Génomique comparée à grande échelle de souches d'Escherichia coli responsables de bactériémies chez l'Homme : implications cliniques et analyse des réseaux métaboliques

Guilhem Royer

université PARIS-SACLAY

Génomique comparée à grande échelle de souches de *Escherichia coli* responsables de bactériémies chez l'Homme : implications cliniques et analyse des réseaux métaboliques

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°577 : Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Sciences de la vie et de la santé
Unité de recherche : Université Paris-Saclay, Univ Évry, CNRS, CEA, Génomique métabolique, 91057, Évry-Courcouronnes, France.
Référent : Université d'Évry Val d'Essonne

Thèse présentée et soutenue à Evry, le 31/03/2021, par

Guilhem ROYER

Composition du Jury

Marie-Frédérique LARTIGUE Professeure des universités – Praticienne hospitalier, Université de Tours – CHRU de Tours	Présidente du jury
Sylvain BRISSE Directeur de recherche, Institut Pasteur	Rapporteur & Examinateur
Catherine SCHOULER Directrice de recherche INRAE, INRAE Val de Loire	Rapporteur & Examinatrice
Thierry NAAS Maître de conférences – Praticien hospitalier, Université Paris-Saclay – Hôpital Bicêtre AP-HP	Examinateur
David VALLENET Directeur de recherche CEA, Genoscope CEA	Directeur de thèse
Jean-Winoc DECOUSSER Professeur des universités – Praticien hospitalier, Université Paris-Est Créteil – Hôpitaux universtiaires Henri-Mondor APHP	Co-Directeur
Claudine MEDIGUE Directrice de recherche CNRS, Genoscope CEA	Invitée
Hélène CHIAPELLO Ingénieure de recherche INRAE, INRAE - Université de Saclay	Invitée

Thèse de doctorat

NNT : 2021UPASL012

“It is a truth universally acknowledged that there are only two kinds of bacteria.
One is *Escherichia coli*, and the other is not.”
Downie & Young, 2001, Nature

Remerciements

Je tiens à remercier les membres du jury de me faire l'honneur de juger mon travail :

A Sylvain Brisse et Catherine Schouler, d'avoir accepté d'en être les rapporteurs,

A Marie-Frédérique Lartigue, Thierry Naas et Hélène Chiapello d'avoir accepté d'en être les examinateurs.

Je tiens à remercier mes deux co-directeurs, Claudine et David, de m'avoir accueilli au sein du LABGeM pour m'initier aux rudiments de l'analyse des génomes bactériens. Merci pour votre gentillesse, votre accessibilité et vos conseils scientifiques toujours pertinents.

Je tiens également à remercier mon troisième co-directeur, Jean-Winoc. Merci pour ta confiance et ton soutien depuis maintenant tant d'années. Merci d'avoir toujours fait ton possible pour m'offrir les conditions de travail idéales.

Merci à Erick pour le temps que vous m'avez accordé, les conseils, la rigueur et les connaissances que vous avez sues me transmettre tout au long de cette thèse. C'est une chance d'avoir pu travailler à vos côtés.

Je remercie également Jean-Michel Pawlotsky de m'avoir accueilli dans le laboratoire de l'hôpital Mondor au sortir de l'internat et d'avoir insisté pour que deux années de cette thèse se déroulent à temps plein.

Je remercie l'ensemble du LABGeM pour tous les moments partagés. Merci pour tout ce que vous m'avez transmis et pour l'accueil chaleureux que vous m'avez réservé. Je retiendrais tous ces moments de discussions et de bonne humeur, les petits-déjeuners, les apéro au Palais sans oublier le Génoboule !

Je remercie également Amin, Christophe et Marco. Je suis fier de notre projet sur l'antibiorésistance, et je nourris secrètement l'espoir d'une nouvelle aventure à vos côtés!

Je remercie l'ensemble de l'équipe IAME qui m'a aussi accueilli au cours de ces quatre années. J'ai appris beaucoup à vos côtés, dans une ambiance toujours aussi agréable. Il est rare de trouver en un même lieu tant de gens si brillants et abordables à la fois.

Merci également à mes collègues biologistes et techniciens du laboratoire de bactériologie de l'hôpital Henri Mondor. Vous m'avez toujours accordé le temps nécessaire pour mon activité universitaire, notamment au cours de ces derniers mois de rédaction parfois éprouvants. C'est chaque jour un plaisir et une fierté de venir de travailler à vos côtés. Merci à Mélanie et Bernadette pour le temps passé à séquencer toutes ces souches, première étape indispensable aux travaux présentés ici. Je remercie également l'ensemble du département de prévention, diagnostic et traitement des infections qui participe au quotidien à cette ambiance agréable et constructive.

Je remercie ma famille, mes parents, mon frère et ma belle-sœur, ma grand-mère. Merci pour votre présence et votre soutien depuis toujours. J'ai une pensée particulière pour mon grand-père, tes histoires me manquent…

Je remercie aussi mes amis de l'époque du lycée, de la fac, de l'internat. Vous avez souvent été une bouffée d'oxygène et m'avez offert une ouverture d'esprit tout au long de ces années. Chaque moment partagé avec vous est une petite pépite de bonheur !

Enfin, merci Candice pour le soutien et le réconfort que tu m'apportes au quotidien. L'année écoulée a parfois été difficile, pour de multiples raisons, mais tu as toujours fait passer notre bonheur au premier plan. Déjà tant de choses construites tous les deux, et encore tant d'autres à venir !

Table des matières

REMERCIEMENTS **5**

LISTE DES ABRÉVIATIONS **9**

LISTE DES FIGURES **10**

LISTE DES TABLEAUX **12**

INTRODUCTION **13**

ÉTAT DE L'ART **15**

I. Les bactériémies **16**
1. Définitions et généralités 16
2. Ecologie microbienne au cours des bactériémies 18
3. Incidence et mortalité associée aux bactériémies 19

II. *Escherichia coli* : généralités **21**
1. Généralités bactériologiques 21
2. Réservoir et mode de vie 22

III. *Escherichia coli* : méthodes d'analyses et structure de la population **27**
1. L'étude des sérotypes 27
2. Le profil électrophorétique des enzymes 28
3. L'amplification des acides nucléiques 30
4. L'apport de la génomique 34

IV. Les souches pathogènes extra-intestinales **38**
1. Comment définir un ExPEC? 38
2. ExPEC et facteurs de virulence 40
3. Résistance aux antibiotiques et ExPEC 53
4. Plasmides et ExPEC 60
5. ExPEC et modèle animale 64

V. Etudes comparées à grande échelle de *Escherichia coli* responsables de bactériémies **67**
1. L'ère pré-génomique 67
2. L'utilisation du génome complet 69

VI. Implication du métabolisme dans la pathogénicité extra-intestinale de *Escherichia coli* **72**
1. Analyse ciblée du métabolisme 73
2. Analyse des réseaux métaboliques 75

OBJECTIFS DE LA THÈSE **78**

RÉSULTATS 80

I. Stratégie d'analyse des génomes de Septicoli 81
1. Stratégie d'analyse des génomes 81
2. Identification de séquences plasmidiques : l'approche PlaScope 86

II. Analyse de génomes de *Escherichia coli* responsables de bactériémies chez l'Homme 100
1. L'étude Septicoli 100
2. Dynamique de la population de *Escherichia coli* responsable de bactériémies sur 12 ans 116

III. Reconstruction des réseaux métaboliques de souches de *Escherichia coli* commensales et pathogènes 161
1. Introduction 161
2. Matériel et méthodes 162
3. Résultats 165
4. Discussion 181

Conclusions et perspectives 184

TRAVAUX ANNEXES 189

RÉFÉRENCES BIBLIOGRAPHIQUES 193

Liste des abréviations

APEC : Avian pathogen *E. coli*

BGN : Bacille à Gram Négatif

BLSE : Bêta-Lactamase à Spectre Etendu

C3G : Céphalosporines de 3ème génération

C4G : Céphalosporines de 4ème génération

CC : Complexe clonal

CNF : Cytotoxic necrotising factor

DAEC : Diffusely adherent *E. coli*

DAF : Decay accelerating factor

EAEC : Enteroaggregative *E. coli*

EHEC : Enterohemorrhagic *E. coli*

EIEC : Enteroinvasive *E. coli*

EPEC : Enteropathogenic *E. coli*

ETEC : Enterotoxinogenic *E. coli*

ExPEC : Extra-intestinal pathogen *E. coli*

HPI : High pathogenicity island

InPEC : Intestinal pathogen *E. coli*

IS : Séquence d'insertion

MLEE : Multi-locus enzyme electrophoresis

MLST : Multi-locus sequence typing

NMEC : Neonatal meningitis-causing *E. coli*

PAI : Îlots de pathogénicité

PAVM : Pneumopathie acquise sous ventilation

RAISIN : Réseau d'alerte, d'investigation et de surveillance des infections nosocomiales

SEPEC : Sepsis-associated *E. coli*

SOFA : Sequential organ failure assessment

SRIS : Syndrome de réponse inflammatoire systémique

ST : Sequence type

STc : Sequence type complex

STEC : Shigatoxin-producing *E. coli*

UPEC : Urinary pathogen *E. coli*

UPGMA : Unweighted pair group method with arithmetic mean

Liste des figures

Figure 1. Principales portes d'entrée/foyers infectieux des bactériémies..... 17

Figure 2. Exemple de galerie d'identification biochimique api20e ensemencée avec une souche de *Escherichia coli.*..... 22

Figure 3. Principaux pathovars de *E. coli*.. 25

Figure 4**.** Dendrogramme construit à partir des profils électrophorétiques de souches isolées en Finlande et en Suède, à partir des fèces, des urines au cours de bactériuries asymptomatiques, cystites et pyélonéphrites et de bactériémies/méningites.. 29

Figure 5. Profils électrophorétiques bi-dimensionnels de carboxylestérases B de souches de *E. coli* commensales et pathogènes extra-intestinales. 30

Figure 6. Arbre phylogénétique non enraciné, obtenu à partir des séquences des 8 gènes du MLST (schéma Pasteur), après élimination des recombinaisons.. 32

Figure 7**.** Taille des génomes et des pangénomes des 8 groupes de *E. coli* sensu stricto. . 36

Figure 8**.** Représentation graphique des interactions entre le fond génétique des souches, leur virulence dans un modèle animal, l'état immunitaire de l'hôte et les formes cliniques observées au cours des infections urinaires..... 39

Figure 9. Structure des fimbriae de type 1 et P fimbriae..... 42

Figure 10. Structure chimiques des principaux sidérophores de *E. coli.*..... 47

Figure 11. Exemple d'îlots de pathogénicité retrouvés au sein de souches ExPEC archétypales.. 52

Figure 12. A) Structure phylogénétique du phylogroupe D. B) Carte génétique de l'intégron de classe 1 identifié dans la souche UMN026 du STc69..... 55

Figure 13. Phylogénie du ST131 après prise en compte des recombinaisons..... 57

Figure 14. Représentation de la carte génétique du plasmide pS88 isolé de la souche *E. coli* S88.. 63

Figure 15. Heatmap à partir des distances euclidiennes calculées en fonction des présences/absences de déterminants génétiques et de la létalité dans un modèle murin... 65

Figure 16. Distribution des différents STs responsables de bactériémies en Angleterre sur une période 11 ans. 70

Figure 17. Adaptation du métabolisme de *E. coli* en fonction des nutriments disponibles, de la pression exercée par l'hôte et de la nécessité de concurrencer les éventuels micro-organismes présents. 73

Figure 18. A) Evolution du coremétabolisme de 29 souches de *E. coli* (a) et du coregénome (b). B) Evolution du panmétabolisme (a) et du pangénome (b).. 75

Figure 19. Analyse des correspondances multiples à partir des occurrences des réactions dans des souches de *E. coli* désignées par leur groupe phylogénétique (A) et leur mode de vie (B).. 76

Figure 20. Schéma de la stratégie PETA'n'C pour l'analyse des génomes de Septicoli.. 82

Figure 21. Distribution des souches de STc131 de l'étude de Kallonen *et al.* (Kallonen et al., 2017) au sein de différents clades en valeur absolue (A) et pourcentage (B). 160

Figure 22. Nombre de voies métaboliques et de réactions au sein de l'ensemble des souches analysées en fonction de leur fréquence.. 167

Figure 23. Distribution des voies dans le métabolisme core (≥ 95%) et variable (< 95%) en fonction de leur classification d'après la base de données MetaCyc. 167

Figure 24. Distribution des voies dans le métabolisme core (≥ 95%) et variable (< 95%) en fonction de leur classification de deuxième niveau d'après la base de données MetaCyc. 168

Figure 25. Arbres construits par la méthode UPGMA à partir des vecteurs de présence/absence A) des voies métaboliques, B) des réactions. 169

Figure 26. Heatmap des valeurs de complétion moyenne des voies métaboliques par phylogroupe. 171

Figure 27. Heatmap des valeurs de complétion moyenne des voies métaboliques par mode de vie. 172

Figure 28. Représentation des deux premiers axes d'une analyse des correspondances multiples à partir des niveaux de complétion des voies métaboliques. 173

Figure 29. Heatmap des valeurs de complétion moyenne des voies métaboliques pour les cinq STc majeurs des bactériémies et le reste de STc.. 174

Figure 30. Voie de dégradation du 3-phenylpropanoate, 3-(3-hydroxyphenyl)propanoate, cinnamate et 3-hydroxycinnamate en 2-hydroxypentadienoate, et voie de dégradation du 2-hydroxypenta-2,4-dienoate..... 176

Figure 31. Alignement des régions entourant l'opéron lactose dans des génomes de différents phylogroupes, sous-groupes B2, *Escherichia* clades, *E. albertii* et *E. fergusonii* 178

Figure 32. A) Représentation schématique de l'environnement génétique de la région contenant le groupe de gènes *mhp*. B) Arbre phylogénétique construit à partir des gènes du coregénome de 27 souches. 180

Liste des tableaux

Tableau 1. Calcul du score Sequential Organ Failure Assessment (SOFA) 16

Tableau 2. Ecologie microbienne observées au cours des sepsis. 18

Tableau 3**.** Estimation du nombre annuel de bactériémies et décès associés. 19

Tableau 4. Principaux sous-groupes au sein de phylogroupe B2 34

Tableau 5. Distribution des principaux facteurs virulence ExPEC au sein des souches isolées de bactériémies de l'adulte et des souches commensales fécales. 41

Tableau 6. Mode de vie et phylogroupes des souches étudiées. 166

Introduction

Escherichia coli est une bactérie commensale du tube digestif de l'Homme et de nombreux autres mammifères. Comme nous le verrons dans une première partie bibliographique, la population bactérienne de cette espèce est très diverse et structurée et comprend aussi des organismes pathogènes responsables d'infections digestives à type de diarrhées, et ou encore d'infections extra-intestinales. Les connaissances accumulées ont permis de mettre en évidence des associations préférentielles entre certaines sous-populations et des modes de vie ou une pathogénicité donnée. Certains facteurs de virulence ont également pu être mis en évidence et leur implication dans la pathogénicité de la bactérie parfois vérifiée expérimentalement. Cependant si les mécanismes sous-jacents à la pathogénicité intestinale sont clairement identifiés, il en est autrement pour les infections extra-intestinales. Aucun profil de virulence unique n'est en effet associé à ces infections. Par ailleurs, outre l'augmentation du nombre de bactériémies à *E. coli* au cours des dernières décennies et leur mortalité toujours élevée, nous avons pu observer l'émergence et la diffusion mondiale de certains clones virulents et parfois multirésistants aux antibiotiques. Les pressions de sélection responsables de telles modifications de l'épidémiologie de ces infections sont pour l'heure toujours discutées, et il semblerait que la résistance antibiotique ne soit qu'un élément parmi d'autres.

Grâce à la démocratisation du séquençage haut débit, il est aujourd'hui possible d'étudier un nombre important de génomes. De telles analyses pourraient permettre d'analyser plus en profondeur l'ensemble des déterminants génétiques associés à la sévérité et/ou à la porte d'entrée des bactériémies. L'un des objectifs principaux de cette thèse est de fournir une description fine de la population responsable de bactériémies de l'adulte en région parisienne. Ces résultats peuvent ensuite être inclus dans l'analyse des facteurs de pronostic de ces infections afin de déterminer la part jouée par l'hôte d'un côté, et la bactérie de l'autre.

Par ailleurs, nous nous intéresserons à la comparaison de deux collections de bactériémies provenant de populations homogènes au sein d'une même zone géographique espacées de 12 années. Nous nous focaliserons plus particulièrement sur la dynamique des principaux sous-groupes et clones au cours de cette période et proposerons des scénarios évolutifs associés à la modification de l'épidémiologie.

Enfin, un dernier objectif de cette thèse est de tirer profit des génomes des souches de *E. coli* isolées de bactériémies ainsi que de souches commensales et de souches

pathogènes/colonisatrices pulmonaires. Pour cela, nous nous intéresserons aux voies métaboliques codées par ces génomes afin de déterminer d'éventuelles associations entre métabolisme et mode de vie.

État de l’art

I. Les bactériémies

1. Définitions et généralités

Une bactériémie correspond à la présence de bactéries dans le sang, généralement associée à des signes systémiques d'infection comme la présence de fièvre et d'un syndrome inflammatoire biologique. On définissait auparavant le sepsis comme un syndrome de réponse inflammatoire systémique (SRIS) en réponse à une infection documentée (Bone et al., 1992). Le SRIS est lui-même défini par la présence d'au moins 2 signes parmi i) une hypo ou une hyperthermie (<36 ou >38°C), ii) une tachycardie (>90 battements par minute), iii) une tachypnée (> 20 respirations par minute ou une hyperventilation avec PaCO2 > 32 mmHg), iv) une hyperleucocytose (> 12000 leucocytes/mm3) ou une leucopénie (<4000 leucocytes/mm3). En fonction de la présence de défaillance d'organes et de l'impossibilité de juguler l'hypotension par un remplissage adéquat, on définissait également deux stades de sévérités successifs, le sepsis sévère et le choc septique, associés à une mortalité croissante. Ces définitions ont été mises à jour et en partie simplifiées en 2016 (Singer et al., 2016). Il convient aujourd'hui de parler de sepsis lorsqu'apparaît une dysfonction d'organe menaçant le pronostic vital et que celle-ci est causée par une réponse inappropriée de l'hôte à une infection. L'entité "sepsis sévère" a disparu alors que la notion de "choc septique" persiste et correspond à un sous-groupe du sepsis nécessitant l'usage de drogues vasopressives pour maintenir une pression artérielle moyenne ≥ 65 mmHg et une concentration sanguine de lactate > 2 mmol/L malgré un remplissage adéquat. Par ailleurs, il existe un score, le Sequential Organ Failure Assessment (SOFA), permettant d'objectiver le sepsis en cas d'augmentation de 2 points en raison de l'infection (Tableau 1).

Tableau 1. Calcul du score Sequential Organ Failure Assessment (SOFA)

Score SOFA	0 point	1 point	2 points	points	4 points
PaO2/FiO2	>400	301-400	201-300	101-200 et VA	≤ 100 et VA
Plaquettes x103 /mm3	>150	101-150	51-100	21-50	≤20
Bilirubine, mg/L (mmol/L)	<12 (<20)	12-19 (20-32)	20-59 (33-101)	60-119 (102-204)	>120 (>204)
Hypotension	PAM ≥70mmHG	PAM < 70mmHG	Dopamine ≤ 5 ou dobutamine (toute dose)	Dopamine > 5 ou adrénaline ≤ 0,1 ou noradrénaline ≤ 0,1	Dopamine > 15 ou adrénaline > 0,1 ou noradrénaline > 0,1
Score de Glasgow	15	13-14	10-12	6-9	<6
Créatinine, mg/L (µmol/L) ou diurèse	<12 (<110)	12-19 (110-170)	20-34 (171-299)	35-49 (300-440) ou <500mL/j	>50 (>440) ou <200mL/j

VA : ventilation assistée. PAM : pression artérielle moyenne [estimée par (PAsystolique + 2 x PAdiastolique) / 3]. Amines : dose en µg/kg/min

D’un point de vue physiopathologique, le passage des bactéries dans le sang fait généralement suite à l’extension d’un foyer infectieux préexistant (*e.g.* : urinaire, pulmonaire), une translocation digestive ou encore une colonisation d’un matériel exogène tel un cathéter. On parle alors de porte d’entrée de la bactériémie. La mise en évidence de cette porte d’entrée est essentielle pour déterminer la source initiale de l’infection et ainsi la contrôler et enrayer la diffusion des bactéries. D’autre part, elle permet également d’orienter le traitement probabiliste, en prenant en compte dans le “pari antibiotique” les germes les plus fréquemment rencontrés pour une porte d’entrée donnée. Les principales portes d’entrée au cours des bactériémies et leur fréquence sont présentées dans la Figure 1. L’épidémiologie est dominée par les portes d’entrée urinaire, les cathéters, les voies respiratoires et le tractus digestif. Chez *E. coli* plus spécifiquement, la porte d’entrée urinaire domine suivie de l’origine digestive (Lefort et al., 2011; Martinez et al., 2006). Cependant ces données dépendent fortement du profil de patients inclus dans les études, certaines retrouvant par exemple une source digestive dans plus de la moitié des cas probablement en raison de patients plus graves cliniquement (Mora-Rillo et al., 2015).

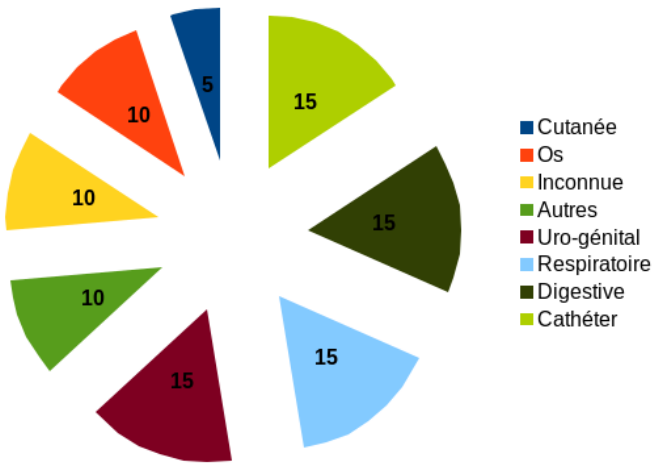


Figure 1. Principales portes d’entrée/foyers infectieux des bactériémies (%) (d’après Collège des Universitaires des Maladies Infectieuses et Tropicales (CMIT), 2018).

2. Ecologie microbienne au cours des bactériémies

L'écologie microbienne des sepsis chez l'Homme se limite généralement à un nombre restreint d'agents bactériens/fongiques (Tableau 2). En effet, une dizaine d'espèces seulement sont responsables d'une majeure partie des infections. Bien qu'il existe des variations géographiques et temporelles, deux espèces dominent largement dans les pays industrialisés : *Escherichia coli* et *Staphylococcus aureus.* Ces données sont également retrouvées dans la revue de Laupland *et al.*, dans laquelle *E. coli* apparaît comme étiologie la plus fréquente des bactériémies au Canada, aux Etats-Unis, en Australie ou encore en Nouvelle-Zélande (Kevin B. Laupland & Church, 2014). Cependant il faut noter que l'épidémiologie de ces infections est bien différente dans d'autres régions du monde, comme en Afrique par exemple où *Salmonella enterica* et *Streptococcus pneumoniae* dominent parmi les bactéries à Gram négatif et positif, respectivement (Reddy et al., 2010).

Tableau 2. Ecologie microbienne observées au cours des sepsis.

Micro-organismes	**Nombre de cas (%)**		
	Javaloyas et al., 2002	Diekema et al., 2003	Bouza et al., 2007
Cocci à Gram positif			
Staphylococcus aureus	68 (8.2)	184 (19.8)	62 (17,5)
S. epidermidis/SCN*	23 (2.7)	114 (12,3)	49 (13,8)
Enterococcus sp.	18 (2.1)	104 (11,2)	31 (8,8)
Streptococcus pneumoniae	83 (10)	41 (4,4)	27 (7,6)
Streptococcus spp.	43 (5.2)	30 (3,3)	25 (7,1)
Bacilles à Gram négatif aéro-anaérobie facultatif et aérobie stricte			
Escherichia coli	335 (40)	154 (16,7)	66 (18,6)
Klebsiella sp.	50 (6)	74 (8,0)	20 (5,6)
Pseudomonas aeruginosa	29 (3.5)	53 (5,7)	13 (3,7)
Proteus sp.	33 (4)	ND**	11 (3,1)
Enterobacter sp.	25 (3)	22 (2,4)	8 (2,3)
Salmonella sp.	26 (3)	ND	7 (1,9)
Bactéries anaérobies			
Bacteroides sp.	22 (2.6)	ND	8 (2,3)
Clostridium sp.	13 (1.5)	ND	4 (1,1)
Champignons	5 (0,6)	39 (4,2)	8 (2,3)
Autres germes	64 (7,7)	75 (8,1)	15 (4,2)
Total	832	929	354

*SCN : Staphylocoque à coagulase négative

**ND : Donnée non disponible

L'écologie est également fonction du caractère communautaire, lié aux soins ou nosocomial de la bactériémie. Ainsi au cours des bactériémies nosocomiales on observe une

augmentation de la fréquence de certains pathogènes comme *S. aureus*, les staphylocoques à coagulase négative ou *P. aeruginosa*, tandis que les infections à *E. coli* ou *Streptococcus pneumoniae* tendent à diminuer (Diekema et al., 2003).

3. Incidence et mortalité associée aux bactériémies

L'incidence de ces infections est élevée, atteignant selon l'estimation de Goto & Hasan 174 à 204 et 166 à 189 cas pour 10^5 habitants/an aux Etats-Unis et en Europe, respectivement (Goto & Al-Hasan, 2013). De plus, la mortalité associée à ces infections reste importante. D'après Goto & Al-Hasan toujours, elle atteint généralement 12 à 20% en Amérique du Nord et en Europe, en faisant une des 7 causes principales de décès, devant toutes les autres causes infectieuses (Tableau 3).

Tableau 3**.** Estimation du nombre annuel de bactériémies et décès associés (d'après Goto & Al-Hasan, 2013).

Pays/Région	**Nombre estimé de bactériémies annuellement**	**Pourcentage de décès liés aux bactériémies (%)**	**Nombre estimé de décès liés aux bactériémies annuellement**
Etats-Unis	535920 - 628320	13,5	72349 - 84823
Canada	39542 - 49069	18	7117 - 8832
Danemark	9130	20,6	1881
Finlande	8736	13	1144
Angleterre	96012	13 - 20	12482 - 19202
Europe	1213460 - 1381590	13 - 20	157750 - 276318

Ici encore ces chiffres cachent une certaine hétérogénéité liée bien évidemment aux comorbidités des patients mais également à l'espèce en cause et/ou au caractère nosocomial ou communautaire. Selon les études, les bactériémies d'origine nosocomiale pourraient être associées à un taux de mortalité de 30 voire 32% (Goto & Al-Hasan, 2013). L'étude française du Réseau d'Alerte, d'Investigation et de Surveillance des Infections Nosocomiales (RAISIN) datant de 2004, met en évidence une mortalité plus importante des bactériémies nosocomiales dues à *S. aureus* (15,6%) et *P. aeruginosa* (21,5%) comparées aux autres agents infectieux (Réseau d'alerte, d'investigation et de surveillance des infections nosocomiales (Raisin), 2004).

Concernant *E. coli*, la mortalité est de 12,9% dans l'étude de Lefort *et al.* (Lefort et al., 2011). Cette valeur est cohérente avec les résultats de Bhattacharya *et al.* qui, à partir de données de surveillance anglaise, estiment que les valeurs en 2020-21 seront de 90,5 cas de bactériémies pour 10^5 habitants et 11,5 décès pour 10^5 habitants (Bhattacharya et al., 2018).

II. *Escherichia coli* : généralités

Le milieu du 19ème siècle voit naître l'âge d'or de la bactériologie, avec la découverte de nombreux pathogènes parmi lesquels *Mycobacterium leprae* (1873), *Bacillus anthracis* (1877), *Neisseria gonorrhoeae* (1879), *Salmonella typhi* (1880), *Mycobacterium tuberculosis* (1882), *Corynebacterium diphtheriae* ou encore *Vibrio chloraea* (1883) (Friedmann, 2014). Le pédiatre allemand Theodor Escherich va lui aussi grandement contribuer au développement de la microbiologie à cette époque (Escherich, 1885). En 1885, au travers de l'étude des selles de nouveaux nés, il met en évidence l'importante diversité bactérienne obtenue par culture des fèces et la prédominance d'une bactérie de forme bacillaire, qu'il nommera alors *Bacterium coli commune* (Escherich, 1885). Plus tard renommée *Escherichia coli* en son honneur, cette bactérie deviendra l'un des organismes vivants les plus étudiés à ce jour, à la fois dans le monde médical, la recherche scientifique en général ou encore l'industrie et l'agronomie.

1. Généralités bactériologiques

E. coli, aussi appelé colibacille, est un bacille à Gram négatif (BGN) appartenant à la famille des *Enterobacteriaceae* au sein de la classe des *Gammaproteobacteria* et du phylum des *Proteobacteria*. Cette bactérie présente un métabolisme respiratoire de type aéro-anaérobie facultatif et peut, en fonction des souches, être mobile grâce à la présence de flagelles ou immobile. Son identification en laboratoire de microbiologie clinique a longtemps reposé sur son profil biochimique, à l'aide par exemple de galeries miniaturisées telles les galeries "api"[1] (Figure 2). Parmi les caractères biochimiques les plus importants, on retiendra : la production de gaz, la fermentation du lactose, du mannitol, du sorbitol, la production d'indole et de pyruvate à partir du tryptophane, l'absence d'uréase, de gélatinase, de désaminase oxydative, l'absence de production de sulfure d'hydrogène, d'utilisation du malonate, de l'inositol et de l'adonitol (Le Minor & Richard, 1993). Par ailleurs, tout comme les autres entérobactéries, elle ne possède pas de cytochrome-oxydase. Si le profil biochimique exhaustif n'est plus utilisé aujourd'hui en routine en raison de l'avènement de l'identification par spectrométrie de masse MALDI-TOF, certaines de ces spécificités restent à la base de tests rapides ou bien de la composition de certains milieux de culture, notamment en raison de leur faible coût et de leur simplicité.

[1] https://www.biomerieux.fr/diagnostic-clinique/galeries-didentification-api

Figure 2. Exemple de galerie d'identification biochimique api20e ensemencée avec une souche de *Escherichia coli.* La souche est ONPG+, ADH-, LDC+, ODC+, CIT-, H2S-, URE-, TDA-, IND+, VP-, GEL-, GLU+, MAN+, INO-, SOR+, RHA+, SAC-, MEL+, AMY-, ARA+.

E. coli possède également des caractères antigéniques particuliers à la base du sérotypage et permettant d'identifier certaines souches par des approches autrefois phénotypiques et désormais génotypiques. On note ainsi les antigènes O (de l'allemand "Ohne Kapsel") dits somatiques, les antigènes H (de l'allemand "Hauch") dits flagellaires, et les antigènes K (de l'allemand "Kapsel") dits capsulaires. Plus rarement, certaines souches possèdent un antigène M leur conférant un aspect muqueux en culture (Cruickshank, 1936; Le Minor & Richard, 1993). Certaines souches célèbres de *E. coli* sont d'ailleurs désignées directement par ces combinaisons de sérotype (cf. paragraphe "L'étude des sérotypes") : les souches O157:H7 responsables d'épidémies d'intoxications alimentaires liées à la consommation de viande mal cuite, ou bien les souches *E. coli* K1 responsable de méningites chez le nouveau-né. Comme nous le verrons par la suite, on observe un déséquilibre dans la diversité de ces antigènes selon le caractère pathogène ou commensal des souches.

2. Réservoir et mode de vie

Bien que moins abondante au sein du microbiote digestif que les anaérobies, *E. coli* est la bactérie aéro-anaérobie facultative dominante au niveau intestinal chez l'Homme (Tenaillon et al., 2010). Les souches historiques de *E. coli* ont d'ailleurs, pour la plupart, été retrouvées à l'état commensal dans des prélèvements fécaux : la souche NCTC86 de Theodor Escherich provenant de selles de nouveaux-nés (Desroches et al., 2018; Méric et al., 2016), la célèbre souche K-12 isolée de selles d'un patient diphtérique convalescent (Bachmann, 1996), et probablement la souche de Félix d'Hérelle *E. coli* B qui a permis les premiers travaux sur les bactériophages (Daegelen et al., 2009). On la retrouve également chez la plupart des vertébrés, aussi bien chez les mammifères, que chez les oiseaux ou encore les reptiles avec des proportions et des structures de population variables (Gordon & Cowling, 2003; Lu et al., 2016; Mohsin et al., 2017; Ochman & Selander, 1984a; Skurnik et al., 2016). Mais le mode de

vie de cette bactérie peut fluctuer selon les souches, l'hôte et l'environnement. On la retrouve ainsi au cours d'infections intestinales, extra-intestinales voire même parfois hors de l'hôte dans des habitats secondaires préférentiellement hydriques (Touchon et al., 2020).

a. Souches commensales

Le tube digestif des mammifères constitue l'habitat primaire de *E. coli*, et plus de 50% d'entre eux abritent cette bactérie (Gordon & Cowling, 2003). La prévalence atteint même 90% chez l'Homme avec une concentration de bactéries retrouvées dans les selles estimée à environ 10^7 à 10^9 UFC par gramme de selles (Tenaillon et al., 2010). Derrière cette fréquence élevée de portage au niveau digestif se cache une importante hétérogénéité en termes de composition de la population. Tout d'abord, chez un individu donné, il existe des clones prédominants et d'autres sous-dominants (Escobar-Páramo et al., 2004). Cette diversité intra-hôte semble liée à différents facteurs, notamment socio-économiques tels l'alimentation et le niveau d'hygiène (Escobar-Páramo et al., 2004). On observe ainsi de fortes disparités à l'échelle mondiale, avec une diversité intra-hôte moins élevée dans des pays ou régions comme la France métropolitaine ou les Etats-Unis par rapport au Bénin ou à la Guyane française. De plus, la distribution des différents groupes phylogénétiques (ou phylogroupes) est également différente entre ces pays. Ces fluctuations ont aussi été mises en évidence sur les souches commensales fécales de français métropolitains expatriés en Guyane française : ceux-ci présentaient une composition intermédiaire, entre celle de la population autochtone et des métropolitains (Skurnik et al., 2008). Cette dynamique de la population est aussi parfois temporelle, comme le montre l'augmentation graduelle des souches de phylogroupe B2 entre trois collections de souches commensales françaises en 1980, 2000 et 2010 (Duriez et al., 2001; Escobar-Páramo et al., 2004; Massot et al., 2016). En outre, la pression antibiotique joue probablement un rôle dans ces remaniements comme en témoigne l'augmentation de la résistance chez les souches commensales au cours du temps (Massot et al., 2016).

Face à ces fluctuations multifactorielles de la population commensale de *E. coli*, il est complexe de définir le profil type d'une souche commensale. On a pour habitude de considérer le phylogroupe A, notamment le Séquence Type (ST) 10 (cf. paragraphe "L'amplification des acides nucléiques"), et parfois le phylogroupe B1 comme principaux pourvoyeurs de souches commensales, et leur absence de virulence intrinsèque en modèle animal valide en partie cette hypothèse (Picard et al., 1999). Pourtant, les rares études récemment entreprises sur des collections de vraies souches commensales dans des pays occidentaux mettent en

évidence une prédominance du phylogroupe B2 (Massot et al., 2016; Raimondi et al., 2019). Parmi ces souches commensales de phylogroupe B2, la plupart sont enrichies en facteurs de virulence, et seulement 5% sont de ST452, décrit comme étant strictement humain et commensal (Clermont et al., 2008; Massot et al., 2016). Mais la comparaison exhaustive de souches commensales et isolées de bactériémies provenant d'une population homogène en termes d'âge, de sexe, de localisation et d'époque pointe toute de même des différences (Clermont et al., 2017) . En effet, bien que l'on observe 34,1% de souches de phylogroupes B2 au sein des souches fécales commensales, elles représentent 64,5% des souches isolées de bactériémies à point de départ urinaire. Concernant les souches isolées de bactériémies à point de départ digestif, la population est comparable à la population fécale commensale (36,7% vs 35% de B2, respectivement) mais les souches de bactériémies présentent une résistance antibiotique accrue. Ces résultats reflètent en réalité les deux grands modèles dans la physiopathologie des bactériémies à *E. coli* : i) celui du pouvoir pathogène spécifique dans lequel des souches essentiellement de phylogroupes B2 et D, porteuses de nombreux facteurs de virulence extra-intestinale, vont pouvoir réaliser une infection par ascension du tractus urinaire, et ii) celui de la prévalence dans lequel des souches abondantes au niveau digestif vont transloquer dans le compartiment sanguin essentiellement à la faveur d'une défaillance de l'hôte sans nécessité de virulence particulière (Clermont et al., 2017; J. R. Johnson & Russo, 2018).

Par ailleurs, la présence de ces souches B2, et dans une moindre mesure D, en portage digestif pose la question des capacités colonisatrices de ces souches. En effet, cette fréquence élevée supporte la théorie selon laquelle la pathogénicité serait en réalité un produit dérivé du commensalisme (Le Gall et al., 2007; Levin & Edén, 1990), les infections invasives étant considérées comme des impasses évolutives puisqu'elles aboutissent généralement à une absence de transmission.

b. Souches pathogènes

E. coli se révèle aussi être un pathogène majeur chez l'Homme, responsable d'infections intestinales et extra-intestinales. On classe généralement ces souches pathogènes sous forme de pathovars ou pathotypes selon une nomenclature assez hétérogène se référant parfois à l'hôte, parfois aux manifestations cliniques ou encore à la présence de facteurs de virulence (Figure 3) (Denamur et al., 2021). On distingue classiquement les souches responsables d'infections intestinales à type de diarrhées, les InPEC ("Intestinal Pathogen *E. coli"*), et par analogie celles responsables d'infections extra-

intestinales, les ExPEC (“Extra-intestinal Pathogen *E. coli*”) (Thomas A. Russo & Johnson, 2000).

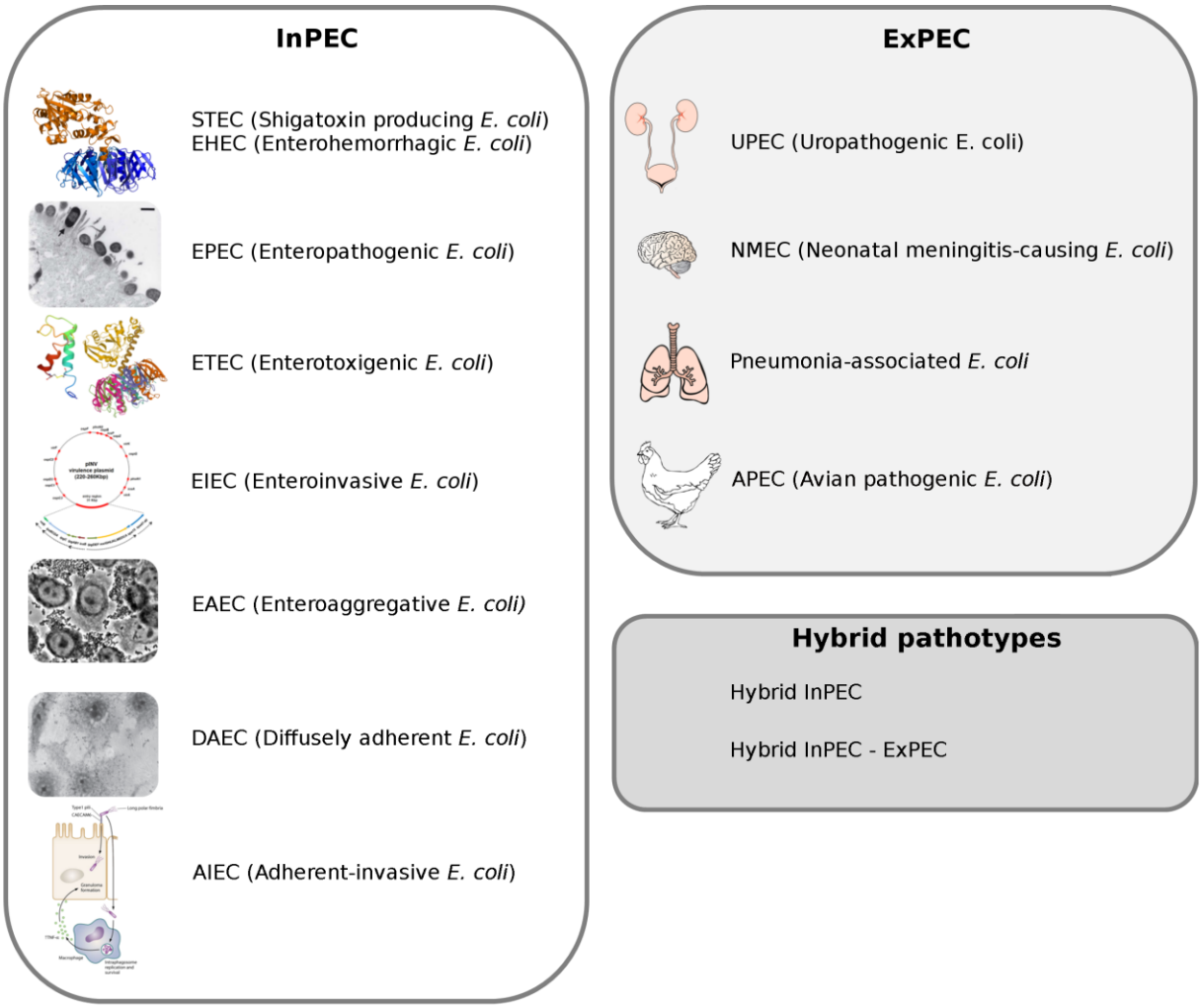


Figure 3. Principaux pathovars de *E. coli*. Cette classification hétérogène repose parfois sur l’organe cible, l’hôte, la présence d’un ou plusieurs facteur(s) de virulence ou encore un phénotype particulier. Photographies EPEC d’après Wales et al., 2005 ; EAEC d’après Prager et al., 2014 ; DAEC d’après Mansan-Almeida et al., 2013 ; illustration AIEC d’après Croxen et al., 2013 ; plasmide pINV d’après Pasqua et al., 2017 ; modélisation de la Shigatoxine et des entérotoxines d’après la banque de données PDP[2] (identifiants 1R4Q, 1EHS et 1TII).

Dans le cas des InPEC, le caractère pathogène est en général associé à un ou quelques facteurs de virulence bien identifiés (caractère pauci-génique) et dont le rôle dans la physiopathologie de l’infection est clairement démontré (Croxen et al., 2013). C’est le cas par

[2] https://www.rcsb.org/

exemple des STEC (“Shigatoxin producing *E. coli*”) et EHEC (“Enterohemorrhagic *E. coli*”) producteurs de Shiga-toxine ou bien encore des EIEC (“Enteroinvasive *E. coli*”) qui possèdent les déterminants de virulence de *Shigella* tel le plasmide de virulence pINV. Ces pathotypes intestinaux ne seront pas abordés en détail dans ce manuscrit. Le cas de *Shigella* ne sera pas non plus traité. Bien que partie intégrante de l’espèce *E. coli,* les spécificités liées notamment à son mode de vie parasitaire en font un sujet à part.

Parmi les ExPEC, il existe des souches responsables d’infections urinaires aussi appelées UPEC (“Uropathogenic *E. coli*”), des souches responsables de méningites, les NMEC (“Neonatal-meningitis-causing *E. coli*”), ou encore, plus récemment décrites, des souches impliquées dans des colonisations et/ou infections pulmonaires notamment au cours de pneumopathies acquises sous ventilation mécanique (Denamur et al., 2021; La Combe et al., 2019). Il est parfois proposé le terme SEPEC (“Sepsis-associated *E. coli*), néanmoins il est plus probable que l’isolement de ces souches dans le sang soit la conséquence de l’extension de l’infection à partir d’un foyer primaire, dû par exemple à une souche UPEC, dans le cas d’une pyélonéphrite bactériémiante. Enfin, on trouve également dans ce groupe des pathogènes aviaires, les APEC (“Avian pathogenic *E. coli*), responsables de la colibacillose aviaire (Guabiraba & Schouler, 2015). La définition de ces ExPEC est plus complexe que dans le cas des InPEC et fait l’objet d’un chapitre particulier.

Enfin, il existe des souches dites hybrides possédant simultanément des caractères associés à différents pathovars. L’une d’entre elles a été particulièrement médiatisée puisque responsable en 2011 d’une épidémie européenne de cas de syndrome hémolytique et urémique chez l’adulte (Frank et al., 2011). La souche responsable, de sérotype O104:H4, provenait probablement de graines germées, et était en réalité un hybride entre deux souches InPEC : une souche EHEC et une souche EAEC. On peut même considérer qu’il s’agit d’un triple hybride puisque la souche possède également des facteurs de virulence généralement observés chez les ExPEC comme la yersiniabactine et l’aérobactine (Mariani-Kurkdjian & Bingen, 2012). Un autre hybride a été décrit récemment et semble émerger notamment en France : il s’agit d’une souche présentant des caractères EHEC, comme la production de la shigatoxine, et un plasmide de virulence ExPEC (Mariani-Kurkdjian et al., 2014; Soysal et al., 2016). Cliniquement, cette souche de sérotype O80:H2 est responsable d’un syndrome hémolytique et urémique associé à une bactériémie.

III. *Escherichia coli* : méthodes d'analyses et structure de la population

1. L'étude des sérotypes

Les connaissances autour de la structure de la population de *E. coli* se sont affinées au cours du temps, tirant profit des évolutions technologiques. Initialement les études ont principalement pris en compte les caractères phénotypiques des souches comme les critères culturaux. Les études reposant sur les données de sérotypage ont ensuite permis de mieux appréhender la population de *E. coli*. En 1947, Kauffmann propose une revue des connaissances sur les méthodes de sérologie de cette espèce incluant les méthodes d'agglutination des antigènes O somatique, H flagellaire et K capsulaire. Dès cette époque apparaît un déséquilibre en termes de distribution de ces déterminants dans la population bactérienne (Kauffmann, 1947). Certains, en effet, sont plus représentés que d'autres, et cette distribution semblent parfois liée à l'origine des souches (*i.e.* fèces *versus* appendicites/péritonites/infections urinaires). Le pouvoir hémolytique et nécrotique de certaines souches est également associé à un nombre restreint de sérotypes O et les souches porteuses d'un antigène K sont plus toxiques et résistent mieux au système immunitaire et aux bactériophages. Par ailleurs, plusieurs isolats de *E. coli* provenant d'un même type de prélèvement pathologique présentent en général le même O-type alors que pour les isolats provenant d'un même échantillon de fèces il est parfois possible d'observer différents sérotypes. De ces données Kaufmann conclut qu'il existe des sérotypes particulièrement pathogènes. Néanmoins, il note déjà qu'il s'agit seulement d'une tendance et que cette pathogénicité n'est pas obligatoire, les souches présentant de tels sérotypes étant parfois isolées des fèces de patients sains.

Quelques années plus tard, Orskov & Orskov confirment ces données à partir d'une collection de 539 souches isolées d'hémocultures au Danemark (Orskov & Orskov, 1975). A nouveau, ils observent un nombre limité de sérotypes : avec seulement 10 sérums anti-O et 10 sérums anti-H ils parviennent à agglutiner 68% des souches. Certaines combinaisons sont particulièrement représentées comme le O4:H5 (qui correspond certainement à des souches de phylogroupe B2 et de ST12). Ils comparent également ces données avec celles de souches fécales. Bien que la fréquence des différents sérotypes varie entre ces deux origines, le classement par ordre de fréquences n'est lui globalement pas modifié. Ainsi, les auteurs concluent que les sérotypes de souches isolées d'infections extra-intestinales sont le reflet du

microbiote intestinal et un nombre restreint de sérotypes est probablement sélectionné au niveau digestif pour envahir ensuite la circulation sanguine.

Ces déséquilibres dans la distribution des combinaisons O:H ne sont pas uniquement observées entre souches pathogènes et commensales mais également en fonction de l'espèce hôte (*e.g.* humain vs mammifères non-humains) (Bettelheim, 1978).

2. Le profil électrophorétique des enzymes

Par la suite la comparaison des profils électrophorétiques d'un nombre croissant d'enzymes du métabolisme va être réalisée, avec la méthode de Multi-Locus Enzyme Electrophoresis (MLEE) (Herzer et al., 1990; Milkman, 1973; Selander et al., 1986). A l'aide de cette technique, il est possible de prendre en compte les différences génotypiques résultant de mutations non synonymes dans les gènes codant les enzymes et non plus uniquement les différences phénotypiques. Cela va considérablement modifier la vision de la population de *E. coli*. Certaines souches ayant l'antigène K1 et appartenant au même sérogroupe O n'apparaissent en réalité pas plus proches entre elles qu'elles ne le sont de n'importe quelle autre souche (Ochman & Selander, 1984b). Les combinaisons de sérotypes prises isolément sont donc de mauvais témoins de la clonalité des souches. En revanche, ici encore avec le MLEE un nombre limité de combinaisons d'allèles est identifié dans la population de *E. coli* mettant à nouveau en évidence un déséquilibre de liaison (Whittam et al., 1983) .

Dans leur revue en 1987, Selander *et al.* offrent une vue détaillée de la structure de la population de *E. coli* obtenue par ces techniques de MLEE (Selander et al., 1987). A l'aide de ces approches, il est désormais possible d'évaluer des distances génétiques et d'obtenir des dendrogrammes afin d'illustrer les relations génétiques entre toutes les souches étudiées. Grâce à ces analyses, les auteurs identifient ce qui constituera par la suite la base de la classification en phylogroupes, et mettent en évidence l'association de certains de ces phylogroupes avec un hôte particulier : B2 et humain ou B1 et mammifères non primates. Par ailleurs, ces phylogroupes présentent parfois des capacités métaboliques particulières comme par exemple la fermentation de certains sucres : raffinose et phylogroupes B1, C et D ; sorbose et phylogroupes B2 et D. En combinant le MLEE avec les données de sérotypage, les auteurs mettent également en lumière des évènements de recombinaison, pourtant considérés jusqu'alors comme rares. Dès lors va se poser la question de l'impact de ces recombinaisons sur la structure de la population de *E. coli*. Selander *et al.* considèrent que

même avec un taux de recombinaison élevé, les déséquilibres de liaison observés peuvent probablement persister en raison de la forte sélection sur certaines combinaisons d'allèles. De la même manière qu'avec les études sérotypiques, ils notent aussi une diversité génétique réduite des souches isolées de patients souffrant de cystites et de pyélonéphrites comparées à des souches commensales fécales ou isolées de bactériuries asymptomatiques. Mieux encore, en comparant des souches obtenues à partir de ces quatre types d'échantillon, provenant de Suède et de Finlande, ils montrent que les souches de pyélonéphrite de ces deux pays sont plus proches entre elles qu'elles ne sont des autres souches de leur pays respectif (Figure 4). Ces données plaident déjà pour l'existence d'un nombre restreint de clones pathogène spécialisés dans les infections extra-intestinales.

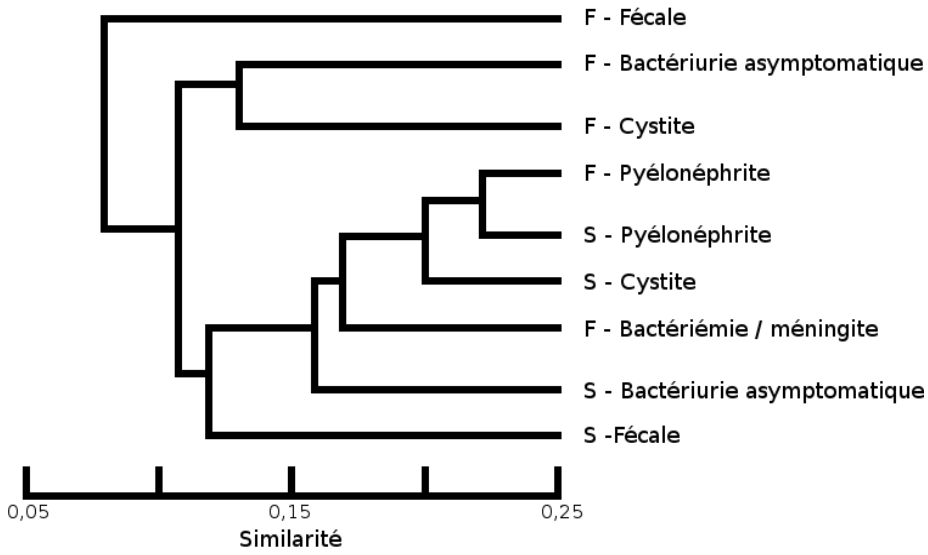


Figure 4**.** Dendrogramme construit à partir des profils électrophorétiques de souches isolées en Finlande (F) et en Suède (S), à partir des fèces, des urines au cours de bactériuries asymptomatiques, cystites et pyélonéphrites et de bactériémies/méningites. L'index de similarité $S = N_{AB} / (N_A + N_B - N_{AB})$ où N_{AB} est le nombre d'électrophorétypes partagés entre 2 échantillons ayant N_A et N_B électrophorétypes, respectivement (d'après Selander et al., 1987).

D'autres équipes ont utilisé ces méthodes d'électrophorèse des enzymes pour étudier la population de *E. coli*. Ainsi, Goullet & Picard vont analyser les profils électrophorétiques des carboxyestérases et notamment la carboxylesterase B. Ils parviennent à séparer la population en deux grands groupes : B_1 et B_2, ce dernier correspondant par un heureux hasard au phylogroupe B2 de Selander et *al.* (Goullet & Picard, 1986). Ils observent une association préférentielle entre les carboxylestérases B de faible mobilité électrophorétique (B_2) et les

souches isolées d'infections extra-intestinales (Figure 5). C'est également dans ce groupe B_2 qu'ils retrouvent le plus d'α-hémolysine et d'hémagglutinines résistantes au mannose. A nouveau, ils soulignent le caractère non exclusif de ces associations puisque certaines souches étiquetées "pathogènes" présentent des profils B_1 et certaines commensales des profils B_2.

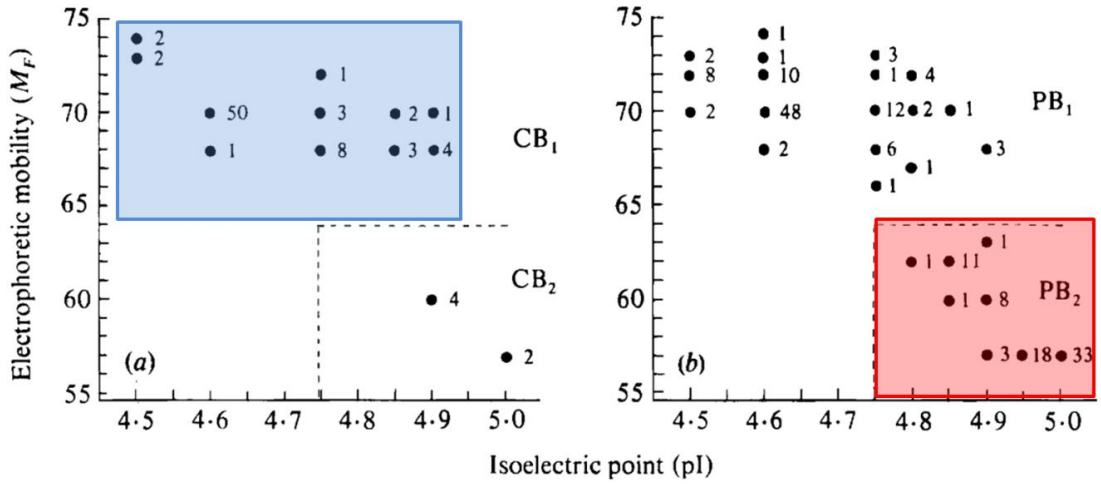


Figure 5. Profils électrophorétiques bi-dimensionnels de carboxylestérases B de souches de *E. coli* commensales et pathogènes extra-intestinales. CB_1 = souches commensales et profil B_1 ; CB_2 = souches commensales et profil B_2 ; PB_1 = souches pathogènes et profil B_1 ; PB_2 = souches pathogènes et profil B_2. En bleu, le profil électrophorétique B_1 regroupe la plupart des commensales. En rouge, le profil électrophorétique B_2 regroupe plus de pathogènes que de commensales. (d'après Goullet & Picard, 1986).

3. L'amplification des acides nucléiques

Le développement des méthodes d'amplification des acides nucléiques va offrir encore un peu plus de puissance à ces analyses populationnelles. En combinant des méthodes de ribotypage (polymorphisme de longueur des fragments de restriction des gènes *rrn* ou "*rrn* RFLP"), d'amplification aléatoire d'ADN polymorphe (RAPD) et les données de MLEE, Desjardins *et al.* retrouvent la classification en 4 grands groupes A, B1, B2 et D (Desjardins et al., 1995). Ces analyses soulignent à nouveau la structure clonale de la population de *E. coli*, particulièrement au sein du phylogroupe B2, et les auteurs concluent que le sexe chez *E. coli* (*i.e.* les recombinaisons) ne perturbent pas cette clonalité. Puis les analyses phylogénétiques réalisées à partir d'un nombre limité de séquences chromosomiques et plasmidiques, retrouvées à la fois chez *E. coli* et *Salmonella*, vont préciser

la position des différents phylogroupes dans la phylogénie : le phylogroupe B2 semble basal, puis vient le D, et enfin les deux phylogroupes A et B1 apparaissent comme des groupes frères (Lecointre et al., 1998).

La démocratisation du séquençage d'ADN par méthode Sanger a lui aussi permis de passer un cap dans la compréhension de la population, notamment avec le Multi-Locus Sequence Typing (MLST) (Maiden et al., 1998). Initialement proposée pour *Neisseria meningitidis,* cette méthode consiste à amplifier un nombre restreint de gènes de ménage (généralement entre 7 et 11 gènes), les séquencer et leur attribuer à chacun un numéro d'allèle à partir d'une base de données stable, commune et accessible dans le monde entier via Internet. La combinaison de ces allèles correspond à un Sequence Type (ST). Pour plus de sens biologique, les ST proches sont parfois regroupés sous la forme de ST complexe (STc) lorsqu'ils présentent moins de 3 allèles de différence, ou en Complexe Clonaux (CC) s'ils diffèrent par un seul allèle (Jauréguy et al., 2008; Wirth et al., 2006). Les souches de *E. coli* au sein de ces complexes partagent souvent une même niche écologique et des traits évolutifs communs (Alm et al., 2014). En reprenant le principe du MLEE mais sans les biais liés à la redondance du code génétique et au caractère synonyme de certaines mutations, le MLST permet à la fois d'avoir une vision globale de la population et une vision locale avec l'identification précise de certains clones. Par ailleurs, en codant l'ensemble des séquences par des numéros d'allèles, le poids des éventuelles recombinaisons dans l'évaluation de la diversité génétique tend à être minimisé. Des schéma MLST[3] sont aujourd'hui disponibles pour 127 organismes, parmi lesquels *E. coli.* Il en existe même trois pour cette espèce, deux adaptés à l'ensemble des souches (Jauréguy et al., 2008; Wirth et al., 2006) et un plus spécialisé et développé pour les souches entéropathogènes (Reid et al., 2000). En utilisant une combinaison différente de gènes de ménage, ces schémas offrent des pouvoirs discriminants variables selon les clones (Clermont et al., 2015). Tout comme avec le MLEE, l'analyse des ST chez *E. coli* fait apparaître une association entre certains ST et le caractère pathogène des souches (*e.g.* STc95 et les souches porteuses de l'antigène capsulaire K1, ST11 et les souches EHEC O157:H7). Mais les pathotypes ne sont pas nécessairement associés à un seul STc. Au contraire, certains sont retrouvés au sein de STc non reliés montrant des phénomènes d'acquisition indépendante et répétée de gènes de virulence dans la population de *E. coli* (Wirth et al., 2006). Dans leur analyse de souches isolées de bactériémies, Jauréguy *et al.* montrent également que parmi les différents phylogroupes certains semblent plus structurés que d'autres (Jauréguy et al., 2008). C'est notamment le cas du phylogroupe B2 au sein

[3] https://pubmlst.org/

duquel les différents complexes clonaux identifiés par MLST apparaissent clairement délimités (Figure 6).

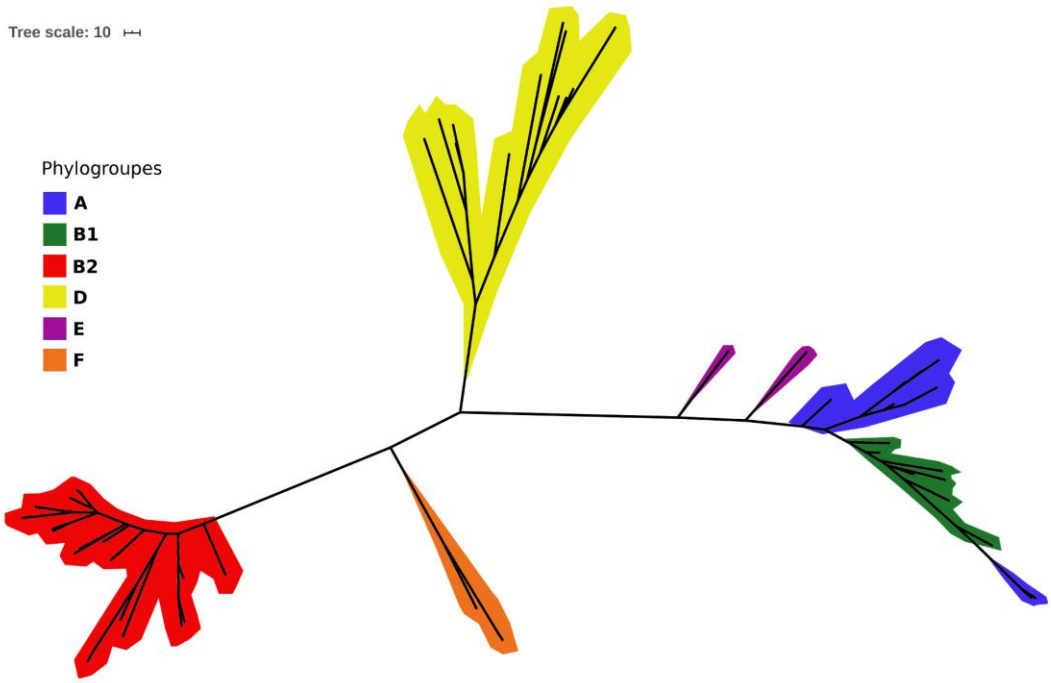


Figure 6. Arbre phylogénétique non enraciné, obtenu à partir des séquences des 8 gènes du MLST (schéma Pasteur), après élimination des recombinaisons. Les phylogroupes alors connus apparaissent en couleur (d'après les données de Jauréguy et al., 2008).

Le pouvoir discriminant à la fois à l'échelle d'une population et des éventuels sous-groupes d'une population ainsi que la portabilité du MLST en ont fait une méthode encore largement utilisée et ce même malgré l'avènement du séquençage de génomes complets. Il permet en effet aux chercheurs et épidémiologistes du monde entier de parler une même langue et d'identifier, par exemple, des clones pandémiques mondiaux ou encore de suivre l'évolution de la population au cours du temps (Kallonen et al., 2017; Nicolas-Chanoine et al., 2014). Bien qu'il existe une certaine hétérogénéité en termes de diversité entre les différents ST, les distances génétiques entre souches à l'intérieur d'un ST chez *E. coli* sont environ 10 fois inférieures aux distances inter-ST (Touchon et al., 2020). Après alignement et concaténation des gènes du MLST, il est possible de réaliser des analyses phylogénétiques et mettre ainsi en lumière l'évolution de la structure de la population. Cela a par ailleurs permis l'identification de souches légèrement divergentes de l'espèce *E. coli*, mais phénotypiquement indistinctes. Regroupées sous le terme "*Escherichia* clade", cinq sous-populations ont été décrites, numérotées de I à V. Parmi elles, le clade I apparaît très proche de l'espèce *E. coli sensu stricto* alors que les autres sont plus éloignées et forment un continuum entre *E. coli* et

l'espèce *E. albertii* (Clermont et al., 2011; Luo et al., 2011; Walk et al., 2009). L'isolement de ces clades est plus fréquent chez les animaux non-humains, principalement les oiseaux, et dans l'environnement notamment aquatique. Ces souches sont rarement retrouvées au cours d'infections extra-intestinales chez l'Homme, en accord avec leur absence de virulence dans un modèle murin d'infection extra-intestinale (Clermont et al., 2011; Ingle et al., 2011).

Parallèlement, des méthodes de typage rapide par PCR amplifiant un nombre limité de cibles ont été développées pour classer les souches au sein des grands phylogroupes. En effet, bien qu'ayant un pouvoir résolutif relativement élevé, le MLST nécessite l'amplification et le séquençage d'un nombre trop élevé de cibles pour en faire une méthode adaptée à de larges jeux de données. De plus, la distribution non aléatoire des souches au sein des phylogroupes selon leur origine (*i.e.* commensale *versus* extra-intestinale) plaide en faveur du développement de telles méthodes (Picard et al., 1999). Ces PCR multiplex ont d'abord reposé sur l'amplification de 3 cibles, puis ont inclus la détection des *Escherichia* clades, suivie d'une mise à jour afin de distinguer les 8 phylogroupes de *E. coli* alors reconnus (A, B1, B2, C, D, E, F pour *E. coli sensu stricto* et le clade I comme 8ème groupe) et les 4 autres *Escherichia* clades (II à V) (Clermont et al., 2000, 2011, 2013).

Sur le même principe Clermont *et al.* ont proposé une PCR pour assigner les souches de phylogroupe B2 à huit des neuf différents sous-groupes décrits, motivé par l'isolement fréquent de ces souches dans les infections extra-intestinales (Tableau 4) (Clermont et al., 2014; Le Gall et al., 2007). Le sous-groupe VIII (STc452) fait figure d'exception puisqu'il n'abrite que des souches commensales, dont la souche ED1a, pouvant faire penser que le fond génétique du sous-groupe B2 est associé au pouvoir commensal (Clermont et al., 2008). Par ailleurs, la diversité génétique au sein de ce phylogroupe très structuré est accrue comparée aux autres phylogroupes : jusqu'à 2% de divergence est observée entre les souches des différents sous-groupes alors que cette diversité n'excède pas 3% pour l'ensemble de l'espèce *E. coli* sensu stricto. Cette diversité pourrait être le reflet d'une adaptation des sous-groupes à une niche particulière ou bien très probablement de l'émergence rapide de clones épidémiques. Cette dernière hypothèse semble notamment corroborée par la fréquence des différents ST dans les infections extra-intestinales : trois des quatre ST les plus fréquents à l'heure actuelle dans les ExPEC appartiennent au phylogroupe B2 (Denamur et al., 2021).

Tableau 4. Principaux sous-groupes au sein de phylogroupe B2 (d'après Clermont et al., 2014).

Sous-groupe	STc	Souches type
I	131	-NA114, EC7372 (O25b:H4) -SE15 (H5:O16)
II	73	CFT073, ECOR57
III	127	536, F11
IV	141	IAI74, H223
V	144	S107, H176
VI	12	J96, ECOR60
VII	14	C1845, ECOR64
VIII	452	ED1a
IX	95	UTI89, S88, RS218
X	372	A034/86, IAI64

Enfin, ces méthodes de typage ont récemment été adaptées pour l'analyse à partir de génomes complets en combinant i) une PCR *in silico* semblable aux méthodes traditionnellement utilisées dans un souci de rétro-compatibilité ; et ii) une approximation des distances génétiques à l'aide du programme Mash et de génomes de références couvrant la diversité du genre *Escherichia* (Beghain et al., 2018; Ondov et al., 2016). Cette approche permet de rattacher les souches aux 8 phylogroupes reconnus : A, B1, B2 et D qui représentent la majorité des souches, et C, E, F et G plus rares. L'adaptation des méthodes d'analyse aux évolutions technologiques est effectivement devenue essentielle, tant le nombre de génomes séquencés a explosé, et ce dès la fin des années 2000. Ce bond technologique a ouvert la porte aux analyses comparatives d'un nombre toujours croissant de souches et notamment l'utilisation d'approches pangénomiques.

4. L'apport de la génomique

Une des premières études de pangénome est proposée par Tettelin *et al.* et porte sur *Streptococcus agalactiae* (Tettelin et al., 2005). Elle a participé à poser les bases de la pangénomique et notamment certaines définitions : le pangénome, composé de l'ensemble des gènes observés au sein des génomes analysés, lui-même composé du coregénome,

l'ensemble des gènes conservés, et du génome "dispensable" aussi appelé génome variable ou accessoire, c'est-à-dire les gènes présents dans un sous-ensemble de génomes voire un seul génome. Cette segmentation a par la suite été affinée et est souvent présentée sous la forme de trois entités : le coregénome (ou le "persistent" pour une définition moins stricte de conservation), le "shell" représentant les familles de gènes orthologues modérément présentes, et le "cloud" représentant les familles de gènes orthologues plus rares et présentes dans un nombre limité de génomes (Collins & Higgs, 2012; Koonin & Wolf, 2008).

Dans les années qui suivent la publication de la première séquence complète de *E. coli* K-12 (Blattner et al., 1997), les études vont tenter de décrypter la complexité de l'espèce *E. coli* par ces approches d'analyse de pangénomes, incluant un nombre toujours croissant de génomes. Grâce à ces analyses, il est aujourd'hui établi que le génome de *E. coli*, à l'exclusion de *Shigella*, a une taille d'environ 5 Mb, pouvant varier entre 4,2 et 6,0 Mb, contient environ 4600 gènes et présente une composition moyenne en GC de 50,6%. Les variations de tailles sont notamment associées aux phylogroupes (Figure 7), comme le suggéraient déjà Bergthorsson & Ochman à l'aide d'enzymes de restriction (Bergthorsson & Ochman, 1998). Les analyses de génomes complets confirment ces données avec des génomes de taille réduite pour les souches des groupes A et B1, comparées aux autres phylogroupes, comme le B2 ou le D (Touchon et al., 2009, 2020). Par ailleurs, la source d'isolement semble également liée à la taille des génomes, comme par exemple la taille réduite observée pour les souches isolées de l'eau (Touchon et al., 2020).

Concernant le pangénome, la taille des différentes partitions est difficile à définir de manière univoque car elle dépend du nombre de génomes analysés, de la diversité et la qualité de ces génomes, et enfin de la méthode utilisée pour définir ces partitions. Sur quatre études récentes analysant plus de 1000 souches, par exemple, la valeur du coregénome pour les gènes présents dans 99% des souches oscille entre 1744 et 2663 familles de gènes (Abram et al., 2019; Kallonen et al., 2017; Park et al., 2019; Touchon et al., 2020). Pour pallier à la rigidité de la définition du coregénome et à la variabilité qui en découle, une étude récente propose une approche différente basée sur une méthode statistique prenant en compte à la fois les présences/absences des familles de gènes et leur information de co-localisation génomique afin de les classer dans le "persistent", le "shell" et le "cloud" sans avoir besoin de définir de seuils de fréquence de présence des familles. Cette approche, plus conservative, met en évidence un génome "persistent" composé de 3710 familles de gènes pour un ensemble de 15141 génomes de *E. coli*, et permet d'éviter les biais liés notamment aux artéfacts techniques (Gautreau et al., 2020). Quoiqu'il en soit, il ressort de toutes ces études qu'une partie importante du génome des souches n'est pas inclue dans ce génome

“persistent”, offrant une grande diversité et plasticité génétique aux différentes souches. Cette partie variable (“shell” et “cloud”) est sur-représentée dans le pangénome de *E. coli* et tend à augmenter à mesure que de nouvelles souches sont ajoutées mettant en évidence un pangénome ouvert pour cette espèce. *In fine*, cela aboutit à des estimations très variables en ce qui concerne la taille des pangénomes, allant de 69274 à parfois plus de 135000 familles de gènes (Abram et al., 2019; Kallonen et al., 2017; Park et al., 2019; Touchon et al., 2020). Le contenu en gènes est lui aussi variable en fonction du phylogroupe avec un phylogroupe B2 se distinguant des autres et une proximité des phylogroupes A et B1, tout comme les phylogroupes D et F.

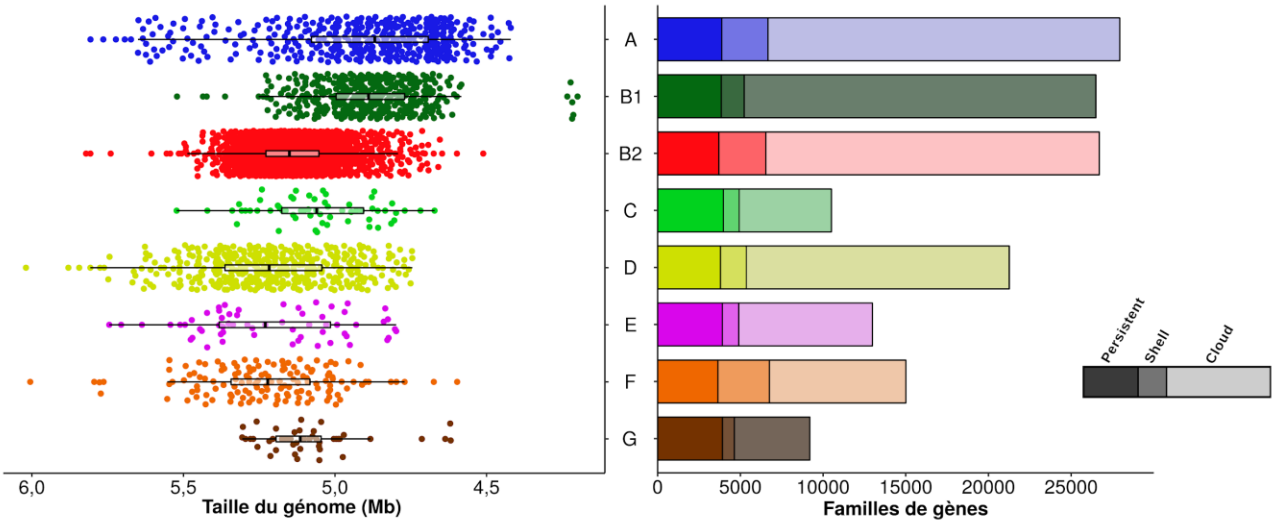


Figure 7**.** Taille des génomes et des pangénomes des 8 groupes de *E. coli* sensu stricto. Construit à partir des données de Kallonen et al., 2017; Touchon et al., 2020.

L’étude des relations entre taille des génomes et pangénome des différents phylogroupes offre des résultats parfois peu intuitifs. Les phylogroupes A et B1, bien que composés de génomes de plus petite taille, possèdent des pangénomes plus importants, témoignant d’une plus grande diversité génétique. De plus, au sein de ces groupes, la fraction de gènes rattachés au “persistent” pour une souche donnée est plus faible, probablement lié à un taux élevé de perte et gain de gènes. Si l’on se concentre sur les éléments génétiques mobiles pour ces groupes A et B1, ce paradoxe est retrouvé avec une fréquence plus faible de gènes associés à des éléments mobiles mais en revanche une plus grande diversité (Touchon et al., 2020).

Cette diversité en termes de contenu en gènes pose la question de l’organisation du génome de *E. coli* et de la capacité à tirer des informations génétiques pertinentes pour la réalisation

d’analyses phylogénétiques. Touchon *et al.* ont observé, que malgré un flux de gènes importants chez *E. coli*, ceux-ci se retrouvent organisés au sein des génomes (Touchon et al., 2009). Cela est rendu possible par le confinement spatial de ce flux de gènes. En effet, si les insertions et délétions de petite taille ont lieu en de nombreux endroits du génome, les événements impliquant un nombre important de gènes (>10 gènes) sont restreints à certains points chauds d’intégration (Touchon et al., 2009; Welch et al., 2002). Parmi ces points chauds, on trouve certains ARN de transfert, au niveau desquels vont s’intégrer des îlots génomiques. Ces îlots ont une structure modulaire avec des motifs de présence/absence non liés à la phylogénie, suggérant de multiples intégrations et/ou des recombinaisons fréquentes. Il existe également deux points chauds de recombinaison : le groupe de gènes *rfb,* impliqué dans la biosynthèse de l’antigène O, et l’opéron *fim* codant le pili de type I. Les éléments mobiles jouent un rôle essentiel dans la diversité génétique de *E. coli* et les échanges de matériel génétique, notamment les prophages et les plasmides qui sont retrouvés dans la quasi-totalité des génomes. A eux deux, ces éléments représentent le tiers du pangenome de chaque phylogroupe (Touchon et al., 2020). Cependant, la fréquence élevée des recombinaisons n’altère pas le signal phylogénétique si la taille des séquences utilisées pour l’analyse est suffisante.

IV. Les souches pathogènes extra-intestinales

1. Comment définir un ExPEC?

Bien que l'acronyme ExPEC soit utilisé par homologie à celui des InPEC, les contours sont en réalité plus complexes à délimiter. De prime abord, la définition de cette entité peut paraître triviale et on pourrait penser qu'il s'agit simplement de souches isolées d'infections extra-intestinales par opposition aux souches commensales d'origine fécale. Mais comme le montrent Johnson *et al.* dans un modèle murin, l'origine écologique (*i.e.* fécale ou clinique) ne reflète pas nécessairement le pouvoir pathogène des souches (J. R. Johnson et al., 2018). Certaines souches peuvent en effet être isolées au cours de bactériémies sans pour autant présenter une virulence accrue dans un modèle animal, témoignant plutôt d'une sensibilité particulière de l'hôte (Figure 8) (Lefort et al., 2011; Tourret & Denamur, 2016). Inversement, certaines souches présentes dans le microbiote digestif de patients sains peuvent se révéler être d'authentiques ExPEC si l'on se fie à leur virulence chez l'animal et/ou aux déterminants génétiques qu'ils portent (Starčič Erjavec & Žgur-Bertok, 2015). Il est d'ailleurs communément admis que la première étape de l'infection extra-intestinale est la colonisation du microbiote digestif, et il est donc logique de retrouver des ExPEC comme colonisant au niveau intestinal. Cela supporte l'hypothèse précédemment évoquée de la virulence comme produit dérivé du commensalisme, faisant des souches pathogènes avant tout de bons commensaux (Le Gall et al., 2007; Levin & Edén, 1990).

Dans leur définition des ExPEC en 2000, Russo et Johnson proposent d'inclure les souches cliniques possédant des facteurs de virulence reconnus et/ou démontrant une virulence accrue dans un modèle animal d'infection extra-intestinale (Thomas A. Russo & Johnson, 2000). La qualification d'ExPEC relève donc d'un faisceau d'arguments épidémiologiques, moléculaires et phénotypiques (Denamur et al., 2021). Si l'on se réfère aux deux modèles de physiopathologie des infections extra-intestinales cités précédemment, les ExPEC correspondent donc à des souches douées d'un pouvoir pathogène particulier. Cependant, le possible portage de ces souches au niveau digestif ne permet pas de les exclure complètement du modèle de prévalence (Clermont et al., 2017; J. R. Johnson & Russo, 2018; Manges et al., 2019; Massot et al., 2016). La Figure 8 illustre, dans les cas des infections du tractus urinaire, le caractère multifactoriel des infections extra-intestinales, impliquant à la fois l'hôte et le fond génétique des souches. Les souches à considérer comme ExPEC "vraie" doivent être capables de réaliser une infection chez un hôte non débilité, grâce à un fond

génétique (phylogroupe, virulence) particulier dont l'impact est démontré sur l'animal. En présence de facteurs de comorbidités importants chez l'hôte, la virulence intrinsèque de la souche ne semble plus être un prérequis, permettant aux souches de réaliser l'infection peu importe leur statut ExPEC ou non.

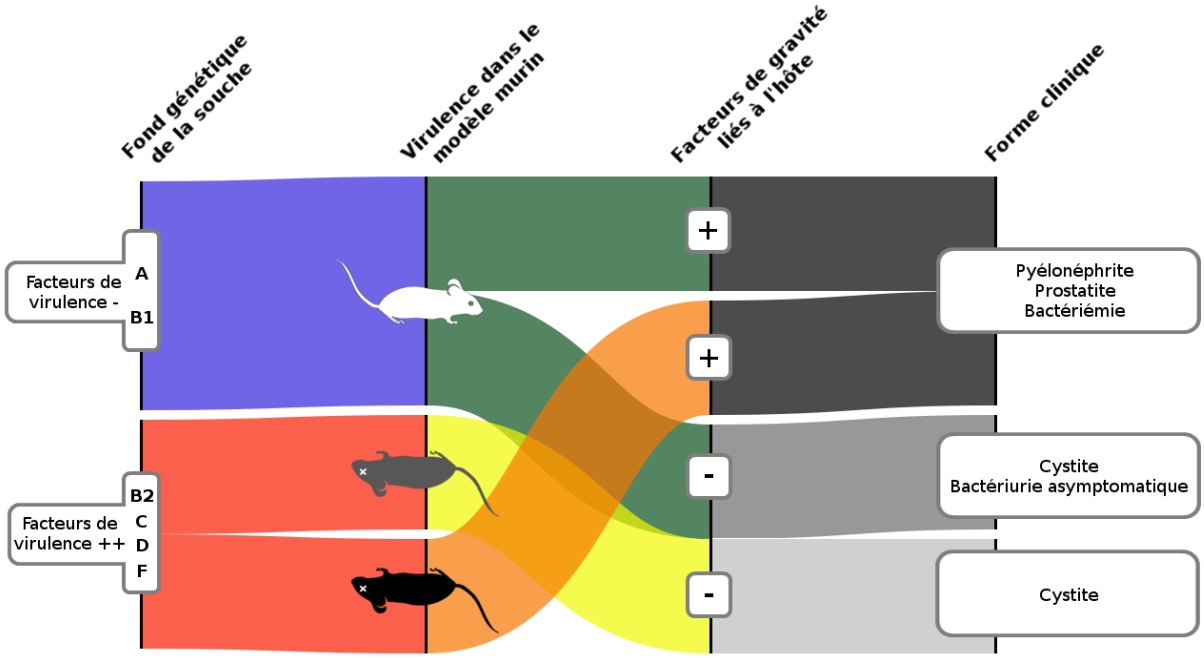


Figure 8**.** Représentation graphique des interactions entre le fond génétique des souches, leur virulence dans un modèle animal, l'état immunitaire de l'hôte et les formes cliniques observées au cours des infections urinaires. La souris de couleur blanche correspond à une absence de pathogénicité dans un modèle animal d'infection extra-intestinale, alors que les couleurs grises et noires correspondent à une virulence intermédiaire et élevée, respectivement (partiellement adaptée de Tourret & Denamur, 2016).

Les formes cliniques d'infections extra-intestinales dues à *E. coli* sont variées et incluent des infections urinaires, des méningites plus volontiers néonatales, des infections intra-abdominales, parfois qualifiées de "paraintestinales", des infections pulmonaires ou bien encore des infections de la peau et des tissu mous. Ces infections sont par ailleurs autant de foyers primaires pour d'éventuelles bactériémies (Thomas A. Russo & Johnson, 2000).

2. ExPEC et facteurs de virulence

Pour être ExPEC une souche doit donc posséder des facteurs de virulence lui permettant de coloniser les muqueuses, de contourner les défenses de l'hôte, de capter des nutriments essentiels comme le fer, d'envahir la cellule hôte et de provoquer une réaction inflammatoire (J. R. Johnson & Russo, 2002). La classification fonctionnelle de ces facteurs de virulence comprend quatre à cinq classes communément admises : les adhésines, les toxines, les systèmes de capture du fer, les polysaccharides de surface et les protectines, les invasines (Dale & Woodford, 2015; J. R. Johnson & Russo, 2018). Une classe dite "miscellanée" est parfois également proposée. Cette dernière ne sera pas détaillée car la fonction des facteurs qui la compose est bien souvent peu ou mal décrite. Parfois, il peut également s'agir de marqueurs de régions associées à la virulence, comme le gène *malX* marqueur d'un îlot de pathogénicité (Östblom et al., 2011; Thomas A. Russo & Johnson, 2000).

Etant donné la diversité et le nombre de facteurs identifiés, sur des données épidémiologiques et/ou expérimentales, il serait peu digeste de tous les détailler ici. Nous n'aborderons donc que les principaux dans chaque classe. Parmi ces déterminants, beaucoup sont associés à la virulence parce que fréquemment retrouvés dans des études épidémiologiques (J. R. Johnson & Russo, 2018). Mais les preuves de leur contribution réelle à la virulence reste parfois faible, de sorte que les postulats moléculaires de Koch ne sont pas toujours vérifiés (Falkow, 2004). Pour chacun des facteurs, je proposerai une estimation de sa prévalence au sein des souches responsables de bactériémies de l'adulte à partir d'une sélection d'études publiées au cours des 10 dernières années (Tableau 5). Les études portant sur des souches volontairement sélectionnées pour leur résistance aux antibiotiques sont exclues, car ces souches sont souvent porteuses d'un plus grand nombre de facteurs de virulence. Par ailleurs, les différentes méthodes utilisées (PCR ou séquençage complet), ainsi que les disparités en termes de population de patients peuvent en partie expliquer la variabilité. Des données obtenues de souches commensales, bien que rares, sont également proposées comme "contrôle".

Tableau 5. Distribution des principaux facteurs virulence ExPEC au sein des souches isolées de bactériémies de l'adulte et des souches commensales fécales.

		Souches isolées de bactériémies									Souches commensales fécales		
		Lefort et al., 2011	Bert et al., 2011	Skjøt-Rasmussen et al., 2012	Williamson et al., 2013	Mora-Rillo et al., 2015	Salipante et al., 2015	Kallonen et al., 2017	Daga et al., 2019	Tao et al., 2020	Bok et al., 2018	Raimondi et al., 2019	Massot et al., 2016
	Pays	France	France	Danemark	Nouvelle Zélande	Espagne	Etats-Unis	Royaume-Uni	Brésil	Chine	Pologne	Italie	France
	Nombre de souches	1051	57	196	101	120	92[d]	1509[d]	58	188	296	51	276
	Population de patients bactériémiques	Tout venant	Transplantés hépatiques	Urosepsis	Urosepsis et post-biopsie de prostate	Tout venant	Tout venant	Tout venant	Tout venant	Patients immunodéprimés	-	-	-
Adhésines	***papGII***	38,6	25			30	35,9	48,6[b]		1,1			11,5
	papGIII	10,0	0			10	14						7,5
	autres gènes *pap*	50,0	37	71	45,6[a]		32 à 46	46,9 à 69,8	35,4		18,2	33,3	23,3
	afa/dr			3		2,5	4,5		10,4	3,7		2,0	
	sfa/focDE	25,4		34	22,8	20	12 à 20	27,6	6,3	8		7,8	
	focG		16	22				16,1		1,6		5,9	
	iha			50	58,4	35	51,7					23,5	33,7
Toxines	***hlyA,B,C ou D***	25,7	14	34	24,8		13,5	27,6		1,6	16,2	5,9	14,7
	cnf1	18,6	12	29	16,8	14,2	20,2	21,8	12,5	ND	21,6	3,9	11,8
	sat	28,8		45	53,4	26,7	53,9	46,6				23,6	25,8
	usp	52,4		68	72,3	57,5	53,9	68,1				37,3	38
Systèmes de capture du fer	***ireA***	28,6		33	19,8	19,2	6,7	24,6			29,4	17,6	20,4
	fyuA	76,4	70	92	98,0	75,8	79,8	89,3	70,8	56,9	76	70,6	60,2
	iroN	57,5		56	29,7	50,8	26,9	47,1	37,5		37,2	25,5	34,1
	iutA/aer	64,5	63	75	65,3	69,2	53,9	68,1	64,3	67	62	47,1	45,9[c]
Protectines	***neuC ou kpsMT K1***	21,1			22,8		29,2		8,3		53,4		23,7
	kpsMTII			83	57,4	55	69,7		45,8	49,5	67,9	52,9	
	traT	63,1		68	80,2	63,3	74,1		77,1	68,6	64,2	68,6	49,8
	iss			23	4,95		84,3				23,6	19,6	
	ompT	71,6			78,2	75,8	76,4		20,8		21	68,6	58,4
Invasine	***ibeA***	8,3		12	10,9	13,3	19	13,2	4,2	8,5		9,8	10,8

[a] Tous gènes *pap* confondus; [b] Tous allèles de *papG* confondus ; [c] *iucC* comme marqueur de l'aérobactine ; [d] Les données de ces études proviennent de génomes complets

Les cases grisées correspondent aux résultats manquants

a. Adhésines

L'adhésion aux cellules de l'hôte est une étape essentielle à la colonisation d'un site par la bactérie. Cette fonction est codée, comme leur nom l'indique, par les adhésines, des protéines exposées directement à la surface ou bien à l'extrémité de structures filamentaires de 2 à 8 nm, appelés "pili" ou "fimbriae" (Thanassi et al., 2007). Parmi ces protéines, certaines sont particulièrement conservées chez *E. coli*. C'est le cas des fimbriae de type I, parfois appelés hémagglutinines mannose-sensibles, codées par l'opéron *fimBEAICDFGH* (Figure 9) (Thanassi et al., 2007). Néanmoins, bien qu'elles soient ubiquitaires elles n'en restent pas moins impliquées dans la pathogénie des souches ExPEC. La présence d'un élément inversible (codé par *fimS*) permet de moduler l'expression de ces fimbriae de type 1 en fonction de la situation, menant par exemple au détachement de la bactérie de l'épithélium de la vessie pour aller coloniser le rein, après ascension de l'uretère (Kaper et al., 2004). De plus, il existe une interaction entre fimbriae de type 1 et mobilité, de sorte qu'en cas d'expression du fimbriae la bactérie perd en partie sa mobilité et la retrouve une fois l'expression inhibée (Lane, Simms, et al., 2007). Cette mobilité est une composante essentielle pour l'ascension du tractus urinaire aboutissant à une pyélonéphrite (Lane, Alteri, et al., 2007).

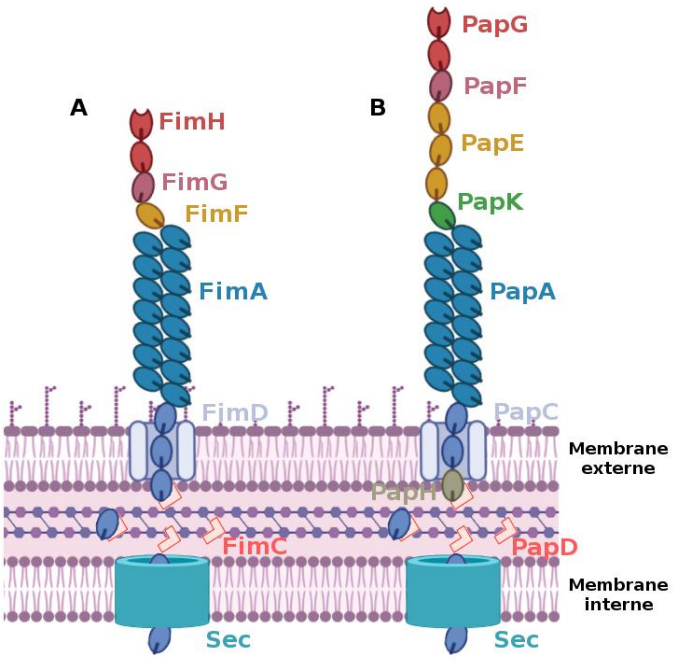


Figure 9. Structure des fimbriae de type 1 (A) et P fimbriae (B). Le transfert du cytoplasme vers la membrane externe est assuré via le système Sec. Dans le cas du fimbriae de type P, une protéine PapH vient coiffer le pili à son extrémité périplasmique (d'après Lillington et al., 2014).

D’autres adhésines sont également impliquées dans le processus infectieux urinaire et sont généralement retrouvées sur des îlots de pathogénicité (PAI). C’est le cas des fimbriae de type P aussi appelés Pap pour “Pyelonephritis-associated pili” (Figure 9). Bien que moins conservés que les fimbriae de type 1, leur prévalence atteint 80% dans les souches isolées de pyélonéphrites (Antão et al., 2009). Cette valeur diminue lorsque l’on inclut les bactériémies de diverses portes d’entrée mais reste néanmoins élevée en comparaison aux souches commensales, notamment pour l’allèle *papGII* (Tableau 5) (Clermont et al., 2017). L’extrémité de ces fimbriae est constituée d’une adhésine, PapG, dont les 3 allèles décrits présentent des récepteurs différents et sont associés à certaines formes cliniques d’infection urinaire (Strömberg et al., 1990). L’adhésine PapGI se fixe préférentiellement au globotriaosylcéramide ou GbO3 présent au niveau des cellules uroépithéliales humaines ; PapGII au globoside ou GbO4 présent au niveau des cellules uroépithéliales humaines ; et PapGIII au globopentaosylceramide ou GbO5 présent chez certains individus uniquement et chez les animaux dont les chiens et les chats par exemple (Lane & Mobley, 2007; Lindstedt et al., 1991; Strömberg et al., 1991). Les souches porteuses de l’allèle *papGII* sont plus fréquemment retrouvées dans les pyélonéphrites, et celles portant *papGIII* au cours des cystites ou des infections urinaires animales. Les allèles *papGI* sont trop rares pour pouvoir les associer à une forme clinique préférentielle.

On trouve également fréquemment des fimbriae de type S chez les souches ExPEC. Ces structures se liant aux acides sialiques sont codées par le groupe de gènes *sfa* (Antão et al., 2009)). Les souches UPEC sont fréquemment porteuses de ces gènes, avec une prévalence atteignant parfois 50% (Ewers et al., 2007). A nouveau, une valeur plus basse est retrouvée dans les souches isolées de bactériémies toutes portes d’entrée confondues, généralement autour de 20 à 30% contre moins de 10% chez les souches commensales (Tableau 5). Ces fimbriae pourraient aussi jouer un rôle important dans la physiopathologie des méningites à *E. coli*, et sont fréquemment retrouvés dans les souches NMEC (Ewers et al., 2007; Wijetunge et al., 2015).

Partageant une forte homologie de séquence avec les fimbriae de type S, on trouve les fimbriae de type F1C. Ces fimbriae non hémagglutinants, codés par le groupe de gènes *focACFGLH*, permettent l’adhésion aux cellules rénales et sont retrouvés chez près de 30% des souches isolées de pyélonéphrites (Antão et al., 2009; Pere et al., 1985; Thanassi et al., 2007). Ils se lient à des glycolipides ayant des motifs lactosylceramide.

D’autres adhésines, nommées Dr, ont pour récepteur une glycoprotéine membranaire régulatrice de la cascade du complément, le DAF ou "Decay Accelerating Factor", et sont

codées par les gènes *drACDE* (Antão et al., 2009). Structurellement différentes des autres fimbriae, et moins fréquemment retrouvées chez les UPEC (6%) et les souches de bactériémies en général, ces adhésines semblent pourtant impliquées dans l'invasion des cellules rénales (Ewers et al., 2007; Goluszko et al., 2001). Des adhésines nom fimbriaires, Afa, codées par les gènes *afaABCDE* peuvent également se lier au même récepteur (Antão et al., 2009).

Enfin nous pouvons également citer une adhésine particulière, codée par le gène *iha* pour "iron-regulated gene homologue adhesin". Cette protéine, initialement décrite chez une souche EHEC O157:H7, est aussi présente chez la souche UPEC CFT073 et leur confère des capacités d'adhésion (J. R. Johnson et al., 2005). Elle est par ailleurs retrouvée chez les souches responsables d'infections urinaires, comme celles appartenant au Clonal Group A (CGA), composé de souches du phylogroupe D résistantes à l'association triméthoprime - sulfaméthoxazole et décrit au début des années 2000 (Léveillé et al., 2006; Manges et al., 2001). L'originalité de cette protéine réside dans le fait qu'elle ne présente pas d'homologie avec les autres adhésines classiquement observées. Elle présente également des particularités fonctionnelles puisqu'en plus de conférer des capacités d'adhésion aux cellules épithéliales urinaires, cette protéine de la membrane externe agit comme récepteur de sidérophores de type catécholate et son expression est étroitement liée à la concentration en fer (Léveillé et al., 2006).

b. Toxines

Facteurs de virulence typiques de certaines souches InPEC, les toxines ont aussi un rôle essentiel dans les infections extra-intestinales. Dans sa revue en 1947, Kaufmann souligne déjà les capacités hémolytiques et nécrotiques régulièrement observées chez les souches isolées d'appendicites, de péritonites et d'infections urinaires (Kauffmann, 1947). Il en conclut qu'il existe un lien entre l'origine des souches, l'inagglutinabilité du sérotype O (liée en fait à la présence d'une capsule), et leur pouvoir hémolytique, nécrotique et toxique.

Les nécroses observées par Kaufmann sont probablement liées aux toxines de type CNF ("Cytotoxic necrotising factor"). Ces toxines, dont 3 variants sont décrits, CNF1, CNF2 et CNF3, ont la capacité de déaminer les résidus glutamine des GTPase Rho de la cellule hôte, les maintenant ainsi activées et aboutissant à de profondes modifications du cytosquelette, de la mobilité, de l'adhésion, du trafic vésiculaire ou encore conduisant à l'apoptose (Boquet,

2001; Kaper et al., 2004; Smith et al., 2015). Le gène *cnf1* codant la toxine du même nom est fréquemment retrouvé dans les souches UPEC et les souches isolées de bactériémies, atteignant parfois 40% contre seulement 10% pour les souches d'origine fécale (Tableau 5) (Landraud et al., 2000; Marrs et al., 2002). Néanmoins sa présence ne semble pas directement associée à la sévérité de l'infection, puisqu'il est trouvé à la fois au cours de bactériuries asymptomatiques, de cystites et de pyélonéphrites (Landraud et al., 2000).

Les souches porteuses de *cnf1* possèdent également souvent une hémolysine α. Cette toxine et son système d'excrétion sont codés par le groupe de gènes *hlyABCD* fréquemment présents sur le même îlot de pathogénicité que *cnf1*. Il s'agit d'une toxine RTX ("Repeat in Toxins") excrétée par un système de sécrétion de type I (Smith et al., 2015). Elle fait partie du groupe des "pore-forming toxin" en raison de son mode d'action sur les membranes des cellules eucaryotes.

D'autres toxines peuvent être retrouvées chez les souches ExPEC, parmi lesquelles Sat (« secreted autotransporter toxin ») et Vat (« vacuolating autotransporter toxin »), deux sérine-protéases. Sat a une activité cytopathique, responsable notamment d'une vacuolisation, sur les cellules du tractus urinaire, rénales et vésiculaires (Guyer et al., 2002). Vat présente aussi cette capacité cytotoxique et semble retrouvée fréquemment chez les UPEC et les APEC (Parreira & Gyles, 2003; Restieri et al., 2007). Cependant, en termes de prévalence Sat est parfois retrouvée autant chez les souches commensales que celles isolées de bactériémies (Tableau 5).

Enfin, on citera également le gène *usp* ("uropathogen specific protein") codant une bactériocine à activité nucléase (Parret & De Mot, 2002; Zaw et al., 2013). Ce gène est particulièrement prévalent dans les souches responsables d'infections urinaires, aussi bien cystite que pyélonéphrite ou prostatite, contrairement aux souches commensales fécales (Kurazono et al., 2000; Yamamoto et al., 2001). Son rôle dans la pathogénicité a été démontré dans un modèle murin de pyélonéphrite.

c. Systèmes de capture du fer

Le fer est un élément essentiel à la survie des bactéries. La quantité de fer libre disponible est bien souvent très limitée et nécessite la présence de systèmes de capture du fer pour pallier ce manque. Chez les vertébrés supérieurs, il peut être stocké sous différentes

formes (ferritine, liaison à l'hème, transferrine, lactoferrine), le rendant plus rares encore pour la bactérie au cours de l'infection (Garénaux et al., 2011). Afin de capter le fer, il existe chez *E. coli* des transporteurs membranaires comme FeoAB ou SitABCD, ou bien encore des récepteurs permettant la récupération à partir de l'hème extracellulaire comme Hma et ChuA (Garénaux et al., 2011). Ce dernier présente par ailleurs un intérêt dans la détermination du phylogroupe par des méthodes de PCR, les souches B2, D, E, F portant ce facteur alors que celles des groupes A, B1 et C non (Clermont et al., 2000, 2013). Il existe également des systèmes de chélation du fer, appelés sidérophores, et dont la prévalence est élevée dans les souches isolées de bactériémies (Clermont et al., 2017). Ces molécules sécrétées ont une très forte affinité pour le fer ferrique et peuvent ainsi capturer le fer de l'hôte lié à la lactoferrine et à la transferrine (Garénaux et al., 2011). Souvent synthétisée par des voies non ribosomales, ces molécules peuvent être classées en fonction de leur structure chimique en catécholates, phénolates, hydroxamates, α-hydroxy-carboxylates et formes mixtes (Figure 10) (Di Lorenzo & Stork, 2014; Garénaux et al., 2011). Ces systèmes sont généralement codés par des groupes de gènes permettant de réaliser l'intégralité du processus : la synthèse du sidérophore dans le cytoplasme, sa sécrétion, la réception du complexe fer/sidérophore, son internalisation et enfin le relargage dans le cytoplasme. Parmi ces sidérophores on peut en retenir quatre particulièrement fréquents : l'entérobactine présente chez la plupart des souches, la salmocheline, l'aérobactine et la yersiniabactine dont le rôle est souvent évoqué dans la virulence des ExPEC.

L'entérobactine, ou entérocheline, est un sidérophore de la classe des catécholates. Sa biosynthèse est codée par le groupe de gènes *ent,* et son transport, import et export par le groupe de gènes *fep.* L'énergie nécessaire est fournie par la force proton motrice, apportée de la membrane interne vers la membrane externe par le complexe TonB-ExbDB (Di Lorenzo & Stork, 2014; Garénaux et al., 2011; Raymond et al., 2003). Ce sidérophore est produit par la plupart des souches de *E. coli*, ExPEC ou non. Chez l'Homme, les polynucléaires neutrophiles sont capables de produire des lipocalines pour capter les complexes enterobactine-Fe^{3+} comme mécanisme de défense (Raymond et al., 2003).

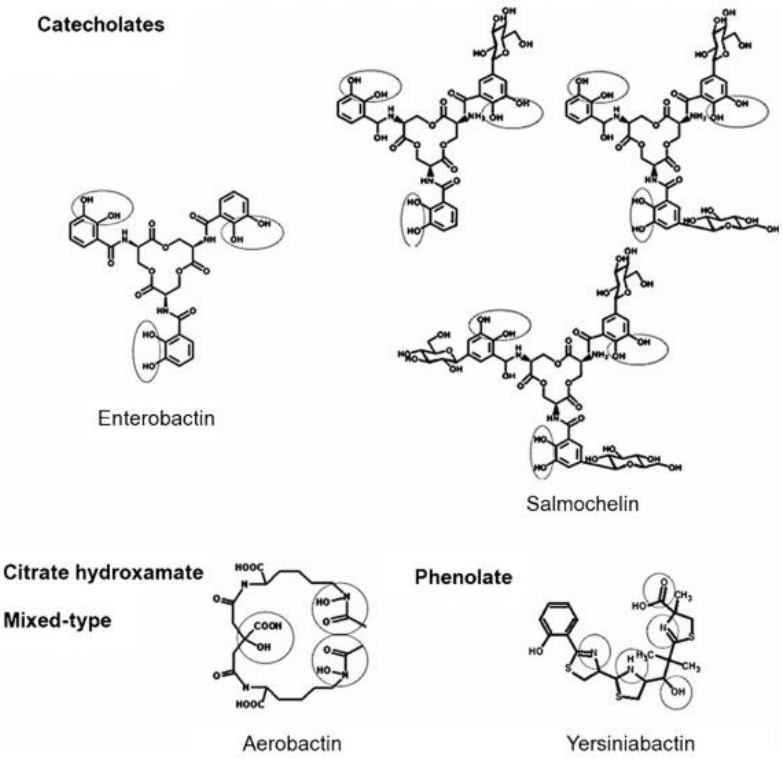


Figure 10. Structure chimiques des principaux sidérophores de *E. coli.* Les sidérophores de type catécholates incluent l'entérobactine et son dérivé glycosylé, la salmocheline. Sont présentés également l'aérobactine de type citrate hydroxamate, et la yersiniabactine appartenant aux phénolates. Les cercles indiquent les sites de fixation du fer (d'après Garénaux et al., 2011).

Cependant certaines souches de *E. coli* sont capables de produire un autre sidérophore de type catécholate insensible à l'activité des lipocalines. Ce sidérophore, appelé salmocheline, est également retrouvé dans plusieurs autres entérobactéries comme *Salmonella* ou *Klebsiella pneumoniae*. Il s'agit en réalité d'une forme glycosylée de l'entérobactine. Cette glycosylation est réalisée par IroB codée par le gène du même nom. Les autres gènes *iroCDE* interviennent dans l'import et l'export de la salmocheline et du complexe salmocheline-Fe^{3+}, alors que *iroN* code le récepteur TonB dépendant (Di Lorenzo & Stork, 2014). On la trouve généralement sur des éléments mobiles, comme sur le plasmide pS88 décrit dans certaines souches NMEC (Peigne et al., 2009). Ce système de capture est observé chez presque 50% des souches isolées de bactériémies (Tableau 5).

Également porté par le même plasmide typique des NMEC, pS88, il existe un autre sidérophore d'intérêt pour les ExPEC. Ce sidérophore de type citrate-hydroxamate, appelé aérobactine, est codé par le groupe de gènes *iucABCD* pour sa synthèse et *iutA* pour le

récepteur TonB dépendant (Di Lorenzo & Stork, 2014). Il est également désigné par le terme Aer et peut aussi être localisé sur le chromosome au niveau d'îlots de pathogénicité. Près de deux tiers des souches isolées de bactériémies portent cette aérobactine, contre seulement 45% voire un tiers des souches d'origine fécale (Tableau 5) (Carbonetti et al., 1986; Lefort et al., 2011; Massot et al., 2016).

La yersiniabactine est un autre sidérophore fréquemment retrouvé chez les souches ExPEC (Tableau 5). Cette molécule de type phénolate, initialement décrite chez les variants pathogènes du genre *Yersinia,* est codée par un ensemble de gènes regroupés sur un îlot désigné HPI pour "High Pathogenicity Island". Ce HPI est découpé en trois groupes fonctionnels : un permettant la biosynthèse de la yersiniabactine (*irp1-5* et *irp9*), un autre pour le transport (*irp6*, *irp7* et *fyuA*), et enfin un pour la régulation (*ybtA*). Ce système de capture du fer a été largement étudié dans un modèle murin. Dès le début des années 2000, il apparaît comme un des facteurs les plus associé à la mortalité (Schubert et al., 2002). Plus récemment, une étude d'association pangénomique a confirmé l'importance de ce facteur dans la virulence chez la souris sur un grand nombre de souches (Galardini et al., 2020).

Enfin on peut également citer le gène *ireA* ("iron responsive element A") qui code la protéine IreA. Celle-ci partage une similarité importante avec les autres récepteurs de sidérophores suggérant une fonction associée à la recapture du fer (T. A. Russo et al., 2001). Les études en modèle murin d'infection urinaire ont montré sa contribution dans l'urovirulence, et les données épidémiologiques viennent conforter ces résultats (T. A. Russo et al., 2001). Bien que considéré comme un facteur de virulence, il apparaît comme facteur de bon pronostic dans les bactériémies à *E. coli*, reflétant probablement son lien important avec la porte d'entrée urinaire, également de bon pronostic (Lefort et al., 2011).

On peut noter une certaine redondance dans ces systèmes de capture du fer puisque certains sont capables de capter les mêmes sidérophores (FepA, IroN, Iha), et que, d'autre part, certaines souches portent plusieurs systèmes de manière concomitante. Une étude dans un modèle murin d'infections urinaires a cependant mis en évidence qu'ils n'étaient pas tous équivalents (Garcia et al., 2011). Ainsi les récepteurs des systèmes de capture de l'hème (Hma et ChuA) et des sidérophores non catécholate (IutA et FyuA) semblent contribuer de manière plus significative à l'infection urinaire par les souches UPEC. Il existe probablement aussi des variations liées à la l'anatomie, avec par exemple une contribution plus importante au niveau vésicale pour IutA et au niveau rénale pour Hma et ChuA (Garcia et al., 2011).

d. Protectines

Certaines souches ont la capacité de produire une capsule polysaccharidique. Celle-ci confère une résistance aux défenses immunitaires de l'hôte et aux conditions environnementales défavorables, permettant d'éviter par exemple la dessiccation (Whitfield, 2006). Très tôt, Kaufmann avait noté l'implication de cet élément dans la pathogénicité des souches, leur conférant une plus grande toxicité et une résistance à l'hôte et aux bactériophages (Kauffmann, 1947). Son rôle dans certaines infections extra-intestinales est largement reconnu, notamment dans les méningites néonatales dont les souches responsables sont souvent porteuses de l'antigène K1, comme c'est le cas de la souche S88 porteuse des gènes codant la capsule sur le plasmide pS88 (Denamur et al., 2021; Peigne et al., 2009). Il existe en réalité 4 groupes de capsules selon le mécanisme de synthèse et d'assemblage. Les groupes 1 et 4 sont proches et sont codés par des gènes *wzx* (flipase) et *wzy* (polymérase) comme les antigènes O (Whitfield, 2006). Cette proximité avec les systèmes de synthèse des antigènes O est d'ailleurs probablement à l'origine d'erreurs dans le sérotypage des souches par des approches reposant sur l'analyse des séquences nucléotidiques. Les groupes 2 et 3, eux, utilisent le groupe de gènes *kps*, organisé en 3 régions principales : une région centrale dont la taille et le contenu en gènes est spécifique du sérotype, et deux régions flanquantes codant la machinerie pour la synthèse et l'export. La plupart des sérotypes capsulaires retrouvés dans les ExPEC appartiennent au groupe 2, notamment les antigènes K1 et K5. La détection du gène *neuC* est souvent utilisée pour mettre en évidence le sérotype K1. Si le rôle de la capsule a été démontré dans la pathogénie des souches NMEC, notamment en conférant une protection vis-à-vis du complément et au sein des cellules endothéliales vasculaires cérébrales, cela apparaît moins évident pour d'autres ExPEC comme les UPEC (Sarkar et al., 2014; Wiles et al., 2008). Néanmoins sa fréquence au sein des souches responsables d'infections extra-intestinales, comme chez la souche UTI89 porteuse de l'antigène K1, laisse supposer qu'elle participe au processus infectieux.

D'autres facteurs sont souvent décrits comme impliqués dans la protection vis-à-vis des défenses de l'hôte. C'est le cas de la protéine Iss pour "Increased Serum Survival" (Chuba et al., 1989). Cette protéine de la membrane externe confère comme son nom l'indique une résistance au sérum, probablement par inhibition de la formation du complexe d'attaque membranaire (J. R. Johnson, 1991). Initialement décrite chez les APEC, elle est aussi fréquemment retrouvée chez les ExPEC humain, sous la forme d'un variant différent, le variant de type 3 (T. J. Johnson et al., 2008). De plus, contrairement aux APEC, chez les ExPEC humain elle est généralement de localisation chromosomique et non plasmidique.

Sur le même plasmide porteur d'*iss*, le gène *traT* code lui aussi une protéine considérée comme protectine (Binns et al., 1979, 1982), Son mode d'action n'est pas complètement élucidé mais elle semble agir également par inhibition de certaines étapes terminales des voies du complément (Miajlovic & Smith, 2014).

Enfin, la protéine OmpT est aussi un mécanisme de défense de certaines souches de *E. coli*. Cette protéine a une activité protéase impliquée dans la dégradation de divers peptides antimicrobiens (Desloges et al., 2019; Stumpe et al., 1998). Le gène chromosomique *ompT* codant cette protéine présente une forte prévalence chez les souches UPEC et pourrait donc participer au processus infectieux au moins chez ce type d'ExPEC. L'activité protéase est d'ailleurs plus importante chez ces souches que chez les souches commensales (Desloges et al., 2019). Par ailleurs, un autre gène partageant une forte similarité est parfois identifié et localisé sur un plasmide (Desloges et al., 2019; McPhee et al., 2014). Ce gène *ompT* épisomal, parfois nommé *arlC*, code également une protéine responsable de la dégradation de peptides antimicrobiens, mais son spectre apparaît légèrement différent de la forme chromosomique.

e. Invasine

Peu de facteurs de virulence appartiennent à cette classe. Certains facteurs sont probablement impliqués dans l'invasion en tant que telle mais de par leur fonction et leur mécanisme d'action, ils sont souvent associés à d'autres classes (adhésines, toxines). L'invasine la plus fréquemment étudiée est une protéine membranaire codée par le gène *ibeA* (Huang et al., 2001). Ce facteur est associé principalement à deux types d'ExPEC : les NMEC et les APEC (Germon et al., 2005; Huang et al., 2001). Sa prévalence est en effet plus importante chez les souches APEC, chez lesquelles le rôle dans la virulence a été démontré. Chez les NMEC, cette protéine semble participer à l'invasion des cellules endothéliales vasculaires cérébrales. Son importance est cependant moins claire pour les autres ExPEC, comme en témoigne sa faible prévalence chez les souches isolées de bactériémies (Tableau 5) (Clermont et al., 2017).

f. Îlots de pathogénicité

Nombre de ces facteurs de virulence sont localisés sur des îlots génomiques appelés îlots de pathogénicité (Figure 11). Ces régions partagent des caractéristiques communes

comme leur taille de 10 à 200 kb, leur association fréquente à un ARNt de transfert, un contenu en GC% différent du reste du génome, des séquences flanquantes répétées, une structure mosaïque avec une multitude de séquences codantes, fonctionnelles ou non et de fonction connue ou non, et enfin la présence de nombreux fragments d'éléments génétiques mobiles (Blum et al., 1994; Dobrindt et al., 2002; Hacker & Kaper, 2000). On parle parfois d'îlots de "fitness" ou d'îlots d'écologie, notamment pour expliquer leur présence dans des souches non pathogènes. Un nombre limité de gènes codant des ARNt chez *E. coli* est impliqué dans l'intégration de ces PAI, et semblent notamment différer en fonction du phylogroupe des souches (Germon et al., 2007). L'organisation de ces ARNt, leur taux de transcription et ou encore la fréquence d'utilisation du codon reconnu semblent être autant d'arguments impliqués dans l'intégration plus ou moins fréquente au niveau de ces sites. Les séquences d'insertion sont rares aux extrémités mais très fréquentes au sein des PAI et pourraient permettre une adaptation rapide par la délétion de certaines parties du PAI (Hacker & Kaper, 2000). Certains auteurs suggèrent que ces PAI jouent un rôle macroévolutif avec l'émergence de souches pathogènes, et microévolutifs quand leurs remaniements et/ou leur délétion permettent une adaptation rapide à certaines conditions environnementales (Groisman & Ochman, 1996; Hacker & Kaper, 2000).

Dans sa revue en 1947, Kauffmann observait d'ailleurs les conséquences de ces PAI au niveau phénotypique avec une forte association entre les capacités hémolytiques et nécrotiques des souches. Ainsi sur 78 souches hémolytiques, 79 % apparaissaient aussi nécrotiques probablement du fait de la présence des gènes *hly* et *cnf* sur un même PAI, alors que sur 464 non hémolytiques seules 27% étaient nécrotiques (Kauffmann, 1947).

Parmi les différents PAI fréquemment observés dans les souches ExPEC, on peut citer le HPI qui porte la yersiniabactine impliquée dans la recapture du fer comme décrit précédemment (Schubert et al., 2002). Ce PAI est particulièrement intéressant car il s'agit du déterminant génétique le plus fortement associé à la virulence extraintestinale chez la souris (Galardini et al., 2020), et d'autre part il n'est pas limité à l'espèce *E. coli* (Schubert et al., 2000). On le trouve en effet dans diverses entérobactéries d'intérêt en clinique humaine, suggérant une implication importante de cet élément dans la virulence extraintestinale (Lawlor et al., 2007; Schubert et al., 2000).

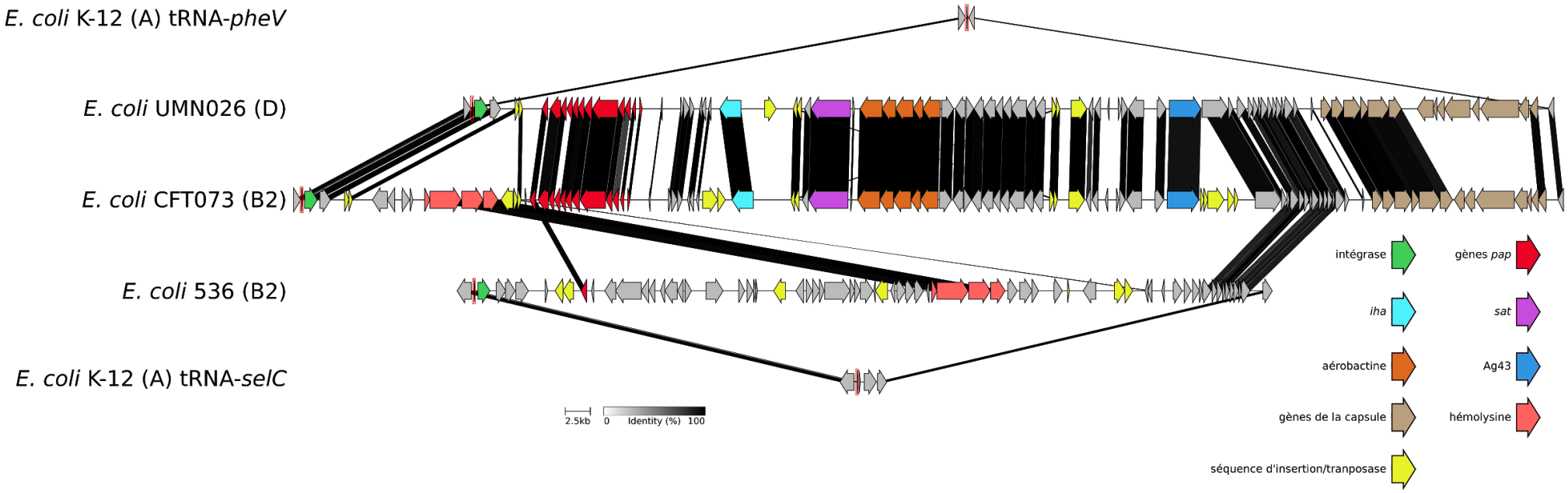


Figure 11. Exemple d'îlots de pathogénicité retrouvés au sein de souches ExPEC archétypales. Les principaux facteurs de virulence connus ainsi que les éléments impliqués dans la mobilité apparaissent en couleur.

3. Résistance aux antibiotiques et ExPEC

Le phénotype “sauvage” de *E. coli* se caractérise par une sensibilité à la plupart des antibiotiques utilisés en clinique contre les bacilles à Gram négatif. Les résistances naturelles observées sont essentiellement liées à la présence de la membrane externe, imperméables à de nombreux composés tels la benzylpénicilline ou pénicilline G, les pénicillines M anti-staphylococcique, les glycopeptides. Bien qu’elle possède une enzyme de type céphalosporinase non inductible, *E. coli* ne présente pas de résistance aux aminopénicillines ni aux céphalosporines de première génération en raison d’une expression trop faible à l’état basal. Les autres classes de bêta-lactamines comme les carboxypénicillines et uréidopénicillines, les céphalosporines de 2ème, 3ème et 4ème générations, ainsi que les carbapénèmes ont une activité sur les souches sauvages.

Les résistances acquises sont en revanche nombreuses et fréquentes, et présentent d’importantes variations spatio-temporelles. A l'échelle même de l’Europe, on observe un gradient Nord/Sud. Ainsi, l’augmentation de la résistance antibiotique, tant en contexte communautaire que nosocomial, en fait une des priorités selon L’OMS[4]. Nous n’aborderons pas ici le détail de ces résistances, mais nous nous focaliserons plutôt sur trois exemples de clones ExPEC d’intérêt clinique car résistants à des antibiotiques largement utilisés, et ce par différents mécanismes. D’après Clermont *et al.*, la résistance antibiotique accrue est une des caractéristiques communes aux souches responsables de bactériémies, aussi bien à point départ urinaire que digestif (Clermont et al., 2017). Ce phénomène est même exacerbé chez certains clones ExPEC ayant récemment émergé, faisant tomber la barrière théorique entre virulence et résistance.

a. Résistance au cotrimoxazole et ST69

Mis en évidence à la toute fin des années 1990, le clone ExPEC ST69 appartient au phylogroupe D et est un modèle de souche à la fois virulente et résistante (Denamur et al., 2021; Manges et al., 2001). Manges *et al.* l’ont identifié en raison de sa forte prévalence dans les infections urinaires féminines dans une université Californienne ainsi que dans le Minnesota et le Michigan, et ont noté sa résistance à l’association triméthoprime-sulfaméthoxazole ou cotrimoxazole. Nommé “Clonal Group A” (CGA) d’après les résultats

[4] https://www.who.int/drugresistance/AMR_Importance/en/

d'amplification de séquences répétées intergéniques (PCR ERIC2), d'autres arguments pointaient clairement son caractère clonal : un nombre limité de sérotypes et de profils en électrophorèse en champs pulsé, une multirésistance antibiotique phénotypique décrite par le profil ACSSuTTp (Ampicilline, Chloramphénicol, Streptomycine, Sulfaméthoxazole, Tétracycline, Triméthoprime) ou encore son profil de virulence génotypique (*papGII*, *iutA*, *kpsMTII*, et *traT*) (Manges et al., 2001). Hormis les gènes plasmidiques *tet* codant la résistance à la tétracycline, l'ensemble des déterminants génétiques associés à ce profil est situé au sein d'une même région chromosomique nommée "module de résistance génomique" contenant notamment un intégron de classe 1 portant les cassettes de gènes *dfrA17,* pour la résistance au triméthoprime-sulfaméthoxazole, et *aadA5* pour la résistance à la streptomycine (Figure 12) (Lescat et al., 2009; Solberg et al., 2006). Ce module de résistance est lui-même présent au sein d'un îlot génomique de plus de 100 kpb, qui présente, en plus de sa taille, toutes les caractéristiques typiques de ces éléments : insertion au niveau d'un ARN de transfert (*leuX*), composition en bases GC différente du reste du génome, présence de nombreux éléments mobiles (séquences d'insertion, transposons, intégrons). Par ailleurs, cette région est un point chaud d'intégration comme en témoignent les variations observées à la même localisation dans d'autres génomes de *E. coli* (Lescat et al., 2009). On notera notamment la présence d'un îlot de pathogénicité à cette même position chez les souches ExPEC de phylogroupe B2 UTI89 (ST95) et 536 (ST127).

L'émergence de ce clone au cours du temps a été étudiée rétrospectivement à l'aide de génomes de souches responsables de bactériémies, et les auteurs ont pu observer son apparition en 2002 (Kallonen et al., 2017). Cette phase correspond, d'après leurs analyses, à une troisième augmentation de ce clone dans la population, plus rapide et plus importante que les deux précédentes qui auraient eu lieu en 1970 et 1990. Sa forte prévalence, notamment dans les pays occidentaux, a bien été démontrée dans les infections extra-intestinales, urinaires ou non, et ce clone est aujourd'hui largement implanté et inclus dans ce que Denamur *et al.* nomment les "big four" (Denamur et al., 2021; J. Johnson et al., 2011).

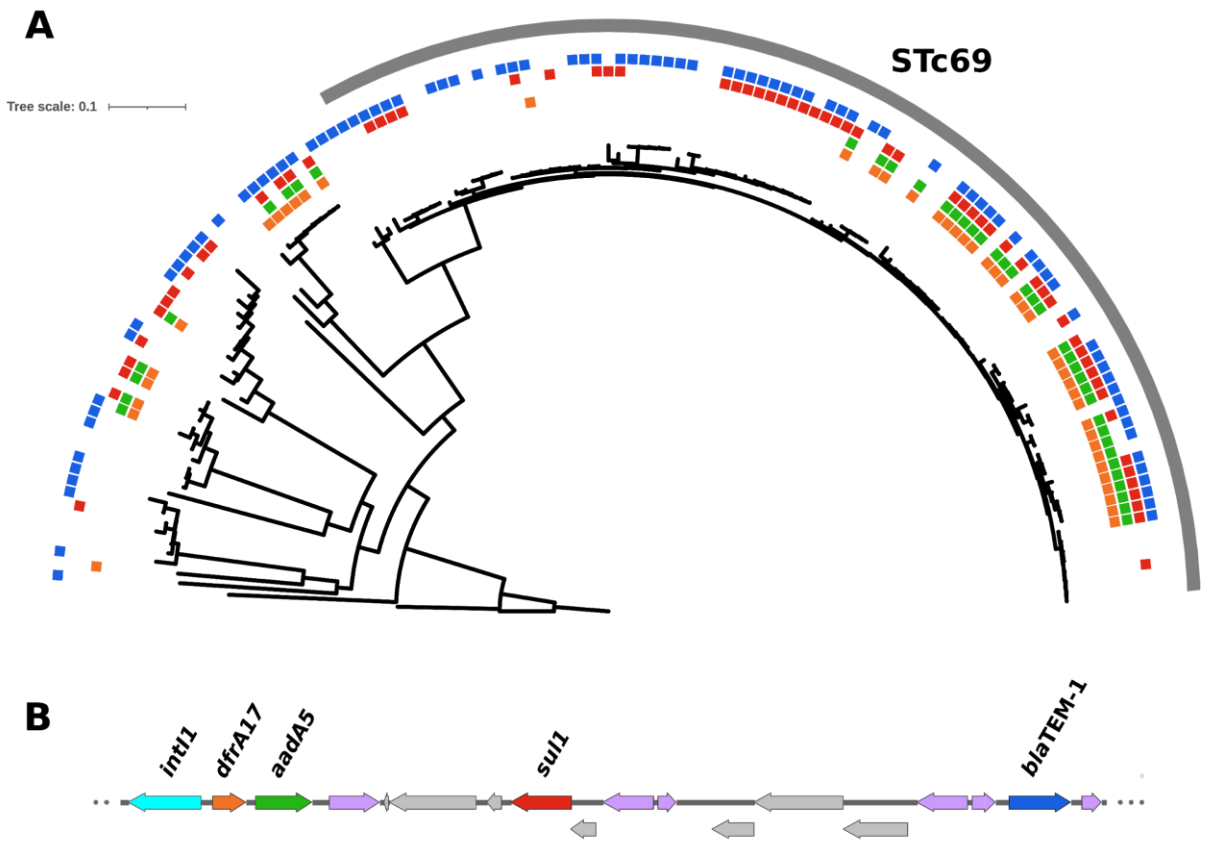


Figure 12. A) Structure phylogénétique du phylogroupe D. Le STc69 ou “Clonal Group A” apparaît cerclé de gris. Les gènes de résistance présents sur l’intégron au sein du module de résistance génomique sont indiqués par des carrés de couleur dont le code couleur correspond aux gènes représentés dans la partie B. Les génomes représentés sur cet arbre proviennent des collections Colibafi (bioproject PRJEB39260) et Septicoli (bioproject PREJB35745). B) Carte génétique de l’intégron de classe 1 identifié dans la souche UMN026 (NC_011751) du STc69. Les principaux gènes de résistance apparaissent en couleur, le gène codant l’intégrase en turquoise et les séquences d’insertion, transposases et résolvases en violet (d’après les annotations de Lescat et al., 2009).

b. Le ST131 et l’expansion de *bla*CTX-M

Des bêta-lactamases à spectre étendu ont été décrites chez le STc69. Dans une étude sur les souches isolées de bactériémies, le STc69 représente même 4% de la population des *E. coli* producteurs de telles enzymes (Roer et al., 2017). Mais, celui-ci arrive loin derrière le clone pandémique du phylogroupe B2 : le ST131. En effet, dans cette même étude danoise, 50% des souches BLSE appartiennent à ce clone ExPEC et portent majoritairement le variant *bla*CTX-M-15. Le ST131 fait figure d’exception au sein du phylogroupe B2, dans lequel la

présence d'enzymes de type CTX-M est habituellement rare (Brisse et al., 2012). Ce clone a été observé au début des années 2000 au Royaume-Uni puis décrit plus précisément et identifié à l'échelle mondiale en 2007 (Nicolas-Chanoine et al., 2007; Woodford et al., 2004). Son augmentation brutale à partir de la moitié des années 2000 a par la suite été rapportée par de nombreuses équipes, notamment dans les souches responsables de bactériémies, et fréquemment en contexte communautaire (J. R. Johnson, Urban, et al., 2012; Kallonen et al., 2017; Matsumura et al., 2017; Peirano et al., 2012). Dans leur revue de 2014, Nicolas-Chanoine *et al.* estiment la prévalence du ST131 entre 12 et 27% au sein des souches de *E. coli* isolées d'infections extra-intestinales humaines (Nicolas-Chanoine et al., 2014). En ne considérant que les isolats résistants, on observe une prévalence encore plus élevée atteignant 20 à 66% des *E. coli* BLSE et 10 à 72% des souches résistantes au fluoroquinolones. Plusieurs clades (*i.e.* A, B et C) sont décrits pour le ST131, en fonction des combinaisons de sérotypes O:H et des allèles *fimH* (Figure 13) : le plus basal correspond au clade A et est composé de souches de sérotype O16:H5 et *fimH*41 ; puis vient le clade B de sérotype O25b:H4 et *fimH*22, voire *fimH*30 comme décrit récemment (Duprilot et al., 2020) ; enfin le clade C également de sérotype O25b:H4 et *fimH*30. Certains clades (C1 et C2) sont particulièrement associés à la résistance aux fluoroquinolones par mutation dans les gènes codant les topoisomérases (*gyrA* et *parC*), ainsi qu'aux céphalosporines de 3ème génération par production d'enzyme de type CTX-M (clade C2) (Figure 13) (Ben Zakour et al., 2016). L'analyse des génomes des différents clades a permis de mettre en évidence une évolution séquentielle avec i) la recombinaison de l'antigène O autour de 1946 séparant le clade A des clades B et C, ii) l'acquisition d'îlots génomiques porteurs de déterminants de virulence dans les années 1980 (sous-clades C0, C1 et C2) suivi de iii) l'acquisition de résistance aux fluoroquinolones (sous-clades C1 et C2) et de la BLSE CTX-M-15 (clade C2) (Ben Zakour et al., 2016). De plus, il semblerait qu'il y ait en réalité eu de nombreux évènements d'acquisition et de perte de *bla*CTX-M-15 dans le clade C2 (Goswami et al., 2018; Kallonen et al., 2017). La résistance aux antibiotiques, qu'il s'agisse des fluoroquinolones ou des céphalosporines de 3ème génération, ne semble pas pour autant être à l'origine de son expansion et aucun profil de virulence spécifique à ce clone n'a pu être identifié pour l'heure (Ben Zakour et al., 2016; Kallonen et al., 2017). Les phénomènes qui ont mené à sa sélection et à son expansion sont donc toujours débattus, mais la virulence équivalente à celles des autres clones ExPEC suggère plutôt des capacités accrues dans les phases initiales du processus infectieux telle la colonisation et/ou la transmission (J. R. Johnson, Porter, et al., 2012; Vimont et al., 2012). La niche écologique de ce clone ExPEC hautement spécialisé semble étroite, comme le suggère le taux de recombinaison plus faible du coregénome comparé aux autres souches de *E. coli* (McNally et al., 2013). D'ailleurs, sa prévalence chez l'animal, dans les aliments et

dans l'environnement apparaît faible, et il semble que le réservoir primaire de ce clone soit très lié à l'Homme (Nicolas-Chanoine et al., 2014).

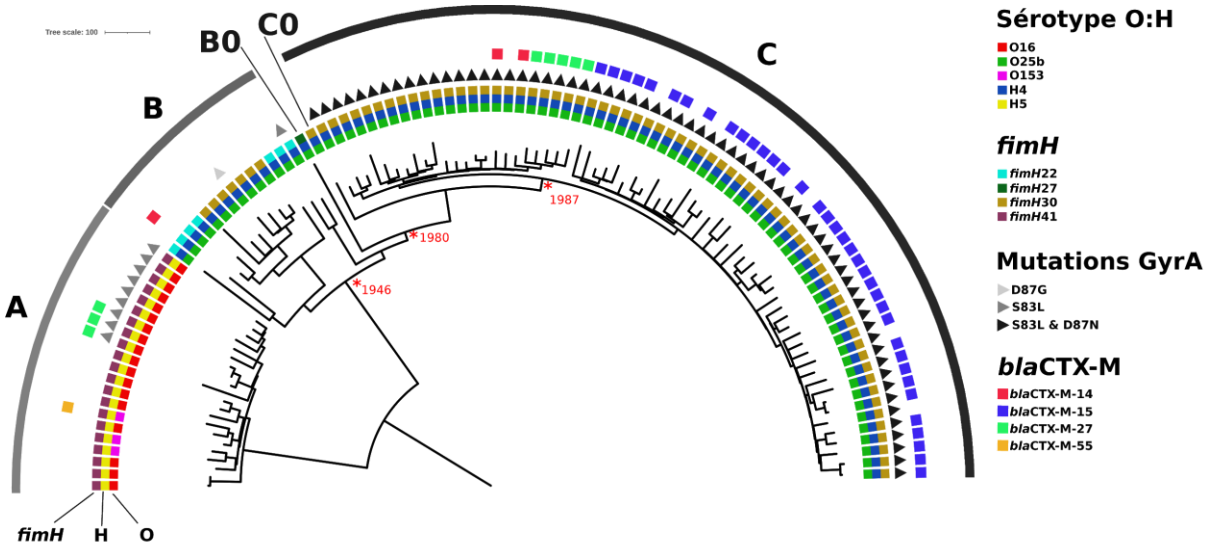


Figure 13. Phylogénie du ST131 après prise en compte des recombinaisons. Les 3 grands clades sont représentés ainsi que les clades intermédiaires B0 et C0. La présence de mutations dans la gyrase GyrA est indiquée par des triangles et les gènes codant les enzymes de type CTX-M par des carrés de couleur. Les étoiles rouges correspondent aux divergences datées par Ben Zakour *et al.* (Ben Zakour et al., 2016). Les génomes représentés sur cet arbre proviennent des collections Colibafi (bioproject PRJEB39260) et Septicoli (bioproject PREJB35745).

Le support génétique des enzymes CTX-M chez le ST131 est plasmidique, bien que des insertions au niveau chromosomique aient été décrites (Hirai et al., 2013; Ludden et al., 2020). Les séquences de type IS*Ecp1* ont joué un rôle essentiel dans la mobilisation de ces enzymes depuis le génome des espèces du genre *Kluyvera* vers des plasmides, tout en apportant un promoteur au gène codant la bêta-lactamase (Poirel et al., 2003). Les plasmides en cause sont variés mais appartiennent bien souvent au groupe d'incompatibilité IncF, faisant évoquer une association particulièrement efficace entre le ST131, l'enzyme CTX-M-15 et ces plasmides (Woodford et al., 2009). De plus, on observe une association préférentielle de certains sous-groupes de plasmides IncF avec les clades C1 ou C2 (T. J. Johnson et al., 2016; Kondratyeva et al., 2020). Par ailleurs, ces plasmides sont souvent porteurs de nombreux autres gènes de résistance parmi lesquels *bla*OXA-1, *bla*TEM-1, *aac6'-Ib-cr, mph*(A), *catB4*. La forte prévalence de ces plasmides au sein du ST131 pourrait en partie s'expliquer par des interactions épistatiques permettant de diminuer le coût associé au portage de telles plasmides (T. J. Johnson et al., 2016).

c. Une voie originale dans l'acquisition de la résistance aux fluoroquinolones : l'exemple du ST1193

Plus récemment, un autre clone ExPEC multirésistant a été identifié et semble diffuser à l'échelle mondiale (Tchesnokova, Rechkina, et al., 2019). Tchesnokova *et al.* ont focalisé leur attention sur les isolats cliniques (urine, hémoculture et plaies) résistants aux fluoroquinolones provenant de 9 laboratoires américains entre 2016 et 2017. Ils ont confirmé d'une part la forte prévalence du ST131 *fimH*30, représentant 44,8% des isolats résistants, et d'autre part ont identifié le ST1193 dans 22,9% des cas, ce qui en fait le second clone en termes de fréquence. Ce clone ExPEC de phylogroupe B2 est un variant mono-allélique du ST14 appartenant au sous-groupe VIII (Tableau 4) et dont l'émergence semble récente, probablement autour de 2005. Contrairement au ST131, le ST1193 est associé à des patients plus jeunes (< 40 ans) et semble plus fréquemment retrouvé dans les urines que les hémocultures. En plus de sa résistance aux fluoroquinolones, il apparaît souvent résistant à l'association triméthoprime-sulfméthoxazole et à la tétracycline dans plus de la moitié des cas.

Comme souvent, la résistance aux fluoroquinolones est médiée chez ce clone par des mutations chromosomiques dans les régions déterminants la résistance aux quinolones (QRDR) dans les gènes *gyrA* et *parC* (Tchesnokova, Rechkina, et al., 2019). Cependant l'originalité réside ici dans les phénomènes ayant mené à ces mutations chez le ST1193. En effet, l'acquisition de cette résistance est habituellement décrite comme séquentielle, comme dans le cas du ST131, avec une accumulation de mutations ponctuelles l'une après l'autre, associée à un niveau de résistance phénotypique croissant (Huseby et al., 2017). Mais dans le cas du ST1193, l'acquisition de la résistance semble avoir eu lieu par des phénomènes de recombinaison (Tchesnokova, Radey, et al., 2019). En effet, Tchesnokova *et al.* ont mis en évidence une diversité nucléotidique élevée dans *gyrA*, *parC* et les régions environnantes par comparaison entre le ST1193 et les autres génomes du STc14. Mais surtout, ils sont parvenus à identifier le progéniteur de ces segments acquis par recombinaison en explorant la base de données Enterobase (Z. Zhou et al., 2020). Il s'agit de souches appartenant au clone ST10-*H*54, très éloigné phylogénétiquement puisque de phylogroupe A et très rarement observé dans les infections urinaires. D'autres segments sont également impliqués dans cette recombinaison aboutissant parfois à des gains ou des pertes de gènes chez le ST1193 par rapport aux autres génomes du STc14. Le phénomène ayant permis l'acquisition initiale de matériel génétique n'a pu être déterminé par les auteurs, bien que la conjugaison soit le plus probable en raison de la taille importante de certains segments, de la multiplicité des segments en cause (9 segments), et du caractère peu ou pas transformable de la bactérie *E.*

coli. Il est difficile d'incriminer cette acquisition de résistance antibiotique comme seule responsable de l'émergence rapide et mondiale de ce clone, néanmoins elle y a sans doute participé au vu du peu d'autres différences observées avec le reste du STc14.

Les analyses épidémiologiques portant sur les souches responsables de bactériémies semblent montrer que la résistance antibiotique n'est pas nécessairement responsable de l'émergence d'un clone ExPEC (Kallonen et al., 2017). En effet, certains clones particulièrement virulents et parmi les plus fréquemment isolés au cours de ces infections présentent une sensibilité à la plupart des antibiotiques d'intérêt clinique, comme dans le cas des ST73 et ST95 (Denamur et al., 2021; Gordon et al., 2017; Kallonen et al., 2017). Néanmoins les clones ST69, ST131 et ST1193, d'émergence relativement récente, présentent une résistance antibiotique à des molécules d'intérêt clinique, et les multiples mécanismes génétiques impliqués (intégrons, mutations ponctuelles, plasmides, recombinaison) pointent vers une certaine convergence dans l'acquisition de la résistance chez les souches ExPEC. S'il est généralement difficile voire impossible de considérer cette résistance comme seule cause à leur émergence, il fait peu de doute qu'elle participe *a minima* à leur maintien dans la population. Par ailleurs, la pression de sélection exercée par les antibiotiques est un phénomène relativement récent comparé à l'adaptation des souches pathogènes à l'Homme, et ces différences d'échelles temporelles peuvent complexifier la mise en évidence du rôle de la résistance et de la virulence dans l'émergence de clones ExPEC (Beceiro et al., 2013).

4. Plasmides et ExPEC

L'évolution des génomes bactériens résulte à la fois de la microévolution du coregénome et de la macroévolution du génome accessoire via des évènements de transfert horizontaux dans lesquels les plasmides jouent un rôle primordial (de Toro et al., 2014). Les plasmides sont des éléments extrachromosomiques ne contenant généralement pas de gènes essentiels à la bactérie, capable de réplication autonome, et qui peuvent porter et disséminer des gènes conférant un avantage sélectif à leur hôte tels des capacités accrues de résistance antibiotique, de virulence, ou encore des traits métaboliques particuliers (Carattoli, 2009). Ils peuvent jouer un rôle primordial dans l'adaptation à une niche écologique particulière. Chez les gamma-protéobactéries près de la moitié des plasmides ne sont pas mobilisables, le reste étant à part égale mobilisable ou conjugatif (Smillie et al., 2010). Bien que le spectre d'hôte de certains plasmides soit réputé large, il semble que pour la majeure partie d'entre eux leurs hôtes naturels soient généralement proches d'un point de vue taxonomique (Redondo-Salvo et al., 2020; Smillie et al., 2010). La densité codante de ces éléments est variable, généralement plus faible que celle du chromosome pour les petits plasmides alors qu'elle tend à s'en rapprocher lorsque la taille devient plus importante. La longueur de la séquence semble aussi influencer le caractère conjugatif ou non des plasmides, avec une taille généralement supérieure à 30 kb pour les éléments capables d'assurer eux-mêmes leur transmission. Les plasmides possèdent des systèmes d'addiction, souvent de type toxine-antitoxine, qui assurent leur maintien dans la population même en l'absence de pression de sélection. L'identification et la classification des plasmides n'est pas chose aisée tant la diversité génétique de ces éléments est grande (Carattoli, 2009; Williams et al., 2013). Une des approches fréquemment employée est d'utiliser des éléments assez conservés et impliqués dans le contrôle de la réplication. Néanmoins aucun schéma ne permet de tous les classer, et il a donc récemment été proposé d'utiliser des approches reposant sur la comparaison de l'identité nucléotidique, d'une manière semblable à celle utilisée pour la définition des espèces bactériennes (Redondo-Salvo et al., 2020).

Chez l'espèce *E. coli*, la plupart des isolats portent des plasmides, généralement au nombre de 2 à 4 (Denamur et al., 2021)(Denamur et al., 2021). Certains semblent particulièrement adaptés à l'espèce comme les plasmides de groupes d'incompatibilité IncF et IncI (de Toro et al., 2014; Williams et al., 2013). Il ne semble pas y avoir d'association entre un phylogroupe particulier et un type de plasmide (Branger et al., 2018).

La résistance antibiotique est un domaine intimement lié aux plasmides et ils jouent indéniablement un rôle essentiel dans sa diffusion, chez *E. coli* comme chez bien d'autres

entérobactéries (Carattoli, 2009; Rozwandowicz et al., 2018). L'un des exemples les plus parlant chez les ExPEC est probablement celui du ST131 cité précédemment. L'acquisition de plasmides de type IncF porteur du gène *bla*CTX-M-15, lui-même fréquemment inséré au sein d'un transposon contenant le gène de la pénicillinase *bla*TEM-1, ainsi que d'autres gènes de résistance, a très probablement participé au succès mondial de ce clone (Branger et al., 2018; Mathers et al., 2015; Woodford et al., 2009). Dans le cas du ST131, il existe une certaine diversité plasmidique et une dissémination de la résistance essentiellement associée à l'expansion d'un clone. Mais dans certains cas, on peut assister à une épidémie d'origine plasmidique avec, comme pour *bla*CTX-M-14 en Corée, la diffusion d'un même plasmide IncF dans différents fonds génétiques de *E. coli* (J. Kim et al., 2011). Des plasmides à spectre d'hôte plus large comme les IncI et IncN sont aussi associés à la diffusion de certains variants comme *bla*CTX-M-1, probablement en lien avec le monde animal, notamment aviaire (Carattoli, 2009; Girlich et al., 2007; Rozwandowicz et al., 2018). De nombreux autres gènes de résistance sont aussi observés sur des plasmides chez *E. coli*, comme des céphalosporinases CMY-2 sur des plasmides IncA/C ou IncI1, la métalloenzymes VIM-1 sur un plasmide IncN, ou encore la carbapénémase de classe D *bla*OXA-48 sur un plasmide IncL/M (Carattoli, 2009; Poirel et al., 2012).

L'utilisation des antibiotiques a probablement en partie façonné l'épidémiologie des plasmides retrouvés chez *E. coli*, et les gènes codant des BLSE sont fréquemment observés au sein de régions de multi-résistance antibiotique. Branger *et al.* observent une association entre les plasmides conjugatifs de grande taille, porteur de gènes codant des BLSE, et l'isolement de la souche porteuse après les premières utilisations de C3G (Branger et al., 2018). A l'inverse, avant utilisation de ces molécules, les plasmides semblent de plus petite taille et sont généralement seulement mobilisables ou non transférables. Ces plasmides peuvent néanmoins, pour certains d'entre eux comme les plasmides MOB_{RNA} de type 4 et les plasmides $RelN_{RNA}$ de type 1, être porteurs de résistance à des antibiotiques d'intérêt en clinique humaine tels les bêta-lactamines, les aminosides et les sulfamides (Branger et al., 2019). Parmi les résistances antibiotiques ayant récemment émergé, on peut également citer la résistance plasmidique à la colistine médiée par les gènes *mcr* (Liu et al., 2016). Cette résistance à un antibiotique de dernière ligne a affolé la communauté scientifique devant la menace potentielle qu'elle représente. Néanmoins elle semble relativement rare chez *E. coli* en France, et associée à de faibles niveaux de résistance (Bourrel et al., 2019). Elle est liée préférentiellement à certains plasmides conjugatifs des groupes IncHI2, IncI2 et IncX4, les deux premiers étant associés aux isolats européens et asiatiques, respectivement (Matamoros et al., 2017).

En association ou non avec ces déterminants de résistance, les plasmides sont aussi de formidables vecteurs de facteurs de virulence (T. J. Johnson & Nolan, 2009). Parmi eux, les plasmides isolés de souches NMEC et APEC ont été particulièrement étudiés (T. J. Johnson et al., 2006; Peigne et al., 2009). C'est le cas de pS88, un plasmide de plus de 130000 pb de groupe IncF porteur 3 systèmes de capture du fer (aérobactine, salmocheline, et gènes *sitABCD*) ainsi que d'autres facteurs de virulence (*iss*, *etsABC*, *ompT*, et *hlyF*) et 2 colicines (Figure 14) (Peigne et al., 2009). La structure de ce plasmide présente de nombreux points communs avec d'autres plasmides isolés de souches APEC. On peut identifier dans ces éléments génétiques une région de virulence conservée et une autre plus variable, ainsi qu'un ensemble de gènes impliqués dans la réplication, la maintenance et la stabilité des plasmides. Les auteurs ont par ailleurs identifié des plasmides semblables chez des souches de phylogroupe B2 isolées au cours d'urosepsis chez l'Homme. Certains groupes de gènes présents sur ces plasmides APEC et NMEC peuvent également être trouvés au niveau chromosomique chez des souches UPEC. C'est le cas par exemple du PAI III_{536} de la souche UPEC 536 qui contient la salmocheline, le gène *tsh* codant une hémagglutinine et un remnant de l'opéron codant la colicine ColV (T. J. Johnson et al., 2006). La présence de gènes de virulence au niveau plasmidique pourrait donc être une étape intermédiaire dans le développement d'îlots de pathogénicité au niveau chromosomique fréquemment retrouvés chez les souches UPEC par exemple. En plus des gènes de virulence ExPEC classiques, sont retrouvés également de nombreux gènes codant des bactériocines, notamment des colicines, sur les plasmides mobilisables MOB_{RNA} avec une association forte entre le type de plasmide et le type de colicine (Branger et al., 2019).

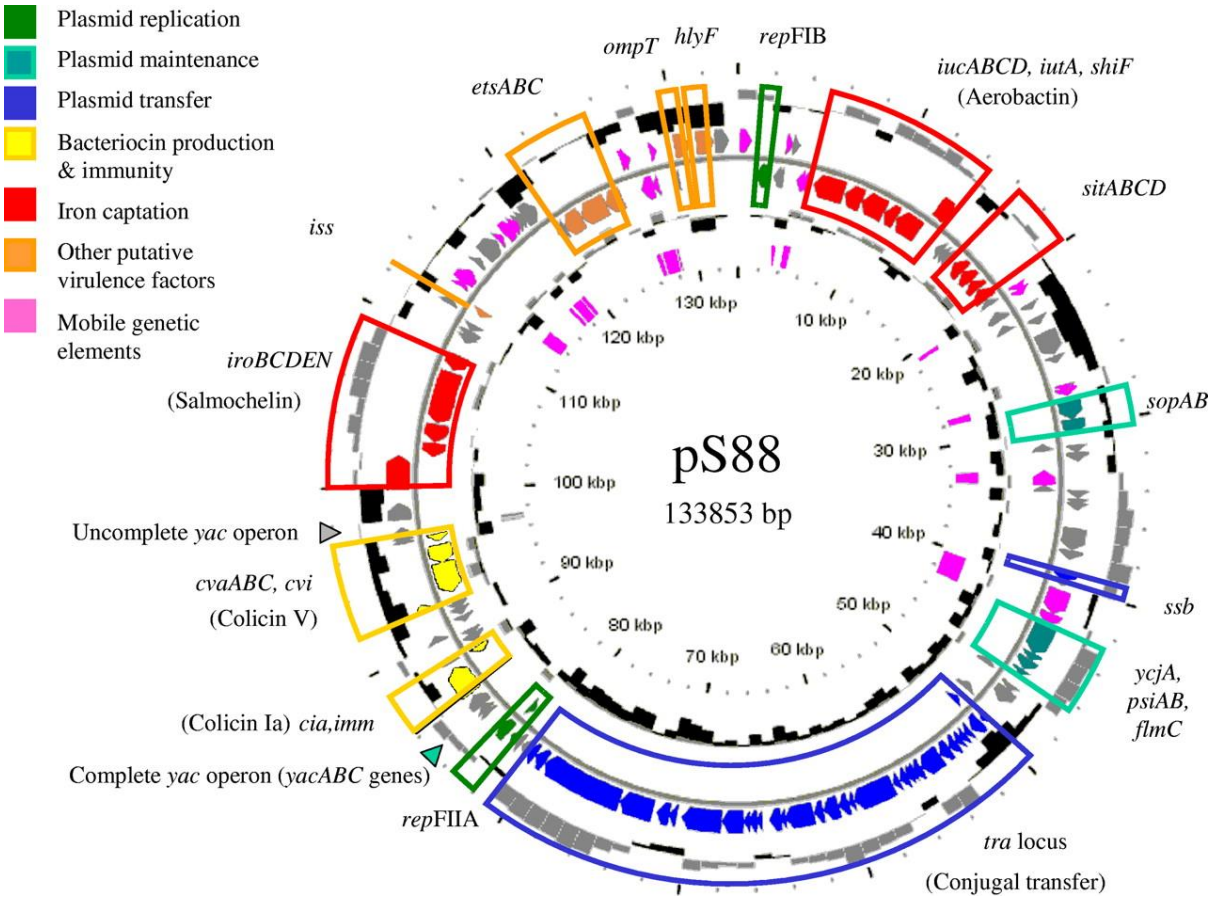


Figure 14. Représentation de la carte génétique du plasmide pS88 isolé de la souche *E. coli* S88. Les cercles indiquent de l'extérieur vers l'intérieur : i) les variations du contenu en GC, ii) les gènes prédits sur le brin direct, iii) les gènes prédits sur le brin complémentaire, iv), la mesure de la proportion de bases G comparées aux bases C (GC skew), v) les éléments transposables en rose, vi) les coordonnées en kpb par rapport à l'origine de réplication. Les gènes d'intérêt sont colorés selon leur fonction (d'après Peigne et al., 2009).

5. ExPEC et modèle animal

Le modèle animal a grandement participé à la caractérisation des souches ExPEC. En effet, le caractère virulent ou non d'une souche ne peut pas toujours être déduit des circonstances dans lesquelles la souche a été isolée, certaines souches ExPEC étant parfois isolées de selles de patients sains. Il est donc apparu indispensable de disposer d'un modèle standardisé pour l'évaluation de la virulence. Le modèle murin par injection sous-cutanée de bactéries s'est très vite imposé comme l'une des références. Ce modèle présente l'avantage d'être reproductible et d'offrir une distribution généralement bimodale des souches (J. R. Johnson et al., 2018), apparaissant soit "non tueuses" (0 - 1 souris sur 10) soit "tueuses" (9 ou 10 souris sur 10). Bien qu'il ne permette pas d'analyser les capacités de certaines souches à utiliser certaines portes d'entrée pour réaliser une bactériémie, il permet d'évaluer d'autres aptitudes essentielles comme la résistance à la phagocytose et au complément, ou encore la chélation du fer.

A l'aide d'un modèle murin, Picard *et al.* ont pu mettre en évidence le lien étroit entre virulence et phylogénie (Picard et al., 1999). Ainsi, les souches de groupes B2, et dans une moindre mesure D, tendent à avoir plus de facteurs de virulence et être plus pathogènes dans le modèle murin (Figure 15). A l'opposé, les souches des groupes A et B1 sont généralement inoffensives dans ce modèle et possèdent peu de facteurs de virulence. Par ailleurs, ces résultats sont également en partie corrélés à l'origine clinique (urine, hémoculture) ou commensale (fèces) des souches. Ce lien entre virulence et phylogénie/fond génétique fait suspecter une interdépendance entre le pouvoir pathogène des souches et des capacités métaboliques particulières.

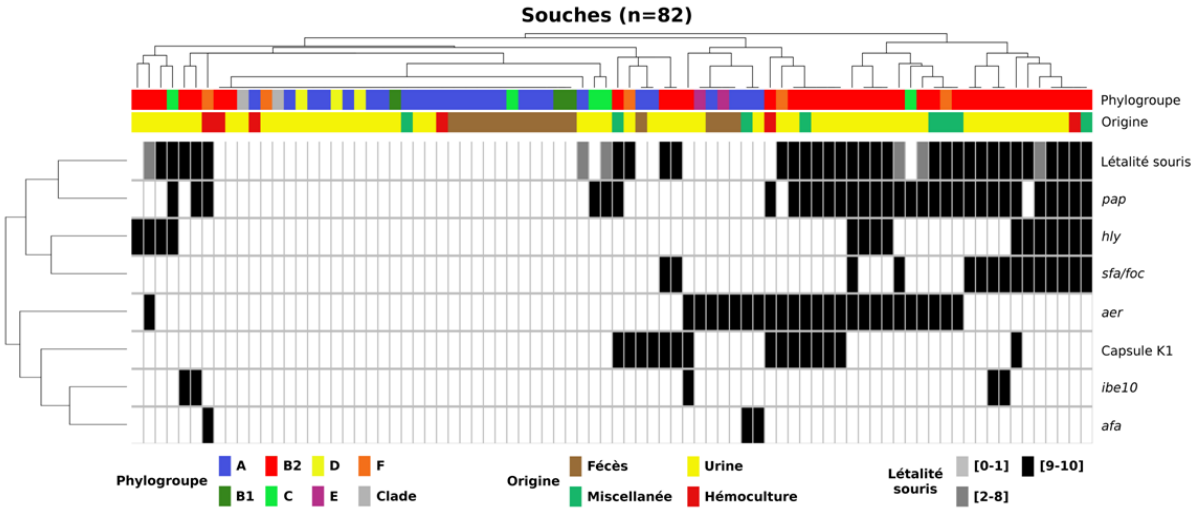


Figure 15. Heatmap à partir des distances euclidiennes calculées en fonction des présences/absences de déterminants génétiques et de la létalité dans un modèle murin (nombre de souris tuées). Le phylogroupe et l'origine des souches est représenté par des rectangles de couleur. Construit à partir des données de Picard *et al.* (Picard et al., 1999).

D'autres modèles animaux ont été utilisés. Parmi eux, le modèle du nématode *Caenorhabditis elegans* offre des résultats proches de ceux du modèle murin et pourrait donc être une alternative moins coûteuse et plus simple à mettre en œuvre(Diard et al., 2007). En utilisant en partie les souches de l'étude de Picard *et al.*, Diard *et al.* ont mis en évidence une association entre la virulence chez les nématodes et : la virulence chez la souris, le phylogroupe B2, certains sérotypes fréquents chez les ExPEC (O2, O6, O18), la résistance au sérum, à la bile et au lysozyme/lactoferrin, la mobilité, la vitesse de croissance rapide et la présence de certains facteurs de virulence d'un autre côté (Diard et al., 2007). Néanmoins, bien que la pathogénicité des souches soit corrélée entre ce modèle et le modèle murin, les facteurs de virulence impliqués sont différents. Le HPI, par exemple, bien que facteur essentiel dans le modèle murin n'apparaît pas associé au décès chez *Caenorhabditis elegans* (Diard et al., 2007; Galardini et al., 2020; Schubert et al., 2002).

Inversement, dans un autre modèle, *Dictyostelium discoideum,* le HPI joue un rôle primordial dans la résistance des bactéries à la prédation par l'amibe (Adiba et al., 2010). Ce modèle de protozoaire permet, lui aussi, d'évaluer la virulence des souches ExPEC. A l'aide de ce modèle, les auteurs ont mis en évidence le rôle potentiel de la virulence des souches dans l'adaptation à une niche écologique. En effet, cette amibe mime en partie les conditions rencontrées au sein des macrophages humains et les auteurs font donc l'hypothèse que les facteurs de virulence impliqués dans la pathogénicité chez l'Homme sont en fait probablement sélectionnés pour leur rôle adaptatif dans d'autres environnements ou pour d'autres fonctions,

comme la résistance à la prédation par les amibes par exemple. On retrouve ici la théorie selon laquelle la virulence serait en réalité un produit dérivé du commensalisme chez *E. coli* (Adiba et al., 2010; Le Gall et al., 2007).

Comme abordé dans le chapitre suivant, le développement des méthodes de séquençage au cours des dernières décennies a permis d'analyser de manière plus exhaustive la génétique des souches de *E. coli*. Si les premières études n'incluaient qu'une sélection de gènes de virulence, du fait de limitations techniques, grâce au séquençage de génomes complets, il est désormais possible d'utiliser la totalité de l'information génétique. L'intégration des données de modèle animal au cours d'analyses de génomes complets permet parfois de mettre en lumière l'importance majeure de certains déterminants bactériens dans la virulence. Galardini *et al.* ont ainsi analysé les données de mortalité dans le modèle murin pour 370 souches (commensales, pathogènes et environnementales) en regard des données génomiques (Galardini et al., 2020). Par une étude d'association à l'échelle du génome (genome-wide association studies : GWAS), les auteurs ont identifié une association entre certains systèmes de capture du fer et le décès chez la souris, plus particulièrement le HPI et dans une moindre mesure l'aérobactine, et SitABCD.

Les modèles animaux apportent donc une validation phénotypique dans la classification des souches ExPEC et peuvent souligner l'importance de certains facteurs de virulence. Néanmoins, la virulence mesurée par ces approches n'est pas corrélée à la sévérité de l'infection chez l'Homme et doit donc uniquement être considérée comme une virulence "intrinsèque" (Landraud et al., 2013). Cela s'explique notamment par la part primordiale jouée par le statut immunitaire et les comorbidités de l'hôte dans l'issue de l'infection (Lefort et al., 2011). Le modèle animal semble plutôt mettre en évidence le caractère pathogène spécialisé de certaines souches de *E. coli* (B2, D, F), sans lien avec la sévérité et par opposition aux souches d'allure commensale (A et B1).

V. Etudes comparées à grande échelle de *Escherichia coli* responsables de bactériémies

1. L'ère pré-génomique

Des études à large échelle ont tenté d'identifier des facteurs bactériens spécifiques chez les souches de *E. coli* responsables de bactériémies. L'utilisation des techniques de MLST et la détermination des phylogroupes a montré que ces souches n'étaient pas regroupées au sein d'une entité unique mais bien distribuées dans toute la population (Jauréguy et al., 2008). Pour autant, il existe tout de même des associations préférentielles. Les souches de phylogroupe B2 représentent généralement la majorité des isolats, suivie des groupes D et A (Jauréguy et al., 2007, 2008; Lefort et al., 2011). L'analyse des ST met en évidence une pauci-clonalité avec certains complexes clonaux particulièrement prévalents et dans lesquels les gènes de virulence ExPEC sont très fréquents (Jauréguy et al., 2007; Lefort et al., 2011; Mora-Rillo et al., 2015). La porte d'entrée de la bactériémie est également en lien étroit avec l'épidémiologie, et l'on observe plus de souches de phylogroupe B2 et plus de facteurs de virulence dans les isolats de bactériémies à point de départ urinaire, alors que les souches de phylogroupe B1 sont parfois associées à la porte d'entrée digestive (Jauréguy et al., 2007; Lefort et al., 2011; Martinez et al., 2006). Certaines études décrivent même cela sous la forme d'un continuum avec i) des souches de phylogroupe B2 hautement virulentes dans les bactériémies à point de départ urinaire ou pulmonaire, ii) des souches de virulence intermédiaire pour les autres portes d'entrée, iii) des souches peu virulentes et plus volontiers de phylogroupes A et B1 dans les fèces (J. R. Johnson et al., 2002).

Lorsqu'ils sont pris en compte dans les analyses, les facteurs associés à l'hôte apparaissent largement impliqués dans le pronostic, comme par exemple l'âge du patient, les comorbidités et notamment les pathologies malignes et l'immunodépression, ou encore une porte d'entrée non urinaire de l'infection (Chung et al., 2012; Hekker et al., 2000; K.B. Laupland et al., 2008; Lefort et al., 2011; Martinez et al., 2006). De même, dans l'étude de Mora-Rillo *et al.,* qui inclut une proportion importante de patients présentant des pathologies malignes (50%), la mortalité observée est particulièrement élevée (27,7%) (Mora-Rillo et al., 2015). Les auteurs considèrent par ailleurs que dans plus de 60% des cas le décès n'est pas attribuable directement à la bactériémie. Parmi les facteurs de bon pronostic, la porte d'entrée urinaire ressort très fréquemment ainsi que le caractère communautaire de l'infection (Chung et al., 2012; K.B. Laupland et al., 2008; Lefort et al., 2011; Mora-Rillo et al., 2015).

Certaines études ont analysé conjointement des données cliniques et bactériologiques exhaustives à la recherche d'association avec la sévérité des infections au sein d'une population homogène et suffisamment grande. Dans leur étude en 2011 incluant 1051 patients adultes bactériémiques, Lefort *et al.* ont analysé les souches en termes de phylogroupes, gènes de virulence et résistance. Le poids des déterminants bactériens était alors faible, et seule la présence du gène *ireA*, impliqué dans la capture du fer et associé préférentiellement aux bactériémies à point de départ urinaire, ressortait associé à un bon pronostic (Lefort et al., 2011)**.** Au sein d'une population de taille plus faible (120 patients) et plus comorbide, Mora-Rillo *et al.* observent également un facteur de virulence de bon pronostic, les P fimbriae, eux aussi plutôt associés au tractus urinaire (Mora-Rillo et al., 2015). Par ailleurs, les résultats peuvent parfois sembler contradictoires, comme dans l'étude de Hekker *et al.* dans laquelle l'absence d'hémolysine est associée à la morbidité et à la mortalité chez les patients présentant des facteurs de risques (Hekker et al., 2000). Mais cela révèle ici à nouveau plutôt des facteurs liés à l'hôte, avec des patients plus fragiles et donc plus sensibles à l'infection par des souches pourtant faiblement virulentes.

De manière assez surprenante, le fait d'avoir une antibiothérapie inadaptée à la sensibilité de la bactérie isolée de l'hémoculture ne ressort généralement pas associé à la mortalité (Jauréguy et al., 2007; Martinez et al., 2006). Cependant, les taux de résistance aux antibiotiques à large spectre comme les C3G sont relativement faibles, bien souvent inférieures à 5%, pour nombre de ces études datant de la première décennie du 21ème siècle (Jauréguy et al., 2007; Lefort et al., 2011; Martinez et al., 2006; Mora-Rillo et al., 2015). Chung *et al.* se sont eux concentrés sur l'analyse de souches productrices de BLSE dans les bactériémies à Taiwan de 2005 à 2010, notamment pour évaluer l'impact clinique du clone ST131 (Chung et al., 2012). Bien qu'il représente presque le tiers de souches BLSE et qu'il soit décrit comme un clone plutôt virulent, le ST131 n'est pas associé à une mortalité plus élevée et est retrouvé essentiellement dans les bactériémies communautaires et à porte d'entrée urinaire.

De manière plus anecdotique, des auteurs ont évalué la capacité de formation de biofilm des souches bactériémiques en fonction des données cliniques. Mais, ils n'ont trouvé aucune association particulière pour expliquer la sévérité de l'infection ni de lien avec un cathétérisme veineux ou des voies urinaires (Martinez et al., 2006).

2. L'utilisation du génome complet

La démocratisation et la diminution du coût du séquençage de génomes complets, depuis la deuxième décennie du 21ème siècle, permet d'envisager des analyses de souches à grande échelle avec une exhaustivité jusqu'alors jamais atteinte. Il devient ainsi possible d'analyser conjointement virulence et résistance génotypique sans limite dans le nombre de gènes recherchés, ainsi que la structure de la population avec une granularité bien plus fine. Ces évolutions technologiques sont d'autant plus intéressantes avec l'expansion mondiale de certains clones pandémiques ExPEC, comme le ST131, et l'augmentation massive de la résistance antibiotique (Nicolas-Chanoine et al., 2014). Elles pourraient permettre d'identifier de nouveaux déterminants bactériens.

Une des premières études sur un échantillon de taille conséquente a été réalisée par Salipante *et al.* (Salipante et al., 2015). Par l'analyse de plus de 300 génomes de souches de *E. coli* isolées d'infections urinaires et d'hémocultures, ils ont pu décrire la diversité de la population de manière plus précise. Ils ont retrouvé la prédominance des phylogroupes B2 et D, ainsi que la pauciclonalité au cours de ces infections extra-intestinales, ST131 et ST95 en tête avec 16,1% et 10,8% des isolats, respectivement. Le calcul du coregénome et du pangenome révèlent des tailles respectivement plus grande et plus petite que celles obtenues par Kaas *et al.* sur une collection de 186 génomes de *E. coli* de toute origine (*i.e.* clinique ou non) (Kaas et al., 2012). Ces résultats illustrent d'une autre manière la structure clonale exacerbée de la population ExPEC, notamment au sein des phylogroupes B2 et D. Grâce à la sensibilité offerte par le génome complet, ils confirment que la transmission entre patients n'est pas un phénomène fréquent au cours de ces infections, même si certains clones dominants pourraient diffuser à plus large échelle au sein d'une communauté donnée. Enfin, la comparaison des présences/absences de gènes peut permettre par des études d'association d'identifier des gènes associés à un trait particulier. Ce type de méthode s'est avéré efficace et les auteurs ont mis en évidence les gènes impliqués dans la résistance antibiotique ainsi que les nombreux éléments mobiles associés (transposases, gènes impliqués dans le transfert et la maintenance des plasmides). Les résultats de cette étude d'association n'ont pas permis d'identifier de nouveaux déterminants, mais doivent plutôt être vus comme une démonstration de faisabilité et un encouragement à l'appliquer sur d'autres variables.

Sur un échantillon encore plus important et composé seulement de souches isolées de bactériémies, Kallonen *et al.* observent les mêmes tendances dans la composition de la

population en termes de phylogroupes et de ST (Kallonen et al., 2017). Les analyses phylogénétiques réalisées soulignent particulièrement bien la différence de structure de population du phylogroupe B2, faite de larges expansions clonales, comparée à celles des A et B1 présentant des longueurs de branche importantes. De plus, leur étude est réalisée sur 11 ans et permet ainsi d'observer l'émergence de deux clones ExPEC majeurs, les ST69 et ST131 cités précédemment pour leur caractère multirésistant (Figure 16). Disposant des génomes de toutes ces souches, les auteurs parviennent à dater la probable émergence des sous-populations présentes au sein de ces 2 ST par des approches phylogénétiques. L'analyse d'un grand nombre de facteurs de virulence (plus de 3500) leur permet également d'être exhaustif et de rechercher d'éventuelles associations avec les principaux clones. En accord avec le caractère probablement multigénique de la virulence ExPEC, ils n'identifient que peu de gènes spécifiquement associés au ST131. Le plus spécifique est le gène *espC* codant un autotransporteur sérine protéase, relativement éloigné d'un point de vue génétique par rapport à ceux observés par exemple chez les souches ExPEC.

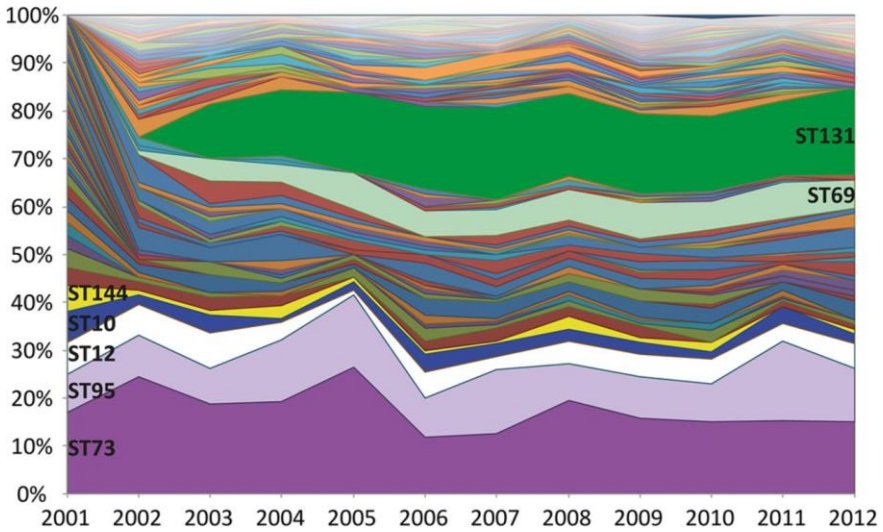


Figure 16. Distribution des différents STs responsables de bactériémies en Angleterre sur une période 11 ans. Les STs sont présentés dans leur ordre de fréquence au début de l'étude (d'après Kallonen et al., 2017).

Cependant il est peu probable que ce gène explique à lui seul l'émergence du ST131. En effet, il n'est retrouvé que dans une des 39 souches de ST131 analysées par Goswami *et al*., et il s'agit donc probablement d'une spécificité locale ou temporelle (Goswami et al., 2018). Ces auteurs ont, eux, inclus des données cliniques et démographiques dans l'analyse de 162 génomes de souches isolées d'hémoculture en Ecosse entre 2013 et 2015. Par une étude d'association, ils identifient un système toxine/antitoxine chromosomique particulièrement

présent chez les souches responsables de bactériémies nosocomiales et ayant un rôle potentiel dans les phénomènes de persistance. Cet état de dormance permet aux bactéries de diminuer leur sensibilité à de nombreux antibiotiques dont l'activité est conditionnée par la réplication des bactéries. Il est donc possible que ce système ait une utilité pour la survie dans le milieu hospitalier où la pression antibiotique est particulièrement forte. Concernant l'issue de l'infection, tout comme l'étude de Lefort *et al.*, aucun déterminant bactérien ne ressort associé positivement au décès (Lefort et al., 2011).

La difficulté voire l'impossibilité de mettre en évidence des associations de gènes de virulence, ou d'autres gènes accessoires, avec l'ensemble des principaux clones isolés de bactériémies suggèrent que ces différents clones existent probablement au sein de niches écologiques différentes (Goswami et al., 2018). L'étude du métabolisme pourrait donc être une autre piste intéressante pour expliquer l'émergence et la prédominance de certaines souches. De même, l'analyse des communautés microbiennes par des approches de métagénomiques pourrait permettre d'identifier des associations préférentielles entre certains microorganismes.

VI. Implication du métabolisme dans la pathogénicité extra-intestinale de *Escherichia coli*

Comme évoqué précédemment, la virulence extra-intestinale est un phénomène complexe et bien que certains facteurs de virulence classiques jouent un rôle dans la physiopathologie de l'infection, aucun profil de virulence unique n'est associé aux bactériémies ou infections urinaires. Il est donc possible que d'autres déterminants bactériens soient associées au processus infectieux. Parmi eux, les gènes impliqués dans le métabolisme pourraient jouer un rôle essentiel dans la colonisation et la survie des souches. On associe au terme métabolisme l'ensemble des processus biochimiques de dégradation, de synthèse de molécules biologiques et de production d'énergie chimique. Comme le soulignent Touchon *et al.*, celui-ci pourrait être impliqué dans l'adaptation à une niche particulière, et une large partie des gènes dont les fonctions sont connues est d'ailleurs reliée au métabolisme (Touchon et al., 2009). Chez *E. coli,* la prévalence et l'abondance des différents phylogroupes est en partie corrélée avec le régime alimentaire de l'hôte (Gordon & Cowling, 2003) et il est donc fort probable que ceci soit la conséquence de capacités métaboliques spécifiques nécessaires à la colonisation d'une niche écologique.

D'après la théorie de Freter, *n* populations bactériennes peuvent cohabiter, à condition qu'il y ait dans l'écosystème *n* substrats différents (Freter et al., 1983). Ainsi, pour que les souches uropathogènes puissent coloniser le tube digestif, première étape nécessaire au processus infectieux, elles doivent soit entrer en compétition avec les souches commensales pour un nutriment, soit utiliser des nutriments différents. Il existe de plus une certaine dynamique dans l'utilisation de ces substrats, dont certains sont préférentiellement utilisés dans les phases de colonisation et d'autres dans le maintien de la population dans le tube digestif (Chang et al., 2004). De même, pour assurer une croissance dans les urines, les souches UPEC doivent avoir les capacités pour survivre dans cet environnement par l'utilisation de voies métaboliques différentes et adaptées aux nutriments disponibles (Figure 17). Au cours de la colonisation intestinale, les souches peuvent utiliser les sucres présents en grande quantité et résultant de la dégradation de polysaccharides complexes par la flore anaérobie résidente. Mais une fois présentes dans le tractus urinaire, elles font face à un milieu très différent, pauvre en fer et en sucre, et contenant certains acides aminés (Alteri & Mobley, 2015). La capacité de croissance élevée et rapide dans des milieux spécifiques comme les urines, essentielle au pouvoir pathogène, peut en soi être considérée comme un facteur de virulence (Alteri & Mobley, 2015; Reitzer & Zimmern, 2019). D'un point de vue plus général, les souches pathogènes disposent de deux grandes stratégies pour entrer en compétition avec le

microbiote digestif résident et/ou coloniser des sites extra-intestinaux : i) adapter leur métabolisme aux changements environnementaux par des phénomènes de régulation des voies métaboliques; ii) acquérir des voies métaboliques supplémentaires et ainsi augmenter le répertoire des nutriments utilisables (Le Bouguénec & Schouler, 2011).

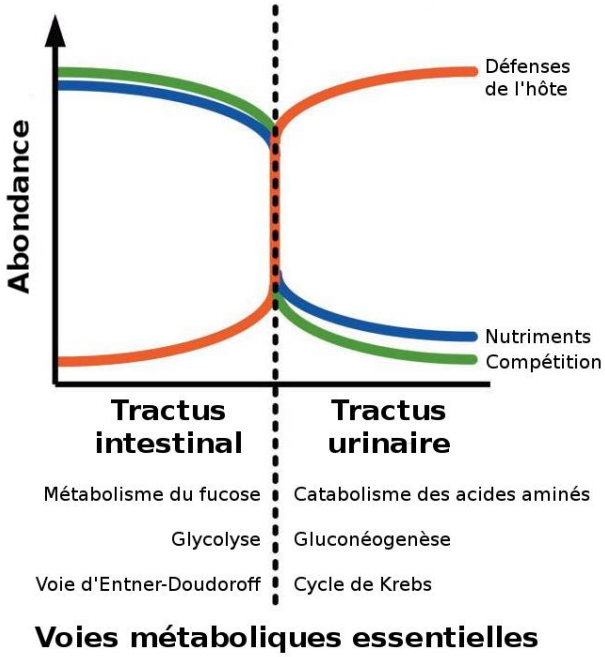


Figure 17. Adaptation du métabolisme de *E. coli* en fonction des nutriments disponibles, de la pression exercée par l'hôte et de la nécessité de concurrencer les éventuels micro-organismes présents. Adapté de Alteri & Mobley, 2015.

1. Analyse ciblée du métabolisme

Plusieurs exemples dans la littérature ont rapporté une implication du métabolisme dans le pouvoir de colonisation ou la virulence des souches ExPEC, notamment chez les souches uropathogènes, par des approches ciblées sur une voie particulière (Reitzer & Zimmern, 2019).

L'une d'entre elles s'est intéressée au métabolisme de la D-sérine, un acide aminé particulièrement abondant dans les urines (Anfora et al., 2007; Reitzer & Zimmern, 2019). Cet acide aminé présente une toxicité pour la plupart des bactéries, avec un effet bactériostatique. Mais la plupart des souches UPEC responsables de pyélonéphrites et d'urosepsis, comme la souche CFT073 (B2, ST73), possèdent une voie de dégradation de la D-sérine, codée par le

groupe de gènes *dsdACX*. Outre la possibilité d'utiliser ce métabolite comme source de carbone et de nitrogène, il semble que son accumulation intracellulaire soit un signal favorisant la colonisation du tractus urinaire dans un modèle d'infection urinaire de souris. Cette accumulation semble associée à l'expression de certains facteurs de virulence comme les fimbriae de type P et F1C ou encore l'hémolysine, mettant en évidence un lien étroit en métabolisme et virulence (Anfora et al., 2007).

Une autre étude s'est intéressée à une voie métabolique parfois colocalisée avec des déterminants de résistance antibiotique. Plus précisément, la voie de l'arginine déiminase a été plus fréquemment observée chez les souches de *E. coli* porteuses de BLSE, essentiellement de type CTX-M (Billard-Pomares et al., 2019). L'opéron *arc* codant cette voie métabolique est effectivement souvent localisé sur des plasmides de type IncF porteurs également du gène codant la BLSE, et est flanqué de séquences d'insertion suggérant une mobilité accrue. Outre son association à des gènes de résistance, les auteurs ont montré son rôle dans le processus infectieux urinaire dans un modèle de souris, dans lequel elle améliore significativement la compétitivité de la souche le possédant. En revanche, de manière intéressante l'avantage procuré par cet opéron est limité à la colonisation du tractus urinaire et devient un poids pour la colonisation du tube digestif chez la souris. Bien que cet opéron à lui seul ne permette pas d'expliquer la diffusion de la résistance antibiotique liée aux BLSE, il s'agit d'un exemple de co-sélection entre métabolisme et résistance avec un impact important sur la virulence des souches.

Chez certaines souches APEC, la présence d'une voie métabolique additionnelle impliquée dans le métabolisme des fructooligosaccharides semble associée à un pouvoir de colonisation accru au niveau de digestif (Schouler et al., 2009). Ce locus comprenant les gènes *fos* est situé au sein d'un îlot génomique flanqué de séquences répétées et inséré au niveau de l'ARNt *selC*, témoignant de la probable mobilité de ce déterminant (Chouikha et al., 2006). La souche APEC dans laquelle cet îlot a été décrit, BEN2908, appartient au phylogroupe B2 et au ST95, l'un des quatre grands clones ExPEC observés actuellement chez l'Homme (Denamur et al., 2021). L'utilisation des fructooligosaccharides comme prébiotique chez l'animal et humain pourrait donc favoriser la sélection de bactéries commensales bénéfiques mais également de souches ExPEC capables de métaboliser de telles substances.

Bien qu'associées à des données expérimentales robustes, nombre de ces études sont réalisées sur un nombre limité de souches, notamment la souche CFT073, un des archétypes de souches responsables de pyélonéphrite (Reitzer & Zimmern, 2019).

2. Analyse des réseaux métaboliques

Une approche plus exploratoire consiste à déterminer l'ensemble des voies métaboliques présentes au sein des souches à partir de l'analyse de leur génome afin, par la suite, d'établir des comparaisons selon des critères de modes de vie ou de phylogénie par exemple. Vieira *et al.* ont réalisé ce genre d'analyses sur 29 souches de *E. coli* comprenant à la fois des génomes de souches commensales, InPEC, ExPEC et de *Shigella* (Vieira et al., 2011). Les résultats obtenus offrent une vision précise du pan- et coremétabolisme chez *E. coli* avec un total de 1545 réactions dont 885 sont conservées. Le nombre de réactions dans le coremétabolisme augmente fortement si les souches de *Shigella* ne sont pas prises en compte, en accord avec leur évolution vers un mode de vie parasite associé à la perte de nombreux gènes et voies métaboliques (Maurelli et al., 1998; Touchon et al., 2009). Ce qui est assez frappant dans l'étude Vieira *et al.*, c'est la faible diversité du métabolisme observée en comparaison à celle observée au sein du pangénome de *E. coli*, qui dans leur étude comprend dont 14986 familles de gènes dont seulement 1957 dans le coregénome (Figure 18). Le panmétabolisme semble atteindre un plateau avec seulement 29 souches ce qui pourrait refléter la nature assez conservée des gènes codant les enzymes, ou bien pourrait être un artéfact lié la difficulté de mettre en évidence de nouvelles fonctions métaboliques par l'annotation des génomes. L'analyse du protéome de plusieurs souches de *E. coli* confirme plutôt l'hypothèse d'un certain niveau de conservation dans le métabolisme (Sabarly et al., 2016).

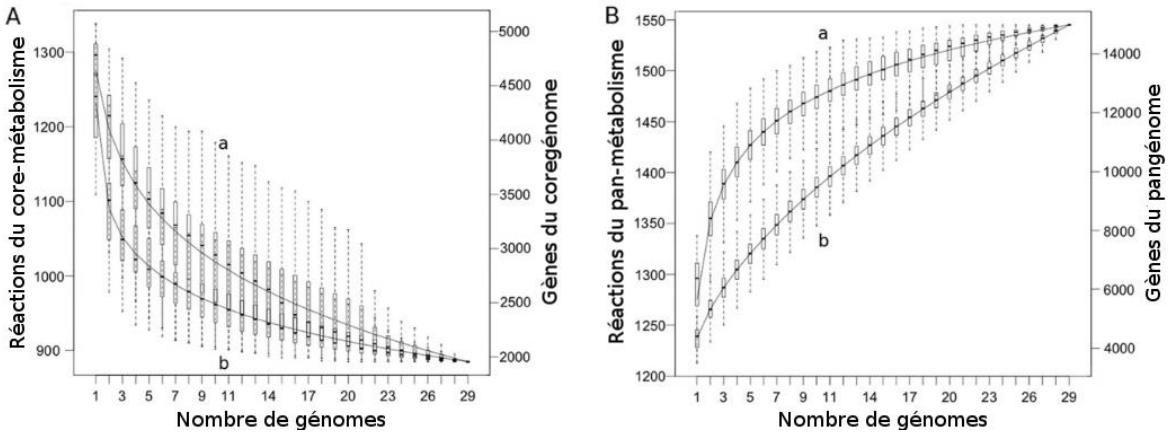


Figure 18. A) Evolution du coremétabolisme de 29 souches de *E. coli* (a) et du coregénome (b). B) Evolution du panmétabolisme (a) et du pangénome (b). Adapté de Vieira et al., 2011.

Dans cette même étude, les distances métaboliques entre les différentes souches étudiées n'apparaissent pas liées à leur mode de vie. Elles sont en réalité fortement associées à la

phylogénie avec des phylogroupes B2, D, E et F bien distincts et des souches des groupes A et B1 plus proches comme en témoignent également les analyses phylogénétiques et les analyses en composantes multiples (Figure 19). Parmi les rares éléments associés aux ExPEC, les auteurs retrouvent l'absence de voies de dégradation de la psicose et psicoselysine, et la présence du gène *kpsT* codant un transporteur de polysaccharide capsulaire et donc associé à un facteur de virulence ExPEC reconnu.

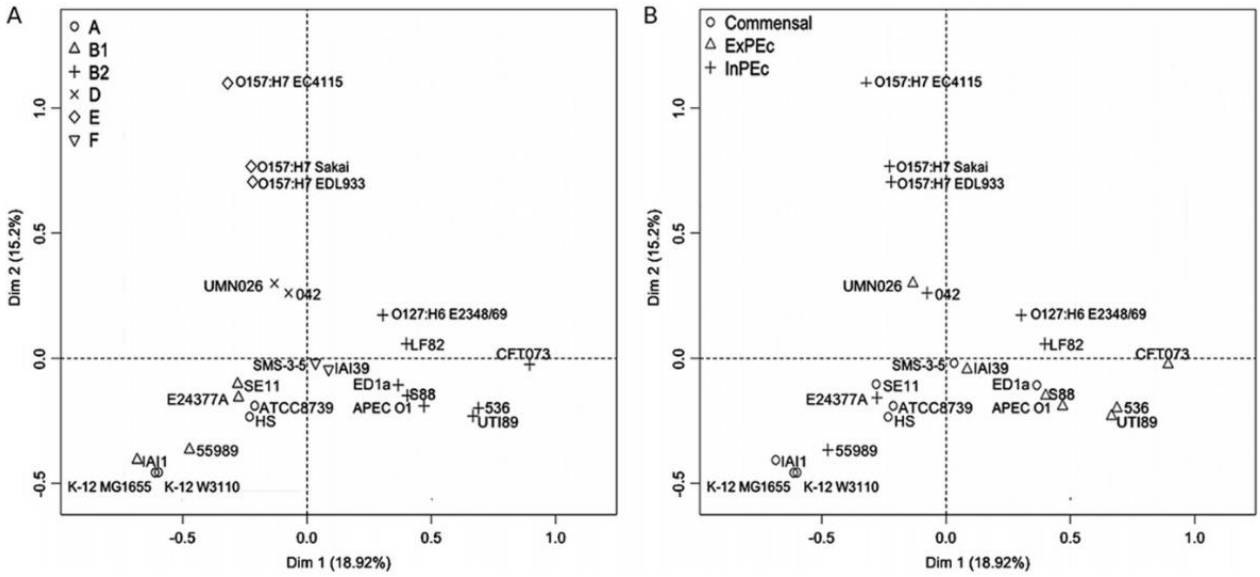


Figure 19. Analyse des correspondances multiples à partir des occurrences des réactions dans des souches de *E. coli* désignées par leur groupe phylogénétique (A) et leur mode de vie (B). (d'après Vieira et al., 2011).

Les différences de mode de vie ne ressortent pas associées à des profils métaboliques particuliers, mais néanmoins on ne peut exclure qu'un nombre limité d'enzymes soit tout de même impliqué. Si tel est le cas, le nombre de souches analysées ne permet probablement pas d'apporter cette réponse.

Sabarly *et al.* ont ajouté dans leur étude des données phénotypiques sur l'utilisation de différentes sources carbonées, et essayé eux aussi d'identifier des associations avec des modes de vie particuliers (caractères anthropogéniques, régime alimentaire, pathogénicité) (Sabarly et al., 2011a). Mais les données phénotypiques ne sont pas non plus liées directement à ces modes de vie. Par ailleurs la corrélation entre données phénotypiques et complétion des voies métaboliques est très faible, probablement en raison d'un rôle important des systèmes de régulation qui viennent brouiller le signal.

Enfin, il est possible de produire des modèles métaboliques afin de simuler la croissance en fonction de la présence ou non de certains nutriments (Gu et al., 2019). C'est ce qu'ont proposé Monk *et al.* sur 55 souches de *E. coli* et de *Shigella.* Ils ont pu mettre en évidence des différences dans le catabolisme de certains composés entre les souches commensales d'une part et les InPEC ou les ExPEC d'autres part (Monk et al., 2013). Les auteurs retrouvent notamment une incapacité fréquente des souches ExPEC à dégrader les composés aromatiques pour assurer leur croissance. Ces données ont été suggérées à l'échelle du pangénome à plusieurs reprises, en raison de l'absence des voies métaboliques correspondantes dans les souches du phylogroupe B2, d'où proviennent de nombreux ExPEC (Touchon et al., 2009, 2020). En revanche, Monk *et al.* n'ont pas pris en compte la structure de la population de *E. coli* dans leurs analyses (phylogroupes ou ST par exemple), rendant parfois difficilement généralisables ces observations sur les liens entre mode de vie et métabolisme.

Nous n'utiliserons pas ces modèles mathématiques au cours de cette thèse, et focaliserons la comparaison des réseaux métaboliques inférés à partir des données de génomiques.

Objectifs de la thèse

Comme présenté au cours de cet état de l'art, les bactériémies sont des pathologies sévères engageant le pronostic vital dans 5 à 30% des cas. *E. coli* est à l'heure actuelle l'agent bactérien majoritairement impliqué dans les pays industrialisés. L'analyse des souches responsables d'infections extra-intestinales a permis de mettre en évidence au cours des dernières années l'apparition de certains clones prédominants, parfois associés à une résistance antibiotique à des molécules de haute importance en clinique, comme les C3G et les fluoroquinolones. Ce constat a en partie motivé l'étude Septicoli[5] (projet financé par l'Agence Nationale de la Recherche, identifiant ANR-15-CE17-0019), portée par le Pr. Agnès Lefort, dont le but était d'étudier les facteurs pronostiques des septicémies à *E. coli* face à l'émergence de clones multirésistants, et les implications thérapeutiques qui en découlent. Cette étude prospective multicentrique s'est déroulée entre octobre 2016 et juillet 2017 au sein de sept centres hospitaliers universitaires de région parisienne. Elle a consisté à inclure sur cette période l'ensemble des épisodes de bactériémies à *E. coli* de l'adulte, avec d'une part l'analyse du phénotype de résistance et le séquençage de la souche bactérienne isolée de l'hémoculture, et d'autre part un recueil de données cliniques grâce à une visite au moment de l'inclusion puis une seconde à la sortie de l'hôpital ou à 28 jours. Cette thèse est intimement liée à ce projet puisqu'elle tire profit des 545 souches recueillies et séquencées au cours de l'étude, et consiste en l'analyse des génomes de ces souches. En effet, l'utilisation des méthodes de séquençage de génomes complets est une autre raison à la mise en place de l'étude Septicoli. Dans le cadre des changements épidémiologiques observés, l'avènement du séquençage haut débit offre la possibilité d'étudier un grand nombre d'isolats de manière concomitante, et ce avec une granularité jusqu'ici rarement atteinte. Il devient possible de déterminer de manière précise la structure de cette population de souches responsables de bactériémies de l'adulte, tout en prenant en compte l'ensemble des facteurs de virulence décrits jusqu'ici ainsi que les déterminants génétiques de la résistance antibiotiques et leur éventuel support. L'objectif global de cette thèse est de montrer l'apport de la génomique comparée dans la compréhension de la physiopathologie des infections extraintestinales à *E. coli*.

Dans cette optique, cette thèse comporte trois chapitres afin de tirer profit de cette quantité d'informations offerte par les séquences génomiques.

[5] https://anr.fr/Projet-ANR-15-CE17-0019

- Le premier objectif est essentiellement méthodologique et consiste à choisir les outils bioinformatiques adaptés à l'étude de ces génomes et de proposer un schéma d'analyse complet. Au cours de ce premier chapitre, je décrirai donc la stratégie d'analyse utilisée, avec un focus particulier sur une combinaison d'outils mise en place dans le but d'identifier les séquences plasmidiques et chromosomique au sein des génomes. J'ai plus particulièrement évalué les performances de cette approche ciblée sur un jeu de données de *E. coli*.

- Le second objectif de cette thèse est de décrire l'épidémiologie des bactériémies de l'adulte à *E. coli*, dans le contexte de l'émergence et la diffusion de certains clones résistants et virulents récemment observés. Une première partie descriptive consistera à décrire les principaux résultats de l'étude princeps Septicoli, dans lequel mon rôle à consister en l'analyse des génomes des 545 souches de *E. coli*. Dans un second temps, nous aborderons la comparaison fine de cette population bactérienne et de celle de l'étude Colibafi, réalisée en 2005 en région parisienne dans des conditions comparables.

- Enfin, dans une dernier chapitre, l'objectif est d'analyser l'implication des réseaux métaboliques dans la pathogénicité de *E. coli.* A l'aide des génomes de collections de souches bactériémiques (Septicoli et Colibafi), commensales (Coliville), et pathogènes/colonisatrices au niveau pulmonaire (Colocoli), nous tenterons d'identifier des voies métaboliques associées à la pathogénicité et à la porte d'entrée des bactériémies. Un objectif supplémentaire sera d'identifier d'éventuelles voies métaboliques candidates pour expliquer le succès du clone pandémique majeur ST131.

Résultats

I. Stratégie d'analyse des génomes de Septicoli

1. Stratégie d'analyse des génomes

Afin d'assurer une analyse rapide et homogène des génomes séquencés dans le cadre de l'étude Septicoli, j'ai débuté ma thèse par l'élaboration d'un plan d'analyse utilisant une combinaison d'outils bioinformatiques qui a conduit au développement de la stratégie PETA'n'C pour "Plasmid-Exploration, Typing, Assembly N'Contig-ordering" (Figure 20). Les fichiers fastq Illumina de chaque génome constituent les fichiers d'entrée et les fichiers de sortie contiennent les données de typage (MLST, sérotype O:H), de phylogroupage et les fichiers fasta contenant les séquences chromosomiques ordonnées, plasmidiques et inclassables. L'ensemble des logiciels utilisés par PETA'n'C est agencé à l'aide de règles interprétées par le gestionnaire de "workflow" Snakemake (Koster & Rahmann, 2012). L'utilisation de Snakemake facilite également la parallélisation des tâches et permet en cas d'erreur de reprendre l'analyse à l'endroit où l'anomalie s'est produite.

Cette stratégie d'analyse a par la suite été adaptée pour l'analyse des génomes de *E. coli* au sein de l'équipe Inserm IAME par Johann Beghain et Bénédicte Condamine, incluant notamment la méthode développée par Beghain *et al.* pour le phylogroupage des souches (Beghain et al., 2018) et la recherche des gènes de résistance et virulence à l'aide d'Abricate (Seemann, 2014/2021). Étant donnée la rapidité d'exécution, la recherche de déterminants de résistance, virulence et de réplicons plasmidiques était initialement réalisée indépendamment à la fin de l'analyse. La base de données de gènes de résistance aux antibiotiques utilisée est Resfinder (Zankari et al., 2012) qui est d'orientation essentiellement clinique. Afin d'avoir une vue exhaustive du contenu en gènes de virulence, nous avons combiné les bases de virulence VirulenceFinder (Joensen et al., 2014) et VFDB (Chen et al., 2016), et ajouté un ensemble de gènes de virulence ExPEC validé par des experts (Erick Denamur, James R. Johnson, David M. Gordon). Enfin la recherche de réplicons plasmidiques utilise la base de données PlasmidFinder (Carattoli et al., 2014).

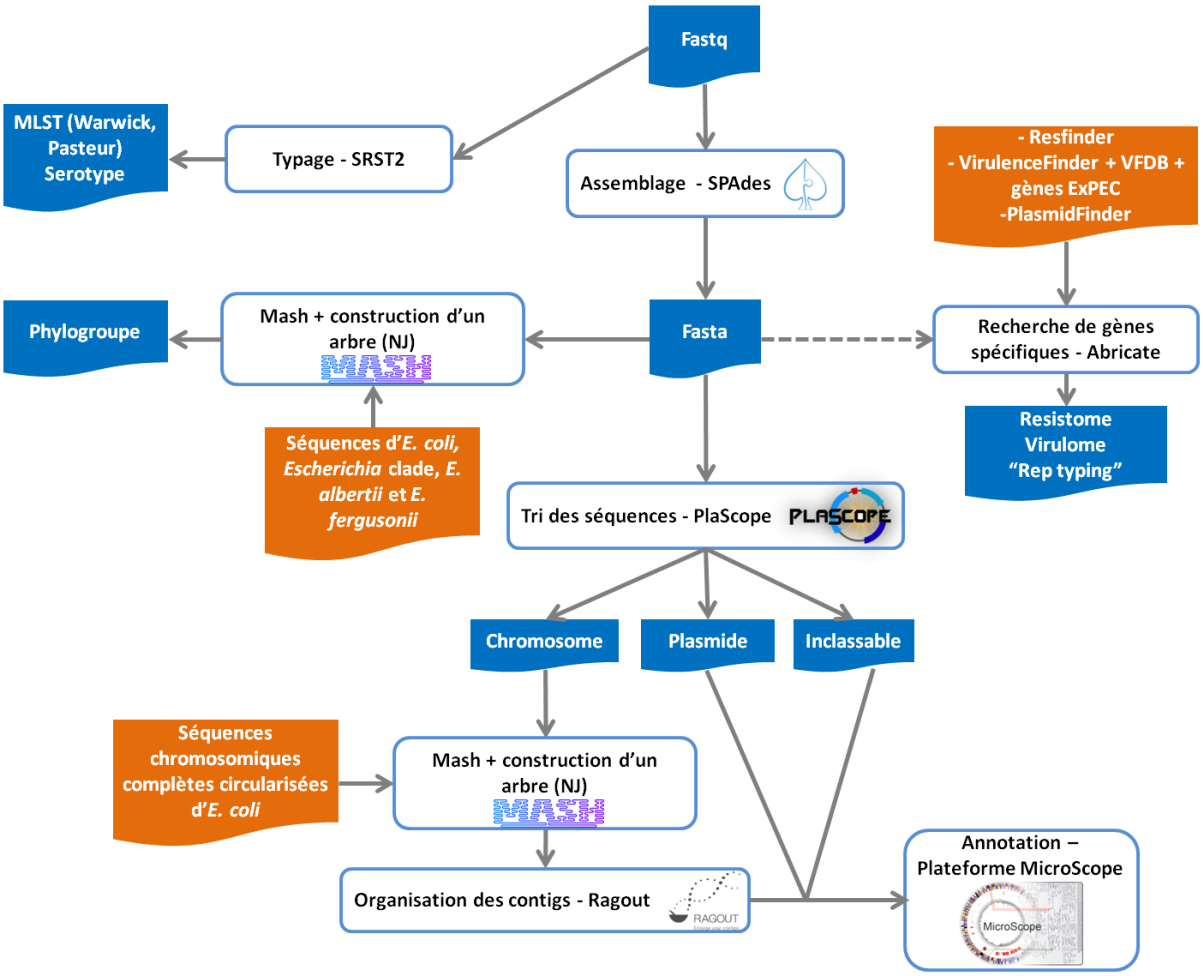


Figure 20. Schéma de la stratégie PETA'n'C pour l'analyse des génomes de Septicoli. La flèche pointillée indique une partie de l'analyse non intégrée initialement. Les encadrés orange correspondent à des données externes utilisées au cours de la caractérisation.

a. MLST et sérotype O:H

La première étape de l'analyse consiste à déterminer le Séquence Type (ST) des souches d'après les schémas de l'université de Warwick (Wirth et al., 2006) et de l'institut Pasteur (Jauréguy et al., 2008), ainsi que les sérotypes O et H. Elle fait appel au programme SRST2 permettant d'utiliser directement les lectures courtes sans étapes d'assemblage (Inouye et al., 2014). La raison initiale de ce choix était la probable sensibilité de ces méthodes aux éventuelles contaminations lors du séquençage. En effet, chaque lecture étant alignée sur les séquences de référence, en cas de mélange de souches par exemple, nous aurions probablement obtenu un mauvais score d'identification invitant à vérifier la qualité de l'assignation.

b. Phylogroupage

Lorsque les outils inclus dans ce schéma ont été choisis, la méthode de détermination des phylogroupes reposait encore sur des techniques de PCR standard. Afin de tirer profit des génomes séquencés et d'éviter les éventuels biais liés à la PCR (*e.g.* délétion de la cible, mésappariement des amorces), nous avons décidé de réaliser l'assignation à un phylogroupe en utilisant les distances génétiques par rapport à un ensemble des génomes de référence. Pour cela, nous réalisons dans un premier temps un assemblage *de novo* des génomes avec SPAdes, l'un des assembleurs basés sur les graphes de de Bruijn les plus utilisé en microbiologie (Prjibelski et al., 2020), puis une estimation des distances génétiques entre nos génomes assemblé et une collection de 413 génomes de *E. coli, Escherichia clade, E. fergusonii* et *E. albertii* de phylogroupe connu qui est obtenue par le programme Mash (Ondov et al., 2016). Ce logiciel compare des ensembles de k-mers représentatifs de chaque génome (obtenus par la commande *mash sketch*) pour calculer une distance entre chaque paire de génomes basée sur un indice de Jaccard (commande *mash dist*). La matrice de distance calculée par Mash, ainsi qu'un fichier tabulé contenant pour chaque génome de référence son phylogroupe, sont par la suite utilisés comme fichier d'entrée d'un script python. Ce script utilise la librairie Phylo de Biopython (Cock et al., 2009). Nous construisons ainsi un arbre par la méthode du "Neighbor Joining" (NJ). Nous déterminons ensuite l'ancêtre commun du phylogroupe auquel appartient le génome de référence le plus proche de notre génome d'intérêt. En parallèle, nous déterminons l'ancêtre commun entre notre génome d'intérêt et l'ensemble du phylogroupe du génome le plus proche. Si ces deux ancêtres communs sont identiques nous attribuons le phylogroupe à notre génome d'intérêt sinon nous concluons uniquement qu'il s'agit du phylogroupe le plus proche. Cette approche s'est révélée concordante avec les résultats de phylogénie à partir de génomes complets, et la méthode proposée par Beghain *et al.* qui repose également sur l'estimation de la distance Mash à l'aide d'un ensemble de génomes bien caractérisés (Beghain et al., 2018).

c. Tri des séquences plasmidiques par l'approche PlaScope

Cette approche de détection de séquences plasmidiques est décrite plus précisément dans la suite de ce chapitre. Brièvement, il s'agit ici d'une utilisation détournée de Centrifuge, un classifieur de lectures de séquençage métagénomique (D. Kim et al., 2016). A la place des lectures, nous utilisons les contigs issus de l'assemblage en fichier d'entrée. Au préalable, nous construisons une base de données pour Centrifuge à partir de séquences complètes

chromosomiques de *E. coli* d'une part et de séquences de plasmides issues de plusieurs sources d'autre part. Une taxonomie artificielle composée uniquement de trois sommets est aussi fournie au programme et nous choisissons d'obtenir une seule assignation taxonomique. Centrifuge réalise ensuite le tri des contigs et leur attribution à une des trois catégories, chromosome, plasmide ou inclassable par détermination d'un score en fonction de la longueur des alignements exactes sur les séquences de référence. Une fois l'analyse terminée, nous créons un tableau par assignation (*i.e.* chromosome, plasmide, inclassable), contenant les noms des contigs en prenant en compte la profondeur calculée par SPAdes, puis nous extrayons les contigs dans un fichier fasta par assignation.

d. Organisation des contigs

Dans une dernière étape, nous réalisons pour chaque génome une organisation des contigs assignés comme chromosomique à l'aide de l'information issue de génomes de référence complets et circularisés de *E. coli, E. albertii, E. fergusonii* (n = 313). Une analyse avec Mash permet d'identifier les 5 génomes de référence les plus proches et ainsi de créer un fichier de "recette" pour le programme Ragout (Kolmogorov et al., 2014). L'utilisation de plusieurs génomes de référence permet à Ragout de prendre en compte les éventuels variants structuraux. Cette méthode découpe tout d'abord les génomes de référence en blocs de synténie conservée. Puis après filtrage des blocs répétés et des blocs absents du génome cible, il construit un graphe dans lequel chaque sommet correspond à l'extrémité d'un bloc de synténie et les arêtes représentent les voisinages entre ces extrémités. Les contigs sont ensuite organisés en scaffold en respectant cette information de voisinage, en utilisant différentes tailles de bloc de synténie afin d'inclure les contigs de petites tailles. Nous utilisons par ailleurs l'option "solid-scaffold" afin de conserver les contigs initialement assemblés par SPAdes et ne pas les découper pendant l'analyse.

Nous n'avons pas évalué en détail les résultats de ces analyses car nous n'avons finalement pas eu besoin de connaître l'organisation des génomes pour l'étude Septicoli. Néanmoins cela a permis de déposer sur la plateforme MicroScope (Vallenet et al., 2019) des génomes de meilleure qualité.

e. Perspectives et améliorations possibles

L'ensemble des analyses réalisées dans la stratégie PETA'n'C permet de typer les souches à partir des outils classiquement utilisés dans les études épidémiologiques de bactéries d'intérêt médical. Mais des étapes supplémentaires pourraient être incluses afin de garantir un résultat de meilleure qualité d'une part et de prendre en compte une plus grande partie de l'information contenue dans les génomes d'autre part. Nous pourrions effectivement intégrer une analyse de la qualité de l'assemblage des génomes en prenant en compte les métriques classiquement utilisées, avec le logiciel Quast par exemple (Gurevich et al., 2013) : taille de l'assemblage, N50 (longueur de séquence du contig le plus court à 50% de la longueur totale du génome), L50 (plus petit nombre de contigs dont la somme des longueurs représente la moitié de la taille du génome), nombres de contigs, GC%. Parallèlement à ces métriques d'assemblage, l'utilisation d'un programme comme CheckM peut être utilisé pour s'assurer de la complétion et de l'absence de contamination (Parks et al., 2015). Pour cela cette méthode s'appuie sur la recherche, en fonction de la position taxonomique du génome analysé, d'un ensemble de gènes conservés et habituellement présents en copie unique.

Concernant le typage des souches, nous avons souhaité utiliser les approches classiques du MLST et du phylogroupe de manière à obtenir des résultats comparables aux études réalisées par PCR. Cependant, ces méthodes n'utilisent qu'une quantité très limitée de l'information disponible. Ils pourraient être intéressant de réaliser des analyses de regroupement à partir des comparaisons des k-mers, comme avec le programme Poppunk par exemple (Lees et al., 2019). De même, puisque nous disposons des séquences complètes, l'intégration d'une méthode pour construire le pangénome des souches analysées serait intéressante. Pour cela, l'approche PPanGGOLiN serait idéale puisqu'elle permet de réaliser l'annotation syntaxique, la création du pangénome et la détection des régions de plasticité génomique grâce à PanRGP (Bazin et al., 2020; Gautreau et al., 2020).

2. Identification de séquences plasmidiques : l'approche PlaScope

a. Contexte

Comme le précise Carattoli dans sa revue en 2009, les méthodes de typage des plasmides par les réplicons ne permettent pas de classer et donc d'identifier tous les plasmides, et le séquençage complet du plasmide reste l'idéal (Carattoli, 2009). Le séquençage complet du génome bactérien donne accès à l'ensemble des séquences, chromosomique et extrachromosomique. Le coût et les capacités de séquençage offertes par les technologies à lecture courte favorisent encore leur utilisation massive en microbiologie au dépend des méthodes à lecture longue. Bien que générant des séquences de bonne qualité, l'assemblage de ces lectures courtes aboutit généralement à des génomes très fragmentés et l'identification des éléments plasmidiques peut donc devenir complexe.

Plusieurs méthodes étaient déjà proposées pour identifier des séquences plasmidiques au sein de génomes bactériens au début de cette thèse mais, comme le montrent Arredondo-Alondo *et al.*, ces méthodes n'étaient pas entièrement satisfaisantes tant du point de vue de la sensibilité que de la spécificité (Arredondo-Alonso et al., 2017). Par ailleurs, elles donnent des résultats hétérogènes tant en termes de nombre de plasmides détectées par génome que de longueur cumulée de séquences plasmidiques (Laczny et al., 2019). Arredondo-Alondo *et al.* ont comparé quatre outils, reposant sur des approches différentes (Arredondo-Alonso et al., 2017). PlasmidFinder permet la détection de séquences de réplicons et présente donc une spécificité élevée, mais une sensibilité très faible en raison de la fragmentation des assemblages à partir de lectures courtes et par conséquent du nombre réduit de contigs porteurs de ces marqueurs de réplication plasmidique (Carattoli et al., 2014). PlasmidSPAdes, lui, propose une approche originale basée sur les différences de profondeur de séquençage qui signeraient la présence d'un élément plasmidique (Antipov et al., 2016). Malheureusement, les plasmides de grande taille sont parfois en copie unique ou en faible nombre de copies et se démarquent donc peu du chromosome en termes de profondeur de séquençage, et les auteurs identifient de nombreux faux positifs par cette méthode (Arredondo-Alonso et al., 2017). Le programme Recycler propose lui aussi d'utiliser les informations de profondeur ainsi que les graphes d'assemblages à la recherche d'éléments circularisés (Rozov et al., 2016). Théoriquement intéressant, les résultats sont plutôt mauvais sur le jeu de données testé tant en termes de sensibilité que de spécificité. Le quatrième outil,

cBar, utilise la fréquence de pentamères préalablement déterminée à partir d'un jeu de données de plasmides et de chromosomes, et est initialement destiné à la recherche de plasmides dans les métagénomes (F. Zhou & Xu, 2010). Bien que pour les plasmides de grande taille la composition en nucléotide souvent proche de celle du chromosome puisse limiter son efficacité, il présente tout de même la meilleure sensibilité derrière PlasmidSPAdes. Enfin, un dernier outil est cité mais non utilisé. Il s'agit de PLACNET, qui combine l'utilisation des informations de profondeur avec la recherche de séquences de réplicons, mais également de relaxases et la comparaison à des bases de données de plasmides (Lanza et al., 2014). Cependant cette méthode requiert une expertise importante puisque l'utilisateur doit lui-même déterminer les séquences qu'il souhaite attribuer comme plasmidiques à partir de l'ensemble de ces informations, raison pour laquelle l'outil n'a pas été inclus dans la comparaison.

Ces résultats nous ont conforté dans l'idée qu'une méthode adaptée à notre jeu de données pouvait offrir de meilleurs résultats. En effet, puisque nous allions par la suite travailler spécifiquement sur l'espèce *E. coli*, l'utilisation d'une approche ciblée nous a semblé idéale, d'autant plus que cette espèce est l'une des mieux décrites et les séquences complètes sont abondantes dans les bases de données.

b. Publication

PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level

Guilhem Royer[1,2,3], Jean-Winoc Decousser[1,2], Catherine Branger[2], Mathieu Dubois[3], Claudine Médigue[3], Erick Denamur[2,4], David Vallenet[3]

1. Département de Microbiologie, Assistance Publique-Hôpitaux de Paris, Hôpital Henri Mondor, Université Paris Est Creteil, F-94000 Creteil, France

2. Université Paris Diderot, INSERM, IAME, UMR 1137, Sorbonne Paris Cité, F-75018 Paris, France

3. LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Universite Paris-Saclay, 91057 Evry, France

4. Assistance Publique-Hôpitaux de Paris, Hôpital Bichat, Laboratoire de Génétique Moléculaire, F-75018 Paris, France

Microbial Genomics 2018 Sep;4(9):e000211. doi: 10.1099/mgen.0.000211.

Résumé

De nombreux déterminants bactériens de résistance et de virulence ont été décrits sur des éléments mobiles. Parmi eux, les plasmides jouent un rôle essentiel dans la diffusion de ces déterminants génétiques, notamment au sein d'espèces d'intérêt médical comme les entérobactéries, *Staphylococcus aureus* ou *Enterococcus*. En effet, en plus des épidémies liées à la diffusion de souches, des épidémies de plasmides ont été décrites, notamment pour certaines bêta-lactamases à spectre étendu ou certaines carbapénémases. Plusieurs outils bioinformatiques existent pour explorer ce "plasmidome" au sein des génomes bactériens, reposant sur différentes approches. Cependant, pour l'heure aucune n'a su combiner sensibilité et spécificité. Dans ce contexte, nous avons développé PlaScope, une approche ciblée permettant d'identifier les séquences plasmidiques au sein d'assemblage à l'échelle de l'espèce voire du genre. Cette approche utilise Centrifuge, un classifieur de séquences métagénomiques, combiné à un jeu de séquences chromosomiques et plasmidiques complètes circularisées provenant de différentes bases de données curées. Elle permet ainsi de classer les contigs en fonction de leur localisation. Comparée aux autres classifieurs plasmidiques, PlasFlow et cBar, elle offre une meilleure sensibilité (0,87), spécificité (0,99), précision (0,96) et exactitude (0,98) sur un échantillon de 70 génomes de *Escherichia coli* contenant des plasmides. Nous avons également utilisé cette méthode sur un ensemble de souches cliniques de *E. coli* et ainsi identifier des phénomènes d'intégration chromosomique de gènes codant des bêta-lactamases à spectre étendu dans 20 cas sur 21. Enfin nous avons également construit une base de données adaptée au genre *Klebsiella* et utilisé celle-ci pour assigner la localisation des gènes de résistance sur une collection de 12 génomes de *K. pneumoniae*. Cette approche est facilement utilisable sur d'autres bactéries bien caractérisées.

PlaScope est disponible à l'adresse suivante : https://github.com/labgem/PlaScope

Ma participation à cette étude a consisté au développement de l'approche et à son évaluation et sa comparaison avec les méthodes disponibles au moment de la publication, et à l'écriture de l'article associé.

MICROBIAL GENOMICS

METHODS PAPER

Royer *et al.*, *Microbial Genomics* 2018;4

DOI 10.1099/mgen.0.000211

PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level

G. Royer,[1,2,3,*] J. W. Decousser,[1,2] C. Branger,[2] M. Dubois,[3] C. Médigue,[3] E. Denamur[2,4] and D. Vallenet[3]

Abstract

Plasmid prediction may be of great interest when studying bacteria of medical importance such as *Enterobacteriaceae* as well as *Staphylococcus aureus* or *Enterococcus*. Indeed, many resistance and virulence genes are located on such replicons with major impact in terms of pathogenicity and spreading capacities. Beyond strain outbreak, plasmid outbreaks have been reported in particular for some extended-spectrum beta-lactamase- or carbapenemase-producing *Enterobacteriaceae*. Several tools are now available to explore the 'plasmidome' from whole-genome sequences with various approaches, but none of them are able to combine high sensitivity and specificity. With this in mind, we developed PlaScope, a targeted approach to recover plasmidic sequences in genome assemblies at the species or genus level. Based on Centrifuge, a metagenomic classifier, and a custom database containing complete sequences of chromosomes and plasmids from various curated databases, PlaScope classifies contigs from an assembly according to their predicted location. Compared to other plasmid classifiers, PlasFlow and cBar, it achieves better recall (0.87), specificity (0.99), precision (0.96) and accuracy (0.98) on a dataset of 70 genomes of *Escherichia coli* containing plasmids. In a second part, we identified 20 of the 21 chromosomal integrations of the extended-spectrum beta-lactamase coding gene in a clinical dataset of *E. coli* strains. In addition, we predicted virulence gene and operon locations in agreement with the literature. We also built a database for *Klebsiella* and correctly assigned the location for the majority of resistance genes from a collection of 12 *Klebsiella pneumoniae* strains. Similar approaches could also be developed for other well-characterized bacteria.

DATA SUMMARY

1. We did not sequence new strains for this study. All the genomes were downloaded from the National Center for Biotechnology Information Sequence Read Archive and Genome database (Tables S1 and S2, available in the online version of this article).

2. The source code of PlaScope is available on Github (https://github.com/GuilhemRoyer/PlaScope).

INTRODUCTION

Recently, several studies have evaluated the effectiveness of *in silico* plasmid prediction tools [1, 2]. In fact, many bioinformatics methods are now available to detect such mobile elements, with different approaches such as read coverage analysis (e.g. PlasmidSPAdes), k-mer-based classification (e.g. cBAR, PlasFlow) and replicon detection (e.g. PlasmidFinder); some of these are fully automated [3–7], others not [8]. Some of them achieve high sensitivity: for example, PlasmidSPAdes and cBar enable plasmid recall of 0.82 and 0.76 on a dataset of 42 genomes, respectively [1]. On the other side, some tools display very high precision, for example PlasmidFinder which reaches 100 % [1]. Unfortunately, none succeeds in finding a good trade-off between sensitivity and specificity, and thus users need to combine different methods to get correct predictions.

Concomitantly, more and more sequences are becoming available in public databases, with various levels of completeness from large sets of contigs to fully circularized genomes and plasmids. Some researchers have made an effort to curate these databases and proposed high-quality datasets. Carattoli *et al.* and Orlek *et al.*, for example, have

Received 22 May 2018; Accepted 24 July 2018

Author affiliations: [1]Département de Microbiologie, Assistance Publique-Hôpitaux de Paris, Hôpital Henri Mondor, Université Paris Est Créteil, F-94000 Créteil, France; [2]Université Paris Diderot, INSERM, IAME, UMR 1137, Sorbonne Paris Cité, F-75018 Paris, France; [3]LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France; [4]Assistance Publique-Hôpitaux de Paris, Hôpital Bichat, Laboratoire de Génétique Moléculaire, F-75018 Paris, France.

***Correspondence:** G. Royer, groyer@genoscope.cns.fr

Keywords: plasmid detection; bioinformatic method; antimicrobial resistance; *Escherichia coli*.

Abbreviations: ESBL, extended-spectrum beta-lactamase; FN, false negative; FP, false positive; NCBI, National Center for Biotechnology Information; TN, true negative; TP, true positive; XDR, extensively drug-resistant.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Four supplementary tables are available with the online version of this article.

published interesting and exhaustive plasmid datasets for *Enterobacteriaceae* [4, 9].

With this in mind, we propose here a workflow, called PlaScope, to assess the plasmidome of genome assemblies. We took advantage of available genomic data to create custom databases of plasmids and chromosomes. These are used as input of the Centrifuge software, a tool originally developed as a metagenomics classifier and that is able to assign sequences based on exact matches against the database [10]. We compared it with other plasmid classifiers, cBar and PlasFlow, and showed that with our specific knowledge-based approach we were able to recover nearly all plasmids of various *Escherichia coli* strains without compromising on specificity. Finally, the usefulness of our approach is illustrated on two datasets: (i) one composed of whole genomes of *E. coli* for which we have sought to identify the location of specific genes involved in virulence or antibiotic resistance, and (ii) the other made up of whole genomes of *Klebsiella pneumoniae* for which we focused on resistance genes and highlighted putative plasmid transmission between strains.

IMPACT STATEMENT

Plasmid exploration could be of great interest because these replicons are pivotal in the adaptation of bacteria to their environment. They are involved in the exchange of many genes within and between species, with a significant impact on antibiotic resistance and virulence in particular. However, plasmid characterization has been a laborious task for many years, requiring complex conjugation or electroporation manipulations, for example. With the advent of whole genome sequencing techniques, access to these sequences is now potentially easier provided that appropriate tools are available. Many software tools have been developed to explore the plasmidome of a large variety of bacteria, but they rarely offer the best compromise in terms of specificity and sensitivity. Here, we focus on single species or genus, and we use the many data available to overcome this problem. With our tool, PlaScope, we achieve high performance compared with two other classifiers, PlasFlow and cBar, and we demonstrate the utility of such an approach to determine the location of virulence or resistance genes. We consider that PlaScope could be very useful in the analysis of specific and well-known bacteria.

THEORY AND IMPLEMENTATION

Workflow description

The PlaScope workflow is illustrated in Fig. 1. First, users have to provide paired end fastq files. Assembly is then run using SPAdes 3.10.1 [11] with the 'careful' option and automatic k-mer size selection to obtain contigs. Subsequently, Centrifuge [10] predicts the location of these contigs thanks to a custom database and sorts them into three classes: plasmid, chromosome and unclassified. The last includes (i) contigs shared by both categories (i.e. matching with plasmid and chromosome sequences from the database) and which are therefore indistinguishable, (ii) contigs without any hit against the database and (iii) contigs with length, hit length or coverage below the defined thresholds. Finally results are sorted based on those three classes and extracted using awk. The complete workflow is available through a unique bash script called PlaScope.sh on github (https://github.com/GuilhemRoyer/PlaScope) or can be installed through BioConda with all dependencies (conda install plascope).

Centrifuge custom database construction

We gathered all the complete genome sequences (chromosomes and plasmids) of *E. coli* from the National Center for Biotechnology Information (NCBI) web site on 10 January 2018. We also added the plasmid sequences that were used to create the PlasmidFinder database [4] and those proposed by Orlek *et al.* [9]. Finally, we added a specific dataset containing *E. coli* plasmids involved in antibiotic resistance [12]. Altogether the database includes 347 chromosome and 3127 plasmid sequences (Table S1 – database available at https://doi.org/10.5281/zenodo.1311641).

We then pooled separately plasmid and chromosome sequences to create a custom database for Centrifuge 1.0.3 [10] with an artificial taxonomy containing only three nodes: 'chromosome', 'plasmid' and 'unclassified' (see README on https://github.com/GuilhemRoyer/PlaScope).

In the same way, we built a *Klebsiella* database. All complete genomes (326 chromosomes and 985 plasmids) of *Klebsiella* species were downloaded from the NCBI web site on 4 July 2018. In addition, the three plasmid databases (PlasmidFinder, Orlek *et al.* and Branger *et al.* datasets) were included (Table S1 – database is available at https://doi.org/10.5281/zenodo.1311647).

Centrifuge classification method

Centrifuge has been developed as a classifier for metagenomic reads. It identifies exact matches between the input sequences and a database originally composed of sequences from several species. It then assigns a score to each of the species that match with the reads and go through a taxonomic tree of these species to output a classification. PlaScope uses this software with a custom database (centrifuge -f --threads 2 -x custom_database -U example.fasta -k 1 --report-file summary.txt -S extendedresult.txt) to classify contigs as 'chromosome', 'plasmid' or 'unclassified', with the option '-k' set to 1 in order to obtain only one taxonomic assignment. Only contigs longer than 500 bp, with a Centrifuge hit longer than 100 bp and with a SPAdes contig coverage higher than 2 are classified as plasmid- or chromosome-related. These parameters were chosen empirically to exclude low-quality contigs and short hits that may not be specific.

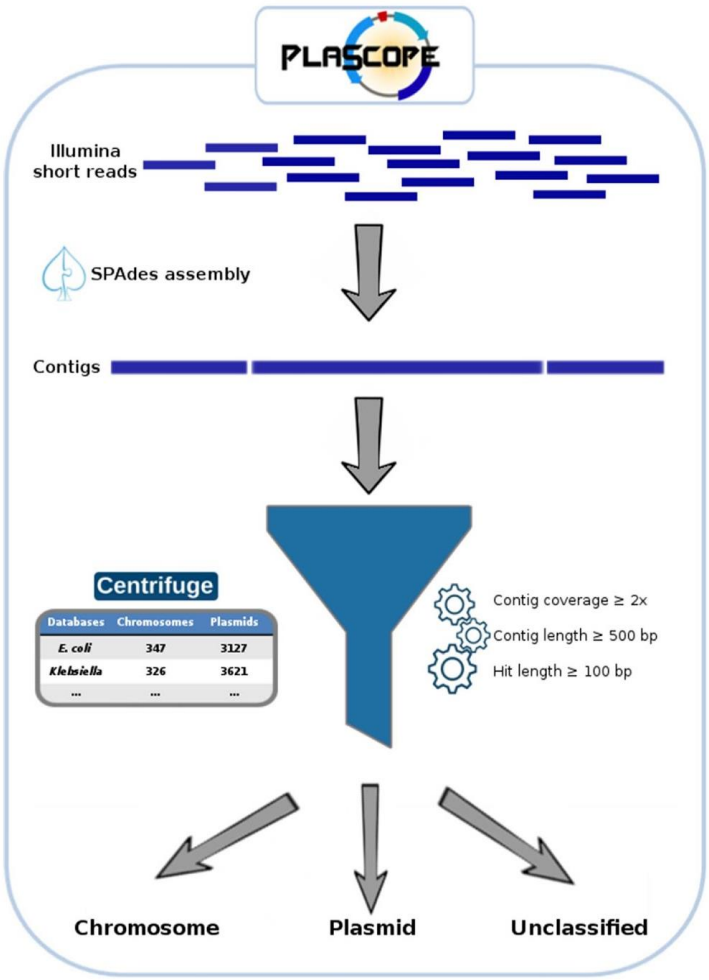


Fig. 1. The PlaScope workflow. After read assembly using SPAdes, contigs are classified into three categories using Centrifuge (i.e. chromosome, plasmid, unclassified) with a custom database containing chromosome and plasmid sequences.

Reference dataset for method evaluation

To evaluate our tool, we searched for completely finished genomes of *E. coli* with Illumina reads available on the NCBI database. All corresponding chromosome and plasmid sequences and Illumina short reads were downloaded from the NCBI on 10 January 2018, and converted into fastq files with fastq-dump from the sra-toolkit (fastq-dump --split-files). For evaluation purposes, these genomes were not included in the centrifuge custom database.

The short reads were assembled with SPAdes 3.10.1 [11] with standard parameters and 'careful' option (spades.py --careful -t 8 -1 read_1.fastq.gz -2 read_2.fastq.gz -o output_directory). After assembly, rapid identification of 16S rRNA sequences was performed on fasta files using ident-16s [13]. Twelve assemblies which did not contain *Escherichia* 16S sequences or with multiple 16S sequences from various organisms were excluded from the subsequent analyses. Finally, we kept 70 genomes containing 183 plasmids and seven genomes with no plasmid according to the NCBI database (Table S2).

We filtered the assemblies based on contig length ($\geq$500 bp) and SPAdes coverage ($\geq$2). Each assembly was then mapped against the corresponding complete chromosome and plasmid sequences from the NCBI database using Quast 4.6 with standard parameters [14]. Contigs that did not align on any sequence (chromosome and plasmid) or aligned on less than 50 % of their length were not considered, as well as contigs that aligned on both sequences.

PlaScope, PlasFlow and cBar benchmark

The PlaScope, PlasFlow [5] and cBar [7] programs with default parameters were run on the reference dataset of 70 *E. coli* genomes containing plasmids. These three methods use different databases and classification approaches to sort contigs as plasmidic or chromosomal. Moreover, PlaScope and PlasFlow may assign contigs as unclassified for ambiguous results.

For each tool, the prediction for each contig was considered as (i) true positive (TP) (plasmid assignment of a plasmidic contig), (ii) true negative (TN) (chromosome or unclassified assignment of a non-plasmidic contig), (iii) false positive (FP) (plasmid assignment of a non-plasmidic contig) or (iv) false negative (FN) (chromosome or unclassified assignment of a plasmidic contig). Detailed counts of these metrics for each assignment type are provided in Table S3. We then calculated recall [TP/(TP+FN)], precision [TP/(TP+FP)], specificity [TN/(FP+TN)], accuracy [(TP+TN)/(TP+TN+FP+FN)] and F1 score [2×(recall×precision)/(recall+precision)]. The results are presented for genomes taken as a whole in Table 1 and individually in Fig. 2.

PlaScope achieves the highest recall on the dataset (0.87) and is closely followed by PlasFlow (0.85), cBar having the lowest value (0.74) (Table 1). At the strain level (Fig. 2), recall values range from 0.50 to 1.00 for the three methods with the lowest median being observed with cBar (0.76) and the highest value with PlaScope (0.90). However, important differences were found for the other assessment criteria. Indeed, with PlaScope we obtained very high precision (0.96), specificity (0.99) and accuracy (0.98) compared to PlasFlow (0.27, 0.68 and 0.70, respectively) and cBar (0.21, 0.60 and 0.62, respectively). Moreover at the strain level, the dispersion of these metrics is high for PlasFlow and cBar compared to PlaScope, in particular for precision (Fig. 2). Clearly, these results are easily explained by the contents of our database, which was built specifically for *E. coli*. PlasFlow and cBar performed well in terms of recall, and their strength relies on their capacity to class many diverse taxonomic groups. Such methods can be particularly useful when working on metagenomes or on single genomes without any prior knowledge of the species, but when focusing on a particular species a targeted approach such as PlaScope drastically limits classification errors. However, 377 contigs from the 10 510 that were analysed remain unclassified with PlaScope. Among them, 248 share hits on both chromosome and plasmid, 117 have no hit and 12 have hits shorter than 100 bp.

Table 1. PlaScope, PlasFlow and cBar benchmark results on contigs from 70 *E. coli* genomes

	PlaScope	PlasFlow	cBar
True positive	1123	1106	954
True negative	9162	6231	5570
False positive	52	2983	3644
False negative	173	190	342
Recall	0.87	0.85	0.74
Precision	0.96	0.27	0.21
Specificity	0.99	0.68	0.6
Accuracy	0.98	0.7	0.62
F1 score	0.91	0.41	0.32

In addition, PlaScope was run on the seven *E. coli* genomes with no plasmids. As expected, no plasmid was predicted for six genomes but, surprisingly, PlaScope predicted two plasmid contigs for *E. coli* KLY (GCA_000725305.1). To assess this result, we aligned these contigs against the NCBI database by BLAST N and obtained perfect alignments with the plasmid F sequence of *E. coli* K-12 C3026 (GenBank accession: CP014273.1). This result suggests that the original assembly of *E. coli* KLY is missing this plasmid.

Application to resistance, virulence gene and operon locations in *E. coli*

In a second step, we evaluated our method on extended-spectrum beta-lactamase (ESBL)-carrying *E. coli* strains sequenced by Illumina MiSeq by Falgenhauer *et al.* [15]. These authors characterized *in silico* the genetic environment and the location of *bla*CTX-M-15 and they found an unusually high rate of chromosomal integration. Indeed, among the 27 isolates of sequence type (ST) 410, 21 carried a *bla*CTX-M-15 gene on their chromosome. We downloaded short reads of these isolates and ran PlaScope to classify the assembled contigs. In parallel, we determined the presence of CTX-M coding genes on the contigs using ResFinder (with a minimal identity of 95 % and a minimal alignment coverage of 90 %) [16].

Using this approach, we accurately identified 20 chromosomally integrated and five plasmid-related CTX-M genes compared to the publication results. In Fig. 3, strains are classified in a neighbour-joining tree (module Phylo from biopython 1.68 [17]) rooted on strain ECOR70 [18] based on genomic similarity distances computed with Mash 2.0 (default parameters) [19]. The tree has been annotated via the Interactive Tree Of Life [20]. We only had a discrepancy with the two isolates of Clade E (strains RS254 and RS371). Indeed, we found a plasmid location of the CTX-M coding gene in strain RS254 whereas it was described as chromosome-related, probably because of an uncommon structure formed by the gene and its adjacent sequences. For the second strain, RS371, the location of the CTX-M coding gene is predicted as unclassified (i.e. hits on both chromosome and plasmid reference sequences) whereas it was stated as plasmid-located. Indeed, BLAST N alignment of the contig carrying this resistance gene against the GenBank database gave perfect hits on plasmid (e.g. CP029575.1) and chromosome (e.g. CP024855.1) sequences which do not allow Centrifuge to differentiate between plasmid or chromosome origin.

In the same publication, the authors also searched for virulence genes and iron metabolism operons. To go further, we

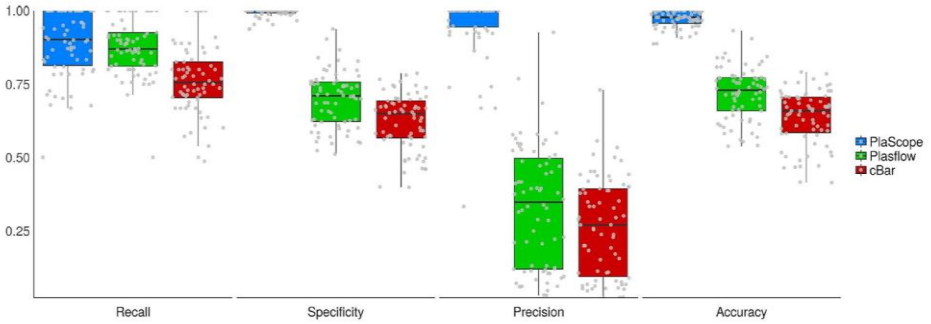


Fig. 2. PlaScope, PlasFlow and cBar performance for each genome taken individually. Recall, specificity, precision and accuracy obtained for each of the 70 genomes containing plasmids are plotted according to the method in blue, green and red for PlaScope, PlasFlow and cBar, respectively. Grey points on box plots represent values for each of these genomes.

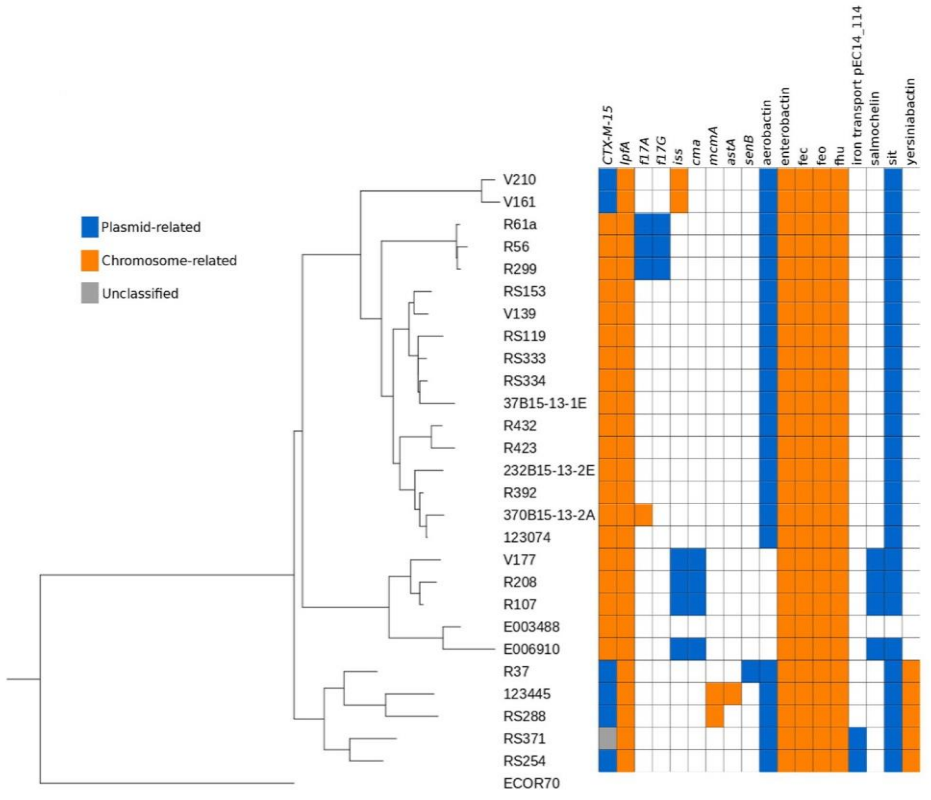


Fig. 3. Genetic distance-based tree with PlaScope-predicted location of *bla*CTX-M-15, virulence genes and operons in the ST410 *E. coli* strains from Falgenhauer *et al.* [15]. Locations of the genes are displayed with coloured squares (blue: plasmid prediction, orange: chromosome prediction, grey: unclassified).

used PlaScope results to determine the location of these genes (Fig. 3). Some of them are exclusively carried by chromosomes (*lpfA*, *mcmA*, *astA*) or plasmids (*f17G*, *cma*, *senB*). Interestingly, *iss* can be found on either type of replicon. For example, *iss* is on a chromosome in Clade A (strains V161 and V210) isolates whereas it is located on plasmids in four out of the five Clade C (strains E006910, R107, R208 and V177) isolates. This illustrates the different genetic background even between closely related strains. In the same way, the gene *f17A* has different locations: on plasmids in three strains (R299, R56, R61a) and on a chromosome in only one (370B15-13-2A, not described in the original publication). These two possible locations of *iss* and *f17A* were previously observed [21, 22]. Regarding the operons, five of them (i.e. enterobactin, fec, feo, fhu and yersiniabactin) were predicted as chromosome-related whereas the others (i.e. aerobactin, salmochellin, sit and the iron transporter pEC14_114) were predicted as plasmidic. These results are in agreement with the literature. Indeed, the first five are known to be chromosome-encoded [23–27] whereas iron transport pEC14_114 is plasmidic [28]. Aerobactin, salmochelin and sit have been found on both types of replicons [29].

Application to resistance gene locations in *Klebsiella pneumoniae*

We applied PlaScope with the *Klebsiella* custom database on a dataset of 12 *K. pneumoniae* strains recovered from a patient and his hospital room environment [30]. PlasmidFinder and ResFinder were then used to identify replicon sequences and resistance genes, respectively. Among the 12 strains, the authors originally described (i) four related strains with one plasmid and no associated resistance genes, (ii) seven extensively drug-resistant (XDR) strains with many plasmids bearing resistance genes and particularly *bla*OXA-181 (a carbapenemase-coding gene) and (iii) a strain close to the four non-XDR strains but with the plasmid carrying *bla*OXA-181. Using PlaScope, we were able to find the correct location of several genes on chromosomes (*bla*SHV-11, *oqxA*, *oqxB*, *fosA*) and plasmids (APH(3″)Ib, APH(6)Id, *bla*OXA-181, *bla*TEM-1B, *catA2*) (Table S4). Furthermore, replicon sequences were detected by PlasmidFinder in 57 contigs predicted as plasmid-related by PlaScope and in only four contigs assigned as unclassified. In addition, some genes were always on unclassified contigs (*dfr*A14, QnrB1, *mph*(A), *arr*-2) as they only contain transposase and resistance genes. These cases cannot be solved by PlaScope due to assemblies being too fragmented and may only be addressed by obtaining finished genomes using long reads as performed by the authors [30]. Nonetheless, we were able to identify the plasmid location of the carbapenemase *bla*OXA-181 in the seven XDR strains and also in the strain that acquired the plasmid during patient hospitalization.

Conclusion

Here, we propose a workflow, called PlaScope, for plasmid and chromosome classification from genome assemblies at the species level. It is based on the assembler SPAdes [11], and Centrifuge [10], a fast metagenomic classifier that uses exact matches between input sequences and a small-sized database to sort these sequences. PlaScope offers high specificity by selecting a unique assignment of contigs to plasmid, chromosome or unclassified. Indeed, we took advantage of the ever growing number of sequences from databases to build a custom database, which combines many high-quality sequences of *Enterobacteriaceae* plasmids and chromosome sequences of *E. coli*. We compared the performance of our tool with cBar and PlasFlow, as these bioinformatic software packages also enable the segregation of plasmid and chromosome contigs. These latter two programs rely on genomic signatures and have been developed to predict plasmid sequences in metagenomic samples.

Compared to PlaScope (recall=0.87), PlasFlow achieves roughly the same recall value on our dataset (recall=0.85), whereas cBar performed less well (recall=0.74). However, regarding other criteria such as precision, specificity and accuracy, PlaScope outperformed the others due to its highly specific database. cBar and PlasFlow are able to identify mobile elements in many bacterial species owing to their very diverse taxonomic database. However, when focusing on a species, the targeted approach of PlaScope gave indisputably better results in terms of both recall and precision as indicated by F1 score (PlaScope: 0.91; PlasFlow: 0.41; cBar: 0.32).

Using PlaScope, we were able to recover almost all plasmids from the analysed strains, with very high precision, specificity and accuracy. Furthermore, in one of the seven strains described as non-bearing plasmid strains in the NCBI database we were able to identify a mobile element: a typical plasmid F in *E. coli* K-12.

In a second analysis, we challenged our approach on more concrete data by looking at specific genes. Analysing clinical or environmental strains, it could be of great interest to detect specific clones with particular genetic backgrounds. Indeed, the plasmid location of a resistance or virulence gene does not have the same impact from an epidemiological point of view and from the capacity of transmission of the strain in a particular environment. For example, plasmid outbreaks can occur when a gene that confers resistance against a wide-spectrum antibiotic is carried by such a mobile element. Conversely, if the same gene integrates in the chromosome of an already highly virulent strain, it can lead to the emergence of a well-adapted and dangerous clone. To highlight this, we chose a genome dataset of *E. coli* wherein many strains exhibited a chromosomal integration of the *bla*CTX-M-15 coding gene, one of the main enzymes responsible for resistance to wide-spectrum antibiotics such as cephalosporins in *E. coli* [15]. Using PlaScope we accurately identified 20/21 of these chromosomal insertions. In addition, we predicted the location of virulence genes and iron metabolism operons in agreement with the literature. This demonstrates that PlaScope may be particularly useful to locate operons such as aerobactin or salmochellin, which

can be plasmidic as well as chromosomal and have, like other iron-metabolism-related systems, major impact on virulence and/or fitness [27, 31].

We also built a *Klebsiella* database and assessed our workflow on *K. pneumoniae* clinical strains [30]. With PlaScope we were able to identify the location of the majority of the resistance genes, notably acquisition of the *bla*OXA-181 gene by a strain through plasmid transmission. However, few contigs carrying resistance genes remain unclassified as they only contain transposase and resistance genes. This is a limitation of our method that requires contigs of sufficient length with specific plasmid or chromosomal regions to make an assignment.

We consider that our approach will be useful when focusing on a well-described species as it makes it possible to decipher the plasmid content of the genomes without overpredicting plasmid sequences. It can highlight integration events or plasmid transmission between isolates. Nonetheless, as it is based on previous knowledge of plasmids found in a specific taxon (e.g. *Enterobacteriaceae*), it will require the database to be enriched to keep it up to date. Finally, it would also be interesting to create other databases for well-known bacteria with many complete genomes available, such as *Staphylococcus aureus*, *Enterococcus* or *Bacillus* species.

Funding information
G.R. was supported by a Poste d'accueil AP-HP/CEA. This work was partially supported by a grant from the 'Fondation pour la Recherche Médicale' to E.D. (Equipe FRM 2016, grant number DEQ20161136698).

Conflicts of interest
The authors declare that there are no conflicts of interest.

Data bibliography
1. Falgenhauer L, Imirzalioglu C, Ghosh H, Gwozdzinski K, Schmiedel J *et al.* Bioproject PRJEB9568 (2016).

2. Simner PJ, Antar AAR, Hao S, Gurtowski J, Tamma PD *et al.* Bioproject PRJNA392824 (2017).

References
1. **Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC.** On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom* 2017;3:e000128.
2. **Laczny CC, Galata V, Plum A, Posch AE, Keller A.** Assessing the heterogeneity of *in silico* plasmid predictions based on whole-genome-sequenced clinical isolates. *Brief Bioinform* 2017;5.
3. **Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A *et al.*** plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* 2016;32:3380–3387.
4. **Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O *et al.*** *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 2014;58:3895–3903.
5. **Krawczyk PS, Lipinski L, Dziembowski A.** PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 2018;46:e35.
6. **Rozov R, Brown Kav A, Bogumil D, Shterzer N, Halperin E *et al.*** Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. *Bioinformatics* 2017;33:475–482.
7. **Zhou F, Xu Y.** cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* 2010;26:2051–2052.
8. **Lanza VF, de Toro M, Garcillán-Barcia MP, Mora A, Blanco J *et al.*** Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genet* 2014;10:e1004766.
9. **Orlek A, Phan H, Sheppard AE, Doumith M, Ellington M *et al.*** A curated dataset of complete Enterobacteriaceae plasmids compiled from the NCBI nucleotide database. *Data Brief* 2017;12:423–426.
10. **Kim D, Song L, Breitwieser FP, Salzberg SL.** Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 2016;26:1721–1729.
11. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al.*** SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
12. **Branger C, Ledda A, Billard-Pomares T, Doublet B, Fouteau S *et al.*** Extended-spectrum β-lactamase-genes are spreading on a wide range of *Escherichia coli* plasmids existing prior the use of third generation cephalosporins. *Microb Genom* 2018;4:000203.
13. **van Zwetselaar M.** 2017. ident-16s Rapid identification of bacterial species from FASTA contigs [Internet]. Github. Available from: https://github.com/zwets/ident-16s [cited 14 January 2018].
14. **Mikheenko A, Valin G, Prjibelski A, Saveliev V, Gurevich A.** Icarus: visualizer for *de novo* assembly evaluation. *Bioinformatics* 2016; 32:3321–3323.
15. **Falgenhauer L, Imirzalioglu C, Ghosh H, Gwozdzinski K, Schmiedel J *et al.*** Circulation of clonal populations of fluoroquinolone-resistant CTX-M-15-producing *Escherichia coli* ST410 in humans and animals in Germany. *Int J Antimicrob Agents* 2016;47: 457–465.
16. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S *et al.*** Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012;67:2640–2644.
17. **Talevich E, Invergo BM, Cock PJ, Chapman BA.** Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics* 2012;13:209.
18. **Galardini M, Koumoutsi A, Herrera-Dominguez L, Cordero Varela JA, Telzerow A *et al.*** Phenotype inference in an *Escherichia coli* strain panel. *eLife* 2017;6:e31035.
19. **Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH *et al.*** Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132.
20. **Letunic I, Bork P.** Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 2016;44:W242–W245.
21. **Johnson TJ, Wannemuehler YM, Nolan LK.** Evolution of the iss gene in *Escherichia coli*. *Appl Environ Microbiol* 2008;74:2360–2369.
22. **Mainil JG, Gérardin J, Jacquemin E.** Identification of the F17 fimbrial subunit- and adhesin-encoding (f17A and f17G) gene variants in necrotoxigenic *Escherichia coli* from cattle, pigs and humans. *Vet Microbiol* 2000;73:327–335.
23. **Burkhardt R, Braun V.** Nucleotide sequence of the fhuC and fhuD genes involved in iron (III) hydroxamate transport: domains in FhuC homologous to ATP-binding proteins. *Mol Gen Genet* 1987; 209:49–55.
24. **Kammler M, Schön C, Hantke K.** Characterization of the ferrous iron uptake system of *Escherichia coli*. *J Bacteriol* 1993;175:6212–6219.
25. **Liu J, Duncan K, Walsh CT.** Nucleotide sequence of a cluster of *Escherichia coli* enterobactin biosynthesis genes: identification of entA and purification of its product 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase. *J Bacteriol* 1989;171:791–798.
26. **Pressler U, Staudenmaier H, Zimmermann L, Braun V.** Genetics of the iron dicitrate transport system of *Escherichia coli*. *J Bacteriol* 1988;170:2716–2724.

27. **Schubert S, Picard B, Gouriou S, Heesemann J, Denamur E.** Yersinia high-pathogenicity island contributes to virulence in *Escherichia coli* causing extraintestinal infections. *Infect Immun* 2002;70:5335–5337.
28. **Debroy C, Sidhu MS, Sarker U, Jayarao BM, Stell AL *et al.*** Complete sequence of pEC14_114, a highly conserved IncFIB/FIIA plasmid associated with uropathogenic *Escherichia coli* cystitis strains. *Plasmid* 2010;63:53–60.
29. **Johnson TJ, Johnson SJ, Nolan LK.** Complete DNA sequence of a ColBM plasmid from avian pathogenic *Escherichia coli* suggests that it evolved from closely related ColV virulence plasmids. *J Bacteriol* 2006;188:5975–5983.
30. **Simner PJ, Antar AAR, Hao S, Gurtowski J, Tamma PD *et al.*** Antibiotic pressure on the acquisition and loss of antibiotic resistance genes in *Klebsiella pneumoniae*. *J Antimicrob Chemother* 2018: 1796–1803.
31. **Gao Q, Wang X, Xu H, Xu Y, Ling J *et al.*** Roles of iron acquisition systems in virulence of extraintestinal pathogenic *Escherichia coli*: salmochelin and aerobactin contribute more to virulence than heme in a chicken infection model. *BMC Microbiol* 2012;12:143.

c. Commentaires et perspectives

Très peu de temps après la publication de ce papier, d'autres outils ont été développés pour la recherche de séquences plasmidiques au sein des génomes bactériens (Rios Miguel et al., 2020). Les approches sont variées : certaines comme MOB-suite combinent l'identification de marqueurs plasmidiques (replicons et relaxase) et le calcul de distances Mash par rapport à un ensemble de séquences plasmidiques de référence (Robertson & Nash, 2018) ; d'autres, comme mlplasmids, utilisent des outils de machine learning pour l'entraînement de classifieur à partir des fréquences de pentamères sur un nombre limité d'espèces (*E. faecium*, *K. pneumoniae*, *E. coli*), et proposent un version en ligne (Arredondo-Alonso et al., 2017). Plus récemment, Schwengers *et al.* ont également proposé une méthode, Platon, combinant les informations de nombreuses sources, et ont comparé leurs résultats à ceux de PlaScope (Schwengers et al., 2020). Pour cela, ils ont établi une nouvelle métrique pour détecter les séquences plasmidiques ("replicon distribution score" ou RDS) basée sur la distribution des groupes de protéines de la base de données UniRef90 entre les séquences plasmidiques et chromosomiques complètes de RefSeq (The UniProt Consortium, 2019). Par ailleurs, pour améliorer la sensibilité de leur méthode, ils prennent également en compte de nombreux autres éléments associés aux plasmides : circularisation, groupe d'incompatibilité, absence d'ARN ribosomaux, gènes de résistance, recherche d'homologie avec des plasmides de référence, détection des origines de réplication *oriT* et des gènes de réplication, de conjugaison et de mobilisation plasmidique. De plus, un tri est effectué sur la taille des contigs et les plasmides trop petits (< 1kpb) ou trop grands (> 500 kpb) sont directement considérés comme chromosomiques. L'un des avantages majeurs de Platon est son indépendance vis-à-vis de la taxonomie, qui lui permet potentiellement d'être utilisé sur une large gamme de génomes bactériens, voire même de métagénomes. Les auteurs réalisent une comparaison avec PlaScope sur un jeu de données de 21 génomes de *E. coli* séquencés en lectures courtes et lectures longues. Les performances des deux programmes sont similaires en tout point, ce qui est rarement le cas des approches indépendantes de la taxonomie généralement désavantagées par rapport aux approches ciblées.

Le développement des technologies de séquençage à lectures longues permet déjà de s'affranchir de ces méthodes et elles sont probablement bien plus adaptées à l'analyse de plasmides. Néanmoins, pour l'heure, peu d'études à grande échelle ont pu être réalisées avec ces techniques de séquençage souvent pour des raisons financières, et les méthodes d'analyse pour les génomes séquencés par lecture courte restent très utilisées, notamment en microbiologie médicale. Les approches ciblées comme PlaScope présentent tout leur intérêt lors de l'analyse de jeu de données de taille importante et centré sur une espèce bien

caractérisée, avec la possibilité de créer ou de mettre à jour facilement la base de données. D'autant que, comme le montrent Laczny *et al.*, il existe d'importantes variations de performances de ces outils de détection de plasmides en fonction du groupe taxonomique des souches analysées (Laczny et al., 2019).

II. Analyse de génomes de *Escherichia coli* responsables de bactériémies chez l'Homme

1. L'étude Septicoli

a. Contexte

Comme nous l'avons vu dans la partie bibliographique, les bactériémies sont des pathologies fréquentes et associées à une mortalité parfois élevée. *E. coli* occupe une place centrale dans l'épidémiologie de ces infections. L'épidémiologie de ces infections n'est pas statique, et comme l'a montré une étude anglaise, certains clones ExPEC comme le ST131 et le ST69, peuvent émerger et atteindre un niveau de prévalence élevé (Kallonen et al., 2017). Ces modifications récentes de l'épidémiologie sont une des raisons de Septicoli. L'étude Colibafi, réalisée dans des conditions similaires une dizaine d'années auparavant, n'avait pu identifier de déterminants bactériens associés à un mauvais pronostic de l'infection (Lefort et al., 2011). Cependant, ces clones virulents et multirésistants aux antibiotiques comme les ST131 et ST69 étaient peu décrits dans les bactériémies. D'autres études se sont intéressées aux facteurs pronostics de ces infections. Mais parmi ces études, certaines n'incluent que des épisodes causés par des souches résistantes (Rodriguez-Bano et al., 2010), ou bien proposent une caractérisation limitée des souches avec la recherche d'un nombre restreint de facteurs de virulence par PCR (Yoon et al., 2018), voire aucune information sur les déterminants génétiques bactériens (J. K. Abernethy et al., 2015). A l'inverse, Kallonen *et al.* proposent une analyse détaillée des génomes de souches responsables mais de manière rétrospective et en l'absence totale de données cliniques essentielles pour identifier d'éventuels facteurs de gravité liés à l'hôte ou au pathogène (Kallonen et al., 2017)**.**

D'autre part, l'épidémiologie de ces infections varie dans le temps mais également géographiquement. Yoon *et al.* rapportent une résistance aux C3G/C4G dépassant 35% en 2016-2017 en Corée (Yoon et al., 2018). A la même date le réseau européen de surveillance de résistance aux antibiotiques (EARS-Net) retrouvait une résistance aux C3G de seulement 11,2% en France (European Centre for Disease Prevention and Control, 2018). Si comme retrouvée dans une méta-analyse en 2007 (Schwaber & Carmeli, 2007), les bactériémies causées par des entérobactéries BLSE sont effectivement associées à un taux de mortalité plus élevé, il semble alors indispensable d'obtenir des données actualisées en terme de résistance pour pouvoir conclure sur les éventuels facteurs pronostics des bactériémies.

Ces considérations épidémiologiques et la possibilité d'obtenir les génomes des souches isolées de bactériémies grâce à la démocratisation des méthodes de séquençage à haut débit ont donc motivé l'étude Septicoli. L'étude Septicoli est un projet financé par l'Agence Nationale de la Recherche (ANR-15-CE-17–0019-01). L'organisation de l'étude reposait sur quatre modules : un module principal de coordination assuré par le Pr. Agnes Lefort, un module "patient" coordonné par le Pr. Victoire de Lastours, un module "laboratoire" coordonné par le Pr. Erick Denamur et enfin un module "statistiques" coordonné par Pr. France Mentré. Par ailleurs, cette nouvelle étude dans des conditions similaires à l'étude Colibafi, tant en termes méthodologiques que démographiques, nous offre également la possibilité d'étudier la dynamique de la population de souches de *E. coli* au sein des bactériémies et fait l'objet de la seconde partie de ce chapitre.

b. Publication

Résumé

Les bactériémies dues à *Escherichia coli* sont associées à une mortalité élevée (5-30%). Les déterminants associés au décès restent peu compris, notamment dans le contexte de l'émergence de souches productrices de bêta-lactamase à spectre étendu (BLSE).

L'objectif de cette étude est de déterminer la part jouée par l'hôte, la virulence et la résistance bactérienne dans le pronostic des bactériémies à *E. coli.*

Pour cela, une étude prospective multicentrique au sein de 7 hôpitaux de région parisienne a consigné l'ensemble des épisodes de bactériémies à *E. coli* consécutifs sur 10 mois. Les isolats ont été séquencés par Illumina Nextseq et leur phylogroupe, ST/STc, virulence génotypique et résistance phénotypique ont été déterminées. Les facteurs de risques associés au décès à la sortie de l'hôpital ou à 28 jours ont été déterminés.

Au total, nous avons inclus 545 patients (âge moyen 68,5 +/- 16,5 ans ; 52,5% de sexe masculin). Le score de comorbidité de Charlson moyen était de 5,6 (+/-3,1) ; 19,6% et 12,8% des patients présentaient un sepsis et un choc septique, respectivement. La porte d'entrée de la bactériémie était principalement urinaire (51,9%), digestive (41,9%) ou pulmonaire (3,5%) ; 98/545 isolats étaient résistants aux céphalosporines de 3ème génération (C3G), dont 86 producteurs de BLSE. Nous avons constaté 52 décès soit 9,5% des épisodes. Les facteurs indépendamment associés au décès étaient la porte d'entrée pulmonaire (Odd ratio ajusté (aOR) = 6,54, intervalle de confiance 95% (IC) = 2,23-19,2, P = 0,0006), le gène de virulence iha_17 (aOR = 4,41, IC = 1,23-15,74, P = 0,022), le STc88 (aOR = 3,62, IC = 1,30-10,09, P = 0,014), le caractère associé aux soins (aOR = 1,98, IC = 1,04-3,76, P = 0,036) et un score de Charlson élevé (aOR = 1,14, IC = 1,04-1,26, P = 0,006), mais pas la résistance aux C3G ni les BLSE.

Les facteurs associés à l'hôte, la porte d'entrée et de rares caractéristiques bactériennes restent les principaux déterminants associés au décès au cours des bactériémies à *E. coli*. Malgré la prévalence élevée des isolats producteurs de BLSE, la résistance antibiotique n'impacte pas le pronostic.

Mon travail au cours de cette étude a constitué en l'analyse des génomes. Plus précisément après séquençage des souches sur la plateforme de l'hôpital Henri Mondor, j'ai réalisé au

GenoScope la caractérisation de ces génomes en déterminant le phylogroupe, le ST, la recherche de gènes de virulence et de résistance à l'aide de la stratégie PETA'n'C décrite au chapitre précédent. Les statistiques ont été réalisées par l'unité de recherche clinique de l'hôpital Bichat.

J Antimicrob Chemother 2020; **75**: 2334–2343
doi:10.1093/jac/dkaa161 Advance Access publication 17 May 2020

Journal of Antimicrobial Chemotherapy

Mortality in *Escherichia coli* bloodstream infections: antibiotic resistance still does not make it

V. de Lastours [1,2]*, C. Laouénan[1,3,4], G. Royer[1,5,6], E. Carbonnelle[1,7], R. Lepeule[6], M. Esposito-Farèse[3,4], O. Clermont[1], X. Duval[1,8], B. Fantin[1,2], F. Mentré[1,3,4], J. W. Decousser[1,6], E. Denamur[1,9] and A. Lefort[1,2] on behalf of the SEPTICOLI Group†

[1]*Université de Paris, IAME, UMR 1137, INSERM, Paris F-75018, France;* [2]*Service de Médecine Interne, Hôpital Beaujon, APHP, F-92100 Clichy, France;* [3]*Département d'épidémiologie, biostatistiques et recherche clinique, Hôpital Bichat, AP-HP, F-75018 Paris, France;* [4]*Unité de recherche clinique, HUPNVS, Hôpital Bichat, AP-HP F-75018 Paris, France;* [5]*LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Paris-Saclay, Evry, France;* [6]*Département de Prévention, Diagnostic et Traitement des Infections, Hôpital Henri Mondor, F-94000 Créteil, France;* [7]*Service de Microbiologie, Hôpital Avicenne, AP-HP, F-93000 Bobigny, France;* [8]*Centre Investigation Clinique INSERM CIC-1425, Bichat Hospital, F-75018 France;* [9]*Laboratoire de Génétique Moléculaire, Hôpital Bichat, AP-HP, F-75018 Paris, France*

*Corresponding author. E-mail: victoire.de-lastours@aphp.fr
†Members of the SEPTICOLI Group are listed in the Acknowledgements section.

Received 13 January 2020; returned 17 February 2020; revised 18 March 2020; accepted 30 March 2020

Background: *Escherichia coli* bloodstream infections (BSIs) account for high mortality rates (5%–30%). Determinants of death are unclear, especially since the emergence of ESBL producers.

Objectives: To determine the relative weight of host characteristics, bacterial virulence and antibiotic resistance in the outcome of patients suffering from *E. coli* BSI.

Methods: All consecutive patients suffering from *E. coli* BSI in seven teaching hospitals around Paris were prospectively included for 10 months. *E. coli* isolates were sequenced using Illumina NextSeq technology to determine the phylogroup, ST/ST complex (STc), virulence and antimicrobial resistance gene content. Risk factors associated with death at discharge or Day 28 were determined.

Results: Overall, 545 patients (mean ± SD age 68.5±16.5 years; 52.5% male) were included. Mean Charlson comorbidity index (CCI) was 5.6 (± 3.1); 19.6% and 12.8% presented with sepsis and septic shock, respectively. Portals of entry were mainly urinary (51.9%), digestive (41.9%) and pulmonary (3.5%); 98/545 isolates (18%) were third-generation cephalosporin resistant (3GC-R), including 86 ESBL producers. In-hospital death (or at Day 28) was 52/545 (9.5%). Factors independently associated with death were a pulmonary portal of entry [adjusted OR (aOR) 6.54, 95% CI 2.23–19.2, $P = 0.0006$], the *iha_17* virulence gene (aOR 4.41, 95% CI 1.23–15.74, $P = 0.022$), the STc88 (aOR 3.62, 95% CI 1.30–10.09, $P = 0.014$), healthcare-associated infections (aOR 1.98, 95% CI 1.04–3.76, $P = 0.036$) and high CCI (aOR 1.14, 95% CI 1.04–1.26, $P = 0.006$), but not ESBL/3GC-R.

Conclusions: Host factors, portal of entry and bacterial characteristics remain major determinants associated with mortality in *E. coli* BSIs. Despite a high prevalence of ESBL producers, antibiotic resistance did not impact mortality. (ClinicalTrials.gov identifier: NCT02890901.)

Introduction

Escherichia coli is a leading causative agent among both community- and hospital-acquired bloodstream infections (BSIs).[1,2] Recent reports from around the world have shown an increase in the incidence of *E. coli* BSI in the last decade.[3,4] Additionally, antibiotic resistance in *E. coli* has increased dramatically due to the global emergence of MDR strains producing ESBL in the healthcare setting and the community.[5] Today, ESBL and/or third-generation cephalosporin-resistant (ESBL/3GC-R, which include ESBL isolates and non-ESBL, 3GC-R isolates) *E. coli* represent the highest burden in terms of deaths and disability-adjusted life-years compared with other MDR organisms in Europe.[6] In-hospital mortality due to *E. coli* BSI varies from 5% to nearly 30%.[7–9] Yet determinants associated with severe outcome and/or death remain poorly understood.[7] Our group previously found host factors and portals of entry to be the most important determinants, outweighing bacterial characteristics, among 1000 episodes of *E. coli* BSI in 2005.[7]

However, in recent years, ESBL/3GC-R strains, which accounted for less than 4% of all *E. coli* in 2005,[10] are now reported to account for 10%–25% of *E. coli* isolates responsible for BSI in France, and up to 50% in some southern European regions.[11] Now that ESBL/3GC-R *E. coli* have become endemic and are found in community patients as well, the impact on mortality of this major epidemiological shift remains unclear. Indeed, ESBL/3GC-R *E. coli* BSI have been associated with increased mortality.[12,13] This may be due to delayed adequate empirical antibiotic therapy, specific host factors (most patients with ESBL/3GC-R *E. coli* had healthcare-related infections) or the intrinsic virulence of the strains.[14] In parallel to these epidemiological evolutions, major progress has occurred in the molecular typing and epidemiology of *E. coli* in the last decade. The availability at reasonable cost of high-quality next-generation genome sequencing and the development of pipelines for sequence analysis allow more precise investigation of bacterial determinants.

In light of the major epidemiological shift, and because determining which patients may be at high risk of dying is crucial in clinical practice, our goal was to determine risk factors for in-hospital mortality from *E. coli* BSI, integrating clinical and genomic data to determine the relative weights of host factors, bacterial factors including resistance determinants, and management on mortality.

Patients and methods

Study design

The SEPTICOLI study is a prospective observational cohort study conducted in seven tertiary-care teaching hospitals in the Paris area. Adult patients with *E. coli* BSI could be included. Only patients previously included in the study and patients receiving vasopressors before the onset of BSI were excluded. *E. coli* BSI was defined as the isolation of *E. coli* from at least one blood culture bottle. Patients were included between October 2016 and July 2017.

Data were prospectively collected by clinicians in each centre on two separate visits: Visit 1 corresponded to the time of BSI (the day the blood culture was drawn; data were collected retrospectively 24–48 h later, once the blood culture had grown) and Visit 2 corresponded to the day of discharge or in-hospital death (or Day 28 if the patient was still hospitalized). For each episode, the first *E. coli* strain collected in the blood culture was identified; antibiotic resistance was determined phenotypically in each hospital laboratory, following a common protocol. Other bacteria present in the blood culture were also identified. Strains corresponding to the portal of entry (when available) were also identified and antibiotic susceptibility was tested. The primary endpoint was vital status at discharge or Day 28 (i.e. Visit 2). In each centre, an infectious diseases clinician and a microbiologist were in charge of including patients and completing the case report form (members of the SEPTICOLI Group are listed in the Acknowledgements section). A steering committee was in charge of implementation and a scientific committee was responsible for scientific overview. Both met twice a year between 2015 and 2019.

Clinical characteristics

Characteristics collected at Visit 1 were demographic data, Charlson comorbidity index (CCI), habitus, previous hospitalizations, antibiotic treatments and past BSI(s). Episodes were nosocomial if the first blood culture was obtained more than 48 h after hospitalization. They were healthcare associated if they were not nosocomial but patients fulfilled at least one of the following criteria: had received intravenous therapy, wound care, specialized nursing care, haemodialysis or intravenous chemotherapy within 30 days prior to the onset of BSI, were hospitalized in an acute-care hospital for ≥2 days within 90 days before or resided in a nursing home or long-term care facility before admission. Otherwise they were considered community acquired.

Immunosuppression was defined as one of the following: patients receiving an immunosuppressive treatment (anticancer chemotherapy or immunomodulating therapies for autoimmune diseases within the last month or ≥10 mg/day prednisone for more than 30 consecutive days), solid-organ transplant recipients, bone marrow transplant recipients, neutropenia (<500 cells/mm^3), congenital immunodeficiency or HIV infection. The portal of entry was established according to clinical and/or radiological characteristics of the episodes and the isolation of *E. coli* from the presumed source of infection. When *E. coli* could be isolated from the source of infection, the portal of entry was considered certain. If not, the presumed portal of entry was assigned on the basis of firm clinical suspicion and considered likely. Patients' charts that had been filled as either undetermined or more than one portal of entry were reviewed by two clinical investigators (V.D.L. and A.L.) to reclassify as best as possible. BSI was polymicrobial if at least one microorganism other than *E. coli* was identified in blood culture.

All antibiotics received by patients between Visits 1 and 2 were prospectively collected with date and time of administration by the hour. All antibiotic regimens were later analysed by a panel of experts (A.L. and E.C.) to determine whether and when the adequate antibiotic regimen had been started to treat the BSI episode (defined as at least one antibiotic effective *in vitro* on the *E. coli* strain, according to the phenotypic antibiotic susceptibilities). Patients who had received no antibiotics were considered to have received inadequate antibiotic treatment.

Bacterial determinants

Phenotypic analysis

For each patient, the first *E. coli* strain collected from the blood culture bottle was identified in each hospital laboratory by MALDI-TOF and antibiotic susceptibility testing was performed (according to EUCAST, www.eucast.org). MDR was defined as acquired non-susceptibility to at least one agent among three or more antimicrobial categories.[15] One *E. coli* isolate per patient as well as any *E. coli* collected from a potential portal of entry were frozen at −80°C in glycerol. Further genomic analyses of isolates were performed at the IAME Laboratory.

Genomic analysis

After DNA extraction, the whole genome of each blood culture strain was sequenced using Illumina NextSeq 2×150 bp after NextEra XT library preparation (Illumina, San Diego, CA, USA). Sequences were analysed with a previously described in-house bioinformatic pipeline.[16] Briefly, assemblies were performed with SPAdes 3.10.0 and genomes were ordered with Ragout after genetic distances were estimated using Mash.[17–19] Phylogroups were determined using the ClermonTyper.[20] Resistance (n = 3077) and virulence (n = 1271) genes/alleles were scanned using ResFinder and custom databases, respectively.[16,21] ST complexes (STcs) were considered only when they contained more than 2% of the total population (n = 11 strains). They were defined as single- or double-locus variants based on the MLST data of the Warwick scheme using Phylovizin congruence with the Mash distances.[22] Sequences are available in the Bioproject PRJEB35745.

Statistical analyses

All patients with available isolates included in the study were analysed. Comparison of variables between groups was performed using Student's t-test or χ^2, Fisher's or Wilcoxon's tests, as appropriate. Clinical and bacterial characteristics were studied using univariate logistic regression in order to

determine their association with the primary endpoint, i.e. in-hospital death secondary to *E. coli* BSI up to 28 days after the first positive blood culture. Factors associated at $P < 0.10$ level were fitted in a multivariate logistic regression. A selection method (backward stepwise) was used to obtain a final model in which all risk factors had a value of $P < 0.05$. All 2×2 interactions between variables in the final model were tested. We tested the goodness of fit of the multivariate model using the Hosmer–Lemeshow test and c-statistics, and assessed the model using the bootstrap method. We also compared clinical characteristics between patients infected with ESBL/3GC-R *E. coli* and those with non-ESBL/3GC-R strains, using the same tests. Subgroup analyses were also performed among patients with a non-urinary portal of entry and among patients who presented with severe sepsis or shock. All analyses were performed with R software (v 3.6) and the significance level was a two-sided type-I error of 0.05.

Ethics

This study was approved by the French Comité de Protection des Personnes Ile de France number IV (IRB 00003835, March 2016). Because of its non-interventional nature, only an oral consent from patients was requested. The study was registered in clinical trials in September 2016 (ClinicalTrials.gov identifier: NCT02890901).

Results

Clinical characteristics

During the study period, 553 *E. coli* BSI patients were included in the study. Eight patients were excluded because *E. coli* isolates were not available. Altogether, 545 patients with *E. coli* BSI were analysed. The patients' main characteristics are summarized in Table 1. Mean (± SD) age was 68.5 (± 16.5) years and 52.5% were men. Portals of entry were predominantly urinary and digestive. Antibiotics received empirically (within the 48 h after the blood culture was drawn) are summarized in Table S1 (available as Supplementary data at *JAC* Online). In terms of severity, 107 (19.6%) and 70 (12.8%) patients presented with sepsis and septic shock, respectively. Fifty-two patients (9.5%, 95% CI 7.0%–12.0%) had died by Visit 2.

Bacterial characteristics

Bacterial characteristics are described in Table 2, Figure 1, Table S2 and Figure S1. ESBL/3GC-R strains accounted for 98 (18.0%) isolates, including 86 (15.8%) ESBL producers; one strain was resistant to carbapenems.

Determinants of death at Visit 2

The in-hospital death rate after *E. coli* BSI up to 28 days after the first positive blood culture was 9.5% (52 patients). The clinical and/or bacterial factors associated with death are shown in Table 3. The variables 'sepsis' and 'septic shock' were excluded from the analysis because by definition they were highly associated with severe outcome. Factors associated with death in the multivariate analysis were the CCI [adjusted OR (aOR) = 1.14, 95% CI 1.04–1.26, $P = 0.006$], a healthcare-related infection (aOR = 1.98, 95% CI 1.04–3.76, $P = 0.03$), a pulmonary portal of entry (aOR = 6.54, 95% CI 2.23–19.2, $P = 0.0006$), being infected with an STc88 strain (aOR = 3.62, 95% CI 1.30–10.09, $P = 0.014$) and the presence of the *iha_17* virulence gene (aOR = 4.41, 95% CI 1.23–15.74, $P = 0.022$). Infections with an ESBL/3GC-R isolate were not associated with a poor outcome.

Of note, carriage of *iha_17* was not different between STc88 isolates and non-STc88 isolates (13/505, 2.6% versus 1/24, 4.2%, $P = 0.46$), none of the STc88 isolates had a pulmonary portal of entry and none of the strains responsible for the pulmonary portal of entry carried *iha_17* (Figure S1). We tested the goodness of fit of the multivariate model using the Hosmer–Lemeshow test ($P = 0.089$) and the c-statistic (0.741) (95% CI 0.68–0.81). We also assessed the model using the bootstrap method. The resulting accuracy was 0.91.

Table 4 shows the comparison of the 98 episodes due to ESBL/3GC-R with the non-ESBL/3GC-R isolates. A major STc shift occurred between ESBL/3GC-R and non-ESBL/3GC-R isolates, from predominantly STc73 ($n = 65$; 14.5%), STc69 ($n = 60$, 13.4%) in non-ESBL/3GC-R to a predominance of STc131 ($n = 44$; 44.9%) and STc88 (9; 9.2%) ($P < 0.001$) (Table S3).

To assess the impact of bacterial resistance on mortality, we excluded urinary portals of entry as they were associated with a more favourable outcome and focused on non-urinary portals of entry (260 episodes, including 33 patient deaths and 33 ESBL/3GC-R isolates). In this subgroup, despite the fact that patients infected with ESBL/3GC-R isolates were more likely to receive adequate antibiotics 48 h or more after onset of BSI compared with non ESBL/3GC-R [16/33 (48.5%) versus 17/227 (7.5%), $P < 0.001$], infection with ESBL/3GC-R isolates was not associated with death [4/33 (12.1%) versus 29/227 (12.7%), $P > 0.999$]. Similarly, we separately analysed 177 patients who presented with severe sepsis or shock, regardless of the portal of entry, among whom 41 died (23.2%). ESBL/3GC-R infections ($n = 38$) were not associated with a worse outcome [9/38 (23.6%) versus 32/139 (23%), $P = 0.53$], even though patients were more likely to receive adequate antibiotics 48 h or more after BSI onset compared with non-ESBL/3GC-R-infected patients [14/38 (36.8%) versus 18/139 (12.9%), $P = 0.001$].

With the limitation of small numbers of patients in the subgroups, among the 98 ESBL/3GC-R infected patients, no difference in terms of mortality was found with respect to empirical antibiotic received in the first 48 h or definitive antibiotic treatment after 48 h.

Discussion

We report here the results of the SEPTICOLI study, which prospectively included 545 episodes of *E. coli* BSI and precisely analysed clinical data and genomes of 545 *E. coli* isolates. Fifty-two patients (9.5%) had died in the hospital by Day 28. The main result of this work is that host-related factors and bacterial virulence determinants, but not antibiotic resistance, were independently associated with mortality.

The most striking result is that infections due to ESBL and/or 3GC-R isolates, albeit frequent (18%), were not associated with a worse outcome despite the fact that patients infected with these strains received adequate antibiotic therapy later than those with susceptible strains. This was found even after excluding patients with a urinary portal of entry, which were more likely to have ESBL/3GC-R infections and had more favourable outcomes, as reported by others.[7–9,23] The absence of impact of 3GC resistance on mortality was also found in the subgroup of the most severe patients

Table 1. Clinical characteristics of the 545 patients included in the SEPTICOLI study, and comparison by outcome (survivors versus non-survivors at discharge or Day 28)

	All patients $N = 545$	Survivors $N = 493$	Non-survivors $N = 52$	*P* value
Demographics				
age, years, mean (SD)[a]	68 (16.5)	68 (16.4)	72 (16.5)	0.088
sex, male	286/545 (52.5)	255/493 (51.7)	31/52 (59.6)	0.278
Habitus				
active smoking	69/537 (12.8)	60/486 (12.3)	9/51 (17.6)	0.282
chronic alcoholism	40/541 (7.4)	35/490 (7.1)	5/51 (9.8)	0.411
Medical history				
CCI, mean (SD)[a]	5.62 (3.1)	5.45 (3.2)	7.17 (2.5)	<0.0001
current malignancy				<0.0001
with metastasis	92/545 (16.9)	72/493 (14.6)	20/52 (38.5)	
without metastasis	98/545 (18.0)	92/493 (18.7)	6/52 (11.5)	
diabetes mellitus				0.963
with complications	33/545 (6.1)	30/493 (6.1)	3/52 (5.8)	
without complications	96/545 (17.6)	88/493 (17.8)	8/52 (15.4)	
chronic renal failure	95/545 (17.4)	87/493 (17.6)	8/52 (15.4)	0.848
chronic peripheral arteritis	60/545 (11.0)	50/493 (10.1)	10/52 (19.2)	0.046
congestive heart failure	51/545 (9.4)	44/493 (8.9)	7/52 (13.5)	0.313
cirrhosis	52/545 (9.5)	47/493 (9.5)	5/52 (9.6)	0.985
dementia	44/545 (8.1)	40/493 (8.1)	4/52 (7.7)	0.916
immunosuppression[b]	147/545 (27.0)	128/493 (26.0)	19/52 (36.5)	0.102
history of bacteraemia in previous 12 months	80/532 (15.0)	72/482 (14.9)	8/50 (16.0)	0.840
hospitalization in the previous 12 months	320/522 (61.3)	285/471 (60.5)	35/51 (68.6)	0.292
pregnancy	9/259 (3.5)	9/238 (3.8)	0/21 (0)	0.364
Current episode				
type of infection				0.075
community acquired	296/545 (54.3)	275/493 (55.8)	21/52 (40.4)	
healthcare associated	82/545 (15.0)	70/493 (14.2)	12/52 (23.1)	
nosocomial	167/545 (30.6)	148/493 (30.0)	19/52 (36.5)	
severity of the initial presentation				<0.0001
sepsis	107/545 (19.6)	90/493 (18.3)	17/52 (32.7)	
septic shock	70/545 (12.8)	46/493 (9.3)	24/52 (46.2)	
Glasgow coma score, mean (SD)[c]	14.5 (1.9)	14.93 (0.7)	14.72 (1.7)	0.005
Portal of entry[d]				
urinary tract	281/541 (51.9)	265/492 (53.9)	16/49 (32.6)	0.006
digestive tract	227/541 (42.0)	199/492 (40.4)	28/49 (57.1)	0.033
pulmonary	19/541 (3.5)	13/492 (2.6)	6/49 (12.2)	0.004
cutaneous	8/541 (1.5)	7/492 (1.4)	1/49 (2.0)	0.535
venous catheter	23/541 (4.2)	20/492 (4.1)	3/49 (6.1)	0.454
surgical site	3/541 (0.5)	3/492 (0.6)	0/49 (0)	>0.999
other[e]	9/541 (1.7)	9/492 (1.8)	0/49 (0)	0.611
undetermined	1/541 (0.2)	1/492 (0.2)	0/49 (0)	>0.999
multiple portals of entry (>1)	26/541 (4.8)	23/492 (4.7)	3/49 (6.1)	0.72
Management				
start of adequate antibiotics more than 48 h after the onset of bacteraemia	99/545 (18.2)	84/493 (17.0)	15/52 (28.8)	0.036

Values represent frequencies, denominator and percentages, *n*/D (%), unless otherwise indicated.
[a]Denominators: total = 545; survivors = 493; non-survivors = 52.
[b]Defined as at least one of the following: (i) patients receiving an immunosuppressive treatment (defined as anticancer chemotherapy or immunomodulating therapies for autoimmune diseases within the last month, or ≥10 mg/day prednisone for more than 30 consecutive days); (ii) solid-organ transplant recipients; (iii) bone marrow transplant recipients; (iv) neutropenia (<500/mm^3); (v) congenital immunodeficiency; or (vi) HIV infection.
[c]Denominators: total = 513; survivors = 465; non-survivors = 48.
[d]Portals of entry are not mutually exclusive. Patients may have multiple portals of entry.
[e]Two osteomyelitis and seven gynaecological portals of entry.

Table 2. Bacterial determinants of the 545 strains responsible for *E. coli* bacteraemia in the SEPTICOLI study, and comparison by outcome at discharge or Day 28 (survivors versus non-survivors)

	All patients $N=545$	Survivors $N=493$	Non-survivors $N=52$	*P* value
Phylogroup				0.257
A	53/545 (9.72)	51/493 (10.3)	2/52 (3.8)	
B1	70/545 (12.84)	62/493 (12.6)	8/52 (15.4)	
B2	279/545 (51.19)	255/493 (51.7)	24/52 (46.1)	
C	24/545 (4.4)	18/493 (3.6)	6/52 (11.5)	
clade I	2/545 (0.37)	2/493 (0.4)	0/52 (0)	
clade V	1/545 (0.18)	1/493 (0.2)	0/52 (0)	
D	87/545 (15.96)	78/493 (15.8)	9/52 (17.3)	
E	4/545 (0.73)	4/493 (0.8)	0/52 (0)	
F	25/545 (4.59)	22/493 (4.5)	3/52 (5.8)	
ST complex				
STc10	33/545 (6.1)	33/493 (6.7)	0/52 (0)	0.612
STc12	18/545 (3.3)	15/493 (3.0)	3/52 (5.8)	0.241
STc131	82/545 (15.0)	76/493 (15.4)	6/52 (11.5)	0.545
STc14	19/545 (3.5)	15/493 (3.0)	4/52 (7.7)	0.097
STc141	13/545 (2.4)	12/493 (2.4)	1/52 (1.9)	>0.999
STc162	12/545 (2.2)	9/493 (1.8)	3/52 (5.8)	0.097
STc58	29/545 (5.3)	26/493 (5.3)	3/52 (5.8)	0.750
STc69	65/545 (11.9)	60/493 (12.2)	5/52 (9.6)	0.821
STc73	68/545 (12.5)	65/493 (13.2)	3/52 (5.8)	0.182
STc88	24/545 (4.4)	18/493 (3.6)	6/52 (11.5)	0.019
STc95	38/545 (7.0)	36/493 (7.3)	2/52 (3.8)	0.565
other	144/545 (26.4)	128/493 (26.0)	16/52 (30.8)	0.508
Virulence factors[a]				
gad_56	10/545 (1.8)	7/493 (1.4)	3/52 (5.8)	0.061
iha_17	15/545 (2.7)	10/493 (2.0)	5/52 (9.6)	0.009
lpfA_5	28/545 (5.1)	21/493 (4.3)	7/52 (13.5)	0.012
papGII	193/545 (35.4)	182/493 (36.9)	11/52 (21.1)	0.023
terC	5/545 (0.9)	3/493 (0.6)	2/52 (3.8)	0.074
Resistance determinants				
WT strain	170/545 (31.2)	151/493 (30.6)	19/52 (36.5)	0.432
resistance to:				
amoxicillin/clavulanic acid	244/545 (44.8)	223/493 (45.2)	21/52 (40.4)	0.559
piperacillin/tazobactam	62/506 (12.3)	55/456 (12.1)	7/50 (14)	0.652
3GC	97/542 (17.9)	87/490 (17.8)	10/52 (19.2)	0.849
fluoroquinolones (any)	135/545 (24.8)	118/493 (23.9)	17/52 (32.7)	0.177
cotrimoxazole	231/543 (42.5)	214/482 (44.4)	17/51 (33.3)	0.182
aminoglycoside (any)	79/545 (14.5)	69/493 (14)	10/52 (19.23)	0.303
carbapenem (any)	1/545 (0.2)	1/493 (0.2)	0/52 (0)	>0.999
tigecycline	1/316 (0.3)	1/292 (0.3)	0/24 (0)	>0.999
fosfomycin	4/544 (0.7)	4/492 (0.8)	0/52 (0)	>0.999
ESBL producers	86/545 (15.8)	79/493 (16.0)	7/52 (13.5)	0.841
ESBL producers and/or 3GC-R	98/545 (18.0)	88/493 (17.8)	10/52 (19.2)	0.849
MDR[b]	234/545 (42.9)	214/493 (43.4)	20/52 (38.5)	0.443
Polymicrobial episodes	70/545 (12.8)	58/493 (11.8)	12/52 (23.1)	0.023

Values represent frequencies, denominators and percentages, *n*/D (%).
[a]Only virulence factors significantly associated with death are noted here. The full list of virulence factors is available in Table S2.
[b]Resistance to more than two classes of antibiotics as defined by Magiorakos *et al.*[15]

suffering sepsis or shock, among whom the death rate was 23%. This result may appear to contrast with previous reports from the literature.[24] Indeed, in a meta-analysis from 2007, an almost 2-fold increase in mortality for ESBL infections and a significant association between ESBL production and delay in onset of an effective antibiotic treatment was found.[24] However, others found

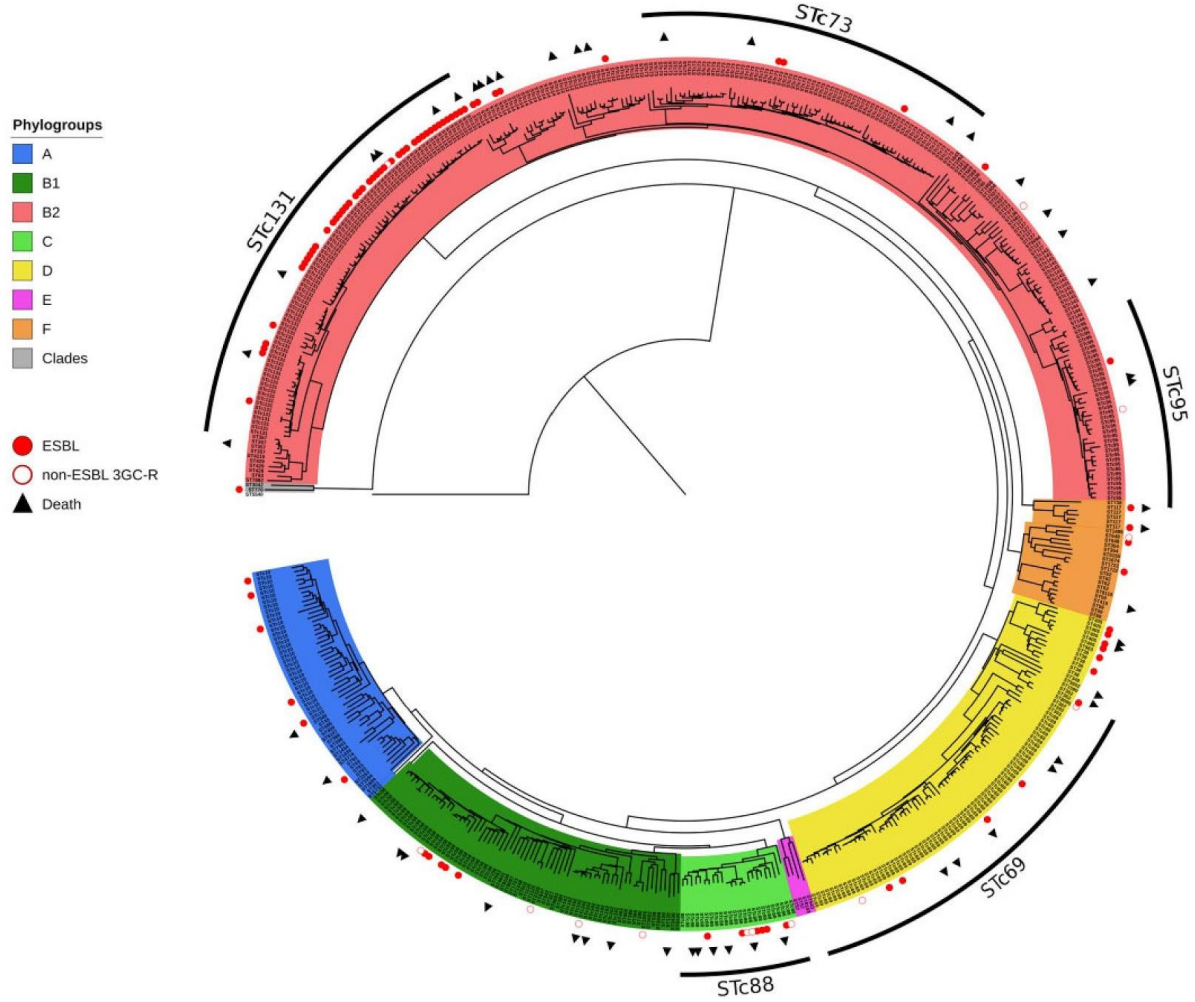


Figure 1. Genetic distance tree of the 545 *Escherichia* clade and *E. coli* strains isolated from bacteraemia in 545 patients. The distance tree was computed from Mash distances using the neighbour-joining method. The tree is rooted on the *Escherichia* clade V strain as it has been shown to be the most divergent from the *E. coli* species.[43] The STcs are displayed only when they contained more than 2% of the total population ($n = 11$ strains), otherwise only STs are presented on the leaf (see Patients and methods section for definition). The seven main *E. coli* phylogroups are shown in colour whereas *Escherichia* clade strains ($n = 3$) are in grey. The presence of ESBL-coding genes is highlighted by red filled circles. Non-ESBL but 3GC-R strains are highlighted with red open circles. The four main STcs (69, 73, 95 and 131) as well as the STc88 associated with death are indicated.

conflicting results:[25,26] nearly half the studies from the above-mentioned meta-analysis found no difference in terms of mortality associated with ESBL presence;[24] mortality among patients with community-acquired *E. coli* BSI was associated with inappropriate empirical therapy irrespective of ESBL production in a Spanish study.[25] Most of these studies are retrospective, do not include only *E. coli* and date back 10 years or more. Both the epidemiology of ESBL isolates and clinicians' therapeutic decisions have evolved in the last 10 years.[27] The results presented here are therefore highly relevant. Two possible explanations can be provided. First, these findings may suggest a decreased virulence of the ESBL/3GC-R strains. Indeed, a difference in terms of the STc responsible for BSI according to the ESBL/3GC-R status of the strains was identified. Among non-ESBL/3GC-R strains, mostly STc73, STc69 and STc95 were found, classical extra-intestinal pathogenic *E. coli* (ExPEC) lineages known for their numerous virulence factors and high intrinsic virulence in the sepsis mouse model.[28,29] On the other hand, among ESBL/3GC-R strains STc131 became highly predominant (nearly 45% of isolates). The intrinsic virulence of STc131 appears to vary widely and as a group it does not appear to be more virulent than other ExPEC in animal models.[30,31] Additionally, STc131 does not appear to be associated with higher mortality than other ESBL-producing non-ST131 clones in patients.[32] However, the second most prevalent ESBL/

Table 3. Univariate and multivariate analyses of determinants of mortality at discharge or Day 28, among the 545 episodes of *E. coli* bacteraemia

Risk factors	Survivors $N = 493$	Non-survivors $N = 52$	Univariate analysis		Multivariate analysis	
			OR (95% CI)	*P* value	aOR (95% CI)	*P* value
Clinical factors						
current malignancy[a]	164/493 (33.3)	26/52 (50)	2.01 (1.13–3.57)	0.018		
CCI, mean (SD)	5.4 (3.2)	7.2 (2.5)	1.19 (1.08–1.30)	0.002	1.14 (1.04–1.26)	0.006
healthcare-associated infection[b]	218/493 (44.2)	31/52 (59.6)	1.86 (1.04–3.33)	0.036	1.98 (1.04–3.76)	0.036
polymicrobial episodes	58/493 (11.8)	12/49 (24.5)	2.25 (1.12–4.53)	0.023		
urinary portal of entry	265/492 (53.9)	16/49 (32.6)	0.42 (0.22–0.77)	0.006		
digestive portal of entry	168/492 (34.1)	19/49 (38.8)	1.96 (1.08–3.55)	0.026		
pulmonary portal of entry	13/492 (2.6)	6/52 (11.5)	5.14 (1.86–14.21)	0.002	6.54 (2.23–19.2)	0.0006
start of adequate antibiotics >48 h after bacteraemia	84/493 (17.0)	15/52 (28.8)	1.97 (1.04–3.76)	0.039		
ESBL producers and/or 3GC-R strains	88/493 (17.85)	10/52 (19.23)	1.10 (0.53–2.27)	0.810		
Bacterial factors						
STc88	18/493 (3.6)	6/52 (11.5)	3.44 (1.30–9.10)	0.013	3.62 (1.30–10.09)	0.014
Virulence factors						
iha_17	10/493 (2.0)	5/52 (9.6)	5.14 (1.69–15.66)	0.004	4.41 (1.23–15.74)	0.022
IpfA_5	21/493 (4.3)	7/52 (13.5)	3.50 (1.41–8.67)	0.007		
papGII	182/493 (36.9)	11/52 (21.1)	0.46 (0.23–0.91)	0.027		

Values represent frequencies, denominators and percentages, *n*/D (%).
[a]Any active malignancies with or without metastasis are included here.
[b]Both 'healthcare-associated' and 'nosocomial' infections are included here versus 'community-acquired' infections.

Table 4. Univariate and multivariate analyses comparing the 98 patients infected with ESBL and/or 3GC-R strains with the 447 others, among the 545 episodes of *E. coli* bacteraemia in the SEPTICOLI study

Risk factors	No ESBL or 3GC-R $N = 447$	ESBL and/or 3GC-R $N = 98$	Univariate analysis		Multivariate analysis	
			OR (95% CI)	*P* value	aOR (95% CI)	*P* value
Renal failure	70/447 (15.7)	25/98 (25.5)	1.84 (1.10–3.11)	0.021	1.87 (1.07–3.28)	0.028
Chronic alcoholism	28/447 (6.3)	12/98 (12.2)	2.10 (1.03–4.29)	0.042	3.04 (1.40–6.60)	0.005
Past history of bacteraemia	56/447 (12.5)	24/98 (24.5)	2.23 (1.30–3.82)	0.004	2.81 (1.57–5.01)	0.0004
Urinary portal of entry	217/447 (48.5)	64/98 (65.3)	2.03 (1.28–3.21)	0.003	2.41 (1.47–3.96)	0.0005
Digestive portal of entry	199/447 (44.5)	28/98 (28.6)	0.50 (0.31–0.81)	0.004		
Start of adequate antibiotics >48 h after the onset of bacteraemia	63/447 (14.1)	36/98 (36.7)	3.54 (2.17–5.77)	<0.0001	3.04 (1.86–4.98)	<0.0001

Values represent frequencies, denominators and percentages, *n*/D (%).

3GC-R group, although accounting only for nine strains, is STc88, which was independently associated with death in the multivariate analysis, and is reported to be highly virulent in the mouse model.[33] Altogether, decreased virulence of some ESBL/3GC-R strains may partly explain these results, but this requires further exploration.

The second hypothesis is that empirical antibiotics used to treat *E. coli* BSI may conserve some activity *in vivo* against ESBL/3GC-R isolates, despite the phenotypic resistance found *in vitro*. A majority of patients were treated empirically with 3GC or piperacillin/tazobactam, both of which have been described to retain some antimicrobial activity in resistant strains *in vivo*.[34,35] From a pharmacokinetics/pharmacodynamics point of view, although deemed 'resistant', the growth of strains may be inhibited *in vivo* by antibiotic concentrations transiently superior to the MIC and sufficient to achieve some activity, depending on the doses used and the MIC for the strains; for instance, piperacillin/tazobactam exposures of ≥55% of time above the MIC have been associated with success in treating ESBL infections.[35] Unfortunately, neither MICs for the ESBL/3GC-R strains nor drug concentrations were measured here to confirm this hypothesis. Additionally, even if the concentrations achieved are subinhibitory, bacteria exposed even to subinhibitory levels of antibiotics may change their adherence properties, cell surface antigen, excretion of enzymes and toxins,

and cell wall thickness.[36] Reports have found that *E. coli* grown in the presence of sub-MIC β-lactams were phagocytosed and killed in numbers significantly higher than untreated bacteria.[37] This phenomenon may be insufficient to cure the infection entirely and prevent recurrence or the emergence of resistant mutants, but sufficient to avoid mortality within the first 48 h until optimal antibiotic treatment is started.

The second important result of this work is that host factors and portal of entry remain important drivers of death in *E. coli* BSI, yet some important new players have emerged.[7] Indeed, we identified the pulmonary portal of entry as highly associated with death (aOR = 6.54, 95% CI 2.23–19.2), albeit only 19 cases, as found in a recent study from South Korea.[9] *E. coli* isolates responsible for pulmonary infections, although infrequent, appear to have a distinct phylogenetic and virulence profile, as described recently by our group in ventilator-associated *E. coli* pneumonia, the prevalence of which is increasing.[38] We are currently analysing the clinical and microbiological specificities of the pneumonia-specific *E. coli*. Second, two specific bacterial factors, STc88 and the virulence gene *ihA_17*, were rare occurrences (4.4% and 2.7%, respectively), yet when present were highly associated with a poor outcome. The STc88, corresponding to the minor C phylogroup, is considered to be a particularly virulent human ExPEC.[33,39] The presence of *iha_17* is not linked to a particular lineage but widespread among all phylogenetic groups. IhA is an IrgA-homologous adhesin with a double function of adhesion and as a siderophore, which has been associated with specific virulence traits, especially in urinary tract infections.[40] The specific role of IhA_17 is unknown and deserves further investigation.

Several limitations need to be addressed. First, we only analysed one *E. coli* isolate for each BSI episode. Within-sample diversity exists in *E. coli* infections, such as those found in intra-abdominal abscesses.[41] However, we are likely to have analysed the single most virulent isolate capable of crossing the organ–blood barrier, which makes it particularly relevant. Second, probably because of small numbers in this subgroup, no difference with respect to antibiotic received was evidenced in terms of outcome among patients infected with ESBL/3GC-R strains. This does not mean all antibiotics are equal, but rather that our data cannot answer this question. Unlike the MERINO study, this work was not designed to compare antibiotics' performance in ESBL/3GC-R infections.[42] Therefore, no conclusions should be drawn from this study concerning the best therapeutic strategy in ESBL/3GC-R infections. Third, the epidemiology of bacterial resistance in the Paris area may not be extrapolated to other parts of the world. Risk factors for ESBL infections are changing in time and space, from mostly healthcare-related to increasingly more community-acquired infections, as found here, and our findings may be rapidly outdated.[4–6]

The major strength of this work is the integration of precise clinical data, prospectively collected by a homogeneous group of clinicians and microbiologists in a short time span, and complete genome analysis of all *E. coli* isolates. Most studies recently published either have little clinical data or used PCR typing to analyse the genome.[7–9] This allowed us to determine that the portal of entry and host factors remain major players in the outcome of *E. coli* BSI but that some bacterial factors may also be involved. Interestingly, we found that 3GC resistance was not associated with higher mortality, even though patients received adequate antibiotic therapy later, because of either the cost of resistance or a possible conserved activity of antibiotics *in vivo*.

Acknowledgements

This work was presented in part at the European Congress of Clinical Microbiology and Infectious Diseases in Amsterdam, the Netherlands in April 2019 (Abstract P2402).

We thank all patients who agreed to take part in this study, as well as the SEPTICOLI research assistants, Naura Gamany and Sarah Zaher, whose contribution to the work was crucial.

Members of the SEPTICOLI Group

Virginie Zarrouk, Fréderic Bert, Marion Duprilot, Véronique Leflon, Naouale Maataoui, Laurence Armand, Liem Luong Nguyen, Rocco Collarino, Anne-Lise Munier, Hervé Jacquier, Emmanuel Lecorché, Laetitia Coutte, Camille Gomart, Ousser Ahmed Fateh, Luce Landraud, Jonathan Messika, Elisabeth Aslangul, Magdalena Gerin, Alexandre Bleibtreu, Mathilde Lescat, Violaine Walewski, Frederic Mechaï, Marion Dollat, Anne-Claire Maherault, Michel Wolff, Mélanie Mercier-Darty and Bernadette Basse.

Funding

This work was funded by a Translational Research Grant from the Agence Nationale de la Recherche (ANR), Ministry of Higher Education and Research, France 2015 (grant no. ANR-15-CE-17–0019-01). E.D. was partially supported by the 'Fondation pour la Recherche Médicale' (Equipe FRM 2016, grant no. DEQ20161136698). G.R. was supported by a 'Poste d'accueil' funded by the Assistance-Publique Hôpitaux de Paris (AP-HP) and the Centre pour l'Energie Atomique (CEA) personal grant for his PhD.

Transparency declarations

None to declare.

Author contributions

V.D.L., E.D. and A.L. wrote the grant proposal, designed the study and wrote the article. V.D.L. and A.L. coordinated patients' inclusion, wrote the clinical report form and controlled the clinical data. E.C. supervised the phenotypic microbiological analysis. G.R., J.W.D. and O.C. were in charge of the genome sequencing and its analysis under the supervision of E.D. C.L. and M.E.-F. were in charge of the statistics under the supervision of F.M. E.C., R.L., X.D. and B.F. sat on the scientific committee and contributed to the scientific input of the work. All authors contributed to the writing of the manuscript.

Supplementary data

Tables S1 to S3 and Figure S1 are available as Supplementary data at *JAC* Online.

References

1 Wisplinghoff H, Bischoff T, Tallent SM *et al*. Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. *Clin Infect Dis* 2004; **39**: 309–17.

2 Laupland KB. Incidence of bloodstream infection: a review of population-based studies. *Clin Microbiol Infect* 2013; **19**: 492–500.

3 Vihta K-D, Stoesser N, Llewelyn MJ *et al*. Trends over time in *Escherichia coli* bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998-2016: a study of electronic health records. *Lancet Infect Dis* 2018; **18**: 1138–49.

4 van der Mee-Marquet NL, Blanc DS, Gbaguidi-Haore H *et al*. Marked increase in incidence for bloodstream infections due to *Escherichia coli*, a side effect of previous antibiotic therapy in the elderly. *Front Microbiol* 2015; **6**: 646.

5 Hogberg LD. Antimicrobial Resistance (AMR) Reporting Protocol. 2018. https://www.ecdc.europa.eu/sites/portal/files/documents/EARS-Net%20reporting%20protocol%202018.%20docx.pdf.

6 Cassini A, Högberg LD, Plachouras D *et al*. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect Dis* 2019; **19**: 56–66.

7 Lefort A, Panhard X, Clermont O *et al*. Host factors and portal of entry outweigh bacterial determinants to predict the severity of *Escherichia coli* bacteremia. *J Clin Microbiol* 2011; **49**: 777–83.

8 Abernethy JK, Johnson AP, Guy R *et al*. Thirty day all-cause mortality in patients with *Escherichia coli* bacteraemia in England. *Clin Microbiol Infect* 2015; **21**: 251.e1–8.

9 Yoon E-J, Choi MH, Park YS *et al*. Impact of host-pathogen-treatment tripartite components on early mortality of patients with *Escherichia coli* bloodstream infection: prospective observational study. *EBioMedicine* 2018; **35**: 76–86.

10 Courpon-Claudinon A, Lefort A, Panhard X *et al*. Bacteraemia caused by third-generation cephalosporin-resistant *Escherichia coli* in France: prevalence, molecular epidemiology and clinical features. *Clin Microbiol Infect* 2011; **17**: 557–65.

11 ECDC. Surveillance Report. Surveillance of Antimicrobial Resistance in Europe. 2016. https://www.ecdc.europa.eu/sites/default/files/documents/AMR-surveillance-Europe-2016.pdf.

12 Kang C-I, Song J-H, Chung DR *et al*. Risk factors and treatment outcomes of community-onset bacteraemia caused by extended-spectrum β-lactamase-producing *Escherichia coli*. *Int J Antimicrob Agents* 2010; **36**: 284–7.

13 Cordery RJ, Roberts CH, Cooper SJ *et al*. Evaluation of risk factors for the acquisition of bloodstream infections with extended-spectrum β-lactamase-producing *Escherichia coli* and *Klebsiella* species in the intensive care unit; antibiotic management and clinical outcome. *J Hosp Infect* 2008; **68**: 108–15.

14 Nicolas-Chanoine M-H, Bertrand X, Madec J-Y. *Escherichia coli* ST131, an intriguing clonal group. *Clin Microbiol Rev* 2014; **27**: 543–74.

15 Magiorakos A-P, Srinivasan A, Carey RB *et al*. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin Microbiol Infect* 2012; **18**: 268–81.

16 Bourrel AS, Poirel L, Royer G *et al*. Colistin resistance in Parisian inpatient faecal *Escherichia coli* as the result of two distinct evolutionary pathways. *J Antimicrob Chemother* 2019; **74**: 1521–30.

17 Bankevich A, Nurk S, Antipov D *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; **19**: 455–77.

18 Ondov BD, Treangen TJ, Melsted P *et al*. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016; **17**: 132.

19 Kolmogorov M, Raney B, Paten B *et al*. Ragout—a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* 2014; **30**: i302–9.

20 Beghain J, Bridier-Nahmias A, Le Nagard H *et al*. ClermonTyping: an easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microb Genomics* 2018; **4**: 207–11.

21 Zankari E, Hasman H, Cosentino S *et al*. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012; **67**: 2640–4.

22 Nascimento M, Sousa A, Ramirez M *et al*. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics* 2017; **33**: 128–9.

23 Jauréguy F, Carbonnelle E, Bonacorsi S *et al*. Host and bacterial determinants of initial severity and outcome of *Escherichia coli* sepsis. *Clin Microbiol Infect* 2007; **13**: 854–62.

24 Schwaber MJ, Carmeli Y. Mortality and delay in effective therapy associated with extended-spectrum β-lactamase production in Enterobacteriaceae bacteraemia: a systematic review and meta-analysis. *J Antimicrob Chemother* 2007; **60**: 913–20.

25 Rodríguez-Baño J, Picón E, Gijón P *et al*. Community-onset bacteremia due to extended-spectrum β-lactamase-producing *Escherichia coli*: risk factors and prognosis. *Clin Infect Dis* 2010; **50**: 40–8.

26 Rottier WC, Ammerlaan HSM, Bonten M. Effects of confounders and intermediates on the association of bacteraemia caused by extended-spectrum β-lactamase-producing Enterobacteriaceae and patient outcome: a meta-analysis. *J Antimicrob Chemother* 2012; **67**: 1311–20.

27 Bevan ER, Jones AM, Hawkey PM. Global epidemiology of CTX-M β-lactamases: temporal and geographical shifts in genotype. *J Antimicrob Chemother* 2017; **72**: 2145–55.

28 Tourret J, Denamur E. Population phylogenomics of extraintestinal pathogenic *Escherichia coli*. *Microbiol Spectr* 2016; **4**: 510–5.

29 Johnson JR, Clermont O, Menard M *et al*. Experimental mouse lethality of *Escherichia coli* isolates, in relation to accessory traits, phylogenetic group, and ecological source. *J Infect Dis* 2006; **194**: 1141–50.

30 Lavigne J-P, Vergunst AC, Goret L *et al*. Virulence potential and genomic mapping of the worldwide clone *Escherichia coli* ST131. *PLoS One* 2012; **7**: e34294.

31 Johnson JR, Porter SB, Zhanel G *et al*. Virulence of *Escherichia coli* clinical isolates in a murine sepsis model in relation to sequence type ST131 status, fluoroquinolone resistance, and virulence genotype. *Infect Immun* 2012; **80**: 1554–2.

32 Chung H-C, Lai C-H, Lin J-N *et al*. Bacteremia caused by extended-spectrum-β-lactamase-producing *Escherichia coli* sequence type ST131 and non-ST131 clones: comparison of demographic data, clinical features, and mortality. *Antimicrob Agents Chemother* 2012; **56**: 618–22.

33 Moissenet D, Salauze B, Clermont O *et al*. Meningitis caused by *Escherichia coli* producing TEM-52 extended-spectrum β-lactamase within an extensive outbreak in a neonatal ward: epidemiological investigation and characterization of the strain. *J Clin Microbiol* 2010; **48**: 2459–63.

34 Brun-Buisson C, Legrand P, Philippon A *et al*. Transferable enzymatic resistance to third-generation cephalosporins during nosocomial outbreak of multiresistant *Klebsiella pneumoniae*. *Lancet* 1987; **2**: 302–6.

35 Abodakpi H, Chang K-T, Gao S *et al*. Optimal piperacillin-tazobactam dosing strategies against extended-spectrum-β-lactamase-producing Enterobacteriaceae. *Antimicrob Agents Chemother* 2019; **63**: e01906–18.

36 Tornqvist IO, Holm SE, Cars O. Pharmacodynamic effects of subinhibitory antibiotic concentrations. *Scand J Infect Dis Suppl* 1990; **74**: 94–101.

37 Adinolfi LE, Bonventre PF. Enhanced phagocytosis, killing, and serum sensitivity of *Escherichia coli* and *Staphylococcus aureus* treated with sub-MICs of imipenem. *Antimicrob Agents Chemother* 1988; **32**: 1012–8.

38 La Combe B, Clermont O, Messika J *et al*. Pneumonia-specific *Escherichia coli* with distinct phylogenetic and virulence profiles, France, 2012-2014. *Emerg Infect Dis* 2019; **25**: 710–8.

39 Clermont O, Couffignal C, Blanco J *et al*. Two levels of specialization in bacteraemic *Escherichia coli* strains revealed by their comparison with commensal strains. *Epidemiol Infect* 2017; **145**: 872–82.

40 Johnson JR, Jelacic S, Schoening LM *et al*. The IrgA homologue adhesin Iha is an *Escherichia coli* virulence factor in murine urinary tract infection. *Infect Immun* 2005; **73**: 965–71.

41 Levert M, Zamfir O, Clermont O *et al*. Molecular and evolutionary bases of within-patient genotypic and phenotypic diversity in *Escherichia coli* extraintestinal infections. *PLoS Pathog* 2010; **6**: e1001125.

42 Harris PNA, Tambyah PA, Lye DC *et al*. Effect of piperacillin-tazobactam vs meropenem on 30-day mortality for patients with *E coli* or *Klebsiella pneumoniae* bloodstream infection and ceftriaxone resistance: a randomized clinical trial. *JAMA* 2018; **320**: 984–94.

43 Clermont O, Christenson JK, Denamur E *et al*. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep* 2013; **5**: 58–65.

c. Commentaires et perspectives

L'étude Colibafi sur les bactériémies de l'adulte à *E. coli* en France en 2005 n'avait pu mettre en évidence de déterminants bactériens associés à un mauvais pronostic, et seul le gène *ireA* apparaissait comme facteur protecteur (Lefort et al., 2011). Les facteurs associés à l'hôte (âge, cirrhose, hospitalisation préalable, immunodépression) et la porte d'entrée cutanée expliquaient en effet principalement les décès. Cependant seuls 19 gènes de virulence avaient été analysés par PCR, et la résistance aux antibiotiques large spectre comme les C3G était de seulement 3,7%.

Dans ce contexte l'étude Septicoli a eu pour objectif d'analyser ces facteurs de pronostic en tirant profit du séquençage des génomes et également dans le contexte d'une augmentation massive de la résistance qui atteint désormais 18%. Mais à nouveau les facteurs de l'hôte sont prépondérants dans la survenue des décès en accord avec plusieurs études (Martinez et al., 2006; Mora-Rillo et al., 2015; Yoon et al., 2018). De manière étonnante, un traitement antibiotique inadapté pendant les 48 premières heures, fréquent au vu des niveaux de résistance observés, n'entraîne pas une aggravation du pronostic, et ce même chez les patients les plus graves (*i.e.* en sepsis et choc septique). D'autres études observent également une absence d'association entre traitement adéquat et mortalité (Rodriguez-Bano et al., 2010; Yoon et al., 2018). En revanche, des auteurs hollandais rapportent une mortalité augmentée dans les bactériémies à *E. coli* BLSE, mais aucune autre donnée clinique n'est utilisée pour pondérer ces résultats et le traitement antibiotique reçu par le patient n'est pas non plus intégré (van Hout et al., 2020). De plus, les épisodes causés par des souches BLSE dans cette étude sont plus fréquemment d'origine nosocomiale, et il ne s'agit pas d'une étude prospective ni d'un recueil consécutif des épisodes de bactériémies.

Nous avons néanmoins pu identifier de rares déterminants bactériens liés au pronostic. Tout d'abord, l'allèle *papGII* ressort comme un facteur protecteur. Il n'est en fait qu'un proxy de la porte d'entrée urinaire puisque 75% des souches porteuses de ce gène ont été isolées au cours de bactériémies associées à un tel foyer primaire. Une observation semblable était faite dans Colibafi avec le gène *ireA* (Lefort et al., 2011). Jauréguy *et al.* en 2007 retrouvaient également *papGII* souvent associé à la voie urinaire (Jauréguy et al., 2007), et Mora-Rillo *et al.* ont également identifié une association de bon pronostic pour les fimbriae de type P (Jauréguy et al., 2007; Mora-Rillo et al., 2015).

Ensuite deux déterminants bactériens semblent associés aux décès. L'un deux ne correspond pas un déterminant génétique unique mais directement un groupe de souches proches, le STc88. Le gène *lpfA_5* qui ressort de l'analyse univariée est en fait un marqueur de STc88 présent dans la totalité des isolats. Ce clone a été mis en évidence au cours de méningite néonatale, et pourrait donc posséder des capacités de virulence particulière (Moissenet et al., 2010). Le second déterminant est un allèle du gène, *iha_17*, retrouvé lui au sein de 6 phylogroupes différents (A, B1, B2, C, D et F). Ce gène code un récepteur de sidérophore de type catécholate possédant également des fonctions d'adhésion (J. R. Johnson et al., 2005). Son rôle a été montré dans un modèle d'infection urinaire mais n'explique pas son association à la sévérité de l'infection. L'analyse des séquences de Septicoli montre que ce gène est retrouvé dans des contigs prédits chromosomiques par PlaScope dans 14/15 cas, avec la présence d'éléments mobiles à proximité (intégrase et tranposase). Des études plus approfondies, notamment phénotypiques, doivent être réalisées si l'on veut pouvoir inférer le rôle réel de ces déterminants bactériens. D'autre part, il faut noter le faible nombre d'observations au cours de notre étude qui n'est pas un signal très fort pour expliquer l'implication de ces éléments dans la physiopathologie.

Enfin, si les génomes complets de ces souches nous ont permis d'obtenir une vision précise de l'épidémiologie en termes de structure de population (phylogroupe et ST/STc) et de contenu en gènes de virulence, il reste énormément de données à exploiter de ces séquences. Ceci est l'objet de l'étude ci-après ainsi que du chapitre suivant sur le métabolisme.

2. Dynamique de la population de *Escherichia coli* responsable de bactériémies sur 12 ans

a. Contexte

Comme nous le laissait entrevoir l'évolution globale de la population de *E. coli* dans les infections extraintestinales, dans l'étude Septicoli nous avons constaté une forte prévalence des souches résistantes aux C3G, essentiellement liée à la présence des souches de STc131, porteuses de BLSE de type CTX-M. Bien que moins importante que dans d'autres études, comme celle de Yoon *et al.* par exemple (Yoon et al., 2018), la résistance aux C3G dans Septicoli apparaît très fortement augmentée en comparaison à l'étude Colibafi réalisée 12 années auparavant, atteignant 17,9% des isolats en 2016-2017 contre seulement 3,7% en 2005 (de Lastours et al., 2020; Lefort et al., 2011). Ces deux études, Septicoli et Colibafi, ont été réalisées dans des conditions similaires, et bien que les hôpitaux parisiens participants soient en partie différents, on peut considérer qu'il existe une certaine homogénéité en termes de distribution géographique. La comparaison des génomes issus de ces deux collections peut donc nous permettre d'obtenir une vision plus dynamique de l'épidémiologie des bactériémies à *E. coli*, à la manière de l'étude proposée par Kallonen *et al.* mais sur une population plus standardisée (Kallonen et al., 2017). Contrairement à l'étude réalisée par ces auteurs nous ne disposons que de deux points dans le temps, mais en revanche un nombre d'isolats bien plus importants pour chacun d'eux. D'autre part, les souches de l'étude Septicoli ont été isolées entre 2016 et 2017 et reflètent donc probablement mieux l'épidémiologie actuelle que le point final de l'étude de Kallonen *et al.*, datant de 2012. Étant donné l'émergence récente et l'expansion rapide de certains clones ExPEC (Ben Zakour et al., 2016; Kallonen et al., 2017; McNally et al., 2019), il est essentiel de disposer de données les plus actualisées possibles si l'on veut pouvoir en tirer des conclusions plus globales. Nous avons donc séquencé 367 génomes de la collection Colibafi, en nous concentrant uniquement sur les épisodes de bactériémies inclus au sein d'hôpitaux de la région parisienne, afin de les comparer aux génomes de l'étude Septicoli.

L'approche de cette seconde étude se veut bien moins clinique que la première, l'objectif étant d'identifier les changements majeurs apparus à 12 années d'écart dans l'épidémiologie des bactériémies à *E. coli* de l'adulte en région parisienne. La comparaison des génomes de nos deux collections permet d'analyser la dynamique de la population avec différents niveaux de granularité, de l'échelle globale représentée par les phylogroupes à une échelle plus fine en considérant les STc et les différents groupes de souches identifiables au sein de ces STc. La prise en compte des données de typage tels les sérotypes O:H, l'allèle *fimH* ainsi que le

contenu en gènes de virulence et de résistance offre une image précise de ces variations dans le temps.

b. Publication

Cet article est actuellement en révision dans le journal Genome Medicine.

Phylogroup stability contrasts with high within sequence type complex dynamics of *Escherichia coli* bacteremia isolates over a 12-year period

Guilhem Royer[1,2,3], Mélanie Mercier Darty[3], Olivier Clermont[1], Cédric Laouenan[1,4], Jean-Winoc Decousser[3], David Vallenet[2], Agnès Lefort[1,5], Victoire de Lastours[1,5], Erick Denamur[1,6], COLIBAFI* and SEPTICOLI** groups

1. Université de Paris, IAME, UMR 1137, INSERM, Paris F-75018, France.
2. LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Paris-Saclay, Evry, France.
3. Département de Prévention, Diagnostic et Traitement des Infections, Hôpital Henri Mondor, F-94000 Créteil, France
4. Département d'épidémiologie, biostatistiques et recherche clinique, Hôpital Bichat, AP-HP, F-75018 Paris, France
5. Service de Médecine Interne, Hôpital Beaujon, APHP, F-92100 Clichy, France
6. Laboratoire de Génétique Moléculaire, Hôpital Bichat, AP-HP, F-75018 Paris, France

* The Colibafi Group: Michel Wolff, Loubna Alavoine, Xavier Duval, David Skurnik, Paul-Louis Woerther, Antoine Andremont, Etienne Carbonnelle, Olivier Lortholary, Xavier Nassif, Sophie Abgrall, Françoise Jauréguy, Bertrand Picard, Véronique Houdouin, Yannick Aujard, Stéphane Bonacorsi, Agnès Meybeck, Guilène Barnaud, Catherine Branger, Agnès Lefort, Bruno Fantin, Claire Bellier, Frédéric Bert, Marie-Hélène Nicolas-Chanoine, Bernard Page, Julie Cremniter, Jean-Louis Gaillard, Françoise Leturdu, Jean-Pierre Sollet, Gaëtan Plantefève, Xavière Panhard, France Mentré, Estelle Marcault, Florence Tubach.

** The Septicoli Group: Virginie Zarrouk, Frederic Bert, Marion Duprilot, Véronique Leflon, Naouale Maataoui, Laurence Armand, Liem Luong Nguyen, Rocco Collarino, Anne Lise Munier, Hervé Jacquier, Emmanuel Lecorché, Laetitia Coutte, Camille Gomart, Ousser Ahmed Fateh, Luce Landraud, Jonathan Messika, Elisabeth Aslangul, Magdalena Gerin, Alexandre Bleibtreu, Mathilde Lescat, Violaine Walewski, Frederic Mechaï, Marion Dollat, Anne-Claire Maherault, Michel Wolff, Mélanie Mercier-Darty, Bernadette Basse.

Résumé

Escherichia coli est la principale cause de bactériémies et est associée à une mortalité importante. Comme le montrent les études récentes, un nombre limité de clones est impliqué dans ces infections et certaines de ces analyses sont parvenues à capturer l'émergence de certaines de ces souches pandémiques. Cependant les données sur la dynamique d'évolution à l'intérieur même de ces séquences types (ST) sont limitées.

Afin de mieux comprendre l'évolution des souches responsables de bactériémies tout en limitant les biais épidémiologiques, nous avons comparé les génomes de 912 isolats de *E. coli* responsables de bactériémies chez l'adulte, collectés au cours de deux études multicentriques prospectives en région parisienne à 12 années d'intervalles. Nous avons analysé ces génomes à différents niveaux de granularité : le phylogroupe, le ST complexe (STc) et le ST tout en prenant en compte la résistance aux antibiotiques et le contenu en gènes de virulence ainsi que la diversité antigénique.

Ces comparaisons mettent en évidence une alternance de stabilité et de modification au cours du temps, en fonction de l'échelle utilisée. Globalement, nous avons observé une augmentation de la résistance antibiotique liée à un nombre restreint de déterminants génétiques alors que les distributions en termes de phylogroupes et de gènes de virulence restent stables. L'analyse des STc souligne la pauci-clonalité de la population, avec seulement 11 STc responsables de 73% des épisodes, parmi lesquels cinq dominent (STc73, STc131, STc95, STc69, STc10). Cependant, au sein de ces STc nous avons observé des changements majeurs comme l'expansion du clone pandémique mondial STc131 au dépend du clone qui dominait jusqu'alors, le STc95. Par ailleurs, l'analyse de la diversité des STc131, 95 et 69 montre une refonte importante de la population avec le remplacement de certains clones parfois couplés à l'acquisition indépendantes de facteurs de virulence parmi lesquels les gènes *pap*, et surtout l'allèle *papGII,* sont fréquemment retrouvés. En outre, le STc10, connu pour son caractère commensal, affiche une diversité antigénique élevée comme en témoigne les nombreuses combinaisons de sérotype O:H/*fimH*, indépendamment de l'année d'isolement des souches.

Ces résultats suggèrent qu'au sein des souches capables de réaliser une bactériémie, il existe une diversité génétique importante mais néanmoins spécifique, et que certains clones pathogènes extra-intestinaux hautement spécialisés subissent des remodelages fréquents de leur population à l'échelle du ST.

Au cours de ce travail, j'ai réalisé l'analyse et la comparaison des génomes de deux collections afin d'identifier les variations de population aux différentes échelles prises en compte, et j'ai participé à la rédaction de la publication.

Abstract

Background

Escherichia coli is the leading cause of bacteremia, associated with a significant mortality. Recent genomic analyses revealed that few clonal lineages are involved in bacteremia and sometimes captured the emergence of some of the most successful ones. However, data on within sequence type (ST) structure evolution are lacking.

Methods

To gain insight into the evolution of bacteremia strain populations over time avoiding epidemiological biases, we compared whole genome sequences of 912 *E. coli* isolates responsible for bacteremia from two multicenter clinical trials that were conducted in the Paris area 12 years apart. We analyzed the strains at different levels of granularity, i.e. the phylogroup, the ST complex (STc) and the within STc clone taking into consideration the phylogeny, the resistance and virulence gene content as well as the antigenic diversity of the strains.

Results

From these analyses we found a mix of stability and changes over time, depending on the level of comparison. Overall, we only observed an increase in antibiotic resistance associated to a restricted number of genetic determinants whereas phylogroup distribution and virulence gene content remained constant. Focusing on STcs highlighted the pauci-clonality of the populations, with only 11 STcs responsible for more than 73% of the cases, dominated by five STcs (STc73, STc131, STc95, STc69, STc10). However, some of them underwent dramatic variations, such as the worldwide pandemic STc131, which replaced the previously predominant STc95. Moreover, within STc131, 95 and 69 diversity analysis revealed high dynamic with an overhaul of the population linked to clonal replacement sometimes coupled with independent acquisitions of virulence factors such as the pap gene cluster bearing *papGII* allele. Besides, the STc10 exhibited huge antigenic diversity evidenced by numerous O:H serotype/*fimH* allele combinations whatever the year of isolation.

Conclusion

Altogether, these data suggest that the bloodstream niche is occupied by a wide but specific phylogenetic diversity and that highly specialized extra-intestinal clones undergo frequent turnover at the within ST level.

Keywords: *Escherichia coli* bacteremia; Antibiotic resistance; Extraintestinal infection; Pandemic clones

Background

Escherichia coli bacteremia represent a considerable and increasing burden in human medicine (Russo and Johnson 2003) due to the increase both in incidence of the disease and antibiotic resistance of the strains (Vihta et al. 2018). The in-hospital mortality of these bloodstream infections is about 10-20% (Lefort et al. 2011; Abernethy et al. 2015; Yoon et al. 2018; de Lastours et al. 2020) and the determinants associated with death are still debated. A major role has been attributed to the host conditions (comorbidities) and the portal of entry (non-urinary) (Martínez et al. 2006; Jauréguy et al. 2007; Lefort et al. 2011; Abernethy et al. 2015; Mora-Rillo et al. 2015; Yoon et al. 2018; de Lastours et al. 2020). However, several studies have also pointed the role of bacterial characteristics such as clonal group belonging and presence of specific virulence genes (Jauréguy et al. 2007; Lefort et al. 2011; Mora-Rillo et al. 2015; Yoon et al. 2018; de Lastours et al. 2020), whereas the impact of antibiotic resistance is unclear (Kang et al. 2010; Abernethy et al. 2015; Yoon et al. 2018; de Lastours et al. 2020).

The reservoir of the strains involved in extra-intestinal infections such as bacteremia is the gut where *E. coli* behaves as a commensal (Tenaillon et al. 2010). The extra-intestinal intrinsic virulence of *E. coli* strains has been evaluated thoroughly using a mouse model of sepsis (Picard et al. 1999; Johnson et al. 2006; Johnson et al. 2019) and depends on the phylogenetic/clonal background (mainly phylogenetic group B2) and the presence of virulence genes encoding for iron capture systems, protectins, invasins, adhesins and toxins. Of note, this intrinsic virulence is not associated with patients' death (Landraud et al. 2013). It has been proposed that these virulence factors (VFs) have been in fact selected for their benefic actions in the commensal niche, virulence being a by-product of commensalism (Le Gall et al. 2007).

E. coli epidemiology has changed during the last 20 years with the emergence of the sequence type (ST) 131 clone/clonal complex (CC) belonging to the B2 phylogroup that has disseminated world-wide, often associated with multidrug resistance (Nicolas-Chanoine et al. 2014). As there is a geographic component in *E. coli* epidemiology, at least for commensal strains (Tenaillon et al. 2010), the understanding of the epidemiologic evolutions of *E. coli* strains requires studies performed at a local level. Such studies are rare. Our group has compared commensal *E. coli* strains in the Paris area between 1980 and 2010 and showed a substantial increase in B2 phylogroup strains, VF content and antibiotic resistance (Massot et al. 2016). Two English studies have described the population structure of *E. coli* causing bacteremia between 2001 and 2012 in the UK and Ireland and have evidenced the rise of the ST131 as well as antibiotic resistance (Day et al. 2016; Kallonen et al. 2017).

In this context, we have thoroughly studied, using whole genome sequencing, two collections of *E. coli* strains responsible for bacteremia gathered in 2005 and 2016-7 in teaching hospitals from the Paris area to evidence population structure dynamics over a 12-year period. First, we studied both collections on a global scale, including phylogroup determination, resistance and virulence content comparison. Second, we went deeper at the ST complex (STc) scale. Lastly, we put special emphasis on the description of within STc evolution, as no data are available at this level of granularity. Through this analysis, we obtained a detailed picture of the evolution of the population involved in bacteremia as well as elements that could explain this phenomenon.

Methods

Clinical studies and strain collections

The studied strains originate from adult patients enrolled in two multicenter clinical trials, Colibafi (Ethics Committee CPP Hôpital Saint Louis, Paris, France: number 2006-4) and Septicoli (ClinicalTrials.gov: identifier NCT02890901) in 2005 and 2016-7, respectively, devoted to the identification of risk factors of mortality in *E. coli* bloodstream infections (Lefort et al. 2011; de Lastours et al. 2020). To limit epidemiologic biases, we focused only on strains isolated in hospitals located in the Paris area (8 among 15 hospitals for the Colibafi collection and all 7 hospitals for the Septicoli collection, 4 hospitals being common between the two studies). All the Paris area hospitals belong to the same institution of University hospitals, the “Assistance Publique-Hôpitaux de Paris” (www.aphp.fr).

A total of 912 strains from 912 patients (one bloodstream strain per patient) were studied, among which 367 among the 374 strains (the remaining seven didn't grow) isolated in the Paris area from the Colibafi collection, hereinafter referred to as “2005” (Lefort et al. 2011) and 545 strains from the Septicoli collection, hereinafter called “2016-7” (de Lastours et al. 2020).

Genome sequencing

Strain genomes were sequenced using Illumina NextSeq technology as previously described (de Lastours et al. 2020). The genomes from the “2005” collection were sequenced in the present work (Bioproject PRJEB39260) whereas the genomes from the “2016-7” collection were previously available (Bioproject PRJEB35745) (de Lastours et al. 2020).

Genome global analysis and typing

All genomes were assembled with shovill version 1.0.4 using SPAdes v3.13.1 and standard parameters, and then annotated with Prokka 1.14.5 (Bankevich et al. 2012; Seemann 2014).

Genome typing was performed as previously described including species identification, phylogrouping, multi-locus sequence type (MLST) determination according to the Warwick scheme and *in silico* serotyping (Ingle et al. 2016; Beghain et al. 2018; Bourrel et al. 2019). Resistance and virulence genes were scanned as previously described, based on Resfinder and a custom database, respectively (Zankari et al. 2012; Bourrel et al. 2019). We also searched for point mutation responsible for betalactam (*ampC* promoter) and fluroquinolone (*gyrA/B*, *parC/E*) resistance (Zankari et al. 2012). Virulence genes were classified into 6 families: invasin, protectin, toxin, adhesin, iron acquisition and miscellaneous (Bourrel et al. 2019). Contig locations, *i.e.* plasmid or chromosome were predicted using PlaScope (Royer et al. 2018). Integrons were searched using Integronfinder with standard parameter (Cury et al. 2016).

A pangenome was computed using Roary v3.12 with default parameters (Page et al. 2015). Then, a phylogeny based on a core genome alignment (Minh et al. 2020) was performed with Iqtree v1.6.12 following the protocol described in (Touchon et al. 2020). We determined STc based on single and double locus variant of MLST profiles in compliance with the phylogenetic tree (Additional file 1: Figure S1). Only the top 10 STcs of each collection were further studied.

Resistance phenotype prediction

Beta-lactamase types and alleles were controlled based on the beta-lactamase database (Naas et al. 2017) and the bacterial antimicrobial resistance reference gene database when necessary (Feldgarden et al. 2019). From these results, we predicted phenotypic resistance to clinically relevant antibiotics of 4 classes: i) betalactams including ampicillin (AMP), piperacillin/tazobactam (TZP), cefotaxime/ceftazidime (CTX/CAZ), cefepime (FEP), carbapenems (CARB); ii) fluoroquinolones (FQ); iii) aminoglycosides including gentamicin (GEN) and amikacin (AMK); and iv) cotrimoxazole (SXT), as previously described (Ruppé et al. 2020). The predicted phenotypes arising from presence of genes or mutations are described in Table S1 (Additional file 2).

Within STc analysis

From the phylogenetic tree we computed patristic distances (the sum of branch lengths in the path between two genomes in the phylogenetic tree) between all strains among the main 11 STcs, as described in (Touchon et al. 2020), using the function cophenetic from R package “ape” (Paradis and Schliep 2019). We also computed the genome fluidity (Kislyuk et al. 2011) for all pairs of strains from the pangenome considering only variables gene families (*i.e.* present in less than 95% of strains). This ratio ranges from 0 to 1: the higher the ratio is, the higher the genome fluidity is *i.e.* the diversity in terms of gene composition.

Moreover, to characterize more thoroughly the five main STcs (STcs 131, 95, 73, 69 and 10), we aligned their genomes to a specific reference (EC958, UTI89, CFT073, UMN026 and K-12 genomes, respectively) using Snippy 4.4.0 with standard parameters (Seemann 2020). Then we constructed a phylogenetic tree using Iqtree v1.6.12 (Minh et al. 2020) after taking into account the recombinations using Gubbins (Croucher et al. 2015) with standard parameters. These SNP-based tree were then visualized and annotated with Itols (Letunic and Bork 2019). We classified strains according to subgroups/clades when possible. STc131 strains were classified among clades A, B and C based on canonical SNPs defined previously (Ben Zakour et al. 2016). For the STc95, we classified strains according to subgroups A, B, C, D, E or unassigned as described in (Gordon et al. 2017). Since no specific classifications are currently described for the STc73, STc69 and STc10, we constructed a SNP-based phylogenetic tree as described above and divided the population into meaningful subgroups (Additional files 3, 4 and 5: Figures S2, S3 and S4).

Pathogenicity island analysis

We performed a detailed analysis of the *papGII* genomic context for the STc69 and STc131 strains. First, we performed blastN alignments on the NCBI website using the contigs containing *papGII* as query to look for the closest circularized *E. coli* genome. Then, we extracted from these complete genomes the nucleic sequences of the PAI containing *pap* genes and aligned the reads of our strains to the corresponding reference sequence using Breseq (Deatherage and Barrick 2014). Finally, we checked if the whole length of the PAI was covered. We also checked for the presence of expected virulence genes depending on the closest PAI we found. Results are available in Table S2 (additional file 6) and Figures S5 and S6 (additional files 7 and 8) drawn using Easyfig (Sullivan et al. 2011).

Statistical analyses

The proportion of strains among each phylogenetic group and each STc was compared between the two collections (2005 and 2016-7) using Chi-2 test. Phylogroup E and *Escherichia* clades were grouped together and only the ten most prevalent STcs of each collection were considered. For each of the main STcs, we compared patristic distances and genome fluidity distributions between both collections using Wilcoxon-Mann-Whitney tests.

We performed an ANalysis Of Variance (ANOVA) to compare the distribution of VFs among the six main functional classes (adhesion, invasion, iron acquisition, miscellaneous, protectin and toxin) between 2005 and 2016-7 both at global scale and for the five STcs we focused on. Likewise, the proportion of strains predicted to be resistant on each of the nine antibiotics were compared at the global scale and for five STcs (STc131, STc95, STc73, STc69 and

STc10) between collections using Fisher exact tests, as well as the proportion of strains carrying complete integrons, clusters of *attC* sites lacking integron-integrases (CALIN) and integron integrase only (In0). Finally, the proportion of strains among the subgroups/clades of STc131, STc95, STc69 and STc10 were compared between collections using Fisher exact tests.

As multiple tests were performed, the p-values were adjusted using the Benjamini and Hochberg method (Benjamini and Hochberg 1995). All statistical analyses were performed using R software (R version 3.4.2). All tests were two-sided with a 5% type I error.

Results

Stability of phylogroup composition and VF content with an increase of antibiotic resistance level between 2005 and 2016-7 collections

We analyzed 367 and 545 strains of *E. coli/Escherichia* clades responsible for bacteremia in adults in 2005 and 2016-7, respectively, which had been collected in two previously published multicentric prospective studies performed in teaching hospitals belonging to the same institution (Assistance Publique-Hôpitaux de Paris) in the Paris area (Lefort et al. 2011; de Lastours et al. 2020). We conducted a global comparison at the phylogroup level, considering the whole 912 strains (Figure 1, additional file 9: Table S3). Only two and three strains (0.5%) were identified as *Escherichia* clades in the 2005 and 2016-7 collections, respectively, confirming the minor role played by these clades in human (Clermont et al. 2011). *E. coli* strains mainly belong to phylogroup B2 (51.2% in both collection) and to a lesser extent to phylogroup D (15.5% and 16%). Then in the 2005 collection, phylogroup A ranked third (11.7%), followed by B1 (7.6%) and C (7.4%) phylogroups, whereas in the 2016-7 collection phylogroup B1 ranked third (12.8%), followed by A (9.7%) and C (4.4%) phylogroups. However, these differences were not significant after multiple testing correction, even when taking into account the main portals of entry (*i.e.* urinary or digestive) of the bacteremia (additional file 9: Table S3). In terms of number of virulence genes classified in main functional categories, we did not observe significant differences either. In contrast, and as expected, when looking at the predicted resistance phenotype, strains from the 2016-7 collection were more resistant to nearly all antibiotic families than the 2005 ones (Figure 1). For example, predicted resistance increased from 4.9% to 17.4% for cefotaxime/ceftazidime, from 21.5% to 31% for fluoroquinolones and 1.6% to 7% for amikacin. In terms of resistance determinants, we found an increase in the number of oxacillinase and ESBL coding genes conferring resistance to wide spectrum betalactams, *qnr*, *aac*(6')-Ib-cr and gyrase mutations conferring

resistance to fluoroquinolones, and *aac*(3)-II/*aac*(3)-IV and *aac*(6')-Ib-cr genes conferring resistance to aminoglycosides (Additional file 10: Figure S7). At the phylogroup level, we observed a significant increase in resistance to piperacillin/tazobactam (TZP), cefotaxime/ceftazidime (CTX/CAZ), cefepime (FEP), fluoroquinolones (FQ), gentamicin (GEN), amikacin (AMK) and cotrimoxazole (SXT) only for B2 strains. We did not find any significant enrichment in integron related sequences.

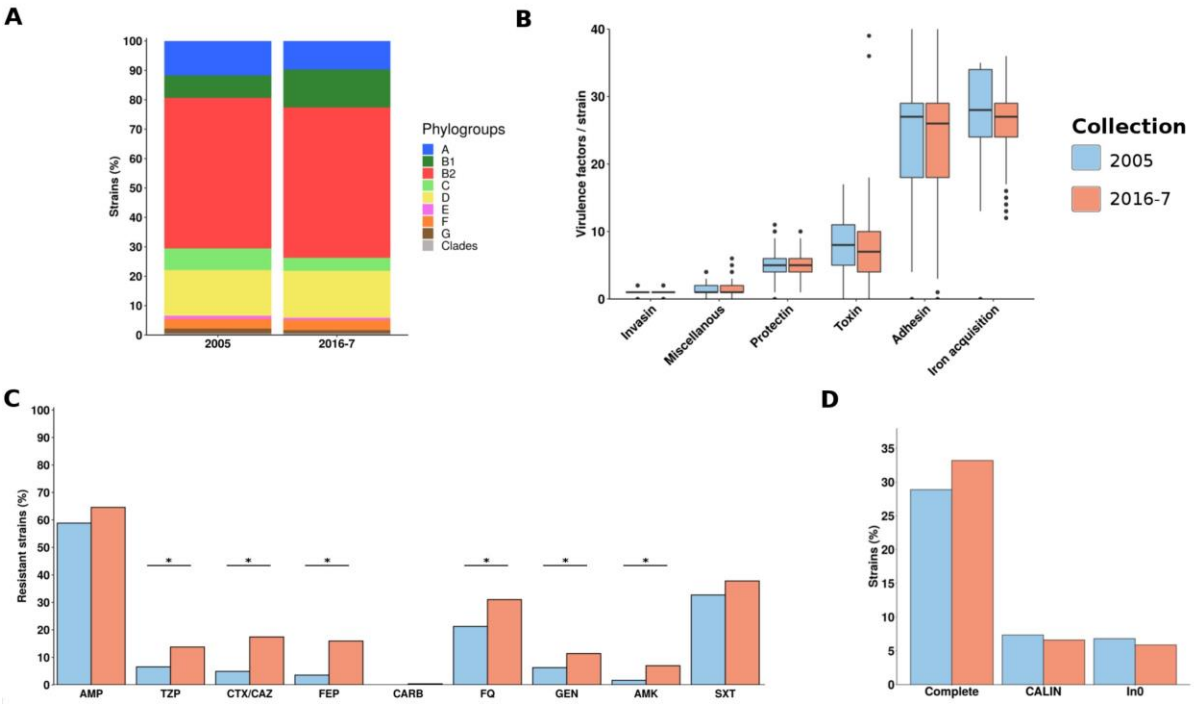


Figure 1. Global comparison of the 2005 and 2016-7 collections. (A) Phylogroup distribution of the strains. (B) Distribution of the number of virulence factors by strain among the six main functional classes of virulence. (C) Bar chart of predicted phenotypes of the strains. The results are presented as percentage of resistant strains for nine antibiotics of clinical importance. (D) Bar chart of the number of strains carrying complete integron, CALIN (clusters of *attC* sites lacking integron-integrases) and In0 (integron integrase only). Significant differences are highlighted by asterisks. AMP = ampicillin; TZP = piperacillin/tazobactam; CTX/CAZ = cefotaxime/ceftazidime; FEP = cefepime; CARB = carbapenems; FQ = fluoroquinolones; GEN= gentamicin; AMK = amikacin; SXT = cotrimoxazole.

The top 10 STcs are similar between the two collections but differ in frequency and in global genetic structure

In a second step, we compared the distribution of the top 10 STcs in both collections, corresponding to a total of 11 STcs (Table 1). These ten STcs represent 73.5% and 73% of

the strains in the 2005 and 2016-7 collections, respectively. Significant variations were observed between collections, notably the increase of the STc131, from 5.7% to 15%, and the decrease of the STc95, from 15.5% to 7%. We also noticed a slight increase of the STc69, from 8.7% to 11.6% that is not statistically significant and has not changed its ranking over years. Conversely, the STc58, which corresponds to the CC 87 according to the Pasteur Institute multilocus sequence typing (MLST) scheme (Skurnik et al. 2016), ranked 9th in 2005 (2.4%) but increased to 5.3% in 2016-7, making it the 6th isolated STc.

Table 1. Distribution of the main STcs in *Escherichia coli* strains from the 2005 and 2016-7 collections

STc* (Phylogroup)	Number of strains (%)	
	2005	2016-7
STc131 (B2)	21 (5.72)	82 (15.05)
STc73 (B2)	51 (13.9)	68 (12.48)
STc69 (D)	32 (8.72)	63 (11.56)
STc95 (B2)	57 (15.53)	38 (6.97)
STc10 (A)	28 (7.63)	33 (6.06)
STc58 (B1)	9 (2.45)	29 (5.32)
STc88 (C)	25 (6.81)	23 (4.22)
STc14 (B2)	17 (4.63)	23 (4.22)
STc12 (B2)	17 (4.63)	18 (3.3)
STc141 (B2)	5 (1.36)	13 (2.39)
STc127 (B2)	8 (2.18)	8 (1.47)

*Only the top 10 STc of each collection are considered
**Benjamini-Hochberg correction

To get a more detailed picture of the evolution of these 11 STcs between 2005 and 2016-7, we computed both patristic distances and genome fluidity between strains of the same STc in a given collection (Figure 2). The first metric reflects the genetic divergence at the nucleotide level whereas the second one indicates the diversity in terms of gene content. Comparison between STcs, whatever the date of isolation of the strains, showed that the STc10 behaved differently from the other STcs as it had a greater genetic diversity with both metrics than the other STcs (p<0.05). When comparing the evolution of the diversity over time between the two collections, we observed an increase in both patristic distances and genome fluidity, especially for STc131, STc69, STc95 and STc10.

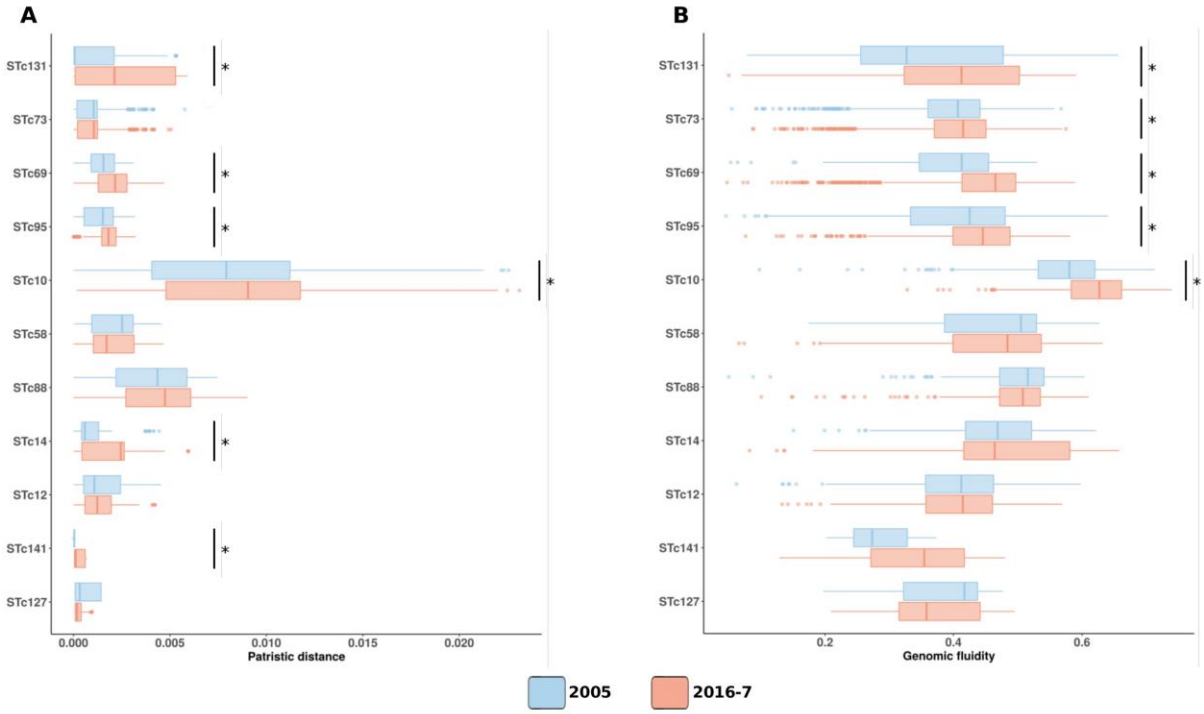


Figure 2. (A) Distribution of the patristic distances between all strains of a given STc in a given collection. (B) Distribution of the genome fluidity between all strains of a given STc in a given collection. Significant differences are highlighted by asterisks (Benjamini-Hochberg corrected p-value < 0.05).

Fine scale analysis of the big four ExPEC STcs

To document more thoroughly the evolution of population structure within STcs, we focused on the four main STcs, namely STc131, STc95, STc73 and STc69. These four STcs, which belong to B2 and D phylogroups, encompass typical ExPEC strains and are currently the major ones involved in bacteremia worldwide (Denamur et al. 2020).

STc131. The STc131, from phylogroup B2, is characterized by a stepwise diversification with two serotypes (O16:H5 and O25:H4), three clades (A, B and C) and three *fimH* alleles (41, 22 and 30), all correlated (Ben Zakour et al. 2016) (Figure 3A, additional files 11 and 12: Figure S8, Figure S9). In our data set, we observed a slight decline of the clade C (O25:H4) in 2016-7 balanced by a slight increase in clade A (O16:H5), which corresponds to the more diverged clade (Figure 4, additional file 11: Figure S8). These changes are reflected by an increase of their genetic diversity (Figure 2). Moreover, clade C strains from the 2005 collection mainly belong to subclade C1, whereas in 2016-7 they are predominantly from subclade C2, which is frequently resistant to both fluoroquinolones and 3GC (Additional file 11: Figure S8). In

2016-7 we also observed several strains in the clade B with the *fimH*30 allele, as previously described (Duprilot et al. 2020).

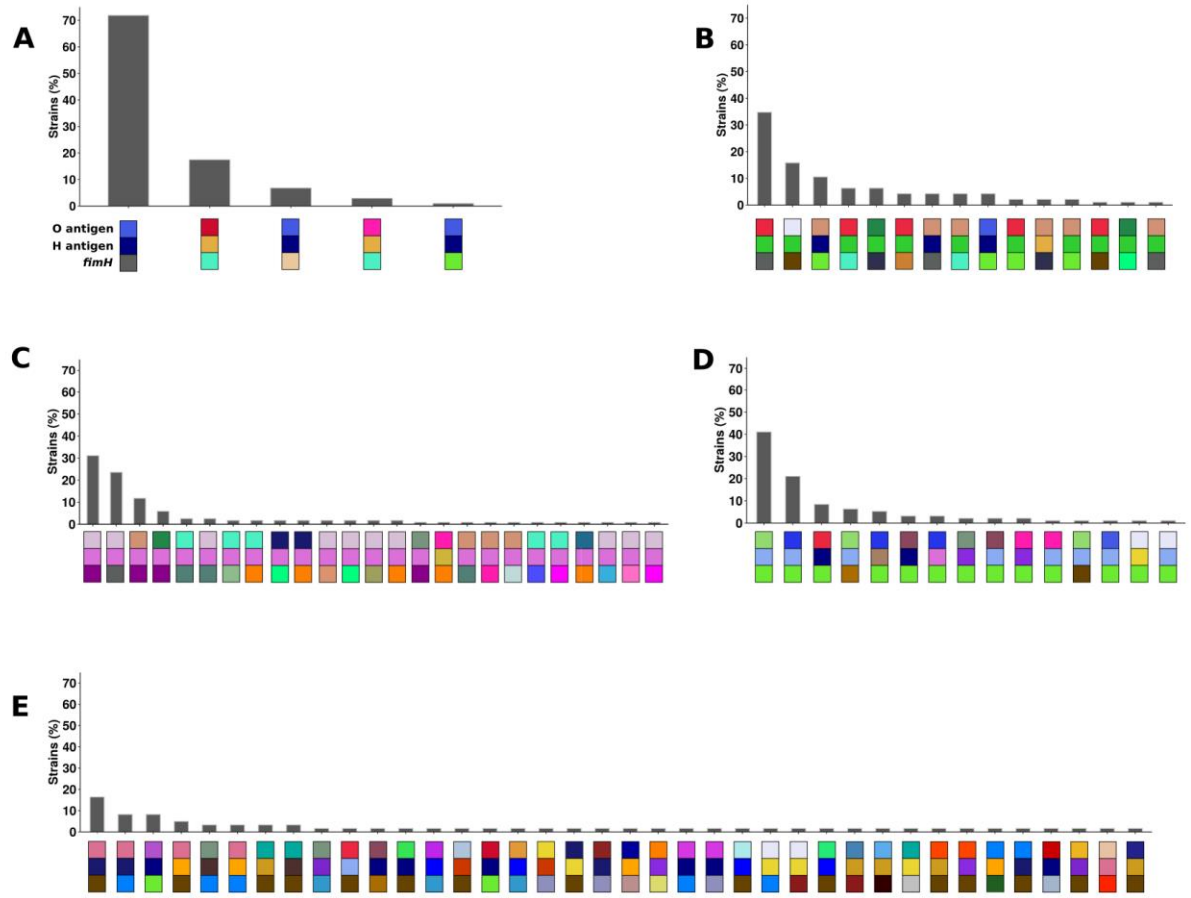


Figure 3. Distribution of the combinations O:H/*fimH* among the big four STcs and the STc10. (A) STc131. (B) STc95. (C) STc73. (D) STc69. (E) STc10. The combinations O:H/*fimH* are schematically represented by colored squares. The exact serotype and *fimH* allele can be found on the Figure S9.

In terms of virulence, only adhesins were significantly increased, and in particular *hek* and the *pap* genes (*papC*, *papD*, *papF*, *papGII*, *papH*, *papJ* and *papK*) which raised from 0 to 9.5% and 41.5 to 46.3%, respectively. All of these VFs were predicted to be located on the strains' chromosomes, and the *pap* genes were always co-localized on the same genomic region. They were distributed mostly in genomes from the subclade C2, but also in three closely related strains from clade A having all *bla*CTX-M-27 coding gene and the mutation GyrA S83L. We further analyzed the genetic context of these *pap* genes as they are usually found on pathogenicity islands (PAIs). In subclade C2, we were able to link the *pap* genes to at least three main PAIs in accordance both with the phylogeny and the virulence gene content of the

strains (Additional file 6: Table S2): one PAI (RHBSTW-00440) being probably a shortened version of another one (Ecol_AZ146) (Additional file 7: Figure S5). In the strains from clade A carrying *papGII*, we found an homology with the PAI WP5-S18-ESBL-09. This PAI also presents a strong similarity with the PAI Ecol_AZ146 (Additional file 7: Figure S5). Of interest, most of these PAIs were inserted next to the tRNA-PheU and few strains present alternative integration sites (tRNA-PheV or between *glnH* and *glnP* genes) (Additional file 6: Table S2). Taken together, these data suggest multiple transfer events of a PAI containing the *pap* gene cluster in the STc131. Besides, only five O:H/*fimH* combinations were evidenced (Figure 3A, additional file 12: Figure S9). In terms of resistance, we observed a tendency towards more resistance including TZP, CTX/CAZ, FEP and AMK, however not statistically significant after correction.

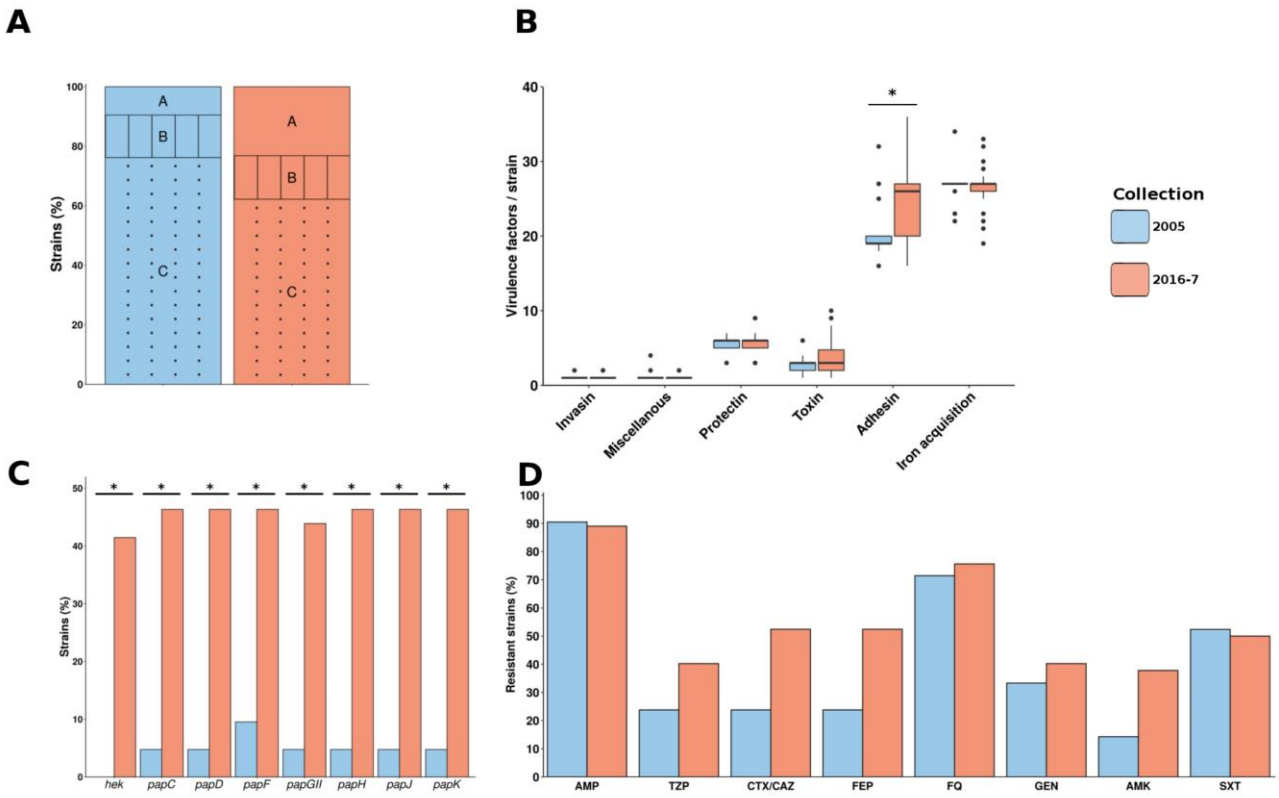


Figure 4. Comparison of STc131 strains in the 2005 and 2016-7 collections. (A) Distribution of strains in the three clades of STc131 described by Ben Zakour *et al*.. (B) Distribution of the number of virulence factors by strain among the six main functional classes of virulence. (C) Distribution of the adhesins in both collections. Only adhesins with a significant difference between 2005 and 2016-7 are presented (Benjamini-Hochberg corrected p-value < 0.05). (D) Predicted phenotypes of the strains. The results are presented as percentage of resistant strains for eight antibiotics of clinical importance (no carbapenem-resistant strain has been found). Significant differences are highlighted by asterisks. AMP = ampicillin; TZP = piperacillin/tazobactam; CTX/CAZ = cefotaxime/ceftazidime; FEP = cefepime; FQ = fluoroquinolones; GEN= gentamicin; AMK = amikacin; SXT = cotrimoxazole.

In summary, we observed an increase in the number of strains from clade A (O16:H5, *fimH*41) within the STc131 including closely related strains isolated in 2016-7 carrying the same PAI and resistance genes. The proportion of clade B remains stable over time, but in 2016-7 we noticed the emergence of uncommon strains exhibiting the *fimH*30 allele. The clade C decreased slightly and we observed a switch to the C2 sub-clade with *bla*CTX-M-15 and GyrA S83L/D87N mutations as well as a high frequency of the *papGII* gene in related PAIs. Thus, the rise of the STc131 over time comes both with the emergence of specific clones and the acquisition of virulence factors (*papGII*) through independent genetic events, while maintaining a very low antigenic diversity.

STc95. The STc95, also from the B2 phylogroup, has diversified rapidly leading to a star-like phylogeny with five subgroups (A to E) and serotypes specific to subgroups (O18:H7 and B, O45:H7 and D) or shared between subgroups (O1:H7 and A, C, D) (Gordon et al. 2017). Between our two collections, a major change in subgroup composition was observed, with a significant decrease of subgroup D strains and an increase of subgroup A strains with the emergence of a O1:H7 *fimH*41 clone and subgroup C strains (O25b:H4/O1:H7 *fimH*27) (Figure 5, additional file 13: Figure S10). The strains from the 2016-7 collection carry less iron acquisition related VFs, especially *iroB, C, D, E, N, iss_12* and *iucA, B, C, D.* These genes are almost exclusively found on plasmidic contigs, *iro* genes and *iss_12* being almost always co-localized on the same contig and *iuc* genes on another. We also found less VFs of the miscellaneous class, partly due to the less frequent presence of *etsC* gene encoding a putative type I secretion outer membrane protein. As these iron acquisition and miscellaneous genes are typically carried by the pS88 plasmid (accession number: CU928146), we searched for its presence in our strains using blastN alignments. The plasmid was detected in nearly all strains, including subgroup D, but was not found in the emerging subgroup A. The antigenic diversity is constrained with 15 O:H/*fimH* combinations (Figure 3B, additional file 12: Figure S9). Finally, we observed a tendency toward greater antibiotic resistance (TZP, CTX/CAZ, FEP, FQ, GEN, SXT), but not statistically significant as in STc131.

In summary, while STc95 strains were the most numerous in the 2005 collection, their decrease in 2016-7 is associated with a more balanced population structure. This is linked to the emergence of subgroups C and A, the latter lacking the pS88-related genes implicated in iron acquisition.

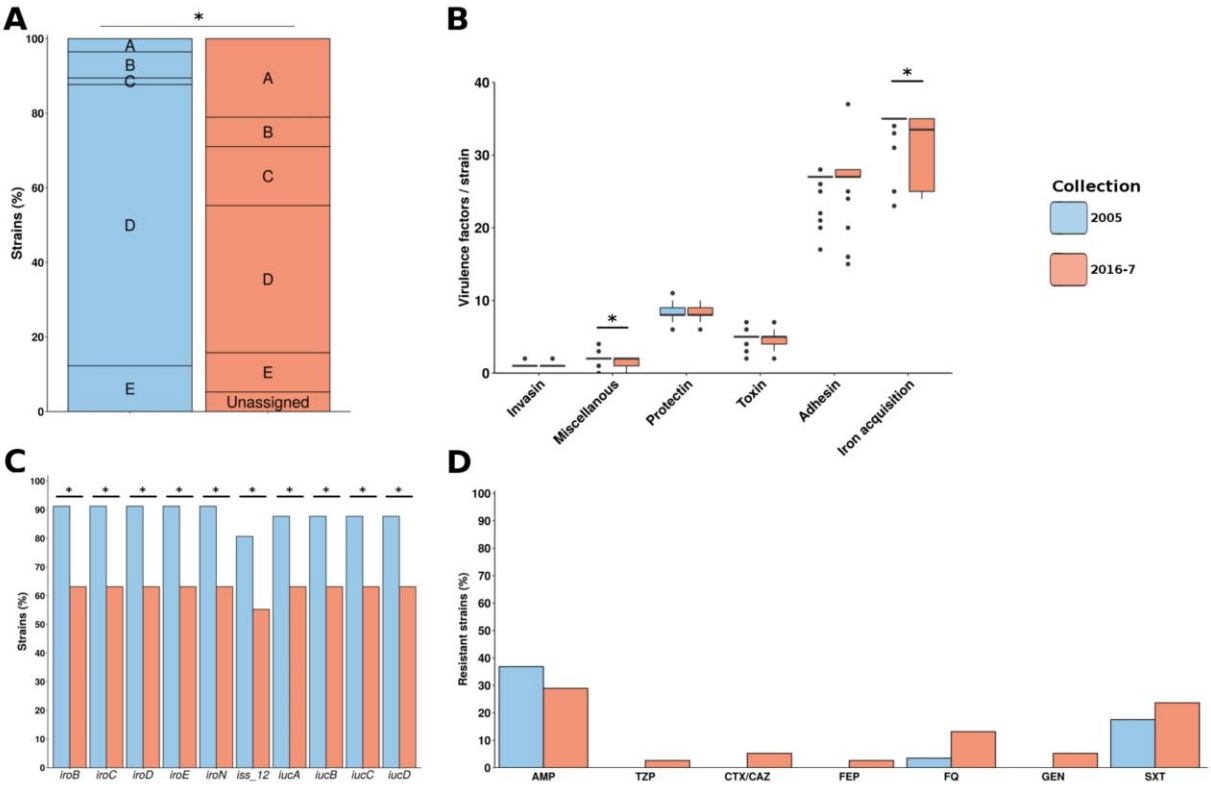


Figure 5. Comparison of STc95 strains in the 2005 and 2016-7 collections. (A) Bar chart of the distribution of strains in the five subgroups of STc95 described by Gordon *et al.*. (B) Distribution of the number of virulence factors by strain among the six main functional classes of virulence. (C) Distribution of the iron acquisition related genes in both collections. Only virulence factors with a significant difference between 2005 and 2016-7 are presented. (D) Predicted phenotypes of the strains. The results are presented as percentage of resistant strains for seven antibiotics of clinical importance (no strain resistant to carbapenems and amikacin has been found). Significant differences are highlighted by asterisks. AMP = ampicillin; TZP = piperacillin/tazobactam; CTX/CAZ = cefotaxime/ceftazidime; FEP = cefepime; FQ = fluoroquinolones; GEN= gentamicin; SXT = cotrimoxazole.

STcs 73 and 69. As no large-scale population genetic structure was available for the STcs 73 and 69 (B2 and D phylogroups, respectively), we merged the strains of the two collections to have an overview of the strain diversity within these STcs. Concerning the STc73, no subgroups were identified in the SNP-based phylogenetic tree. Indeed, the tree showed a polytomy (*i.e.* a multifurcation) with low values of patristic distances (Figure 3, additional file 3: Figure S2), indicating very few diversification. Moreover, a total of 25 O:H/*fimH* combinations were observed (Figure 3C, additional file 12: Figure S9), preventing any clustering based on these elements. We did not evidence any cluster based on collection origin but observed intermixed strains (Additional file 3: Figure S2). In terms of virulence and

resistance, we did not identify any significant difference. All these elements point to the absence of emerging clone over the years but a relatively high antigenic diversity.

Concerning the STc69, the SNP-based phylogenetic tree reconstructed from the strains of both collections delineated at least two subgroups (namely A and B) with short internal branches (Additional file 4: Figure S3). Almost all the strains exhibit a *fimH*27 allele (n = 88/95) and we did not find significant variations between 2005 and 2016-7 subgroup repartition (Figure 6). At the STc scale, the strains of the 2016-7 collection have a higher number of adhesins than the strains from 2005, but not statistically significant. As in the STc131, these additional adhesins are part of the *pap* gene cluster (*papC*, *papD*, *papF*, *papGII*, *papH*, *papJ*, *papK* and *tia*). This increase in adhesins is partly linked to a clone of 14 strains of the 2016-7 collection in subgroup A. These strains exhibit an O15:H18 serotype and carry for almost all the *sul1*, *sul2*, *dfrA*, *papGII* genes and a truncated *hlyC* gene (Additional file 4: Figure S3). We also observed the emergence of a O117:H4 *fimH*27 clone in the B subgroup exhibiting the *pap* genes. The analysis of the genetic context of *papGII* showed at least two different paths of acquisition of the adhesins (Additional file 6: Table S2). On one hand, in subgroup B we found the typical PAI of STc69 archetypal strain UMN026 (NC_011751.1) inserted in tRNA-PheU, sometimes partly deleted, which usually carries *pap* genes, *iha*, *sat*, *iutA*, *iucA* and the capsule coding genes *kpsMDE* (Additional file 8: Figure S6). This PAI is closely related to the ATCC25922 PAI found in the STc131 at the difference of the *hly* genes that are absent (Additional file 7: Figure S5). On the other hand, in subgroup A we found a PAI carrying *papGII*, *tia*, *ireA* and inserted in tRNA-PheU (Additional file 8: Figure S6). The level of antigenic diversity is similar to the STc95 one with 15 O:H/*fimH* combinations (Figure 3D, additional file 12: Figure S9). Strains tend to be slightly more resistant to some antibiotics (AMP, TZP, CTX/CAZ, FEP, FQ), but not significantly.

In summary, we observed a slight increase in the number of STc69 strains over time, although not significantly. This change is partly explained by the emergence of two clones: the first in subgroup A with a O15:H18 serotype and carrying *papGII* on a uncommon PAI; the second in the subgroup B with a O117:H4 serotype and carrying the archetypal PAI of ST69.

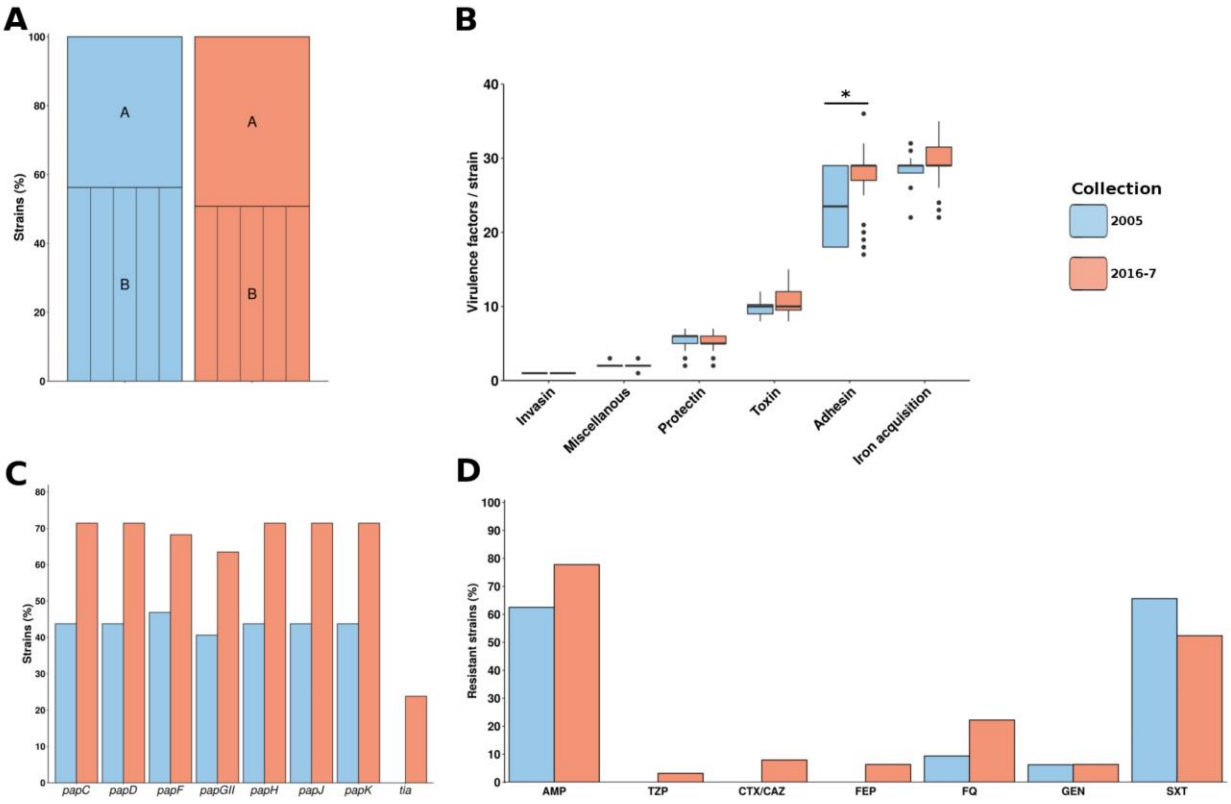


Figure 6. Comparison of STc69 strains in 2005 and 2016-7 collections. (A) Bar-chart of the distribution of strains in the two subgroups of STc69 defined by the phylogenetic analysis. (B) Distribution of the number of virulence factors by strain among the six main functional classes of virulence. (C) Distribution of the adhesins in both collections. Only virulence factors with the most significant differences (*i.e.* significant before multiple test correction) between 2005 and 2016-7 are presented. (D) Predicted phenotypes of the strains. The results are presented as percentage of resistant strains for seven antibiotics of clinical importance (no strain resistant to carbapenems and amikacin has been found). Significant differences are highlighted by asterisks. AMP = ampicillin; TZP = piperacillin/tazobactam; CTX/CAZ = cefotaxime/ceftazidime; FEP = cefepime; FQ = fluoroquinolones; CARB = Carbapenems; GEN= gentamicin; SXT = cotrimoxazole.

The particular case of the STc10

In terms of prevalence, the fifth STc is the STc10, which encompasses typically commensal strains devoid of intrinsic extra-intestinal virulence (Picard et al. 1999). We observed for this STc a high level of diversity with both patristic distance and genome fluidity metrics (Figure 2). Two subgroups can be distinguished (Additional file 5: Figure S4), the subgroup A corresponding to the ST48 and the subgroup B corresponding to the ST10. Within the subgroup B, strains from a recently emerged lineage exhibit gyrase mutations and ESBL production for some of them. A huge antigenic diversity was observed with 38 widespread O:H/*fimH* combinations (Figure 3E, additional file 12: Figure S9). No difference was evidenced

between the collections in terms of subgroup repartition, virulence factors and antibiotic resistance and no clear pattern of emerging clone was observed (Additional file 5: Figure S4).

In summary, the STc10 exhibits a unique pattern of diversification with a huge antigenic diversity present in both collections and not linked to specific emerging clones.

Discussion

A recent study has highlighted the dynamic structure of the *E. coli* population in bacteremia over time (Kallonen et al. 2017) and captured the emergence of some of the most prevalent lineages nowadays, *i.e.* STc131, STc69. Other studies focused on the diversification of specific STs as the ST131 and gave clues on the process leading to within ST clade formation (Ben Zakour et al. 2016; McNally et al. 2019). However, to our knowledge, no studies have analyzed diversification within STc using time-series data. From two epidemiologically comparable collections of *E. coli* strains isolated from bacteremia during two multicentric clinical trials conducted 12 years apart in the same institution in the Paris area, we were able to describe the evolution of the bacterial population structure over time at different levels of granularity.

The first striking result of our study is the remarkable stability in phylogroup composition and virulence gene content between the 2005 and 2016-7 collections (Figure 1). The proportion of B2 phylogroup strains (53%) is surprisingly identical in the two collections despite an observed increase in the frequency of B2 and their VF content for commensal strains isolated in the same area during the period 1980-2010 (Massot et al. 2016). The incidence of B2 strains in bacteremia is dependent of the portal of entry, urinary tract infections being associated with the higher proportion, *i.e.* 60% (Clermont et al. 2017) (Additional file 9: Table S3). Of note, this stability of B2 phylogroup strain proportion was already observed in a collection of 34 bacteremia strains from urinary portal of entry isolated in the 1980's in one of the hospital of the present study where carboxyl esterase of B_2 type (corresponding to B2 phylogroup) strains represented 56% (Picard and Goullet 1989).

This stability suggests that a specific pattern of phylogroup diversity is adapted to the bacteremia lifestyle of the strains, probably due to phylogroup specific characteristics. Such characteristics could be linked to metabolic processes, as genes involved in metabolism were found differentially represented at the phylogroup level (Touchon et al. 2020). For example, genes involved in aromatic compound degradation are negatively and positively associated with B2 and B1 phylogroup strains, respectively (Touchon et al. 2020). Metabolic functions are fundamental for adaptation to different nutritional niches (Monk et al. 2013) and survival in the

face of bactericidal defense mechanisms (Alteri and Mobley 2012). According to the portal of entry (urinary, digestive, pulmonary), different phylogroups could be selected due to their ability to grow in these environments.

Beyond this phylogroup repartition stability, we observed, as already reported in commensal strains (Massot et al. 2016), a major increase in antibiotic resistance of many classes, *i.e.* betalactams, fluroquinolones and aminoglycosides. This increase in antibiotic resistance is only statistically significant in B2 strains, due mostly to the increase of the STc131, but there is nevertheless a slight tendency in all the studied STcs. Integron analyses failed to identify a significant differences suggesting that such mobile genetic element are not the main factor of this elevated antibiotic resistance.

The second striking result is the stability of the global prevalence of the 10 main STcs that represent almost ¾ of the strains associated to modification of prevalence for some of them. The pauci-clonality of the strains responsible for bacteremia has also been observed in a study from UK (Kallonen et al. 2017), where our defined top 10 STcs represent at least 67% of their isolates. These data contrast with the commensal fecal ones where 23 STc/ST are necessary to reach 73.5% of the whole population (n=206/280 strains) (Massot et al. 2016; Clermont et al. 2017). However, among these highly prevalent STcs, we found a major switch in the B2 phylogroup corresponding to an increase of the STc131 compensated by a decrease of the STc95 (Table 1). One of the main differences between these two STcs is their antibiotic resistance phenotypes, the STc131 being multi-resistant (Nicolas-Chanoine et al. 2014; Ben Zakour et al. 2016), whereas the STc95 is multi-sensitive, possibly due to the presence of restriction modification systems precluding the gain of foreign DNA (Stephens et al. 2017). Of interest, the intrinsic extra-intestinal virulence in a mouse model of sepsis was reported similar for both STcs (Johnson et al. 2012; Gordon et al. 2017). It is also interesting to note the increase of STc58 (CC87), although not significant, which now ranks 6th in the 2016-7 collection, together with a slight increase in resistance, although also not significant (data not shown). This STc of the B1 phylogroup, rarely isolated in humans, has been shown to originate from animals and spread in humans, carrying antibiotic resistance determinants (Skurnik et al. 2016). Until now, this lineage was considered as a harmless commensal devoid of intrinsic extra-intestinal virulence (Skurnik et al. 2016). Its increase in bacteremia prevalence could indicate that STc58 has acquired additional virulence determinants. Further epidemiological studies are needed in the future to monitor this potential emerging group.

The third striking result is that, regardless of changes in STc prevalence over the study period, a change in the genetic structure of their population is underway, with increased diversity over

time in several STcs (Figure 2), according to various evolutionary scenarios discussed below.

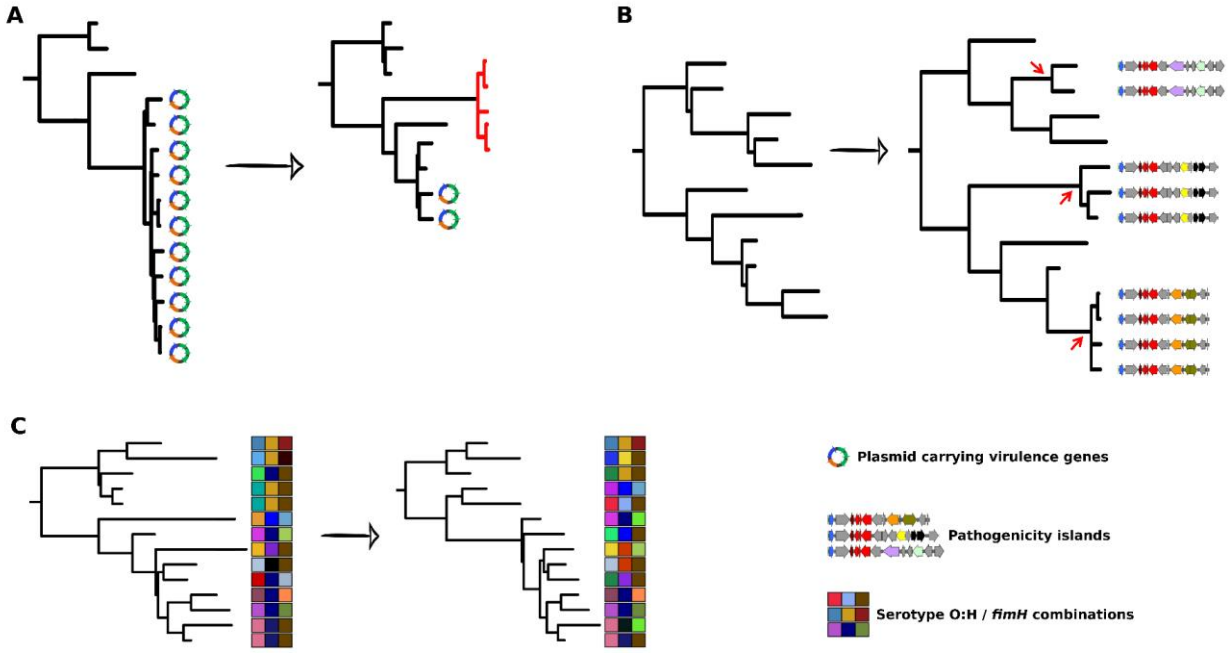


Figure 7. Schematic representation of the different scenarios leading to within STc dynamic. (A) Example of clonal replacement as observed in STc95. The represented plasmid corresponds to pS88 whereas the red terminal branches correspond to the emerging O1:H7 *fimH*41 subgroup A clone. (B) Multiple acquisitions of related PAIs associated to clonal expansion as observed in STc69 and STc131. The *pap* gene cluster with the *papGII* allele is represented in red on genetic maps. Red arrows indicate the acquisition of PAIs. (C) High antigenic diversity at a given time and over time, as observed in STc10. This pattern corresponds probably to multiple recombination events at the main chromosomal hot spots (*rfb*, *fimH*).

(i) Clonal replacement. This scenario is observed in both declining (STc95) and successful (STc131) lineages. Within the STc95, clonal replacement (subgroup A and C strain increase while subgroup D strain decrease) is associated with the decrease of plasmid borne genes implicated in iron acquisition (Figure 7A). It can be hypothesized that these virulence genes have contributed to the past success of the STc95. Within the STc131, the emergence of the clade B *fimH*30 clone, which shares some genetic features with the C clade, exhibits a reduced virulence in mice as compared to other B clade strains (Duprilot et al. 2020) and is not antibiotic resistant. This clone needs to be monitored to assess its fate.

(ii) Clonal replacement associated to convergent evolution. We observed an increase in the frequency of specific VFs, especially adhesins with the *papGII* allele, due to multiple PAI arrivals in distinct clones, in both STc131 and STc69, and linked to clonal expansion (Figure 7B). Such convergent evolution is a strong sign of selection (Tenaillon et al. 2012) and has been involved frequently in the evolution of pathogenic *E. coli* (Reid et al. 2000; Denamur et

al. 2020). Nonetheless according to the level of clone divergence, this selection most likely occurred at more ancient times than the increase in frequency that took place over 2005-2017. In accordance with our data, *pap* gene increase was also observed in Kallonen *et al.* data where *papG* in ST131 raises from 8% in 2003 to 44% in 2012, as well as in a Spanish study where the authors found an increase of strains from STc131/clade C carrying *papGII* between 2006 and 2011 (Mamani et al. 2019).

Furthermore, within the ST131, resistance acquisition sometimes co-occurred with virulence increase. In clade A, the emerging clone exhibiting *papGII* virulence factor also harbors the GyrA S83L mutation and sometimes *bla*CTX-M-27. It will be interesting to see if this clone is expanding as observed for clade C2 that acquired the same genetic attributes (Ben Zakour et al. 2016).

(iii) Antigenic variation. The O polysaccharide and the H flagellin are major surface antigens (Wang et al. 2003; DebRoy et al. 2011). The fimbrial tip-positioned adhesive protein FimH is also a surface antigen (Tchesnokova et al. 2011). The O-antigen biosynthesis gene cluster and the *fim* operon are known as the two major hotspots of recombination on the *E. coli* chromosome that are under diversifying selection (Touchon et al. 2009). The diversity of these antigens is variable according to the STcs (Denamur et al. 2020). Variable patterns of serotype/*fimH* allele combinations can be evidenced in the five main STcs (Figure 3, additional file 12: Figure S9): very few combinations (STc131, n=5) the main one representing 70% of the isolates, few combinations (STc95, n=15; STc69, n=15) with the main ones representing 30-40% of the isolates, intermediate number of combinations (STc73, n=25) and high number of combinations widely distributed (STc10, n=38) with the main one representing less than 17% of the isolates. This indicates that the STc10 has a very specific pattern of diversification as it remains polyclonal and exhibits a huge antigenic diversity whatever the year of isolation (Figure 7C), in line with its commensal ecology (Kauffmann 1947). This diversity could help it to resist to the immune system.

Surprisingly, the STc73 that is remarkably stable in frequency is not affected by clonal replacement over time. Moreover, this stability is also observed both in terms of resistance and virulence. Although far below the antigenic diversity observed in the STc10 (Figure 3), the multiple combinations of O:H/*fim*H of the STc73 as well as the high frequency of *papGII* could participate to its evolutionary success.

In conclusion, our results suggest that, depending of the level of granularity considered, contrasting results are obtained when comparing bacteremia *E. coli* strains over a 12-year period: a remarkable stability in terms of phylogroup distribution, a global stability in terms of main STc distribution with an increase or a decrease of some specific STcs and huge

modifications within STcs with clonal interference, characterized by competition between variant clones in each STc, and large variation in frequency of virulence and sometimes resistance genes. This indicates a global evolutionary constraint at the phylogroup level, and to a lesser extent, at the STc level that is associated to a diversifying selection within STcs. The intra-STc dynamics could result from negative frequency-dependent selection, as suggested previously (Kallonen et al. 2017). This selection is probably anterior to the period studied in this work and could occur in the commensal niche. Indeed, the gut is the primary habitat of *E. coli* (Tenaillon et al. 2010; Denamur et al. 2020) and the reservoir of ExPEC strains (Denamur et al. 2020). Extra-intestinal infections, especially bacteremia with a high level of mortality, can be considered as dead-ends, the “virulence determinants” being in fact selected for allowing a more successful gut colonisation (Le Gall et al. 2007; Diard et al. 2010). Whatever the type and place of selection, it will be of interest to perform the same analysis in 10 years to see the fate of the actual clones.

Declarations

Ethics approval and consent to participate

Both multicenter clinical trials were approved by ethic committees: Colibafi: Ethics Committee CPP Hôpital Saint Louis, Paris, France (number 2006-4); Septicoli: ClinicalTrials.gov: identifier NCT02890901.

Consent for publication

All the consent have been obtained from the patient according the ethic committee requirement.

Availability of data and materials

The whole-genome sequences of the 912 strains studied have been deposited under the Bioprojects PRJEB35745 and PRJEB39260.

Competing interests

Nothing to declare.

Funding

This work was partially funded by a Translational Research Grant from the Agence Nationale de la Recherche (ANR), Ministry of Higher Education and Research, France 2015 (grant n°ANR-15-CE-17-0019-01). ED was partially supported by the “Fondation pour la Recherche Médicale” (Equipe FRM 2016, grant number DEQ20161136698). GR was supported by a “Poste d’accueil” funded by the “Assistance Publique-Hôpitaux de Paris” (AP-HP) and the

“Commissariat à l'énergie atomique et aux énergies alternatives ”(CEA) personal grant for his PhD.

Authors' contributions

G.R and E.D contributed to the conception and design of the work, analysis and interpretation of data and the drafting of the work. M.M-D contributed to the acquisition of data. OC contributed to the acquisition, analysis and interpretation of data. C.L, J-W.C, A.L contributed to the acquisition of data. V.dL contributed to the acquisition of data and drafting of the work. D.V contributed to the interpretation and the drafting of the work.

Acknowledgements

We are particularly grateful to François Blanquart, Julie Marin and Marie-Hélène Nicolas-Chanoine for their useful comments on the manuscript.

Rerences

Abernethy JK, Johnson AP, Guy R, Hinton N, Sheridan EA, Hope RJ. 2015. Thirty day all-cause mortality in patients with *Escherichia coli* bacteraemia in England. *Clin Microbiol Infect* 21:251.e1-8.

Alteri CJ, Mobley HLT. 2012. *Escherichia coli* physiology and metabolism dictates adaptation to diverse host microenvironments. *Curr Opin Microbiol* 15:3–9.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477.

Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. 2018. ClermonTyping: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. *Microb Genom* 4.

Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Khanh Nhu NT, Roberts LW, Stanton-Cook M, Schembri MA, Beatson SA. 2016. Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. *mBio* 7:e00347-00316.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57:289–300.

Bourrel AS, Poirel L, Royer G, Darty M, Vuillemin X, Kieffer N, Clermont O, Denamur E, Nordmann P, Decousser J-W. 2019. Colistin resistance in Parisian inpatient faecal

Escherichia coli as the result of two distinct evolutionary pathways. *J Antimicrob Chemother* 74:1521–1530.

Clermont O, Couffignal C, Blanco J, Mentré F, Picard B, Denamur E. 2017. Two levels of specialization in bacteraemic *Escherichia coli* strains revealed by their comparison with commensal strains. *Epidemiol Infect* 145:872–882.

Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E. 2011. Characterization of the cryptic *Escherichia* lineages: rapid identification and prevalence. *Environ Microbiol* 13:2468–2477.

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 43:e15.

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

Day MJ, Doumith M, Abernethy J, Hope R, Reynolds R, Wain J, Livermore DM, Woodford N. 2016. Population structure of *Escherichia coli* causing bacteraemia in the UK and Ireland between 2001 and 2010. *J Antimicrob Chemother* 71:2139–2142.

Deatherage DE, Barrick JE. 2014. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol* 1151:165–188.

DebRoy C, Roberts E, Fratamico PM. 2011. Detection of O antigens in *Escherichia coli*. *Anim Health Res Rev* 12:169–185.

Denamur E, Clermont O, Bonacorsi S, Gordon D. 2020. The population genetics of pathogenic *Escherichia coli*. *Nat Rev Microbiol*.

Diard M, Garry L, Selva M, Mosser T, Denamur E, Matic I. 2010. Pathogenicity-associated islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal colonization. *J Bacteriol* 192:4885–4893.

Duprilot M, Baron A, Blanquart F, Dion S, Pouget C, Lettéron P, Flament-Simon S-C, Clermont O, Denamur E, Nicolas-Chanoine M-H. 2020. Success of *Escherichia coli* O25b:H4 ST131 clade C associated with a decrease in virulence. *Infect Immun*.

Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu C-H, McDermott PF, et al. 2019. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob Agents Chemother* 63.

Gordon DM, Geyik S, Clermont O, O'Brien CL, Huang S, Abayasekara C, Rajesh A, Kennedy K, Collignon P, Pavli P, et al. 2017. Fine-scale structure analysis shows epidemic patterns of clonal complex 95, a cosmopolitan *Escherichia coli* lineage responsible for extraintestinal infection. *mSphere* 2.

Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, Levine MM, Robins-Browne RM, Holt KE. 2016. In silico serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb Genom* 2:e000064.

Jauréguy F, Carbonnelle E, Bonacorsi S, Clec'h C, Casassus P, Bingen E, Picard B, Nassif X, Lortholary O. 2007. Host and bacterial determinants of initial severity and outcome of *Escherichia coli* sepsis. *Clin Microbiol Infect* 13:854–862.

Johnson JR, Clermont O, Menard M, Kuskowski MA, Picard B, Denamur E. 2006. Experimental mouse lethality of *Escherichia coli* isolates, in relation to accessory traits, phylogenetic group, and ecological source. *J Infect Dis* 194:1141–1150.

Johnson JR, Johnston BD, Porter S, Thuras P, Aziz M, Price LB. 2019. Accessory traits and phylogenetic background predict *Escherichia coli* extraintestinal virulence better than does ecological source. *J Infect Dis* 219:121–132.

Johnson JR, Porter SB, Zhanel G, Kuskowski MA, Denamur E. 2012. Virulence of *Escherichia coli* clinical isolates in a murine sepsis model in relation to sequence type ST131 status, fluoroquinolone resistance, and virulence genotype. *Infect Immun* 80:1554–1562.

Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, Peacock SJ, Parkhill J. 2017. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res* 27:1437–1449.

Kang C-I, Song J-H, Chung DR, Peck KR, Ko KS, Yeom J-S, Ki HK, Son JS, Lee SS, Kim Y-S, et al. 2010. Risk factors and treatment outcomes of community-onset bacteraemia caused by extended-spectrum beta-lactamase-producing *Escherichia coli*. *Int J Antimicrob Agents* 36:284–287.

Kauffmann F. 1947. The serology of the coli group. *J Immunol* 57:71–100.

Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. 2011. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* 12:32.

Landraud L, Jauréguy F, Frapy E, Guigon G, Gouriou S, Carbonnelle E, Clermont O, Denamur E, Picard B, Lemichez E, et al. 2013. Severity of *Escherichia coli* bacteraemia is independent of the intrinsic virulence of the strains assessed in a mouse model. *Clin Microbiol Infect* 19:85–90.

de Lastours V, Laouénan C, Royer G, Carbonnelle E, Lepeule R, Esposito-Farèse M, Clermont O, Duval X, Fantin B, Mentré F, et al. 2020. Mortality in *Escherichia coli* bloodstream infections: antibiotic resistance still does not make it. *J Antimicrob Chemother* 75:2334–2343.

Le Gall T, Clermont O, Gouriou S, Picard B, Nassif X, Denamur E, Tenaillon O. 2007. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. *Mol Biol Evol* 24:2373–2384.

Lefort A, Panhard X, Clermont O, Woerther P-L, Branger C, Mentré F, Fantin B, Wolff M, Denamur E. 2011. Host factors and portal of entry outweigh bacterial determinants to predict the severity of *Escherichia coli* bacteremia. *J Clin Microbiol* 49:777–783.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 47:W256–W259.

Mamani R, Flament-Simon SC, García V, Mora A, Alonso MP, López C, García-Meniño I, Díaz-Jiménez D, Blanco JE, Blanco M, et al. 2019. Sequence types, clonotypes, serotypes, and virotypes of extended-spectrum β-lactamase-producing *Escherichia coli* causing bacteraemia in a spanish hospital over a 12-year period (2000 to 2011). *Front Microbiol* 10:1530.

Martínez JA, Soto S, Fabrega A, Almela M, Mensa J, Soriano A, Marco F, Jimenez de Anta MT, Vila J. 2006. Relationship of phylogenetic background, biofilm production, and time to detection of growth in blood culture vials with clinical variables and prognosis associated with *Escherichia coli* bacteremia. *J Clin Microbiol* 44:1468–1474.

Massot M, Daubié A-S, Clermont O, Jauréguy F, Couffignal C, Dahbi G, Mora A, Blanco J, Branger C, Mentré F, et al. 2016. Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of *Escherichia coli* from community subjects in the Paris area in 2010 and evolution over 30 years. *Microbiology (Reading)* 162:642–650.

McNally A, Kallonen T, Connor C, Abudahab K, Aanensen DM, Horner C, Peacock SJ, Parkhill J, Croucher NJ, Corander J. 2019. Diversification of Colonization Factors in a Multidrug-Resistant *Escherichia coli* Lineage Evolving under Negative Frequency-Dependent Selection. *mBio* 10.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 37:1530–1534.

Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, Feist AM, Palsson BØ. 2013. Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci U S A* 110:20338–20343.

Mora-Rillo M, Fernández-Romero N, Navarro-San Francisco C, Díez-Sebastián J, Romero-Gómez MP, Fernández FA, López JRA, Mingorance J. 2015. Impact of virulence genes on sepsis severity and survival in *Escherichia coli* bacteremia. *Virulence* 6:93–100.

Naas T, Oueslati S, Bonnin RA, Dabos ML, Zavala A, Dortet L, Retailleau P, Iorga BI. 2017. Beta-lactamase database (BLDB) - structure and function. *J Enzyme Inhib Med Chem* 32:917–919.

Nicolas-Chanoine M-H, Bertrand X, Madec J-Y. 2014. *Escherichia coli* ST131, an intriguing clonal group. *Clin Microbiol Rev* 27:543–574.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691–3693.

Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528.

Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, Elion J, Denamur E. 1999. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect Immun* 67:546–553.

Picard B, Goullet P. 1989. Correlation between electrophoretic types B1 and B2 of carboxylesterase B and sex of patients in *Escherichia coli* urinary tract infections. *Epidemiol Infect* 103:97–103.

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* 406:64–67.

Royer G, Decousser JW, Branger C, Dubois M, Médigue C, Denamur E, Vallenet D. 2018. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genom* 4.

Ruppé E, Cherkaoui A, Charretier Y, Girard M, Schicklin S, Lazarevic V, Schrenzel J. 2020. From genotype to antibiotic susceptibility phenotype in the order Enterobacterales: a clinical perspective. *Clin Microbiol Infect* 26:643.e1-643.e7.

Russo TA, Johnson JR. 2003. Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem. *Microbes Infect* 5:449–456.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069.

Seemann T. 2020. tseemann/snippy. Available from: https://github.com/tseemann/snippy

Skurnik D, Clermont O, Guillard T, Launay A, Danilchanka O, Pons S, Diancourt L, Lebreton F, Kadlec K, Roux D, et al. 2016. Emergence of antimicrobial-resistant *Escherichia coli* of animal origin spreading in humans. *Mol Biol Evol* 33:898–914.

Stephens CM, Adams-Sapper S, Sekhon M, Johnson JR, Riley LW. 2017. Genomic analysis of factors associated with low prevalence of antibiotic resistance in extraintestinal pathogenic *Escherichia coli* Sequence Type 95 strains. *mSphere* 2.

Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics* 27:1009–1010.

Tchesnokova V, Aprikian P, Kisiela D, Gowey S, Korotkova N, Thomas W, Sokurenko E. 2011. Type 1 fimbrial adhesin FimH elicits an immune response that enhances cell adhesion of *Escherichia coli. Infect Immun* 79:3895–3904.

Tenaillon O, Rodríguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. 2012. The molecular diversity of adaptive convergence. *Science* 335:457–461.

Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* 8:207–217.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 5:e1000344.

Touchon M, Perrin A, de Sousa JAM, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon D, Rocha EP. 2020. Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLoS Genet* 16:e1008866.

Vihta K-D, Stoesser N, Llewelyn MJ, Quan TP, Davies T, Fawcett NJ, Dunn L, Jeffery K, Butler CC, Hayward G, et al. 2018. Trends over time in *Escherichia coli* bloodstream

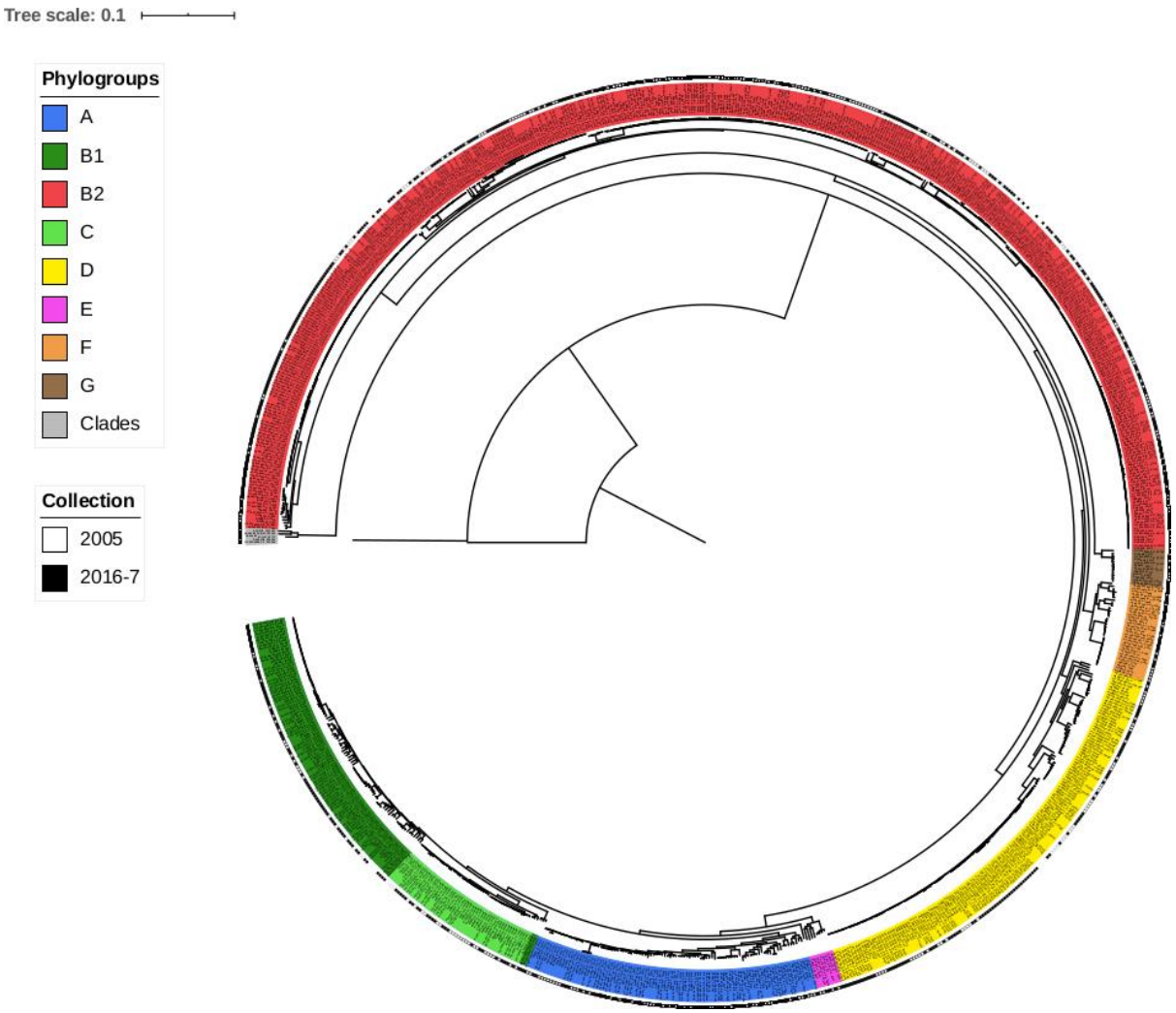
infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998-2016: a study of electronic health records. *Lancet Infect Dis* 18:1138–1149.

Wang L, Rothemund D, Curd H, Reeves PR. 2003. Species-wide variation in the *Escherichia coli* flagellin (H-antigen) gene. *J Bacteriol* 185:2936–2943.

Yoon E-J, Choi MH, Park YS, Lee HS, Kim D, Lee H, Shin KS, Shin Jong Hee, Uh Y, Kim YA, et al. 2018. Impact of host-pathogen-treatment tripartite components on early mortality of patients with *Escherichia coli* bloodstream infection: Prospective observational study. *EBioMedicine* 35:76–86.

Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 67:2640–2644.

Additional files

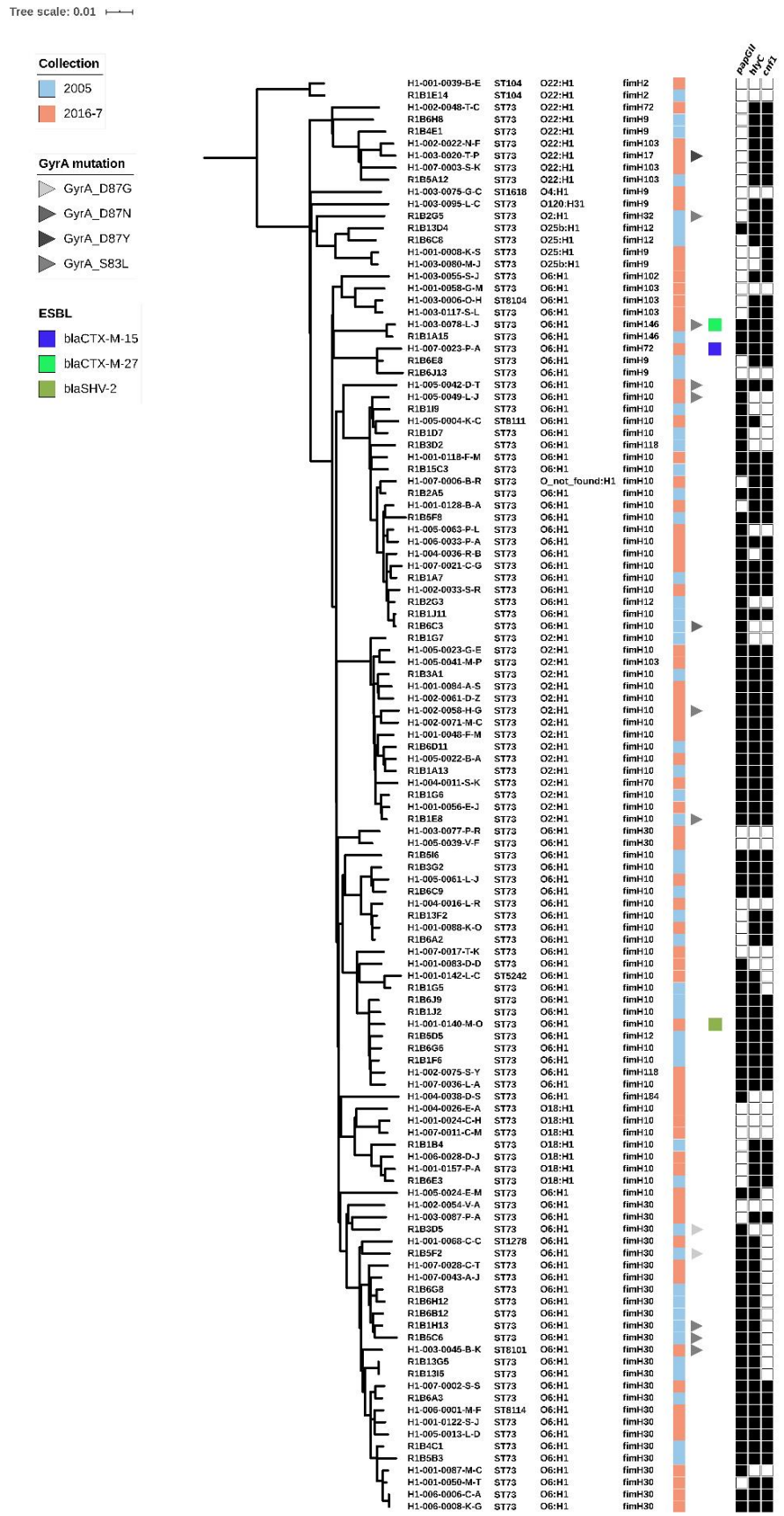


Additional file 1: Figure S1. Core-genome SNP based phylogenetic tree of the 912 strains from collections 2005 and 2016-7. The tree is rooted on an Escherichia clade V strain. The eight main phylogroups are highlighted in color. On the outermost circle appear white and black squares according to the origin of the strains (i.e. 2005 or 2016-7).

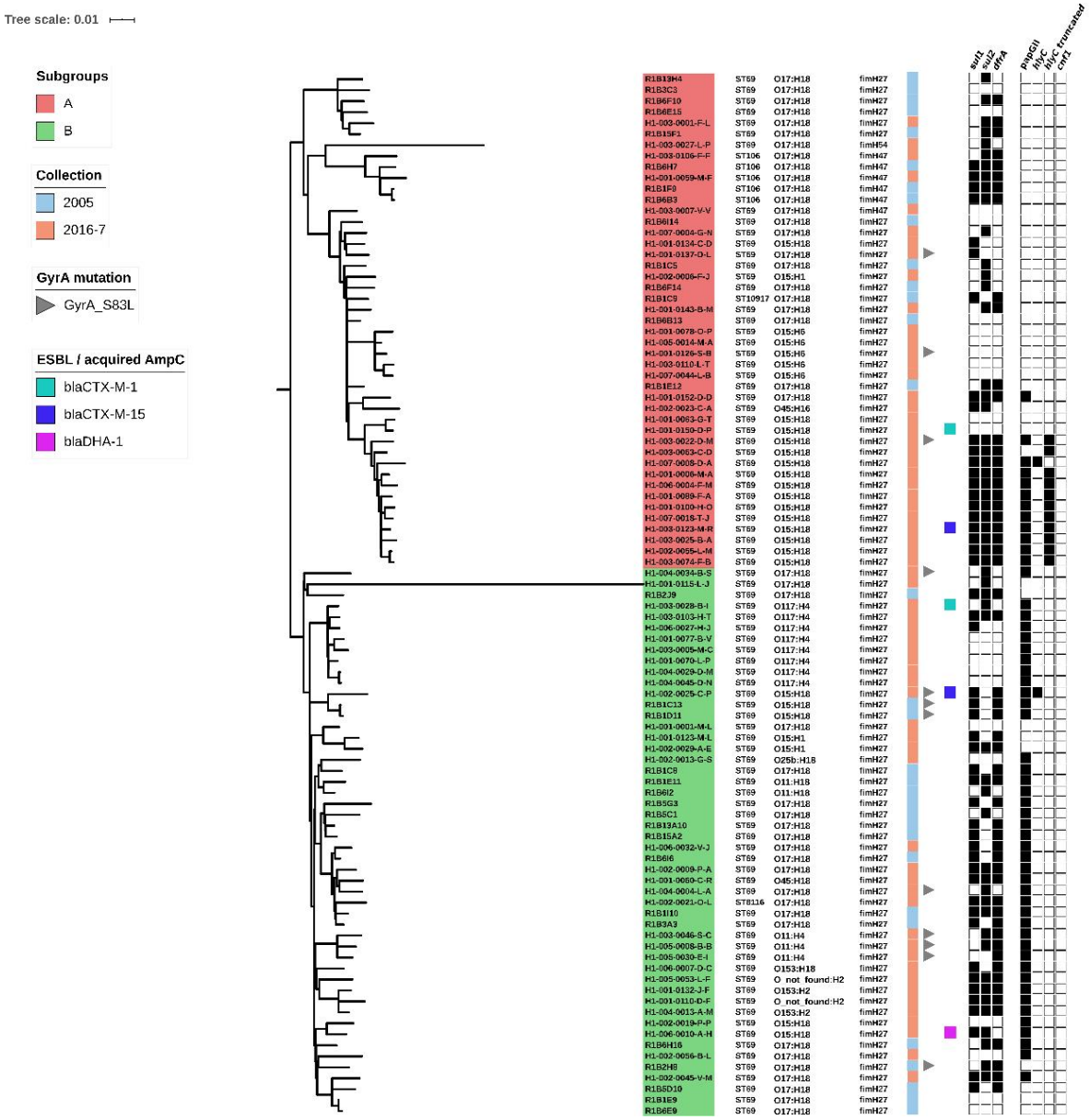
Additional file 2: Table S1. Antibiotic resistance prediction according to genes/mutations.

Resistance determinant	Gene or mutation	AMP*	TZP*	CTX/CAZ*	FEP*	CARB*	GEN*	AMK*	FQ*	SXT*
blaCMY-2	gene	R	R	R	S	S	NA	NA	NA	NA
blaCTX-M	gene	R	S	R	R	S	NA	NA	NA	NA
blaDHA-1	gene	R	R	R	S	S	NA	NA	NA	NA
blaOXA-1	gene	R	R	S	S	S	NA	NA	NA	NA
blaOXA-2	gene	R	R	S	S	S	NA	NA	NA	NA
blaOXA-48	gene	R	R	S	S	R	NA	NA	NA	NA
blaOXA-244	gene	R	R	S	S	R	NA	NA	NA	NA
blaSHV-2	gene	R	S	R	R	S	NA	NA	NA	NA
blaSHV-12	gene	R	S	R	R	S	NA	NA	NA	NA
blaTEM-1	gene	R	S	S	S	S	NA	NA	NA	NA
blaTEM-1_variant	gene	R	S	S	S	S	NA	NA	NA	NA
blaTEM-30	gene	R	R	S	S	S	NA	NA	NA	NA
blaTEM-33	gene	R	R	S	S	S	NA	NA	NA	NA
blaTEM-35	gene	R	R	S	S	S	NA	NA	NA	NA
blaTEM-36_variant	gene	R	R	S	S	S	NA	NA	NA	NA
blaTEM-40	gene	R	R	S	S	S	NA	NA	NA	NA
blaTEM-135	gene	R	S	S	S	S	NA	NA	NA	NA
blaTEM-163	gene	R	R	S	S	S	NA	NA	NA	NA
blaTEM-166	gene	R	S	S	S	S	NA	NA	NA	NA
qnrB4	gene	NA	NA	NA	NA	NA	NA	NA	R	NA
qnrB19	gene	NA	NA	NA	NA	NA	NA	NA	R	NA
qnrS1	gene	NA	NA	NA	NA	NA	NA	NA	R	NA
GyrA83	mutation	NA	NA	NA	NA	NA	NA	NA	R	NA
GyrA87	mutation	NA	NA	NA	NA	NA	NA	NA	R	NA
aac(3)-IIa	gene	NA	NA	NA	NA	NA	R	S	NA	NA
aac(3)-IId	gene	NA	NA	NA	NA	NA	R	S	NA	NA
aac(3)-IVa	gene	NA	NA	NA	NA	NA	R	S	NA	NA
aac(3)-VIa	gene	NA	NA	NA	NA	NA	R	S	NA	NA
aac(6')-Ib-cr	gene	NA	NA	NA	NA	NA	S	R	R	NA
aadA2	gene	NA	NA	NA	NA	NA	S	S	NA	NA
aadA5	gene	NA	NA	NA	NA	NA	S	S	NA	NA
ant(2")-Ia	gene	NA	NA	NA	NA	NA	R	S	NA	NA
ant(3")-Ia_1	gene	NA	NA	NA	NA	NA	S	S	NA	NA
aph(3')-Ia	gene	NA	NA	NA	NA	NA	S	S	NA	NA
aph(3")-Ib	gene	NA	NA	NA	NA	NA	S	S	NA	NA
aph(4)-Ia	gene	NA	NA	NA	NA	NA	S	S	NA	NA
aph(6)-Id	gene	NA	NA	NA	NA	NA	S	S	NA	NA

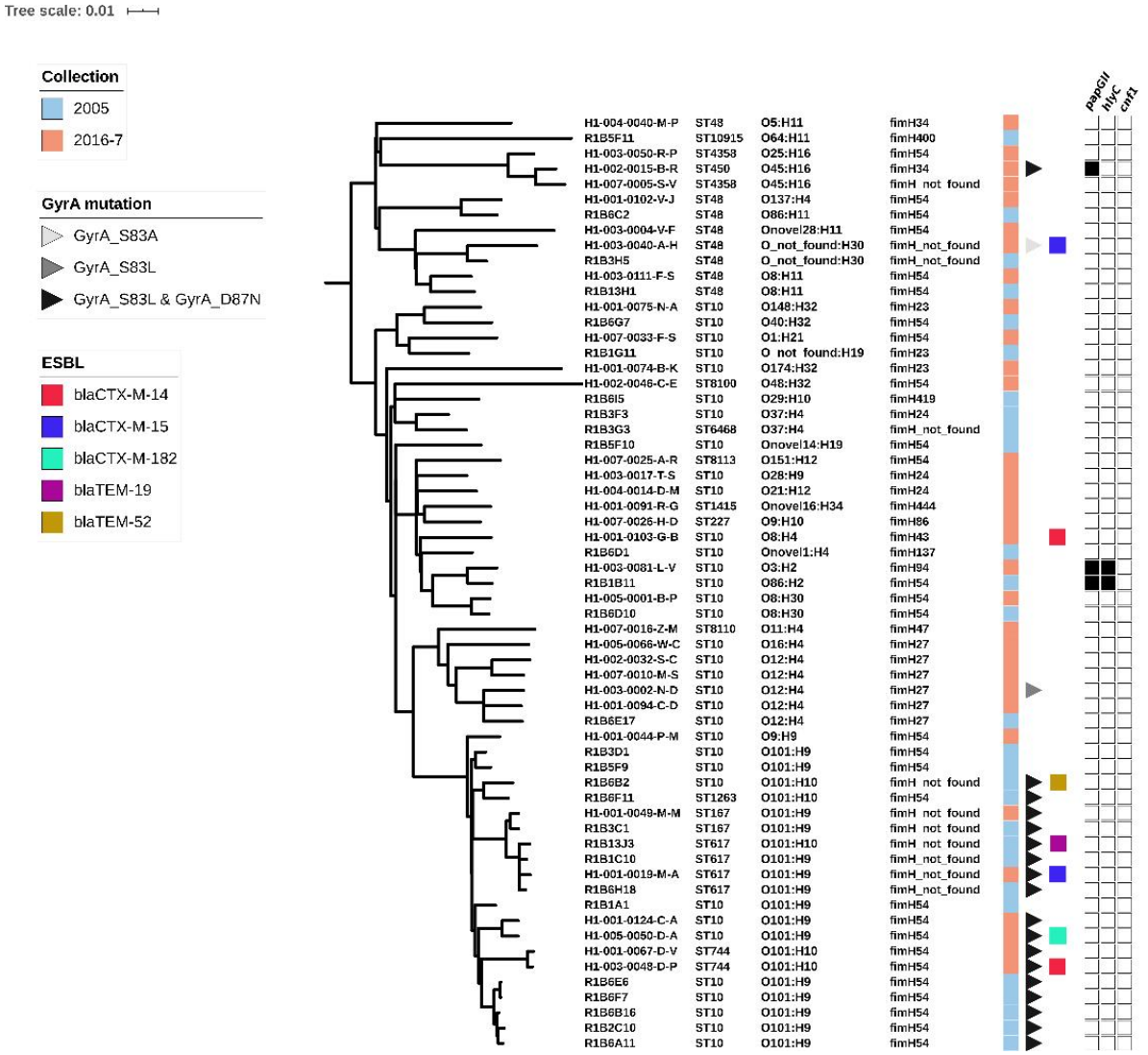
*AMP = ampicillin, TZP = piperacillin/tazobactam, CTX/CAZ = cefotaxime/ceftazidime, FEP = cefepime, CARB = carbapenems, GEN = gentamicin, AMK = amikacin, FQ = fluoroquinolones, SXT = cotrimoxazole



Additional file 3: Figure S2. SNP-based phylogenetic tree of STc73 strains. Mutations in GyrA are reported with triangle. ESBL coding genes are shown with colored squares and presence of *papGII*, *hlyC and cnf1* with black squares. The tree is mid-point rooted.



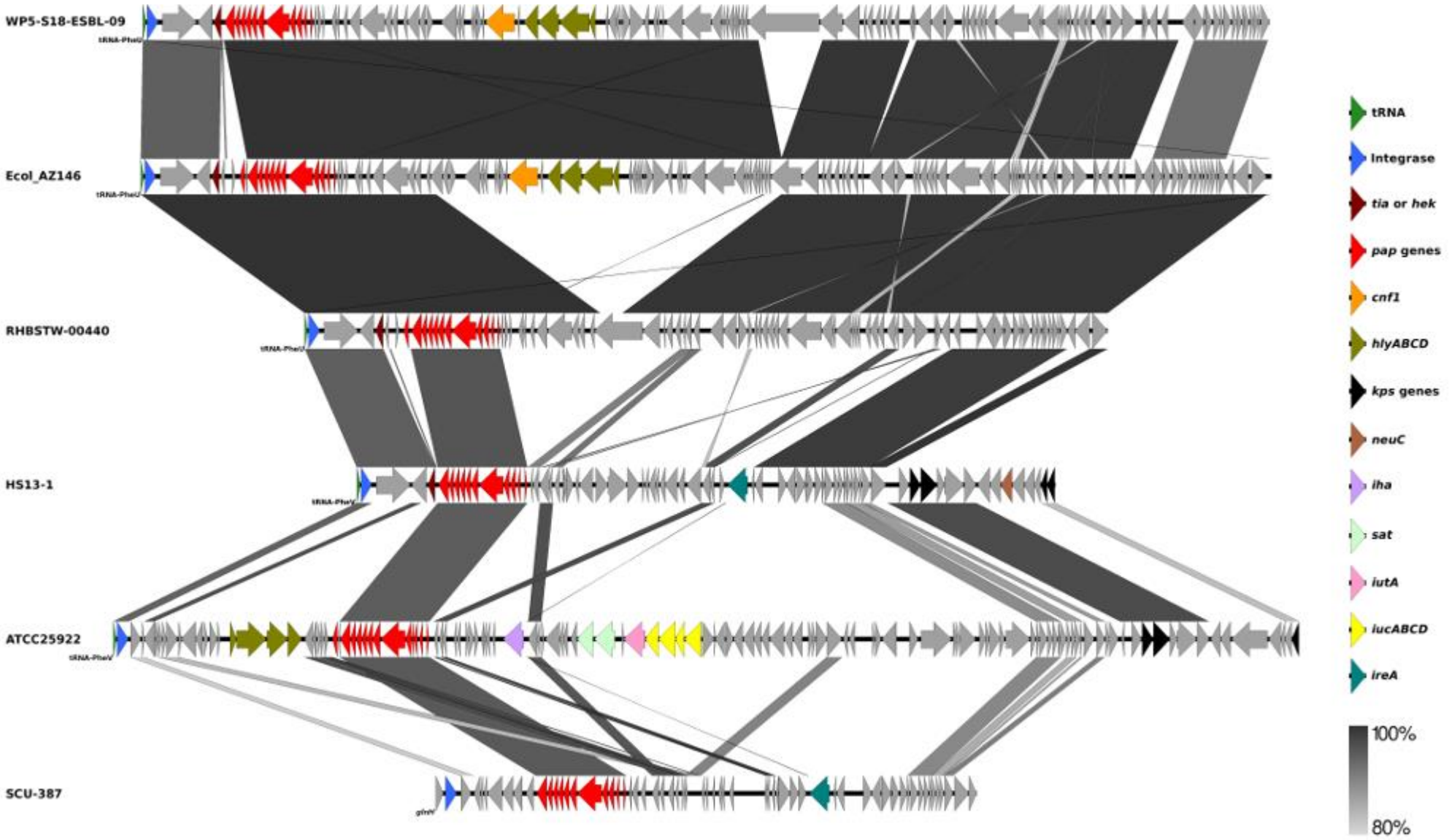
Additional file 4: Figure S3. SNP based phylogenetic tree of STc69 strains. Clades A and B are highlighted in color. Mutations in GyrA are reported with triangle and *bla*CTX-M coding genes with colored squares. The presence of *sul1*, *sul2, dfrA, papGII*, *hlyC, hlyC* truncated and *cnf1* is shown with black squares. One strain (H1-001-0115-L-J) exhibits a very long branch which could be related to a mutator phenotype as a non-synonymous mutation was found in MutS (A60T). The tree is mid-point rooted.



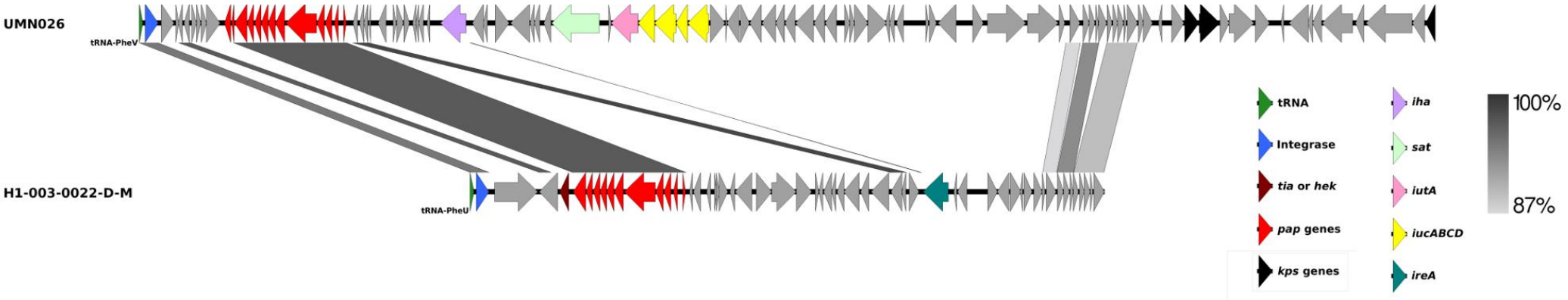
Additional file 5: Figure S4. SNP-based phylogenetic tree of STc10 strains. Mutations in GyrA are reported with triangle. ESBL coding genes are shown with colored squares and presence of *papGII*, *hlyC and cnf1* with black squares. The tree is mid-point rooted.

Additional file 6: Table S2. Pathogenicity island analysis.

STc	Isolate	Subgroup/Clade	Reference strain	Accession number and coordinates	VF on the reference PAI	insertion site	Comments
131	H1-002-0039-R-E	A					No match found
131	H1-001-0011-M-Y	A	WP5-S18-ESBL-09	AP022171.1:4401443-4517387	*papGII, tia, hlyABCD, cnf1*	tRNA-PheU	
131	H1-003-0032-M-L	A	WP5-S18-ESBL-09	AP022171.1:4401443-4517387	*papGII, tia, hlyABCD, cnf1*	tRNA-PheU	
131	H1-003-0066-D-R	A	WP5-S18-ESBL-09	AP022171.1:4401443-4517387	*papGII, tia, hlyABCD, cnf1*	tRNA-PheU	
131	R1B5I4	B0	ATCC25922	CP037449.1:1072107-1194187	*papGII, iha, sat, iutA, iucABCD, hlyABCD, kpsMDE*	tRNA-PheV	
131	H1-005-0017-S-M	C1	RHBSTW-00440	CP056524.1:329738-413067	*papGII, hek*	tRNA-PheU	
131	H1-003-0069-M-R	C1				tRNA-PheU	No match found
131	H1-003-0122-K-R	C1	HS13-1	CP026491.1:1912022-1983914	*papGII, tia, ireA, kpsMDET, neuC*	tRNA-PheV	Closest reference is an ST95 strain. Deletion at the end of the PAI. H1-003-0122-K-R lacks *kpsT* and *neuc*
131	H1-006-0005-A-H	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-001-0141-G-Y	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	Match only with the beginning of the PAI.
131	H1-002-0026-M-N	C2	SCU-387	CP051688.1:1563956-1619596	*papGII, ireA*	Insertion between *glnH* and *glnP*	
131	H1-003-0052-K-I	C2	RHBSTW-00440	CP056524.1:329738-413067	*papGII, hek*	tRNA-PheU	Deletion into PAI
131	H1-005-0028-K-P	C2	RHBSTW-00440	CP056524.1:329738-413067	*papGII, hek*	tRNA-PheU	Deletion into PAI
131	H1-004-0022-M-L	C2	RHBSTW-00440	CP056524.1:329738-413067	*papGII, hek*	tRNA-PheU	Deletion into PAI
131	H1-003-0058-P-M	C2	SCU-387	CP051688.1:1563956-1619596	*papGII, ireA*	Insertion between *glnH* and *glnP*	
131	H1-001-0079-V-O	C2	RHBSTW-00440	CP056524.1:329738-413067	*papGII, hek*	tRNA-PheU	Deletion into PAI
131	H1-003-0056-S-G	C2	RHBSTW-00440	CP056524.1:329738-413067	*papGII, hek*	tRNA-PheU	Deletion into PAI
131	H1-006-0029-A-Y	C2	RHBSTW-00440	CP056524.1:329738-413067	*papGII, hek*	tRNA-PheU	
131	H1-002-0049-D-R	C2	RHBSTW-00440	CP056524.1:329738-413067	*papGII, hek*	tRNA-PheU	
131	H1-003-0033-A-J	C2	RHBSTW-00440	CP056524.1:329738-413067	*papGII, hek*	tRNA-PheU	
131	H1-002-0001-M-C	C2	RHBSTW-00440	CP056524.1:329738-413067	*papGII, hek*	tRNA-PheU	
131	H1-001-0036-E-A	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	Match only with the beginning of the PAI. H1-001-0036-E-A lacks hlyABCD and cnf1
131	H1-007-0027-K-T	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	Match only with the beginning of the PAI. H1-007-0027-K-T lacks hlyABCD and cnf1
131	H1-003-0107-C-T	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-003-0119-G-R	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-005-0011-B-M	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-005-0057-S-L	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-002-0070-M-J	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-004-0001-B-J	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-007-0001-L-D	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	Deletion into PAI
131	H1-002-0040-D-F	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-001-0151-L-P	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-004-0037-C-M	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-002-0004-T-A	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-003-0039-D-O	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-003-0014-B-N	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
131	H1-003-0083-B-J	C2	Ecol_AZ146	CP018991.1:3438092-3554476	*papGII, hek, hlyABCD, cnf1*	tRNA-PheU	
69	H1-004-0034-B-S	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. H1-004-0034-B-S lacks *iutA*
69	H1-003-0028-B-I	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-003-0103-H-T	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-006-0027-H-J	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-001-0077-B-V	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-003-0005-M-C	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-001-0070-L-P	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion into PAI. H1-001-0070-L-P lacks *iha*
69	H1-004-0029-D-M	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-004-0045-D-N	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-002-0025-C-P	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI
69	R1B1C13	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	R1B1D11	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-002-0013-G-S	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. H1-002-0013-G-S lacks *iutA*
69	R1B1C8	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	R1B1E11	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	R1B6I2	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	R1B5G3	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	R1B5C1	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	R1B13A10	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	R1B15A2	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-006-0032-V-J	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. H1-006-0032-V-J lacks *kpsM*
69	R1B6I6	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. R1B6I6 lacks *kpsM*
69	H1-002-0009-P-A	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. H1-002-0009-P-A lacks *kpsM*
69	H1-001-0060-C-R	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. H1-001-0060-C-R lacks *kpsM*
69	H1-004-0004-L-A	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. H1-004-0004-C-R lacks *kpsM*
69	H1-002-0021-O-L	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. H1-002-0021-O-L lacks *kpsM*
69	R1B1I10	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. R1B1I10 lacks *kpsM*
69	R1B3A3	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. R1B3A3 lacks *kpsM*
69	H1-003-0046-S-C	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-005-0008-B-B	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-005-0030-E-I	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-006-0007-D-C	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-005-0053-L-F	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-001-0132-J-F	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-001-0110-D-F	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-004-0013-A-M	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-002-0019-P-P	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-006-0010-A-H	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	R1B6H16	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	
69	H1-002-0056-B-L	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. H1-002-0056-B-L lacks *kpsM*
69	H1-002-0045-V-M	B	UMN026	NC_011751.1:3445924-3551321	*papGII, iha, sat, iutA, iucABCD, kpsMDE*	tRNA-PheV	Deletion at the end of the PAI. H1-002-0045-V-M lacks *iutA*
69	H1-001-0152-D-D	A					No match found
69	H1-003-0022-D-M	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	
69	H1-007-0008-D-A	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	Deletion at the end of the PAI
69	H1-001-0006-M-A	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	Deletion at the end of the PAI
69	H1-006-0004-F-M	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	Deletion at the end of the PAI
69	H1-001-0089-F-A	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	Deletion at the end of the PAI
69	H1-001-0100-H-O	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	Deletion at the end of the PAI
69	H1-007-0018-T-J	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	Deletion at the end of the PAI
69	H1-003-0123-M-R	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	Deletion at the end of the PAI
69	H1-003-0025-B-A	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	Deletion at the end of the PAI
69	H1-002-0055-L-M	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	Deletion at the end of the PAI
69	H1-003-0074-F-B	A	H1-003-0022-D-M (GCA_902709125.2)	CACSUY020000004.1:146692-198942	*papGII, tia, ireA*	tRNA-PheU	Deletion at the end of the PAI



Additional file 7: Figure S5. Genetic map of the reference PAIs found in the STc131 strains. The virulence genes of interest are highlighted in color as well as tRNA-PheU and tRNA-PheV. The result of the BlastN comparison between PAI is represented by grey to black blocks depending on the shared similarity of the sequences.

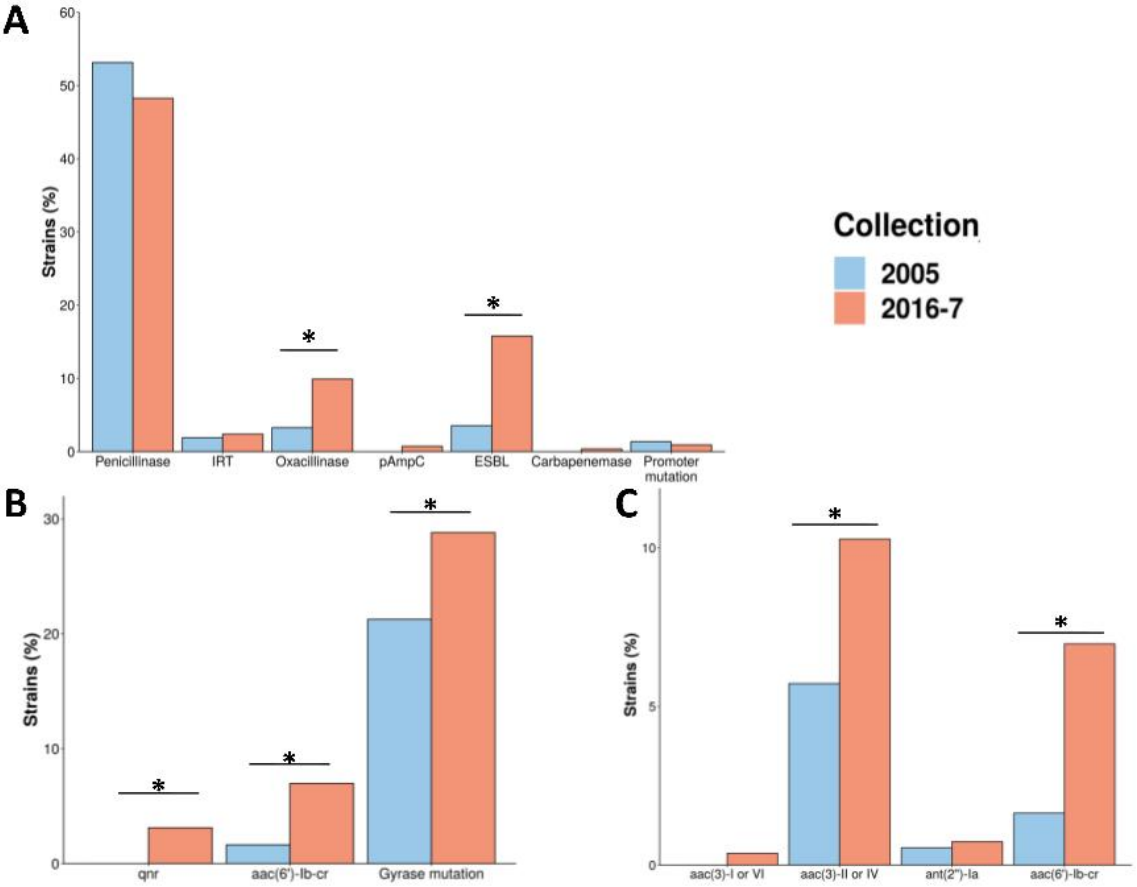


Additional file 8: Figure S6. Genetic map of the reference PAIs found in the STc69 strains. The virulence genes of interest are highlighted in color as well as tRNA-PheU and tRNA-PheV. The result of the BlastN comparison between PAI is represented by grey to black blocks depending on the shared similarity of the sequences.

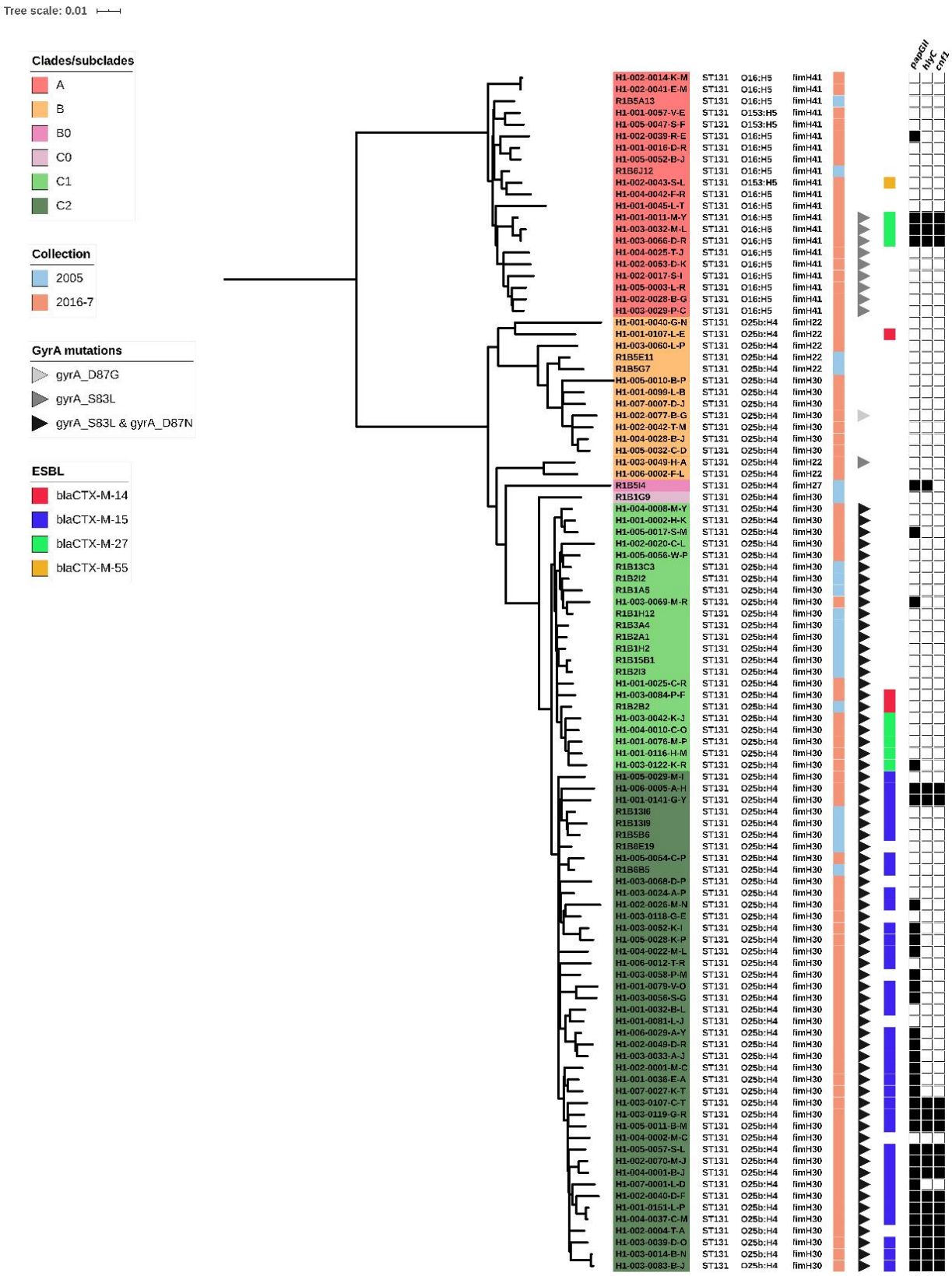
Additional file 9: Table S3. Phylogroup distribution among 2005 and 2016-7 collections, overall and according to urinary and digestive portal of entry

	All strains			Urinary source			Digestive source		
Phylogroups	2005 (%)	2016-7 (%)	*P-value**	2005 (%)	2016-7 (%)	*P-value**	2005 (%)	2016-7 (%)	*P-value**
A	43 (11.72)	53 (9.72)	1.00	11 (5.31)	15 (5.7)	1.00	22 (28.57)	30 (14.71)	0.42
B1	28 (7.63)	70 (12.84)	0.11	8 (3.86)	20 (7.6)	0.56	10 (12.99)	39 (19.12)	0.79
B2	188 (51.23)	279 (51.19)	1.00	130 (62.8)	160 (60.84)	1.00	26 (33.77)	82 (40.2)	0.79
C	27 (7.36)	24 (4.4)	0.35	15 (7.25)	12 (4.56)	0.96	3 (3.9)	9 (4.41)	1.00
D	57 (15.53)	87 (15.96)	1.00	38 (18.36)	44 (16.73)	1.00	8 (10.39)	31 (15.2)	0.79
E	4 (1.09)	4 (0.73)	1.00	1 (0.48)	2 (0.76)	1.00	1 (1.3)	0 (0)	0.79
F	12 (3.27)	19 (3.49)	1.00	2 (0.97)	9 (3.42)	0.56	4 (5.19)	6 (2.94)	0.79
G	6 (1.63)	6 (1.1)	1.00	1 (0.48)	1 (0.38)	1.00	2 (2.6)	5 (2.45)	1.00
Clades	2 (0.54)	3 (0.55)	1.00	1 (0.48)	0 (0)	0.99	1 (1.3)	2 (0.98)	1.00

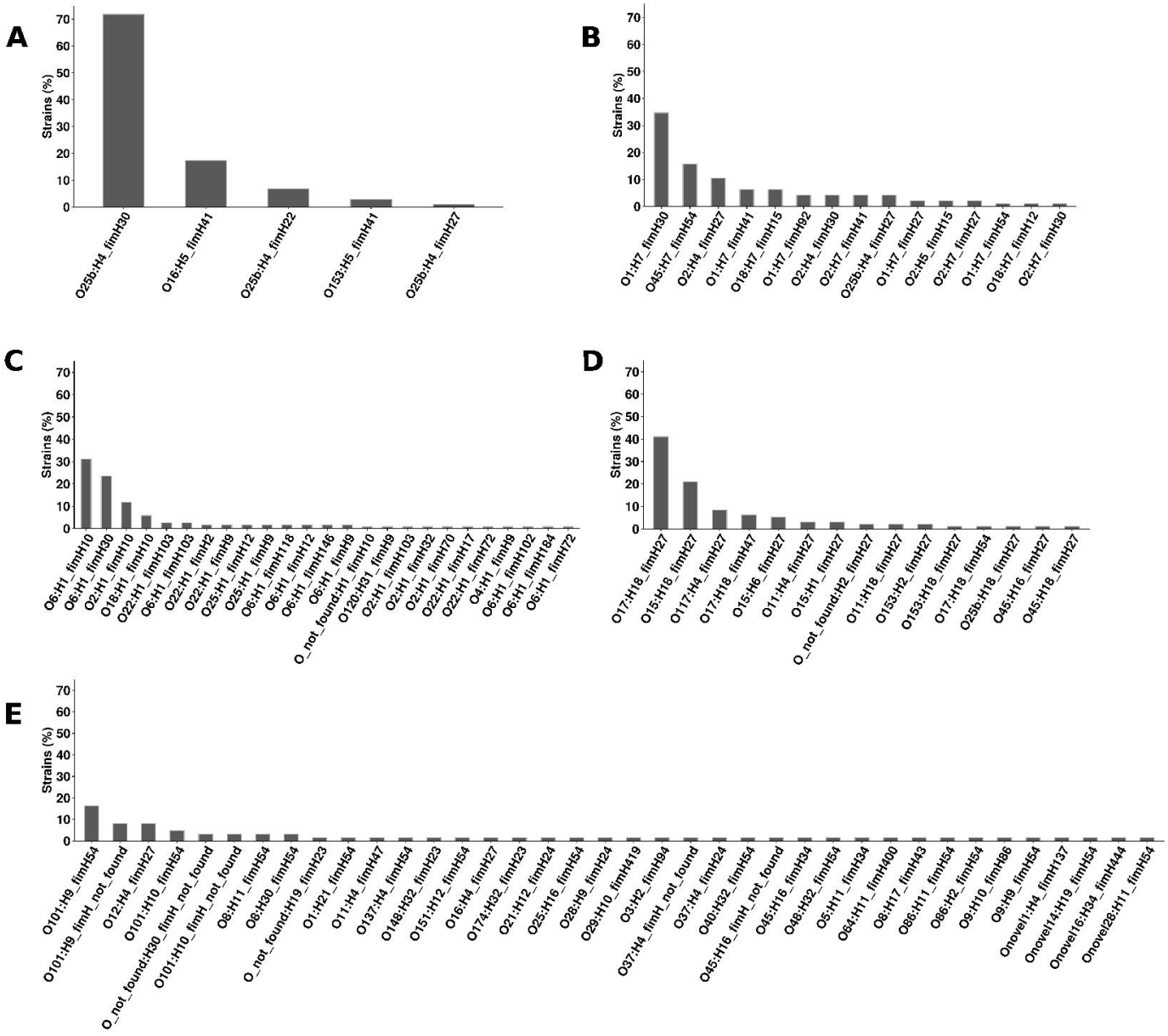
* Corrected p-value (Benjamini-Hochberg)



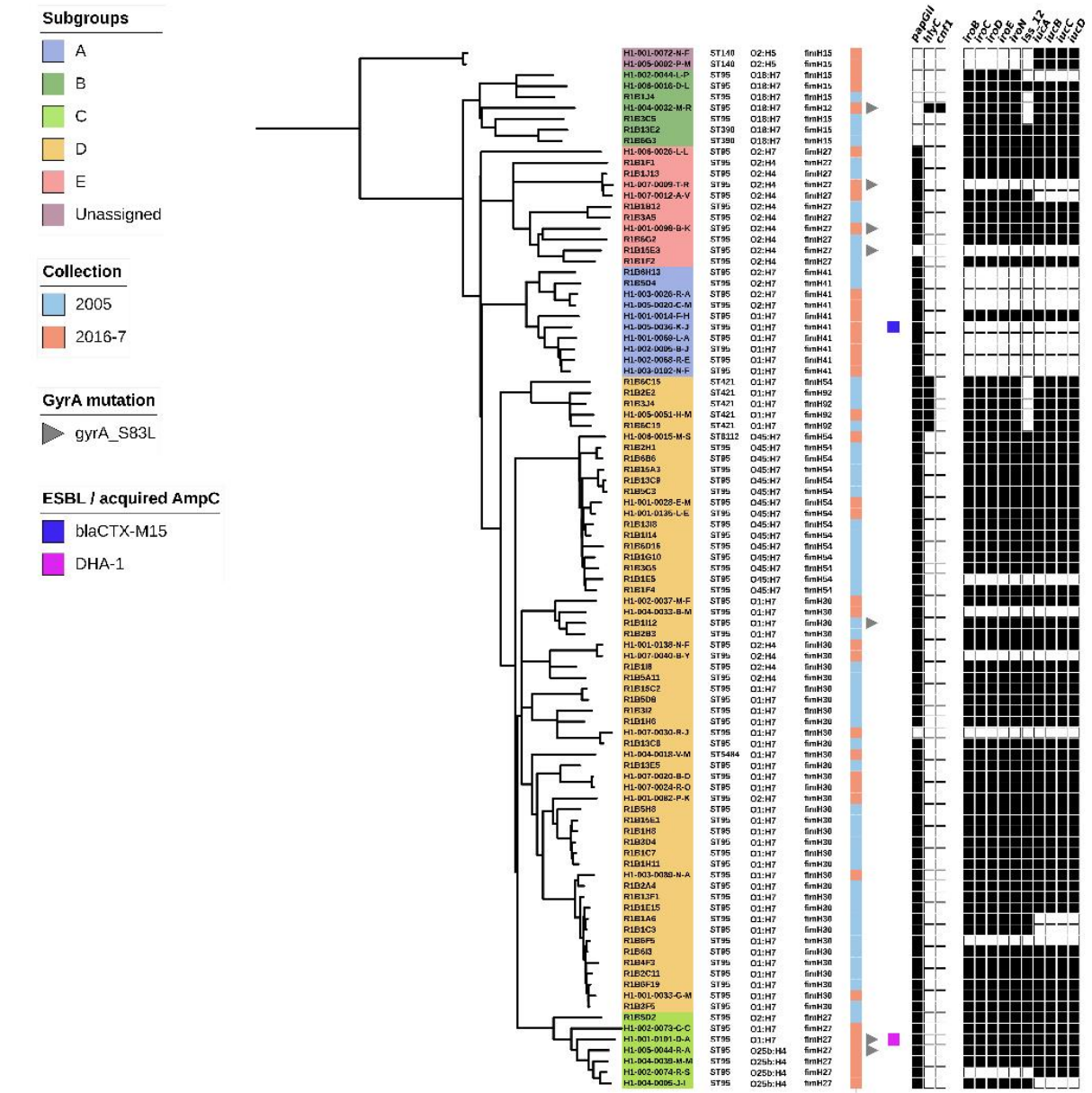
Additional file 10: Figure S7. Distribution of genes and mutations responsible for resistance to beta-lactams (A), fluroquinolones (B) and aminoglycosides (C) among strains from the 2005 and 2016-7 collections. Betalactam resistance coding genes are grouped by phenotype. The gyrase mutations include only GyrA mutations for fluoroquinolone resistance. Significant differences are highlighted by asterisks. IRT = Inhibitor Resistant TEM; pAmpC = plasmidic ampC; ESBL = extended spectrum betalactamase.



Additional file 11: Figure S8. SNP-based phylogenetic tree of STc131 strains. The previously described clades/sub-clades A, B, B0, C0, C1 and C2 are highlighted in color. Mutations in GyrA are reported with triangle. *bla*CTX-M coding genes are shown with colored squares and presence of *papGII*, *hlyC* and *cnf1* with black squares. The tree is mid-point rooted.



Additional file 12: Figure S9. O:H/*fimH* combinations for the five main STcs. Results are shown as percentage of strains belonging to (A) STc131, (B) STc95, (C), STc73, (D) STc69 and (E) STc10.



Additional file 13: Figure S10. SNP-based phylogenetic tree of STc95 strains. The previously described subgroups A, B, C, D and E are highlighted in color. Mutations in GyrA are reported with triangle. ESBL and acquired AmpC coding genes are shown with colored squares and presence of *papGII*, *hlyC*, *cnf1, iro*B/C/D/E/N, *iss*_12 and *iuc*A/B/C/D with black squares. The tree is mid-point rooted.

c. Commentaires et perspectives

A travers cette étude nous avons observé une stabilité sur une échelle globale, avec un phylogroupe B2 dominant et présent en proportion identique entre les deux études. De plus, nous avons pu confirmer la pauci-clonalité des bactériémies à *E. coli*, comme l'ont constaté d'autres auteurs (Kallonen et al., 2017; Yoon et al., 2018), bien que la prévalence de certains STc comme les STc131 et STc95 varie fortement dans le temps. Dans le cas de ces deux derniers STc, nous avons constaté l'expansion de certaines lignées au dépend d'autres. De telles expansions ont été décrites récemment par Ludden *et al.* (Ludden et al., 2020). Ces auteurs mettent en évidence l'émergence d'un groupe de souches au sein du STc131, présentant une insertion chromosomique particulière du gène *bla*CTX-M-15. Cependant l'expansion de cette lignée est essentiellement observée à l'échelle locale, et est différente de la lignée dominante à l'échelle mondiale. Cela souligne l'importance de prendre en compte les spécificités spatio-temporelles dans l'interprétation des résultats d'analyse épidémiologique de telles infections.

Parmi les expansions clonales observées dans notre étude, certaines sont associées à l'acquisition indépendante d'un même facteur de virulence, à savoir le gène *papGII,* par le biais de différents PAI. Cette convergence évolutive est particulièrement intéressante au regard des données publiées très récemment par Biggel *et al.* (Biggel et al., 2020). Les auteurs de cette étude ont en effet identifié ce même facteur de virulence impliqué dans l'adhésion comme étant hautement associé aux souches pathogènes urinaires invasives (pyélonéphrites et bactériémies à point de départ urinaire). Dater l'émergence des groupes de souches que nous avons identifiés pourrait permettre un interprétation plus fine de ces évènements et d'identifier les éventuelles pressions de sélection responsables. Cependant, nous ne disposons que de deux dates d'isolement des souches dans notre étude, espacées d'un faible nombre d'années, ce qui rend difficile de telles analyses. Idéalement, il nous faudrait alors inclure dans cette analyse un éventail de génomes, en conservant si possible une homogénéité dans la population de patients considérée.

De manière surprenante, l'une de ces émergences concerne un groupe de souches de clade A au sein du ST131. Les récentes études sur le ST131 se sont généralement concentrées sur le clade B et surtout sur le clade C en raison de sa forte association à la multirésistance aux antibiotiques (Ben Zakour et al., 2016; McNally et al., 2019). Néanmoins, l'analyse rétrospective des données de Kallonen *et al.* nous montre qu'il y a une légère tendance à l'augmentation du clade A en 2012 (Figure 21). Bien que la taille de l'échantillon soit ici très faible, et que ces souches anglaises ne portent pas les mêmes déterminants de virulence et

résistance que nos souches, cela montre le caractère dynamique de ce STc131, et suggère que d'autres bouleversements pourraient survenir dans le temps.

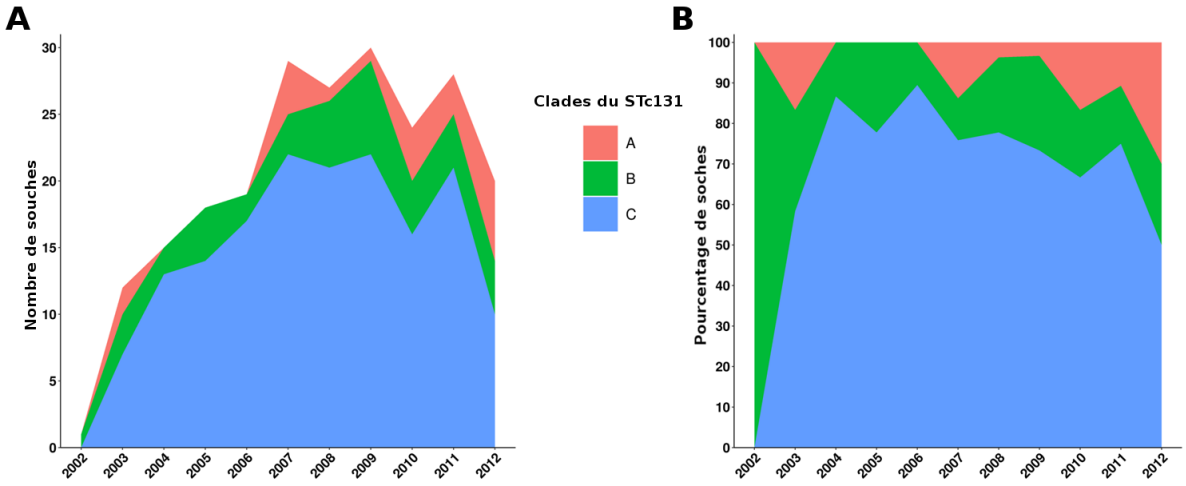


Figure 21. Distribution des souches de STc131 de l'étude de Kallonen *et al.* (Kallonen et al., 2017) au sein de différents clades en valeur absolue (A) et pourcentage (B).

Par ailleurs, le groupe de souches de clade A isolées en 2016-2017 dans Septicoli présente à la fois des déterminants de résistance (*bla*CTX-M-27 et mutation dans la *gyrA*) et de virulence avec l'allèle *papGII*. La question de l'évolution future d'une telle lignée se pose donc puisqu'elle mime en partie celle observée précédemment dans le clade C2 aujourd'hui prédominant. La surveillance épidémiologique dans les années à venir devrait permettre de répondre à une telle question.

Enfin, la virulence et la résistance ne nous permettent pas d'expliquer l'ensemble de ces mouvements au sein de la population ExPEC. On observe par exemple une stabilité remarquable du STc73 pourtant peu porteur de gènes de résistance. D'autres part, la présence élevée de gènes de virulence est une caractéristique très fréquente dans l'ensemble des phylogroupes B2 et D. L'analyse du métabolisme est donc un autre axe à envisager, considérant la possibilité que ces clones pandémiques sont peut-être avant tout de bons colonisateurs intestinaux. Il a d'ailleurs été mis en évidence que l'abondance au niveau intestinal est une étape essentielle pour les uropathogènes avant la réalisation d'une infection (Magruder et al., 2019).

III. Reconstruction des réseaux métaboliques de souches de *Escherichia coli* commensales et pathogènes

1. Introduction

Les gènes de virulence des souches ExPEC ne suffisent pas à expliquer la sévérité des infections. Landraud *et al.* ont fait ce constat en analysant un nombre limité de facteurs de virulence dont la présence est bien associée à la virulence chez la souris, mais ne se traduit pas par une aggravation du pronostic chez l'humain (Landraud et al., 2013). L'étude princeps de Septicoli semble corroborer ces observations (de Lastours et al., 2020). En effet, tout comme dans l'étude Colibafi (Lefort et al., 2011), les facteurs associés à l'hôte jouent un rôle primordial dans la sévérité de l'infection. Seul un nombre limité de déterminants bactériens ressortent associés au décès, et soit reflètent en réalité la porte d'entrée de l'infection soit sont retrouvés avec une très faible prévalence (de Lastours et al., 2020). Cependant, bien que disposant de l'intégralité des séquences des souches, nous avons procédé à une analyse ciblée sur un nombre restreint de gènes de virulence au cours de cette étude. Cela n'exclut donc pas pour autant le rôle d'autres facteurs microbiens dans la pathogénicité des souches et, notamment, dans la porte d'entrée des bactériémies. D'un point de vue théorique, le métabolisme semble être un bon candidat pour analyser des facteurs bactériens associés à la capacité de coloniser ou de survivre dans le milieu extraintestinal comme le tractus urinaire ou l'arbre respiratoire par exemple. Certains auteurs ont d'ailleurs mis en évidence les capacités métaboliques particulières de certaines souches responsables d'infection urinaire (Anfora et al., 2007). Par ailleurs, comme le suggèrent certaines études, la virulence pourrait être un produit dérivé du commensalisme (Adiba et al., 2010; Le Gall et al., 2007) et il est donc probable que des clones ExPEC particulièrement virulents et/ou actuellement en pleine expansion à l'échelle mondiale présentent avant tout un fort pouvoir colonisateur au niveau digestif.

Afin de mieux comprendre les liens entre métabolisme et mode de vie et tenter d'identifier des voies métaboliques spécifiques de clones pandémiques comme le STc131, nous avons reconstruit les réseaux métaboliques à partir de l'analyse des génomes d'un ensemble de souches commensales fécales, de souches issues de bactériémies à différentes portes d'entrée, et de souches pathogènes ou colonisatrice de l'arbre respiratoire.

2. Matériel et méthodes

a. Collections de souches

Nous avons analysé les génomes de souches provenant de quatre collections. La collection Coliville (n=280) est constituée de souches commensales fécales isolées en 2010 en région parisienne (Massot et al., 2016). Les critères d'inclusion stricts au cours de cette étude en garantissent le caractère purement commensal : absence de pathologie gastrointestinale, absence d'immunosupression, aucune antibiothérapie au cours du mois précédent et absence d'hospitalisation dans les trois mois précédents. Nous avons également inclus les souches recueillies au cours de bactériémies de l'adulte en région parisienne en 2005 et 2016-17 et qui correspondent aux études Colibafi (n=367) et Septicoli (n=545) (de Lastours et al., 2020; Lefort et al., 2011). Enfin, sont incluses des souches isolées de pneumopathies acquises sous ventilation ou de colonisation respiratoire entre 2012 et 2014 en France grâce à l'étude Colocoli (n=216) (La Combe et al., 2019).

b. Séquençage et typage des souches

Les souches ont été séquencées par Illumina NextSeq 2x150 pb, après préparation de librairie NextEra XT (Illumina, San Diego, CA, USA). Tous les génomes ont été assemblés avec le programme shovill v1.0.4 (Seemann, 2016/2021) combinant plusieurs logiciels (SPAdes v3.13.1, Mash v2.1.1, FLASH v1.2.11, bwa-index 0.7.17-r1188, samtools v1.9, pilon v1.23). L'absence de contamination des séquences a été vérifiée à l'aide du logiciel CheckM v1.0.11 (Parks et al., 2015). Nous avons réalisé le typage des souches en déterminant le phylogroupe et le MLST d'après le schéma de l'université de Warwick (Beghain et al., 2018).

c. Construction du pangénome, phylogénie

Dans un second temps nous avons réalisé l'annotation syntaxique, la reconstruction du pangénome de nos souches et la détection des régions plasticités génomiques avec le logiciel PPanGGOLiN v1.1.81, en utilisant les paramètres standards (*i.e.* 80% de couverture et 80% d'identité protéique pour la création des familles de gènes homologues) (Bazin et al., 2020; Gautreau et al., 2020). Puis nous avons construit un arbre phylogénétique à partir du coregénome de nos souches en considérant uniquement les familles monogéniques présentes dans 99% de nos génomes (2565 gènes) à l'aide du logiciel IQ-TREE v1.6.12 avec le modèle GTR+F+I+G4 (Nguyen et al., 2015).

d. Reconstruction des réseaux métaboliques

Afin d'optimiser l'analyse, nous avons réalisé la reconstruction des réseaux métaboliques à l'échelle du pangénome puis, dans un second temps, nous somme revenus à l'échelle du génome pour inférer les réactions et les voies métaboliques présentes dans chacun de nos isolats. Nous avons procédé en trois étapes pour l'identification des réactions associées aux protéines. Dans un premier temps, nous avons aligné l'ensemble des séquences des protéines représentatives de chaque famille du pangénome sur celles des protéines de la souche type *E. coli* K-12 qui sont associées à des réactions dans la base de données Ecocyc (Keseler et al., 2017). Pour cela, nous avons utilisé le logiciel Diamond v0.9.30 avec comme paramètres 80% d'identité et de couverture (Buchfink et al., 2015). Etant donné que la base EcoCyc est uniquement centrée sur *E. coli* K-12, nous avons enrichi les annotations par deux autres approches : tout d'abord le logiciel KofamKOALA en ne considérant que les familles associées à un numéro E.C. (Enzyme Commission) (Aramaki et al., 2020); puis le logiciel DeepEC (Ryu et al., 2019) sur les protéines encore non annotées par les approches EcoCyc et KofamKOALA. Nous avons ensuite reconstruit les réseaux métaboliques à partir de ces annotations à l'aide du logiciel Pathway-tools v23.5 qui utilise la base de données généraliste MetaCyc comme référence (Karp et al., 2021). Ce réseau à l'échelle du pangénome regroupe l'ensemble des réactions et voies métaboliques pouvant être prédites à partir des génomes étudiés. Dans la dernière étape, nous somme revenus à l'échelle du génome en attribuant les réactions et les voies métaboliques présentes dans chaque isolat. Une réaction est considérée comme présente dans un isolat si son génome contient au moins un gène de la ou les famille(s) associée(s) à cette réaction. Concernant les voies métaboliques, nous avons au préalable calculé, à l'échelle du pangénome, la complétion de chaque voie qui correspond au nombre de réactions prédites (*i.e.* associées à une famille du pangénome) divisé par le nombre total de réactions non spontanées de la voie métabolique. Pour chaque génome, des complétions relatives ont ensuite été calculées qui correspondent, pour une voie métabolique, au nombre de réactions prédites dans un génome divisé par le nombre total de réactions prédites à l'échelle du pangénome. Pour certaines analyses, ces données de complétion C ont été traduites en trois catégories : absence si $C = 0$, incomplète si $0 < C =< 0,5$, présence si $C > 0,5$.

e. Recherche d'associations entre mode de vie, phylogroupe, STc et métabolisme

Dans l'objectif d'identifier des voies métaboliques associées à un mode de vie particulier, nous avons analysé les motifs de présence/absence des réactions et des voies métaboliques en fonction de ces modes de vie et, parallèlement, en fonction des phylogroupes. Après calcul des distances entre les souches ("Average Manhattan Distance") à partir des vecteurs de présence/absence en considérant une voie incomplète comme absente, nous avons réalisé un regroupement des souches à partir de cette matrice de distance (méthode UPGMA - Unweighted Pair Group Method with Arithmetic mean) puis représenté les résultats sous la forme d'un arbre (librairie R *ape*).

Dans un second temps, nous nous sommes intéressés à la complétion des voies à l'échelle du phylogroupe et du mode de vie. Pour chaque voie, nous avons comparé les données de complétion pour les souches en fonction du phylogroupe ou du mode de vie par ANOVA, puis conservé uniquement les plus significatives (p-value corrigée par Benjamini-Hochberg < 0,05). Nous avons ensuite réalisé une classification des voies à partir des valeurs de complétion moyenne de chaque phylogroupe ou mode de vie en utilisant des distances euclidiennes et la méthode "Ward" de classification hiérarchique. Étant donné le nombre parfois important de voies obtenues, nous nous sommes concentrées sur celles dont les différences de complétions moyennes étaient les plus variables c'est-à-dire supérieures à 0,5 entre deux phylogroupes, et supérieures à 0,2 entre deux modes de vie.

Enfin, nous avons réalisé une analyse identique en prenant en compte les cinq STc majeurs de bactériémies (STc131, STc73, STc69, STc95, STc10) et l'ensemble des autres ST.

f. Analyse des correspondances multiples

Dans le but d'explorer conjointement les données de phylogroupes, STc et mode de vie, nous avons réalisé une analyse des correspondances multiples à l'aide de la librairie FactoMineR (Lê et al., 2008). Pour cela, nous avons utilisé comme variables explicatives les niveaux de complétion (absence, incomplète, présence) des voies métaboliques, après exclusion des voies présentes dans plus de 95% ou moins de 5% des souches. Nous avons ensuite identifié les voies dont la contribution aux deux premiers axes était la plus importante (p-value corrigée < 0.05) afin de déterminer leur classification d'après la base de données MetaCyc.

g. Analyse d’une voie métabolique d’intérêt

Nous avons analysé plus en détail une voie métabolique impliquée dans la dégradation des composés aromatiques absente des souches de groupe B2 à l’exception des STc131, STc452, ST136/ST681. Pour cela, nous avons extrait les séquences correspondant aux gènes codant cette voie et leur voisinage au sein de génomes de référence issus de GenBank pour chacun des phylogroupes et sous-groupes pour le groupe B2, en privilégiant les génomes complets circularisés dans le but d’écarter les séquences fragmentées liées aux assemblages de lectures courtes. Nous avons également extrait les séquences de chacun des gènes de la voie d’intérêt au sein des souches des différentes collections analysées, puis les avons alignées et enfin concaténées afin de réaliser une analyse phylogénétique avec le logiciel IQ-TREE v1.6.12 (modèle GTR+F+I+G4) dans le but d’identifier d’éventuelles incongruences par rapport à la phylogénie obtenue à partir du génome complet. Enfin, nous avons réalisé une analyse phylogénétique à partir des génomes complets de référence préalablement sélectionnés afin d’identifier à un éventuel scénario évolutif associé à la présence de notre voie d’intérêt.

3. Résultats

a. Modes de vie et phylogroupes

Les différents modes de vie et la distribution des phylogroupes au sein des collections sont présentés dans le Tableau 6. Au total, parmi les 1408 souches de *E. coli* et *Escherichia* clades incluses dans cette étude, 280 (19,9%) sont des souches commensales fécales, 470 (33,4%) sont issues de bactériémies à porte d’entrée urinaire, 281 (20,0%) à porte d’entrée digestives et 235 (16,7%) sont liées à une origine pulmonaire (bactériémies, pneumopathie ou colonisation). Par ailleurs, 142 souches (10,1%) sont issues de bactériémies de portes d’entrée plus rares (cutanées, cathéter, infection de site opératoire) ou bien non définies (multiple ou inconnue). On observe un enrichissement des souches de phylogroupes B2 dans les souches isolées de bactériémies à porte d’entrée urinaire et les souches pulmonaires comparées aux souches commensales.

Tableau 6. Mode de vie et phylogroupes des souches étudiées.

Mode de vie	Commensal	Bactériémies				Pulmonaire (bactériémies, PAVM*, colonisation)	Total
		Urinaire	Digestive	Autre	Inconnue ou multiple		
Nombre de génomes	280	470	281	46	96	235	1408
Phylogroupes (%)							
A	76 (27.1)	26 (5.5)	52 (18.5)	8 (17.4)	8 (8.3)	25 (10.6)	195 (13.8)
B1	38 (13.6)	28 (6)	49 (17.4)	6 (13)	11 (11.5)	31 (13.2)	163 (11.6)
B2	94 (33.6)	290 (61.7)	108 (38.4)	15 (32.6)	48 (50)	124 (52.8)	679 (48.2)
C	8 (2.9)	27 (5.7)	12 (4.3)	4 (8.7)	7 (7.3)	20 (8.5)	78 (5.5)
D	27 (9.6)	82 (17.4)	39 (13.9)	9 (19.6)	12 (12.5)	23 (9.8)	192 (13.6)
E	10 (3.6)	3 (0.6)	1 (0.4)	1 (2.2)	2 (2.1)	1 (0.4)	18 (1.3)
F	21 (7.5)	11 (2.3)	10 (3.6)	3 (6.5)	5 (5.2)	6 (2.6)	56 (4)
G	4 (1.4)	2 (0.4)	7 (2.5)	0 (0)	2 (2.1)	5 (2.1)	20 (1.4)
Escherichia clades	2 (0.7)	1 (0.2)	3 (1.1)	0 (0)	1 (1)	0 (0)	7 (0.5)

*PAVM : pneumopathie acquise sous ventilation

b. Pangénome, voies métaboliques et réactions

La construction du pangénome à l'aide du logiciel PPanGGOLiN a abouti à un total de 39613 familles de gènes, découpées en trois entités : un génome persistant de 3455 familles, un "shell" de 4558 familles et un "cloud" de 31600 familles. En termes de réactions métaboliques, sur les 3655 réactions retrouvées, la part de réactions conservées dans presque toutes les souches (*i.e.* dont la fréquence est > 95%) atteint 1202 réactions (32,9%), alors que 1982 réactions (54,2%) sont présentes dans moins de 15% des souches et 471 (12,9%) entre 15 et 95% (Figure 22). On retrouve donc une distribution en "U" des réactions et des familles de gènes, comme décrit précédemment chez *E. coli* (Touchon et al., 2009; Vieira et al., 2011). Concernant les voies métaboliques, elles sont plus conservées avec 365 des 477 voies (76,5%) présentes dans plus de 95% des souches (Figure 22).

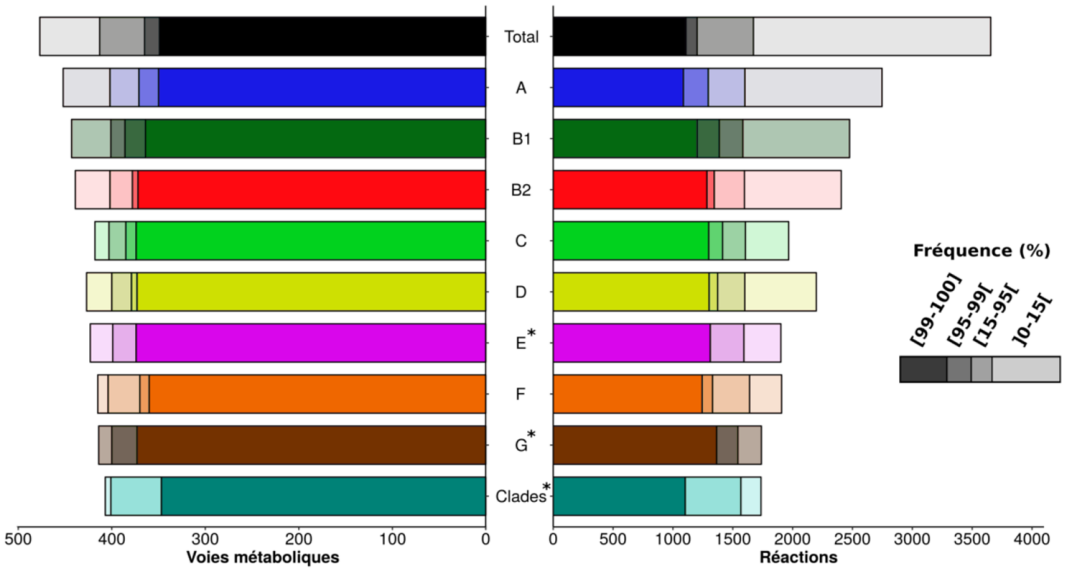


Figure 22. Nombre de voies métaboliques et de réactions au sein de l'ensemble des souches analysées en fonction de leur fréquence. Les phylogroupes/Clades marqués d'un astérisque n'ont aucune voie/réaction retrouvée à une fréquence comprise entre 95 et 99% des souches.

En termes de classification des voies métaboliques, les voies de biosynthèse et de dégradation sont les plus nombreuses, représentant 43,2% et 38,2% des voies, respectivement. Les voies de biosynthèse sont plus conservées que les voies de dégradation et les voies du métabolisme des glycanes (p-value corrigée < 0.05) (Figure 23). De même, les voies impliquées dans le métabolisme énergétique sont plus conservées que les voies de dégradation, du métabolisme des glycanes ou de détoxification (p-value corrigée < 0.05).

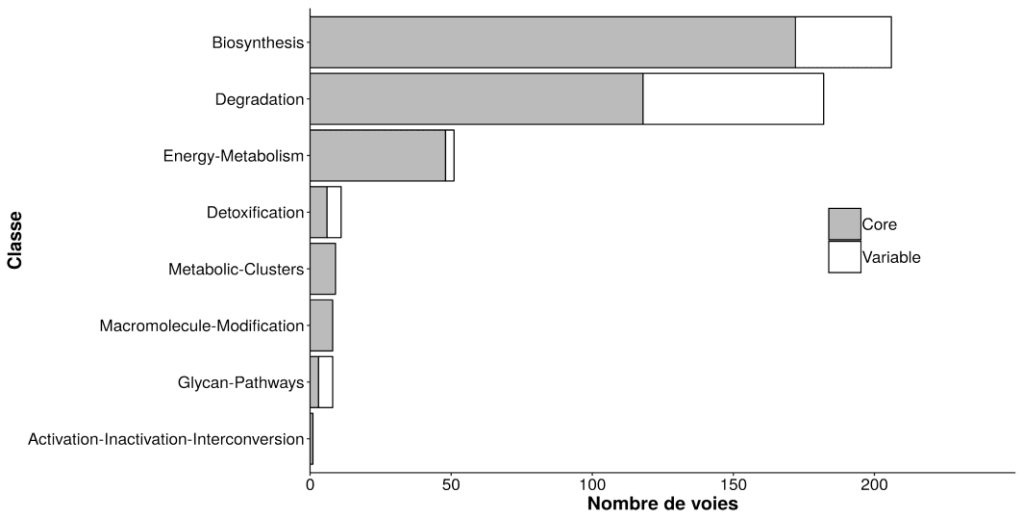


Figure 23. Distribution des voies dans le métabolisme core (≥ 95%) et variable (< 95%) en fonction de leur classification d'après la base de données MetaCyc.

Si l'on considère le second niveau de classification des voies proposée par dans la base de données MetaCyc, on observe des différences marquantes comme par exemple une très forte conservation des voies de biosynthèse des acides aminés et des nucléotides, des voies impliqués dans le transfert d'électrons et la respiration, alors que les voies de dégradation des composés aromatiques, par exemple, sont toutes variables (Figure 24).

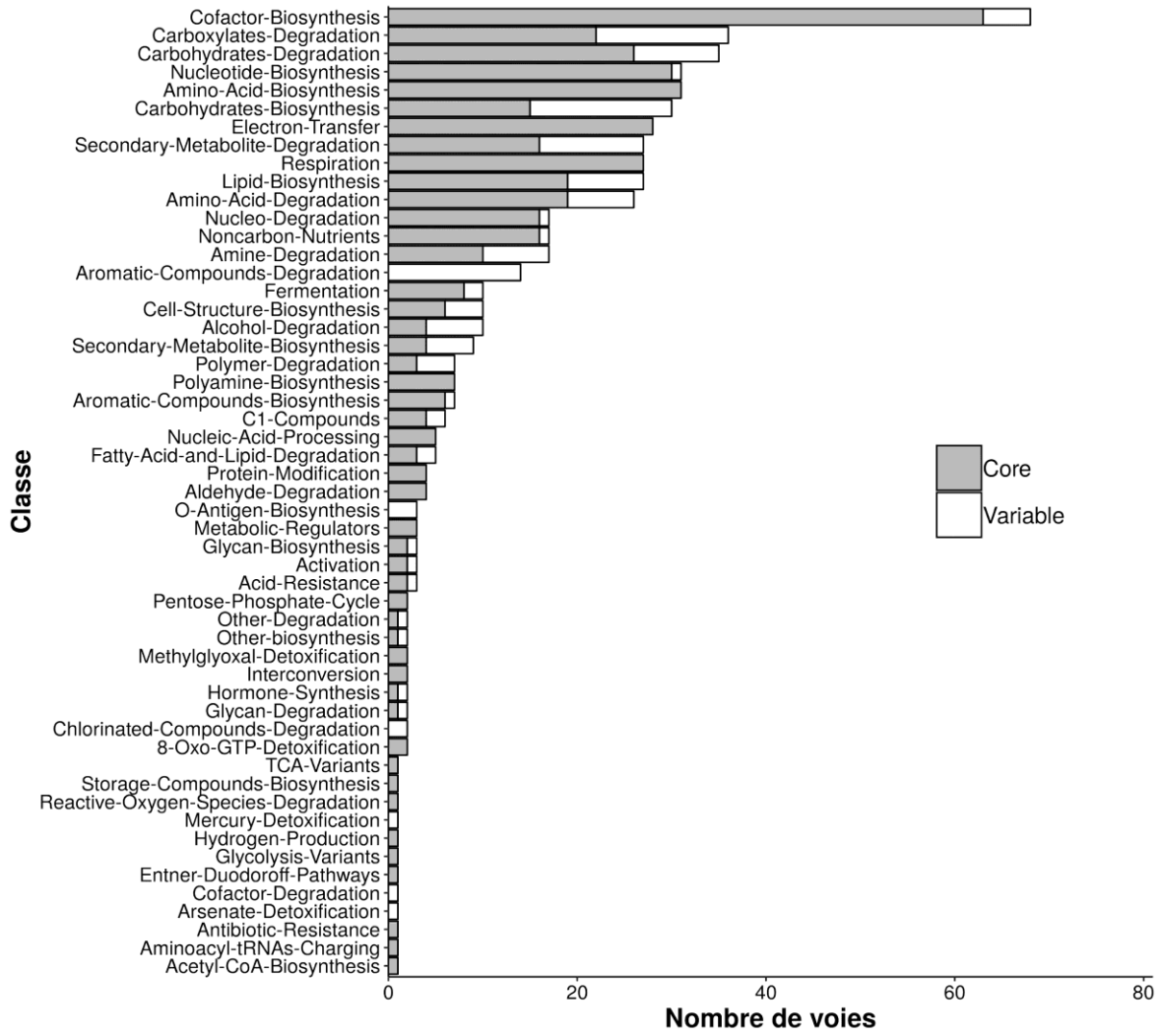


Figure 24. Distribution des voies dans le métabolisme core (≥ 95%) et variable (< 95%) en fonction de leur classification de deuxième niveau d'après la base de données MetaCyc.

c. Associations aux phylogroupes et modes de vie

Nous avons réalisé une classification hiérarchique des souches à partir des vecteurs de présence/absence des voies métaboliques et des réactions (Figure 25). Les arbres obtenus montrent une topologie fortement liée à la phylogénie de l'espèce. En effet, on retrouve une séparation en fonction des phylogroupes qui est plus nette pour les présences/absences de réactions que pour les voies métaboliques (Figure 25). Ceci s'explique notamment par le fait que la présence des réactions découle directement de la présence des gènes les codant au sein de souches. Pour les voies métaboliques, la séparation est moins précise et fait apparaître trois groupes (Figure 25) : i) le phylogroupe B2, ii) les phylogroupes A, B1 et C, iii) les phylogroupes D, E, F et G. Les souches de clade I, clade le plus proche de l'espèce *E. coli* sensu stricto, sont groupées à proximité des souches de groupe F.

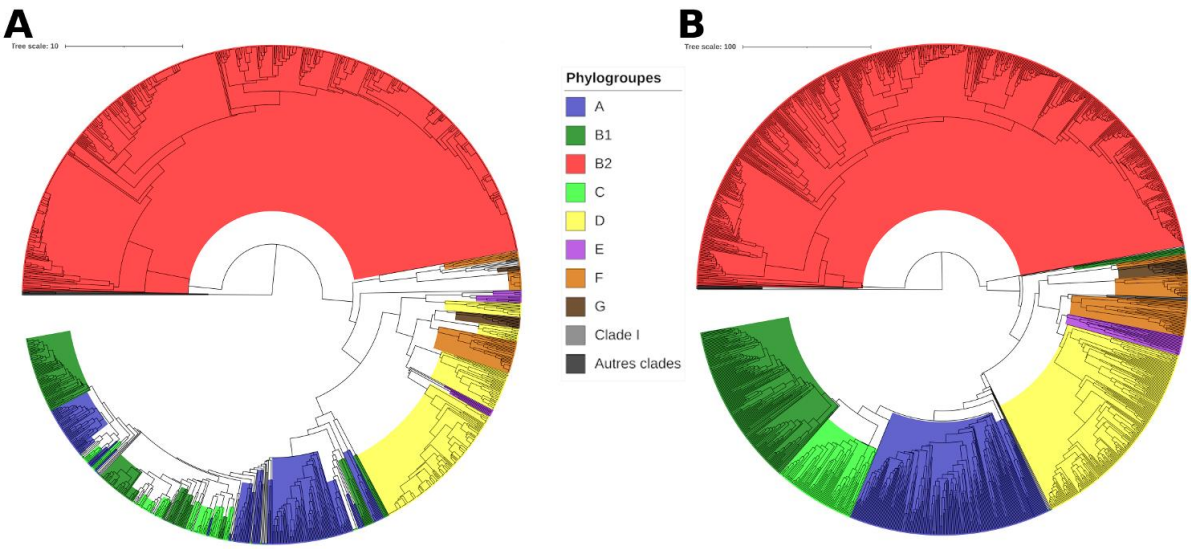


Figure 25. Arbres construits par la méthode UPGMA à partir des vecteurs de présence/absence A) des voies métaboliques, B) des réactions. Les phylogroupes des souches sont indiqués par un code couleur.

L'analyse de la complétion des voies métaboliques en fonction des phylogroupes et des modes de vie tend également à regrouper les souches en fonction de la phylogénie de l'espèce. En se concentrant sur les voies métaboliques dont la complétion est la plus variable entre les phylogroupes, nous obtenons un regroupement identique à l'arbre de présence/absence, avec de nouveau un phylogroupe B2 très distinct (Figure 26). Concernant les modes de vie, ils sont regroupés entre origine commensale, digestive et autre, d'un côté, et urinaire, pulmonaire et multiple ou inconnu de l'autre (Figure 27). Parmi les 23 voies pour

lesquelles la différence de complétion est supérieure à 0,2 entre les différents modes de vie, 22 ont une différence de complétion supérieure à 0,5 entre les différents phylogroupes, soulignant à nouveau l'empreinte de la phylogénie. La seule exception est la voie métabolique de l'aérobactine, un sidérophore fréquemment décrit dans les souches ExPEC, qui semble enrichie dans les souches de bactériémie à point de départ urinaire. Néanmoins cette voie reste plus fréquemment retrouvée dans les groupes B2, C, D, F et G que A et B1.

Figure 26. Heatmap des valeurs de complétion moyenne des voies métaboliques par phylogroupe. La classification hiérarchique des phylogroupes a été obtenue par la méthode Ward avec des distances euclidiennes. Les voies pour lesquelles la différence de complétion est > 0,5 entre deux phylogroupes sont indiquées en rouge.

Figure 27. Heatmap des valeurs de complétion moyenne des voies métaboliques par mode de vie. La classification hiérarchique des modes de vie a été obtenue par la méthode Ward avec des distances euclidiennes. Les voies pour lesquelles la différence de complétion est > 0,2 entre deux modes de vie sont indiquées en rouge.

Afin d'avoir une vision plus globale de ces données, nous avons réalisé une analyse des correspondances multiples en utilisant trois niveaux de complétion (présence, incomplète, absence) des voies métaboliques les plus variables comme variables explicatives et les phylogroupes, STc et modes de vie comme variables illustratives. La Figure 28 représente les deux premiers axes de cette analyse, qui expliquent respectivement 25,7% et 11,2%. Les voies qui contribuent significativement au premier axe sont au nombre de 61 et sont essentiellement associées à la dégradation (n=47) et à la biosynthèse (n=18), et, notamment, la dégradation des composés aromatiques (n=13). Pour le second axe, ce sont 58 voies qui contribuent significativement, à nouveau essentiellement impliquées dans la dégradation (n=45) et la biosynthèse (n=17). Les modes de vie sont peu séparés par cette analyse : les bactériémies à porte d'entrée urinaire et les souches d'origine pulmonaire ont tendance à se placer sur la partie négative du premier axe à l'opposé des portes d'entrée digestive et "autre" ainsi que des souches commensales. Ce sont essentiellement les phylogroupes et les STc qui sont séparés par ces axes, avec les souches de phylogroupe B2 sur la partie négative du premier axe à l'opposé des souches des autres phylogroupes. Les phylogroupes D, E et F sont également séparés des phylogroupes A, B1 et C selon le deuxième axe.

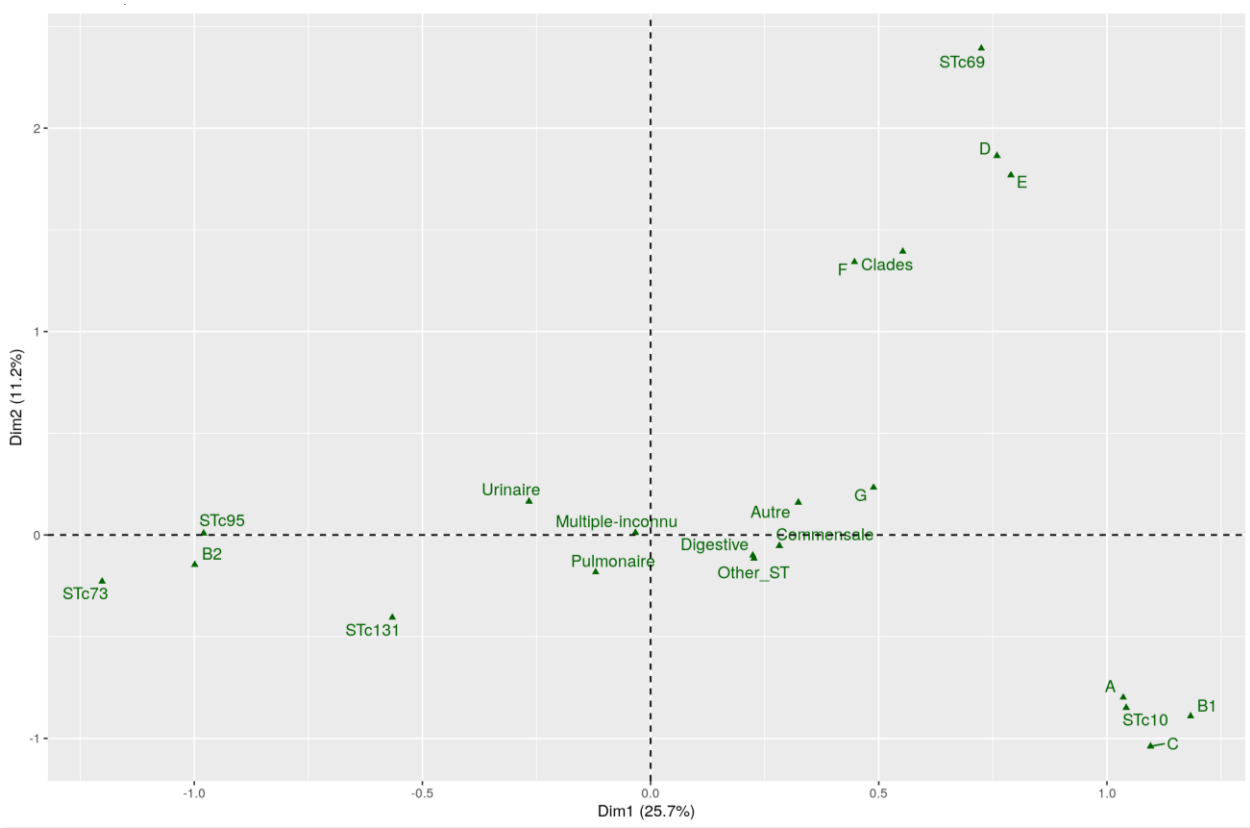


Figure 28. Représentation des deux premiers axes d'une analyse des correspondances multiples à partir des niveaux de complétion des voies métaboliques. Les phylogroupes, les STc et le mode de vie sont utilisés comme variables illustratives.

d. Voies métaboliques et STc majeurs des bactériémies

Puisque les analyses précédentes ne nous ont pas permis d'identifier de voies métaboliques spécifiquement associées à un mode de vie, nous avons souhaité nous concentrer sur les liens entre métabolisme et phylogénie. Pour cela, nous avons analysé plus en détail les voies métaboliques des cinq STc majoritairement retrouvés dans les bactériémies à l'heure actuelle en France et dont la dynamique a été présentée au cours du chapitre précédent. Plus particulièrement, nous avons cherché à identifier des voies métaboliques spécifiques du STc131 comparé aux autres STc de groupe B2. Les études sur la virulence, qu'il s'agisse de la prévalence des gènes de virulence ou bien d'expérimentation en modèle animal, n'ont pas montré de profil particulier pour le STc131. La présence de voies métaboliques particulières pourrait donc être une raison alternative à son succès. A nouveau, nous avons analysé les complétions des voies métaboliques les plus variables entre ces STc (Figure 29).

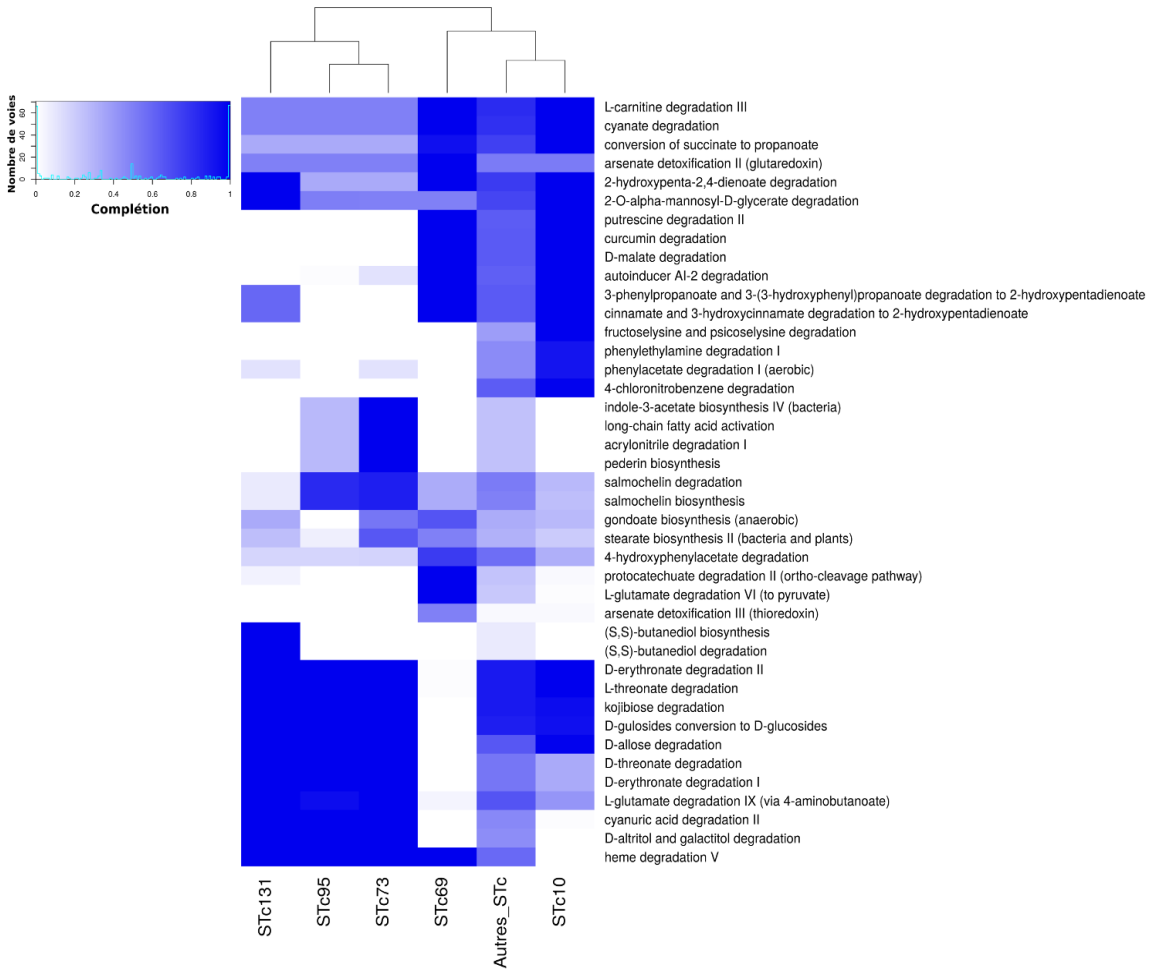


Figure 29. Heatmap des valeurs de complétion moyenne des voies métaboliques pour les cinq STc majeurs des bactériémies et le reste de STc. La classification hiérarchique des STc a été obtenue par la méthode Ward avec des distances euclidiennes.

Comme on peut le voir, les STc de phylogroupe B2 sont regroupés ensemble, mais le STc131 apparaît légèrement divergent des STc95 et STc73, ce que l'on observait également lors de l'analyse des correspondances multiples (Figure 28). Parmi les voies qui expliquent ces divergences, on peut noter l'absence des voies de biosynthèse et de dégradation de la salmocheline dans le STc131. Le groupe de gènes *iro* codant cette voie de la salmochelline est fréquemment décrit comme plasmidique, notamment dans le STc95, comme on a pu le constater au cours du chapitre précédent. Concernant les voies spécifiquement présentes dans STc131, on trouve les voies de dégradation et de biosynthèse du (S,S)-butanediol. Ces deux voies correspondent en réalité à une seule et même réaction réversible, habituellement non observées chez *E. coli*, et sont prédites en raison de la présence d'un gène qui code une protéine présentant une faible identité (35%) avec celles codées par les gènes *budC* de *Klebsiella pneumoniae* et *butA* de *Corynebacterium glutamicum*. L'annotation de la protéine d'après le logiciel InterProScan (Jones et al., 2014) correspond à la famille des "Short-chain dehydrogenase/reductase" et ne permet donc pas de conclure réellement sur sa fonction précise. Un autre ensemble de voies métaboliques est associé avec le STc131 mais absent des STc73 et STc95. Ces voies sont impliquées dans la dégradation de composés aromatiques. La première est la voie de dégradation du 3-phenylpropanoate et du 3-(3-hydroxyphenyl)propanoate en 2-hydroxypentadienoate, qui permet également la dégradation du cinnamate et du 3-hydroxycinnamate en 2-hydroxypentadienoate (Figure 30). En réalité cette voie est composée de deux branches, l'une pour le 3-phenylpropanoate ou le cinnamate et l'autre pour le 3-(3-hydroxyphenyl)propanoate ou le 3-hydroxycinnamate. Seule cette dernière est présente dans le STc131 ce qui explique le niveau de complétion imparfait observé, contrairement aux souches des phylogroupes A, B1, C, D, E, F, G et des *Escherichia* clade I et V. La seconde voie utilise comme substrat le produit de la précédente, il s'agit de la voie de dégradation du 2-hydroxypenta-2,4-dienoate qui aboutit à la formation d'acétyl-coA qui intègre ensuite le cycle de Krebs. L'une des trois réactions de cette voie peut être réalisée par une enzyme alternative codée par le gène *eutE*, un gène du coregénome ce qui explique que la complétion de cette voie ne soit pas égale à zéro pour les souches de STc73 et STc95.

Nous avons ensuite exploré plus en détail ces dernières voies qui sont codées par le groupe de gènes mhp*RABCDFET* et impliquées dans le métabolisme des produits de dégradation de la lignine (Díaz et al., 2001). Parmi nos 1408 souches, les souches possédant ce groupe de gènes et appartenant au phylogroupe B2 sont de STc131 ou proches (ST83, ST91 et ST7882), de STc452, ou bien de ST136 ou ST681. Ces deux derniers ST ne diffèrent entre eux que d'un allèle. Le STc452 est remarquable au sein du phylogroupe B2 car il est considéré comme un commensal strictement humain (Clermont et al., 2008).

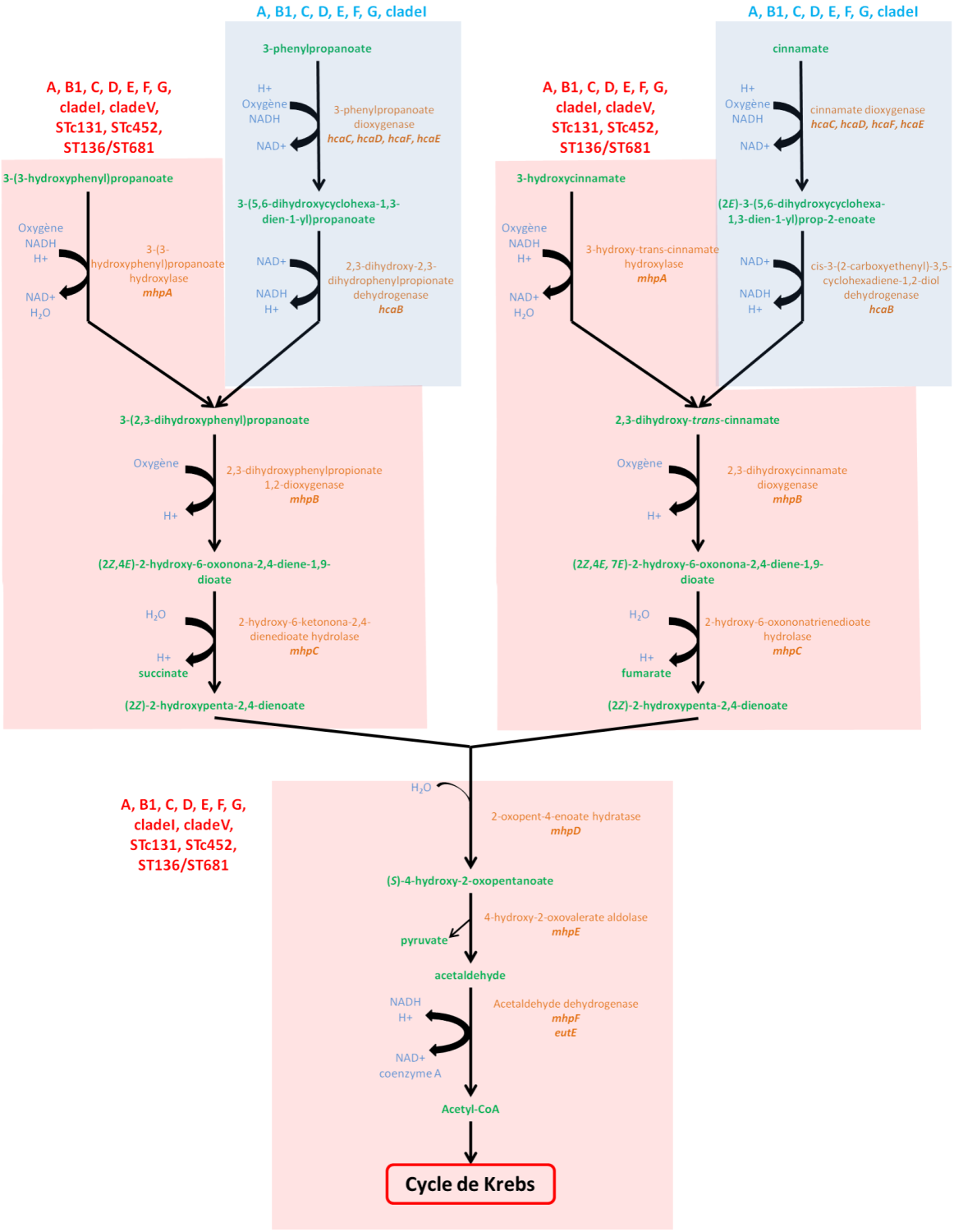


Figure 30. Voie de dégradation du 3-phenylpropanoate, 3-(3-hydroxyphenyl)propanoate, cinnamate et 3-hydroxycinnamate en 2-hydroxypentadienoate, et voie de dégradation du 2-hydroxypenta-2,4-dienoate. Les zones colorées en rouge sont présentes dans les phylogroupes A, B1, B2, C, D, E, F, G, *Escherichia* clade I et clade V, STc131 et STc452 ainsi que de rares ST proches. Les zones colorées en bleu sont absentes de toutes les souches de groupe B2 (adapté des voies métaboliques HCAMHPDEG-PWY et PWY-5162 de la base de donnée MetaCyc).

Dans l'ensemble des souches possédant les gènes *mhpRABCDFET*, on retrouve ces éléments à proximité directe de l'opéron lactose. La Figure 31 illustre un alignement des régions entourant l'opéron lactose, avec un génome représentant de chaque phylogroupe, de chaque sous-groupe B2, des *Escherichia* clades et des espèces *E. albertii* et *E. fergusonii*. Nous avons privilégié l'utilisation de génomes de référence complets lorsqu'ils étaient disponibles afin d'éviter les biais liés à la fragmentation en contigs. On peut noter la structure mosaïque de cette région avec des groupes entiers de gènes spécifiquement présents dans certains phylogroupes, laissant supposer de multiples remaniements probablement par le biais de recombinaisons homologues qu'il est difficile d'identifier. En amont des gènes *mhp*, on observe deux gènes dans les souches de phylogroupe E et des clades I et V (Figure 31, Figure 32). Ces gènes codent un régulateur transcriptionnel pour l'un, et une protéine hypothétique pour l'autre.

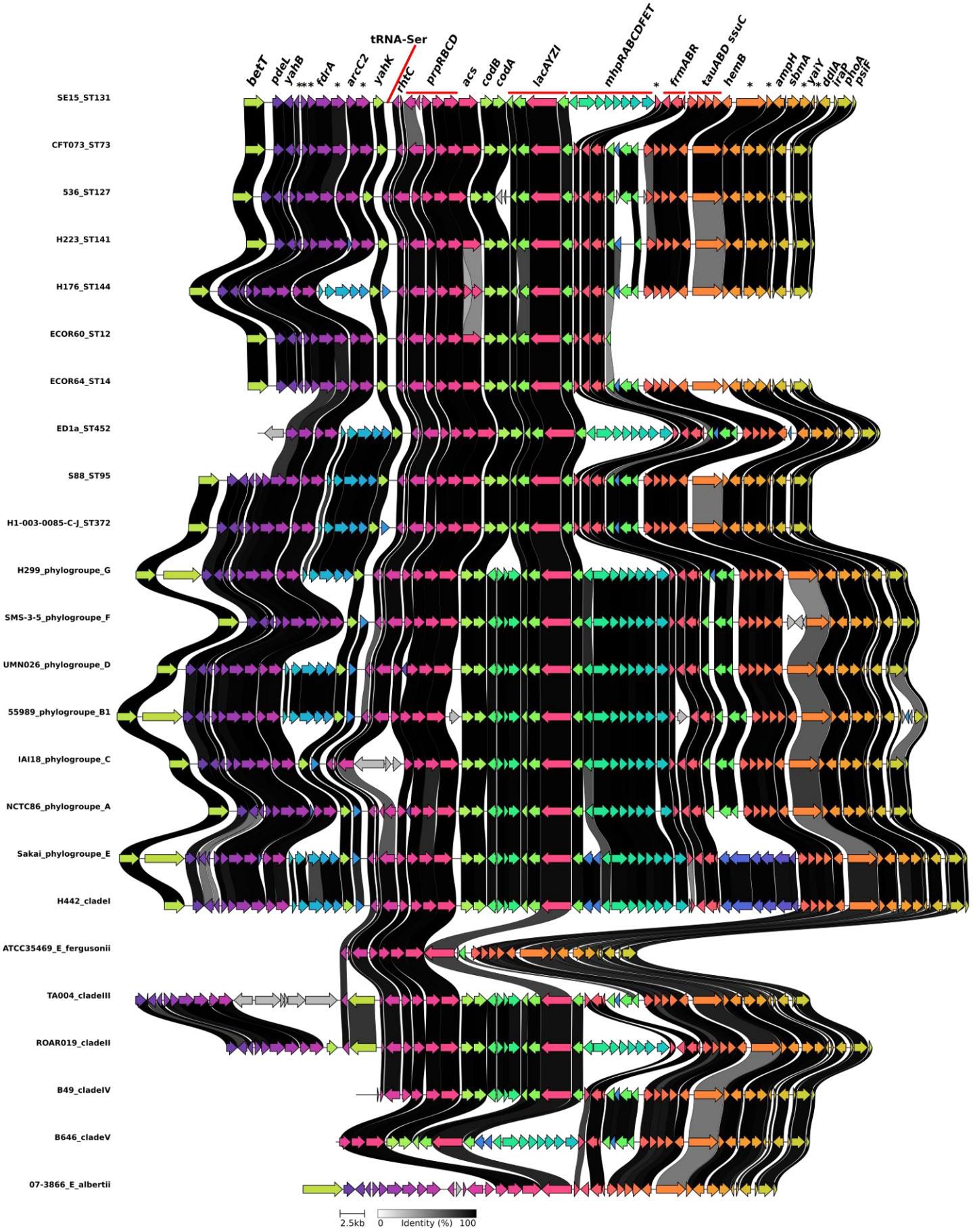


Figure 31. Alignement des régions entourant l'opéron lactose dans des génomes de différents phylogroupes, sous-groupes B2, *Escherichia* clades, *E. albertii* et *E. fergusonii.* L'annotation des séquences codantes dans la souche SE15 (ST131) est détaillée en haut de la figure. Les astérisques correspondent à des protéines hypothétiques.

Dans le but d'identifier les évènements ayant abouti à la présence de ce groupe de gènes *mhp* dans les STc131, STc452, ST136 et ST681, comme par exemple une recombinaison avec des souches de phylogroupe différent, nous avons réalisé un alignement des gènes puis une analyse phylogénétique. Mais les souches de groupes B2 apparaissent proches dans l'arbre ne permettant pas d'objectiver un quelconque échange de matériel génétique avec des souches génétiquement distantes. Nous avons ensuite analysé plus précisément les régions intergéniques de part et d'autre du groupe de gènes (Figure 32). Cette analyse met en évidence un fragment du gène codant le régulateur transcriptionnel des souches de phylogroupe E et clade I/V, chez les souches de groupe B2 non porteuses des gènes *mhp*. Un fragment plus court de ce gène est présent dans les souches de groupes A, B1, C, D, F, G, clade II et STc131/STc452/ST136/ST681. A l'aide de cette structure intergénique et de l'arbre phylogénétique obtenu à partir de l'alignement des gènes du coregénome de 27 souches de référence, nous proposons un scénario évolutif ayant abouti au motif de présence/absence observé chez *E. coli* pour ce groupe de gènes Figure 32. Selon cette hypothèse, le groupe de gènes *mhp* est présent chez l'ancêtre commun de *E. coli* et des *Escherichia* clades. S'en suivent plusieurs évènements de délétions de la région intergénique associée ou non à la délétion des gènes *mhp.* Toujours d'après cette hypothèse, nos gènes d'intérêt seraient présents chez l'ancêtre commun des souches de phylogroupe B2, tout comme dans le ST131 qui diverge le plus tôt dans les B2. Les gènes seraient ensuite délétés, probablement par un évènement de recombinaison homologue puis réacquis indépendamment par le STc452 et les ST136/ST681. D'autres hypothèses sont possibles, impliquant notamment des régions plus larges étant donné la structure mosaïque retrouvée autour l'opéron lactose (Figure 31). Néanmoins d'après l'éloignement phylogénétique des souches de phylogroupe B2 qui possèdent les gènes *mhp*, l'hypothèse de multiples réacquisitions indépendantes semble plus parcimonieuse que celle faisant intervenir des délétions pour chacun des sous-groupes B2.

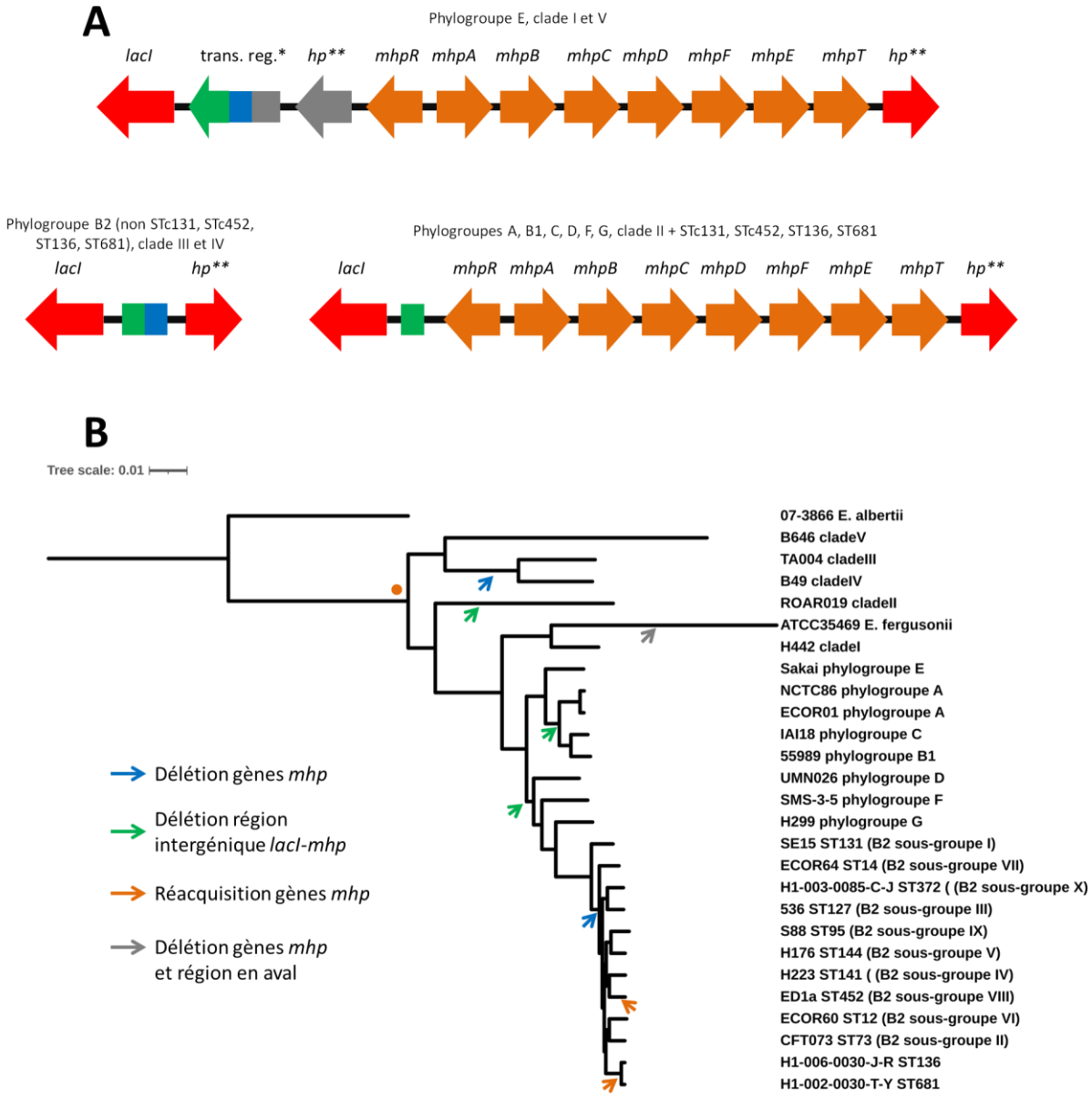


Figure 32. A) Représentation schématique de l'environnement génétique de la région contenant le groupe de gènes *mhp*. Les régions similaires entre les différents groupes de souches en amont de *lacI* sont notées en couleur. * : gène codant un régulateur transcriptionnel ; ** gènes codant des protéines hypothétiques. B) Arbre phylogénétique construit à partir des gènes du coregénome de 27 souches. Les flèches de couleur indiquent les possibles évènements de délétion et réacquisition. Le point orange indique la présence supposée des gènes *mhp* chez l'ancêtre commun des souches de *Escherichia* hors *E. albertii*. L'arbre est enraciné sur une souche de *E. albertii.*

4. Discussion

A l'aide de ces différentes collections de souches de *E. coli,* nous avons pu étudier à grande échelle le contenu métabolique de souches humaines commensales et pathogènes extraintestinales. Nous avons observé un nombre plus élevé de réactions qu'au cours de l'étude de Vieira *et al.* (Vieira et al., 2011), avec plus de 3600 réactions dans nos souches contre 1545 dans leur étude. Plusieurs raisons peuvent être évoquées pour expliquer de telles discordances. Tout d'abord, nous avons inclus dans notre étude plus de 1400 souches contre seulement 29 en 2011. De plus, bien que nos stratégies d'analyse soient relativement proches, nous avons pris le parti d'utiliser des bases de données additionnelles par rapport à Ecocyc. Une telle approche expose évidemment à un risque de "surprédiction" des réactions, mais permet en contrepartie de ne pas être centré uniquement sur *E. coli* K-12. Ce choix nous semblait cohérent, d'autant plus que les souches ExPEC sont généralement de phylogroupe B2 et D, et donc relativement éloignées des souches de phylogroupe A comme *E. coli* K-12. Pour autant nous retrouvons néanmoins certains points communs avec les études précédentes, notamment le niveau de conservation élevé (>75%) des voies métaboliques en comparaison au contenu en gènes des souches (Monk et al., 2013; Sabarly et al., 2016; Vieira et al., 2011). De même, nous observons une plus grande variabilité des voies de dégradations, notamment des composés aromatiques.

Nos analyses révèlent également une empreinte très forte de la phylogénie sur le métabolisme, rendant difficile l'identification de voies métaboliques spécifiquement associées à un mode de vie indépendamment de la structure de la population. L'une des hypothèses initiales de cette étude était la possible présence de voies spécifiquement associées, par exemple, à une porte d'entrée de la bactériémie, et essentielles à la colonisation de niches écologiques particulières comme le tractus urinaire ou respiratoire. Si aucune voie ne ressort spécifiquement, le rôle du métabolisme n'est pour autant pas à exclure et pourrait résulter de niveaux de régulation variables plutôt que du contenu en gènes comme précédemment proposé (Sabarly et al., 2011a). Cependant ces informations ne sont pas accessibles à partir des génomes et restent complexes à intégrer à une telle échelle.

Par ailleurs, certaines souches ExPEC pourraient posséder des capacités métaboliques impliquées dans l'étape initiale du processus infectieux, à savoir la colonisation intestinale. De même qu'il est évoqué un rôle potentiel dans le commensalisme de certains déterminants bactériens impliqués dans la virulence extraintestinale (Adiba et al., 2010; Le Gall et al., 2007), certaines voies métaboliques pourraient ainsi offrir un avantage sélectif à ces souches ExPEC

en leur garantissant un fort pouvoir colonisateur. Dans cette optique, nous avons poursuivi notre analyse en nous concentrant sur les 5 STc majoritairement retrouvés dans les bactériémies de l'adulte au cours du chapitre précédent. Nous avons pu identifier une voie impliquée dans la dégradation de composés aromatiques, le 3-(3-hydroxyphenyl)propanoate et le 3-hydroxycinnamate, codée par le groupe de gènes *mhp*, présente dans le STc131 mais absente des autres grands clones ExPEC de phylogroupe B2 (STc73 et STc95). Cette voie est habituellement décrite comme absente des souches de phylogroupes B2, à l'exception du STc452, un clone commensal et strictement humain (Clermont et al., 2008; Touchon et al., 2009, 2020). La recherche des gènes *mhp* dans les souches de phylogroupe B2 de notre jeu de données confirme cette prévalence limitée au STc131, STc452 et à deux souches de ST136 et ST681 proches phylogénétiquement. Lors de la première description de cette voie chez *E. coli,* Burlingame & Chapman mettaient déjà en évidence des différences en termes de présence/absence des voies de dégradation des aromatiques au sein d'isolats cliniques (Burlingame & Chapman, 1983), reflétant très probablement les différences liées à la phylogénie. Les substrats de cette voie proviennent essentiellement de la dégradation de la lignine par les bactéries anaérobies du tube digestif (Díaz et al., 2001), un composant des végétaux présent dans l'alimentation, et notamment dans l'élevage avec le son de riz, les pépins de raisin ou encore les graines de tournesol[6]. Le ST131 et STc452 sont fortement liées à l'Homme et cette voie pourrait leur procurer un avantage sélectif pour la colonisation intestinale humaine. Cependant les ST136/ST681 retrouvés dans la collection de Touchon *et al.* proviennent d'oiseaux (n=2), de mammifères non humains (n=7) et de l'eau (n=4) (Touchon et al., 2020). Cette voie est donc peut-être utile à une plus large échelle, au sein de différentes niches écologiques. L'enrichissement en voies de dégradation des aromatiques observé dans le phylogroupe B1 a d'ailleurs été proposé comme facteurs d'adaptation de ces souches à des niches environnementales comme l'eau, les sols ou les plantes (Touchon et al., 2020).

La présence de cette voie dans au sein du STc131, clone pandémique mondial dont l'émergence semble récente (Nicolas-Chanoine et al., 2014), pose par ailleurs la question de l'acquisition des gènes *mhp*. Nous avons recherché ces gènes au sein des génomes plus anciens décrits pour le ST131 (Ben Zakour et al., 2016), et avons bien retrouvé cette voie dans ces souches de 1967 et des années 1980. D'après notre hypothèse évolutive, cette voie serait en réalité ancestrale dans l'espèce *E. coli,* avec des évènements de délétions puis de réacquisition indépendants. Si cette hypothèse se confirme, ces réacquisitions indépendantes par des souches phylogénétiquement distinctes au sein du phylogroupe B2 représenterait un

[6] https://www.feedtables.com/fr/content/lignine

signal fort de convergence évolutive suggérant un avantage sélectif lié à l'utilisation de ces composés aromatiques.

L'émergence récente et la prévalence désormais élevée du ST131 dans les infections extraintestinales comme les bactériémies, suscitent des interrogations quant aux raisons d'un tel succès. L'implication de la multirésistance aux antibiotiques fréquemment observée chez ce clone a parfois été avancée puis réfutée en raison de la sensibilité aux antibiotiques des autres clones ExPEC toujours largement prévalents, comme les STc73 et STc95 (Ben Zakour et al., 2016; Kallonen et al., 2017; Nicolas-Chanoine et al., 2014). Les facteurs de virulence n'expliquent pas non plus à eux seuls ce phénomène, étant donné leur forte prévalence de manière générale chez les ExPEC. Il est donc probable que le succès de ce clone soit multifactoriel, et les capacités métaboliques offertes par le groupe de gènes *mhp* pourraient ainsi elles aussi participer. Nos résultats sont observationnels et ne prédisent en rien la fonctionnalité et l'intérêt d'une telle voie métabolique, et il est nécessaire de les évaluer à l'aide d'approches expérimentales, par exemple dans un modèle animal de colonisation intestinale. Néanmoins si son rôle dans le commensalisme venait à être confirmé, le ST131 bénéficierait à la fois d'un avantage sélectif dans la phase de colonisation, dans l'étape d'infection proprement dite par le biais de facteurs de virulence et dans la survie face à la pression antibiotique par le biais des déterminants génétiques de résistance.

Conclusions et perspectives

Les bactériémies sont des pathologies sévères et relativement fréquentes dans les pays industrialisés. *E. coli* occupe une place particulière dans ces infections qui tendent à augmenter depuis plusieurs années (van der Mee-Marquet et al., 2015; Vihta et al., 2018) et motivent parfois la mise en place de systèmes de surveillance dans certains pays comme le Royaume-Uni (J. Abernethy et al., 2017). Ces bactériémies surviennent préférentiellement chez des patients âgés (> 65 ans) (J. Abernethy et al., 2017; de Lastours et al., 2020; Lefort et al., 2011; Yoon et al., 2018), et le vieillissement de la population dans les pays occidentaux, comme en France[7] par exemple, laisse supposer que la prévalence de ces infections ne devrait pas diminuer dans les années à venir. Comme nous l'avons vu, l'espèce *E. coli* prise dans sa globalité est versatile, comprenant à la fois d'authentiques souches commensales mais également des souches pathogènes opportunistes intestinales et extraintestinales. Pour autant, cette ségrégation entre les modes de vie est complexe à appréhender dans les bactériémies, certaines souches peu virulentes pouvant être responsables d'une infection à la faveur d'une immunodépression ou de facteurs débilitants chez l'hôte. Les approches ciblées par des méthodes de PCR ont permis de proposer une définition des souches ExPEC en fonction du contenu en gènes de virulence (Thomas A. Russo & Johnson, 2000), mais se sont montrées peu efficaces pour la mise en évidence de déterminants bactériens associés à la sévérité des bactériémies (Lefort et al., 2011). Dans ce contexte, le développement massif des méthodes de séquençage semble prometteur en offrant virtuellement accès à l'intégralité de l'information génétique chromosomique et extrachromosomique des souches. Néanmoins, malgré l'augmentation du nombre de gènes de virulence détectés grâce à ces approches de génomiques, nous ne sommes parvenus à identifier qu'un faible nombre de facteurs bactériens associés à la gravité. De plus, le nombre réduit de souches porteuses de certains de ces déterminants dans notre collection est une limite certaine à l'interprétation de leur rôle dans la pathologie. La confirmation sur d'autres jeux de données et l'évaluation dans un modèle animal pourraient permettre de conclure de manière plus robuste. Parmi les perspectives envisageables pour analyser de manière plus exploratoire les déterminants associés au pronostic des bactériémies, nous pourrions envisager de réaliser une étude d'association en prenant en compte directement les séquences nucléotidiques, et non plus un nombre fini de gènes de fonction connue. Dans cette optique, l'utilisation d'un logiciel comme Pyseer (Lees et al., 2018), prenant en compte la distribution des k-mers, est une solution qui semble particulièrement adaptée.

[7] https://www.ined.fr/fr/tout-savoir-population/chiffres/france/evolution-population/projections/

Par ces approches de génomique comparée à grande échelle, nous avons également vu qu'il est possible d'analyser finement la dynamique des populations de souches responsables de ces infections (Kallonen et al., 2017), et parfois même de dater l'émergence de certains clones présents à l'échelle mondiale ou à un niveau plus local (Ben Zakour et al., 2016; Kallonen et al., 2017; Ludden et al., 2020). De tels résultats sont essentiels dans la compréhension et l'identification des phénomènes qui s'exercent sur ces populations pour aboutir à de telles expansions clonales. Les mécanismes qui sous tendent ces émergences sont complexes et pourraient être variables dans le temps et dans l'espace, justifiant le suivi de ces infections localement au sein de populations relativement homogènes comme pour les études Colibafi et Septicoli (de Lastours et al., 2020; Lefort et al., 2011). En étudiant les souches de ces deux collections, nous avons pu constater une impressionnante stabilité de la population en termes de phylogroupes au cours du temps, suggérant une adaptation spécifique du phylogroupe B2, et dans une moindre mesure du D, aux conditions rencontrées notamment dans le milieu extraintestinal. Certains auteurs évoquent d'ailleurs les différences de capacités métaboliques pour expliquer les déséquilibres dans la diversité de la population observés dans certaines niches écologiques (Touchon et al., 2020). La reconstruction des réseaux métaboliques que nous avons réalisée à partir de plus de 1400 souches montre effectivement avant tout une association forte avec la phylogénie.

A une échelle plus fine, nous avons constaté la pauci-clonalité de cette population de souches issues de bactériémies avec seulement 11 STc responsables de plus de 70% des cas. Pour autant, la prévalence de ces clones subit parfois des modifications drastiques, certains clones multirésistants, comme le STc131, devenant largement dominants aux dépens d'autres sensibles aux antibiotiques comme le STc95. De plus, au sein même des cinq STc majoritaires, des remodelages de la population surviennent, parfois associés à l'acquisition indépendante des facteurs de virulence. C'est le cas de certaines souches des clones pandémiques STc131 et STc69 porteuses du gène de virulence *papGII*, localisé sur différents îlots de pathogénicité, soulignant au passage l'importance des éléments génétiques mobiles dans l'évolution des souches pathogènes (Hacker & Kaper, 2000). De futures analyses sont nécessaires pour tenter de dater plus précisément l'expansion de ces groupes de souches, en utilisant par exemple des approches bayésiennes comme celle de la méthode BEAST (Suchard et al., 2018). Cependant, notre collection de souches issues de bactériémies ne comprend que deux points peu espacés dans le temps et nécessitera donc probablement l'intégration de données supplémentaires pour obtenir des résultats robustes à partir de ces analyses.

L'identification d'acquisitions indépendantes du facteur de virulence *papGII* via différents éléments mobiles, nous permet également de mieux comprendre une partie de la physiopathologie des bactériémies à *E. coli*. En effet, le rôle de ce fimbriae dans les pyélonéphrites a été démontré *in vitro* par le passé (Lane & Mobley, 2007) et se traduit par une prévalence élevée dans les bactériémies à porte d'entrée urinaire dans Septicoli (de Lastours et al., 2020)**.** Nos résultats, ainsi que ceux récemment publiés par Biggel *et al.* qui montrent une association forte entre *papGII* et le caractère invasif des souches pathogènes urinaires (Biggel et al., 2020), confirment l'avantage sélectif qu'offre la présence d'un tel déterminant à certaines souches ExPEC pour la colonisation d'une niche particulière, l'arbre urinaire haut. Néanmoins, les données cliniques intégrées dans Septicoli montrent que la virulence des souches et la gravité des bactériémies sont souvent découplées chez l'Homme (Landraud et al., 2013), *papGII* apparaissant comme un facteur de bon pronostic.

Les variations intra-STc constatées pourraient aussi résulter de phénomènes de sélection fréquence-dépendante, favorisant les phénotypes rares et expliquant la diversification en continue au sein de ces clones pandémiques majeurs (McNally et al., 2019). Un échantillonnage régulier de la population dans les années à venir serait donc indispensable pour mieux comprendre les évolutions parfois rapides de ces souches pathogènes et en tirer les informations les plus pertinentes dans la pratique clinique quotidienne.

L'une des questions qui reste en suspens est le rôle de la résistance aux antibiotiques dans ces phénomènes épidémiologiques. Certains auteurs ne la considèrent pas réellement responsable des modifications récentes (Kallonen et al., 2017). Pourtant, l'un des points communs des clones nouvellement émergents est bien cette résistance à des antibiotiques d'intérêt clinique largement utilisés, qu'il s'agisse du STc131, du STc69 ou plus récemment du STc1193 (Manges et al., 2001; Nicolas-Chanoine et al., 2014; Tchesnokova, Rechkina, et al., 2019). Dans la collection Septicoli, nous avons même constaté un groupe de souches porteuses des gènes *bla*CTX-M-27 et de mutations de résistance aux fluoroquinolones dans le clade A du STc131, habituellement plutôt sensible aux antibiotiques. Cette émergence mime en partie celle du clade C2 aujourd'hui largement dominant et fait donc redouter une issue identique à terme. La sélection de gènes de virulence pour la colonisation du tube digestif et/ou des sites extraintestinaux humains s'est exercée sur une échelle de temps autrement plus importante que celle correspondant au début de l'utilisation massive des antibiotiques en clinique (Beceiro et al., 2013). Cela pourrait, par exemple, expliquer la stabilité quasi-parfaite du STc73 dans nos deux collections du fait d'une longue coévolution avec l'hôte. Quoi qu'il en soit, dans ces conditions il est complexe d'estimer la part jouée par chacun de ces phénomènes, et si pendant longtemps virulence et résistance ont été considérées

comme deux phénomènes mutuellement exclusifs, ce dogme ne s'applique pas à ces clones récemment introduits dans la population.

Enfin, au cours de cette thèse, nous avons souhaité étendre l'analyse des génomes en explorant le lien potentiel entre métabolisme et mode de vie ou phylogénie. Les facteurs de virulence classiques montrent un intérêt certain dans la compréhension de la physiopathologie des infections extraintestinales, mais ne permettent pas toujours d'expliquer l'adaptation à une niche écologique particulière. Comme le suggère la théorie de Freter (Freter et al., 1983), l'utilisation de substrats particuliers ou bien une capacité plus importante d'utilisation des nutriments disponibles comparé aux populations résidentes peut être un avantage majeur pour la survie et l'implantation des souches ExPEC. C'est, par exemple, le cas des souches UPEC dont certaines peuvent utiliser des acides aminés présents dans l'urine, telle la sérine, comme source de carbone (Alteri & Mobley, 2015; Anfora et al., 2007). Un processus identique a probablement lieu pour la colonisation du tube digestif, comme le montre le lien entre abondance au niveau digestif et infection urinaire (Magruder et al., 2019). En analysant les génomes de *E. coli* et de *Escherichia* clades de quatre collections, contenant à la fois des souches fécales commensales et des souches d'infections extraintestinales, nous n'avons pas pu mettre en évidence de voies métaboliques spécifiquement associées à un mode de vie, en raison d'une forte empreinte de la phylogénie. Néanmoins, pour compléter ce type d'analyse, il faudrait pouvoir prendre en compte les phénomènes de régulation et d'épistasie dont le rôle est parfois prépondérant (Rousset et al., 2021; Sabarly et al., 2011b). Cela requiert la réalisation d'études expérimentales supplémentaires, comme, par exemple, l'utilisation de la technologie Biolog pour évaluer des phénotypes de croissance sur un large panel de substrats et la transcriptomique pour repérer des opérons métaboliques différentiellement exprimés entre les souches.

Malgré tout, en nous concentrant sur les principaux clones responsables de bactériémies (Denamur et al., 2021), nous avons identifié deux voies métaboliques d'intérêt, codées par le groupe de gènes *mhp*. Ces deux voies permettent la dégradation de composés aromatiques provenant de la dégradation de la lignine (Burlingame & Chapman, 1983) et sont absentes des souches de phylogroupes B2 à l'exception de trois groupes de souches non reliées phylogénétiquement : le STc131, le STc452 et les ST136/681. Cette distribution particulière au sein du phylogroupe B2 pourrait être le signe d'une forte pression sélection pour leur acquisition. Si tel est le cas, de futurs travaux sont nécessaires pour identifier l'avantage sélectif lié à la présence de ces voies. Le caractère commensal strictement humain du STc452 peut faire évoquer un bénéfice en termes de colonisation intestinale, mais cela reste purement hypothétique. Pour répondre à une telle interrogation, nous pourrions réaliser des expériences

de compétition de croissance, entre des souches de différents fonds génétiques possédant ou non les gènes *mhp*, dans un milieu minimum supplémenté des substrats de ces voies. En parallèle, une autre approche est l'évaluation des capacités de colonisation intestinale dans un modèle animal. Pour cela nous pourrions comparer le pouvoir colonisateur de ces souches délétées ou non des gènes d'intérêt, et également en compétition avec des souches commensales typiques et des souches ExPEC. Cependant, bien qu'une telle approche est utile pour évaluer la compétitivité d'une souche de *E. coli* vis-à-vis d'une autre, elle ne prend pas en compte l'impact du microbiote digestif dans sa globalité. Pour cela, des modèles reproduisant plus précisément les communautés microbiennes intestinales humaines (Auchtung et al., 2015) pourraient permettre d'intégrer un niveau de complexité supérieur, et ainsi des résultats plus fidèle à la réalité. Enfin, l'abondance des substrats de cette voie dans les végétaux (Díaz et al., 2001) fait suspecter un rôle de l'alimentation qui nécessite lui aussi d'être analysé plus en détail.

Travaux annexes

Au cours de ces 4 années de thèse, j'ai pu me former à l'analyse des génomes bactériens. Bien que les bactériémies à *E. coli* soient au centre de cette thèse, j'ai eu l'occasion de travailler sur d'autres aspects de cette bactérie et même d'autres microorganismes d'intérêt médical. Les publications en premier ou deuxième auteur liées à ces études et parues au cours de ma thèse sont listées ci-après. Les publications en premier co-auteur sont indiquées par un astérisque.

1. M. Desroches, G. Royer*, D. Roche, M. Mercier-Darty, D. Vallenet, C. Médigue, K. Bastard, C. Rodriguez, O. Clermont, E. Denamur, J.-W. Decousser. The odyssey of the ancestral Escherich strain through culture collections: an example of allopatric diversification. *mSphere*. 2018 Jan 31;3(1).

Ce travail a porté sur l'analyse des génomes de la souche ancestrale de *E. coli* provenant de quatre collections (anglaise, française, américaine et allemande). Par des comparaisons à différents niveaux de granularité (MLST, recherche de SNP, coregenome MLST), nous avons mis en évidence des différences majeures entre ces quatre isolats. Ces différences s'expliquent par la présence de mutations dans les systèmes de réparation de l'ADN responsables d'un phénotype mutateur et associé à des profils mutationnels typiques. De plus, l'accumulation de ces mutations est corrélée avec l'histoire de cette souche et l'envoi au sein des différentes collections, évoquant une diversification allopatrique. Enfin, les différents isolats présentent des profils d'hypersensibilité inhabituels à certains antibiotiques comme la pénicilline G en lien avec la présence de mutations dans certaines porines. Ces mutations semblent avoir été sélectionnées au cours de la conservation de la souche et permettent aux isolats une adaptation de leur balance entre préservation de la cellule et capacités nutritionnelles accrues.

2. A.-S. Bourrel, L. Poirel, G. Royer*, M. Darty, X. Vuillemin, N. Kieffer, O. Clermont, E. Denamur, P. Nordmann, J.-W. Decousser, IAME Resistance Group. Colistin resistance in Parisian inpatient faecal *Escherichia coli* as the result of two distinct evolutionary pathways. *J Antimicrob Chemother*. 2019 Jun 1;74(6):1521-1530.

Au cours de ce travail, nous avons évalué le portage rectal de souches de *E. coli* résistantes à la colistine en région parisienne. Le séquençage complet des souches a permis leur typage (MLST, phylogroupe) ainsi que la recherche de déterminants de résistance. Les gènes plasmidiques *mcr* sont rarement retrouvés, appartiennent au variant *mcr-1* et sont localisés sur les plasmides classiquement décrits. Le fond génétique particulier de ces souches évoque une probable origine animale. A l'inverse, les souches négatives pour *mcr* présentent pour la plupart des résistances chromosomiques par mutations dans les gènes *prmA/pmrB* impliqués dans la modification du lipide A. L'épidémiologie des ces souches porteuses de mutations chromosomiques miment celle des populations commensales fécales et cliniques observées en région parisienne et évoque donc une pression de sélection différente encore non identifiée. Par ailleurs, les niveaux de résistance, en termes de concentration minimale inhibitrice de la colistine, apparaissent plus faibles dans le cas des résistances par acquisition de gènes plasmidiques. Ma participation au cours de cette étude a consisté à analyser l'ensemble des génomes de *E. coli* résistants à la colistine préalablement isolés des prélèvements fécaux.

3. N. Kieffer, G. Royer, J.-W. Decousser, A.-S. Bourrel, M. Palmieri, J.-M. Ortiz De La Rosa, H. Jacquier, E. Denamur, P. Nordmann, L. Poirel. *mcr-9*, an inducible gene encoding an acquired phosphoethanolamine transferase in *Escherichia coli*, and its origin. *Antimicrob Agents Chemother*. 2019 Jun 17. pii: AAC.00965-19.

Cette étude décrit un nouveau variant du gène *mcr* isolé au cours de l'étude précédente. Ce gène codant une phosphoéthanolamine transférase présente une faible identité avec les variants précédemment décrits, et une identité plus importante avec des gènes chromosomiques retrouvés dans le genre *Buttiauxella.* Le niveau de résistance procuré par ce gène après insertion dans E. coli K-12 est faible, mais peut-être induit en présence de concentration sub-inhibitrice de colistine grâce à un système à deux composants codé par les gènes *qseC/qseB.* Ma participation à ce travail a consisté à analyser la souche porteuse de ce variant, et rechercher d'autres souches dans les bases de données afin d'analyser la distribution de ce gène et son contexte génétique.

4. G. Royer, E. Melloul, L. Roisin, V. Courbin, H. Jacquier, R. Lepeule, L. Coutte, M. Darty, V. Fihman, P. Lim, J.-W. Decousser, C. Rodriguez, P.L. Woerther. Complete genome sequencing of *Enterococcus faecalis* strains suggests role of Ebp deletion in

infective endocarditis relapse. *Clin Microbiol Infect*. 2019 Jul 12. pii: S1198-743X(19)30395-7.

Au cours de ce travail nous avons analysé des souches de *E. faecalis* responsables d'endocardite infectieuse. Nous disposions de l'isolat initialement obtenu à partir de la première hémoculture et ayant permis de poser le diagnostic d'endocardite chez un patient. D'autre part, nous avions également un second isolat provenant de la valve cardiaque, dont l'ablation cinq mois après le diagnostic a été jugée indispensable en raison d'une réapparition des signes cliniques, et ce malgré un traitement antibiotique bien conduit. La comparaison de deux isolats nous a permis d'objectiver une rechute de l'infection. Cependant par l'analyse détaillée des ces isolats nous avons identifié une large délétion de plus de 47 kpb dans le second isolat. Ce segment délété contient les gènes *ebpABC et srtC*, codant des facteurs de virulence reconnus dans l'endocardite, notamment pour la formation de biofilm, et sont des cibles du système immunitaire de l'hôte. Nous avons pu en parallèle observer une diminution des capacités de croissance et de formation de biofilm dans la souche présentant la délétion. Ces résultats suggèrent un possible mécanisme d'échappement de la bactérie vis-à-vis du système immunitaire par la perte de ces déterminants, associé néanmoins à un coût important pour la bactérie. Au cours de cette étude mon rôle a été d'analyser les génomes et de les comparer à la recherche de variations génétiques (mutations, insertions/délétions).

5. G. Royer, F. Fourreau, B. Boulanger, M. Darty, D. Ducellier, F. Cizeau, A. Potron, I. Podglajen, N. Mongardon, J.-W. Decousser. Local outbreak of extended spectrum beta lactamase SHV2a producing *Pseudomonas aeruginosa* reveals the emergence of a new specific sub-lineage of the international ST235 high-risk clone. *J Hosp Infect*. 2019 Jul 29. pii: S0195-6701(19)30308-1.

Au cours de cette étude nous avons décrits une épidémie d'infections causées par une souche de *P. aeruginosa* BLSE à l'hôpital Henri Mondor (AP-HP, Créteil). L'analyse des génomes des quatre isolats a permis d'identifier le gène codant la betalactamase *bla*SHV-2a, rarement observée dans cette espèce. La comparaison des souches par coregénome MLST a montré la proximité de ces souches de *P. aeruginosa* appartenant au ST235. Une analyse phylogénétique comprenant l'ensemble des génomes de ST montre des acquisitions fréquentes de différentes betalactamases par des souches proches génétiquement suggérant le caractère épidémique à haut risque de ce ST. Par ailleurs, nous avons pu identifier quatre autres souches porteuses du gène *bla*SHV-2a au sein de ce ST, formant un groupe avec nos quatre souches épidémiques et constituant probablement un sous-lignée de ce clone

épidémique. Ma participation dans ce travail a consisté à caractériser et comparer l'ensemble des génomes.

6. G. Royer, F. Fourreau, C. Gomart, A. Maurand, B. Hacquin, D. Ducellier, F. Cizeau, S. Lo, C. Cordonnier-Jourdin, M. Mercier-Darty, J.-W. Decousser. Outbreak of an uncommon rifampin-resistant *bla*NDM-1 *Citrobacter amalonaticus* strain in a digestive rehabilitation center: the putative role of rifaximin. *Clin Infect Dis.* 2019 Dec 7.

Nous avons décrit une épidémie de souches de *Citrobacter amalonaticus bla*NDM-1 dans un service de rééducation digestive. Le séquençage et la comparaison des isolats ont confirmé qu'il s'agissait bien de la même souche. Par ailleurs l'analyse du résistome a mis en évidence un gène impliqué dans la résistance à la rifampicine, *arr*-3, rarement retrouvé chez les entérobactéries. Les cinq patients colonisés par cette souche multirésistante étaient traités au long cours par rifaximine en prévention de l'encéphalopathie hépatique, évoquant une possible implication de cette antibiotique dans la sélection de cette bactérie. J'ai réalisé l'analyse des génomes de ces cinq isolats.

7. M. Mercier-Darty, G. Royer*, B. Lamy, C. Charron, O. Lemenand, C. Gomart, F. Fourreau, J.-Y. Madec, E. Jumas-Bilak, J.-W. Decousser, the RESAPATH Network, the ColBVH Network. Comparative whole genome phylogeny of animal, environmental, and human strains confirms the genogroup organization and diversity of the *Stenotrophomonas maltophilia* complex. *Appl Environ Microbiol.* 2020 Mar 20.

Nous avons étudié des souches de Stenotrophomonas maltophilia isolées d'humains et d'animaux non humains. L'analyse de leur génome combinés à ceux disponibles dans les bases de données a montré une population organisée sous forme d'un complexe plutôt qu'une espèce, comme le montre les résultats des comparaisons de l'identité nucléotidique moyenne. Combiné à une analyse phylogénétique nous avons pu observer l'organisation en génogroupes, dont certains sont particulièrement associés à l'homme et d'autres aux animaux non-humains. La comparaison du contenu en gènes en fonction de l'hôte montre un enrichissement des souches animales dans certains gènes de résistance. Peu de gènes sont associés spécifiquement aux souches humaines hormis un gène potentiellement impliqué dans le métabolisme du tréhalose. Des analyses supplémentaires sont nécessaires pour identifier un éventuel rôle dans l'adaptation à une niche écologique particulière. J'ai réalisé l'analyse des génomes au cours de cette étude.

Références bibliographiques

Abernethy, J., Guy, R., Sheridan, E. A., Hopkins, S., Kiernan, M., Wilcox, M. H., Johnson, A. P., Hope, R., Sen, R. A., Mifsud, A., O'Driscoll, J., Brown, N., Trundle, C., Allison, D., Twagira, M., Gnanarajah, Awad-El Kariem, F., Rajendran, R., Umashankar, S., … Pasztor, M. (2017). Epidemiology of *Escherichia coli* bacteraemia in England: results of an enhanced sentinel surveillance programme. *Journal of Hospital Infection*, *95*(4), 365-375. https://doi.org/10.1016/j.jhin.2016.12.008

Abernethy, J. K., Johnson, A. P., Guy, R., Hinton, N., Sheridan, E. A., & Hope, R. J. (2015). Thirty day all-cause mortality in patients with *Escherichia coli* bacteraemia in England. *Clinical Microbiology and Infection*, *21*(3), 251.e1-251.e8. https://doi.org/10.1016/j.cmi.2015.01.001

Abram, K., Udaondo, Z., Bleker, C., Wanchai, V., Wassenaar, T. M., Robeson, M. S., & Ussery, D. W. (2019). *What can we learn from over 100,000* Escherichia coli *genomes?* [Preprint]. Genomics. https://doi.org/10.1101/708131

Adiba, S., Nizak, C., van Baalen, M., Denamur, E., & Depaulis, F. (2010). From grazing resistance to pathogenesis: The coincidental evolution of virulence factors. *PLoS ONE*, *5*(8), e11882. https://doi.org/10.1371/journal.pone.0011882

Alm, E. W., Walk, S. T., & Gordon, D. M. (2014). The niche of *Escherichia coli*. In S. T. Walk & P. C. H. Feng (Éds.), *Population Genetics of Bacteria* (p. 67-89). ASM Press. https://doi.org/10.1128/9781555817114.ch6

Alteri, C. J., & Mobley, H. L. T. (2015). Metabolism and Fitness of Urinary Tract Pathogens. *Microbiology Spectrum*, *3*(3). https://doi.org/10.1128/microbiolspec.MBP-0016-2015

Anfora, A. T., Haugen, B. J., Roesch, P., Redford, P., & Welch, R. A. (2007). Roles of serine accumulation and catabolism in the colonization of the murine urinary tract by *Escherichia coli* CFT073. *Infection and Immunity*, *75*(11), 5298-5304. https://doi.org/10.1128/IAI.00652-07

Antão, E.-M., Wieler, L. H., & Ewers, C. (2009). Adhesive threads of extraintestinal pathogenic *Escherichia coli*. *Gut Pathogens*, *1*(1), 22. https://doi.org/10.1186/1757-4749-1-22

Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., & Pevzner, P. A. (2016). plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*, btw493. https://doi.org/10.1093/bioinformatics/btw493

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics (Oxford, England)*, *36*(7), 2251-2252. https://doi.org/10.1093/bioinformatics/btz859

Arredondo-Alonso, S., Willems, R. J., van Schaik, W., & Schürch, A. C. (2017). On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics*, *3*(10). https://doi.org/10.1099/mgen.0.000128

Auchtung, J. M., Robinson, C. D., & Britton, R. A. (2015). Cultivation of stable, reproducible microbial communities from different fecal donors using minibioreactor arrays (MBRAs). *Microbiome*, *3*, 42. https://doi.org/10.1186/s40168-015-0106-5

Bachmann, B. J. (1996). Derivations and genotypes of some mutant derivatives of *Escherichia coli* K-12. *Escherichia coli and Salmonella: cellular and molecular biology, 2nd ed. ASM Press, Washington, DC*, 2460-2488.

Bazin, A., Gautreau, G., Médigue, C., Vallenet, D., & Calteau, A. (2020). panRGP: a pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics*, *36*(Supplement_2), i651-i658. https://doi.org/10.1093/bioinformatics/btaa792

Beceiro, A., Tomas, M., & Bou, G. (2013). Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clinical Microbiology Reviews*, *26*(2), 185-230. https://doi.org/10.1128/CMR.00059-12

Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E., & Clermont, O. (2018). ClermonTyping: an easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microbial Genomics*, *4*(7). https://doi.org/10.1099/mgen.0.000192

Ben Zakour, N. L., Alsheikh-Hussain, A. S., Ashcroft, M. M., Khanh Nhu, N. T., Roberts, L. W., Stanton-Cook, M., Schembri, M. A., & Beatson, S. A. (2016). Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. *MBio*, *7*(2), e00347-00316. https://doi.org/10.1128/mBio.00347-16

Bergthorsson, U., & Ochman, H. (1998). Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Molecular Biology and Evolution*, *15*(1), 6-16. https://doi.org/10.1093/oxfordjournals.molbev.a025847

Bert, F., Huynh, B., Dondero, F., Johnson, J. R., Paugam-Burtz, C., Durand, F., Belghiti, J., Valla, D., Moreau, R., & Nicolas-Chanoine, M.-H. (2011). Molecular epidemiology of *Escherichia coli* bacteremia in liver transplant recipients. *Transplant Infectious Disease*, *13*(4), 359-365. https://doi.org/10.1111/j.1399-3062.2011.00618.x

Bettelheim, K. A. (1978). The sources of 'OH' serotypes of *Escherichia coli*. *Journal of Hygiene*, *80*(1), 83-113. https://doi.org/10.1017/S0022172400053420

Bhattacharya, A., Nsonwu, O., Johnson, A. P., & Hope, R. (2018). Estimating the incidence and 30-day all-cause mortality rate of *Escherichia coli* bacteraemia in England by 2020/21. *Journal of Hospital Infection*, *98*(3), 228-231. https://doi.org/10.1016/j.jhin.2017.09.021

Biggel, M., Xavier, B. B., Johnson, J. R., Nielsen, K. L., Frimodt-Møller, N., Matheeussen, V., Goossens, H., Moons, P., & Van Puyvelde, S. (2020). Horizontally acquired *papGII-*

containing pathogenicity islands underlie the emergence of invasive uropathogenic *Escherichia coli* lineages. *Nature Communications*, *11*(1), 5968. https://doi.org/10.1038/s41467-020-19714-9

Billard-Pomares, T., Clermont, O., Castellanos, M., Magdoud, F., Royer, G., Condamine, B., Fouteau, S., Barbe, V., Roche, D., Cruveiller, S., Médigue, C., Pognard, D., Glodt, J., Dion, S., Rigal, O., Picard, B., Denamur, E., & Branger, C. (2019). The arginine deiminase operon is responsible for a fitness trade-off in extended-spectrum-β-lactamase-producing strains of *Escherichia coli*. *Antimicrobial Agents and Chemotherapy*, *63*(8). https://doi.org/10.1128/AAC.00635-19

Binns, M. M., Davies, D. L., & Hardy, K. G. (1979). Cloned fragments of the plasmid ColV,I-K94 specifying virulence and serum resistance. *Nature*, *279*(5716), 778-781. https://doi.org/10.1038/279778a0

Binns, M. M., Mayden, J., & Levine, R. P. (1982). Further characterization of complement resistance conferred on *Escherichia coli* by the plasmid genes *traT* of R100 and *iss* of ColV,I-K94. *Infection and Immunity*, *35*(2), 654-659. https://doi.org/10.1128/IAI.35.2.654-659.1982

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., & Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science (New York, N.Y.)*, *277*(5331), 1453-1462. https://doi.org/10.1126/science.277.5331.1453

Blum, G., Ott, M., Lischewski, A., Ritter, A., Imrich, H., Tschäpe, H., & Hacker, J. (1994). Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. *Infection and Immunity*, *62*(2), 606-614. https://doi.org/10.1128/IAI.62.2.606-614.1994

Bok, E., Mazurek, J., Myc, A., Stosik, M., Wojciech, M., & Baldy-Chudzik, K. (2018). Comparison of commensal *Escherichia coli* isolates from adults and young children in Lubuskie province, Poland: virulence potential, phylogeny and antimicrobial resistance. *International Journal of Environmental Research and Public Health*, *15*(4), 617. https://doi.org/10.3390/ijerph15040617

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M. H., & Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, *101*(6), 1644-1655. https://doi.org/10.1378/chest.101.6.1644

Boquet, P. (2001). The cytotoxic necrotizing factor 1 (CNF1) from *Escherichia coli*. *Toxicon*, *39*(11), 1673-1680. https://doi.org/10.1016/S0041-0101(01)00154-4

Bourrel, A. S., Poirel, L., Royer, G., Darty, M., Vuillemin, X., Kieffer, N., Clermont, O., Denamur, E., Nordmann, P., Decousser, J.-W., IAME Resistance Group, Lafaurie, M., Bercot, B., Walewski, V., Lescat, M., Carbonnelle, E., Ousser, F., Idri, N., Ricard, J.-D., … Gomart, C. (2019). Colistin resistance in Parisian inpatient faecal *Escherichia*

coli as the result of two distinct evolutionary pathways. *Journal of Antimicrobial Chemotherapy*, *74*(6), 1521-1530. https://doi.org/10.1093/jac/dkz090

Bouza, E., Sousa, D., Rodriguez-Creixems, M., Lechuz, J. G., & Munoz, P. (2007). Is the volume of blood cultured still a significant factor in the diagnosis of bloodstream infections? *Journal of Clinical Microbiology*, *45*(9), 2765-2769. https://doi.org/10.1128/JCM.00140-07

Branger, C., Ledda, A., Billard-Pomares, T., Doublet, B., Barbe, V., Roche, D., Médigue, C., Arlet, G., & Denamur, E. (2019). Specialization of small non-conjugative plasmids in *Escherichia coli* according to their family types. *Microbial Genomics*, *5*(9). https://doi.org/10.1099/mgen.0.000281

Branger, C., Ledda, A., Billard-Pomares, T., Doublet, B., Fouteau, S., Barbe, V., Roche, D., Cruveiller, S., Médigue, C., Castellanos, M., Decré, D., Drieux-Rouze, L., Clermont, O., Glodt, J., Tenaillon, O., Cloeckaert, A., Arlet, G., & Denamur, E. (2018). Extended-spectrum β-lactamase-encoding genes are spreading on a wide range of *Escherichia coli* plasmids existing prior to the use of third-generation cephalosporins. *Microbial Genomics*, *4*(9). https://doi.org/10.1099/mgen.0.000203

Brisse, S., Diancourt, L., Laouenan, C., Vigan, M., Caro, V., Arlet, G., Drieux, L., Leflon-Guibout, V., Mentre, F., Jarlier, V., Nicolas-Chanoine, M.-H., & the Coli ss Study Group. (2012). Phylogenetic distribution of CTX-M- and non-extended-spectrum-β-lactamase-producing *Escherichia coli* isolates: group B2 isolates, except clone ST131, rarely produce CTX-M-enzymes. *Journal of Clinical Microbiology*, *50*(9), 2974-2981. https://doi.org/10.1128/JCM.00919-12

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59-60. https://doi.org/10.1038/nmeth.3176

Burlingame, R., & Chapman, P. J. (1983). Catabolism of phenylpropionic acid and its 3-hydroxy derivative by *Escherichia coli*. *Journal of Bacteriology*, *155*(1), 113-121. https://doi.org/10.1128/JB.155.1.113-121.1983

Carattoli, A. (2009). Resistance plasmid families in *Enterobacteriaceae*. *Antimicrobial Agents and Chemotherapy*, *53*(6), 2227-2238. https://doi.org/10.1128/AAC.01707-08

Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., Møller Aarestrup, F., & Hasman, H. (2014). *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*, *58*(7), 3895-3903. https://doi.org/10.1128/AAC.02412-14

Carbonetti, N. H., Boonchai, S., Parry, S. H., Väisänen-Rhen, V., Korhonen, T. K., & Williams, P. H. (1986). Aerobactin-mediated iron uptake by *Escherichia coli* isolates from human extraintestinal infections. *Infection and Immunity*, *51*(3), 966-968. https://doi.org/10.1128/IAI.51.3.966-968.1986

Chang, D.-E., Smalley, D. J., Tucker, D. L., Leatham, M. P., Norris, W. E., Stevenson, S. J., Anderson, A. B., Grissom, J. E., Laux, D. C., Cohen, P. S., & Conway, T. (2004).

Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proceedings of the National Academy of Sciences*, *101*(19), 7427-7432. https://doi.org/10.1073/pnas.0307888101

Chen, L., Zheng, D., Liu, B., Yang, J., & Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. *Nucleic Acids Research*, *44*(D1), D694-D697. https://doi.org/10.1093/nar/gkv1239

Chouikha, I., Germon, P., Brée, A., Gilot, P., Moulin-Schouleur, M., & Schouler, C. (2006). A *selC* -associated genomic isIsland of the extraintestinal avian pathogenic *Escherichia coli* strain BEN2908 is involved in carbohydrate uptake and virulence. *Journal of Bacteriology*, *188*(3), 977-987. https://doi.org/10.1128/JB.188.3.977-987.2006

Chuba, P. J., Leon, M. A., Banerjee, A., & Palchaudhuri, S. (1989). Cloning and DNA sequence of plasmid determinant *iss*, coding for increased serum survival and surface exclusion, which has homology with lambda DNA. *Molecular and General Genetics MGG*, *216*(2-3), 287-292. https://doi.org/10.1007/BF00334367

Chung, H.-C., Lai, C.-H., Lin, J.-N., Huang, C.-K., Liang, S.-H., Chen, W.-F., Shih, Y.-C., Lin, H.-H., & Wang, J.-L. (2012). Bacteremia caused by extended-spectrum-β-lactamase-producing *Escherichia coli* sequence type ST131 and non-ST131 clones: comparison of demographic data, clinical features, and mortality. *Antimicrobial Agents and Chemotherapy*, *56*(2), 618-622. https://doi.org/10.1128/AAC.05753-11

Clermont, O., Bonacorsi, S., & Bingen, E. (2000). Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Applied and Environmental Microbiology*, *66*(10), 4555-4558. https://doi.org/10.1128/AEM.66.10.4555-4558.2000

Clermont, O., Christenson, J. K., Daubié, A.-S., Gordon, D. M., & Denamur, E. (2014). Development of an allele-specific PCR for *Escherichia coli* B2 sub-typing, a rapid and easy to perform substitute of multilocus sequence typing. *Journal of Microbiological Methods*, *101*, 24-27. https://doi.org/10.1016/j.mimet.2014.03.008

Clermont, O., Christenson, J. K., Denamur, E., & Gordon, D. M. (2013). The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups: A new *E. coli* phylo-typing method. *Environmental Microbiology Reports*, *5*(1), 58-65. https://doi.org/10.1111/1758-2229.12019

Clermont, O., Couffignal, C., Blanco, J., Mentré, F., Picard, B., Denamur, E., & the COLIVILLE and COLIBAFI groups. (2017). Two levels of specialization in bacteraemic *Escherichia coli* strains revealed by their comparison with commensal strains. *Epidemiology and Infection*, *145*(5), 872-882. https://doi.org/10.1017/S0950268816003010

Clermont, O., Gordon, D., & Denamur, E. (2015). Guide to the various phylogenetic classification schemes for *Escherichia coli* and the correspondence among schemes. *Microbiology*, *161*(5), 980-988. https://doi.org/10.1099/mic.0.000063

Clermont, O., Gordon, D. M., Brisse, S., Walk, S. T., & Denamur, E. (2011). Characterization of the cryptic *Escherichia* lineages: rapid identification and prevalence. *Environmental Microbiology*, *13*(9), 2468-2477. https://doi.org/10.1111/j.1462-2920.2011.02519.x

Clermont, O., Lescat, M., O'Brien, C. L., Gordon, D. M., Tenaillon, O., & Denamur, E. (2008). Evidence for a human-specific *Escherichia coli* clone. *Environmental Microbiology*, *10*(4), 1000-1006. https://doi.org/10.1111/j.1462-2920.2007.01520.x

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422-1423. https://doi.org/10.1093/bioinformatics/btp163

Collège des Universitaires des Maladies Infectieuses et Tropicales (CMIT). (2018). *ECN PILLY: maladies infectieuses et et tropicales.* MED-LINE EDITIONS - EDUC.

Collins, R. E., & Higgs, P. G. (2012). Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Molecular Biology and Evolution*, *29*(11), 3413-3425. https://doi.org/10.1093/molbev/mss163

Croxen, M. A., Law, R. J., Scholz, R., Keeney, K. M., Wlodarska, M., & Finlay, B. B. (2013). Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clinical Microbiology Reviews*, *26*(4), 822-880. https://doi.org/10.1128/CMR.00022-13

Cruickshank, J. C. (1936). Modern Methods in Agglutination. *Proceedings of the Royal Society of Medicine*, *29*(7), 841-854. https://doi.org/10.1177/003591573602900736

Daegelen, P., Studier, F. W., Lenski, R. E., Cure, S., & Kim, J. F. (2009). Tracing ancestors and relatives of *Escherichia coli* B, and the derivation of B strains REL606 and BL21(DE3). *Journal of Molecular Biology*, *394*(4), 634-643. https://doi.org/10.1016/j.jmb.2009.09.022

Daga, A. P., Koga, V. L., Soncini, J. G. M., de Matos, C. M., Perugini, M. R. E., Pelisson, M., Kobayashi, R. K. T., & Vespero, E. C. (2019). *Escherichia coli* bloodstream infections in patients at a university hospital: virulence factors and clinical characteristics. *Frontiers in Cellular and Infection Microbiology*, *9*. https://doi.org/10.3389/fcimb.2019.00191

Dale, A. P., & Woodford, N. (2015). Extra-intestinal pathogenic *Escherichia coli* (ExPEC): disease, carriage and clones. *Journal of Infection*, *71*(6), 615-626. https://doi.org/10.1016/j.jinf.2015.09.009

de Lastours, V., Laouénan, C., Royer, G., Carbonnelle, E., Lepeule, R., Esposito-Farèse, M., Clermont, O., Duval, X., Fantin, B., Mentré, F., Decousser, J. W., Denamur, E., & Lefort, A. (2020). Mortality in *Escherichia coli* bloodstream infections: antibiotic resistance still does not make it. *Journal of Antimicrobial Chemotherapy*, *75*(8), 2334-2343. https://doi.org/10.1093/jac/dkaa161

de Toro, M., Garcilláon-Barcia, M. P., & De La Cruz, F. (2014). Plasmid diversity and adaptation analyzed by massive sequencing of *Escherichia coli* plasmids. *Microbiology Spectrum*, *2*(6). https://doi.org/10.1128/microbiolspec.PLAS-0031-2014

Denamur, E., Clermont, O., Bonacorsi, S., & Gordon, D. (2021). The population genetics of pathogenic *Escherichia coli*. *Nature Reviews Microbiology*, *19*(1), 37-54. https://doi.org/10.1038/s41579-020-0416-x

Desjardins, P., Picard, B., Kaltenböck, B., Elion, J., & Denamurl, E. (1995). Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. *Journal of Molecular Evolution*, *41*(4), 440-448. https://doi.org/10.1007/BF00160315

Desloges, I., Taylor, J. A., Leclerc, J.-M., Brannon, J. R., Portt, A., Spencer, J. D., Dewar, K., Marczynski, G. T., Manges, A., Gruenheid, S., Le Moual, H., & Thomassin, J.-L. (2019). Identification and characterization of OmpT-like proteases in uropathogenic *Escherichia coli* clinical isolates. *MicrobiologyOpen*, *8*(11), e915. https://doi.org/10.1002/mbo3.915

Desroches, M., Royer, G., Roche, D., Mercier-Darty, M., Vallenet, D., Médigue, C., Bastard, K., Rodriguez, C., Clermont, O., Denamur, E., & Decousser, J.-W. (2018). The odyssey of the ancestral Escherich strain through culture collections: an example of allopatric diversification. *MSphere*, *3*(1). https://doi.org/10.1128/mSphere.00553-17

Di Lorenzo, M., & Stork, M. (2014). Plasmid-Encoded Iron Uptake Systems. *Microbiology Spectrum*, *2*(6). https://doi.org/10.1128/microbiolspec.PLAS-0030-2014

Diard, M., Baeriswyl, S., Clermont, O., Gouriou, S., Picard, B., Taddei, F., Denamur, E., & Matic, I. (2007). Caenorhabditis elegans as a simple model to study phenotypic and genetic virulence determinants of extraintestinal pathogenic *Escherichia coli*. *Microbes and Infection*, *9*(2), 214-223. https://doi.org/10.1016/j.micinf.2006.11.009

Diekema, D. J., Beekmann, S. E., Chapin, K. C., Morel, K. A., Munson, E., & Doern, G. V. (2003). Epidemiology and outcome of nosocomial and community-onset bloodstream infection. *Journal of Clinical Microbiology*, *41*(8), 3655-3660. https://doi.org/10.1128/JCM.41.8.3655-3660.2003

Díaz, E., Ferrández, A., Prieto, M. A., & García, J. L. (2001). Biodegradation of aromatic compounds by *Escherichia coli*. *Microbiology and Molecular Biology Reviews*, *65*(4), 523-569. https://doi.org/10.1128/MMBR.65.4.523-569.2001

Dobrindt, U., Blum-Oehler, G., Nagy, G., Schneider, G., Johann, A., Gottschalk, G., & Hacker, J. (2002). Genetic structure and distribution of four pathogenicity islands (PAI I536 to PAI IV536) of uropathogenic *Escherichia coli* strain 536. *Infection and Immunity*, *70*(11), 6365-6372. https://doi.org/10.1128/IAI.70.11.6365-6372.2002

Duprilot, M., Baron, A., Blanquart, F., Dion, S., Pouget, C., Lettéron, P., Flament-Simon, S.-C., Clermont, O., Denamur, E., & Nicolas-Chanoine, M.-H. (2020). Success of *Escherichia coli* O25b:H4 Sequence Type 131 clade C associated with a decrease in virulence. *Infection and Immunity*, *88*(12). https://doi.org/10.1128/IAI.00576-20

Duriez, P., Clermont, O., Bonacorsi, S., Bingen, E., Chaventré, A., Elion, J., Picard, B., & Denamur, E. (2001). Commensal *Escherichia coli* isolates are phylogenetically

distributed among geographically distinct human populations. *Microbiology*, *147*(6), 1671-1676. https://doi.org/10.1099/00221287-147-6-1671

Escherich, T. (1885). Die darmbakterien des neugeboreiien und säuglings. *DMW-Deutsche Medizinische Wochenschrift*, *11*(43), 740-741.

Escobar-Páramo, P., Grenet, K., Le Menac'h, A., Rode, L., Salgado, E., Amorin, C., Gouriou, S., Picard, B., Rahimy, M. C., Andremont, A., Denamur, E., & Ruimy, R. (2004). Large-scale population structure of human commensal *Escherichia coli* isolates. *Applied and Environmental Microbiology*, *70*(9), 5698-5700. https://doi.org/10.1128/AEM.70.9.5698-5700.2004

European Centre for Disease Prevention and Control. (2018). *External quality assessment of laboratory performance: European Antimicrobial Resistance Surveillance Network (EARSNet), 2017.* http://publications.europa.eu/publication/manifestation_identifier/PUB_TQ0418959ENN

Ewers, C., Li, G., Wilking, H., Kiebling, S., Alt, K., Antao, E., Laturnus, C., Diehl, I., Glodde, S., & Homeier, T. (2007). Avian pathogenic, uropathogenic, and newborn meningitis-causing *Escherichia coli*: How closely related are they? *International Journal of Medical Microbiology*, *297*(3), 163-176. https://doi.org/10.1016/j.ijmm.2007.01.003

Falkow, S. (2004). Molecular Koch's postulates applied to bacterial pathogenicity — a personal recollection 15 years later. *Nature Reviews Microbiology*, *2*(1), 67-72. https://doi.org/10.1038/nrmicro799

Frank, C., Werber, D., Cramer, J. P., Askar, M., Faber, M., an der Heiden, M., Bernard, H., Fruth, A., Prager, R., Spode, A., Wadl, M., Zoufaly, A., Jordan, S., Kemper, M. J., Follin, P., Müller, L., King, L. A., Rosner, B., Buchholz, U., … Krause, G. (2011). Epidemic profile of Shiga-toxin–producing *Escherichia coli* O104:H4 outbreak in Germany. *New England Journal of Medicine*, *365*(19), 1771-1780. https://doi.org/10.1056/NEJMoa1106483

Freter, R., Brickner, H., Botney, M., Cleven, D., & Aranki, A. (1983). Mechanisms that control bacterial populations in continuous-flow culture models of mouse large intestinal flora. *Infection and Immunity*, *39*(2), 676-685. https://doi.org/10.1128/IAI.39.2.676-685.1983

Friedmann, H. C. (2014). Escherich and *Escherichia*. *EcoSal Plus*, *6*(1). https://doi.org/10.1128/ecosalplus.ESP-0025-2013

Galardini, M., Clermont, O., Baron, A., Busby, B., Dion, S., Schubert, S., Beltrao, P., & Denamur, E. (2020). Major role of iron uptake systems in the intrinsic extra-intestinal virulence of the genus *Escherichia* revealed by a genome-wide association study. *PLOS Genetics*, *16*(10), e1009065. https://doi.org/10.1371/journal.pgen.1009065

Garcia, E. C., Brumbaugh, A. R., & Mobley, H. L. T. (2011). Redundancy and specificity of *Escherichia coli* iron acquisition systems during urinary tract infection. *Infection and Immunity*, *79*(3), 1225-1235. https://doi.org/10.1128/IAI.01222-10

Garénaux, A., Caza, M., & Dozois, C. M. (2011). The Ins and Outs of siderophore mediated iron uptake by extra-intestinal pathogenic *Escherichia coli*. *Veterinary Microbiology*, *153*(1-2), 89-98. https://doi.org/10.1016/j.vetmic.2011.05.023

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., Perrin, A., Médigue, C., Calteau, A., Cruveiller, S., Matias, C., Ambroise, C., Rocha, E. P. C., & Vallenet, D. (2020). PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph. *PLOS Computational Biology*, *16*(3), e1007732. https://doi.org/10.1371/journal.pcbi.1007732

Germon, P., Chen, Y.-H., He, L., Blanco, J. E., Brée, A., Schouler, C., Huang, S.-H., & Moulin-Schouleur, M. (2005). *ibeA*, a virulence factor of avian pathogenic *Escherichia coli*. *Microbiology*, *151*(4), 1179-1186. https://doi.org/10.1099/mic.0.27809-0

Germon, P., Roche, D., Melo, S., Mignon-Grasteau, S., Dobrindt, U., Hacker, J., Schouler, C., & Moulin-Schouleur, M. (2007). tDNA locus polymorphism and ecto-chromosomal DNA insertion hot-spots are related to the phylogenetic group of *Escherichia coli* strains. *Microbiology*, *153*(3), 826-837. https://doi.org/10.1099/mic.0.2006/001958-0

Girlich, D., Poirel, L., Carattoli, A., Kempf, I., Lartigue, M.-F., Bertini, A., & Nordmann, P. (2007). Extended-spectrum β-lactamase CTX-M-1 in *Escherichia coli* isolates from healthy poultry in France. *Applied and Environmental Microbiology*, *73*(14), 4681-4685. https://doi.org/10.1128/AEM.02491-06

Goluszko, P., Niesel, D., Nowicki, B., Selvarangan, R., Nowicki, S., Hart, A., Pawelczyk, E., Das, M., Urvil, P., & Hasan, R. (2001). Dr operon-associated invasiveness of *Escherichia coli* from pregnant patients with pyelonephritis. *Infection and Immunity*, *69*(7), 4678-4680. https://doi.org/10.1128/IAI.69.7.4678-4680.2001

Gordon, D. M., & Cowling, A. (2003). The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology*, *149*(12), 3575-3586. https://doi.org/10.1099/mic.0.26486-0

Gordon, D. M., Geyik, S., Clermont, O., O'Brien, C. L., Huang, S., Abayasekara, C., Rajesh, A., Kennedy, K., Collignon, P., Pavli, P., Rodriguez, C., Johnston, B. D., Johnson, J. R., Decousser, J.-W., & Denamur, E. (2017). Fine-scale structure analysis shows epidemic patterns of clonal complex 95, a cosmopolitan *Escherichia coli* lineage responsible for extraintestinal infection. *MSphere*, *2*(3). https://doi.org/10.1128/mSphere.00168-17

Goswami, C., Fox, S., Holden, M., Connor, M., Leanord, A., & Evans, T. J. (2018). Genetic analysis of invasive *Escherichia coli* in Scotland reveals determinants of healthcare-associated versus community-acquired infections. *Microbial Genomics*, *4*(6). https://doi.org/10.1099/mgen.0.000190

Goto, M., & Al-Hasan, M. N. (2013). Overall burden of bloodstream infection and nosocomial bloodstream infection in North America and Europe. *Clinical Microbiology and Infection*, *19*(6), 501-509. https://doi.org/10.1111/1469-0691.12195

Goullet, P., & Picard, B. (1986). Highly pathogenic strains of *Escherichia coli* revealed by the distinct electrophoretic patterns of carboxylesterase B. *Journal of General Microbiology*, *132*(7), 1853-1858. https://doi.org/10.1099/00221287-132-7-1853

Groisman, E. A., & Ochman, H. (1996). Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, *87*(5), 791-794. https://doi.org/10.1016/S0092-8674(00)81985-6

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1730-3

Guabiraba, R., & Schouler, C. (2015). Avian colibacillosis: still many black holes. *FEMS Microbiology Letters*, *362*(15), fnv118. https://doi.org/10.1093/femsle/fnv118

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075. https://doi.org/10.1093/bioinformatics/btt086

Guyer, D. M., Radulovic, S., Jones, F.-E., & Mobley, H. L. T. (2002). Sat, the secreted autotransporter toxin of uropathogenic *Escherichia coli*, is a vacuolating cytotoxin for bladder and kidney epithelial cells. *Infection and Immunity*, *70*(8), 4539-4546. https://doi.org/10.1128/IAI.70.8.4539-4546.2002

Hacker, J., & Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annual Review of Microbiology*, *54*(1), 641-679. https://doi.org/10.1146/annurev.micro.54.1.641

Hekker, T. A. M., Groeneveld, A. B. J., Simoons-Smit, A. M., de Man, P., Connell, H., & MacLaren, D. M. (2000). Role of bacterial virulence factors and host factors in the outcome of *Escherichia coli* bacteraemia. *European Journal of Clinical Microbiology & Infectious Diseases*, *19*(4), 312-316. https://doi.org/10.1007/s100960050483

Herzer, P. J., Inouye, S., Inouye, M., & Whittam, T. S. (1990). Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *Journal of Bacteriology*, *172*(11), 6175-6181. https://doi.org/10.1128/JB.172.11.6175-6181.1990

Hirai, I., Fukui, N., Taguchi, M., Yamauchi, K., Nakamura, T., Okano, S., & Yamamoto, Y. (2013). Detection of chromosomal *bla*CTX-M-15 in *Escherichia coli* O25b-B2-ST131 isolates from the Kinki region of Japan. *International Journal of Antimicrobial Agents*, *42*(6), 500-506. https://doi.org/10.1016/j.ijantimicag.2013.08.005

Huang, S., Wan, Z., Chen, Y., Jong, A. Y., & Kim, K. S. (2001). Further characterization of *Escherichia coli* brain microvascular endothelial cell invasion gene *ibeA* by deletion, complementation, and protein expression. *The Journal of Infectious Diseases*, *183*(7), 1071-1078. https://doi.org/10.1086/319290

Huseby, D. L., Pietsch, F., Brandis, G., Garoff, L., Tegehall, A., & Hughes, D. (2017). Mutation supply and relative fitness shape the genotypes of ciprofloxacin-resistant *Escherichia coli*. *Molecular Biology and Evolution*, msx052. https://doi.org/10.1093/molbev/msx052

Ingle, D. J., Clermont, O., Skurnik, D., Denamur, E., Walk, S. T., & Gordon, D. M. (2011). Biofilm formation by and thermal niche and virulence characteristics of *Escherichia* spp. *Applied and Environmental Microbiology*, *77*(8), 2695-2700. https://doi.org/10.1128/AEM.02401-10

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M. B., Pope, B. J., Tomita, T., Zobel, J., & Holt, K. E. (2014). *SRST2: Rapid genomic surveillance for public health and hospital microbiology labs*. 16.

Jauréguy, F., Carbonnelle, E., Bonacorsi, S., Clec'h, C., Casassus, P., Bingen, E., Picard, B., Nassif, X., & Lortholary, O. (2007). Host and bacterial determinants of initial severity and outcome of *Escherichia coli* sepsis. *Clinical Microbiology and Infection*, *13*(9), 854-862. https://doi.org/10.1111/j.1469-0691.2007.01775.x

Jauréguy, F., Landraud, L., Passet, V., Diancourt, L., Frapy, E., Guigon, G., Carbonnelle, E., Lortholary, O., Clermont, O., Denamur, E., Picard, B., Nassif, X., & Brisse, S. (2008). Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics*, *9*(1), 560. https://doi.org/10.1186/1471-2164-9-560

Javaloyas, M. ., Garcia-Somoza, D., & Gudiol, F. (2002). Epidemiology and prognosis of bacteremia: a 10-y study in a community hospital. *Scandinavian Journal of Infectious Diseases*, *34*(6), 436-441. https://doi.org/10.1080/00365540110080629

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., & Aarestrup, F. M. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *Journal of Clinical Microbiology*, *52*(5), 1501-1510. https://doi.org/10.1128/JCM.03617-13

Johnson, J., Menard, M., Lauderdale, T.-L., Lauderdale, T.-L., Kosmidis, C., Gordon, D., Collignon, P., Maslow, J., Andrasević, A., & Kuskowski, M. (2011). Global distribution and epidemiologic associations of *Escherichia coli* clonal group A, 1998–2007. *Emerging Infectious Diseases*, *17*(11). https://doi.org/10.3201/eid1711.110488

Johnson, J. R. (1991). Virulence factors in *Escherichia coli* urinary tract infection. *CLIN. MICROBIOL. REV.*, *4*, 49.

Johnson, J. R., Jelacic, S., Schoening, L. M., Clabots, C., Shaikh, N., Mobley, H. L. T., & Tarr, P. I. (2005). The IrgA homologue adhesin Iha is an *Escherichia coli* virulence factor in murine urinary tract infection. *Infection and Immunity*, *73*(2), 965-971. https://doi.org/10.1128/IAI.73.2.965-971.2005

Johnson, J. R., Johnston, B. D., Porter, S., Thuras, P., Aziz, M., & Price, L. B. (2018). Accessory Traits and Phylogenetic Background Predict Escherichia coli Extraintestinal Virulence Better Than Does Ecological Source. *The Journal of Infectious Diseases*. https://doi.org/10.1093/infdis/jiy459

Johnson, J. R., Kuskowski, M. A., O'Bryan, T. T., & Maslow, J. N. (2002). Epidemiological correlates of virulence genotype and phylogenetic background among *Escherichia coli* blood isolates from adults with diverse-source bacteremia. *The Journal of Infectious Diseases*, *185*(10), 1439-1447. https://doi.org/10.1086/340506

Johnson, J. R., Porter, S. B., Zhanel, G., Kuskowski, M. A., & Denamur, E. (2012). Virulence of *Escherichia coli* clinical isolates in a murine sepsis model in relation to sequence type ST131 status, fluoroquinolone resistance, and virulence genotype. *Infection and Immunity*, *80*(4), 1554-1562. https://doi.org/10.1128/IAI.06388-11

Johnson, J. R., & Russo, T. A. (2002). Extraintestinal pathogenic *Escherichia coli* : "The other bad *E*. *coli*". *Journal of Laboratory and Clinical Medicine*, *139*(3), 155-162. https://doi.org/10.1067/mlc.2002.121550

Johnson, J. R., & Russo, T. A. (2018). Molecular epidemiology of extraintestinal pathogenic *Escherichia coli*. *EcoSal Plus*, *8*(1). https://doi.org/10.1128/ecosalplus.ESP-0004-2017

Johnson, J. R., Urban, C., Weissman, S. J., Jorgensen, J. H., Lewis, J. S., Hansen, G., Edelstein, P. H., Robicsek, A., Cleary, T., Adachi, J., Paterson, D., Quinn, J., Hanson, N. D., Johnston, B. D., Clabots, C., Kuskowski, M. A., & the AMERECUS Investigators. (2012). Molecular epidemiological analysis of *Escherichia coli* sequence type ST131 (O25:H4) and i>bla</i>$_{CTX\text{-}M\text{-}15}$ among extended-spectrum-β-lactamase-producing *E*. *coli* from the United States, 2000 to 2009. *Antimicrobial Agents and Chemotherapy*, *56*(5), 2364-2370. https://doi.org/10.1128/AAC.05824-11

Johnson, T. J., Danzeisen, J. L., Youmans, B., Case, K., Llop, K., Munoz-Aguayo, J., Flores-Figueroa, C., Aziz, M., Stoesser, N., Sokurenko, E., Price, L. B., & Johnson, J. R. (2016). Separate F-type plasmids have shaped the evolution of the *H*30 subclone of *Escherichia coli* sequence type 131. *MSphere*, *1*(4). https://doi.org/10.1128/mSphere.00121-16

Johnson, T. J., & Nolan, L. K. (2009). Pathogenomics of the Virulence Plasmids of Escherichia coli. *Microbiology and Molecular Biology Reviews*, *73*(4), 750-774. https://doi.org/10.1128/MMBR.00015-09

Johnson, T. J., Siek, K. E., Johnson, S. J., & Nolan, L. K. (2006). DNA sequence of a ColV plasmid and prevalence of selected plasmid-encoded virulence genes among avian *Escherichia coli* strains. *Journal of Bacteriology*, *188*(2), 745-758. https://doi.org/10.1128/JB.188.2.745-758.2006

Johnson, T. J., Wannemuehler, Y. M., & Nolan, L. K. (2008). Evolution of the *iss* gene in *Escherichia coli*. *Applied and Environmental Microbiology*, *74*(8), 2360-2369. https://doi.org/10.1128/AEM.02634-07

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5:

genome-scale protein function classification. *Bioinformatics (Oxford, England)*, *30*(9), 1236-1240. https://doi.org/10.1093/bioinformatics/btu031

Kaas, R. S., Friis, C., Ussery, D. W., & Aarestrup, F. M. (2012). Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics*, *13*(1), 577. https://doi.org/10.1186/1471-2164-13-577

Kallonen, T., Brodrick, H. J., Harris, S. R., Corander, J., Brown, N. M., Martin, V., Peacock, S. J., & Parkhill, J. (2017). Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Research*, *27*(8), 1437-1449. https://doi.org/10.1101/gr.216606.116

Kaper, J. B., Nataro, J. P., & Mobley, H. L. T. (2004). Pathogenic *Escherichia coli*. *Nature Reviews Microbiology*, *2*(2), 123-140. https://doi.org/10.1038/nrmicro818

Karp, P. D., Midford, P. E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., Ong, W. K., Subhraveti, P., Caspi, R., Fulcher, C., Keseler, I. M., & Paley, S. M. (2021). Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, *22*(1), 109-126. https://doi.org/10.1093/bib/bbz104

Kauffmann, F. (1947). The serology of the *coli* group. *The Journal of Immunology*, *57*(1), 71-100.

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velázquez-Ramírez, D. A., Weaver, D., Collado-Vides, J., … Karp, P. D. (2017). The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Research*, *45*(D1), D543-D550. https://doi.org/10.1093/nar/gkw1003

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721-1729. https://doi.org/10.1101/gr.210641.116

Kim, J., Bae, I. K., Jeong, S. H., Chang, C. L., Lee, C. H., & Lee, K. (2011). Characterization of IncF plasmids carrying the *bla*CTX-M-14 gene in clinical isolates of *Escherichia coli* coli from Korea. *Journal of Antimicrobial Chemotherapy*, *66*(6), 1263-1268. https://doi.org/10.1093/jac/dkr106

Kolmogorov, M., Raney, B., Paten, B., & Pham, S. (2014). Ragout--a reference-assisted assembly tool for bacterial genomes. *Bioinformatics*, *30*(12), i302-i309. https://doi.org/10.1093/bioinformatics/btu280

Kondratyeva, K., Salmon-Divon, M., & Navon-Venezia, S. (2020). Meta-analysis of pandemic *Escherichia coli* ST131 plasmidome proves restricted plasmid-clade associations. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-019-56763-7

Koonin, E. V., & Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, *36*(21), 6688-6719. https://doi.org/10.1093/nar/gkn668

Koster, J., & Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520-2522. https://doi.org/10.1093/bioinformatics/bts480

Kurazono, H., Yamamoto, S., Nakano, M., Nair, G. B., Terai, A., Chaicumpa, W., & Hayashi, H. (2000). Characterization of a putative virulence island in the chromosome of uropathogenic *Escherichia coli* possessing a gene encoding a uropathogenic-specific protein. *Microbial Pathogenesis*, *28*(3), 183-189. https://doi.org/10.1006/mpat.1999.0331

La Combe, B., Clermont, O., Messika, J., Eveillard, M., Kouatchet, A., Lasocki, S., Corvec, S., Lakhal, K., Billard-Pomares, T., Fernandes, R., Armand-Lefevre, L., Bourdon, S., Reignier, J., Fihman, V., de Prost, N., Bador, J., Goret, J., Wallet, F., Denamur, E., … on behalf of the COLOCOLI group. (2019). Pneumonia-specific *Escherichia coli* with distinct phylogenetic and virulence profiles, France, 2012–2014. *Emerging Infectious Diseases*, *25*(4), 710-718. https://doi.org/10.3201/eid2504.180944

Laczny, C. C., Galata, V., Plum, A., Posch, A. E., & Keller, A. (2019). Assessing the heterogeneity of *in silico* plasmid predictions based on whole-genome-sequenced clinical isolates. *Briefings in Bioinformatics*, *20*(3), 857-865. https://doi.org/10.1093/bib/bbx162

Landraud, L., Gauthier, M., Fosse, T., & Boquet, P. (2000). Frequency of *Escherichia coli* strains producing the cytotoxic necrotizing factor (CNF1) in nosocomial urinary tract infections. *Letters in Applied Microbiology*, *30*(3), 213-216. https://doi.org/10.1046/j.1472-765x.2000.00698.x

Landraud, L., Jauréguy, F., Frapy, E., Guigon, G., Gouriou, S., Carbonnelle, E., Clermont, O., Denamur, E., Picard, B., Lemichez, E., Brisse, S., & Nassif, X. (2013). Severity of *Escherichia coli* bacteraemia is independent of the intrinsic virulence of the strains assessed in a mouse model. *Clinical Microbiology and Infection*, *19*(1), 85-90. https://doi.org/10.1111/j.1469-0691.2011.03750.x

Lane, M. C., Alteri, C. J., Smith, S. N., & Mobley, H. L. T. (2007). Expression of flagella is coincident with uropathogenic *Escherichia coli* ascension to the upper urinary tract. *Proceedings of the National Academy of Sciences*, *104*(42), 16669-16674. https://doi.org/10.1073/pnas.0607898104

Lane, M. C., & Mobley, H. L. T. (2007). Role of P-fimbrial-mediated adherence in pyelonephritis and persistence of uropathogenic *Escherichia coli* (UPEC) in the mammalian kidney. *Kidney International*, *72*(1), 19-25. https://doi.org/10.1038/sj.ki.5002230

Lane, M. C., Simms, A. N., & Mobley, H. L. T. (2007). Complex interplay between type 1 fimbrial expression and flagellum-mediated motility of uropathogenic *Escherichia coli*. *Journal of Bacteriology*, *189*(15), 5523-5533. https://doi.org/10.1128/JB.00434-07

Lanza, V. F., de Toro, M., Garcillán-Barcia, M. P., Mora, A., Blanco, J., Coque, T. M., & de la Cruz, F. (2014). Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genetics*, *10*(12), e1004766. https://doi.org/10.1371/journal.pgen.1004766

Laupland, K.B., Gregson, D. B., Church, D. L., Ross, T., & Pitout, J. D. D. (2008). Incidence, risk factors and outcomes of *Escherichia coli* bloodstream infections in a large Canadian region. *Clinical Microbiology and Infection*, *14*(11), 1041-1047. https://doi.org/10.1111/j.1469-0691.2008.02089.x

Laupland, Kevin B., & Church, D. L. (2014). Population-based epidemiology and microbiology of community-onset bloodstream infections. *Clinical Microbiology Reviews*, *27*(4), 647-664. https://doi.org/10.1128/CMR.00002-14

Lawlor, M. S., O'Connor, C., & Miller, V. L. (2007). *Yersiniabactin* is a virulence factor for *Klebsiella pneumoniae* during pulmonary infection. *Infection and Immunity*, *75*(3), 1463-1472. https://doi.org/10.1128/IAI.00372-06

Le Bouguénec, C., & Schouler, C. (2011). Sugar metabolism, an additional virulence factor in enterobacteria. *International Journal of Medical Microbiology*, *301*(1), 1-6. https://doi.org/10.1016/j.ijmm.2010.04.021

Le Gall, T., Clermont, O., Gouriou, S., Picard, B., Nassif, X., Denamur, E., & Tenaillon, O. (2007). Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. *Molecular Biology and Evolution*, *24*(11), 2373-2384. https://doi.org/10.1093/molbev/msm172

Le Minor, L., & Richard, C. (1993). *Méthodes de laboratoire pour l'identification des entérobactéries*. Institut Pasteur.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR : An *R* Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1). https://doi.org/10.18637/jss.v025.i01

Lecointre, G., Rachdi, L., Darlu, P., & Denamur, E. (1998). Escherichia coli molecular phylogeny using the incongruence length difference test. *Molecular Biology and Evolution*, *15*(12), 1685-1695. https://doi.org/10.1093/oxfordjournals.molbev.a025895

Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N., & Corander, J. (2018). pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, *34*(24), 4310-4312. https://doi.org/10.1093/bioinformatics/bty539

Lees, J. A., Harris, S. R., Tonkin-Hill, G., Gladstone, R. A., Lo, S. W., Weiser, J. N., Corander, J., Bentley, S. D., & Croucher, N. J. (2019). Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Research*, *29*(2), 304-316. https://doi.org/10.1101/gr.241455.118

Lefort, A., Panhard, X., Clermont, O., Woerther, P.-L., Branger, C., Mentre, F., Fantin, B., Wolff, M., Denamur, E., & for the COLIBAFI Group. (2011). Host factors and portal of

entry outweigh bacterial determinants to predict the severity of *Escherichia coli* bacteremia. *Journal of Clinical Microbiology*, *49*(3), 777-783. https://doi.org/10.1128/JCM.01902-10

Lescat, M., Calteau, A., Hoede, C., Barbe, V., Touchon, M., Rocha, E., Tenaillon, O., Médigue, C., Johnson, J. R., & Denamur, E. (2009). A module located at a chromosomal integration hot spot is responsible for the multidrug resistance of a reference strain from *Escherichia coli* clonal group A. *Antimicrobial Agents and Chemotherapy*, *53*(6), 2283-2288. https://doi.org/10.1128/AAC.00123-09

Léveillé, S., Caza, M., Johnson, J. R., Clabots, C., Sabri, M., & Dozois, C. M. (2006). Iha from an *Escherichia coli* urinary tract infection outbreak clonal group A strain is expressed in vivo in the mouse urinary tract and functions as a catecholate siderophore receptor. *Infection and Immunity*, *74*(6), 3427-3436. https://doi.org/10.1128/IAI.00107-06

Levin, B. R., & Edén, C. S. (1990). Selection and evolution of virulence in bacteria: an ecumenical excursion and modest suggestion. *Parasitology*, *100*(S1), S103-S115. https://doi.org/10.1017/S0031182000073054

Lillington, J., Geibel, S., & Waksman, G. (2014). Biogenesis and adhesion of type 1 and P pili. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1840*(9), 2783-2793. https://doi.org/10.1016/j.bbagen.2014.04.021

Lindstedt, R., Larson, G., Falk, P., Jodal, U., Leffler, H., & Svanborg, C. (1991). The receptor repertoire defines the host range for attaching *Escherichia coli* strains that recognize globo-A. *Infection and Immunity*, *59*(3), 1086-1092. https://doi.org/10.1128/IAI.59.3.1086-1092.1991

Liu, Y.-Y., Wang, Y., Walsh, T. R., Yi, L.-X., Zhang, R., Spencer, J., Doi, Y., Tian, G., Dong, B., Huang, X., Yu, L.-F., Gu, D., Ren, H., Chen, X., Lv, L., He, D., Zhou, H., Liang, Z., Liu, J.-H., & Shen, J. (2016). Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *The Lancet Infectious Diseases*, *16*(2), 161-168. https://doi.org/10.1016/S1473-3099(15)00424-7

Lu, S., Jin, D., Wu, S., Yang, J., Lan, R., Bai, X., Liu, S., Meng, Q., Yuan, X., Zhou, J., Pu, J., Chen, Q., Dai, H., Hu, Y., Xiong, Y., Ye, C., & Xu, J. (2016). Insights into the evolution of pathogenicity of *Escherichia coli* from genomic analysis of intestinal *E. coli* of *Marmota himalayana* in Qinghai–Tibet plateau of China. *Emerging Microbes & Infections*, *5*(1), 1-9. https://doi.org/10.1038/emi.2016.122

Ludden, C., Decano, A. G., Jamrozy, D., Pickard, D., Morris, D., Parkhill, J., Peacock, S. J., Cormican, M., & Downing, T. (2020). Genomic surveillance of *Escherichia coli* ST131 identifies local expansion and serial replacement of subclones. *Microbial Genomics*, *6*(4). https://doi.org/10.1099/mgen.0.000352

Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M., & Konstantinidis, K. T. (2011). Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the*

National Academy of Sciences, *108*(17), 7200-7205. https://doi.org/10.1073/pnas.1015622108

Magruder, M., Sholi, A. N., Gong, C., Zhang, L., Edusei, E., Huang, J., Albakry, S., Satlin, M. J., Westblade, L. F., Crawford, C., Dadhania, D. M., Lubetzky, M., Taur, Y., Littman, E., Ling, L., Burnham, P., De Vlaminck, I., Pamer, E., Suthanthiran, M., & Lee, J. R. (2019). Gut uropathogen abundance is a risk factor for development of bacteriuria and urinary tract infection. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-13467-w

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., & Spratt, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, *95*(6), 3140-3145. https://doi.org/10.1073/pnas.95.6.3140

Manges, A. R., Geum, H. M., Guo, A., Edens, T. J., Fibke, C. D., & Pitout, J. D. D. (2019). Global extraintestinal pathogenic *Escherichia coli* (ExPEC) lineages. *Clinical Microbiology Reviews*, *32*(3). https://doi.org/10.1128/CMR.00135-18

Manges, A. R., Johnson, J. R., Foxman, B., O'Bryan, T. T., Fullerton, K. E., & Riley, L. W. (2001). Widespread distribution of urinary tract infections caused by a multidrug-resistant *Escherichia coli* clonal group. *New England Journal of Medicine*, *345*(14), 1007-1013. https://doi.org/10.1056/NEJMoa011265

Mansan-Almeida, R., Pereira, A., & Giugliano, L. (2013). Diffusely adherent *Escherichia coli* strains isolated from children and adults constitute two different populations. *BMC Microbiology*, *13*(1), 22. https://doi.org/10.1186/1471-2180-13-22

Mariani-Kurkdjian, P., & Bingen, E. (2012). *Escherichia coli* O104:H4 : un pathotype hybride. *Archives de Pédiatrie*, *19*, S97-S100. https://doi.org/10.1016/S0929-693X(12)71281-2

Mariani-Kurkdjian, P., Lemaître, C., Bidet, P., Perez, D., Boggini, L., Kwon, T., & Bonacorsi, S. (2014). Haemolytic-uraemic syndrome with bacteraemia caused by a new hybrid *Escherichia coli* pathotype. *New Microbes and New Infections*, *2*(4), 127-131. https://doi.org/10.1002/nmi2.49

Marrs, C. F., Zhang, L., Tallman, P., Manning, S. D., Somsel, P., Raz, P., Colodner, R., Jantunen, M. E., Siitonen, A., Saxen, H., & Foxman, B. (2002). Variations in 10 putative uropathogen virulence genes among urinary, faecal and peri-urethral Escherichia coli. *Journal of Medical Microbiology*, *51*(2), 138-142. https://doi.org/10.1099/0022-1317-51-2-138

Martinez, J. A., Soto, S., Fabrega, A., Almela, M., Mensa, J., Soriano, A., Marco, F., Jimenez de Anta, M. T., & Vila, J. (2006). Relationship of phylogenetic background, biofilm production, and time to detection of growth in blood culture vials with clinical variables and prognosis associated with *Escherichia coli* bacteremia. *Journal of Clinical Microbiology*, *44*(4), 1468-1474. https://doi.org/10.1128/JCM.44.4.1468-1474.2006

Massot, M., Daubié, A.-S., Clermont, O., Jauréguy, F., Couffignal, C., Dahbi, G., Mora, A., Blanco, J., Branger, C., Mentré, F., Eddi, A., Picard, B., Denamur, E., & the COLIVILLE Group. (2016). Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of *Escherichia coli* from community subjects in the Paris area in 2010 and evolution over 30 years. *Microbiology*, *162*(4), 642-650. https://doi.org/10.1099/mic.0.000242

Matamoros, S., van Hattem, J. M., Arcilla, M. S., Willemse, N., Melles, D. C., Penders, J., Vinh, T. N., Thi Hoa, N., Bootsma, M. C. J., van Genderen, P. J., Goorhuis, A., Grobusch, M., Molhoek, N., Oude Lashof, A. M. L., Stobberingh, E. E., Verbrugh, H. A., de Jong, M. D., & Schultsz, C. (2017). Global phylogenetic analysis of *Escherichia coli* and plasmids carrying the *mcr-1* gene indicates bacterial diversity but plasmid restriction. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-15539-7

Mathers, A. J., Peirano, G., & Pitout, J. D. D. (2015). The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant *Enterobacteriaceae*. *Clinical Microbiology Reviews*, *28*(3), 565-591. https://doi.org/10.1128/CMR.00116-14

Matsumura, Y., Noguchi, T., Tanaka, M., Kanahashi, T., Yamamoto, M., Nagao, M., Takakura, S., Ichiyama, S., & on behalf of the 89th JAID BRG. (2017). Population structure of Japanese extraintestinal pathogenic *Escherichia coli* and its relationship with antimicrobial resistance. *Journal of Antimicrobial Chemotherapy*, dkw530. https://doi.org/10.1093/jac/dkw530

Maurelli, A. T., Fernandez, R. E., Bloch, C. A., Rode, C. K., & Fasano, A. (1998). « Black holes » and bacterial pathogenicity: A large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proceedings of the National Academy of Sciences*, *95*(7), 3943-3948. https://doi.org/10.1073/pnas.95.7.3943

McNally, A., Alhashash, F., Collins, M., Alqasim, A., Paszckiewicz, K., Weston, V., & Diggle, M. (2013). Genomic analysis of extra-intestinal pathogenic *Escherichia coli* urosepsis. *Clinical Microbiology and Infection*, *19*(8), e328-e334. https://doi.org/10.1111/1469-0691.12202

McNally, A., Kallonen, T., Connor, C., Abudahab, K., Aanensen, D. M., Horner, C., Peacock, S. J., Parkhill, J., Croucher, N. J., & Corander, J. (2019). Diversification of colonization factors in a multidrug-resistant *Escherichia coli* lineage evolving under negative frequency-dependent selection. *MBio*, *10*(2). https://doi.org/10.1128/mBio.00644-19

McPhee, J. B., Small, C. L., Reid-Yu, S. A., Brannon, J. R., Le Moual, H., & Coombes, B. K. (2014). Host defense peptide resistance contributes to colonization and maximal intestinal pathology by Crohn's disease-associated adherent-invasive *Escherichia coli*. *Infection and Immunity*, *82*(8), 3383-3393. https://doi.org/10.1128/IAI.01888-14

Méric, G., Hitchings, M. D., Pascoe, B., & Sheppard, S. K. (2016). From Escherich to the *Escherichia coli* genome. *The Lancet Infectious Diseases*, *16*(6), 634-636. https://doi.org/10.1016/S1473-3099(16)30066-4

Miajlovic, H., & Smith, S. G. (2014). Bacterial self-defence: how *Escherichia coli* evades serum killing. *FEMS Microbiology Letters*, *354*(1), 1-9. https://doi.org/10.1111/1574-6968.12419

Milkman, R. (1973). Electrophoretic variation in *Escherichia coli* from natural sources. *Science*, *182*(4116), 1024-1026.

Mohsin, M., Raza, S., Schaufler, K., Roschanski, N., Sarwar, F., Semmler, T., Schierack, P., & Guenther, S. (2017). High prevalence of CTX-M-15-type ESBL-producing *E. coli* from migratory avian species in Pakistan. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.02476

Moissenet, D., Salauze, B., Clermont, O., Bingen, E., Arlet, G., Denamur, E., Mérens, A., Mitanchez, D., & Vu-Thien, H. (2010). Meningitis caused by *Escherichia coli* producing TEM-52 extended-spectrum beta-lactamase within an extensive outbreak in a neonatal ward: epidemiological investigation and characterization of the strain. *Journal of Clinical Microbiology*, *48*(7), 2459-2463. https://doi.org/10.1128/JCM.00529-10

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M., & Palsson, B. O. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences*, *110*(50), 20338-20343. https://doi.org/10.1073/pnas.1307797110

Mora-Rillo, M., Fernández-Romero, N., Navarro-San Francisco, C., Díez-Sebastián, J., Romero-Gómez, M. P., Arnalich Fernández, F., Arribas López, J. R., & Mingorance, J. (2015). Impact of virulence genes on sepsis severity and survival in *Escherichia coli* bacteremia. *Virulence*, *6*(1), 93-100. https://doi.org/10.4161/21505594.2014.991234

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268-274. https://doi.org/10.1093/molbev/msu300

Nicolas-Chanoine, M.-H., Bertrand, X., & Madec, J.-Y. (2014). *Escherichia coli* ST131, an intriguing clonal group. *Clinical Microbiology Reviews*, *27*(3), 543-574. https://doi.org/10.1128/CMR.00125-13

Nicolas-Chanoine, M.-H., Blanco, J., Leflon-Guibout, V., Demarty, R., Alonso, M. P., Canica, M. M., Park, Y.-J., Lavigne, J.-P., Pitout, J., & Johnson, J. R. (2007). Intercontinental emergence of *Escherichia coli* clone O25:H4-ST131 producing CTX-M-15. *Journal of Antimicrobial Chemotherapy*, *61*(2), 273-281. https://doi.org/10.1093/jac/dkm464

Ochman, H., & Selander, R. K. (1984a). Standard reference strains of Escherichia coli from natural populations. *Journal of Bacteriology*, *157*(2), 690-693. https://doi.org/10.1128/JB.157.2.690-693.1984

Ochman, H., & Selander, R. K. (1984b). Evidence for clonal population structure in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, *81*(1), 198-201. https://doi.org/10.1073/pnas.81.1.198

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0997-x

Orskov, F., & Orskov, I. (1975). *Escherichia coli* O:H serotypes isolated from human blood: prevalence of the K1 antigen with technical details of O and H antigenic determination. *Acta pathologica et microbiologica Scandinavica. Supplement*, *83*(6), 595-600.

Östblom, A., Adlerberth, I., Wold, A. E., & Nowrouzian, F. L. (2011). Pathogenicity island markers, virulence determinants *malX* and *usp* , and the capacity of *Escherichia coli* to persist in infants' commensal microbiotas. *Applied and Environmental Microbiology*, *77*(7), 2303-2308. https://doi.org/10.1128/AEM.02405-10

Park, S.-C., Lee, K., Kim, Y. O., Won, S., & Chun, J. (2019). Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.00834

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

Parreira, V. R., & Gyles, C. L. (2003). A novel pathogenicity island integrated adjacent to the *thrw* trna gene of avian pathogenic *Escherichia coli* encodes a vacuolating autotransporter toxin. *Infection and Immunity*, *71*(9), 5087-5096. https://doi.org/10.1128/IAI.71.9.5087-5096.2003

Parret, A. H. A., & De Mot, R. (2002). *Escherichia coli*'s uropathogenic-specific protein: a bacteriocin promoting infectivity? *Microbiology (Reading, England)*, *148*(Pt 6), 1604-1606. https://doi.org/10.1099/00221287-148-6-1604

Pasqua, M., Michelacci, V., Di Martino, M. L., Tozzoli, R., Grossi, M., Colonna, B., Morabito, S., & Prosseda, G. (2017). The intriguing evolutionary journey of enteroinvasive *E. coli* (EIEC) toward pathogenicity. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.02390

Peigne, C., Bidet, P., Mahjoub-Messai, F., Plainvert, C., Barbe, V., Médigue, C., Frapy, E., Nassif, X., Denamur, E., Bingen, E., & Bonacorsi, S. (2009). The plasmid of *Escherichia coli* strain S88 (O45:K1:H7) that causes neonatal meningitis is closely related to avian pathogenic *E. coli* plasmids and is associated with high-level bacteremia in a neonatal rat meningitis model. *Infection and Immunity*, *77*(6), 2272-2284. https://doi.org/10.1128/IAI.01333-08

Peirano, G., van der Bij, A. K., Gregson, D. B., & Pitout, J. D. D. (2012). Molecular epidemiology over an 11-year period (2000 to 2010) of extended-spectrum β-lactamase-producing *Escherichia coli* causing bacteremia in a centralized Canadian region. *Journal of Clinical Microbiology*, *50*(2), 294-299. https://doi.org/10.1128/JCM.06025-11

Pere, A., Leinonen, M., Vaisanen-Rhen, V., Rhen, M., & Korhonen, T. K. (1985). Occurrence of type-1C fimbriae on *Escherichia coli* strains isolated from human extraintestinal infections. *Microbiology*, *131*(7), 1705-1711. https://doi.org/10.1099/00221287-131-7-1705

Picard, B., Garcia, J. S., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., Elion, J., & Denamur, E. (1999). The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infection and Immunity*, *67*(2), 546-553. https://doi.org/10.1128/IAI.67.2.546-553.1999

Poirel, L., Bonnin, R. A., & Nordmann, P. (2012). Genetic features of the widespread plasmid coding for the carbapenemase OXA-48. *Antimicrobial Agents and Chemotherapy*, *56*(1), 559-562. https://doi.org/10.1128/AAC.05289-11

Poirel, L., Decousser, J.-W., & Nordmann, P. (2003). Insertion sequence IS*ecp1B* is involved in expression and mobilization of a *bla*CTX-M β-lactamase gene. *Antimicrobial Agents and Chemotherapy*, *47*(9), 2938-2945. https://doi.org/10.1128/AAC.47.9.2938-2945.2003

Prager, R., Lang, C., Aurass, P., Fruth, A., Tietze, E., & Flieger, A. (2014). Two novel EHEC/EAEC hybrid strains isolated from human infections. *PLoS ONE*, *9*(4), e95379. https://doi.org/10.1371/journal.pone.0095379

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes *de novo* assembler. *Current Protocols in Bioinformatics*, *70*(1). https://doi.org/10.1002/cpbi.102

Raimondi, S., Righini, L., Candeliere, F., Musmeci, E., Bonvicini, F., Gentilomi, G., Starčič Erjavec, M., Amaretti, A., & Rossi, M. (2019). Antibiotic resistance, virulence factors, phenotyping, and genotyping of *E. coli* isolated from the feces of healthy subjects. *Microorganisms*, *7*(8), 251. https://doi.org/10.3390/microorganisms7080251

Raymond, K. N., Dertz, E. A., & Kim, S. S. (2003). Enterobactin: An archetype for microbial iron transport. *Proceedings of the National Academy of Sciences*, *100*(7), 3584-3588. https://doi.org/10.1073/pnas.0630018100

Reddy, E. A., Shaw, A. V., & Crump, J. A. (2010). Community-acquired bloodstream infections in Africa: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, *10*(6), 417-432. https://doi.org/10.1016/S1473-3099(10)70072-4

Redondo-Salvo, S., Fernández-López, R., Ruiz, R., Vielva, L., de Toro, M., Rocha, E. P. C., Garcillán-Barcia, M. P., & de la Cruz, F. (2020). Pathways for horizontal gene transfer

in bacteria revealed by a global map of their plasmids. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17278-2

Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K., & Whittam, T. S. (2000). Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature*, *406*(6791), 64-67. https://doi.org/10.1038/35017546

Reitzer, L., & Zimmern, P. (2019). Rapid growth and metabolism of uropathogenic *Escherichia coli* in relation to urine composition. *Clinical Microbiology Reviews*, *33*(1). https://doi.org/10.1128/CMR.00101-19

Réseau d'alerte, d'investigation et de surveillance des infections nosocomiales (Raisin). (2004). *Surveillance des bactériémies nosocomiales en France - Réseau BN-Raisin - Résultats 2004*. https://www.santepubliquefrance.fr/content/download/184722/2313790#:~:text=Il%20s'agit%20d'une,microbiologiques%20par%20des%20donn%C3%A9es%20cliniques.

Restieri, C., Garriss, G., Locas, M.-C., & Dozois, C. M. (2007). Autotransporter-encoding sequences are phylogenetically distributed among *Escherichia coli* clinical isolates and reference strains. *Applied and Environmental Microbiology*, *73*(5), 1553-1562. https://doi.org/10.1128/AEM.01542-06

Rios Miguel, A. B., Jetten, M. S. M., & Welte, C. U. (2020). The role of mobile genetic elements in organic micropollutant degradation during biological wastewater treatment. *Water Research X*, *9*, 100065. https://doi.org/10.1016/j.wroa.2020.100065

Robertson, J., & Nash, J. H. E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics*, *4*(8). https://doi.org/10.1099/mgen.0.000206

Rodriguez-Bano, J., Picon, E., Gijon, P., Hernandez, J. R., Cisneros, J. M., Pena, C., Almela, M., Almirante, B., Grill, F., Colomina, J., Molinos, S., Oliver, A., Fernandez-Mazarrasa, C., Navarro, G., Coloma, A., Lopez-Cerero, L., & Pascual, A. (2010). Risk factors and prognosis of nosocomial bloodstream infections caused by extended-spectrum-β-lactamase-producing *Escherichia coli*. *Journal of Clinical Microbiology*, *48*(5), 1726-1731. https://doi.org/10.1128/JCM.02353-09

Roer, L., Hansen, F., Thomsen, M. C. F., Knudsen, J. D., Hansen, D. S., Wang, M., Samulioniené, J., Justesen, U. S., Røder, B. L., Schumacher, H., Østergaard, C., Andersen, L. P., Dzajic, E., Søndergaard, T. S., Stegger, M., Hammerum, A. M., & Hasman, H. (2017). WGS-based surveillance of third-generation cephalosporin-resistant *Escherichia coli* from bloodstream infections in Denmark. *Journal of Antimicrobial Chemotherapy*, *72*(7), 1922-1929. https://doi.org/10.1093/jac/dkx092

Rousset, F., Cabezas-Caballero, J., Piastra-Facon, F., Fernández-Rodríguez, J., Clermont, O., Denamur, E., Rocha, E. P. C., & Bikard, D. (2021). The impact of genetic diversity on gene essentiality within the *Escherichia coli* species. *Nature Microbiology*. https://doi.org/10.1038/s41564-020-00839-y

Rozov, R., Brown Kav, A., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., & Shamir, R. (2016). Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. *Bioinformatics*, btw651. https://doi.org/10.1093/bioinformatics/btw651

Rozwandowicz, M., Brouwer, M. S. M., Fischer, J., Wagenaar, J. A., Gonzalez-Zorn, B., Guerra, B., Mevius, D. J., & Hordijk, J. (2018). Plasmids carrying antimicrobial resistance genes in *Enterobacteriaceae*. *Journal of Antimicrobial Chemotherapy*, *73*(5), 1121-1137. https://doi.org/10.1093/jac/dkx488

Russo, T. A., Carlino, U. B., & Johnson, J. R. (2001). Identification of a new iron-regulated virulence gene, *ireA*, in an extraintestinal pathogenic isolate of *Escherichia coli*. *Infection and Immunity*, *69*(10), 6209-6216. https://doi.org/10.1128/IAI.69.10.6209-6216.2001

Russo, Thomas A., & Johnson, J. R. (2000). Proposal for a new inclusive designation for extraintestinal pathogenic isolates of *Escherichia coli:* ExPEC. *The Journal of Infectious Diseases*, *181*(5), 1753-1754. https://doi.org/10.1086/315418

Ryu, J. Y., Kim, H. U., & Lee, S. Y. (2019). Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proceedings of the National Academy of Sciences*, *116*(28), 13996-14001. https://doi.org/10.1073/pnas.1821905116

Sabarly, V., Aubron, C., Glodt, J., Balliau, T., Langella, O., Chevret, D., Rigal, O., Bourgais, A., Picard, B., de Vienne, D., Denamur, E., Bouvet, O., & Dillmann, C. (2016). Interactions between genotype and environment drive the metabolic phenotype within *Escherichia coli* isolates: metabolic diversity within *Escherichia coli* isolates. *Environmental Microbiology*, *18*(1), 100-117. https://doi.org/10.1111/1462-2920.12855

Sabarly, V., Bouvet, O., Glodt, J., Clermont, O., Skurnik, D., Diancourt, L., De VIENNE, D., Denamur, E., & Dillmann, C. (2011a). The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity: metabolic diversity in *E. coli*. *Journal of Evolutionary Biology*, *24*(7), 1559-1571. https://doi.org/10.1111/j.1420-9101.2011.02287.x

Sabarly, V., Bouvet, O., Glodt, J., Clermont, O., Skurnik, D., Diancourt, L., De VIENNE, D., Denamur, E., & Dillmann, C. (2011b). The decoupling between genetic structure and metabolic phenotypes in Escherichia coli leads to continuous phenotypic diversity: Metabolic diversity in E. coli. *Journal of Evolutionary Biology*, *24*(7), 1559-1571. https://doi.org/10.1111/j.1420-9101.2011.02287.x

Salipante, S. J., Roach, D. J., Kitzman, J. O., Snyder, M. W., Stackhouse, B., Butler-Wu, S. M., Lee, C., Cookson, B. T., & Shendure, J. (2015). Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome Research*, *25*(1), 119-128. https://doi.org/10.1101/gr.180190.114

Sarkar, S., Ulett, G. C., Totsika, M., Phan, M.-D., & Schembri, M. A. (2014). Role of capsule and O antigen in the virulence of uropathogenic *Escherichia coli*. *PLoS ONE*, *9*(4), e94786. https://doi.org/10.1371/journal.pone.0094786

Schouler, C., Taki, A., Chouikha, I., Moulin-Schouleur, M., & Gilot, P. (2009). A genomic island of an extraintestinal pathogenic *Escherichia coli* strain enables the metabolism of fructooligosaccharides, which improves intestinal colonization. *Journal of Bacteriology*, *191*(1), 388-393. https://doi.org/10.1128/JB.01052-08

Schubert, S., Cuenca, S., Fischer, D., & Heesemann, J. (2000). High-pathogenicity island of *Yersinia pestis* in *Enterobacteriaceae* isolated from blood cultures and urine samples: prevalence and functional expression. *The Journal of Infectious Diseases*, *182*(4), 1268-1271. https://doi.org/10.1086/315831

Schubert, S., Picard, B., Gouriou, S., Heesemann, J., & Denamur, E. (2002). *Yersinia* high-pathogenicity island contributes to virulence in *Escherichia coli* causing extraintestinal infections. *Infection and Immunity*, *70*(9), 5335-5337. https://doi.org/10.1128/IAI.70.9.5335-5337.2002

Schwaber, M. J., & Carmeli, Y. (2007). Mortality and delay in effective therapy associated with extended-spectrum β-lactamase production in *Enterobacteriaceae* bacteraemia: a systematic review and meta-analysis. *Journal of Antimicrobial Chemotherapy*, *60*(5), 913-920. https://doi.org/10.1093/jac/dkm318

Schwengers, O., Barth, P., Falgenhauer, L., Hain, T., Chakraborty, T., & Goesmann, A. (2020). Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microbial Genomics*, *6*(10). https://doi.org/10.1099/mgen.0.000398

Seemann, T. (2021). *Shovill* [Perl]. https://github.com/tseemann/shovill (Original work published 2016)

Seemann, T. (2021). *ABRicate* [Perl]. https://github.com/tseemann/abricate (Original work published 2014)

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N., & Whittam, T. S. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Applied and Environmental Microbiology*, *51*(5), 873-884. https://doi.org/10.1128/AEM.51.5.873-884.1986

Selander, R. K., Caugant, D. A., & Whittam, T. S. (1987). Genetic structure and variation in natural populations of *Escherichia coli*. In *Escherichia coli and Salmonella typhimurium: cellular and molecular biology* (American Society for Microbiology, p. 1625-1648). Neidhardt F.C. et al.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J.-L., & Angus, D. C. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, *315*(8), 801. https://doi.org/10.1001/jama.2016.0287

Skjøt-Rasmussen, L., Ejrnæs, K., Lundgren, B., Hammerum, A. M., & Frimodt-Møller, N. (2012). Virulence factors and phylogenetic grouping of *Escherichia coli* isolates from patients with bacteraemia of urinary tract origin relate to sex and hospital- vs. community-acquired origin. *International Journal of Medical Microbiology*, *302*(3), 129-134. https://doi.org/10.1016/j.ijmm.2012.03.002

Skurnik, D., Bonnet, D., Bernde-Bauduin, C., Michel, R., Guette, C., Becker, J.-M., Balaire, C., Chau, F., Mohler, J., Jarlier, V., Boutin, J.-P., Moreau, B., Guillemot, D., Denamur, E., Andremont, A., & Ruimy, R. (2008). Characteristics of human intestinal *Escherichia coli* with changing environments. *Environmental Microbiology*, *10*(8), 2132-2137. https://doi.org/10.1111/j.1462-2920.2008.01636.x

Skurnik, D., Clermont, O., Guillard, T., Launay, A., Danilchanka, O., Pons, S., Diancourt, L., Lebreton, F., Kadlec, K., Roux, D., Jiang, D., Dion, S., Aschard, H., Denamur, M., Cywes-Bentley, C., Schwarz, S., Tenaillon, O., Andremont, A., Picard, B., … Denamur, E. (2016). Emergence of antimicrobial-resistant *Escherichia coli* of animal origin spreading in humans. *Molecular Biology and Evolution*, *33*(4), 898-914. https://doi.org/10.1093/molbev/msv280

Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C., & de la Cruz, F. (2010). Mobility of plasmids. *Microbiology and Molecular Biology Reviews*, *74*(3), 434-452. https://doi.org/10.1128/MMBR.00020-10

Smith, M. A., Weingarten, R. A., Russo, L. M., Ventura, C. L., & O'Brien, A. D. (2015). Antibodies against hemolysin and cytotoxic necrotizing factor type 1 (CNF1) reduce bladder inflammation in a mouse model of urinary tract infection with toxigenic uropathogenic *Escherichia coli*. *Infection and Immunity*, *83*(4), 1661-1673. https://doi.org/10.1128/IAI.02848-14

Solberg, O. D., Ajiboye, R. M., & Riley, L. W. (2006). Origin of class 1 and 2 integrons and gene cassettes in a population-based sample of uropathogenic *Escherichia coli*. *Journal of Clinical Microbiology*, *44*(4), 1347-1351. https://doi.org/10.1128/JCM.44.4.1347-1351.2006

Soysal, N., Mariani-Kurkdjian, P., Smail, Y., Liguori, S., Gouali, M., Loukiadis, E., Fach, P., Bruyand, M., Blanco, J., Bidet, P., & Bonacorsi, S. (2016). Enterohemorrhagic *Escherichia coli* hybrid pathotype O80:H2 as a new therapeutic challenge. *Emerging Infectious Diseases*, *22*(9), 1604-1612. https://doi.org/10.3201/eid2209.160304

Starčič Erjavec, M., & Žgur-Bertok, D. (2015). Virulence potential for extraintestinal infections among commensal *Escherichia coli* isolated from healthy humans—the Trojan horse within our gut. *FEMS Microbiology Letters*, *362*(5). https://doi.org/10.1093/femsle/fnu061

Strömberg, N., Marklund, B. I., Lund, B., Ilver, D., Hamers, A., Gaastra, W., Karlsson, K. A., & Normark, S. (1990). Host-specificity of uropathogenic *Escherichia coli* depends on differences in binding specificity to Galα1-4Gal-containing isoreceptors. *The EMBO Journal*, *9*(6), 2001-2010.

Strömberg, N., Nyholm, P. G., Pascher, I., & Normark, S. (1991). Saccharide orientation at the cell surface affects glycolipid receptor function. *Proceedings of the National Academy of Sciences*, *88*(20), 9340-9344. https://doi.org/10.1073/pnas.88.20.9340

Stumpe, S., Schmid, R., Stephens, D. L., Georgiou, G., & Bakker, E. P. (1998). Identification of OmpT as the protease that hydrolyzes the antimicrobial peptide protamine before it enters growing cells of *Escherichia coli*. *Journal of Bacteriology*, *180*(15), 4002-4006. https://doi.org/10.1128/JB.180.15.4002-4006.1998

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1). https://doi.org/10.1093/ve/vey016

Tao, X., Wang, H., Min, C., Yu, T., Luo, Y., Li, J., Hu, Y., Yan, Q., Liu, W. en, & Zou, M. (2020). A retrospective study on *Escherichia coli* bacteremia in immunocompromised patients: Microbiological features, clinical characteristics, and risk factors for shock and death. *Journal of Clinical Laboratory Analysis*, *34*(8). https://doi.org/10.1002/jcla.23319

Tchesnokova, V., Radey, M., Chattopadhyay, S., Larson, L., Weaver, J. L., Kisiela, D., & Sokurenko, E. V. (2019). Pandemic fluoroquinolone resistant *Escherichia coli* clone ST1193 emerged via simultaneous homologous recombinations in 11 gene loci. *Proceedings of the National Academy of Sciences*, *116*(29), 14740-14748. https://doi.org/10.1073/pnas.1903002116

Tchesnokova, V., Rechkina, E., Larson, L., Ferrier, K., Weaver, J. L., Schroeder, D. W., She, R., Butler-Wu, S. M., Aguero-Rosenfeld, M. E., Zerr, D., Fang, F. C., Ralston, J., Riddell, K., Scholes, D., Weissman, S., Parker, K., Spellberg, B., Johnson, J. R., & Sokurenko, E. V. (2019). Rapid and extensive expansion in the United States of a new multidrug-resistant *Escherichia coli* clonal group, sequence type 1193. *Clinical Infectious Diseases*, *68*(2), 334-337. https://doi.org/10.1093/cid/ciy525

Tenaillon, O., Skurnik, D., Picard, B., & Denamur, E. (2010). The population genetics of commensal *Escherichia coli*. *Nature Reviews Microbiology*, *8*(3), 207-217. https://doi.org/10.1038/nrmicro2298

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial « pan-genome ». *Proceedings of the National Academy of Sciences*, *102*(39), 13950-13955. https://doi.org/10.1073/pnas.0506758102

Thanassi, D. G., Nuccio, S.-P., Shu Kin So, S., & Baümler, A. J. (2007). Fimbriae: classification and biochemistry. *EcoSal Plus*, *2*(2). https://doi.org/10.1128/ecosalplus.2.4.2.1

The UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506-D515. https://doi.org/10.1093/nar/gky1049

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M. E., Frapy, E., … Denamur, E. (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genetics*, *5*(1), e1000344. https://doi.org/10.1371/journal.pgen.1000344

Touchon, M., Perrin, A., de Sousa, J. A. M., Vangchhia, B., Burn, S., O'Brien, C. L., Denamur, E., Gordon, D., & Rocha, E. P. (2020). Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLOS Genetics*, *16*(6), e1008866. https://doi.org/10.1371/journal.pgen.1008866

Tourret, J., & Denamur, E. (2016). Population phylogenomics of extraintestinal pathogenic *Escherichia coli*. *Microbiology Spectrum*, *4*(1). https://doi.org/10.1128/microbiolspec.UTI-0010-2012

Vallenet, D., Calteau, A., Dubois, M., Amours, P., Bazin, A., Beuvin, M., Burlot, L., Bussell, X., Fouteau, S., Gautreau, G., Lajus, A., Langlois, J., Planel, R., Roche, D., Rollin, J., Rouy, Z., Sabatet, V., & Médigue, C. (2019). MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz926

van der Mee-Marquet, N. L., Blanc, D. S., Gbaguidi-Haore, H., Dos Santos Borges, S., Viboud, Q., Bertrand, X., & Quentin, R. (2015). Marked increase in incidence for bloodstream infections due to *Escherichia coli*, a side effect of previous antibiotic therapy in the elderly. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.00646

van Hout, D., Verschuuren, T. D., Bruijning-Verhagen, P. C. J., Bosch, T., Schürch, A. C., Willems, R. J. L., Bonten, M. J. M., & Kluytmans, J. A. J. W. (2020). Extended-spectrum beta-lactamase (ESBL)-producing and non-ESBL-producing *Escherichia coli* isolates causing bacteremia in the Netherlands (2014 – 2016) differ in clonal distribution, antimicrobial resistance gene and virulence gene content. *PLOS ONE*, *15*(1), e0227604. https://doi.org/10.1371/journal.pone.0227604

Vieira, G., Sabarly, V., Bourguignon, P.-Y., Durot, M., Le Fevre, F., Mornico, D., Vallenet, D., Bouvet, O., Denamur, E., Schachter, V., & Medigue, C. (2011). Core and panmetabolism in *Escherichia coli*. *Journal of Bacteriology*, *193*(6), 1461-1472. https://doi.org/10.1128/JB.01192-10

Vihta, K.-D., Stoesser, N., Llewelyn, M. J., Quan, T. P., Davies, T., Fawcett, N. J., Dunn, L., Jeffery, K., Butler, C. C., Hayward, G., Andersson, M., Morgan, M., Oakley, S., Mason, A., Hopkins, S., Wyllie, D. H., Crook, D. W., Wilcox, M. H., Johnson, A. P., … Walker, A. S. (2018). Trends over time in *Escherichia coli* bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998–2016: a study of electronic health records. *The Lancet Infectious Diseases*, *18*(10), 1138-1149. https://doi.org/10.1016/S1473-3099(18)30353-0

Vimont, S., Boyd, A., Bleibtreu, A., Bens, M., Goujon, J.-M., Garry, L., Clermont, O., Denamur, E., Arlet, G., & Vandewalle, A. (2012). The CTX-M-15-producing *Escherichia coli* clone O25b: H4-ST131 has high intestine colonization and urinary tract infection abilities. *PLoS ONE*, *7*(9), e46547. https://doi.org/10.1371/journal.pone.0046547

Wales, A. D., Woodward, M. J., & Pearson, G. R. (2005). Attaching-effacing bacteria in animals. *Journal of Comparative Pathology*, *132*(1), 1-26. https://doi.org/10.1016/j.jcpa.2004.09.005

Walk, S. T., Alm, E. W., Gordon, D. M., Ram, J. L., Toranzos, G. A., Tiedje, J. M., & Whittam, T. S. (2009). Cryptic lineages of the genus *Escherichia*. *Applied and Environmental Microbiology*, *75*(20), 6534-6544. https://doi.org/10.1128/AEM.01262-09

Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S.-R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G. F., Rose, D. J., Zhou, S., Schwartz, D. C., Perna, N. T., Mobley, H. L. T., Donnenberg, M. S., & Blattner, F. R. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences*, *99*(26), 17020-17024. https://doi.org/10.1073/pnas.252529799

Whitfield, C. (2006). Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli*. *Annual Review of Biochemistry*, *75*(1), 39-68. https://doi.org/10.1146/annurev.biochem.75.103004.142545

Whittam, T. S., Ochman, H., & Selander, R. K. (1983). Multilocus genetic structure in natural populations of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, *80*(6), 1751-1755. https://doi.org/10.1073/pnas.80.6.1751

Wijetunge, D. S. S., Gongati, S., DebRoy, C., Kim, K. S., Couraud, P. O., Romero, I. A., Weksler, B., & Kariyawasam, S. (2015). Characterizing the pathotype of neonatal meningitis causing *Escherichia coli* (NMEC). *BMC Microbiology*, *15*(1). https://doi.org/10.1186/s12866-015-0547-9

Wiles, T. J., Kulesus, R. R., & Mulvey, M. A. (2008). Origins and virulence mechanisms of uropathogenic *Escherichia coli*. *Experimental and Molecular Pathology*, *85*(1), 11-19. https://doi.org/10.1016/j.yexmp.2008.03.007

Williams, L. E., Wireman, J., Hilliard, V. C., & Summers, A. O. (2013). Large plasmids of *Escherichia coli* and *Salmonella* encode highly diverse arrays of accessory genes on common replicon families. *Plasmid*, *69*(1), 36-48. https://doi.org/10.1016/j.plasmid.2012.08.002

Williamson, D. A., Freeman, J. T., Porter, S., Roberts, S., Wiles, S., Paterson, D. L., & Johnson, J. R. (2013). Clinical and molecular correlates of virulence in *Escherichia coli* causing bloodstream infection following transrectal ultrasound-guided (TRUS) prostate biopsy. *Journal of Antimicrobial Chemotherapy*, *68*(12), 2898-2906. https://doi.org/10.1093/jac/dkt276

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., Karch, H., Reeves, P. R., Maiden, M. C. J., Ochman, H., & Achtman, M. (2006). Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology*, *60*(5), 1136-1151. https://doi.org/10.1111/j.1365-2958.2006.05172.x

Woodford, N., Carattoli, A., Karisik, E., Underwood, A., Ellington, M. J., & Livermore, D. M. (2009). Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major *Escherichia coli* lineages from the United Kingdom, all belonging to the international O25:H4-ST131 clone. *Antimicrobial Agents and Chemotherapy*, *53*(10), 4472-4482. https://doi.org/10.1128/AAC.00688-09

Woodford, N., Ward, M. E., Kaufmann, M. E., Turton, J., Fagan, E. J., James, D., Johnson, A. P., Pike, R., Warner, M., Cheasty, T., Pearson, A., Harry, S., Leach, J. B., Loughrey, A., Lowes, J. A., Warren, R. E., & Livermore, D. M. (2004). Community and hospital spread of *Escherichia coli* producing CTX-M extended-spectrum β-lactamases in the UK. *Journal of Antimicrobial Chemotherapy*, *54*(4), 735-743. https://doi.org/10.1093/jac/dkh424

Yamamoto, S., Nakano, M., Terai, A., Yuri, K., Nakata, K., Nair, G. B., Kurazono, H., & Ogawa, O. (2001). The presence of the virulence island containing the *usp* gene in uropathogenic *Escherichia coli* is associated with urinary tract infection in an experimental mouse model. *The Journal of Urology*, *165*(4), 1347-1351.

Yoon, E.-J., Choi, M. H., Park, Y. S., Lee, H. S., Kim, D., Lee, H., Shin, K. S., Shin, J. H., Uh, Y., Kim, Y. A., Shin, J. H., & Jeong, S. H. (2018). Impact of host-pathogen-treatment tripartite components on early mortality of patients with *Escherichia coli* bloodstream infection: prospective observational study. *EBioMedicine*, *35*, 76-86. https://doi.org/10.1016/j.ebiom.2018.08.029

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, *67*(11), 2640-2644. https://doi.org/10.1093/jac/dks261

Zaw, M., Yamasaki, E., Yamamoto, S., Nair, G., Kawamoto, K., & Kurazono, H. (2013). Uropathogenic specific protein gene, highly distributed in extraintestinal uropathogenic *Escherichia coli*, encodes a new member of H-N-H nuclease superfamily. *Gut Pathogens*, *5*(1), 13. https://doi.org/10.1186/1757-4749-5-13

Zhou, F., & Xu, Y. (2010). cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, *26*(16), 2051-2052. https://doi.org/10.1093/bioinformatics/btq299

Zhou, Z., Alikhan, N.-F., Mohamed, K., Fan, Y., the Agama Study Group, & Achtman, M. (2020). The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Research*, *30*(1), 138-152. https://doi.org/10.1101/gr.251678.119

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Génomique comparée à grande échelle de souches de *Escherichia coli* responsables de bactériémies chez l'Homme : implications cliniques et analyse des réseaux métaboliques

Mots clés : *Escherichia coli*, bactériémie, génomique comparée, réseaux métaboliques

Résumé : *Escherichia coli* est la bactérie aéro-anaérobie facultative majoritaire du tube digestif de l'Homme et, également, l'espèce la plus fréquemment isolée au cours de bactériémies dans les pays industrialisés. La population de *E. coli* présente une diversité importante mais néanmoins structurée, avec certains groupes phylogénétiques préférentiellement associés à un mode de vie (*e.g.* A/B1 et commensal, B2/D et pathogène extraintestinal). De nombreux facteurs de virulence ont été décrits chez les souches pathogènes extraintestinales (ExPEC), mais le pronostic des bactériémies à *E. coli* semble dépendre des facteurs associés à l'hôte. Cependant, certains clones virulents et multirésistants aux antibiotiques ont récemment émergé, modifiant drastiquement l'épidémiologie de ces infections. En parallèle, la démocratisation des méthodes de séquençage offre aujourd'hui une granularité encore jamais atteinte dans l'analyse des génomes bactériens. L'objectif de ce travail de thèse est de tirer profit des méthodes de génomique comparée pour améliorer notre compréhension de la physiopathologie des bactériémies à *E. coli*, tout en prenant en compte les modifications épidémiologiques majeures qui sont observées. La réalisation de telles comparaisons génomiques a nécessité, tout d'abord, la mise en place d'une stratégie d'analyse, incluant notamment le développement d'une approche ciblée pour l'identification des séquences plasmidiques au sein d'assemblages de génomes.

Cette stratégie, appliquée à 545 souches recueillies au cours de l'étude Septicoli en 2016-2017, a permis de confirmer le rôle mineur des déterminants bactériens dans l'issue des bactériémies à *E. coli*. De plus, par la comparaison de ces souches à celles de l'étude Colibafi, nous avons étudié la dynamique de la population sur 12 ans. Si celle-ci apparaît globalement stable, l'exploration plus fine des principaux clones montre d'importants remaniements, parfois associés à des facteurs de virulence typiques des ExPEC. D'autre part, la diversité antigénique de ces clones est variable et suggère des pressions de sélection différentes en fonction de leur niche écologique respective. Enfin, dans une dernière partie nous avons réalisé la reconstruction des réseaux métaboliques de plus de 1400 souches de *E. coli* afin d'étudier les liens entre métabolisme et mode de vie. Les résultats soulignent la conservation du métabolisme chez *E. coli* et son association forte avec la phylogénie. Par ailleurs, une analyse détaillée des principaux clones retrouvés dans les bactériémies, dont notamment le ST131, met en évidence une voie métabolique impliquée dans la dégradation de composés aromatiques dérivés de la lignine. Cette voie, habituellement absente des souches de phylogroupe B2, pourrait procurer un avantage sélectif à ce clone pandémique mondial d'émergence récente.

Title: Large-scale comparative genomics of *Escherichia coli* strains responsible for bacteremia in humans: clinical implications and role of metabolic networks

Keywords: *Escherichia coli,* bacteremia, comparative genomic, metabolic networks

Abstract: Escherichia coli is the predominant aero-anaerobic bacterium of the human gut and also the leading cause of bacteremia in industrialized countries. The E. coli population presents a high but structured diversity, with certain phylogenetic groups preferentially associated with a given lifestyle (e.g. A/B1 and commensal, B2/D and extraintestinal pathogen). Many virulence factors have been described in extraintestinal pathogenic strains (ExPEC), but the prognosis of E. coli bacteremia appears to be linked mainly to host-associated factors. However, some virulent and multi-resistant clones have recently emerged and dramatically modified the epidemiology of these infections. At the same time, the democratization of sequencing methods now offers a granularity yet never reached in the analysis of bacterial genomes. The objective of this thesis is to take advantage of comparative genomic methods to improve our understanding of the physiopathology of E. coli bacteremia, while taking into account the major epidemiological changes that are observed. Carrying out such genomic comparisons required, first of all, the implementation of an analysis strategy, including in particular the development of a targeted approach for the identification of plasmid sequences within genome assemblies.

This strategy, applied on 545 strains collected during the Septicoli study in 2016-2017, confirmed the minor role of bacterial determinants in the outcome of E. coli bacteremia. In addition, by comparing these strains to those of the Colibafi study, we studied the population dynamics over 12 years. While the population appears to be stable overall, further exploration of the main clones shows significant changes, sometimes associated with virulence factors typical of ExPEC. On the other hand, the antigenic diversity of these clones is variable and suggests different selective pressures according to their respective ecological niches. Finally, in a last part we reconstructed the metabolic networks of more than 1400 E. coli strains in order to study the links between metabolism and lifestyle. The results highlight the conservation of metabolism in E. coli and its strong association with phylogeny. In addition, a detailed analysis of the main clones found in bacteremia, including the ST131, highlights a metabolic pathway involved in the degradation of aromatic compounds derived from lignin. This pathway, usually absent in phylogroup B2 strains, could provide a selective advantage to this recently emerging global pandemic clone.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France