponding bacterial receptors and shape the host-range of the phage and the sensitivity of the bacterium. By contrast, conjugation only requires close contact between cells and does not seem to rely on a specific receptor at the cell envelope [82]. These mechanistic differences contribute to explain why phages tend to have narrower host ranges than conjugative elements.

Structures located at the cell envelope, like the bacterial capsule, provide additional control over the access of MGEs to the cell. Capsules are composed of membrane-bound polysaccharide chains and constitute the first point of contact of MGEs with the cell [83]. They can be very large, creating exclusion zones thicker than the cell diameter, and protect bacteria from agents like macrophages or antibiotics [84]. They can also protect from phages, because capsules can hide phage receptors [85]. Capsules were thus thought to decrease gene flow [86]. However, phages that infect bacteria that constitutively express their capsule, like *Klebsiella pneumoniae* and *Acinetobacter baumannii*, have evolved to use the capsule to adsorb to the cell [87]. The RBPs of these types of phages are endowed with capsule depolymerases, specific to one or a few capsular serotypes, granting them access to the outer membrane after adsorption at the capsule (figure 3*a*). But this adaptation comes at a cost: such phages may become dependent on a specific capsule to adsorb efficiently to the cell envelope, and are unable to infect non-capsulated cells, or even cells with a different capsular serotype. This is not a rare occurrence since the temperate phage infection networks of *K. pneumoniae* show clear serotype-specific clusters [88], resulting in more frequent phage-driven gene flow between strains with similar serotypes [89] (figure 3*b*). The requirement for a capsule for phage adsorption implies that phage pressure may lead to selection for capsule inactivation, because non-capsulated bacteria are resistant to these phages [88]. Interestingly, such non-capsulated cells are not sexually isolated because

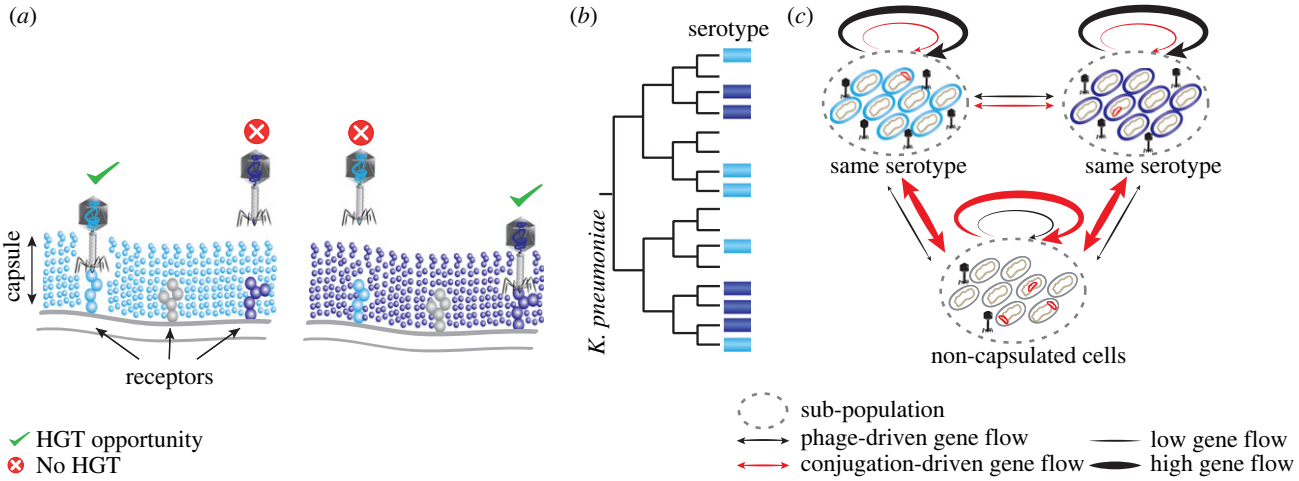


Figure 3. (*a*) The capsule is a barrier to phage infection when it hides phage receptors. But some phages have evolved to degrade the capsule and can thus use it for adsorption. (*b*) The capsule is frequently lost and gained by HGT during *K. pneumoniae* evolution resulting in frequent serotype switching. (*c*) Because of the capsular specificity of temperate phages in *K. pneumoniae*, phage-driven HGT is much more frequent within than between serotypes. By contrast, non-capsulated cells are more permissive to conjugation. Hence, gene flow depends on the presence of the capsule, its serotype, and the type of MGE driving HGT.

even if phage-driven transfer may be diminished, they are much more receptive to conjugative elements [89]. Hence, variations in the capsule composition or expression change both phage and conjugation-driven gene flow. The consequences of these changes are very different and somewhat complementary (figure 3*c*): phage-driven transfer is particularly high between strains of the same serotype and conjugation is more frequent towards non-capsulated strains.

Many other components of the cell envelope are involved in complex interplays with MGEs and affect their rates of transfer. The O-antigen of LPS is often targeted by phages, and it displays high genetic and chemical variability within and across species [90]. Switching from smooth to rough LPS type is usually associated with phage resistance and altered LPS structures. Since LPS-related rough phenotypes are also associated with modified virulence in pathogens [91], phage predation also impacts the evolution of virulence in these strains. The dependence of MGE transfer on the physiological traits of cells means that changes in envelope composition can reshape networks of gene flow and this will eventually also affect the HGT of components of the envelope. In conclusion, bacterial physiology, and the different selective pressures impacting it, are a strong determinant of both the frequency and type of MGE-driven HGT.

7. Cell and mobile genetic element defenses and counter-defenses constrain gene flow

Once the DNA enters the recipient cell cytoplasm, defense systems can still block its expression. Microbes and their MGEs have evolved numerous specialized defense and counter-defense systems that are frequently gained and lost. Their genetic diversification is caused by the antagonistic coevolution between microbial cells and MGEs. These defense systems are currently being uncovered at a fast pace and have recently been reviewed [92–94]. Interestingly, recent data suggest that most such 'cellular' defense systems are actually encoded in MGEs and not in conserved sections of the host chromosome [95]. The available evidence is thus that MGE-encoded defense systems are protecting their host cell as a side-effect of their action to protect the MGE from other MGEs [96]. Antagonistic coevolution between MGEs could thus be at least as important as that between MGEs and the host.

One might think that there is a trade-off between maintaining many defense systems and allowing the genome to acquire adaptive genes by HGT. Since defense systems block some MGEs from certain lineages, they carve preferential pathways of gene flow in microbial populations. Notably, there is more HGT and homologous (allelic) recombination between pairs of strains with compatible restriction modification (R-M) systems, by far the most abundant specialized defense systems, than between other strains. This is because MGEs transferred between strains with compatible R-M systems carry the same methylation patterns and thus are able escape restriction that would otherwise exclude their DNA from establishing in the cell [97] (figure 2*d*). While defense systems tend to limit the income of new DNA, in certain circumstances they may even facilitate HGT [98]. Many defense systems, like viperins or retrons [99,100], target very specific functions and may not impact the transfer of most MGEs. Hence, defense systems shape but do not abolish gene flow in microbial populations. MGEs, being both targets and producers of defense systems, are both vectors of and barriers to HGT.

8. Mobile genetic elements turnover

MGEs represent a large fraction of the accessory genome of many species, but they are rarely maintained in a lineage for a long period of time [95,101]. These rapid dynamics of gene gain and loss contribute to the U-shaped distribution of the frequency of gene families in pangenomes, typically resulting in a large majority of gene families being either very frequent (persistent genome) or quite rare (usually acquired in MGEs) [102]. The high turnover of MGEs means that closely related strains can have very different MGE contents. This is the case in *E. coli* and *K. pneumoniae*, where epidemiologically indistinguishable strains (from the same sequence types) differ in the many different MGEs they carry [103,104]. A high MGE turnover also means that

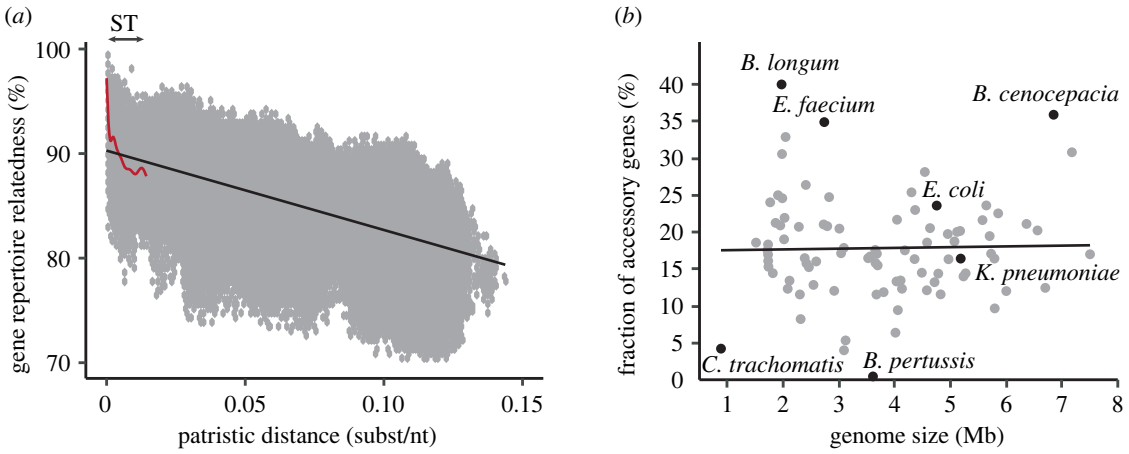


Figure 4. Impact of the high turn-over of MGEs on gene repertoire (left) and genome size of the host. (*a*) Gene repertoire relatedness decreases quickly with the patristic distance in *E. coli* (red spline fit line) at short evolutionary distances, i.e. between genomes of the same sequence types (ST). The subsequent changes are more moderated and approximately linear with time (black linear fit line) [103]. Of note, the variance around these average trends is very large. This figure was simplified and redrawn from the data in [103]. (*b*) Linear regression of the fraction of accessory genes per genome in function of the average species genome size for the 90 most represented species in GenBank. Figure redrawn and simplified from the results presented in [105].

while MGEs are a sizeable part of bacterial genomes (*ca* 10% in *E. coli* for phages plus plasmids) they account for most of its variation in size [103]. This rapid flux of MGEs explains why relatedness between gene repertoires decreases very quickly with phylogenetic distance for closely related genomes (figure 4*a*).

Many forces drive the rapid turnover of MGEs and their genes in bacterial genomes [37,106]. Foremost, MGEs can be very costly and their hosts counter-selected [107]. Induction of temperate phages kills the host, and even plasmids and transposons may involve lower, but not necessarily negligible, costs [108,109]. The rapid loss of MGEs could thus be interpreted as the result of their negative contribution to the host fitness. In this view, the ubiquitous presence of MGEs in microbial populations could be explained by their selfish spread.

However, extensive data suggest a more nuanced view of the costs and benefits of HGT driven by MGEs [110]. The costs of MGEs can decrease rapidly after their acquisition by a host, as frequently observed in plasmids. The acquisition of novel plasmids is usually associated with an elevated physiological burden, but purifying selection does not necessarily lead to plasmid loss or chromosomal integration of beneficial genes [106], especially when the element carries adaptive traits under positive selection [111]. In such cases, there is rapid emergence of compensatory mutations, either in the chromosome or in the plasmid themselves, that alleviate the cost of the element [112], e.g. by resolving specific genetic conflicts [113]. Amelioration contributes to lower the cost of MGEs as parasites and increases their stability in microbial lineages.

Many MGEs carry genes that are adaptive under specific and potentially transient conditions [2]. The linkage between these adaptive genes and the MGE may provide the ensemble with positive net fitness advantage to the host for some time. The MGE would be selectively maintained as long as these genes provide a sufficient fitness advantage, but could be quickly lost when its positive impact on fitness ceases. Many accessory genes in MGEs may be adaptive for only short periods. For example, antibiotic resistance genes tend to be costly and are typically lost when individuals are no longer subject to antibiotics [114]. Genes under negative frequency dependent selection, e.g. toxins encoded by MGEs associated with inter or intra-specific competition [115], are also expected to be rapidly replaced. The presence of genes adaptive only in particular contexts means that the associated MGEs may endure fluctuating types of selection, i.e. they are adaptive in certain contexts and parasites in others.

Finally, neutral processes may accelerate the loss of MGEs. Adaptive genes may escape costly MGEs by translocating into the chromosome [116], thereby turning an adaptive MGE into a costly one that becomes counter-selected even if the host fitness has not changed. MGEs may also be affected by the pervasive bias toward deletions in bacteria [117] that may be more pronounced in MGEs because they have many transposable elements [118] and repeated DNA [119]. Therefore, the high turnover of MGEs is probably the result of multiple selective pressures and mutational biases that operate at different scales: the gene, the MGE and the host genome.

9. Impact of mobile genetic element turnover on pangenome evolution

The rapid turnover of MGEs implies that high rates of HGT do not necessarily result in larger microbial genomes. Except for very small genomes, that sometimes show little or no evidence of MGEs and HGT, there is extensive variation in the frequency of accessory genes per microbial genome. This frequency varies from a few percent to close to 40% [105], with many species showing values between 10 and 25% (figure 4*b*). Species with large genomes tend to have higher effective population sizes [120], but they do not necessarily have very high rates of HGT [121], nor of homologous recombination [120]. The fraction of the genome that corresponds to the accessory genome is also not correlated with the average species genome size [105]. Hence, the fraction of accessory genes, most of which are acquired by HGT, does not seem to result from the same selection processes that result in larger genomes. Instead, it may reflect the rates and costs of gene gain and loss. Since most HGT seems to be driven by MGEs, the persistence of novel genes in bacterial lineages will be dependent on deletion biases, on the fitness effect of the gene and on its direct genetic environment (the MGE). If the MGEs have high horizontal transmission rates, they are also more likely to be costly. Hence, genomes with high rates of HGT might only have an average amount of

accessory genes because most acquired genes are in costly MGEs that are rapidly lost from the genome (or the genome is purged from populations by purifying selection).

Extreme reductions in genome size have been observed in endo-mutualists that are sexually isolated, endure population bottlenecks, and live in constant environments [122]. But similar processes of genome reduction have been found in free-living bacteria that are able to exchange DNA, presumably due to selection for genome streamlining [123]. Surprisingly, bacterial genomes can shrink despite being under the influence of high rates of HGT. The phylogroups of *E. coli* with the smallest genomes have the highest rates of gene repertoire diversification and fewer but more diverse MGEs [103]. Many of these small *E. coli* genomes are from freshwater isolates, lack antibiotic resistance genes and virulence factors, and have a large pangenome. They seem to be locally adapted to their nutrient poor environment. This example illustrates how ecological opportunities can shape the number, the type, and the distribution of MGEs in a population. In this case, while high gene flow may have facilitated parallel adaptation to an environment that is very different from the mammalian gut, selection for streamlining in such nutrient-poor environments [123] has likely resulted in genome reduction.

10. Outlook and unsolved mysteries

The identification of the pertinent levels of selection—genes, MGEs or/and genomes—can be extremely complicated when populations have many MGEs that are prone to genetic conflicts. Because a lot of HGT is driven by MGEs, many of the most recent genes in the genome may be neutral or deleterious to the host cell, while being selected due to the benefits they confer to the MGE itself. Still, genes in MGEs can sometimes be adaptive to the host as a by-product of their selection by the MGE, typically because higher host fitness increases the fitness of the MGE encoding the trait. This is the case for many traits in plasmids and phages, like antibiotic resistance, toxins and defense systems which are adaptive both to the MGE and to its host. Many such genes may be adaptive under certain situations and not in others. For example, phage satellites can block phage infections and thus favour the bacterial host, but may be costly when the specific helper phages are absent. Likewise, prophages without genes that are adaptive to the host might still provide resistance to other similar phages. While the qualitative understanding of these processes has much progressed, there is a paucity of quantitative data to understand how much of the HGT is potentially of adaptive value for the recipient cell.

MGEs can be costly and reproduce selfishly across population but may also occasionally provide adaptive genomic changes by increasing genome evolvability [124]. Many studies revealed the roles of transposable elements in shuttling adaptive genes between replicons thereby favouring their transfer in plasmids or their stabilization in the chromosome [118]. But transposition of these elements also results in frequent pseudogenization of useful genes. How frequently the gains in evolvability provided by MGEs compensate the costs of these elements is poorly known. These indirect selective effects (i.e. higher order selection), are hard to measure in the laboratory because they depend on the genetic diversity of communities and the frequencies and types of ecological challenges faced by Bacteria and Archaea. Further work will be needed to disentangle how and when such elements contribute, or not, to host adaptation. Such studies should account for the fact that recipient cells have little control over the rates of HGT and that MGEs have their own evolutionary interests, meaning that it is difficult to interpret changes in the rates of HGT in the light of selection for microbial evolvability.

The availability of low-cost sequencing and the current focus on the worrisome spread of antibiotic resistance genes by MGEs may provide crucial data to quantify how rates of HGT depend on the type of MGE and its mechanisms of horizontal transmission. For example, phages encode many toxins, but few antibiotic resistance genes [125]. The latter are much more frequent in conjugative elements, especially in plasmids [75]. The genetic plasticity, range of interactions and mode of transfer of MGEs might explain why certain MGEs are preferentially associated with certain traits.

Finally, it is important to stress that many MGEs might still be unknown and many of the known ones have yet unknown mechanisms of transfer. For example, over 50% of known plasmids do not encode a conjugative apparatus nor a known relaxase [10]. They may be transferred by one of many processes: conjugation using a relaxase from another plasmid [126], generalized transduction [30,127], natural transformation [128] or vesicles [129]. The current lack of information on the mechanisms of transfer of many MGEs raises questions about their origins, mechanisms of dissemination and impact on microbial evolution. Rough estimates suggest that most large contiguous stretches of non-homologous sequences integrated in genomes by integrases, presumably MGEs, remain to be characterized [130]. The identification of these elements and their interactions with hosts and other MGEs will certainly contribute to a better understanding of gene flow in microbial populations.

Data accessibility. This article has no additional data.

Authors' contributions. M.H.: writing—original draft, writing—review and editing; J.M.d.S.: writing—original draft, writing—review and editing; M.T.: writing—original draft, writing—review and editing; E.P.C.R.: conceptualization, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. We acknowledge the financial support of Equipe FRM (EQU201903007835), Laboratoire d'Excellence IBEID (ANR-10-LABX-62-IBEID), the INCEPTION program (PIA/ANR-16-CONV-0005), the ANR (SALMOPROPHAGE ANR-16-CE16-0029; ENCAPSULATION ANR-18-CE12-0001-01)..

Acknowledgements. We thank the GEM laboratory for discussions on these topics along the years.

References

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

2. Brockhurst MA, Harrison E, Hall JP, Richards T, Mcnally A, Maclean C. 2019 The ecology and

evolution of pangenomes. *Curr. Biol.* **29**, R1094–RR103. (doi:10.1016/j.cub.2019.08.012)

3. García-Aljaro C, Balleste E, Muniesa M. 2017 Beyond the canonical strategies of horizontal gene transfer in prokaryotes. *Curr. Opin. Microbiol.* **38**, 95–105. (doi:10.1016/j.mib.2017.04.011)
4. Johnston C, Martin B, Fichant G, Polard P, Claverys JP. 2014 Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.* **12**, 181–196. (doi:10.1038/nrmicro3199)
5. Wang Y, Lu J, Engelstädter J, Zhang S, Ding P, Mao L, Yuan Z, Bond PL, Guo J. 2020 Non-antibiotic pharmaceuticals enhance the transmission of exogenous antibiotic resistance genes through bacterial transformation. *ISME J.* **14**, 2179–2196. (doi:10.1038/s41396-020-0679-2)
6. De La Cruz F, Frost LS, Meyer RJ, Zechner E. 2010 Conjugative DNA metabolism in Gram-negative bacteria. *FEMS Microbiol. Rev.* **34**, 18–40. (doi:10.1111/j.1574-6976.2009.00195.x)
7. Guglielmini J, Quintais L, Pilar Garcillan-Barcia M, De La Cruz F, Rocha EPC. 2011 The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* **7**, e1002222. (doi:10.1371/journal.pgen.1002222)
8. Cavalli-Sforza L, Lederberg J. 1956 Isolation of pre-adaptive mutants in bacteria by sib selection. *Genetics* **41**, 367. (doi:10.1093/genetics/41.3.367)
9. Lee CA, Babic A, Grossman AD. 2010 Autonomous plasmid-like replication of a conjugative transposon. *Mol. Microbiol.* **75**, 268–279. (doi:10.1111/j.1365-2958.2009.06985.x)
10. Smillie C, Pilar Garcillan-Barcia M, Victoria Francia M, Rocha EPC, De La Cruz F. 2010 Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452. (doi:10.1128/MMBR.00020-10)
11. Touchon M, Bernheim A, Rocha EP. 2016 Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J.* **10**, 2744–2754. (doi:10.1038/ismej.2016.47)
12. Taylor VL, Fitzpatrick AD, Islam Z, Maxwell KL. 2019 The diverse impacts of phage morons on bacterial fitness and virulence. *Adv. Virus Res.* **103**, 1–31. (doi:10.1016/bs.aivir.2018.08.001)
13. Paul JH. 2008 Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J.* **2**, 579–589. (doi:10.1038/ismej.2008.35)
14. Touchon M, De Sousa JAM, Rocha EP. 2017 Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr. Opin. Microbiol.* **38**, 66–73. (doi:10.1016/j.mib.2017.04.010)
15. Chiang YN, Penadés JR, Chen J. 2019 Genetic transduction by phages and chromosomal islands: the new and noncanonical. *PLoS Pathog.* **15**, e1007878. (doi:10.1371/journal.ppat.1007878)
16. Chen J, Quiles-Puchalt N, Chiang YN, Bacigalupe R, Fillol-Salom A, Chee MSJ, Fitzgerald JR, Penadés JR. 2018 Genome hypermobility by lateral transduction. *Science* **362**, 207–212. (doi:10.1126/science.aat5867)
17. Fillol-Salom A, Alsaadi A, Sousa JAM, Zhong L, Foster KR, Rocha EPC, Penadés JR, Ingmer H, Haaber J. 2019 Bacteriophages benefit from generalized transduction. *PLoS Pathog.* **15**, e1007888. (doi:10.1371/journal.ppat.1007888)
18. Ibarra-Chávez R, Hansen MF, Pinilla-Redondo R, Seed KD, Trivedi U. 2021 Phage satellites and their emerging applications in biotechnology. *FEMS Microbiol. Rev.* **45**, fuab031. (doi:10.1093/femsre/fuab031)
19. De Sousa JM, Rocha EP. 2021 To catch a hijacker: abundance, evolution and genetic diversity of P4-like bacteriophage satellites. *bioRxiv*. (doi:10.1101/2021.03.30.437493)
20. Brady A, Felipe-Ruiz A, Gallego Del Sol F, Marina A, Quiles-Puchalt N, Penadés JR. 2021 Molecular Basis of Lysis–lysogeny decisions in Gram-positive phages. *Annu. Rev. Microbiol.* **75**, 563–581. (doi:10.1146/annurev-micro-033121-020757)
21. Barth ZK, Netter Z, Angermeyer A, Bhardwaj P, Seed KD. 2020 A family of viral satellites manipulates invading virus gene expression and can affect cholera toxin mobilization. *Msystems* **5**. (doi:10.1128/mSystems.00358-20)
22. Novick RP, Ram G. 2017 Staphylococcal pathogenicity islands—movers and shakers in the genomic firmament. *Curr. Opin. Microbiol.* **38**, 197–204. (doi:10.1016/j.mib.2017.08.001)
23. Rousset F, Dowding J, Bernheim A, Rocha E, Bikard D. 2021 Prophage-encoded hotspots of bacterial immune systems. *bioRxiv* 2021.01.21.427644. (doi:10.1101/2021.01.21.427644)
24. Hays SG, Seed KD. 2020 Dominant vibrio cholerae phage exhibits lysis inhibition sensitive to disruption by a defensive phage satellite. *eLife* **9**, e53200. (doi:10.7554/eLife.53200)
25. Lindqvist BH, Deho G, Calendar R. 1993 Mechanisms of genome propagation and helper exploitation by satellite phage P4. *Microbiol. Rev.* **57**, 683–702. (doi:10.1128/mr.57.3.683-702.1993)
26. Haag AF *et al.* 2021 A regulatory cascade controls *Staphylococcus aureus* pathogenicity island activation. *Nat. Microbiol.* 1–9.
27. Garcillan-Barcia MP, Cuartas-Lanza R, Cuevas A, De La Cruz F. 2019 Cis-acting relaxases guarantee independent mobilization of MOBQ4 plasmids. *Front. Microbiol.* **10**, 2557. (doi:10.3389/fmicb.2019.02557)
28. Bondy-Denomy J, Qian J, Westra ER, Buckling A, Guttman DS, Davidson AR, Maxwell KL. 2016 Prophages mediate defense against phage infection through diverse mechanisms. *ISME J.* **10**, 2854–2866. (doi:10.1038/ismej.2016.79)
29. León LM, Park AE, Borges AL, Zhang JY, Bondy-Denomy J. 2021 Mobile element warfare via CRISPR and anti-CRISPR in *Pseudomonas aeruginosa*. *Nucleic Acids Res.* **49**, 2114–2125. (doi:10.1093/nar/gkab006)
30. RodrãGuez-Rubio L *et al.* 2020 Extensive antimicrobial resistance mobilization via multicopy plasmid encapsidation mediated by temperate phages. *J. Antimicrob. Chemother.* **75**, 3173–3180. (doi:10.1093/jac/dkaa311)
31. Loftie-Eaton W *et al.* 2016 Evolutionary paths that expand plasmid host-range: implications for spread of antibiotic resistance. *Mol. Biol. Evol.* **33**, 885–897. (doi:10.1093/molbev/msv339)
32. Moura De Sousa JA, Pfeifer E, Touchon M, Rocha EPC. 2021 Causes and consequences of bacteriophage diversification via genetic exchanges across lifestyles and bacterial taxa. *Mol. Biol. Evol.* **38**, 2497–2512. (doi:10.1093/molbev/msab044)
33. Van Der Zee A, Kraak WB, Burggraaf A, Goessens WHF, Pirovano W, Ossewaarde JM, Tommassen J. 2018 Spread of carbapenem resistance by transposition and conjugation among *Pseudomonas aeruginosa*. *Front. Microbiol.* **9**, 2057. (doi:10.3389/fmicb.2018.02057)
34. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011 Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244. (doi:10.1038/nature10571)
35. Brito IL *et al.* 2016 Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435–439. (doi:10.1038/nature18927)
36. Hooper SD, Mavromatis K, Kyrpides NC. 2009 Microbial co-habitation and lateral gene transfer: what transposases can tell us. *Genome Biol.* **10**, R45. (doi:10.1186/gb-2009-10-4-r45)
37. Popa O, Dagan T. 2011 Trends and barriers to lateral gene transfer in prokaryotes. *Curr. Opin. Microbiol.* **14**, 615–623. (doi:10.1016/j.mib.2011.07.027)
38. Skippington E, Ragan MA. 2011 Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol. Rev.* **35**, 707–735. (doi:10.1111/j.1574-6976.2010.00261.x)
39. Liao J, Guo X, Weller DL, Pollak S, Buckley DH, Wiedmann M, Cordero OX. 2021 Nationwide genomic atlas of soil-dwelling Listeria reveals effects of selection and population ecology on pangenome evolution. *Nat. Microbiol.* **6**, 1021–1030. (doi:10.1038/s41564-021-00935-7)
40. Frazão N, Sousa A, Lässig M, Gordo I. 2019 Horizontal gene transfer overrides mutation in *Escherichia coli* colonizing the mammalian gut. *Proc. Natl Acad. Sci. USA* **116**, 17 906–17 915. (doi:10.1073/pnas.1906958116)
41. Muniesa M, Lucena F, Jofre J. 1999 Comparative survival of free shiga toxin 2-encoding phages and *Escherichia coli* strains outside the gut. *Appl. Environ. Microbiol.* **65**, 5615–5618. (doi:10.1128/AEM.65.12.5615-5618.1999)
42. Flemming H-C, Wuertz S. 2019 Bacteria and archaea on Earth and their abundance in biofilms. *Nat. Rev. Microbiol.* **17**, 247–260. (doi:10.1038/s41579-019-0158-9)
43. Madsen JS, Burmølle M, Hansen LH, Sørensen SJ. 2012 The interconnection between biofilm formation and horizontal gene transfer. *FEMS Immunol. Med. Microbiol.* **65**, 183–195. (doi:10.1111/j.1574-695X.2012.00960.x)
44. Bradley DE. 1984 Characteristics and function of thick and thin conjugative pili determined by transfer-derepressed plasmids of incompatibility groups I1, I2, I5, B, K and Z. *J. Gen. Microbiol.* **130**, 1489–1502. (doi:10.1099/00221287-130-6-1489)
45. Sheppard RJ, Beddis AE, Barraclough TG. 2020 The role of hosts, plasmids and environment in determining plasmid transfer rates: a meta-analysis.

Plasmid **108**, 102489. (doi:10.1016/j.plasmid.2020.102489)

46. Hausner M, Wuertz S. 1999 High rates of conjugation in bacterial biofilms as determined by quantitative *in situ* analysis. *Appl. Environ. Microbiol.* **65**, 3710–3713. (doi:10.1128/AEM.65.8.3710-3713.1999)
47. Stalder T, Top E. 2016 Plasmid transfer in biofilms: a perspective on limitations and opportunities. *NPJ Biofilms Microbiomes* **2**, 16022. (doi:10.1038/npjbiofilms.2016.22)
48. Lili LN, Britton NF, Feil EJ. 2007 The persistence of parasitic plasmids. *Genetics* **177**, 399–405. (doi:10.1534/genetics.107.077420)
49. Ghigo JM. 2001 Natural conjugative plasmids induce bacterial biofilm development. *Nature* **412**, 442–445. (doi:10.1038/35086581)
50. Simmons M, Drescher K, Nadell CD, Bucci V. 2017 Phage mobility is a core determinant of phage–bacteria coexistence in biofilms. *ISME J.* **12**, 531–543. (doi:10.1038/ismej.2017.190)
51. De Sousa JAM, Rocha EP. 2019 Environmental structure drives resistance to phages and antibiotics during phage therapy and to invading lysogens during colonisation. *Sci. Rep.* **9**, 3149. (doi:10.1038/s41598-019-39773-3)
52. Lourenço M *et al.* 2020 The spatial heterogeneity of the gut limits predation and fosters coexistence of bacteria and bacteriophages. *Cell Host Microbe* **28**, 390–401.e5. (doi:10.1016/j.chom.2020.06.002)
53. Konkol MA, Blair KM, Kearns DB. 2013 Plasmid-encoded ComI inhibits competence in the ancestral 3610 strain of *Bacillus subtilis*. *J. Bacteriol.* **195**, 4085–4093. (doi:10.1128/JB.00696-13)
54. Durieux I, Ginevra C, Attaiech L, Picq K, Juan P-A, Jarraud S, Charpentier X. 2019 Diverse conjugative elements silence natural transformation in Legionella species. *Proc. Natl Acad. Sci. USA* **116**, 18 613–18 618. (doi:10.1073/pnas.1909374116)
55. Tazzyman SJ, Bonhoeffer S. 2013 Fixation probability of mobile genetic elements such as plasmids. *Theor. Popul. Biol.* **90**, 49–55. (doi:10.1016/j.tpb.2013.09.012)
56. Oppenheim AB, Kobiler O, Stavans J, Court DL, Adhya S. 2005 Switches in bacteriophage lambda development. *Annu. Rev. Genet.* **39**, 409–429. (doi:10.1146/annurev.genet.39.073003.113656)
57. Nanda AM, Heyer A, Krämer C, Grünberger A, Kohlheyer D, Frunzke J. 2014 Analysis of SOS-induced spontaneous prophage induction in *Corynebacterium glutamicum* at the single-cell level. *J. Bacteriol.* **196**, 180–188. (doi:10.1128/JB.01018-13)
58. Beaber JW, Hochhut B, Waldor MK. 2004 SOS response promotes horizontal dissemination of antibiotic resistance genes. *Nature* **427**, 72–74. (doi:10.1038/nature02241)
59. Allen HK, Looft T, Bayles DO, Humphrey S, Levine UY, Alt D, Stanton TB. 2011 Antibiotics in feed induce prophages in swine fecal microbiomes. *MBio* **2**, 00 260–00 211. (doi:10.1128/mBio.00260-11)
60. Jutkina J, Marathe N, Flach C-F, Larsson D. 2018 Antibiotics and common antibacterial biocides stimulate horizontal transfer of resistance at low concentrations. *Sci. Total Environ.* **616**, 172–178. (doi:10.1016/j.scitotenv.2017.10.312)
61. Stecher B *et al.* 2012 Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. *Proc. Natl Acad. Sci. USA* **109**, 1269–1274. (doi:10.1073/pnas.1113246109)
62. Diard M *et al.* 2017 Inflammation boosts bacteriophage transfer between *Salmonella* spp. *Science* **355**, 1211–1215. (doi:10.1126/science.aaf8451)
63. Van Gestel J *et al.* 2021 Short-range quorum sensing controls horizontal gene transfer at micron scale in bacterial communities. *Nat. Commun.* **12**, 2324. (doi:10.1038/s41467-021-22649-4)
64. Auchtung JM, Lee CA, Monson RE, Lehman AP, Grossman AD. 2005 Regulation of a Bacillus subtilis mobile genetic element by intercellular signaling and the global DNA damage response. *Proc. Natl Acad. Sci. USA* **102**, 12 554–12 559. (doi:10.1073/pnas.0505835102)
65. Kohler V, Keller W, Grohmann E. 2019 Regulation of Gram-positive conjugation. *Front. Microbiol.* **10**, 1134. (doi:10.3389/fmicb.2019.01134)
66. Erez Z *et al.* 2017 Communication between viruses guides lysis-lysogeny decisions. *Nature* **541**, 488–493. (doi:10.1038/nature21049)
67. Brady A *et al.* 2021 The arbitrium system controls prophage induction. *Curr. Biol.* **31**, 5037–5045.
68. Bruce JB, Lion S, Buckling A, Westra ER, Gandon S. 2021 Regulation of prophage induction and lysogenization by phage communication systems. *Curr. Biol.* **31**, 5046–5051. (doi:10.1016/j.cub.2021.08.073)
69. Bernard C, Li Y, Lopez P, Bapteste E. 2021 Beyond arbitrium: identification of a second communication system in *Bacillus* phage phi3T that may regulate host defense mechanisms. *ISME J.* **15**, 545–549. (doi:10.1038/s41396-020-00795-9)
70. 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 Res.* **21**, 599–609. (doi:10.1101/gr.115592.110)
71. Trieu-Cuot P, Carlier C, Martin P, Courvalin P. 1987 Plasmid transfer by conjugation from *Escherichia coli* to Gram-positive bacteria. *FEMS Microbiol. Lett.* **48**, 289–294. (doi:10.1111/j.1574-6968.1987.tb02558.x)
72. Hyman P, Abedon ST. 2010 Bacteriophage host range and bacterial resistance. *Adv. Appl. Microbiol.* **70**, 217–248. (doi:10.1016/S0065-2164(10)70007-1)
73. Guglielmini J, De La Cruz F, Rocha EPC. 2013 Evolution of conjugation and Type IV secretion systems. *Mol. Biol. Evol.* **30**, 315–331. (doi:10.1093/molbev/mss221)
74. Williams KP. 2002 Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.* **30**, 866–875. (doi:10.1093/nar/30.4.866)
75. Cury J, Oliveira PH, De La Cruz F, Rocha EPC. 2018 Host range and genetic plasticity explain the coexistence of integrative and extrachromosomal mobile genetic elements. *Mol. Biol. Evol.* **35**, 2230–2239. (doi:10.1093/molbev/msy123)
76. Jiang X, Ellabaan MMH, Charusanti P, Munck C, Blin K, Tong Y, Weber T, Sommer MOA, Lee SY. 2017 Dissemination of antibiotic resistance genes from antibiotic producers to pathogens. *Nat. Commun.* **8**, 15784. (doi:10.1038/ncomms15784)
77. Vulic M, Dionisio F, Taddei F, Radman M. 1997 Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl Acad. Sci. USA* **94**, 9763–9767. (doi:10.1073/pnas.94.18.9763)
78. Vos M, Didelot X. 2009 A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* **3**, 199–208. (doi:10.1038/ismej.2008.93)
79. Humphrey S, Fillol-Salom A, Quiles-Puchalt N, Ibarra-Chávez R, Haag AF, Chen J, Penadés JR. 2021 Bacterial chromosomal mobility via lateral transduction exceeds that of classical mobile genetic elements. *Nat. Commun.* **12**, 6509. (doi:10.1038/s41467-021-26004-5)
80. Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD, Fraser C. 2016 Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. *PLoS Biol.* **14**, e1002394. (doi:10.1371/journal.pbio.1002394)
81. Bertozzi Silva J, Storms Z, Sauvageau D. 2016 Host receptors for bacteriophage adsorption. *FEMS Microbiol. Lett.* **363**. (doi:10.1093/femsle/fnw002)
82. Perez-Mendoza D, De La Cruz F. 2009 Escherichia coli genes affecting recipient ability in plasmid conjugation: are there any? *BMC Genomics* **10**, 71. (doi:10.1186/1471-2164-10-71)
83. Whitfield C, Wear SS, Sande C. 2020 Assembly of bacterial capsular polysaccharides and exopolysaccharides. *Annu. Rev. Microbiol.* **74**, 521–543. (doi:10.1146/annurev-micro-011420-075607)
84. Cress BF, Englaender JA, He W, Kasper D, Linhardt RJ, Koffas MA. 2014 Masquerading microbial pathogens: capsular polysaccharides mimic host-tissue molecules. *FEMS Microbiol. Rev.* **38**, 660–697. (doi:10.1111/1574-6976.12056)
85. Labrie SJ, Samson JE, Moineau S. 2010 Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **8**, 317–327. (doi:10.1038/nrmicro2315)
86. Chewapreecha C *et al.* 2014 Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46**, 305–309. (doi:10.1038/ng.2895)
87. Knecht LE, Veljkovic M, Fieseler L. 2019 Diversity and function of phage encoded depolymerases. *Front. Microbiol.* **10**, 2949. (doi:10.3389/fmicb.2019.02949)
88. De Sousa JAM, Buffet A, Haudiquet M, Rocha EPC, Rendueles O. 2020 Modular prophage interactions driven by capsule serotype select for capsule loss under phage predation. *ISME J.* **14**, 2980–2996. (doi:10.1038/s41396-020-0726-z)
89. Haudiquet M, Buffet A, Rendueles O, Rocha EPC. 2021 Interplay between the cell envelope and

mobile genetic elements shapes gene flow in populations of the nosocomial pathogen Klebsiella pneumoniae. *PLoS Biol.* **19**, e3001276. (doi:10.1371/journal.pbio.3001276)

90. Whitfield C, Williams DM, Kelly SD. 2020 Lipopolysaccharide O-antigens-bacterial glycans made to measure. *J. Biol. Chem.* **295**, 10 593–10 609. (doi:10.1074/jbc.REV120.009402)
91. Sandlin RC, Lampel KA, Keasler SP, Goldberg MB, Stolzer AL, Maurelli AT. 1995 Avirulence of rough mutants of Shigella flexneri: requirement of O antigen for correct unipolar localization of IcsA in the bacterial outer membrane. *Infect. Immun.* **63**, 229–237. (doi:10.1128/iai.63.1.229-237.1995)
92. Van Houte S, Buckling A, Westra ER. 2016 Evolutionary ecology of prokaryotic immune mechanisms. *Microbiol. Mol. Biol. Rev.* **80**, 745–763. (doi:10.1128/MMBR.00011-16)
93. Koonin EV, Makarova KS, Wolf YI, Krupovic M. 2019 Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat. Rev. Genet.* 1–13.
94. Bernheim A, Sorek R. 2019 The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.* **18**, 113–119. (doi:10.1038/s41579-019-0278-2)
95. Hussain FA *et al.* 2021 Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science* **374**, 488–492. (doi:10.1126/science.abb1083)
96. Rocha E, Bikard D. 2021 Many defense systems in microbial genomes, but which is defending whom from what? *EcoEvoXRiv*. (doi:10.32942/osf.io/zuh4c)
97. Oliveira PH, Touchon M, Rocha EP. 2016 Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl Acad. Sci. USA* **113**, 5658–5663. (doi:10.1073/pnas.1603257113)
98. Watson BNJ, Staals RHJ, Fineran PC. 2018 CRISPR-Cas-mediated phage resistance enhances horizontal gene transfer by transduction. *MBio* **9**, e02406-17. (doi:10.1128/mbio.02406-17)
99. Millman A, Bernheim A, Stokar-Avihail A, Fedorenko T, Voichek M, Leavitt A, Oppenheimer-Shaanan Y, Sorek R. 2020 Bacterial retrons function in anti-phage defense. *Cell* **183**, 1551–1561. (doi:10.1016/j.cell.2020.09.065)
100. Bernheim A *et al.* 2020 Prokaryotic viperins produce diverse antiviral molecules. *Nature* **589**, 120–124. (doi:10.1038/s41586-020-2762-2)
101. De Toro M, Garcilláon-Barcia MP, De La Cruz F. 2014 Plasmid diversity and adaptation analyzed by massive sequencing of *Escherichia coli* plasmids. *Microbiol. Spectr.* **2**, 32. (doi:10.1128/microbiolspec.PLAS-0031-2014)
102. Collins RE, Higgs PG. 2012 Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol. Biol. Evol.* **29**, 3413–3425. (doi:10.1093/molbev/mss163)
103. Touchon M, Perrin A, De Sousa JAM, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon D, Rocha EPC. 2020 Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLoS Genet.* **16**, e1008866. (doi:10.1371/journal.pgen.1008866)
104. Wyres KL, Lam MM, Holt KE. 2020 Population genomics of Klebsiella pneumoniae. *Nat. Rev. Microbiol.* **18**, 344–359. (doi:10.1038/s41579-019-0315-1)
105. Gautreau G *et al.* 2020 PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput. Biol.* **16**, e1007732. (doi:10.1371/journal.pcbi.1007732)
106. Harrison E, Brockhurst MA. 2012 Plasmid-mediated horizontal gene transfer is a coevolutionary process. *Trends Microbiol.* **20**, 262–267. (doi:10.1016/j.tim.2012.04.003)
107. Iranzo J, Cuesta JA, Manrubia S, Katsnelson MI, Koonin EV. 2017 Disentangling the effects of selection and loss bias on gene dynamics. *Proc. Natl Acad. Sci. USA* **114**, E5616–E5624. (doi:10.1073/pnas.1704925114)
108. Carroll AC, Wong A. 2018 Plasmid persistence: costs, benefits, and the plasmid paradox. *Can. J. Microbiol.* **64**, 293–304. (doi:10.1139/cjm-2017-0609)
109. San Millan A, Craig Maclean R. 2019 Fitness costs of plasmids: a limit to plasmid transmission. *Microbial Transm.* 65–79. (doi:10.1128/9781555819743.ch4)
110. Mcinerney JO, Mcnally A, O'connell MJ. 2017 Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040. (doi:10.1038/nmicrobiol.2017.40)
111. San Millan A, Peña-Miller R, Toll-Riera M, Halbert Z, Mclean A, Cooper B, MacLean RC. 2014 Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. *Nat. Commun.* **5**, 1–11. (doi:10.1038/ncomms6208)
112. Loftie-Eaton W *et al.* 2017 Compensatory mutations improve general permissiveness to antibiotic resistance plasmids. *Nature Ecol. Evol.* **1**, 1354–1363. (doi:10.1038/s41559-017-0243-2)
113. Hall JPJ, Wright RCT, Harrison E, Muddiman KJ, Wood AJ, Paterson S, Brockhurst MA. 2021 Plasmid fitness costs are caused by specific genetic conflicts enabling resolution by compensatory mutation. *PLoS Biol.* **19**, e3001225. (doi:10.1371/journal.pbio.3001225)
114. Andersson DI, Hughes D. 2010 Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat. Rev. Microbiol.* **8**, 260–271. (doi:10.1038/nrmicro2319)
115. Levin BR, Antonovics J, Sharma H. 1988 Frequency-dependent selection in bacterial populations. *Phil. Trans. R Soc. Lond. B* **319**, 459–472. (doi:10.1098/rstb.1988.0059)
116. Hall JP, Wood AJ, Harrison E, Brockhurst MA. 2016 Source-sink plasmid transfer dynamics maintain gene mobility in soil bacterial communities. *Proc. Natl Acad. Sci. USA* **113**, 8260–8265. (doi:10.1073/pnas.1600974113)
117. Kuo CH, Ochman H. 2009 Deletional bias across the three domains of life. *Genome Biol. Evol.* **1**, 145–152. (doi:10.1093/gbe/evp016)
118. Vandecraen J, Chandler M, Aertsen A, Van Houdt R. 2017 The impact of insertion sequences on bacterial genome plasticity and adaptability. *Crit. Rev. Microbiol.* **43**, 709–730. (doi:10.1080/1040841X.2017.1303661)
119. Oliveira PH, Lemos F, Monteiro GA, Prazeres DM. 2008 Recombination frequency in plasmid DNA containing direct repeats–predictive correlation with repeat and intervening sequence length. *Plasmid* **60**, 159–165. (doi:10.1016/j.plasmid.2008.06.004)
120. Bobay L-M, Ochman H. 2018 Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol. Biol.* **18**, 153. (doi:10.1186/s12862-018-1272-4)
121. Andreani NA, Hesse E, Vos M. 2017 Prokaryote genome fluidity is dependent on effective population size. *ISME J.* **11**, 1719–1721. (doi:10.1038/ismej.2017.36)
122. Ochman H, Moran NA. 2001 Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1099. (doi:10.1126/science.1058543)
123. Giovannoni SJ, Cameron Thrash J, Temperton B. 2014 Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565. (doi:10.1038/ismej.2014.60)
124. Werren JH. 2011 Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proc. Natl Acad. Sci. USA* **108**(Suppl. 2), 10 863–10 870. (doi:10.1073/pnas.1102343108)
125. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit MA. 2017 Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* **11**, 237–247. (doi:10.1038/ismej.2016.90)
126. Ramsay JP, Firth N. 2017 Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr. Opin. Microbiol.* **38**, 1–9. (doi:10.1016/j.mib.2017.03.003)
127. Humphrey S, San Millán Á, Toll-Riera M, Connolly J, Flor-Duro A, Chen J, Ubeda C, Maclean RC, Penadés JR, 2021 Staphylococcal phages and pathogenicity islands drive plasmid evolution. *Nat. Commun.* **12**, 5845. (doi:10.1038/s41467-021-26101-5)
128. Frischer ME, Stewart GJ, Paul JH. 1994 Plasmid transfer to indigenous marine bacterial populations by natural transformation. *FEMS Microbiol. Ecol.* **15**, 127–135. (doi:10.1111/j.1574-6941.1994.tb00237.x)
129. Forterre P, Da Cunha V, Catchpole R. 2017 Plasmid vesicles mimicking virions. *Nat. Microbiol.* **2**, 1340–1341. (doi:10.1038/s41564-017-0032-3)
130. Oliveira PH, Touchon M, Cury J, Rocha EPC. 2017 The chromosomal organization of horizontal gene transfer in bacteria. *Nat. Commun.* **8**, 841. (doi:10.1038/s41467-017-00808-w)

Annex #3

Research article: IntegronFinder 2.0: identification and analysis of integrons across Bacteria, with a focus on antibiotic resistance in Klebsiella.

Bertrand Néron[*], Eloi Littner[*], Matthieu Haudiquet, Amandine Perrin, Jean Cury[†] and Eduardo P. C. Rocha[†]

Published in Microorganisms.

For this work, I provided the *Klebsiella pneumoniae* genomics dataset and genomic annotations like the ST, capsule locus type, antibiotic resistance genes, earlier versions of IntegronFinder annotations, and rarefaction curves.

microorganisms

MDPI

Article

IntegronFinder 2.0: Identification and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in Klebsiella

Bertrand Néron [1,†], Eloi Littner [2,3,4,†], Matthieu Haudiquet [2,5], Amandine Perrin [1,2,4], Jean Cury [2,6,*] and Eduardo P. C. Rocha [2,*]

1 Bioinformatics and Biostatistics Hub, Institut Pasteur, Université de Paris Cité, 75015 Paris, France; bertrand.neron@pasteur.fr (B.N.); amandine.perrin@pasteur.fr (A.P.)
2 Microbial Evolutionary Genomics, Institut Pasteur, Université de Paris Cité, CNRS UMR3525, 75015 Paris, France; eloi.littner@pasteur.fr (E.L.); matthieu.haudiquet@pasteur.fr (M.H.)
3 DGA CBRN Defence, 91710 Vert-le-Petit, France
4 Collège Doctoral, Sorbonne Université, 75005 Paris, France
5 Ecole Doctorale FIRE–Programme Bettencourt, CRI, 75004 Paris, France
6 Laboratoire Interdisciplinaire des Sciences du Numérique, Université Paris-Saclay, CNRS UMR 9015, INRIA, 91400 Orsay, France
* Correspondence: jean.cury@normalesup.org (J.C.); erocha@pasteur.fr (E.P.C.R.)
† These authors contributed equally to this work.

Abstract: Integrons are flexible gene-exchanging platforms that contain multiple cassettes encoding accessory genes whose order is shuffled by a specific integrase. Integrons embedded within mobile genetic elements often contain multiple antibiotic resistance genes that they spread among nosocomial pathogens and contribute to the current antibiotic resistance crisis. However, most integrons are presumably sedentary and encode a much broader diversity of functions. IntegronFinder is a widely used software to identify novel integrons in bacterial genomes, but has aged and lacks some useful functionalities to handle very large datasets of draft genomes or metagenomes. Here, we present IntegronFinder version 2. We have updated the code, improved its efficiency and usability, adapted the output to incomplete genome data, and added a few novel functions. We describe these changes and illustrate the relevance of the program by analyzing the distribution of integrons across more than 20,000 fully sequenced genomes. We also take full advantage of its novel capabilities to analyze close to 4000 *Klebsiella pneumoniae* genomes for the presence of integrons and antibiotic resistance genes within them. Our data show that *K. pneumoniae* has a large diversity of integrons and the largest mobile integron in our database of plasmids. The pangenome of these integrons contains a total of 165 different gene families with most of the largest families being related with resistance to numerous types of antibiotics. IntegronFinder is a free and open-source software available on multiple public platforms.

Keywords: integron; antibiotic resistance; bioinformatics; genomics

Citation: Néron, B.; Littner, E.; Haudiquet, M.; Perrin, A.; Cury, J.; Rocha, E.P.C. IntegronFinder 2.0: Identification and Analysis of Integrons across Bacteria, with a Focus on Antibiotic Resistance in Klebsiella. *Microorganisms* **2022**, *10*, 700. https://doi.org/10.3390/microorganisms10040700

Academic Editors: Jose A. Escudero and Céline Loot

Received: 3 March 2022
Accepted: 22 March 2022
Published: 24 March 2022

1. Introduction

Bacteria evolve novel traits by many processes, including horizontal gene transfer (HGT) driven by phages, conjugative elements, or natural transformation. The impact of human activities has challenged bacteria in numerous ways, including antibiotic therapy and stress caused by pollution and this has spurred the adaptation of bacteria by HGT [1,2]. Once novel genes are acquired by the abovementioned processes, their integration in the novel host genome and subsequent expression is facilitated by the action of mobile genetic elements (MGEs) that rearrange genetic information within genomes [3,4]. Among elements reshaping bacterial genomes, integrons have a particularly important role in gene shuffling and allow the concentration of certain genes in compact genetic regions, which

may result in the co-transfer of related functions (reviewed in [5]). The action of integrons also provides a genetic mechanism to quickly vary the expression of these genes [6]. As a result, integrons have become notorious drivers of the transfer and expression of antibiotic resistance genes [5].

Integrons are an assembly of several genetic elements that include an integrase (IntI) and a succession of gene cassettes flanked by recombination attachment sites (*att*) (Figure 1A). The integrase evolved from the typical tyrosine recombinases to mediate ssDNA recombination by the means of a folded substrate [7]. This unusual activity for a tyrosine recombinase is associated with an extra domain in the protein close to the C-terminus of the protein [7] that allows IntI to be easily distinguished from other integrases [8]. IntI can interact with recombination sites (attachment sites) and lead to the excision or integration of gene cassettes [7]. Integrons have typically one initial attachment site (*attI*) close to the integrase where integration of gene cassettes takes place. The cassettes are flanked by another type of site (*attC*) whose recognition by IntI leads to cassette excision. As a result, the integrase promotes the excision of internal cassettes in the integron and its integration at the *attI* site [9–11]. The *attC* sites are very variable in sequence but conserved in secondary structure [9]. Covariance models using this information can be used to find them efficiently [10]. Since near the *attI* site there is a promoter and most cassettes are devoid of promoters, the cassettes closest to the *attI* site are more likely to be expressed, even if internal promoters may also exist [11]. Hence, integrons include several cassettes and the action of the integrase promotes their shuffling, thereby varying those expressed at a given moment.

One usually distinguishes sedentary from mobile integrons [6]. Sedentary integrons are encoded in the chromosome and can be very large, up to hundreds of cassettes in certain strains of *Vibrio* spp. [12]. In contrast, mobile integrons have few cassettes and are often contained in larger MGEs, such as plasmids. Mobile integrons have been grouped in five classes defined by their similarity in terms of the sequence of IntI, which seem to have emerged recently [13]. Class I integrons are particularly abundant and have the capacity of recombining more diverse types of cassettes (i.e., a broader variety of *attC* sites) than sedentary integrons [14]. This flexibility and their association with MGEs have probably contributed to the presence of many antibiotic resistance genes in integrons and their subsequent spread across nosocomial bacteria. Yet, not all clades of bacteria carry integrons. While these elements are very abundant in Gamma and Beta Proteobacteria [10,15,16], they are only very occasionally identified in strains of Alpha Proteobacteria, Actinobacteria, or Firmicutes [10,17], even among those that are also human-associated. The reasons for this remain unclear. While the frequency of antibiotic resistance genes in integrons has driven a lot of interest in these elements [18], it is likely that in natural environments, they are linked with other traits. The peculiar ability of integrons to shuffle genetic information is also being leveraged to develop synthetic biology approaches [19].

Six years ago, we published an extensive analysis of the distribution of integrons in bacteria [10]. In that work, we searched for integrons, but also for integron-integrases lacking cassettes (In0) and clusters of *attC* sites lacking integron-integrases (CALIN) (Figure 1B–D). For this, we developed a software called IntegronFinder that has become popular to identify integrons in bacterial genomes. The program searches for the integrase using hidden Markov models protein profiles with the software HMMer [20], and for the *attC* sites using covariance models with the software Infernal [21]. These and other genetic components, e.g., promoters, are then clustered if they are co-localized in the DNA sequence. IntegronFinder is thus able to identify integrons with novel cassettes and is complementary to the use of databases such as Integrall [22] that store known integrons and allow to search for homology of the cassettes by sequence similarity. The program is particularly useful to study the wealth of integrons found in environmental bacteria and which often encode unknown functions [23]. Another software to identify *attC* sites, HattCI, was made available at approximately the same time [24]. More recently I-VIP was published to identify class 1 integrons using sequence similarity to a database of integrases and

our own covariance model [16]. However, no software seems to be currently maintained and developed. This is a problem because IntegronFinder version 1 has some limitations. Notably, it was designed to analyze complete genomes. The analysis of draft genomes and metagenomes was possible but required scripting skills and managing an output that was not designed for such data. This is a limitation since recent studies have uncovered novel integrons in metagenomes and in metagenome-assembled genomes [25,26]. IntegronFinder version 1 has also aged poorly in that it was written in Python 2 (now deprecated) and lacks modern tools to facilitate its maintenance. Finally, some of its annotation databases are no longer updated. We have therefore refactored the program, changed the outputs, added flexibility, and updated reference databases to come up with a novel version that is much better adapted to the analysis of incomplete genome data and to stand the passage of time (and pervasive lack of funding for software maintenance). Key changes are indicated in Figure 1E.

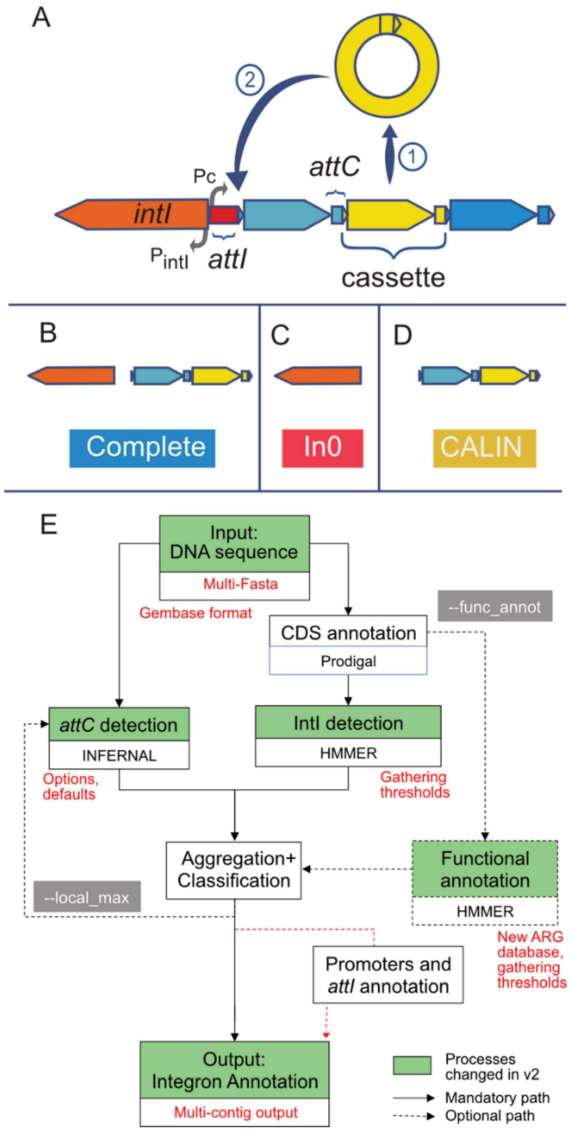


Figure 1. (**A**) Key genetic elements of integrons. (**B**–**D**) Different types of elements searched by IntegronFinder v2. (**E**) Diagram describing the different steps to identify and annotate integrons with in green the processes that were changed in some way (changes in red). Figure modified from [10].

2. Materials and Methods

2.1. Refactoring of IntegronFinder

The first version of IntegronFinder was coded in Python v2.7, which is now deprecated. The program was ported to Python 3.7 (but currently works with versions up to Python 3.10) and the code was refactored to improve efficiency and correct minor bugs. Promoters and *attI1* sites are no longer detected by default to increase speed, especially as *attI* sites are poorly described (and hence not detected by IntegronFinder) outside class I integrons. We have added unit (or non-regression) tests, which make the software more robust and more attractive for future contributors.

IntegronFinder can now be easily run in parallel with a Nextflow script [27] that we provide ready to use. We have also diversified the installation methods, so it can be easily deployed on a variety of machines. Notably, we built a Singularity container which will allow a smooth installation on clusters. We have also updated and improved the documentation, especially on the developer part so that anyone can contribute to the code, to add novel features or fix bugs.

2.2. Novel Functionalities

IntegronFinder v2 has some novel functions in relation to the previous version. A key novel functionality is the systematic use of the gathering thresholds (–cut_ga option in hmmer) that allows to identify hits to the protein profiles with better accuracy than arbitrary e-values. For this, we introduced a gathering threshold in the protein profile specific to the IntI delivered in IntegronFinder that was calculated using the hits of our previous analysis [10].

The reference database for annotating antibiotic resistance genes in the first version of IntegronFinder was RESFAMS (http://www.dantaslab.org/resfams, accessed on 11 November 2018) [28]. However, this database has not been updated recently and while it fits the needs of researchers in metagenomics aiming at identifying distant homologs of antibiotic resistance genes, it is less appropriate to identify antibiotic resistance genes identified in clinical samples. Hence, IntegronFinder now includes the AMRFinderPlus HMM profiles by default [29]. However, any other HMM database, including RESFAMS, can be input in the program. For example, one can input the entire PFAM database [30] for broader functional annotation of the cassettes (at the cost of significant increase in running time).

2.3. Input

IntegronFinder now accepts multifasta files as input. This is a major change since it allows the analysis of genome drafts or metagenome data without the need for scripting an analysis (i.e., calling IntegronFinder recurrently and managing the output). Draft genomes should be used with the linear replicon type (corresponding to the fact that they are contigs). By default, IntegronFinder uses the linear replicon type if there are multiple contigs in the input file. If there are multiple replicons in the input file with different topologies, this can be specified by providing a topology file. IntegronFinder also accepts an option –gembase that corresponds to a particular gene identifier that allows to put multiple replicons in the same file and inform the program that they should be regarded as different replicons. This format can be created by PanACoTA, a pipeline to automatically build the basic bricks of comparative genomics automatically, including download, annotation, and formatting of genome data [31]. The sequence files are initially annotated using Prodigal [32] to provide accurate and uniform annotations of all the sequences.

As described above the input can also include files for annotation of the integron, e.g., for antibiotic resistance genes, using any kind of database compatible with hmmer (http://hmmer.org/documentation.html, accessed on 5 February 2022). Interestingly, a recent study has produced diverse *attC* covariance models for different clades [33]. These can be used in IntegronFinder by invoking the option –attc-model to use another covariance model instead of the one used by default.

2.4. Output

The output of IntegronFinder has been changed to integrate the use of draft genomes. Only three files are now created by default (see the "output" section of the documentation for details about the other possible output files). These three files will be sufficient for most researchers. They include the main output file that contains the position of the different elements (*attC* site, integrase, CDS, etc.). A summary file contains the information of the number and type of elements (complete, In0, and CALIN, see Figure 1B–D) that were found in the input sequence(s). Finally, one file contains a copy of the standard output. This latter file can be excluded using the –mute option. Temporary files can be kept using the –keep-tmp option. They include all the outputs from the different intermediate steps (prodigal, infernal, hmmer). These files can be very useful for certain tasks, notably if one wants to fine tune some parameters such as the aggregation distance threshold. In that case, IntegronFinder will not try to identify the *attC* sites or the integrase again, as they were already detected, but rather allows to only change the way the genetic elements are clustered. Since the temporary files take a significant amount of disk space, they are not kept by default.

Some minor changes were made to the information outputted by IntegronFinder. CALIN are now reported when they have at least 2 *attC* sites (instead of just 1 as before). This value can be changed by the user with –calin-threshold. This is to diminish the probability of false positives when the CALIN are reported. Of note, CALIN with a single *attC* can be true positives, but we let the user decide what to do in that case.

We also provide an alignment of *attC* sites to facilitate their analysis. Standard multiple alignment tools fail to align correctly palindromic sequences. Hence, IntegronFinder now outputs the alignment of *attC* sites built with Infernal, in which the different features of the *attC* sites (R and L boxes and unpaired central spacer) are correctly aligned. This alignment is available with the –keep-tmp option.

2.5. Availability

The code of IntegronFinder is available under the free and open-source license GPLv3 at https://github.com/gem-pasteur/Integron_Finder (accessed on 5 February 2022). The code has been packaged and is available on pypi https://pypi.org/project/integron-finder/ (accessed on 5 February 2022) to make it easy to install with the python packager installer *pip*. We also provide a bioconda package and a container solution https://hub.docker.com/r/gempasteur/integron_finder (accessed on 5 February 2022). A user-friendly Galaxy implementation can be found at https://galaxy.pasteur.fr/tool_runner?tool_id=toolshed.pasteur.fr%2Frepos%2Fkhillion%2Fintegron_finder%2Fintegron_finder%2F2.0%2Bgalaxy0 (accessed on 5 February 2022) [34]. The documentation is available at https://integronfinder.readthedocs.io/en/latest/ (accessed on 5 February 2022).

2.6. RefSeq Complete Genomes

The first dataset used in this study consists of 21,105 complete genomes retrieved from NCBI RefSeq database of high quality complete non-redundant prokaryotic genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/, last accessed on 30 March 2021) [35]. The complete list of genomes can be found in Table S1.

2.7. Klebsiella Data

The Klebsiella genomes were downloaded and annotated with PanACoTA [31] as explained in [36]. Briefly, we downloaded all the 5805 genome assemblies labeled as *Klebsiella pneumoniae sensu stricto* (*Kpn*) from NCBI RefSeq (accessed on 10 October 2018). We removed lower quality assemblies (L90 >100 and number of contigs >990) and strains that were too divergent from the reference strain (Mash distance > 0.06, [37]) or too similar (Mash distance < 0.0001) to other strains. The resulting set of 3980 sequences were consistently re-annotated with Prokka (v1.13.3) [38], and species assignments corrected with Kleborate (https://github.com/katholt/Kleborate, accessed on 5 February 2022) [39]. The accession

numbers for the genomes are in Table S2, along with all the annotations identified in this study.

We built Klebsiella integron pangenomes using PanACoTA v1.3.1 pangenome module with default parameters (-i 0.8, -c 1). The latter clustered integron proteins (defined as the proteins encoded by genes lying between two *attC* or less than 200 bp away from the last *attC*) using MMSeqs2 13-45111 [40] with single-linkage and a threshold of 80% of identity and bidirectional coverage.

We computed the species phylogenetic tree with IQ-TREE, as explained in [36]. Briefly, the 1431 protein families present in more than 99% of the strains in a single copy were aligned with Mafft (v7.407) [41], back-translated to DNA, and concatenated. We used this alignment containing 220,912 parsimony-informative sites over a total of 1,455,396 bp to infer a phylogenetic tree with IQ-TREE (v1.6.7.2) using ModelFinder (-m TEST) [42,43]. We used the best-fit model (GTR+F+I) without gamma correction, because of branch length scaling issues with our large dataset, and assessed the robustness of the phylogenetic inference by calculating 1000 ultra-fast bootstraps (-bb 1000) [44]. The mean support value was 97.6%. We placed the *Kpn* species root according to the outgroup formed by 22 misannotated *Klebsiella quasipneumoniae* subspecies *similipneumoniae* identified by Kleborate.

2.8. Analysis of Antibiotic Resistance Genes (ARGs)

To annotate ARGs we used four databases: ARG-ANNOT (v5) [45], CARD [46], AMRFinderPlus [28] (v3.0.5) (v2019-09-07), and ResFinder (v2019-07-17) [47]. The annotations were made by searching for homology with Abricate (v0.9.8) (https://github.com/tseemann/abricate, accessed on 5 February 2022). While these databases are largely concordant, they do not all include the same types of functions. Hence, we compared them by looking at the number of databases reporting at least one hit (coverage > 80%, identity > 90%) for all the proteins identified in *Klebsiella pneumoniae* integrons. A large majority of proteins reported either 0 or 4 matches, but many had hits in only 1 to 3 databases. Noticeably, the second most abundant family in the pangenome, later labeled as efflux-pump multidrug transporters, hit only the AMRFinderPlus database. Hence, we decided to classify as ARG all the proteins reporting at least one hit in any of the four databases.

We classified each of the integron pangenome families as ARG or non-ARG, which resulted in homogeneous results within families (i.e., same number of databases reporting a hit), except for two families where a small fraction of genes (resp. 6% and 2%) had fewer hits than the other members. As a final criterion, we classified as ARG all the pangenome families for which at least 90% of the members had at least one hit in any of the four databases.

2.9. Saturation Curves

Pangenome saturation curves were computed with the function *specaccum* of the R package *vegan* with default parameters [48], calculating the expected (mean) gene richness using a sample-based rarefaction method that has been independently developed numerous times [49] and is often known as Mao Tau estimate.

2.10. Graphics and Visualizations

Graphics and visualizations were computed in Python with the library *seaborn* [50].

3. Results and Discussion

3.1. Distribution of Integrons across Bacteria

We used IntegronFinder v2 to search for integrons in the complete genomes of the RefSeq NCBI database, including 22,301 replicons named chromosomes and 21,520 replicons named plasmids (the few remaining were phages or not classed). The time that takes to run IntegronFinder is extremely dependent on the existence of *attC* sites in the genome. Genomes lacking *attC* sites run very fast, e.g., 3 s for a genome of *Bacillus subtilis* 168 with a genome size of ~4 Mb on a laptop. If the genome contains *attC* sites, then the use of

the most accurate covariance model can slow down analyses, up to almost 4 min for the genome of *Vibrio cholerae* El Tor N16961 where there is one super-integron [12]. The entire database used in this study is almost ten times larger than the one used for the publication of version 1 (2484 genomes, [10]) and took around 150 CPU hours on a standard laptop. Given the difference in size to the previous database, we started by assessing integron distribution in the new much broader dataset (Figure 2A). We found 3116 genes coding for IntI and 56,994 *attC* sites. The classification of the elements by IntegronFinder v2 indicated the presence of 2760 complete integrons, 356 In0, and 2961 CALIN. These numbers are around ten times larger than those obtained 6 years ago, which is in rough proportion to the difference of size of the two datasets. However, we found a ratio CALIN/integron of almost 1, whereas it was almost 2 six years ago. This is caused by the change in the default parameters to identify CALIN that now require the presence of two *attC* sites. Among genomes that encode an integron, the frequency of In0 elements relative to integrons is also smaller in the new version. This is because the IntI protein is searched using appropriate thresholds (gathering thresholds) that diminish the rates of false positives, and this affects much more the counts of In0 than those of complete integrons (which also have neighboring *attC* sites and are therefore much less likely to be false positives). For example, 6 years ago, we identified two In0 in Actinobacteria. Since we found no complete integron in the clade, this suggested that our hits were false positives. We now find only one In0 (and still no complete integrons) with a much larger dataset that should have resulted in proportionally more false positives.

Among genomes encoding at least one IntI, 22% encode more than one. This result confirms the observations made at the time of IntegronFinder v1.0, 6 years ago, where 20% of the genomes had more than one of the three [10]. One key difference between the current data and the previous data concerns the frequency of complete integrons encoded in plasmids. Six years ago, they accounted for a small minority of all the integrons detected, while they are now more numerous than those in chromosomes (53% of the total). The recent increase in the intensity of sampling and sequencing of nosocomial pathogens, which often have antibiotic resistance genes in mobile integrons (in plasmids), may explain these differences.

We found an average of 14 *attC* sites per complete integron and almost 6 per CALIN. CALINs were proposed to be degraded integrons [10], and it is expected that they have fewer cassettes than integrons that encode an integrase allowing the integration of novel cassettes. We then separated integrons and CALIN with more than 10 *attC* (typically corresponding to sedentary elements) from the others. This shows, as expected, the presence of many more *attC* sites in complete sedentary integrons than in large CALIN (Figure 2B). When the analysis was restricted to clades known to have sedentary integrons (*Vibrio* and *Xanthomonas* spp.), we also found that CALIN tend to have fewer *attC* sites (Figure S1). Finally, as expected, integrons or CALIN in plasmids have fewer (respectively, 11 and 2.5 times less) *attC* sites than those in chromosomes.

The distribution of integrons in plasmids revealed three elements with more than 10 *attC* sites (Figure 2C), which we studied in detail. The two largest elements were in very large replicons (1.1 Mb in *Gemmatirosa kalamazoonesis* KBS708 and 599 kb in *Vibrio* sp. HDW18) that are probably secondary chromosomes. Hence, the largest integron we could identify in replicons compatible with the typical sizes of mobile plasmids (1–300 kb, [51]) was a 10 *attC* integron in *Klebsiella* sp. RHBSTW-00464 plasmid 5 (46 kb). The second largest plasmid integron was in *Enterobacter hormaechei* subsp. steigerwaltii BD-50-Eh plasmid pBD-50-Eh_VIM-1 and was reported before [52]. Hence, while we had previously used 19 *attC* sites as a threshold to delimit mobile from sedentary integrons, our present results suggest that using a 10 *attC* sites threshold could be more adequate.

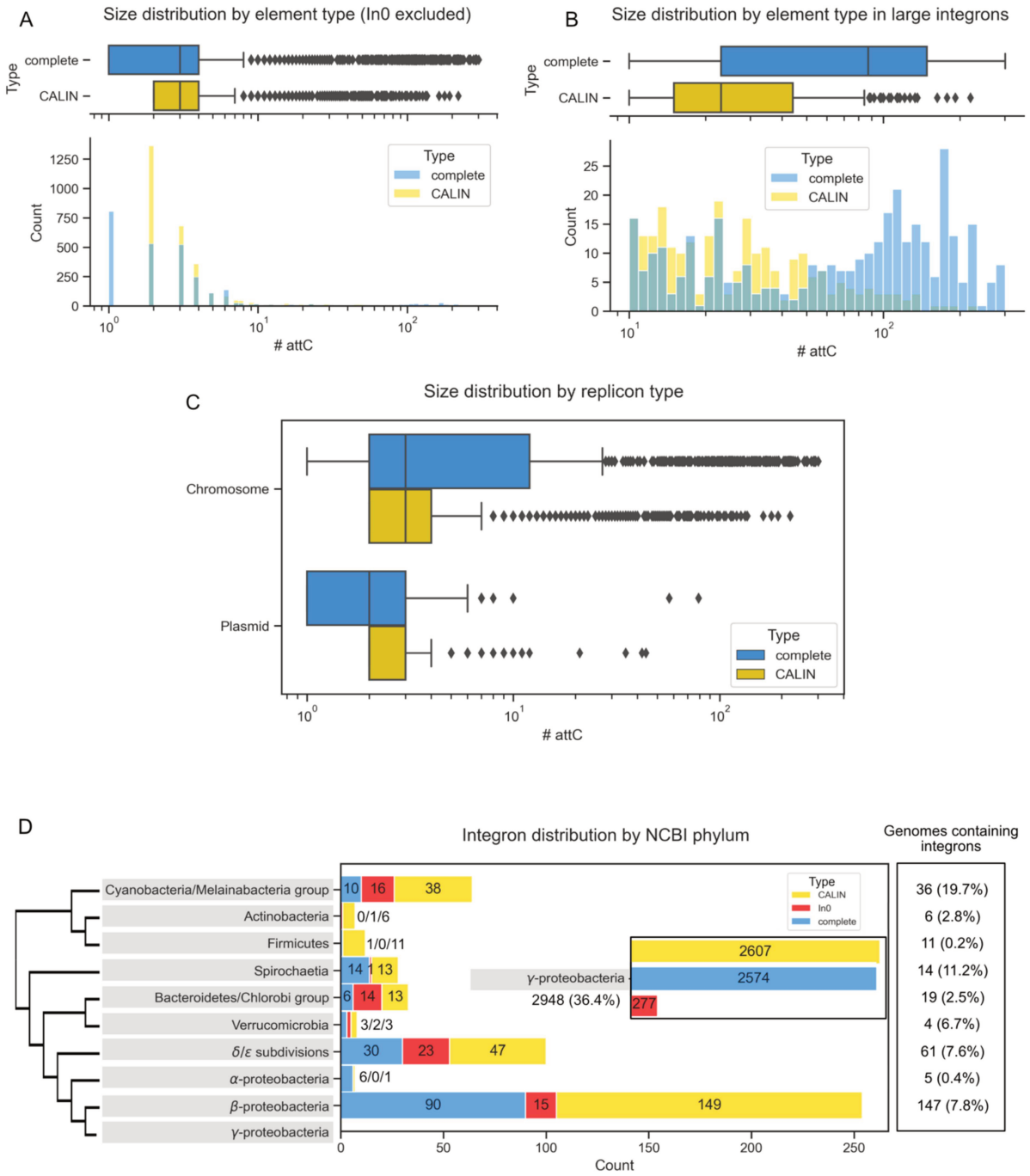


Figure 2. Statistics concerning the distribution of integrons in the RefSeq NCBI database. Distribution of the number of *attC* sites found per element (complete integron or CALIN) (**A**) and zoom on the distribution of large elements (>10 sites, **B**). Distribution of number of *attC* sites per type of integron and replicon (**C**). Distribution of integrons across major bacterial phyla. Only phyla comprising integrons are shown. The percentage in the last box is the proportion of the genomes in our database that contain at least one integron (**D**).

The taxonomic distribution of integrons follows previously described trends [10,16,33]. The vast majority of complete integrons (93%) are found in γ-Proteobacteria, a percentage that largely exceeds the representation of these genomes in the database (38%) (Figure 2D). In line with previous work, we did not identify integrons in *Chlamydiae* nor in *Tenericutes*, despite the good representation of these two phyla in our dataset (respectively, 170 and 400 genomes). However, the much larger dataset analyzed in this work, as well as some

changes in taxonomy and species classification, led to the identification of several clades that lacked integrons six years ago and now have a few. For example, we found 2 CALIN elements in the phylum Nitrospira, both occurring in the genome of *Candidatus Nitrospira inopinata*, a bacterium known to perform complete ammonia oxidation to nitrate [53]. The newly defined phylum Acidithiobacillia has 2 complete integrons and 8 CALIN over a total of 11 genomes. Finally, while only In0 had been detected in Bacteroidetes in our previous study, we identified complete integrons in the Bacteroidetes/Chlorobi group. More precisely, on top of the ones found previously in Chlorobi and Ignavibacteriae, we identified a complete integron in *Salinibacter ruber* M1 [54], a halophilic bacteria that belongs to Bacteroidetes.

Our previous study failed to identify complete integrons in Actinobacteria, Firmicutes, and Alpha Proteobacteria. We now identify a few elements in these clades. However, their frequency is extremely small relative to the high frequency of such genomes in the database. Firmicutes account for 22% of the genomes, Alpha Proteobacteria for 6%, and Actinobacteria for 10%. The only complete integron in Firmicutes was found in *Limnochorda pilosa* strain HC45, and a few CALIN were found in different genera across the phylum. The 6 complete integrons in Alpha Proteobacteria were found in species of Sphingobium and Agrobacterium. As mentioned above, Actinobacteria lack complete integrons, have only one In0, and a few CALIN. These results show that while integrons, or their components, are not altogether missing in these clades, in line with previous works [55,56], they are indeed extremely rare and the observed occurrences may represent recent acquisitions. Considering the presence of integrons in very diverse clades and their ability to spread within plasmids and transposons, the lack of integrons in certain clades remains intriguing as it suggests the existence of some incompatibility between integrons and the genetic background of these large bacterial clades.

A recent preprint revealed the presence of integrons in metagenome assembled genomes (MAGs) of Archaea, but not in complete genomes [25]. Since this study used a combination of a pre-release of IntegronFinder 2.0 and HattCI [24], we wished to understand if the novel version of IntegronFinder alone, which is better suited to study MAGs, could identify these systems. We also failed to identify complete integrons, CALIN, or In0 in the genomes of Archaea present in our database. This is not unexpected since the clades where the integrons have been identified are very poorly represented in RefSeq. Therefore, we downloaded five MAGs of the study cited above and searched them for integrons. We confirmed the presence of two complete integrons, one CALIN, and one In0 whereas one complete integron in [25] was detected as a CALIN in our analysis. Hence, if Archaea are found to encode integrons like those suggested by the recent analysis of MAGs, IntegronFinder version 2 is expected to be able to identify them.

3.2. Recent Spread of Integrons in K. pneumoniae

To exemplify the use of the novel version of IntegronFinder, we analyzed a large collection of draft genomes of *Klebsiella pneumoniae* (*Kpn*). *Kpn* is a Gram-negative bacterial species belonging to the Enterobacteriaceae family, naturally occurring in soil, freshwater, and mammalian gut and is considered an opportunistic pathogen [57]. It is a member of the ESKAPE pathogens, the list of high-risk multi-resistant nosocomial pathogens from the World Health Organization. Aside from having the largest known mobile integron in the RefSeq database (see above), *Kpn* is also an interesting example of a nosocomial bacteria that is thought to have recently acquired integrons carrying antibiotic resistance genes. While the presence of integrons in the species is well known [58], its characterization has not been done at the species level. Hence, we collected 3980 genomes of the species, re-annotated them uniformly, and searched for complete integrons in the draft genomes. The analysis by IntegronFinder took 14 h using 5 CPUs on a standard desktop computer.

We identified 1855 complete integrons, 2590 CALIN, and 405 In0 in 2709 of the 3980 genomes, resulting in a (CALIN+In0)/complete ratio of 1.61 (Table S4). As our dataset mostly comprises draft genomes, it was expected that some complete integrons

would be divided in several contigs, leading to an increase of the proportion of CALIN and In0. This phenomenon may be enhanced by the presence of *attC* sites very similar in sequence, which complicates the assembly of contigs harboring integrons. For comparison, among the 730 *K. pneumoniae* complete genomes comprised in our database extracted from RefSeq (see previous section), 492 complete integrons, 282 CALIN, and 21 In0 were identified in 435 genomes. The (CALIN+In0)/complete ratio is 0.62 in this dataset, which seems to confirm that analyses of draft genomes will tend to overestimate the frequency of In0 and CALIN. To further investigate how many complete integrons may have been split between contigs, we looked for occurrences of In0, complete integrons, and CALIN where the integrase or a gene cassette was the first gene in a contig. We found 819 isolates harboring at least one CALIN together with one complete integron or In0 at the border of a contig. Among these, we found 339 In0 and 532 complete integrons. These might correspond to true complete integrons that were split between contigs during assembling. Hence, the analysis of genomes that are not fully assembled may result in the underestimation of the number of cassettes, incomplete integrons, and spurious identification of CALIN elements.

For the following analysis, we decided to focus on the 1855 complete integrons identified in 1677 genomes. In other words, around 42% of the genomes of our *Kpn* in the database have an integron (all data in Table S3). In genomes carrying integrons, the number of integrons per genome is usually one but can be higher. Notably, the genomes of strains AR_0039 (GCF_001874875.1) and DHQP1002001 (GCF_001704235.1) carry four complete integrons each. To study the distribution of integrons across the species, we built a phylogenetic tree and plotted the presence of integrons in function of the sequence type (ST) (Figure 3). This analysis revealed no clear pattern of aggregation, since integrons are found across the species. This suggests multiple independent acquisitions of integrons in the species recent past.

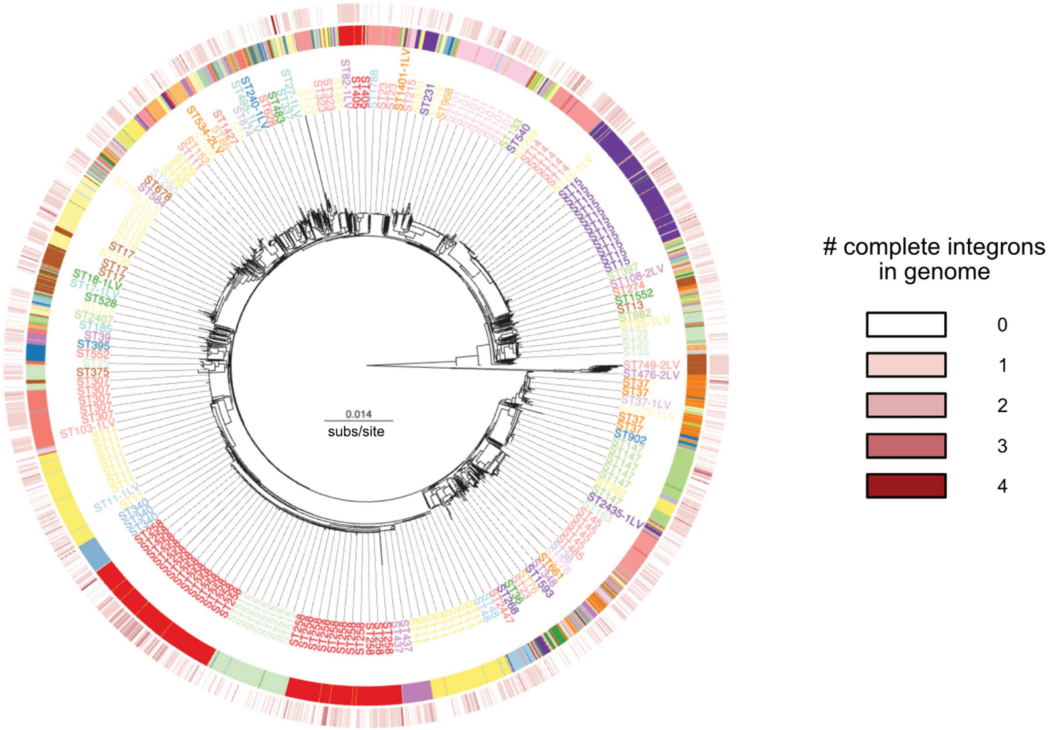


Figure 3. Phylogenetic tree of the core genome of *Klebsiella pneumoniae*, with an indication of the sequence type (ST) in the ribbon (and mention to the most frequent in the intermediate circle). The outer circle indicates the number of complete integrons in each genome. Tree was built as indicated in Methods and drawn using Microreact [59].

Integrons in *Kpn* have been studied because they have a few well-known antibiotic resistance genes [59,60]. One might thus have expected to find little genetic diversity among them. To assess the diversity of these integrons, we analyzed their gene repertoires. The integrons in *Kpn* encoded a total of 5763 protein coding genes, with the largest elements having up to 15 genes or 7 *attC* sites. Hence, we could not find in this dataset integrons as large as the one found in the complete genomes of RefSeq. This may be caused by different samplings or because integrons in draft genomes may be split in several contigs. We used PanACoTA [31] to compute the pangenome of the integrons and identified 165 different gene families. Hence, the integrons of *Kpn* are small, as expected from mobile integrons, but carry a large diversity of gene families. We computed saturation curves for the integrons' pangenome and observed that it is open, i.e., after analyzing almost 2000 elements, the curve does not show evidence of saturation (Figure 4). Further genomic sampling of *Kpn* will thus likely reveal novel gene cassettes.

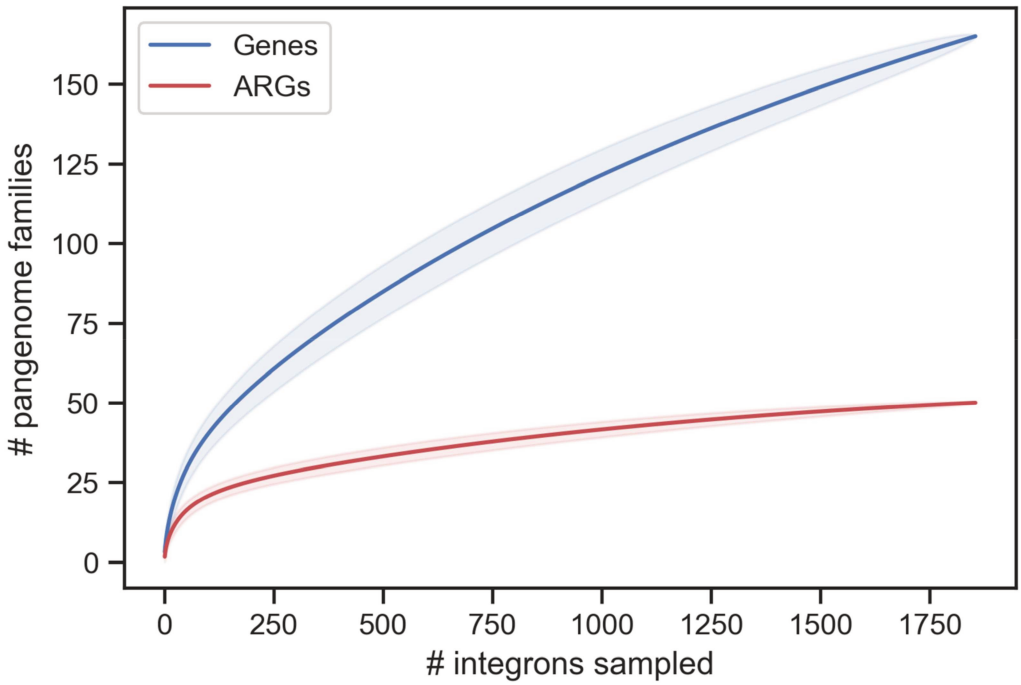


Figure 4. Analysis of the pangenome of the integrons of *Kpn*. The figure displays two saturation curves, representing the expected (mean) gene richness when increasing the number of integrons. They were computed using the Mao Tau estimate (see Methods for more details). The shadowed regions correspond to one standard deviation.

We then detailed the functions of the genes encoded in the cassettes of integrons (IntI being obviously the most frequent protein of integrons). Expectedly, most of the largest gene families in cassettes are associated with antibiotic resistance genes (ARG). To study them in detail we used four databases of ARGs: CARD, ARG-Annot, ResFinder, and AMRFinderPlus. The integrated analysis of the results (see Methods) resulted in the identification of 50 families of ARG. An analysis of the 20 most frequent gene families reveals 12 that are ARGs, including multidrug transporters and enzymes associated with resistance to antiseptics, aminoglycosides, diaminopyrimidine (e.g., trimethoprim), fluoroquinolones, rifamycin, chloramphenicol, or beta-lactams. The second largest family of the cassettes codes for the protein AadA2, an aminoglycoside nucleotidyltransferase, and is present in more than a third of the integrons (654). This family is also present in multiple copies within an integron in 50 genomes. The list of resistances to beta-lactams was very diverse, including OXA-1, OXA-2, OXA-10, OXA-18, and metallo-beta-lactamase. To assess the diversity of ARG families, we computed a saturation curve for this fraction of the pangenome (Figure 4). The results show that while this subset is necessarily less

diverse, it still accounts for up to 50 different gene families, and that further sampling of *Kpn* genomes will most likely reveal novel genes in extant *Kpn* populations. Some of the other large families encode functions implicated in recombination, notably DDE or other types of recombinases. However, most families of genes are of unknown function, suggesting that even in species such as *Klebsiella*, where integrons seem to be recent and driven by selection for antibiotic resistance, there may be many other relevant functions encoded in integrons.

4. Conclusions

Many of the improvements in IntegronFinder version 2 will be almost invisible to the user because they involve software engineering, but they will facilitate its maintenance in the next decade. This is important because funding for software maintenance is almost inexistent. The novel most visible capacities of IntegronFinder will allow it to better tackle large datasets of genome, which tend to be composed of draft genomes, and metagenomes. Here, we have exemplified its use with the database of complete genomes and a large dataset of *Kpn* draft genomes. As it stands, we regard speed as a major limitation to the use of the program in exceptionally large datasets, e.g., in metagenomes or in datasets with hundreds of thousands of genomes. The current deadlock is the use of the covariance model that is computationally expensive in its most accurate version. Significant advances in the speed of IntegronFinder will require the development of novel methods to use such models.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/microorganisms10040700/s1, Figure S1: Statistics concerning the distribution of integrons in two genera known for harboring sedentary integrons in RefSeq NCBI database; Table S1: List of complete genomes from RefSeq used in the study; Table S2: Accession numbers for the *K. pneumoniae* genomes; Table S3: IntegronFinder results for complete genomes; Table S4: IntegronFinder results for *K. pneumoniae*.

Author Contributions: Conceptualization, B.N., E.L., J.C. and E.P.C.R.; data curation, E.L. and M.H.; formal analysis, E.L. and M.H.; funding acquisition, E.P.C.R.; methodology, B.N., A.P. and J.C.; supervision, E.P.C.R.; writing—original draft, E.L. and E.P.C.R.; writing—review and editing, B.N., E.L., M.H., J.C. and E.P.C.R. All authors have read and agreed to the published version of the manuscript.

Funding: Work allowing the development of IntegronFinder was funded by INCEPTION project (Agence National de la Recherche: ANR-16-CONV-0005), the Fondation pour la Recherche Médicale (EQU201903007835), and the Laboratoire d'Excellence IBEID Integrative Biology of Emerging Infectious Diseases (Agence National de la Recherche: ANR-10-LABX-62-IBEID). E.L. is supported by the Direction Générale de l'Armement (DGA).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data of this study will be made available by the authors, without reservation, to any qualified researcher. The software code is publicly available in the Git.

Acknowledgments: We thank input on IntegronFinder from the many users that have alerted us to issues or made suggestions of improvement. This work used the computational and storage services (TARS cluster) provided by the IT department at Institut Pasteur, Paris. We thank Fabien Mareuil for the IntegronFinder integration in galaxy@pasteur.

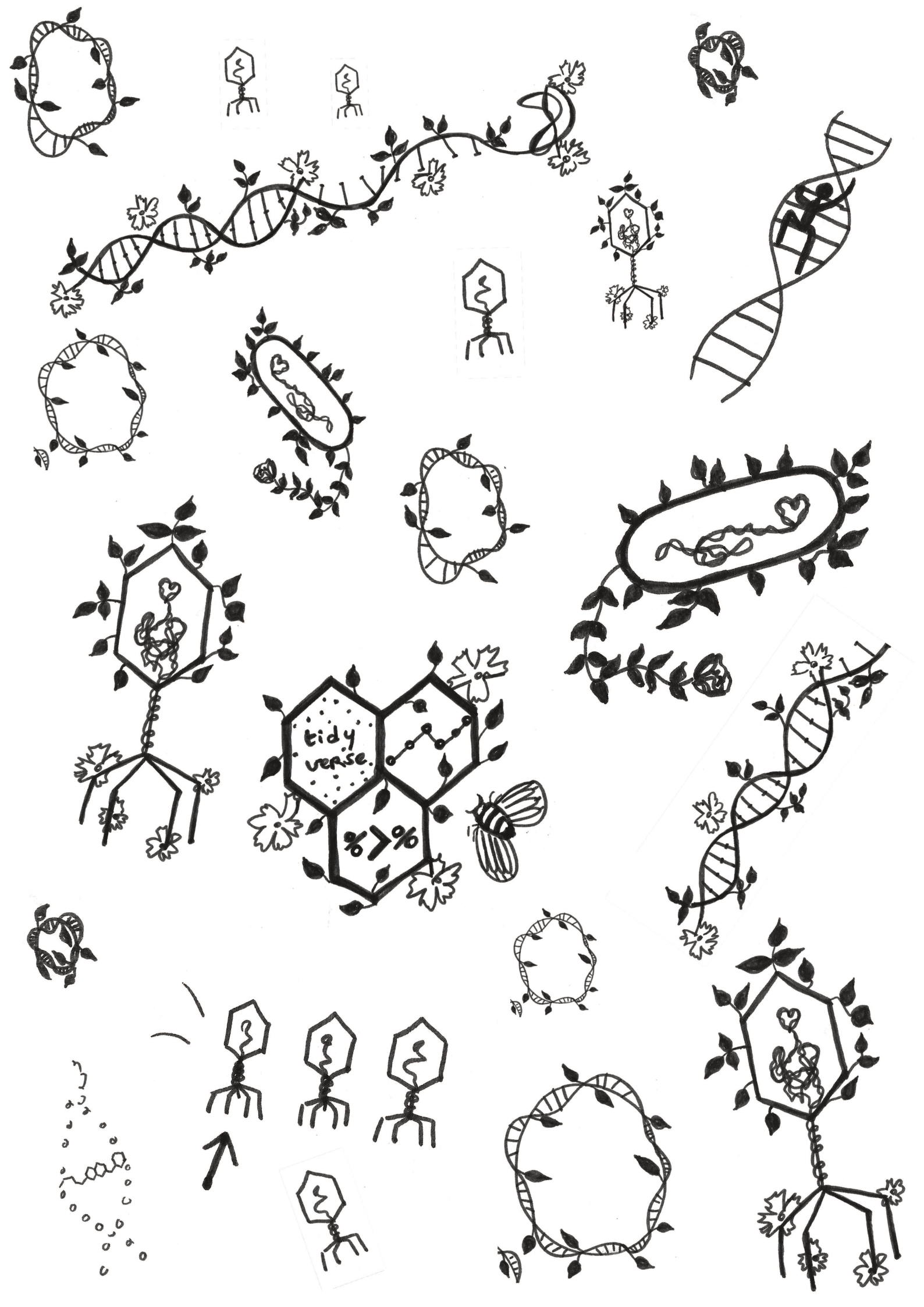
Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Haudiquet, M.; de Sousa, J.M.; Touchon, M.; Rocha, E. Selfish, promiscuous, and sometimes useful: How mobile genetic elements drive horizontal gene transfer in microbial populations. *EcoEvoRxiv* **2021**. preprint. [CrossRef]
2. Arnold, B.J.; Huang, I.-T.; Hanage, W.P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Genet.* **2021**, *20*, 206–218. [CrossRef]

3. Cerveau, N.; Leclercq, S.; Bouchon, D.; Cordaux, R. Evolutionary Dynamics and Genomic Impact of Prokaryote Transposable Elements. In *Evolutionary Biology—Concepts, Biodiversity, Macroevolution and Genome Evolution*; Pontarotti, P., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 291–312.
4. Bourque, G.; Burns, K.H.; Gehring, M.; Gorbunova, V.; Seluanov, A.; Hammell, M.; Imbeault, M.; Izsvák, Z.; Levin, H.L.; Macfarlan, T.S.; et al. Ten things you should know about transposable elements. *Genome Biol.* **2018**, *19*, 199. [CrossRef] [PubMed]
5. Cambray, G.; Guerout, A.M.; Mazel, D. Integrons. *Annu. Rev. Genet.* **2010**, *44*, 141–166.
6. Escudero, J.A.; Loot, C.; Nivina, A.; Mazel, D. The Integron: Adaptation on Demand. *Microbiol. Spectr.* **2015**, *3*, 139–161. [CrossRef]
7. Bouvier, M.; Demarre, G.; Mazel, D. Integron cassette insertion: A recombination process involving a folded single strand substrate. *EMBO J.* **2005**, *24*, 4356–4367. [CrossRef]
8. Smyshlyaev, G.; Bateman, A.; Barabas, O. Sequence analysis of tyrosine recombinases allows annotation of mobile genetic elements in prokaryotic genomes. *Mol. Syst. Biol.* **2021**, *17*, e9880. [CrossRef]
9. Nivina, A.; Grieb, M.S.; Loot, C.; Bikard, D.; Cury, J.; Shehata, L.; Bernardes, J.; Mazel, D. Structure-specific DNA recombination sites: Design, validation, and machine learning–based refinement. *Sci. Adv.* **2020**, *6*, eaay2922. [CrossRef]
10. Cury, J.; Jové, T.; Touchon, M.; Néron, B.; Rocha, E. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucl. Acids Res.* **2016**, *44*, 4539–4550. [CrossRef]
11. Michael, C.A.; Labbate, M. Gene cassette transcription in a large integron-associated array. *BMC Genet.* **2010**, *11*, 82. [CrossRef]
12. Mazel, D.; Dychinco, B.; Webb, V.A.; Davies, J. A Distinctive Class of Integron in the Vibrio cholerae Genome. *Science* **1998**, *280*, 605–608. [CrossRef] [PubMed]
13. Recchia, G.D.; Hall, R.M. Gene cassettes: A new class of mobile element. *Microbiology* **1995**, *141*, 3015–3027. [CrossRef] [PubMed]
14. Biskri, L.; Bouvier, M.; Guérout, A.-M.; Boisnard, S.; Mazel, D. Comparative Study of Class 1 Integron and *Vibrio cholerae* Superintegron Integrase Activities. *J. Bacteriol.* **2005**, *187*, 1740–1750. [CrossRef] [PubMed]
15. Nemergut, D.R.; Robeson, M.S.; Kysela, R.F.; Martin, A.P.; Schmidt, S.K.; Knight, R. Insights and inferences about integron evolution from genomic data. *BMC Genomics* **2008**, *9*, 261. [CrossRef] [PubMed]
16. Ni Zhang, A.; Li, L.-G.; Ma, L.; Gillings, M.; Tiedje, J.M.; Zhang, T. Conserved phylogenetic distribution and limited antibiotic resistance of class 1 integrons revealed by assessing the bacterial genome and plasmid collection. *Microbiome* **2018**, *6*, 130. [CrossRef]
17. Li, Y.; Yang, L.; Fu, J.; Yan, M.; Chen, D.; Zhang, L. Microbial pathogenicity and virulence mediated by integrons on Gram-positive microorganisms. *Microb. Pathog.* **2017**, *111*, 481–486. [CrossRef]
18. Stalder, T.; Barraud, O.; Casellas, M.; Dagot, C.; Ploy, M.-C. Integron involvement in environmental spread of antibiotic resistance. *Front Microbiol.* **2012**, *3*, 119.
19. Bikard, D.; Julié-Galau, S.; Cambray, G.; Mazel, D. The synthetic integron: An in vivo genetic shuffling device. *Nucl. Acids Res.* **2010**, *38*, e153. [CrossRef]
20. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195.
21. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **2013**, *29*, 2933–2935. [CrossRef]
22. Moura, A.; Soares, M.; Pereira, C.; Leitão, N.; Henriques, I.; Correia, A. INTEGRALL: A database and search engine for integrons, integrases and gene cassettes. *Bioinformatics* **2009**, *25*, 1096–1098. [CrossRef] [PubMed]
23. Sandoval-Quintana, E.; Lauga, B.; Cagnon, C. Environmental integrons: The dark side of the integron world. *Trends Microbiol.* **2022**, in press. [CrossRef]
24. Pereira, M.B.; Wallroth, M.; Kristiansson, E.; Axelson-Fisk, M. HattCI: Fast and Accurate attC site Identification Using Hidden Markov Models. *J. Comput. Biol.* **2016**, *23*, 891–902.
25. Ghaly, T.M.; Tetu, S.G.; Penesyan, A.; Qi, Q.; Rajabal, V.; Gillings, M.R. Discovery of integrons in Archaea: Platforms for cross-domain gene transfer. *bioRxiv* **2022**. bioRxiv:2022.02.06.479319.
26. Buongermino Pereira, M.; Österlund, T.; Eriksson, K.M. A comprehensive survey of integron-associated genes present in meta-genomes. *BMC Genomics* **2020**, *21*, 495.
27. Di Tommaso, P.; Chatzou, M.; Floden, E.W.; Barja, P.P.; Palumbo, E.; Notredame, C. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **2017**, *35*, 316–319.
28. Gibson, M.K.; Forsberg, K.; Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **2015**, *9*, 207–216. [CrossRef]
29. Feldgarden, M.; Brover, V.; Gonzalez-Escalona, N.; Frye, J.G.; Haendiges, J.; Haft, D.H.; Hoffmann, M.; Pettengill, J.B.; Prasad, A.B.; Tillman, G.E.; et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.* **2021**, *11*, 12728. [CrossRef]
30. Finn, R.D.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L. The Pfam protein families database: Towards a more sus-tainable future. *Nucl. Acids Res.* **2016**, *44*, D279–D285.
31. Perrin, A.; Rocha, E.P.C. PanACoTA: A modular tool for massive microbial comparative genomics. *NAR Genomics Bioinform.* **2021**, *3*, lqaa106.
32. Hyatt, D.; Chen, G.-L.; Locascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef]
33. Ghaly, T.M.; Tetu, S.G.; Gillings, M.R. Predicting the taxonomic and environmental sources of integron gene cassettes using structural and sequence homology of attC sites. *Commun. Biol.* **2021**, *4*, 946. [CrossRef] [PubMed]

34. Afgan, E.; Baker, D.; Batut, B.; van den Beek, M.; Bouvier, D.; Čech, M.; Chilton, J.; Clements, D.; Coraor, N.; Grüning, B.A.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucl. Acids Res.* **2018**, *46*, W537–W544. [CrossRef] [PubMed]
35. Haft, D.H.; DiCuccio, M.; Badretdin, A.; Brover, V.; Chetvernin, V.; O'Neill, K.; Li, W.; Chitsaz, F.; Derbyshire, M.K.; Gonzales, N.R.; et al. RefSeq: An update on prokaryotic genome annotation and curation. *Nucl. Acids Res.* **2018**, *46*, D851–D860. [CrossRef] [PubMed]
36. Haudiquet, M.; Buffet, A.; Rendueles, O.; Rocha, E.P.C. Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen *Klebsiella pneumoniae*. *PLoS Biol.* **2021**, *19*, e3001276. [CrossRef]
37. Ondov, B.D.; Treangen, T.J.; Melsted, P.; Mallonee, A.B.; Bergman, N.H.; Koren, S.; Phillippy, A.M. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **2016**, *17*, 132. [CrossRef]
38. Seemann, T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* **2014**, *30*, 2068–2069. [CrossRef]
39. Lam, M.M.C.; Wick, R.R.; Watts, S.C.; Cerdeira, L.T.; Wyres, K.L.; Holt, K.E. A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. *Nat. Commun.* **2021**, *12*, 4188. [CrossRef]
40. Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028. [CrossRef]
41. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780.
42. Nguyen, L.-T.; Schmidt, H.A.; Von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef] [PubMed]
43. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; Von Haeseler, A.; Jermiin, L.S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **2017**, *14*, 587–589. [CrossRef] [PubMed]
44. Hoang, D.T.; Chernomor, O.; Von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **2017**, *35*, 518–522. [CrossRef]
45. Gupta, S.K.; Padmanabhan, B.R.; Diene, S.M.; Lopez-Rojas, R.; Kempf, M.; Landraud, L.; Rolain, J.-M. ARG-ANNOT, a New Bioinformatic Tool to Discover Antibiotic Resistance Genes in Bacterial Genomes. *Antimicrob. Agents Chemother.* **2014**, *58*, 212–220. [CrossRef]
46. Jia, B.; Raphenya, A.R.; Alcock, B.; Waglechner, N.; Guo, P.; Tsang, K.K.; Lago, B.A.; Dave, B.M.; Pereira, S.; Sharma, A.N.; et al. CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucl. Acids Res.* **2017**, *45*, D566–D573. [CrossRef]
47. Kleinheinz, K.A.; Joensen, K.G.; Larsen, M.V. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage* **2014**, *4*, e27943.
48. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **2003**, *14*, 927–930. [CrossRef]
49. Chiarucci, A.; Bacaro, G.; Rocchini, D.; Fattorini, L. Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. *Commun. Ecol.* **2008**, *9*, 121–123. [CrossRef]
50. Waskom, M.L. Seaborn: Statistical data visualization. *J. Open Source Softw.* **2021**, *6*, 3021. [CrossRef]
51. Smillie, C.; Pilar Garcillan-Barcia, M.; Victoria Francia, M.; Rocha, E.P.C.; de la Cruz, F. Mobility of Plasmids. *Microbiol. Mol. Biol. Rev.* **2010**, *74*, 434–452.
52. Campos-Madueno, E.I.; Gmuer, C.; Risch, M.; Bodmer, T.; Endimiani, A. Characterisation of a new blaVIM-1-carrying IncN2 plasmid from an *Enterobacter hormaechei* subsp. steigerwaltii. *J. Glob. Antimicrob. Resist.* **2021**, *24*, 325–327. [CrossRef] [PubMed]
53. Daims, H.; Lebedeva, E.V.; Pjevac, P.; Han, P.; Herbold, C.; Albertsen, M.; Jehmlich, N.; Palatinszky, M.; Vierheilig, J.; Bulaev, A.; et al. Complete nitrification by *Nitrospira* bacteria. *Nature* **2015**, *528*, 504–509. [CrossRef] [PubMed]
54. González-Torres, P.; Gabaldón, T. Genome Variation in the Model Halophilic Bacterium *Salinibacter ruber*. *Front. Microbiol.* **2018**, *9*, 1499. [CrossRef] [PubMed]
55. Nešvera, J.; Hochmannová, J.; Pátek, M. An integron of class 1 is present on the plasmid pCG4 from Gram-positive bacterium *Corynebacterium glutamicum*. *FEMS Microbiol. Lett.* **1998**, *169*, 391–395. [CrossRef]
56. Nandi, S.; Maurer, J.J.; Hofacre, C.; Summers, A.O. Gram-positive bacteria are a major reservoir of Class 1 antibiotic resistance integrons in poultry litter. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 7118–7122. [CrossRef]
57. Paczosa, M.K.; Mecsas, J. Klebsiella pneumoniae: Going on the Offense with a Strong Defense. *Microbiol. Mol. Biol. Rev.* **2016**, *80*, 629–661. [CrossRef]
58. Poirel, L.; Le Thomas, I.; Naas, T.; Karim, A.; Nordmann, P. Biochemical Sequence Analyses of GES-1, a Novel Class A Extended-Spectrum β-Lactamase, and the Class 1 Integron In52 from *Klebsiella pneumoniae*. *Antimicrob. Agents Chemother.* **2000**, *44*, 622–632. [CrossRef]
59. Argimón, S.; AbuDahab, K.; Goater, R.J.E.; Fedosejev, A.; Bhai, J.; Glasner, C.; Feil, E.J.; Holden, M.T.G.; Yeats, C.A.; Grundmann, H.; et al. Microreact: Visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genomics* **2016**, *2*, e000093. [CrossRef]
60. Kaushik, M.; Kumar, S.; Kapoor, R.K.; Virdi, J.S.; Gulati, P. Integrons in *Enterobacteriaceae*: Diversity, distribution and epidemiology. *Int. J. Antimicrob. Agents* **2018**, *51*, 167–176.


tidy
verse
%>%