Thèse de doctorat

De l'Université Paris 13

Université Sorbonne Paris Cité

Ecole doctorale Galilée (ED 146)

Infection, Antimicrobiens, Modélisation, Evolution (IAME)

UMR 1137, INSERM, Université de Paris et Paris 13

Impact de l'alimentation et de l'effet souche sur l'adaptation de *Escherichia coli* dans son environnement naturel: le tube digestif

Par Mohamed GHALAYINI

Thèse de doctorat en

Science de la Vie et de la Santé

Présentée et soutenue publiquement à Paris le 13 décembre 2019

Membres du jury

Président du jury : Pr Etienne Carbonnelle, IAME, Inserm, Université Paris 13

Directeur de thèse : Dr Olivier Tenaillon, Inserm, Université de Paris et Paris 13

Co-encadrante de thèse : Dr Mathilde Lescat, Inserm, Université Paris 13

Rapporteur : Dr Marianne De Paepe, INRA, Université Paris-Saclay

Rapporteur : Dr Isabelle Kempf, ANSES, Université Bretagne Loire

Examinateur : Dr Philippe Glaser, Institut Pasteur et CNRS, Université Paris Sud

RESUME

Escherichia coli est une bactérie commensale du tube digestif qui, dans certaines conditions dépendantes de la souche et de l'hôte, peut devenir un redoutable pathogène intestinal ou extra-intestinal responsable d'une morbi-mortalité significative. L'espèce *E. coli* possède une structure clonale au sein de laquelle on peut différencier des groupes phylogénétiques. De façon intéressante cette structure est partiellement prédictive de la virulence. En particulier, le groupe phylogénétique B2, est composé de souches souvent isolées de pathologies extra-intestinales mais qui sont aussi de bonnes commensales. En effet, de façon inquiétante, la prévalence de ce groupe augmente dans le tube digestif des individus des pays industrialisés. Cela suggère un potentiel couplage entre des facteurs œuvrant à l'échelle des populations humaines comme l'alimentation et le fond génétique de la souche bactérienne. Plusieurs expériences d'évolution *in vivo* d'*E. coli* dans le tube digestif de souris ont déjà été menées et ont permis de montrer une forte adaptation notamment liée aux ressources énergétiques et au métabolisme de sucres. Ces études souffraient cependant de plusieurs limitations comme l'utilisation d'une souche de laboratoire, d'un milieu non naturel ou encore d'un seul régime alimentaire. L'objectif de ma thèse a donc été d'étudier les modes d'adaptation d'isolats naturels de *E. coli* dans le tube digestif. Cette étude a été effectuée à plusieurs niveaux: à la fois *in natura*, mais aussi *in vivo* en fonction du fond génétique des souches et de l'alimentation de l'hôte.

Dans un premier temps, nous avons eu l'opportunité de suivre l'adaptation *in natura* d'un isolat de *E. coli* ED1a dominant dans le tube digestif d'un individu vivant en région parisienne et sans problème de santé. Nous avons pu observer que *E. coli* ED1a, appartenant au groupe phylogénétique B2, évoluait de façon neutre dans le tube digestif. Le recours aux expérimentations *in vivo* sur modèle murin au long court (plus d'un an) s'est avéré nécessaire pour étudier l'impact de l'alimentation et du fond génétique de la souche dans l'adaptation de *E. coli* dans le tube digestif. Le propos de la première expérimentation était d'étudier l'adaptation d'un isolat naturel de *E. coli* 536 (appartenant au groupe phylogénétique B2 et initialement issu d'infection urinaire), dans un tube digestif de souris se rapprochant le plus possible des conditions naturelles (en utilisant le principe de transmission du microbiote digestif de la mère à la progéniture) et soumises à deux régimes alimentaires : standard ou riche en gras et en sucre. Nous avons pu mettre en évidence une activation constitutive de l'opéron lactose survenant lors de la période d'allaitement des progénitures pour ensuite évoluer de manière neutre quelque soit le régime alimentaire. Le propos de la deuxième expérimentation était d'étudier l'impact du fond génétique de la souche combinée à l'alimentation en étudiant l'évolution de deux isolats naturels de *E. coli* 536 et HS (appartenant au groupe phylogénétique A et issu d'une situation commensale chez un humain) dans le tube digestif de souris traitées à la streptomycine. Nous avons pu distinguer deux profils d'adaptation en fonction de la souche et de l'alimentation. Il semblerait que l'intensité de l'adaptation dépende préférentiellement de la souche et probablement de son fond génétique, et que les cibles de l'adaptation dépendent plutôt du régime alimentaire. Plus précisément, ces cibles semblent principalement être associées au métabolisme de sucres. Ces études constituent un progrès important dans la compréhension des pressions sélectives agissant sur *E. coli in vivo*.

Mots-clés : *Escherichia coli*, groupe phylogénétique, adaptation, expérimentation *in vivo* au long court

ABSTRACT

Escherichia coli is a commensal bacteria of the digestive tract. Under certain strain and host conditions, it may become a serious intestinal or extra-intestinal pathogen responsible for a significant morbidity and mortality. The *E. coli* species has a clonal structure within which phylogenetic groups can be distinguished. Interestingly this structure is partially predictive of virulence. In particular, phylogenetic group B2, is composed of strains often isolated in extra-digestive disease but are also good commensals. Indeed, disturbingly, the prevalence of this group is increasing in the digestive tract of individuals of industrialized countries. This suggests a potential coupling between factors working at the level of human populations such as food and the genetic background of the bacterial strain. Several in vivo evolution experiments of *E. coli* in the digestive tract of mice have already been conducted and allowed to show a strong adaptation particularly related to energy resources and sugar metabolism. These studies, however, suffered from several limitations such as the use of a laboratory strain, in a non-natural environment and subject to a single diet. The objective of my thesis was to study the adaptation of natural isolates of *E. coli* in digestive tract. This study was carried out at several levels: both *in natura*, but also *in vivo* depending on the genetic background of the strains and host diet.

First, we had the opportunity to follow the adaptation *in natura* of a dominant ED1a *E. coli* isolate in the digestive tract of an individual living in the Paris region and without any health problem. We observed that *E. coli* ED1a, belonging to phylogenetic group B2, evolved neutrally in the digestive tract. Subsequently, *in vivo* long-term experiment (over a year) mouse model was used to study the impact of diet and genetic background of strain in the adaptation of *E. coli* to the digestive tract. The purpose of the first experiment was to study the adaptation of a natural isolate of *E. coli*, strain 536 (belonging to the phylogenetic group B2 and initially isolated in urinary tract infection), in a mouse digestive tract as close as possible to the natural conditions (using the principle of transmission of digestive microbiota from mother to offspring) and two diets: standard or high in fat and sugar. We were able to demonstrate a constitutive activation of the lactose operon occurring during the lactation period of offspring and then a neutral evolution regardless of the diet. The purpose of the second mice experiment was to study the impact of the host diet on the evolution of two natural isolates of *E. coli*, strain 536 and strain HS, the latter belonging to the phylogenetic group A, a group known to be mostly commensal in human. We used this time mice treated with streptomycin. We were able to distinguish two adaptation profiles according to the strain and diet. The intensity of adaptation depended on the strain and probably its genetic background, and the target of adaptation depended on the diet. More precisely, the targets of adaptation were mainly associated with the sugar metabolism.

These studies represent an important advance in the understanding of the selective pressures acting on *E. coli in vivo*.

Keywords: *Escherichia coli*, phylogenetic group, adaptation, *in vivo* long-term experimentation evolution

REMERCIEMENTS

TABLE DES MATIERES

LISTE DES TABLEAUX

LISTE DES FIGURES

Figure 8. Functional and genetic support of constitutive activation of *cpx* pathway in WD.

LISTE DES EQUATIONS

INTRODUCTION : PROBLEMATIQUE GENERALE ET OBJECTIFS DE LA THESE

En tant que médecin en réanimation, j'ai été confronté aux pathologies graves tout au long de mon internat puis lors de mes différents postes. Il s'avère que les pathologies infectieuses constituent d'une manière ou d'une autre une part importante de mon activité si ce n'est la majorité. *Escherichia coli* et les entérobactéries en général, sont des pathogènes parmi les plus fréquents dont j'ai pu mesurer toute la gravité durant mon exercice. Il est fascinant de penser qu'*E. coli* qui est susceptible de provoquer une morbi-mortalité telle qu'elle constitue un véritable problème de santé public est, en fait, la plupart du temps un paisible commensal du tube digestif. Évidemment, le commensalisme et la pathogénie de *E. coli* sont liées par le passage de l'un à l'autre comme si le tube digestif constituait un réservoir à pathogène potentiel. Mais la situation semble en réalité plus complexe.

E. coli est une bactérie intégrée dans un système comprenant le reste du microbiote digestif interagissant avec l'épithélium et le système immunitaire de l'hôte. Par ailleurs, le tube digestif est un environnement en interaction avec des éléments extérieurs à l'hôte tels que l'alimentation ou même les antibiotiques. Comment *E. coli* se comporte au vu de tous ces éléments ? Comment cette espèce fait-elle pour persister ? Comment s'adapte-t-elle ? La manifestation la plus criante de son évolution est bien sûr l'émergence de l'antibiorésistance mais il ne s'agit pas du seul facteur auquel doit se confronter *E. coli* dans son milieu naturel. Le perfectionnement et le moindre coût des techniques de séquençage, nous ont permis d'adopter une approche génomique pour répondre à la question de la dynamique d'adaptation de *E. coli* dans son milieu naturel : le tube digestif.

L'objectif global de cette thèse est d'étudier les stratégies déployées par *E. coli* pour s'adapter à son milieu naturel, le rôle de son fond génétique et de l'alimentation dans un contexte commensal. Dans une première partie, nous présenterons d'abord l'état des

connaissances sur *E. coli* intégré au sein d'un système d'interaction avec le reste du microbiote et l'hôte avec lequel elle peut entretenir une relation mutualiste ou pathogène. Nous présenterons aussi l'état des connaissances sur les stratégies adoptées par *E. coli* pour justement s'adapter à son milieu en utilisant une approche génomique. Dans une deuxième partie, nous présenterons les études réalisées au cours de cette thèse : un travail ayant porté sur le suivi *in natura* d'une souche de *E. coli* ED1a prédominante dans le tube digestif d'un homme occidental et d'autre part deux expérimentations *in vivo* de colonisation du tube digestif de souris pendant plus d'un an. Dans la première expérimentation, nous avons tenté de reproduire un environnement le plus naturel possible et suivi un isolat naturel de *E. coli* 536 en fonction du régime alimentaire. Dans la deuxième étude, nous avons comparé l'évolution de deux isolats naturels 536 et HS en fonction du régime alimentaire. Et enfin, dans la troisième partie, nous mettrons en perspective les différents résultats présentés dans cette thèse afin de discuter des implications et proposer des pistes d'explorations pour des travaux ultérieurs.

PARTIE BIBLIOGRAPHIQUE

CHAPITRE I : *E. COLI* INTEGRE ET INTERAGISSANT DANS UN ENVIRONNEMENT :

LE TUBE DIGESTIF

Escherichia coli est une bactérie commensale du tube digestif. Cette espèce est aussi responsable de pathologies aigues digestives ou extra-digestives responsable d'une morbidité importante mais aussi d'une mortalité de l'ordre d'un million de mort par an (Olivier Tenaillon *et al.* 2010a). Dans le tube digestif, il est important de noter qu'*E. coli* n'évolue pas seul mais au sein d'un microbiote complexe lui-même en interaction avec l'hôte, son système immunitaire, son alimentation...

Figure 1. Le microbiote digestif, un sujet d'actualité
Illustration du regain d'intérêt scientifique avec un histogramme du nombre de publications referencés dans pubmed évoquant le microbiote digestif, et du regain d'intérêt grand public par la parution d'articles de journaux ou d'ouvrages de vulgarisation.

L'étude du microbiote digestif semble être la cible d'un regain d'intérêt scientifique depuis le début des années 2000, s'étendant largement au grand public (Figure 1) avec la parution de nombreux ouvrages de vulgarisation sur le sujet. Cet intérêt est lié à l'implication potentielle du microbiote dans des maladies aigues mais aussi chroniques très variées comme des pathologies métaboliques (Ley *et al.* 2005; Bäckhed *et al.* 2007; Seki *et al.* 2007; Miura *et al.* 2010) mais aussi carcinologique (Raisch *et al.* 2014), auto-immunes (Vieira, Pagovich, and Kriegel 2014), inflammatoires digestives (Wu, Bushmanc, and Lewis 2013), et même

neuronales et comportementales comme l'autisme (Li *et al.* 2017). Le microbiote digestif offre potentiellement des pistes de réflexion sur la compréhension physiopathologique de ces maladies et offre surtout la perspective de diagnostics et de stratégies thérapeutiques présentant un coût et une invasibilité réduite (Bibbò *et al.* 2017; Ferrarese *et al.* 2018). Pour appréhender le rôle joué par le microbiote digestif, il convient d'établir sa fonction et les interactions avec son hôte. Nous nous intéresserons dans cette première partie à l'étude du microbiote digestif et de ses interactions avec le système immunitaire et l'épithélium digestif et la place qu'y occupe *E. coli*.

I/Interaction entre le microbiote digestif et l'hôte (épithélium digestif et système immunitaire)

Le microbiote digestif entretient un rôle bénéfique significatif pour l'hôte en remplissant des fonctions métaboliques (production d'énergie pour l'hôte, production de vitamines, fermentations des aliments non digestibles), de protection (production de facteurs antimicrobiens, compétitions avec les pathogènes), de barrière (entretien des jonctions serrées) et immunitaire (développement et maturation). L'accomplissement de ces différentes fonctions est étroitement lié à la composition du microbiote et à son interaction avec l'épithélium digestif et le système immunitaire (Sekirov *et al.* 2010).

1/Les caractéristiques du microbiote digestif

Le microbiote digestif représente une biomasse considérable au sein du corps humain avec près de $3.8x10^{13}$ bactéries, du même ordre que le nombre de cellules dans le corps humain (Sender, Fuchs, and Milo 2016). Initialement peu étudié à cause du caractère peu ou non cultivable de ces bactéries, l'avènement du séquençage haut de débit a permis grâce l'analyse de l'ADN codant pour l'ARN 16S (Fraher, O'Toole, and Quigley 2012) d'avoir une vision quantitative de sa composition avec un coût qui n'a cessé de réduire au court des dernières années. L'ADN 16S est une séquence qui combine des régions très conservées permettant une amplification avec des oligonucléotides faiblement dégénérés et 9 régions variables dont la

séquence est spécifique d'une espèce bactérienne et permet donc son identification après séquençage. Le séquençage étant quantitatif : le nombre de lectures correspondant à une espèce reflète sa fréquence dans l'échantillon, l'analyse du 16S génère donc une image quantitative du microbiote.

En appliquant cette approche à de nombreux individus, une composition récurrente du microbiote digestif retrouvé chez tous les humains (Eckburg *et al.* 2005) a pu être établie. Le microbiote se caractérise par une grande diversité de plusieurs centaines d'espèces bactériennes différentes. Néanmoins, certaines espèces présentes chez la majorité des individus semblent représenter des piliers du microbiote. Deux phyla bactériens représentent la grande majorité des bactéries fécales : le phylum Firmicutes, comprenant exclusivement des bactéries Gram positives, représente en général autour de 51% du microbiote ; le phylum Bacteroidetes comprenant exclusivement des bacilles à Gram négatif anaérobie représente entre 40% et 48% du microbiote. Les phyla Actinobacteria et Proteobacteria, respectivement composés de bactéries Gram positives et de bactéries Gram négatives, arrivent ensuite et sont particulièrement peu représentés (Eckburg *et al.* 2005). La composition en grands groupes bactériens et en espèces dominantes semble relativement stable dans le temps et fait preuve de résilience en cas de perturbation du microbiote (Lozupone *et al.* 2012). Cependant, une dysbiose durable peut d'installer en cas d'antibiothérapie au long court (Vangay *et al.* 2015) ou de changements d'alimentation (Ley *et al.* 2005).

2/Epithélium digestif

L'épithélium intestinal a la délicate fonction d'absorber les nutriments tout en étant une ligne de défense vis-à-vis des agressions réalisées par les bactéries tant pathogènes que commensales qui peuvent coloniser le tube digestif. De façon intéressante, bien qu'étant une menace potentielle, le microbiote lui-même contribue à façonner la composante physique et chimique de cette barrière (Natividad and Verdu 2013). La composante physique est constituée

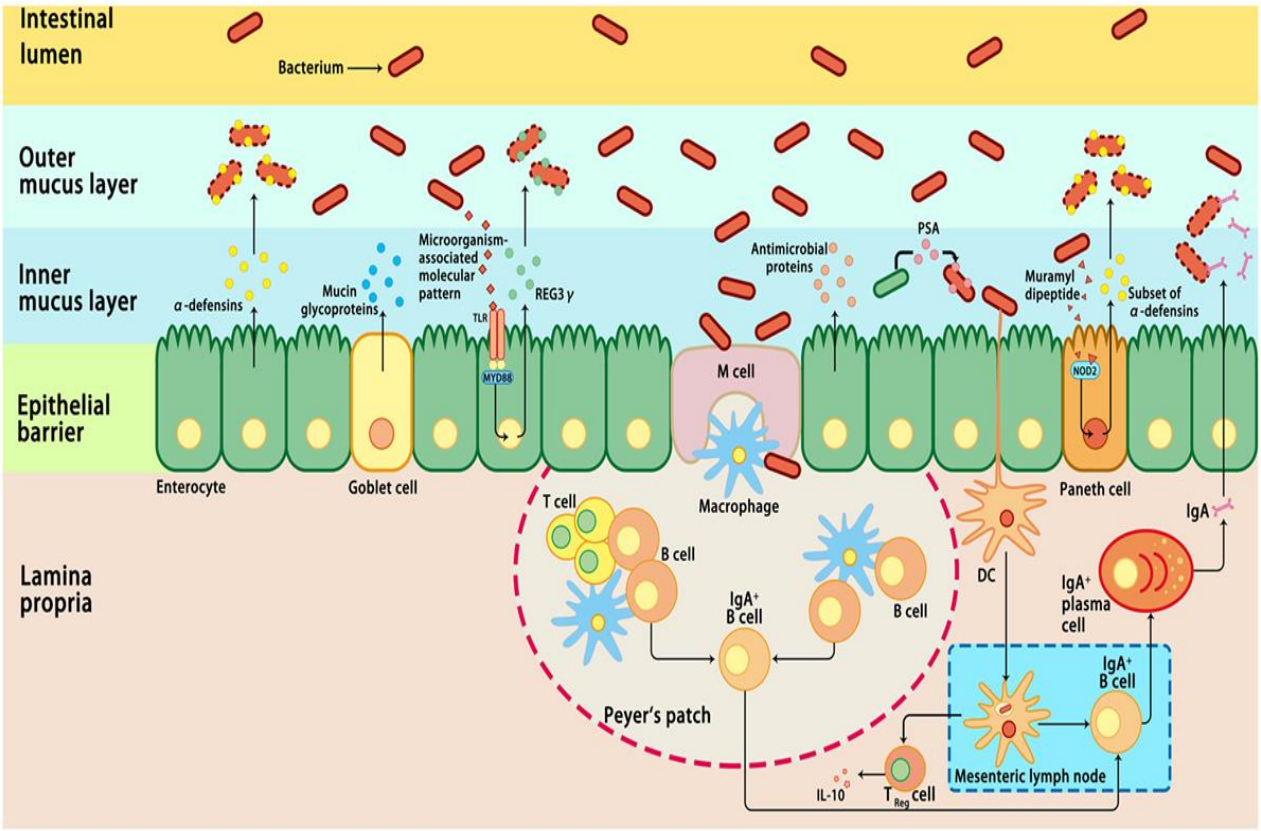
de deux éléments principaux interagissant avec le microbiote : (i) les jonctions serrées qui sont des jonctions étanches entre les cellules épithéliales empêchant la diffusion de molécules et de pathogènes (Powell 1981), et (ii) la couche de mucus qui est fabriquée par les cellules caliciformes (Jakobsson *et al.* 2015). La composante chimique est constituée principalement de molécules antimicrobiennes qui sont synthétisées essentiellement par les cellules épithéliales, dont certaines sont inductibles par la composition du microbiote (Putsep *et al.* 2000), et qui détruisent ou inhibent la croissance microbienne (Figure 2).

3/Système immunitaire

Le système immunitaire, avec l'épithélium digestif, a pour rôle de contenir les bactéries pathogènes et commensales dans la lumière digestive. En retour, le microbiote a pour fonction de maintenir une certaine « tension » dans le système immunitaire, un équilibre entre les forces inflammatoires et anti-inflammatoires (Hooper, Littman, and Macpherson 2012). Le système immunitaire peut se voir sous deux aspects inné et adaptatif.

L'immunité innée, au-delà de la fonction barrière physico-chimique de l'épithélium digestif, peut prendre un aspect de production de peptides antimicrobiens ou d'intermédiaire pour l'immunité adaptative et ainsi jouer un rôle dans la régulation de l'inflammation (M. Zhang *et al.* 2017). La sécrétion de peptides antimicrobiens (α-défensine, REG3γ) est réalisée par des cellules épithéliales ou des cellules spécialisées telles que les cellules de Paneth, de manière constitutive ou induite (Cunliffe and Mahida 2004) via des récepteurs de reconnaissance de motifs bactériens Toll ou NOD2 (Figure 2)(Cario 2005). Les bactéries ou des motifs bactériens (LPS, ADN, …) se lient aux récepteurs de reconnaissance de type Toll ou NOD2 (récepteurs intracellulaires). Elles se trouvent sur les cellules de présentations de l'antigènes (macrophages et cellules dendritiques), les lymphocytes et les cellules épithéliales. Elles provoquent une cascade inflammatoire responsable de l'amplification, la prolifération et la migration des cellules du système immunitaire qu'il soit inné ou adaptatif (Cario 2005).

Après présentation des antigènes par les cellules présentatrices de l'antigène aux lymphocytes T résidents (principalement CD4+) de la lamina propria, les lymphocytes T sont activés. En fonction de l'environnement inflammatoire (notamment la présence de cytokines), les lymphocytes T naïfs prendront un phénotype pro-inflammatoire (ou effecteur) ou anti-inflammatoire (ou régulateur). Il existe une activation des lymphocytes T de manière basale, même en l'absence d'infection. Cette activation est en grande partie sous la dépendance de



<u>Figure 2.</u> Réponses immunitaires de l'hôte au microbiote intestinal.
Plusieurs mécanismes immunitaires agissent de concert sur le microbiote intestinal et contribuent à l'homéostasie intestinale englobant la reconnaissance des bactéries, leur contrôle et la régulation de l'immunité. Les cellules caliciformes (Goblet cell) sécrètent des glycoprotéines de mucine, les plasmocytes sécrètent des IgA, et les cellules épithéliales sécrètent des protéines antimicrobiennes après reconnaissance de motifs bactérien (Microorganism associated molecular pattern, Muramyl dipeptide) par des récepteurs de récepteurs membranaire Toll ou intracellulaire NOD2. Les cellules dendritiques (DC) absorbent les bactéries puis migrent vers les plaques de Peyer et les ganglions mésentériques où les lymphocytes B se différencient en cellules plasmatiques sécrétant de l'IgA. Les lymphocytes T, en fonctions de l'environnement inflammatoire et des informations fournies par les cellules dendritiques, peuvent exprimer un phénotype inflammatoire ou anti-inflammatoire. Le prélèvement du polysaccharide A (PSA) de *Bacteroides fragilis* par les cellules dendritiques intestinales conduit à l'induction de T régulatif (Treg), responsable de la production d'IL-10 régulatrice anti-inflammatoire. **Figure extraite de Zhang et al, 2017.**

bactéries du microbiote intestinal et jouerait un rôle dans le maintien de l'homéostasie intestinale. Il existe notamment un équilibre fin entre les populations Th17 (pro-inflammatoire) et Treg (anti-inflammatoire – régulatrice). Les lymphocytes B activés quelques heures

auparavant au contact de l'antigène terminent à ce niveau leur différenciation en plasmocytes et produisent des immunoglobulines A (IgA) spécifiques de cet antigène. Les IgA sont captées puis internalisées au niveau basolatéral des cellules épithéliales et libérées au pôle apical. Elles tapissent la surface des muqueuses, peuvent capter les antigènes et empêcher leur entrée dans le tissu sous-jacent (Kamada, Seo, *et al.* 2013).

II/Place de *E. coli* dans le microbiote digestif: Mutualiste (ou commensale), Pathogénicité aigue et potentiellement chronique

E. coli est une bactérie versatile pouvant être une commensale du tube digestif des mammifères ou un pathogène redoutable responsable d'une morbi-mortalité importante. Bien que son implication dans la survenue d'infection aigue ne fait plus aucun doute, il ne faut pas oublier qu'il s'agit d'une bactérie jouant un rôle important comme commensale pour l'hôte de par son interaction avec celui-ci (épithélium digestif et système immunitaire) et le reste du microbiote.

1/*E. coli* est un pathogène redoutable

La pathogénicité de *E. coli* peut avoir deux manifestations cliniques : digestive et extra-digestive. La pathogénicité digestive est classiquement regroupée en 6 cadres nosologiques : les entérotoxinogènes, entérohémorragiques, entéro-invasives, entéro-pathogènes, entéro-agrégatives et les souches à adhésion diffuse (Nataro and Kaper 1998). Les souches de pathogénicité digestive ne sont généralement pas des commensales naturelles des humains et leur acquisition par un individu naïf est suffisante pour provoquer la maladie. Elles peuvent cependant être des commensales dans d'autres espèces comme les EHEC qui sont commensales des bovins (Bertin *et al.* 2011). Les manifestations intestinales (diarrhées) dues à *E. coli* sont considérées comme sporadiques dans les pays industrialisés, plutôt associées aux voyages dans des régions à fort risque de transmission oro-fécale (Fletcher, McLaws, and Ellis 2013).

Les manifestations extra-digestives de *E. coli* sont bien plus fréquentes et sont surtout dominées par les infections urinaires avec un impact de santé public évalué de manière importante dans les sociétés occidentales industrialisées en terme de coût et d'incapacité de travail (Foxman 2003). *E. coli* est responsable de près de 74% des infections urinaires en ambulatoire et de 65% dans un contexte nosocomial (Foxman 2010). Le nombre de cas a été estimé entre 6 et 8 millions par an en France (Elkharrat *et al.* 2007). Par ailleurs *E. coli* peut être responsable d'infections intra-abdominales (de Ruiter *et al.* 2009) comme les péritonites, cholécystites, angiocholites, dont la sanction thérapeutique peut être chirurgicale entrainant une morbi-mortalité importante. Spécifiquement le taux de mortalité des péritonites varie en France autour de 15% dans les infections communautaires (Gauzit *et al.* 2009). *E. coli* est aussi une des deux principales causes de méningites néo-natales dont la mortalité est de 14% avec des séquelles neurologiques sévères chez près de 10% des survivants (Stevens *et al.* 2003). Ces différentes pathologies peuvent donner secondairement une bactériémie responsable de 13% de mortalité (Lefort A, Panhard X, Clermont O, Woerther PL, Branger C, Mentré F, Fantin B, Wolff M, Denamur E; COLIBAFI Group 2011) .

Cette bactérie possède déjà un profil pathogène redoutable mais deux points semblent particulièrement inquiétants. (i) Les entérobactéries, *E. coli* en particulier, sont devenues les bactéries les plus fréquemment isolées dans les infections communautaires et nosocomiales chez les patients hospitalisés en unité de soins intensifs et de réanimation (Javaloyas, Garcia-Somoza, and Gudiol 2002; Diekema *et al.* 2003; Laupland *et al.* 2007; Lyytikäinen *et al.* 2002). Notamment, *E. coli* semblerait devenir l'espèce bactérienne la plus régulièrement impliquée dans les pneumopathies infectieuses acquises sous ventilation mécanique dépassant *Pseudomonas aeruginosa* et *Staphylococcus aureus* (Fihman *et al.* 2015; Hamet *et al.* 2012). (ii) Les entérobactéries sont le support de l'émergence de souches résistantes aux antibiotiques dans le monde au point que l'Organisation Mondiale de la Santé les ait figurées comme étant

un objectif critique sur la liste des pathogènes prioritaires pour la recherche et le développement de nouveaux antibiotiques en 2017 (World Health Organization 2017). L'utilisation des antibiotiques de spectres de plus en plus larges a engendré des résistances en retour depuis son début dans le milieu des années 1940. De nos jours, nous assistons à l'émergence de résistances aux carbapénèmes et BetaLactamases à Spectre Elargi (BLSE) pouvant potentiellement engendrer des impasses thérapeutiques (Robinson *et al.* 2016). Les souches de *E. coli* productrices de BLSE de types CTX-M sont fréquemment isolées, avec une prévalence oscillant entre 10% et 60% en fonction des études, et en constante augmentation depuis les années 2000 (Chong, Shimoda, and Shimono 2018). La dissémination spectaculaire des souches productrices de BLSE dans la communauté est associée à l'émergence relativement récente du clone d'*E. coli* ST131 producteur de CTX-M-15, qui porte souvent de nombreux autres gènes de résistance aux antibiotiques (incluant les quinolones)(Chong, Shimoda, and Shimono 2018).

Même si les facteurs liés à l'hôte sont déterminants (âge, comorbidité, statut immunitaire) (Lefort A, Panhard X, Clermont O, Woerther PL, Branger C, Mentré F, Fantin B, Wolff M, Denamur E; COLIBAFI Group 2011) pour la pathogénicité de *E. coli*, celle-ci dépend aussi en grande partie de facteurs génétiques liés à la bactérie. *E. coli* a une structure clonale permettant de différencier plusieurs groupes phylogénétiques (Picard *et al.* 1999). Cette différenciation n'est pas anodine car il existe un lien entre le groupe phylogénétique et la pathogénicité de la souche (Picard *et al.* 1999). Par exemple, les souches isolées dans des situations pathogènes extra-intestinales appartiennent principalement au groupe phylogénétique B2 alors que les souches de *E. coli* impliquées dans les pathologies intra-intestinales appartiennent principalement aux groupes A, et B1. De plus, les *E. coli* du groupe phylogénétique B2 sont porteurs de facteurs de virulence extra intestinales dans des îlots de pathogénicité. La délétion de ces îlots se caractérisant par une baisse de colonisation de la niche

commensale (Diard *et al.* 2010), il est probable que ces facteurs de virulence soient en fait sélectionnés pour leur rôle dans les environnements commensaux (Bingen-Bidois *et al.* 2002) et que leur impact sur la pathogénèse soit coïncident. Le groupe phylogénétique B2 abrite de nombreuses souches virulentes, et il est en conséquence des plus inquiétant de le voir émerger en commensal du tube digestif des individus habitant dans les pays occidentaux industrialisés (Olivier Tenaillon *et al.* 2010a)(Figure 3a) depuis ces 30 dernières années (Massot *et al.* 2016a)(Figure 3b).

Ce changement épidémiologique, couplé à l'émergence de souches multirésistantes aux antibiotiques comme ST131 appartenant groupe phylogénétique B2 contribue probablement en partie à l'augmentation des pathologies liées à *E. coli* observée dans les pays industrialisés (Nicolas-Chanoine, Bertrand, and Madec 2014).

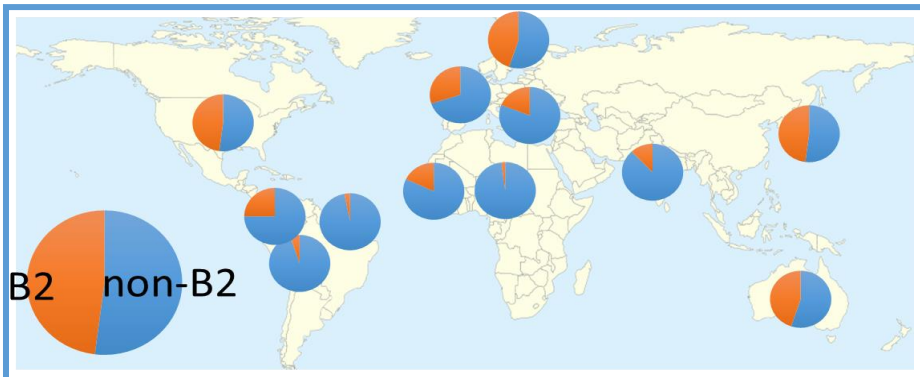


Figure 3a. Distribution de la population de *E. coli* commensal du tube digestif en fonction de leur appartenance au groupe phylogénétique B2

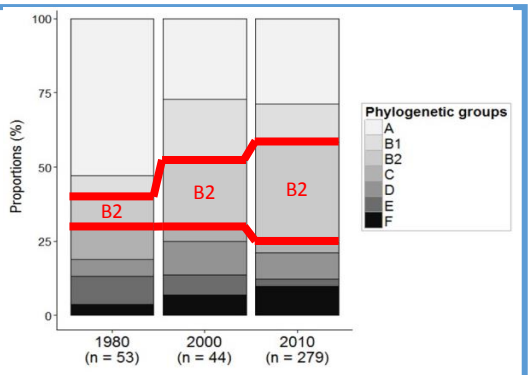


Figure 3b. Evolution dans le temps de la proportion de *E. coli* B2 commensale du tube digestif à Paris, adapté de Massot *et al.* 2016

2/Rôle mutualiste et implication dans les pathologies

2-1/*E. coli* : premier colonisateur du tube digestif

Le nouveau-né, stérile *in utero*, se trouve à la naissance brutalement plongé dans un univers bactérien riche et est rapidement colonisé par une flore simple à partir des flores de sa

mère et de l'environnement proche. La flore fécale maternelle apparaît être le déterminant essentiel des premières bactéries s'implantant chez l'enfant, les nouveau-nés étant colonisés plutôt par les entérobactéries et bifidobactéries, d'origine fécale, que par les lactobacilles, d'origine vaginale (Tannock *et al.* 1990). Le nouveau-né est ensuite continuellement exposé à de nouvelles bactéries provenant de l'environnement, de la nourriture et des bactéries cutanées des adultes via les tétées, les caresses ou les baisers. Une flore complexe et stable, proche de celle de l'adulte, ne semble être obtenue qu'entre 2 et 4 ans. Chez l'enfant à terme, les premières bactéries implantées sont des organismes aérobies-anaérobies facultatifs : les entérobactéries (principalement l'espèce *E. coli*), les entérocoques et les staphylocoques (Ducluzeau 1993; Hill *et al.* 2017). Les bactéries aérobies-anaérobies facultatives, dont le niveau atteint rapidement 10^9 à 10^{10} unité formant des colonies (UFC)/g de contenu colique, consomment l'oxygène, diminuant ainsi le potentiel redox de la lumière du tube digestif, ce qui permet l'implantation des genres anaérobies stricts (*Bifidobacterium, Clostridium, Bacteroides*) ainsi que celle des lactobacilles.

L'implantation des bactéries semble une étape cruciale conditionnant la santé de l'enfant et son développement vers l'âge adulte (Al Nabhani *et al.* 2019a). L'impact de l'implantation du microbiote digestif a été largement suspecté en individualisant la césarienne (Rutayisire *et al.* 2016) ou le mode d'allaitement (Gregory *et al.* 2016), principaux facteurs conditionnant la composition du microbiote dans les premiers instants de la vie, comme facteur de risque pour les enfants de développer des allergies (Bager, Wohlfahrt, and Westergaard 2008; Thavagnanam *et al.* 2008; Kull *et al.* 2002; N. J. Friedman and Zeiger 2005) ou des pathologies métaboliques comme l'obésité infantile (Armstrong, Reilly, and Child Health Information Team 2002; Mueller *et al.* 2015) ou le développement d'un diabète de type 2 (Owen *et al.* 2006). L'interaction entre le microbiote, l'épithélium digestif et le système immunitaire semble être la cause du développement de ces maladies chez l'enfant. Le changement dans le

microbiote intestinal au moment du sevrage, secondaire à une antibiothérapie (Al Nabhani *et al.* 2019a) ou un apport calorique excessif chez les souris nouveau-nées (Al Nabhani, *et al.* 2019b, publication en cours de révision favorable), induit une réponse immunitaire intense associé à l'induction de lymphocytes T régulateurs et sa perturbation entraîne une susceptibilité accrue aux pathologies à composante immunologique plus tard dans la vie.

2-2/*E. coli* commensale oscillant entre mutualisme et pathologie chronique

E. coli est une bactérie commensale dont l'habitat primaire est le tube digestif des vertébrés (Gordon and Cowling 2003) mais qui est aussi retrouvée dans un habitat secondaire comme le sol (Olivier Tenaillon *et al.* 2010). Cette bactérie est loin d'être passive, il serait plus juste de la qualifier de bactérie mutualiste. Effectivement, si *E. coli* bénéficie une niche de colonisation et des ressources de son envirronement, elle joue aussi un rôle actif. *E. coli* possède des (i) fonctions métaboliques utiles pour l'hôte comme la synthèse de la vitamine K (Bentley and Meganathan 1982), la fermentation des résidus alimentaires non digestibles, la production d'énergie (Niba *et al.* 2009) et (ii) des fonctions de protection en participant à la compétition avec les pathogènes ou la production de facteurs antimicrobiens (Salomón and Farías 1992).

Un faisceau d'arguments suggère que le microbiote, et les entérobactéries en particulier, sont potentiellement impliquées dans la survenue de pathologies chroniques et notamment la survenue de maladies métaboliques comme l'obésité, le diabète de type 2 ou même la stéatose hépatique non alcoolique (NASH). L'observation que les souris axéniques soient résistantes au gain de poids corporel et à l'accumulation de masse grasse induits par un régime alimentaire riche en gras et en sucre suggère que le microbiote intestinal est une des clés permettant le stockage des graisses dans les adipocytes (Bäckhed *et al.* 2007; Rabot *et al.* 2010).

Le changement de microbiote contribuerait à la perturbation des protéines de jonctions serrées (Zonula Occludens et Occludin) impliquées dans la fonction barrière de l'intestin. Cet effet dépend directement du microbiote intestinal car le traitement antibiotique abolit la

perméabilité intestinale induite par le régime alimentaire (P. D. Cani *et al.* 2009; Patrice D. Cani *et al.* 2008). Cette perméabilité digestive permettrait le passage d'éléments bactériens tels que le lipopolyssacharide (LPS), un des constituants de l'enveloppe de membrane externe des bactéries Gram négatif. Différentes études ont mis en évidence une inflammation chronique de bas grade caractérisé par l'infiltration de macrophages dans les adipocytes (Weisberg *et al.* 2003) et des cellules de Kupffer (Seki *et al.* 2007; Miura *et al.* 2010) dans le foie associé à une activation de récepteurs membranaires de reconnaissance de motifs moléculaires, les récepteurs de type Toll, par cette endotoxémie par le LPS.

Quels types de modifications dans le microbiote sont susceptibles de provoquer cette perturbation métabolique ? L'alimentation riche en gras et en sucre provoque une modification du microbiote digestif (Ley *et al.* 2005) vers une augmentation du ratio Firmicutes/Bacteroidetes (Ley *et al.* 2005) et une augmentation des Protéobactéries (Shin, Whon, and Bae 2015). Il est difficle de supposer que *E. coli* puisse avoir un impact déterminant alors qu'elle ne représente que moins de 1% du microbiote (Kasai *et al.* 2016). En utilisant une technique de micro-dissection des tissus au laser et le séquençage de l'ADN 16S, il s'avère que les Protéobactéries, dont fait partie *E. coli*, représente 47,6% en moyenne du microbiote des cryptes digestives contre 2,7% au niveau de la lumière intestinale dans un modèle murin (Pédron *et al.* 2012). Cela suggère que les Protéobactéries, au plus près de la muqueuse digestive, sont les plus susceptibles d'interagir avec l'épithélium digestif et le système immunitaire local. Par ailleurs, les souris recevant un régime riche en graisses et recevant des antibiotiques, touchant surtout les entérobactéries (ampicilline et néomycine), présentaient une réduction de l'endotoxémie métabolique, de l'inflammation, de la résistance à l'insuline et de la prise de poids (Patrice D. Cani *et al.* 2008).

Certaines souches d*'E. coli* semblent aussi impliquées dans l'émergence du cancer du côlon en contribuant activement à l'accumulation de mutations résultants de lésions de l'ADN

induites par les génotoxines codées par l'ilot génomique *pks*, ou en dérégulant la capacité de réparation de l'ADN de l'hôte (Arthur *et al.* 2012; Maddocks *et al.* 2009; Nougayrède *et al.* 2006). Il s'avère que le groupe phylogénétique B2 a une prévalence plus élevé chez les patients atteints d'un cancer du côlon plus susceptible de persister, de provoquer une inflammation chronique de bas grade et une prolifération cellulaire, mettant en avant son potentiel oncogène (Raisch *et al.* 2014).

CONCLUSION DU CHAPITRE I

E. coli est un organisme de première importance au vu de son impact en terme de problème de santé public majeur, non seulement à cause des pathologies infectieuses digestives et extra-digestives mais potentiellement aussi pour son rôle dans la survenue de maladies métaboliques ou carcinologiques. A cet égard, l'étude du groupe phylogénétique B2 est intéressante d'autant plus que l'on observe son émergence dans les pays occidentaux industrialisés en rapport avec le mode de vie, en particulier l'hygiène et l'alimentation. Quels sont les déterminants de la colonisation de *E. coli* dans le tube digestif ? Quels sont les déterminants permettant l'émergence de ce phylogroupe ? Quelles stratégies adopte *E. coli* pour persister dans le tube digestif ? Quel impact a le fond génétique de la souche et l'alimentation dans la persistance de *E. coli* dans le tube digestif ? Les progrès du séquençage et son faible coût nous ont permis d'adopter une approche génomique nous permettant de déterminer les stratégies d'adaptations en fonction du groupe phylogénétique de *E. coli* et du régime alimentaire.

CHAPITRE II : *E. COLI*, MECANISMES ET STRATEGIES D'ADAPTATIONS

I/Les processus à la base de l'évolution

L'évolution est un processus continu conduisant à la prédominance ou la diminution de fréquence d'un trait phénotypique au sein d'une population permettant sa persistance dans un

environnement donné. C'est l'expression des allèles (correspondant à des versions d'un ou plusieurs gènes), qui est à l'origine d'une certaine variabilité phénotypique. L'émergence d'un allèle au sein d'une population dépend de processus basiques telles que la mutation, la dérive génétique, la migration, les recombinaisons ou la sélection naturelle (Olson-Manning, Wagner, and Mitchell-Olds 2012).

1/Les mutations

Une mutation est une modification aléatoire rare, accidentelle ou provoquée, de l'information génétique dans le génome. Selon la partie du génome touchée, les conséquences d'une mutation peuvent varier (Loewe and Hill 2010). Elle est un élément majeur de la diversité de par la création de nouveaux allèles et l'un des nombreux facteurs pouvant éventuellement participer dans l'évolution de l'espèce.

Les mutations ponctuelles sont les types de mutations les plus simples et parmi les plus répandues. Une mutation ponctuelle se traduit par le remplacement d'un nucléotide par un autre engendrant la modification du codon (mutations non-synonymes) ou non de par la redondance du code génétique (mutations synonymes). Les « Indels » est un mot générique pour désigner une modification par un gain ou une perte de matériel génétique par rapport à une séquence de référence. Le terme « indel » a été introduit parce que la notion d'insertion ou de délétion est relative suivant le choix de la séquence utilisée comme référence : à une insertion dans une séquence, correspond une délétion dans la séquence qui lui est comparée. « Indel » permet ainsi de désigner globalement la variation biologique, sans préjuger de quelle séquence constitue la référence. Dans le cas des insertions, on peut distinguer plusieurs mécanismes dont la duplication d'une séquence de nucléotides, une inversion de nucléotides ou une insertion d'un élément transposable. Un élément transposable est une séquence d'ADN de provenance diverse (élément génétique provenant du même génome, d'un génome différent comme une autre

bactérie ou un bactériophage) qui a la capacité de se transposer d'un endroit à un autre dans le génome.

2/La sélection naturelle

La sélection naturelle, fruit de facteurs environnementaux, est un phénomène crucial de l'évolution permettant de modifier la fréquence des allèles en fonction de leur impact sur le phénotype et de leurs conséquences sur la survie et la reproduction (fitness ou valeur sélective). En effet, les allèles délétères seront éliminés par une sélection négative ou purificatrice, alors que les allèles bénéfiques bénéficieront d'une sélection positive et d'une augmentation de leur fréquence dans la population (Loewe and Hill 2010). Nous pouvons distinguer plusieurs régimes de sélection en fonction de la diversité présente dans la population et l'intensité des différentiels de valeur sélective (Barrick and Lenski 2013)(Figure 4).

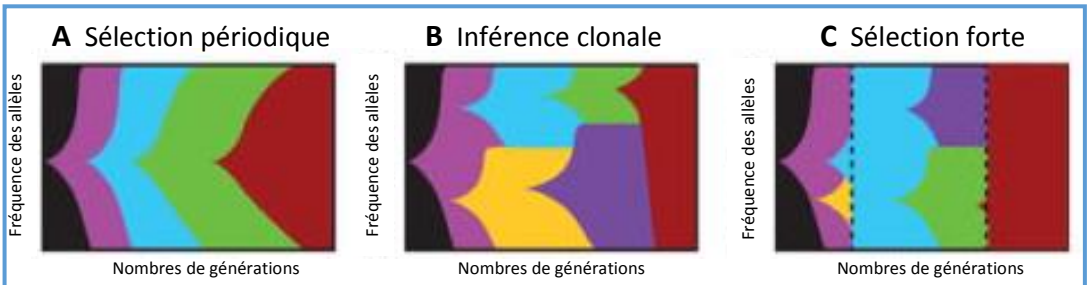


Figure 4. Dynamique génétique des différents modes de sélection.
Les fréquences des génotypes dans le temps sont représentés par des plages de couleurs. (A) Dans la sélection périodique, une mutation bénéfique se fixe périodiquement balayant le reste de la population. (B) Dans l'inférence clonale, plusieurs mutations bénéfiques apparaissent et entrent en compétition pour la fixation. (C) Dans la sélection forte, une mutation bénéfique émerge par un goulot d'étranglement (effet bottleneck) suite à un phénomène environnemental majeur (en pointillé). Figure adaptée de Barrick and Lenski, 2013.

La sélection périodique intervient si le taux d'apparition de nouvelles mutations bénéfiques est faible et que les avantages liés à chacune de ces mutations sont importants. Dans ces conditions, une seule mutation à la fois envahira la population, on parle de fixation d'une mutation. Ces dynamiques entraînent des modifications par à-coup des traits phénotypiques qui s'accumulent avec le temps. Les processus de fixation, appelés balayages sélectifs prennent

généralement de plus en plus de temps car l'avantage marginal escompté d'une mutation ultérieure diminue si l'évolution se situe dans le régime optimal (Figure 4A).

L'interférence clonale intervient si le taux d'apparition de mutations bénéfiques est supérieur, du fait que la taille de la population ou le taux de mutation global est augmenté, de multiples mutations bénéfiques peuvent alors se produire avant que l'une d'elles n'atteigne la fixation. Dans les populations asexuées, la compétition entre les mutations en conflit ralentit leur progression vers la fixation, laissant le temps à de nouvelles mutations bénéfiques d'apparaitre et de donner lieu à des trajectoires plus complexes (Figure 4B). Dans ce régime, l'effet intrinsèque de la mutation dès qu'elle est suffisamment avantageuse peut importer moins que celui des mutations auxquelles elle se trouvera associée par chance.

Alors que les deux premiers modes de sélection dépendaient principalement de la diversité offerte par la population, la sélection forte dépend de l'intensité des facteurs environnementaux s'exerçant sur la population. Si une sélection forte est imposée périodiquement, d'une manière qui peut même être mortelle pour la plupart de la population, seul un ou plusieurs génotypes peuvent subsister et ils peuvent ensuite obtenir rapidement une fixation après ce goulot d'étranglement induit par la sélection. Ce scénario peut entraîner des changements importants et soudains dans un phénotype, tels que la résistance à un antibiotique (Q. Zhang *et al.* 2011) (Figure 4C).

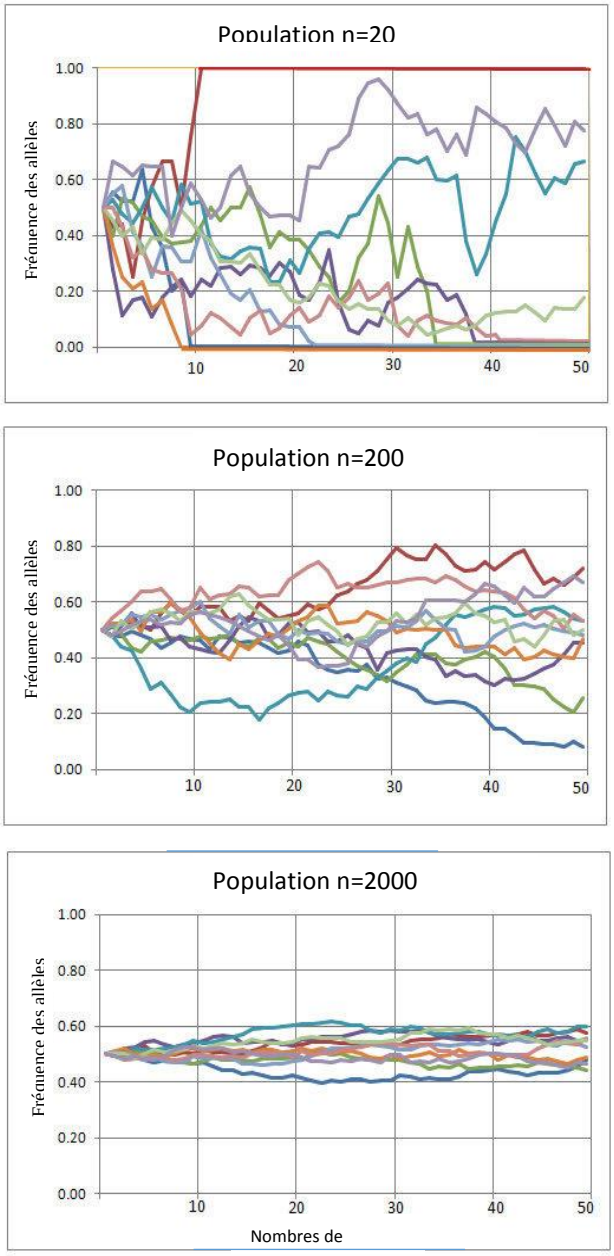
3/La dérive génétique

L'oscillation de la fréquence dans le temps des allèles créés par mutation peut se produire en dehors de la sélection naturelle. La dérive génétique est le deuxième mécanisme majeur régissant l'émergence ou la disparition des allèles dans une population.

La théorie neutre de l'évolution moléculaire, établi par un biologiste japonais Motoo Kimura, stipule que la plupart des changements évolutifs au niveau moléculaire, et la plupart des variations au sein des espèces et entre espèces, sont dus à la dérive génétique aléatoire des

allèles qui sont sélectivement neutres (Kimura and Ota 1969). Selon cette théorie, les mutations (créant de nouveaux allèles) délétères sont rapidement éliminées par la sélection naturelle et les autres mutations sont principalement neutres plutôt que bénéfiques. Ainsi, la théorie de l'évolution neutre suggère qu'un allèle mutant peut émerger au sein d'une population et atteindre la fixation par hasard plutôt que par avantage sélectif (Kimura 1979).

Dans un contexte d'évolution neutre, la probabilité qu'un allèle se fixe dans une population dépend principalement de sa fréquence initiale qui est d'autant plus grande que la population est petite (Charlesworth 2009) car un mutant apparaît initialement avec une fréquence de 1/N (N étant la taille de la population). Par ailleurs, d'une génération à une autre, les changement de fréquence sont d'autant plus grand que la population est petite (figure 5) (Otto and Whitlock 1997). D'ailleurs, de grands changements de fréquence peuvent s'observer lors d'une réduction importante de la taille de la population (effet goulot d'étranglement, « bottleneck effect ») (Nei, Maruyama, and Chakraborty 1975) ou l'isolation d'une partie de celle-ci (effet des fondateurs, « founder effect ») conduisant à une réduction de sa diversité allélique (Matute 2013).



<u>Figure 5,</u> Effet de la taille de la population sur la dérive génétique. Trois simulations d'un changement aléatoire de la distribution de fréquence d'un seul allèle hypothétique sur 50 générations pour des populations de tailles différentes; première population n = 20, deuxième population n = 200 et troisième population n = 2000. Illustration adaptée du travail de thèse *de* Kryvokhyzha Dmytro, *2018.*

Loin d'être incompatible avec la théorie de l'évolution Darwinienne qui consacre la sélection naturelle comme facteur de modification de proportion des allèles dans une population, la théorie d'évolution neutre de Kimura permet de la compléter et d'expliquer des phénomènes de fixation de certains allèles en dehors de la sélection. Il existe encore d'autres phénomènes permettant d'expliquer l'émergence d'un allèle au sein d'une population.

4/Autres mécanismes contribuant au flux d'allèles et aux changements de fréquence: la recombinaison et la migration

Si la sélection naturelle et la dérive génétique sont les deux principaux processus à l'origine des changements des fréquences alléliques dans une population isolée, l'arrivée de nouveaux individus dans une population par migration ou de nouveau matériel génétique par recombinaison peut aussi jouer un rôle important dans la diversification des populations.

La migration est l'introduction de nouveaux individus dans la population. Ceux-ci peuvent porter de nouveaux allèles ou porter des allèles présents dans la population focale, mais avec une fréquence différente. Leur introduction dans la population peut en conséquence entrainer des changements rapides des fréquences alléliques qui sont à cumuler aux effets sur les fréquences liées à la sélection ou la dérive génétique.

La recombinaison, notamment chez les bactéries, peut permettre l'introduction de matériel génétique nouveau au sein d'une population. Trois mécanismes peuvent être impliqués. (i) La conjugaison est un phénomène observé chez les bactéries qui aboutit à l'échange de matériel génétique. Elle consiste en une transmission par l'intermédiaire du pilus, d'ADN plasmidique ou d'ADN chromosomique d'une bactérie donneuse, (porteuse de plasmide) à une bactérie receveuse et, potentiellement, en son intégration dans le génome de celle-ci. (ii) La transduction est un processus qui consiste en un transfert de matériel génétique (ADN bactérien), d'une bactérie donneuse à une bactérie receveuse, par l'intermédiaire d'un vecteur viral (un bactériophage). Enfin (iii) la transformation est l'altération génétique d'une cellule résultant de l'absorption directe et de l'incorporation de matériel génétique exogène provenant de son environnement à travers la ou les membranes cellulaires. Pour que la transformation se produise, les bactéries réceptrices doivent être dans un état de compétence, ce qui peut se produire dans la nature en tant que réponse limitée dans le temps à des conditions environnementales telles que la famine et la densité cellulaire (Johnston *et al.* 2014), et peut également être induit en laboratoire (Datsenko and Wanner 2000). Tous ces mécanismes

influencent l'évolution bactérienne de deux façons. Tout d'abord ces transferts peuvent avoir lieu entre espèce et aboutir à l'incorporation d'une diversité génique importante au sein de la population. Deuxièmement, au sein de la population ces échanges génèrent des nouvelles combinaisons d'allèles qui peuvent influencer le processus évolutif.

II/Les enseignements des expérimentations évolutives

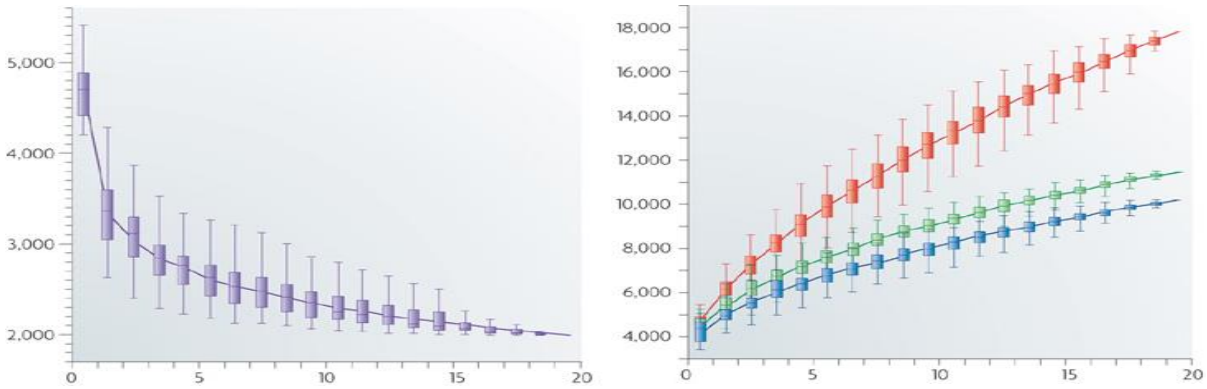
Comme nous venons de le voir, l'évolution est un processus de transformation du génome au fil des générations. Au cours des dernières décennies, l'étude de l'évolution microbienne a permis de mieux caractériser ce processus *in vitro*, *in vivo* et l'observation *in natura*. Au sein de ces études, l'espèce modèle *E. coli* a joué un rôle central.

1/Structure clonale et plasticité du génome de *E. coli*

Comme nous l'avons vu, *E. coli* est avant tout un organisme d'intérêt médical. En plus de cela, la facilité d'isolation et de mise en culture, son temps de génération court et la connaissance approfondie de son métabolisme en font un organisme de choix pour l'étude de l'adaptation. La connaissance de sa structure génétique et la plasticité de *E. coli* nous semble un préambule nécessaire à l'étude de son adaptation.

Afin de comprendre comment le génome de *E. coli* évolue il est important de connaître sa composition et sa structure à l'échelle de l'espèce. L'information génétique de *E. coli* repose un chromosome dont la taille est comprise entre 4.000.000 et 5.500.000 paires de bases (pb) pouvant porter autour de 5000 gènes (Bergthorsson and Ochman 1998). *E. coli* est une espèce très diversifiée pouvant adopter une grande variabilité de combinaisons de gènes prises à partir d'une liste de génomes différents, et dont l'ensemble des gènes constitue le pan génome. Le séquençage de 20, 61 puis 491 génomes de *E. coli* a permis de montrer que le pan-génome augmentait avec le nombre de génomes étudiés atteignait 43.415 gènes (Touchon *et al.* 2009; Yang *et al.* 2019; Lukjancenko, Wassenaar, and Ussery 2010)(Figure 6). Une analyse en cours

au laboratoire sur plus de 60.000 génomes suggère même que le pan génome se composerait de plus de 250 000 gènes.



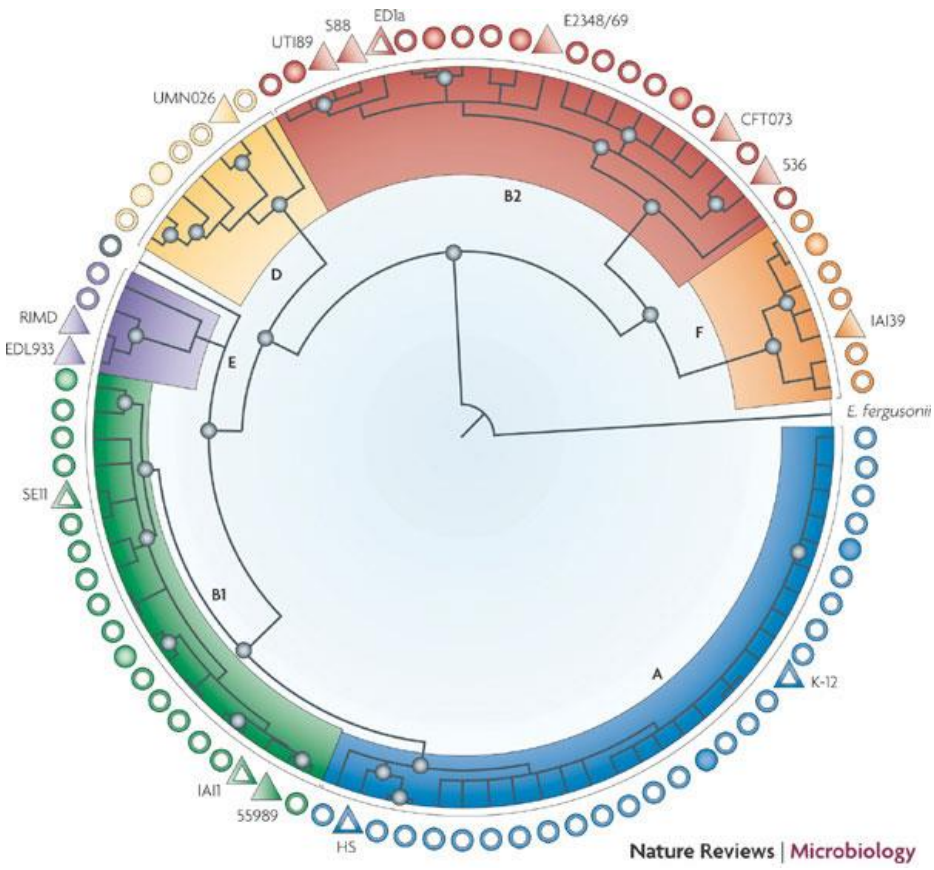
<u>Figure 6.</u> Evolution du « core » (à gauche) et « pan » génome (à droite) de *E. coli* en fonction du nombre de génomes séquencés.

A gauche évolution du nombre de gènes présents chez toutes les souches. A droite, évolution du nombre de gènes du « pan » génome. En rouge, tous les gènes (17838 gènes), en bleu lorsque les gènes présentent plus de 80% d'homologie (11432 gènes) et en vert lorsqu'on retire les séquences d'insertions et les séquences de prophages (10131 gènes). Figure adaptée de Touchon *et al.*, Plos one, 2009.

Le pan génome est donc une liste de gènes susceptibles de figurer dans le génome de *E. coli*. On peut distinguer deux fractions du génome en fonction de la conservation de celui-ci au sein de l'espèce. (i) Le génome variable est constitué de l'ensemble des gènes qui ne sont pas communs aux génomes étudiés et est le théâtre d'un flux intense d'acquisition et de perte de gènes. L'acquisition d'un nouveau matériel génétique se fait par transfert horizontal entre bactéries d'une même espèce ou non. Cette acquisition chez *E. coli* peut se faire par conjugaison (transfert d'un élément génétique mobile, un plasmide codant ses propres éléments permettant le transfert, via des pili) ou par transduction (Transfert de matériels génétiques via un bactériophage). Ces gènes « accessoires », acquis à partir de génomes de bactéries environnantes, semble en partie conférer un avantage sélectif telles que l'adaptation à une niche écologique, la capacité à coloniser un hôte spécifique ou la résistance aux antibiotiques (Yang *et al.* 2019). (ii) Le core génome est défini par la liste de gènes partagés entre quasiment toutes les souches de *E. coli,* en rapport avec un mode de reproduction clonal. Le caractère clonal de *E. coli* a pu être déterminé assez précocement (Ochman and Selander 1984), servant de base

aux techniques d'études génétiques de la population de *E. coli* utilisant des gènes conservés entre souches de *E. coli*, tels que le sérotypage (identification basée sur les combinaisons des antigènes O, K et H)(Orskov *et al.* 1976), le *multilocus enzyme electrophoresis* (identification basée sur la différence de migrations de plusieurs enzymes)(Selander *et al.* 1986), le *multilocus sequence typing* (identification d'un profil allélique spécifique d'une population sur un ensemble de fragments d'ADN)(Enright and Spratt 1999), la PCR multiplex Clermont (identification basée sur la présence ou l'absence de gènes amplifiés par PCR spécifique d'une population)(Clermont, Gordon, and Denamur 2015) et enfin le séquençage haut débit (Touchon *et al.* 2009). Le core génome est composé d'environ 2000 gènes (nombre légèrement variable en fonction du nombre de souches séquencées et du critère retenu (présent dans 100% des souches ou 95%) occupant ainsi moins de la moitié du génome de *E. coli* (Touchon *et al.* 2009; Yang *et al.* 2019). Le core génome est lui aussi affecté par la recombinaison mais cette fois il s'agit de recombinaison homologue qui entraine la création de nouvelles combinaisons d'allèles le long du chromosome. Ce processus pourrait aussi être impliqué dans l'adaptation d'une souche à son environnement (Rodríguez-Beltrán *et al.* 2015). Si l'existence de recombinaison est peu compatible avec la clonalité de l'espèce, il semble que pour *E. coli*, la recombinaison ne soit pas suffisamment abondante pour brouiller massivement la structure clonale de l'espèce. Ainsi dès qu'une région suffisamment grande du génome est séquencée un signal phylogénétique robuste permettant de déterminer une structure à la population émerge. Ainsi, les clones qui composent l'espèce se regroupent dans des structures plus larges, les groupes phylogénétiques qui comptent 6 groupes principaux A, B1,B2, D, E et F (Olivier Tenaillon *et al.* 2012a) (Figure 7) ainsi que quelques autres groupes marginaux (Clermont *et al.* 2013) et dont des études épidémiologiques ont montré qu'ils présentaient un certain degré de spécialisation en termes d'écologie et de pouvoir pathogène. Ainsi, les souches du groupe

phylogénétiques B2 sont plus fréquemment impliquées dans les infections extra intestinales que celles des autres groupes (Picard *et al.* 1999; Escobar-Páramo *et al.* 2004).



<u>Figure 7.</u> Arbre phylogénétique de l'espèce *E. coli* réalisé à partir du séquençage de 8 gènes de ménage, soit 4095 nucléotides au total. Soixante-douze souches de la collection ECOR (cercles) et 15 souches dont le génome a été séquencé (triangles) ont été intégrées dans l'analyse. Les souches commensales sont représentées par un symbole vide, les pathogènes par un symbole plein. Les couleurs indiquent les groupes phylogénétiques principaux. Les ronds gris sur les nœuds signifient que le nœud est monophylétique à plus de 80%. Figure adaptée de Tenaillon *et al.,* Nat Rev Microbiol, 2010.

L'analyse des génomes peut donc se faire à partir de l'étude de la diversité présente au sein du core génome, ou à partir de la composition en gènes. Si les deux mesures (distance sur le core, ou distance en composition génique) sont corrélées, il est possible d'avoir des génomes quasiment identiques au niveau du core génome dont la composition en gène varie de plus de 20% (Figure 8). Des dynamiques évolutives complexes sont donc à l'œuvre.

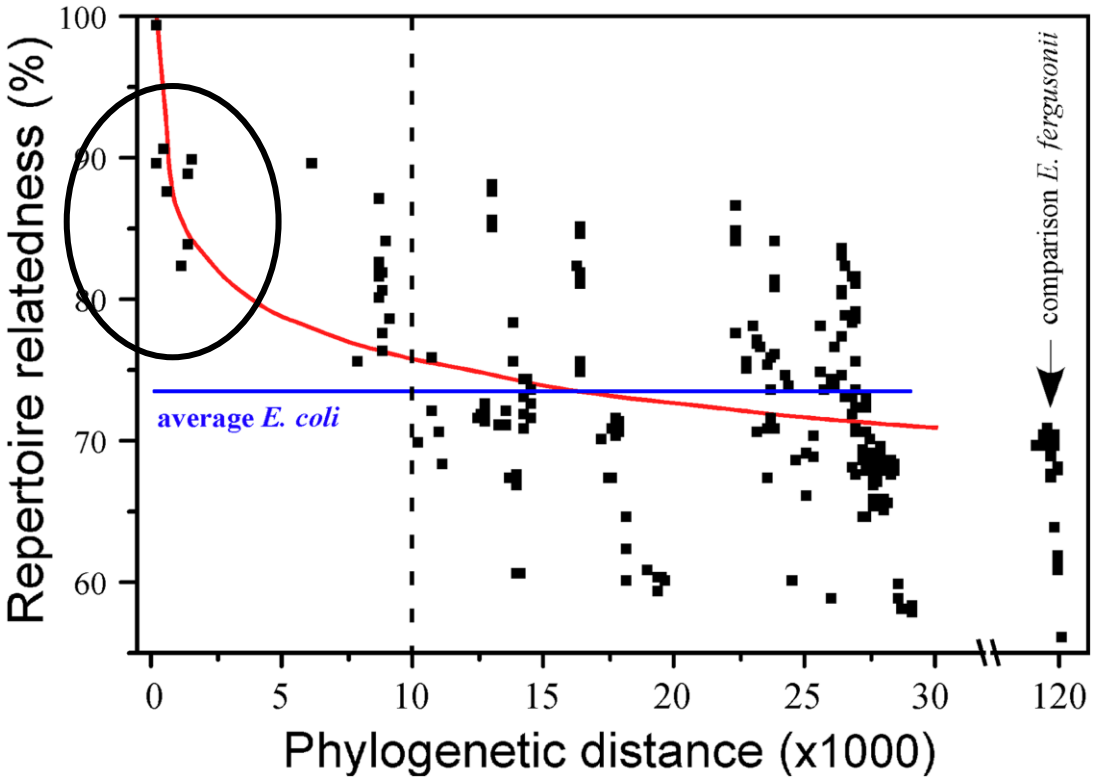


Figure 8. Association entre la proportion de gènes partagés entre deux souches et la distance phylogénétique. La ligne horizontale bleue correspond à la proportion moyenne de gènes partagées parmi les souches de *E. coli/Shigelles*. Les points entourés représentent des souches, qui malgré une faible distance phylogénétique, ne partagent qu'entre 80% et 90% de leurs génomes. Figure adaptée de Touchon *et al.* 2009.

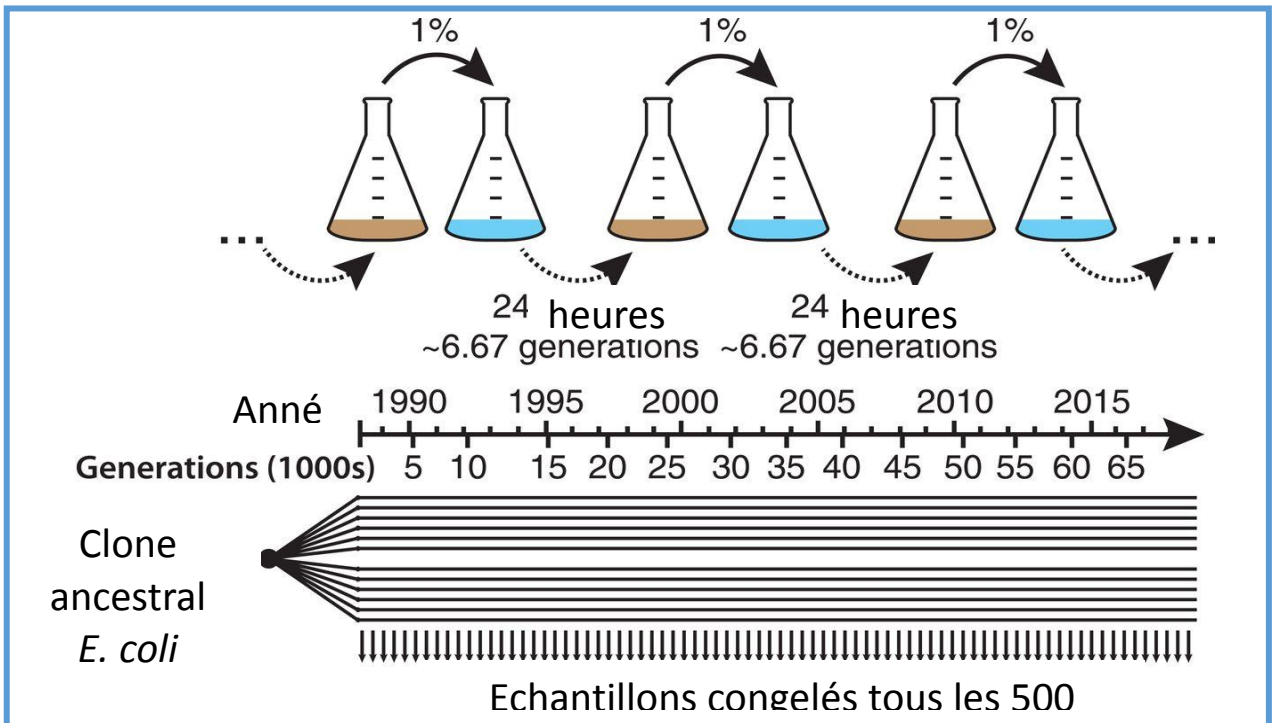
2/Les expérimentations évolutives *in vivo*

Pour mieux comprendre l'évolution du d'*E. coli*, une stratégie proposée dès les années 1950 consiste à recourir à l'évolution expérimentale. Avec l'avènement de la génomique à bas coûts, il est maintenant possible de coupler cette évolution au séquençage de génomes complets et donc de déterminer à la fois les dynamiques d'accumulation de mutations mais aussi l'identité des mutations recrutées au cours de l'adaptation.

2-1/Une expérimentation emblématique : Long Term Evolution Experimentation (LTEE) de Richard Lenski (Figure 9)

Si de nombreuses expérimentations évolutives *in vitro* ont été menées, aucune n'a eu l'ampleur celle développée par Richard Lenski. Douze populations ont été fondées à partir d'un ancêtre commun en 1988 et évoluent depuis au moins plus de 65.000 générations, dans un

milieu pauvre en nutriment à 37°C, avec des échantillons congelés toutes les 500 générations. Richard Lenski et son équipe ont fait évoluer 6 populations indépendantes de *E. coli* REL606, appartenant au groupe phylogénétique A, et 6 populations indépendantes de *E. coli* REL607 qui diffèrent de *E. coli* REL 606 par des mutations dans *araA* (conférant la capacité à croître dans un milieu contenant du L-arabinose) et *recD*. Couplée au séquençage à haut débit de génomes complets, cette expérimentation avec d'autres ont permis à partir de la fin des années 2000 de faire d'énormes progrès dans la compréhension la dynamique de l'évolution ainsi que dans la caractérisation des bases génétiques de l'adaptation (20,30).



<u>Figure 9.</u> Expérimentation évolutive à long terme de Richard Lenski
Il s'agit d'une expérimentation évolutive parallèle de 12 populations, dérivant toutes d'une souche ancestrale commune de *E. coli*, pendant plus de 65 000 générations dans des conditions de culture liquide pauvre en nutriment. Toutes les 24 heures, 1/100 de la population est dilué dans un milieu DM25 frais. Des échantillons de chaque population sont congelés toutes les 500 générations. Figure adapté de Blount *et al.*, 2018.

2-2/Les enseignements des expérimentations *in vitro*

Cette expérimentation emblématique et d'autres nous ont permis de dégager quelques principes de l'adaptation microbienne au niveau génomique. Tout d'abord et nous reviendrons dessus dans le chapitre suivant, comment détecter que des mutations sont impliquées dans l'adaptation.

2-2-2/Les marqueurs de la sélection

2-2-2-1/La convergence génétique

L'observation de l'apparition de génotypes aux expressions phénotypiques semblables sur des lignées évoluant indépendamment mais dans un milieu identique, ou convergence phénotypique, suggère l'action répétée de la sélection naturelle. Le séquençage à haut débit couplé à l'évolution expérimentale ont permis d'avoir une base moléculaire de la convergence à l'échelle génotypique et phénotypique (grâce aux informations sur les gènes impliqués) en quantifiant la récurrence de mutations dans les différentes lignées indépendantes ayant évolué dans le même milieu. Ce phénomène a été parfaitement illustré dans les expérimentations évolutives *in vitro*. Dans la LTEE, 3 gènes (*nadR* codant un régulateur transcriptionnel, *pykF* codant une pyruvate kinase et l'opéron *rbs* codant le catabolisme du ribose) ont été mutés dans les 12 lignées. Dans presque toutes les lignées, les allèles de ces 3 gènes différaient entre chaque population, excluant une contamination croisée. Par ailleurs, ces mutations confèrent un avantage en terme de croissance bactérienne et donc un avantage adaptatif (Barrick *et al.* 2009). Dans une autre étude, la convergence a été analysée en comparant 115 lignées s'adaptant à haute température pendant 2000 générations. Là encore, des convergences ont été observées et cette fois-ci, des centaines de mutations adaptatives différentes ont été identifiées, suggérant que de très nombreuses mutations et de nombreux gènes peuvent contribuer à l'adaptation (Olivier Tenaillon *et al.* 2012a).

2-2-2-2/L'excès de mutations non-synonymes

Un autre marqueur d'évolution est le rapport entre le taux de substitution synonyme (Ks) et non-synonyme (Ka), le Ka/Ks. Théoriquement, quelque soit le régime de sélection le taux de mutation synonyme est constant car la survenue de telles mutations n'entraine pas de changements d'acides aminés et donc de structures de la protéine vecteur de l'expression phénotypique. Les mutations non-synonymes s'accumulent quant à elles en fonction du taux de mutation mais aussi de la sélection auxquelles elles sont soumises. Si l'ensemble des mutations non-synonymes d'un gène n'ont pas d'effet sélectif, le gène est certainement non fonctionnel et le ratio Ka/Ks=1. Si les mutations sont majoritairement délétères la sélection purificatrice réduira le taux de substitutions non-synonymes en purgeant ces mutations de la population et, le ratio Ka/Ks sera inférieur à 1. Ce n'est que lorsque les changements d'acide aminé offrent un avantage sélectif, que les mutations non-synonymes se fixent à un taux supérieur à celui d'une mutation synonyme, résultant en un ratio Ka/Ks supérieur à 1. Mais comme nous le verrons certains éléments sont à prendre en compte pour déterminer avec exactitude le Ka/Ks. Néanmoins cette approche permet même avec potentiellement une seule lignée d'avoir une idée du rôle que la sélection naturelle a joué dans l'émergence des mutations observées. Ainsi dans la LTEE au début de l'évolution le ratio est entre 15 et 30 et même après 50000 génération d'adaptation il reste largement supérieur à 1 (environ 4) suggérant que plus de 80% des mutations non-synonymes observées sont le fruit de la sélection.

2-2-1/Stratégies d'adaptation

D'autres phénomènes récurrents ont été observés lors des expériences d'évolution *in vitro*.

2-2-1-1/Emergence de souches hypermutatrices

Le système de réparation des mésappariements est un mécanisme de surveillance de l'ADN lors de la réplication qui permet de réparer préférentiellement les erreurs introduites par la polymérase dans le brin nouvellement synthétisé. Ce processus fait intervenir plusieurs protéines codées par les gènes *mutS, mutL, mutH*. L'inactivation de ces gènes conduit à une

accumulation des erreurs de réplication (en d'autres termes, des mutations) et résulte en des souche qualifiées d'hypermutateur.

La sélection peut favoriser une augmentation du taux de mutation, et améliorer le taux d'adaptation de la population (Olivier Tenaillon *et al.* 2016a). La présence de souches hypermutatrices entraine une augmentation de la diversité allélique dans la population. Les allèles hypermutateurs générant une diversité génétique plus grande ont plus de chance de produire des mutations bénéfiques et d'augmenter en fréquence grâce à la sélection de ces allèles auxquels ils restent associés en l'absence d'échanges génétiques. De cette sélection de souches hypermutatrices résulte particulièrement suite à un changement d'environnement, une réponse adaptative plus rapide (Sniegowski, Gerrish, and Lenski 1997) (Figure 4C).

Bien que les lignées de mutants puissent connaître des taux plus élevés d'amélioration de leur valeur sélective (O. Tenaillon *et al.* 2001; Wiser, Ribeck, and Lenski 2013), l'effet est généralement faible en raison de l'interférence clonale entre les mutations bénéfiques concurrentes (Gerrish and Lenski 1998a; Barrick and Lenski 2013) et de l'accumulation de mutations délétères (Good and Desai 2014; Wielgoss *et al.* 2011). Des mutations compensatrices ultérieures peuvent survenir pour réduire (Olivier Tenaillon *et al.* 2016a) et même retourner au taux de mutation de base (Turrientes *et al.* 2013) permettant ainsi de limiter l'accumulation de mutations délétères dans un milieu où « le gros » de l'adaptation est réalisé.

2-2-1-1/Mutations des régulateurs globaux

Dans leur environnement naturel, les bactéries doivent survivre à des pressions multiples et doivent faire face à des épisodes répétés d'abondance de nutriments et de famine. Les bactéries détectent ces changements environnementaux en utilisant des réseaux de régulation complexes et fortement interconnectés, ce qui permet aux cellules d'adapter l'expression de leurs gènes aux nouvelles conditions (McAdams, Srinivasan, and Arkin 2004). Ces réseaux de régulation sont caractérisés par une structure multicouche allant du contrôle

local de gènes et d'opérons individuels au contrôle global de l'expression des gènes dans l'ensemble du génome (Hatfield and Benham 2002).

Il a été constaté dans la LTEE et de nombreuses autres expériences, que l'on retrouvait parmi les premières mutations sélectionnées des mutations dans ces régulateurs globaux. Si ces mutations affectant les régulateurs globaux peuvent conférer d'importants avantages sélectifs, elles présentent néanmoins des effets pléiotropes et sont susceptibles d'avoir des effets secondaires maladaptatifs. Une adaptation ultérieure implique donc des changements compensatoires qui contribuent à la réduction de ces effets secondaires (Lamrabet *et al.* 2019). L'adaptation de *E. coli* passant d'un milieu à un autre (Fong, Joyce, and Palsson 2005) révèle donc d'un processus adaptatif en deux temps, avec des mutations ayant de forts effets pléiotropes se produisant dans un premier temps entrainant notamment de nombreux changements transcriptionnels, suivi de modifications compensatoires qui assurent un retour vers un taux de transcription basal (Fong, Joyce, and Palsson 2005; Lamrabet *et al.* 2019).

3/Observation de l'adaptation dans leur environnement naturel

3-1/Adaptation intra-hôte des bactéries pathogènes

Une bactérie pathogène pendant une infection est soumise à un environnement nouveau (passage d'un site où la bactérie est commensale vers le site infecté, par exemple) dans lequel elle sera soumise à plusieurs pressions de sélection intrinsèques d'une part avec la gestion des nutriments et le système immunitaire de l'hôte et extrinsèques d'autre part avec l'antibiothérapie. Les études longitudinales *in vivo* de bactéries pathogènes dans un contexte d'infection ont été largement menées chez l'homme, permettant d'en apprendre un peu plus sur leurs stratégies adaptives.

La diversification de la population de bactérie est un phénomène précoce et permet de produire un éventail d'allèles augmentant les chances qu'une sélection naturelle se fasse sur une sous population. La majorité des études ont pu mettre en évidence une forte diversité

allélique dans les populations bactériennes infectantes. La création de cette diversité peut être obtenue par plusieurs mécanismes comme un taux de mutation de base fort pouvant aller jusqu'à environ 30 mutations par génome par an pour *Helicobacter pylori* (Kennemann *et al.* 2011) ou environ 10 mutations par génome par an chez *Klebsiella pneumoniae* (Mathers *et al.* 2015). Certaines bactéries peuvent même développer un phénotype hypermutateur comme *Burkholderia dolosa* (Lieberman *et al.* 2014a) ou *Pseudomonas aeruginosa* (Marvig *et al.* 2013) dans les infections du poumon chez les patients atteints de mucoviscidose dans la phase initiale de l'infection ou lors d'infections aigue avec *E. coli* (Levert *et al.* 2010). Un autre facteur permettant la diversification allélique des bactéries est la perte ou l'acquisition d'éléments génétiques mobiles par transfert horizontal de gènes (Ochman, Lawrence, and Groisman 2000) vu chez les espèces *Helicobacter pylori* (Kennemann *et al.* 2011) et *Pseudomonas aeruginosa* (Rau *et al.* 2012).

La diversité allélique est façonnée par la dérive génétique et/ou la sélection naturelle. Une sélection très spécifique de certaines fonctions a été observée. La plus emblématique étant l'émergence de résistance en réponse à des traitements antibiotiques comme par exemple *Staphyloccoccus aureus* résistant au linézolide par une mutation ponctuelle sur le gène *relA* (Gao *et al.* 2010) ou *K. pneumoniae* résistant à la colistine par une mutation ponctuelle sur le gène *pmrK* (Cannatelli *et al.* 2014) dans un contexte d'infection évolutive. La sélection favorise aussi l'émergence de mutations permettant l'échappement de la bactérie au système immunitaire du patient par la modification de structures de membranes (Kennemann *et al.* 2011; Lieberman *et al.* 2014a), affectant la balance entre la préservation et les compétences de captations nutritionnelles (Self-preservation And Nutritionnal Compétence, SPANC), notamment chez *E. coli* (Levert *et al.* 2010) ou les gènes de virulence dans les cas de persistance de la bactérie chez l'hôte (Price *et al.* 2013; Young and Wilson 2012; Zdziarski *et al.* 2010). En dehors de ces sites privilégiés de la mutation, le reste du génome est en général conservé faisant

même l'objet d'une sélection purificatrice (Ka/Ks <1) (Marvig *et al.* 2013; Lieberman *et al.* 2014; Didelot *et al.* 2016) suggérant que l'effet dominant de la sélection naturelle sur la bactérie pendant une infection est de conserver la fonctionnalité de la plupart des gènes, l'adaptation n'agissant que sur certaines positions génomiques d'un sous-ensemble de gènes. La diversité de la population bactérienne est aussi façonnée par la dérive génétique amplifiée par la fluctuation de la taille de la population associée à une migration (transmission hôte ou environnement à hôte ou développement de foyer infectieux secondaires) ou à une diminution de la population illustrant respectivement un effet fondateur et goulot d'étranglement (Didelot *et al.* 2016). Ainsi, des études d'observations d'adaptation au cours de la colonisation et de l'infection de *Burkholderia dolosa, Helicobacter pylori, Staphylococcus aureus* et *Pseudomonas aeruginosa* et bien sûr *E. coli* chez l'hôte infecté ont montré des stratégies d'évolution adaptatives communes.

Kisiela *et al.* ont suivi plusieurs isolats de *E. coli* ST131-H30 particulièrement virulents et résistants aux antibiotiques dans deux foyers distincts de 2 et 5 membres atteints d'infections urinaires, impliquant une transmission inter-hôte, allant d'une simple cystite à une pyélonéphrite abcédée nécessitant un drainage. Ils ont montré une forte divergence clonale entre les isolats d'un même foyer probablement façonnée par la sélection positive. La plupart des mutations intéressaient préférentiellement des régulateurs transcriptionnels comme *lacI* ou le régulateur global *lrhA* (Kisiela *et al.* 2017). La transmission de la souche ST131-H30 nous questionne sur le contexte d'adaptation des isolats se faisant soit lors de la colonisation pour s'adapter au tube digestif soit lors de l'infection urinaire en situation pathogène.

3-2/Adaptation intra-hôte des bactéries commensales

Certaines de ces bactéries sont cependant considérées comme commensales du pharynx et du nez pour *Staphylococcus aureus* ou du tube digestif pour *E. coli*. Le commensalisme implique la mise en œuvre d'interactions avec l'hôte et les autres bactéries commensales devant

contribuer au maintien d'un équilibre entre les différents acteurs. Ce contexte différent nécessite probablement une stratégie d'adaptation différente pour les bactéries commensales, notamment pour *E. coli* dans son environnement naturel : le tube digestif.

Comme nous l'avons vu, de par une structure en partie clonale, *E. coli* peut être définie par plusieurs groupes phylogénétiques. La variabilité de distribution des groupes phylogénétiques observée au sein d'un même hôte ou d'hôtes différents en fonction de leur espèce, de leur alimentation ou de leur situation géographique suggère un rôle dans l'adaptation. Par exemple, le groupe phylogénétique B2 est associé à une pathogénicité extra-intestinale (Johnson *et al.* 2006) en rapport avec la présence de facteurs de virulences impliqués dans l'adhésion, la production de toxines, la capture de fer et d'échappement au système immunitaire (Picard *et al.* 1999; Bingen 1998; Johnson and Russo 2018; Bonacorsi *et al.* 2003). Or ces souches s'avèrent être de très bonnes colonisatrices du tube digestif suggérant que la virulence des souches appartenant au groupe phylogénétique B2 soit un sous-produit du commensalisme résultant d'une évolution intra-hôte à long terme (Diard *et al.* 2010).

Pour étudier la dynamique évolutive dans des conditions de commensalisme, une approche consiste à échantillonner *E. coli* dans les selles d'individus asymptomatiques à différents intervalles de temps. Reeves *et al.* ont mené une étude de 3 ans sur l'évolution et la transmission d'un clone de *E. coli* dans un foyer de 6 membres, dont un chien. Ils ont suivi un clone B2 (nommé clone D) qui n'était pas dominant mais isolé à 14 reprises à différents moments et chez différents hôtes. Ils ont montré un taux de transmission élevé du clone au sein du foyer, un faible taux de mutation, une diversité limitée et aucune trace d'adaptation claire (Reeves *et al.* 2011). Cela semble concordant avec Golubchik *et al.* (Golubchik *et al.* 2013) qui ont suivi longitudinalement l'adaptation de *Staphylococcus aureus* en portage nasal asymptomatique. Ils avaient aussi pu mettre en évidence une faible diversité, une évolution marquée par une sélection purificatrice suggérant une bactérie bien adaptée au commensalisme

(Couce and Tenaillon 2015). Mais cette démarche a ses limites, en effet, nous ne pouvons rigoureusement pas conclure pour l'adaptation de *E. coli* qui change constamment de milieu et inconstamment présente dans le tube digestif. Dans l'article n°1, nous nous concentrerons sur l'évolution d'un clone dominant isolé sur une période d'un an chez un individu en bonne santé.

4/Expérimentation évolutive *in vivo*

Les expérimentations évolutives *in vitro* nous ont permis de tirer de précieux enseignements et une base pour l'observation de l'évolution de bactéries dans leur milieu naturel. Cependant, la détection d'une absence ou d'un faible signal de sélection (faible puissance avec une faible taille de population efficace) couplé à l'impossibilité de reproduire les conditions d'évolution de manière rigoureuse afin de tester si l'adaptation a réellement eu lieu illustrent les limites de l'observation *in natura* et révèle l'intérêt pour une stratégie alternative plus proche des conditions naturelles que l'évolution *in vitro*, mais plus contrôlé que l'évolution *in natura* : l'expérimentation *in vivo* en modèle animal.

4-1/Modèles d'études *in vivo* pour étudier l'interaction hôte-microbe

L'étude de l'impact de certains facteurs implique le contrôle d'autres facteurs explicatifs pouvant créer une confusion dans l'interprétation des résultats de l'expérimentation. L'expérimentation *in vivo* sur modèle animal permet de contrôler le fond génétique, particulièrement chez les souris qu'elles soient consanguines ou non (Rice and O'Brien 1980), le régime alimentaire, l'environnement , l'âge et l'état de santé. Le choix commun du modèle murin en laboratoire résulte d'un raisonnement pragmatique du fait de la docilité, la tolérance au confinement, un haut taux de fécondité et des installations relativement peu encombrantes ou onéreuses par rapport à d'autres modèles. Utiliser un modèle murin pour pouvoir extrapoler semble assez contre intuitif en raison de la différence de taille, de rythme métabolique, de physiologie intestinale ou de régime (Perlman 2016). Cependant, au-delà d'une proximité

génétique (Mouse Genome Sequencing Consortium *et al.* 2002), il semble qu'il existe une certaine similarité entre la souris et l'homme (Vanhooren and Libert 2013). La composition du microbiote intestinal est relativement similaire entre les deux espèces, ainsi que les modifications induites par l'alimentation (Swanson *et al.* 2011) et leur réponse immunitaire (Takao and Miyakawa 2015).

4-2/Evolution expérimentale de *E. coli* dans le tube digestif de souris

Plusieurs études expérimentales utilisant un modèle de colonisation intestinale de souris ont donc été réalisées. Fait intéressant, les données génomiques ont révélé des convergences dans l'inactivation (De Paepe *et al.* 2011) ou l'activation constitutive (Barroso-Batista *et al.* 2014) d'opérons impliqués dans le métabolisme des sucres. En effet, l'adaptation observée chez *E. coli* K-12 dans le tube digestif de souris semble être, en partie, façonnée par une stratégie d'utilisation des sucres. De Paepe *et al.*, dans un modèle d'expérimentation évolutive de *E. coli* K-12 dans le tube digestif de souris axéniques, ont rapporté une convergence de mutations dans le gène *ompBSG1*, dans l'opéron *flhDC* et le gène *malT*. La mutation *ompBSG1* entraînait une diminution significative du taux de croissance maximal de tous les sucres, quant aux mutants Δ*flhDC* et Δ*malT*, ils se développaient également lentement sur certains de ces sucres (L-arabinose, D-galactose, D-mannose et bien sûr le maltose pour Δ*malT*) (De Paepe *et al.* 2011). La sélection ciblant les mutations de l'opéron *flhDC* chez *E. coli* K-12 a aussi été retrouvée après une colonisation digestive par cette souche de souris traitées par la streptomycine, suggérant une association étroite entre la perte de mobilité (un changement possible de niche de colonisation) et le changement de stratégie de captation des sucres et cela indépendamment du microbiote (Leatham *et al.* 2005; Gauger *et al.* 2007). Plus tard, Barroso-Batista *et al.* ont suivi l'adaptation de *E. coli* K-12 MG1655 au tube digestif de souris traitées par la streptomycine et ont observé des mutations dans le gène *srlR* codant pour le répresseur de l'opéron *srl* et plusieurs gènes de l'opéron *gat* dont les opérons sont respectivement impliqués

dans le métabolisme du sorbitol et du galactitol. Ces observations presque constantes révèlent que les ajustements du métabolisme du sucre sont nécessaires dans l'intestin. Dans cette thèse, nous nous proposons d'étudier l'impact du régime alimentaire (Articles n°2 et 3), susceptible de délivrer qualitativement et quantitativement des sucres différents, sur l'adaptation de *E. coli* dans le tube digestif de souris au long court.

Cependant, ces études souffrent de limites inhérent aux milieux d'évolution et aux souches utilisées. Premièrement, ces études utilisent des souris axéniques ou traitées à la streptomycine pour permettre la colonisation et le maintien de *E. coli* dans le tube digestif. Or ces conditions non naturelles pourraient influencer l'adaptation des souches. Par exemple, Lescat *et al.* ont observé des mutations répétées dans les gènes *rluD* et *gidB* impliqués dans la maturation des ribosomes et associés à l'adaptation à la streptomycine (Lescat *et al.* 2017). Afin d'étudier l'adaptation d'un isolat naturel *de E. coli* 536 dans des conditions naturelles, nous nous sommes efforcés d'élaborer un modèle de colonisation de souris en mimant la transmission du microbiote de la mère à l'enfant et en se passant le plus possible de la streptomycine (Article n°2).

Deuxièmement, *E. coli* K-12 utilisée dans de nombreuses études d'évolutions *in vivo* est une souche isolée dans les fèces d'un patient convalescent après une dysenterie dans les années 1920 et qui s'est depuis adaptée aux conditions de conservations de laboratoire durant toutes ces années. Comme mentionné précédemment, le taux d'adaptation de *E. coli* dépend de sa mal-adaptation initiale (Couce and Tenaillon 2015), et il est donc probable qu'une souche de laboratoire soit moins adaptée au tube digestif qu'une souche commensale et donc que le processus adaptatif en soit affecté et ne représente que partiellement le processus d'adaptation à un nouvel hôte. Par ailleurs, Barroso-Batista *et al.* ont suivi une souche dérivée de *E. coli* K-12 dans des intestins de souris traitées à la streptomycine. La souche utilisée pourrait également contribuer au modèle d'évolution observé. Ils ont observé une inactivation récurrente de

l'opéron *gat*. Pour garantir l'expression constitutive des protéines fluorescentes (utilisé pour suivre la dynamique d'adaptation par le changement de fréquence de ces marqueurs «neutres»), le promoteur *lac* a été utilisé et une suppression de l'opéron *lac* était nécessaire (Barroso-Batista *et al.* 2014). L'absence d'un opéron *lac* fonctionnel ne permettait pas *à E. coli* de dégrader le lactose en glucose et en galactose et donc sa forme réduite, le galactitol. Ainsi, l'inactivation constitutive de l'opéron *gat* pourrait éventuellement être liée à l'absence de son substrat, le galactitol. Dans le travail présenté nous avons fait le choix d'étudier l'adaptation de *E. coli* dans le tube digestif, en utilisant des isolats naturels. Nous avons choisi d'utiliser la souche 536 isolée dans les urines d'une patiente atteinte d'une infection urinaire et appartenant au groupe phylogénétique responsable de pathologies extra-intestinales et réputée être une bonne colonisatrice du tube digestif. Afin de déterminer le rôle du fond génétique de la souche (Article n°3), nous avons aussi utilisé la souche *E. coli* HS. Il s'agit d'une souche commensale isolée dans les selles sans pathogénicité intestinale ou extra-intestinale. En outre, *E. coli* HS appartient au groupe phylogénétique A comme K-12, mais les deux n'ont que 3640 gènes en commun (Dunne *et al.* 2017).

CONCLUSION DU CHAPITRE II

E. coli possède un génome clonal lui permettant de conserver, entre autre, les gènes impliqués dans l'adaptation à sa niche écologique comme le groupe phylogénétique B2. *E. coli* a aussi un génome malléable par le flux d'entrée et d'élimination de gènes provenant de l'environnement, la recombinaison et les mutations dont la persistance est guidée par la dérive génétique ou la sélection. *E. coli* est donc adaptée à son environnement mais aussi façonnée par lui. De nombreuses expérimentations ont été menées pour comprendre les stratégies d'adaptations de *E. coli* de manière absolue *in vitro*, en situation pathogène ou en situation commensale confrontée à son environnement naturel : le tube digestif. Alors que les expérimentations *in vivo* sur modèle animal ne nous permettaient pas de reproduire de façon

satisfaisante le cadre d'évolution naturel de *E. coli* (souris axénique, souris traitées à la streptomycine, colonisation par une souche *de E. coli* de laboratoire), les observations *in natura* ne permettaient pas, en général, de suivre une lignée au long court et d'identifier les déterminants de l'adaptation de manière rigoureuse. Nous avons conduit une première étude afin de déterminer le comportement adaptatif de *E. coli* dans le tube digestif. En effet, nous avons pu avoir la possibilité de suivre un clone dominant de *E. coli* ED1a pendant 1 an dans le tube digestif d'un homme occidental. Nous avons ensuite conduit une expérimentation d'évolution *in vivo* sur modèle murin pendant un an en étudiant l'impact de l'alimentation sur l'adaptation d'un isolat naturel de *E. coli* 536 appartenant au groupe phylogénétique B2. Nous avons essayé de reproduire un environnement digestif le plus proche possible de la réalité en mimant la colonisation du tube digestif de nouveau-nés par la transmission du microbiote de la mère à l'enfant. En outre, afin de déterminer le rôle du fond génétique de la souche couplée à l'alimentation dans l'adaptation de *E. coli* dans le tube digestif, nous avons mené une troisième expérimentation suivant l'adaptation de *E. coli* 536 ou *HS*, appartenant au groupe phylogénétique A, en fonction du régime alimentaire pendant plus d'un an. L'étude de l'adaptation de manière quantitative, en d'autres termes, l'étude de la diversité (dépendant du la taille de population efficace et du taux de mutation) ou du régime de sélection nécessite la maitrise de certains outils que nous allons voir dans le prochain chapitre.

CHAPITRE III : ETUDE DES MUTATION APPARAISSANT *IN VIVO*

I/Introduction

L'évolution est un processus dont résulte la modification des espèces au fil des générations. Des traits phénotypiques nouveaux s'imposent au sein d'une population puis d'une espèce par la sélection naturelle ou par le fruit du hasard. Ces nouveaux traits phénotypiques sont l'expression d'un nouveau génotype issu d'une ou plusieurs mutations. Le taux de mutation rythme ainsi en partie l'émergence de ces nouveaux traits phénotypiques.

Le processus évolutif est illustré de manière tout à fait concrète en terme de santé publique avec, notamment, l'apparition de l'antibio-résistance et de la résistance aux anti-infectieux en général. La pression de sélection imposée par les antibiotiques aux bactéries, abondamment utilisés dans le milieu vétérinaire et hospitalier, ont permis l'avènement de bactéries résistantes à des antibiotiques de plus en plus large spectre au point que l'on se pose la question de leur efficacité future. On observe un ensemble de génotypes différents pouvant aboutir à un même spectre de résistance. Prenons l'exemple des BLSE qui comptent plus de 500 variantes incluant les plus significatives en terme clinique tel que CTX-M, OXA, TEM et SHV (Bush and Jacoby 2010). Les différentes enzymes BLSE de type TEM, SHV et OXA dérivent respectivement de mutations ponctuelles apparues dans des enzymes de spectre étroit à médiation plasmidique TEM-1, TEM-2, SHV-1 et OXA-10 (Paterson and Bonomo 2005; Bush and Jacoby 2010). Bien que les enzymes BLSE de type CTX-M trouvent leur origine dans un transfert d'éléments génétique mobiles provenant de *Kluyvera spp* (Cantón, González-Alba, and Galán 2012), on observe plusieurs allèles divergents se différenciant par des mutations ponctuelles (Bonnet 2004; Poirel, Naas, and Nordmann 2008). Une telle diversité allélique pour l'expression d'un même phénotype suggère un moteur de création de la diversité qui puisse en partie être quantifiée par l'intensité du taux de mutation.

Il convient cependant de ne pas réduire la diversité allélique présente au sein d'une espèce à son seul taux de mutation. La taille de la population efficace se définit, en pratique, par le nombre d'individus au sein d'une espèce susceptible de se reproduire, en d'autres termes, de transmettre leurs mutations et de garantir la diversité de l'espèce. En l'absence de sélection, les principaux déterminants de la diversité au sein d'une population (θ_w) s'avèrent être le taux de mutation par génération (μ_g), la taille de la population efficace (N_e) et la taille du génome de l'espèce (L), de telle sorte que $\theta_w = 2N_eL\mu_g$ (Equation 1. Calcul de la diversité génétique par le θ de Watterson) (Watterson 1975). La taille du génome étant relativement fixe et facilement

mesurable, les deux variables contrôlant la diversité génétique s'avèrent donc être la taille de la population efficace et le taux de mutation au sein de l'espèce. Plus la population est grande et/ou plus le taux de mutation sera élevé, plus la population disposera d'une grande diversité génétique (Ellegren and Galtier 2016). Une fois le taux de mutation quantifié, la taille de la population efficace peut potentiellement être estimée simplement par la diversité génétique observée. Une faible diversité génétique suggère une taille de population réduite alors qu'une forte diversité génétique, en l'absence d'hypermutateur, suggère plutôt une population de taille importante et un potentiel adaptatif conséquent. La prédiction d'une taille efficace réduite alors que la densité de *E. coli* dans les fèces est assez importante peut avoir plusieurs origines. Soit cela peut être lié à une mortalité importante de la population, soit au fait que seulement une sous-population se maintient et contribue à l'avenir évolutif de la population alors qu'une partie importante de la population, celle dans la lumière de l'intestin semble-t-il (Poulsen *et al.* 1995), n'y contribue pas. Il se peut aussi qu'une colonisation récente soit à l'origine de cette faible diversité, l'état d'équilibre sous-jacent au modèle n'étant pas encore atteint, ou pour des raisons similaires qu'un balayage sélectif récent se soit produit et que la reconstruction de la diversité suite à cet événement soit encore en cours.

Le rythme d'apparition des mutations est donc un enjeu capital au sein de la stratégie d'adaptation des bactéries. La détermination du taux de mutation d'une bactérie dans un milieu donné nous permettrait ainsi de déterminer son potentiel d'adaptation. Dans ce chapitre, nous souhaitons étudier les différentes stratégies de variabilité du taux de mutation, les éléments pouvant biaiser son calcul et les différents modes de calculs.

II/Le taux de mutation est variable au niveau local et global

1/Les points chauds mutationnels

Le taux de mutation ne peut pas simplement être considéré de manière globale. Nous pouvons observer des variations des taux de mutation en fonction du type de mutation et de la

région du génome. Les points chauds mutationnels sont des régions circonscrites du génome faisant l'objet d'une diversification allélique intense. Le taux de mutation de ces régions peut être 1000 fois plus élevé que sur le reste du génome (Lourenço *et al.* 2016). Ces régions peuvent passer inaperçues dans le cadre d'une évolution neutre ou une sélection purificatrice, car la diversité créée reste malgré tout à faible fréquence mais peuvent être mise en avant dans le contexte de la sélection positive (Chattopadhyay *et al.* 2009a) ou certains allèles dont la fréquence initiale est non négligeable vont rapidement répondre à la sélection. De façon général, la diversité allélique issue des points chauds mutationnels s'illustre par des indels (duplication/insertion/délétion) se produisant sur de courtes séquences répétées (Chattopadhyay *et al.* 2009a). Ces insertions ou délétions résultent d'un glissement de l'ADN polymérase lors de la réplication suite à un appariement décalé entre matrice et brin synthétisé dans la séquence répétée (Maki 2002a).

Les points chauds mutationnels semblent ainsi le fait de circonstances en grande partie structurelle liées à la composition du génome. Leurs conséquences fonctionnelles dépendent néanmoins de leur localisation. Par exemple l'opéron lactose est un ensemble composé d'un gène régulateur *lacI* et de 3 gènes *lacZ, lacY* et *lacA* concourant à la métabolisation du lactose en une molécule de glucose et une molécule de galactose. L'activité de cet opéron peut être régulée en fonction de la présence ou l'absence de glucose et de lactose dans l'environnement. En l'absence de lactose, le répresseur exprimé constitutivement par *lacI* va se fixer sur la séquence *lacO* dans la région intergénique entre *lacI* et *lacZ* créant ainsi une boucle bloquant la transcription du reste de l'opéron. Dans des conditions de sélection favorisant l'activation constitutive de l'opéron lactose, on observe des mutations dans un point chaud au sein d'une séquence répétée de *lacI* mais aussi quelques mutations non-synonymes dans le gène *lacI* et quelques mutations ponctuelles sur l'opérateur *lacO* correspondant au site de fixation du répresseur (Swerdlow and Schaaper 2014a). Le point chaud de mutation dans *lacI*, assure la

présence à une fréquence non négligeable de bactéries ayant une activation constitutive de l'opéron lactose. Dès que les conditions favorisant cette expression sont rencontrées, la population évoluera rapidement vers l'expression constitutive car les allèles sont déjà présents dans la population. Da façon intéressante, de nombreux régulateurs de voies métaboliques, *lacI, malT, gat, srlR* ou *dgoR* semblent porter de tels points chauds et donc permettre de rapides changement de régulation métabolique (De Paepe *et al.* 2011; Barroso-Batista *et al.* 2014; Lescat *et al.* 2017).

2/Variabilité du taux de mutation en fonction du temps et du milieu

Nous venons de voir qu'il existe un phénomène d'hypermutabilité locale dû à la fois à des incapacités des ADN polymérases sur les séquences répétées mais aussi à des limites du système de réparation des mésappariements dans la réparations d'indels de plusieurs bases. Le taux de mutation d'une souche peut cependant aussi augmenter globalement du fait de la perte de la capacité à réparer ou prévenir les dommages sur l'ADN (Foster *et al.* 2015). Ce phénotype appelé hypermuteur est principalement provoqué par la perte d'enzymes codés par les gènes *mutT, mutY, mutM* (Foster *et al.* 2015) ou *mutH, mutL ou mutS* (Olivier Tenaillon *et al.* 2016). Les trois premiers gènes sont impliqués dans le contrôle des mutations crées par le 8-oxo guanine, un agent mutagène puissant, et les trois derniers sont les gènes majeurs du système de réparation des mésappariements.

Les souches hypermutatrices ont d'abord été décrites lors d'expérimentations *in vitro* notamment dans le LTEE de Richard Lenski. Dans cette expérience, sur les 12 lignées propagées, la moitié a développé un phénotype hypermutateur entrainant une rapide accumulation de mutations dont une partie de mutations bénéfiques. Ces lignées se sont adaptées aux conditions imposées par l'expérience légèrement plus rapidement que les lignées de faible taux de mutation (Olivier Tenaillon *et al.* 2016). A terme, dans certaines populations, l'hypermutabilité a décliné au cours des derniers milliers de générations car la production à

grande échelle de mutations délétères favorisent la sélection d'allèles anti-mutateurs (Wielgoss *et al.* 2013). Le phénotype d'hypermutabilité présente un avantage sélectif initialement grâce au grand nombre d'allèles bénéfiques accessibles dont la sélection compense les coûts associés à la production de mutations délétères. Cependant avec la raréfaction des mutations bénéfiques d'effet large, l'accumulation de mutations délétères rend par la suite le phénotype hypermutateur coûteux.

Dans la nature, nous avons vu que certaines bactéries comme *P. aeruginosa* (Marvig *et al.* 2013) ou *E. coli* (Levert *et al.* 2010) pouvaient présenter un fort taux de mutation dans le cadre d'infections chroniques ou aigue. Il s'agit probablement d'une stratégie sélectionnée par les fortes pressions de sélection (antibiothérapie, système immunitaire de l'hôte, nouvelle niche de colonisation …) et favorisant une adaptation intra-hôte rapide (Didelot *et al.* 2016).

En conclusion de cette partie, le taux de mutation est un élément déterminant dans l'adaptation des bactéries à leur environnement. Il peut varier de manière globale ou ciblée dans le génome. L'hétérogénéité du taux de mutation le long du génome ou dans le temps et donc entre différentes souches d'une même espèce doit donc être proprement quantifiée pour identifier le comportement global de la bactérie face à son environnement.

III/Les éléments pouvant biaiser le taux de mutation

1/Rôle de la sélection dans la perception du taux de mutation

Les gènes sont codés par une série de triplet de nucléotide (codons) codant pour un acide aminé dont l'enchainement produit une protéine. Le code génétique est universel et redondant, il se compose de 64 combinaisons de triplet de nucléotides susceptibles de traduire 20 acides aminés (Osawa *et al.* 1992) et trois codons stops. En conséquence un acide aminé peut être codé par plusieurs combinaisons possibles de triplet. On peut donc caractériser les mutations en fonction de leur impact sur la protéine codée. Une mutation synonyme n'engendre pas de

changement d'acide aminé. Au contraire, une mutation non-synonyme donne lieu par définition à un changement d'acide aminé.

La sélection Darwinienne menant au changement d'un trait phénotypique dans une population favorise l'émergence de mutations non-synonyme modifiant le génotype. A l'inverse, une sélection purificatrice (agissant contre le changement phénotypique d'une population) aura tendance à éliminer les mutations non-synonymes non bénéfiques. Selon l'un ou l'autre régime de sélection, le taux de mutation peut être surestimé ou sous-estimé respectivement. Le rythme d'accumulation des mutations non-synonymes est ainsi soumis à l'action de la sélection naturelle.

Les mutations synonymes qui ne modifient pas la séquence d'une protéine, sont potentiellement moins affectées par l'action de la sélection naturelle et peuvent au premier ordre être considérées comme neutres n'ayant pas a priori d'impact sur le phénotype. Etant moins affectée par la sélection, l'accumulation des mutations synonymes dans le temps se rapproche au mieux du processus mutationnel.

2/Les mutations synonymes peuvent possiblement être soumises à la sélection (biais d'usage de codon)

Cependant l'affirmation d'une accumulation neutre des mutations synonymes mérite d'être grandement nuancé. Dans une grande variété d'organismes, des codons synonymes sont utilisés à différentes fréquences, un phénomène connu sous le nom de biais de codon. Il semble que la cause principale de la sélection sur le biais des codons soit que certains codons, notamment ceux dont les ARNt sont fréquents, sont traduits de manière plus efficace et plus fidèle (Hershberg and Petrov 2008). L'usage d'un codon synonyme par rapport à un autre pour coder le même acide aminé dépend de la composition du génome, la proportion de GC dans le génome et donc des biais mutationnels, de la quantité relative d'ARNt, de l'intensité de l'expression des gènes et de l'efficacité de la sélection naturelle au sein de l'espèce. La combinaison de ces

facteurs génère un biais d'usage de codon qui peut être spécifique. La sélection a donc un impact sur l'accumulation des mutations synonymes.

En pratique, cependant les effets sélectifs associés au biais de codon sont faibles, et d'autres types de sélection liée à l'impact des mutations synonymes sur la structuration des ARN messagers sont peu fréquents. Il en résulte que ces mutations peuvent être considérées au premier ordre comme neutres, mais il faut bien sûr rester vigilant.

IV/Biais liés au calcul du taux de mutation

Comment peut-on calculer le taux de mutation ? Pour ce calcul, nous devons à la fois prendre en compte la nature des mutations (synonymes/ non synonymes) pour éviter si possible l'effet confondant de la sélection naturelle mais aussi le type de mutations car celle-ci n'ont pas la même fréquence notamment les transitions sont plus fréquentes que les transversions et par ailleurs, elles entrainent des changements synonymes ou non synonymes avec des fréquences différentes.

L'observation du code génétique nous permet de déduire que la probabilité d'apparition de mutations non-synonymes est à peu près 4 fois plus élevée que la survenue de mutations synonymes. On observe qu'une mutation ponctuelle survenant au niveau du premier ou du deuxième nucléotide du codon aboutissait à une mutation non-synonyme dans la grande majorité des cas et dans 1 tiers des cas lorsque la mutation ponctuelle survenait au niveau du troisième nucléotide. Théoriquement, pour chaque codon, une mutation ponctuelle peut aboutir à 9 codons différents entrainant de 0 (codons ATG et TGG) à 4 mutations synonymes possibles (codons CGA, CTG, CTA et CGG). Ainsi, la fréquence des mutations synonymes dépend du codon. Pour estimer un taux de mutation synonyme, il faut déterminer un nombre de sites où de telles mutations peuvent se produite et en faire le ratio avec le nombre de mutations observées. Le premier enjeu est donc de déterminer, dans un génome donné, le nombre de sites nucléotidiques pouvant possiblement aboutir à une mutation synonyme.

Type de mutation ponctuelle	i	Type de substitution
AT=>CG	1	Transversion
AT=>GC	2	Transversion
AT=>TA	3	Transversion
CG=>GC	4	Transversion
CG=>TA	5	Transition
CG=>TA	6	Transversion

Table 1. Liste des types de mutations ponctuelles

Pour calculer le taux de mutation, nous nous intéressons à une mesure réalisée sur des temps courts, ce qui nous permet de négliger les évènements multiples par site. En pratique nous sommes intéressés par l'accumulation des mutations à l'échelle génomique sur des centaines ou des milliers de générations. Dans ce cadre, pour déterminer le nombre de sites nucléotidiques pouvant possiblement aboutir à une mutation synonyme (S), il faut préalablement établir pour chacun des codons et chacun des 6 types de mutations le nombre du mutations synonymes (s_{ij}) et recenser le nombre d'emplacements sur le génome pour chacun des 61 codons codant pour des acides aminés (E_j). i étant un type de mutation ponctuelle et j étant un des 61 codons. On a alors

$$S = \sum_{i=1}^{6} S_i = \sum_{i=1}^{6} \sum_{j=1}^{61} (E_j \ . \ s_{ij})$$. (Equation 2. Détermination du nombre de sites nucléotidiques susceptible d'aboutir à une mutation synonyme)

Par exemple, pour le génome de *E. coli* 536, nous avons recensé 13 185 emplacements pour le triplet GAG ; ce triplet peut donner une mutation synonyme par une substitution par transition G->A en troisième position du codon. En d'autres termes, en ne tenant compte que du codon GAG, une mutation synonyme G->A ne peut se faire que sur 13 185 sites. Mais en prenant compte de l'ensemble du génome, et donc des 60 autres codons, une mutation synonyme G->A peut se faire sur 769 338 sites.

L'étape suivante consiste à estimer les biais de mutation. Sur les 6 types possibles de changement d'un nucléotide vers un autre, 4 sont des transversions entre une purine (A ou G)

et une pyrimidine (C ou T), et les 2 autres sont des transitions. Chaque nucléotide (par exemple le nucléotide G) peut muter par un type de transition (G->A) et deux types de transversion (G->C ou G->T). Paradoxalement, nous observons de manière constante dans les études d'évolution expérimentale une proportion au moins deux fois plus importante de transition que de transversion que soit *in vitro* (Wielgoss *et al.* 2011) ou *in vivo* (Lescat *et al.* 2017).

La cause de ce biais de transition/transversion peut être multiple. Deux hypothèses principales ont émergé pour expliquer ce phénomène : l'hypothèse de sélection et l'hypothèse mutationnel. L'hypothèse sélective repose sur l'observation selon laquelle, les transitions non synonymes sont plus susceptibles de conserver des propriétés biochimiques importantes de l'acide aminé d'origine (Miyata, Miyazawa, and Yasunaga 1979; Vogel and Kopun 1977; J. Zhang 2000). Même si cette hypothèse est plausible (Lyons and Lauring 2017), c'est oublier que les mutations ponctuelles synonymes, moins susceptibles d'être soumises à la sélection, sont aussi soumises au biais de transition/transversion (Wielgoss *et al.* 2011; Lescat *et al.* 2017). Par ailleurs, ce biais touche indifféremment les régions codantes et non codantes faisant émerger l'hypothèse mutationnelles (Jiang and Zhao 2006; Z. Zhang and Gerstein 2003) suggérant que la cause de ce biais se trouve au niveau de l'ADN à travers tout le génome. Les transitions seraient moins susceptibles de perturber la conformation en double hélice de l'ADN que les transversions de par la différence de taille entre les purines et les pyrimidines (Guo *et al.* 2017). L'excès de transitions proviendrait donc de la similarité stérique des purines d'une part et des pyrimidines d'autre part. Cette similarité limite la capacité de la polymérase à reconnaître les transitions et résulte donc dans un taux de transitions plus élevé que pour les transversions qui sont plus à même d'être corrigées par l'activité de relecture de la polymérase.

S'il a été initialement proposé de ne considérer que deux taux de mutation différents : un pour les transitions l'autres pour les transversions, des études plus poussées et notamment l'étude de lignées d'évolution expérimentale dans lesquelles le sens des mutations ne fait pas

de doute, ont permis de constater que chaque type de mutation a son propre taux. Ainsi, plusieurs études ont observé un large biais mutationnel en faveur la transition GC vers AT par rapport à la transition AT vers GC (Wielgoss *et al.* 2011; Lescat *et al.* 2017) confortant l'idée d'une forte pression mutationnelle vers des séquences riches en AT (Rocha and Feil 2010a).

L'existence de biais de mutation est connu depuis longtemps, et plusieurs taux peuvent être pris en compte dans les analyses phylogénétiques (Wakeley 1996). Dans de tels modèles, les mutations les moins fréquentes comme les transversions ont plus de poids (Sinsheimer, Lake, and Little 1997). Il est aussi important de prendre en compte ces biais dans le calcul du taux de mutation pour deux raisons. La première est évidemment la précision, et la seconde est que les différents types de mutations ne produisent pas avec la même probabilité des mutations synonymes. Par exemple la transition AT vers GC produit dans le code génétique 34% de mutations synonymes alors que la transversion GC vers CG n'en produit que 17%.

Prendre en compte les biais de mutation dans le calcul du taux de mutation est donc nécessaire. Une façon d'opérer est de ramener les différents taux à celui d'une des transitions. Nous avons calculé pour chaque type de mutation, i, une pondération par le biais ε_i entre le taux de mutation de type i par rapport à la transition GC à AT, transition la plus fréquente dans les données. Pour étudier le taux de mutation sur une durée courte, prendre en compte les mutations multiples au même site n'est pas important, par contre il est fondamental de corriger pour le biais de mutation. Cependant, il est possible de n'avoir que peu de mutations synonymes pour estimer ces biais. On peut dans ce cas prendre l'ensembles des mutations indépendantes du jeu de données pour estimer les biais en faisant l'hypothèse que le panel de mutations observées correspond au biais de mutation.

Nous pouvons dans ce cas résumer la pondération relative par l'équation :

$$\varepsilon_i = \frac{\sum_{\text{all genotype}}(n_i + s_i)/(N_i + S_i)}{\sum_{\text{all genotype}}(n_5 + s_5)/(N_5 + S_5)}$$ (Equation 3. Pondération ε pour le biais de mutation) avec n_i et s_i le nombre de mutations non synonymes et synonymes d'un type de mutation i et N_i et S_i le nombre de site du génome sur lesquels peuvent survenir les mutations non synonymes et synonymes de type i (Tableau 1). On remarquera que la pondération pour les mutations de type 5 est de 1. Cette transition, la plus fréquente (GC -> TA) sert de référence. L'intégration de la pondération dans les calculs se fera en divisant le nombre de mutations dans un isolat appartenant au type de mutation i par son pondérateur ε_i.

V/Mode de calcul du taux de mutation

Pour déterminer le mode de calcul du taux de mutation, il convient de poser l'hypothèse de la structure de la population directement liée au design de l'étude. On peut distinguer 2 types de conception dans les études évolutives. Le premier type d'études expérimentales évolutives fait évoluer les lignées de manière indépendante dans le temps. Ces lignées évoluent sans aucun lien généalogique hormis celui de la souche ancestrale. Le deuxième type d'étude est la plupart du temps observationnelle suivant une lignée en prélevant plusieurs isolats à plusieurs points de temps. Ces isolats prélevés dans un environnement donné ont un lien généalogique.

1/Mode de calcul pour les lignées indépendantes

Le choix du mode de calcul du taux de mutation repose sur l'hypothèse de distribution de l'apparition des mutations dans le temps. On peut considérer que les mutations apparaissent fréquemment ou de manière régulière dans le temps suivant le rythme d'une horloge moléculaire dont l'accumulation dans le temps est linéaire. Similairement, on peut considérer que les mutations sont des phénomènes rares dont l'apparition est indépendante du temps écoulé depuis la dernière mutation suivant ainsi une loi de Poisson dont le paramètre devra être déterminé par un maximum de vraisemblance.

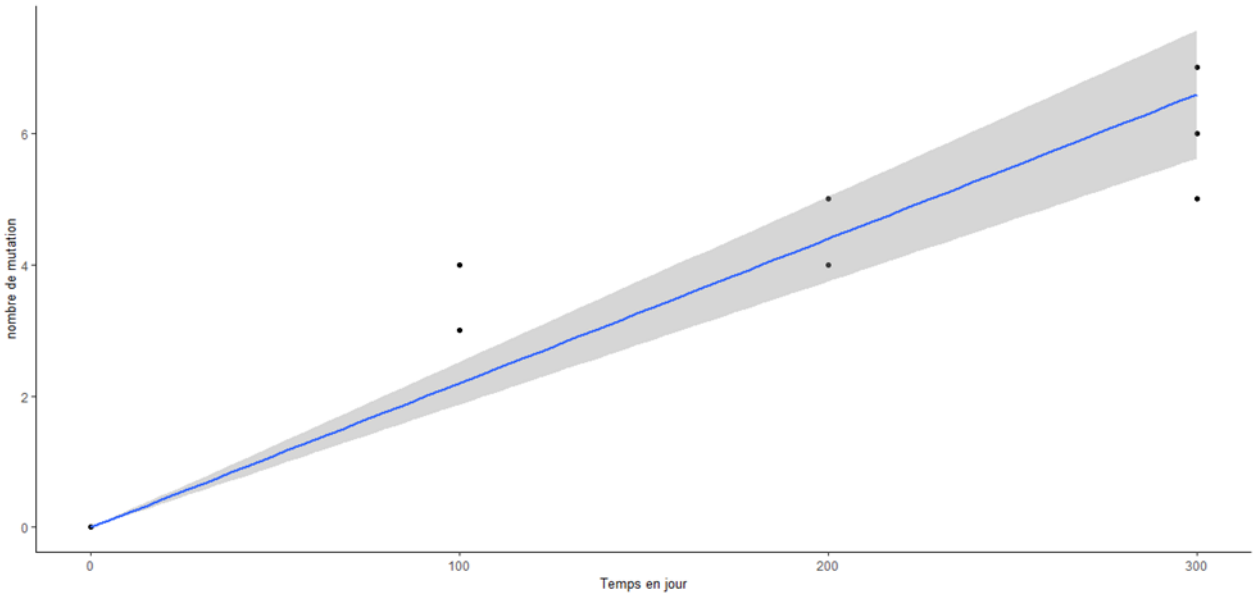
1-1/Approche par régression linéaire

Le modèle de régression linéaire permet d'établir une relation linéaire entre une variable expliquée (dans notre cas le taux de mutation) une variable explicative (dans notre cas le temps). Nous obtenons ainsi un nuage de points à travers lequel nous pouvons tracer une droite de formule y=bx+a avec y le nombre de mutation, x le temps de suivi des isolats, b la pente de la droite et a une constante égale à 0 dans notre cas dans la mesure où la souche ancestrale à l'origine de toutes les lignées qui, par définition, n'a pas de mutation.

Ainsi on peut calculer le taux de mutation en calculant la pente de la droite :

$b = [n(\sum_{i=1}^{n} y_i \quad x_i) - (\sum_{i=1}^{n} x_i)\,(\sum_{i=1}^{n} y_i)]/[n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2]$ (Equation 4. Calcul du taux de mutation par la pente de la droite de régression)

x_i et y_i les coordonnées de la mutation i pour un nombre total de n isolats. La pente sera ensuite normalisée par le nombre de sites et le temps pour un taux de mutation par unité de temps par site.



lignées	NA	1	2	3	4	1	2	3	4	1	2	3	4
isolat ID	souche ancestrale	1	2	3	4	5	6	7	8	9	10	11	12
temps en jour (x)	0	100	100	100	100	200	200	200	200	300	300	300	300
nombre de mutation (y)	0	3	2	0	4	5	4	4	5	5	6	6	7

Figure 10. Illustration de l'approche par régression logistique

Soit 4 lignées dont on prélève un isolat à 3 points de temps différents (J100, J200 et J300), soit 12 isolats séquencés pour lesquels le nombre absolu (la même mutation peut être partagée entre deux isolats d'une même lignée) de mutations synonymes représenté sur la figure et reporté dans le tableau ci-dessus (pour des raisons de simplicité, nous n'avons pas tenu compte du biais mutationnel de transition et transversion).

Si le jeu de donné est suffisamment conséquent, une régression pourrait être effectuée pour chaque type i de mutation. Cependant dans de nombreux cas, nous avons peu de mutations car nous travaillons sur des durées d'évolution relativement courtes. Afin de n'estimer qu'un seul taux de mutation, nous fixons les biais de mutation qui peuvent être estimés sur les synonymes uniquement ou sur l'ensemble de mutations observées. Une régression unique peut alors être faite mais chacune des mutations devra être pondérée en fonction des biais mutationnels. On écrit alors :

$\overline{s_{inx}} = \frac{s_{inx}}{(S_i \ . \ \varepsilon_i)}$ (Equation 5. Pondération de mutations synonymes en fonction du nombre de site nucléotidique susceptible d'aboutir à une mutation synonyme et le biais

mutationnel) avec $\overline{s_{inx}}$ le nombre de mutations de type i sur un isolat n à un temps x, pondéré par le nombre de sites S_i susceptible de voir émerger une mutation de type i et ε_i, la pondération relative du au biais mutaiton.

1-2/Approche par maximum de vraisemblance suivant la loi de Poisson

Le taux de mutation, en d'autres termes, l'émergence de mutations dans le temps est un phénomène aléatoire. Afin de décrire le comportement du phénomène d'apparition de ces mutations, de les modéliser, il convient d'avoir recours à une loi de probabilité. L'apparition de mutations dans le temps, illustré par le taux de mutation, est un phénomène rare indépendant du temps écoulé depuis la dernière mutation dans un intervalle de temps fixé qui suit en conséquence une loi de Poisson. La loi de Poisson peut être modélisée par l'équation $\mathrm{p(k)} = \frac{\lambda^k}{k!} e^{-\lambda}$ (Equation 6. Modélisation de la loi de Poisson), p(k) étant la probabilité que k occurrences du phénomène se produisent dans l'intervalle de temps et λ l'occurrence moyenne du phénomène pendant ce délai. Ce modèle permet de décrire la densité de probabilité d'un taux de mutation connu en faisant varier λ et permet d'obtenir des intervalles de confiance sur la valeur estimée. Or, le problème est que nous disposons des variations de k (nombre de mutations par isolat dans un intervalle de temps donné) et que nous cherchons à déterminer λ (le nombre moyen d'occurrences de mutation dans le temps fixé, le taux de mutation). Quel serait la valeur de λ pour laquelle la densité de probabilité serait la plus vraisemblable au vu des valeurs de k dont nous disposons ?

L'estimateur du maximum de vraisemblance est un estimateur statistique utilisé pour inférer les paramètres de la loi de probabilité d'un échantillon donné en recherchant les valeurs des paramètres maximisant la fonction de vraisemblance (L). L'équation peut s'écrire ainsi : $L(\mathrm{s};\lambda) = e^{-\lambda} \frac{\lambda^s}{s\ !}$ (Equation 7. Estimation du maximum de vraisemblance d'un paramètre λ suivant la loi de Poisson). Avec s le nombre total de mutations synonymes

observées. La valeur de λ estimée représente alors $\lambda = \mu St$ (Equation 8. Estimation λ du nombre de mutation survenant durant un intervalle t). Avec μ le nombre de mutations par génération, S le nombre de sites synonymes et t le temps total sur lequel l'ensemble des clones ont évolué.

Nous souhaitons maintenant intégrer le biais lié aux biais de mutations ε_i dans l'estimation du taux de mutation par le maximum de vraisemblance suivant la loi de Poisson. Nous pouvons décomposer l'équation suivante en fonction des types de mutations i et estimer un taux pour chaque type de mutation

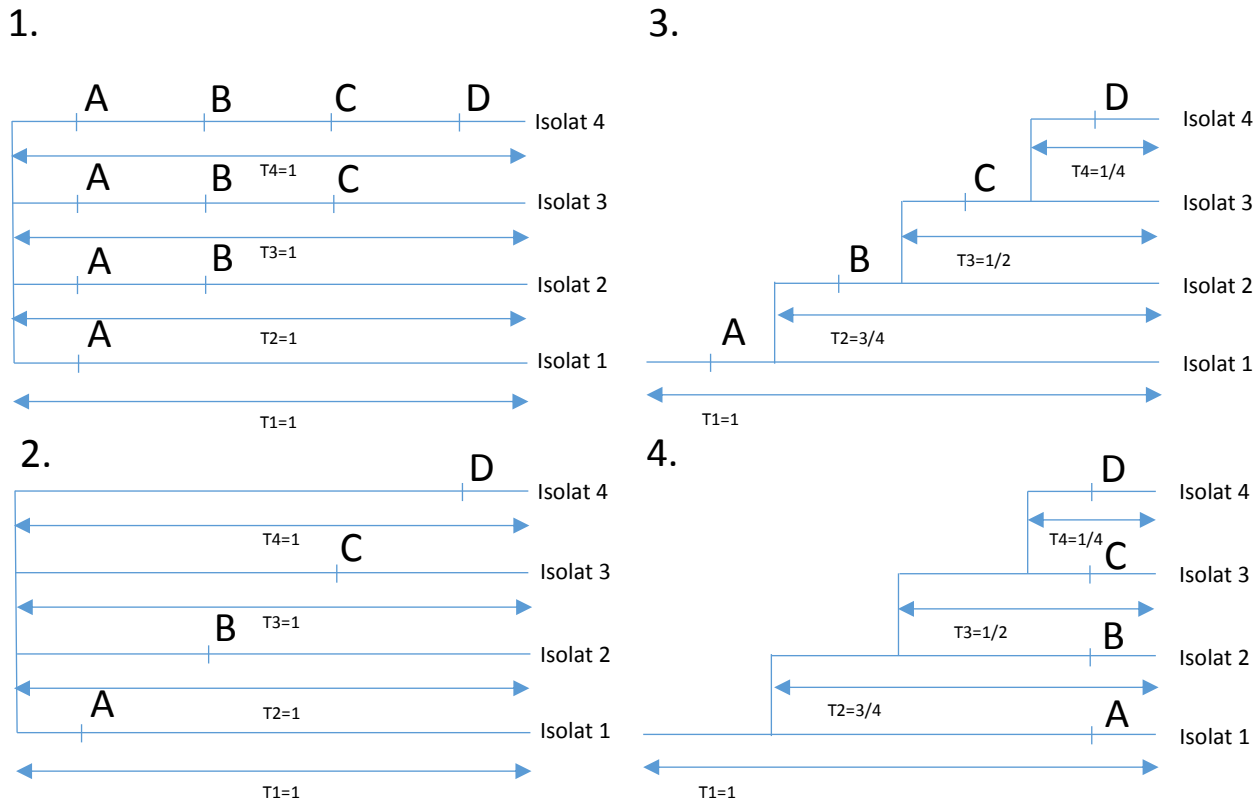
$L(s_1, \dots, s_i, \dots, s_6; \lambda_1, \dots, \lambda_i, \dots, \lambda_6) = \prod_{i=1}^{6} e^{-\lambda_i} \frac{{\lambda_i}^{s_i}}{s_i\ !}$ (Equation 9. Estimation du taux de mutation par le maximum de vraisemblance suivant la loi de Poisson)

Ainsi $\lambda_i = \mu_i S_i t$ μ_i étant le taux de mutation de type i par site par unité de temps, S_i étant le nombre de sites synonymes pour le type de mutation i et t le temps d'accumulation des mutations observées. Si nous n'avons que peu de mutations, il peut être préférable de n'estimer qu'un taux et de prendre en compte les biais précédemment estimés.

$L(s_1, \dots, s_i, \dots, s_6; \mu) = \prod_{i=1}^{6} e^{-\mu\varepsilon_i S_i t} \frac{(\mu\varepsilon_i S_i t)^{s_i}}{s_i\ !}$ (Equation 10. Estimation de chaque taux de mutation i par le maximum de vraisemblance suivant la loi de Poisson intégrant la pondération par le biais mutationnel ε)

2/Mode de calcul pour les lignées non indépendantes (approche Bayésienne)

Pour calculer le taux de mutation à partir d'isolats avec un lien généalogique, il convient d'utiliser une autre approche que celle décrite précédemment. Les approches précédentes sont basées sur le calcul d'une tendance à l'accumulation de mutations dans des lignées qui sont indépendantes. Si nous appliquions cette approche, en considérant que les mutations n'apparaissent qu'entre les deux points de prélèvements, nous pourrions sous-estimer ou surestimer le taux de mutation. Si on considère un lien généalogique, il faut déterminer la



<u>Figure 11.</u> Illustration du biais de méconnaissance de la structure de la population
Soit 4 mutations (A, B, C et D) apparues sur 4 isolats prélevés à deux points de temps séparé par une unité de temps Ti (temps de suivi de l'isolat i). Dans le cas de figure 1. et 3., nous observons des mutations partagées entre les isolats mais en considérant leur évolution indépendante (1.) ou avec un lien généalogique respectivement (3.). Nous obtenons un taux de mutation (μ) respectivement de μ=10/4=2,5 et μ=4/(1+1/2+3/4+1/4)=1,6. Dans le cas de figure 2. et 4., nous observons des mutations non partagées probablement terminale entre les isolats, en considérant encore une fois leur évolution indépendante (2.) ou avec un lien généalogique respectivement (4.). Nous obtenons un taux de mutation (μ) respectivement de μ=4/4=1 et μ=4/(1+1/2+3/4+1/4)=1,6. La méconnaissance de la structure de la population conduit à une mésestimation du taux de mutation.

généalogie sous-tendant les liens entre isolats et donc calibrer l'âge des nœuds (ancêtres) dans la généalogie car la distribution de ces âges contrôle la longueur totale de la généalogie et donc le nombre de mutations qui peuvent y être apparu. Une approche Bayésienne reposant sur la théorie de la coalescence peut être utilisée. Celle-ci infère l'âge des nœuds de la généalogie, inconnue *a priori*, à partir d'une hypothèse de distribution a priori inférée par la théorie de la coalescence auquel on ajoute une hypothèse d'horloge moléculaire pour l'accumulation des mutations.

Les méthodes phylogénétiques bayésiennes ont été introduites dans les années 1990 (Rannala and Yang 1996) et ont depuis révolutionné la façon dont nous analysons les données de séquence génomique (Huelsenbeck *et al.* 2001). Son utilisation s'est dès lors largement répandue avec entre autres l'analyse phylogéographique de la propagation de virus chez l'homme (Wilfert *et al.* 2016; Pybus *et al.* 2012; Faria *et al.* 2014), l'histoire phylogéographique et la migration entre espèces (Bloomquist, Lemey, and Suchard 2010; Nascimento *et al.* 2013), l'analyse des taux de diversification des espèces (Hoorn *et al.* 2010) ou l'estimation du temps de divergence (dos Reis *et al.* 2015; Prum *et al.* 2015). La méthode bayésienne est une méthodologie d'inférence statistique. Sa principale caractéristique est l'utilisation de distributions de probabilité pour décrire l'incertitude de toutes les inconnues, y compris le ou les paramètres du modèle. Soit D les données observées, f(D) la probabilité d'observer ces données et θ le paramètre inconnu. Nous affectons une distribution f (θ), appelée distribution *à priori*, basée sur notre connaissance de θ avant l'analyse des données. Une fois les données observées, nous utilisons le théorème de Bayes pour calculer la distribution *a posteriori* de θ à partir des données:

$f(\theta|D) = \frac{1}{z} f(D)f(D|\theta)$ où la probabilité des données étant donné le paramètre, f (D | θ), est appelée la vraisemblance (likelihood). Ceci résume les informations sur θ dans les données. La constante de normalisation $z = \int f(\theta)\, f(D \mid \theta)\, d\theta$ garantit que f (θ | D) s'intègre à 1 et constitue une distribution statistique appropriée. L'équation indique que le postérieur est proportionnel aux temps antérieurs à la vraisemblance, ou que le postérieur combine les informations du *prior* dans les données (Nascimento, Reis, and Yang 2017).

En pratique, l'analyse Bayésienne se fait en utilisant la chaine de Monte Carlo Markov (Monte Carlo Markov Chain, MCMC), une technique de simulation pour échantillonner à partir d'une

distribution de probabilité connue jusqu'à une constante de normalisation. Il est à noter que tous les termes du côté droit de l'équation sont faciles à calculer, à l'exception de la constante de normalisation z, qui implique des intégrales multidimensionnelles et peut être trop lourd à calculer et réclame donc un outil bio-informatique performant.

VI/Recherche de traces de sélections : le Ka/Ks

1/Principe et forme classique du calcul Ka/Ks

Le rapport entre le taux de substitution non-synonyme (Ks) et synonyme (Ka), le Ka/Ks, parfois noté dN/dS, est un marqueur de la sélection. Comme nous l'avons vu plus haut, théoriquement, quelque soit le régime de sélection le taux de mutations synonymes est constant car la survenue de telles mutations n'entraine pas de changements d'acides aminés et donc de structures de la protéine vecteur de l'expression phénotypique. L'expression parfaitement neutre des mutations synonymes est à nuancer, mais l'effet sur la sélection semble toutefois assez marginale. Concernant le taux de mutations non-synonymes, si elles entrainent un changement d'acide aminé neutre, il sera fixé au même taux qu'une mutation synonyme, le Ka/Ks sera égal à 1. Si le changement d'acide aminé est délétère, la sélection purificatrice réduira le taux de substitutions non-synonymes, le Ka/Ks sera inférieur à 1. Ce n'est que lorsque les changements d'acide aminé offre un avantage sélectif qu'il est fixé à un taux supérieur à celui d'une mutation synonyme, avec un Ka/Ks supérieur à 1. Nous nous attendons donc que l'accumulation de mutations synonymes illustre le comportement mutationnel d'un individu et que l'accumulation des mutations non-synonymes illustre plutôt les pressions de sélections. Le calcul de ce ratio n'est pas aussi simple et nécessite de prendre en compte les biais mutationnels vu plus haut en les intégrants à différentes approches.

Une première méthode est de faire le rapport brut, à partir du jeu de donnée total, entre l'accumulation de mutations non-synonymes n et synonyme s en pondérant simplement par le

nombre de sites de mutations non-synonymes N et synonyme S pour chaque type de mutation i (Table). On a alors :

$\frac{Ka}{Ks} = \frac{\sum_{i=1}^{6} n_i \quad \frac{S_i}{N_i}}{\sum_{i=1}^{6} s_i}$ (Equation 11. Rapport Ka/Ks brut pondéré par le nombre de sites non-synonyme et synonyme)

Mais ce taux qui ne prend pas en compte les biais est source d'erreur, car comme nous l'avons vu les transitions sont plus fréquemment synonymes que les transversions. Une façon plus rigoureuse de mesurer de taux est de contrôler pour chaque type de mutation avec la formule suivante :

$\frac{Ka}{Ks} = \frac{\sum_{i=1}^{6} n_i}{\sum_{i=1}^{6} s_i \frac{N_i}{S_i}}$ (Equation 12. Rapport Ka/Ks dont l'accumulation des mutations synonyme est pondéré par le rapport du nombre de site non-synonyme sur le nombre de site synonyme)

Le numérateur représente le nombre de mutations non synonymes observées alors que le dénominateur fait la somme pour chaque type de mutations de nombre de mutations non synonymes attendu étant donné les taux de mutations synonymes (s_i/S_i). Si peu de mutations synonymes sont observées on peut calculer

$$\frac{Ka}{Ks} = \frac{\sum_{i=1}^{6} n_i \frac{S_i}{N_i}}{\sum_{i=1}^{6} s_i}$$

Ou l'on compare cette fois le nombre de mutations synonymes observées au nombre de mutations synonymes attendu étant donnes les taux de mutations non synonymes mesurés.

L'intervalle de confiance à 95% peut être obtenu, par boostrap. Si $K = \sum_{i=1}^{6}(s_i \quad + \quad n_i)$ mutations ont été observées au total, on peut en échantillonner avec remise K mutations parmi les mutations observées en fonction de leur statut (synonyme non synonyme) et de leur type. La probabilité de survenue p_i d'une mutation de type i est $p_i = \frac{s_i \quad ou \quad n_i}{\sum_{i=1}^{6}(s_i \quad + \quad n_i)}$, et le processus

d'échantillonage suit en fait une loi multinomiale. En pratique, cet intervalle est déterminé par des procédés bio-informatiques en utilisant par exemple la fonction *rmultinom* dans R. Ce calcul a le mérite de la simplicité et est particulièrement bien adapté pour de faibles nombres de mutations observées.

2/Distribution des mutations suivant la loi de Poisson et estimation du Ka/Ks par le maximum de vraisemblance

Le Ka/Ks peut aussi être estimé en recherchant les occurrences de mutations suivant une loi de poisson et maximisant la fonction de vraisemblance (L). Si nous notons $\omega = Ka/Ks$ alors

$$L(\mathrm{s}_1, \dots, \mathrm{s}_\mathrm{i}, \dots, \mathrm{s}_6, \mathrm{n}_1, \dots, \mathrm{n}_\mathrm{i}, \dots, \mathrm{n}_6; \mu, \omega) = \prod_{i=1}^{6} e^{-\mu\varepsilon_i(\mathrm{S}_i+\omega\mathrm{N}_i)\mathrm{t}} \frac{(\mu\varepsilon_i(\mathrm{S}_i + \omega\mathrm{N}_i)\mathrm{t})^{s_i+\mathrm{n_i}}}{(s_i + \mathrm{n_i}) \quad !}$$

(Equation 13. Estimation du Ka/Ks par maximum de vraisemblance dont l'accumulation des mutations suit une loi de Poisson)

Nous avons utilisé une méthode d'optimisation pour minimiser le log(L) pour les deux paramètres μ et ω. Il s'agit de rechercher la valeur minimale de la fonction log(L) sur plusieurs itérations pour différentes valeurs de μ et ω. Il est cependant nécessaire d'inférer des valeurs de départ pour les deux paramètres en fonction de l'hypothèse posé en début de l'étude. Plus simplement, l'optimisation de la fonction log(L), en posant l'hypothèse de départ pour les valeurs μ et ω peut se faire à partir de la fonction *optim* dans R.

3/Calcul du Ka/Ks par approche par régression linéaire

Si nous pouvons faire l'hypothèse que les mutations s'accumulent dans le temps, nous pouvons aussi utiliser une approche par régression pour estimer les taux d'accumulations des mutations synonymes et non synonymes et d'en inférer le ratio. Il suffit de faire le rapport des taux de mutations non-synonymes sur le taux de mutation synonyme en pondérant par le nombre de sites impliqués.

$$\omega = \frac{[6\left(\sum_{i=1}^{6} \frac{n_{it}}{(N_i \cdot \varepsilon_i)} \quad t_i\right) - (\sum_{i=1}^{6} t_i)(\sum_{i=1}^{6} \frac{n_{it}}{(N_i \cdot \varepsilon_i)})]/[6(\sum_{i=1}^{6} t_i^2) - (\sum_{i=1}^{6} t_i)^2]}{[6\left(\sum_{i=1}^{6} \frac{s_{it}}{(S_i \cdot \varepsilon_i)} \quad t_i\right) - (\sum_{i=1}^{6} t_i)(\sum_{i=1}^{6} \frac{s_{it}}{(S_i \cdot \varepsilon_i)})]/[6(\sum_{i=1}^{6} t_i^2) - (\sum_{i=1}^{6} t_i)^2]}$$ (Equation 14. Estimation du Ka/Ks utilisant une approche par régression linéaire) avec n_{it} et s_{it} le nombre de mutation non-synonyme et synonyme de type i (Table) au temps t et t_i le temps de survenue des mutations s_i et n_i correspondantes.

CONCLUSION DU CHAPITRE III

Nous avons vu qu'il existe une première approche qualitative de l'étude l'adaptation à la recherche de mutations convergentes au niveau du gène ou d'un opéron avec un impact fonctionnel. La recherche de la diversité et de trace de sélection peut se faire par une approche quantitative. Avec la taille de population efficace, le taux de mutation est curseur de la diversité permettant d'illustrer la sélection en jeu subie par la souche dans un milieu donné. Mais le calcul du taux de mutation est adapté au mode d'observation de ces mutations et nécessite de connaitre certains biais qu'il convient de corriger.

Le taux de mutation se calcule avec les mutations ponctuelles synonymes car elles sont censées être peu touchées par la sélection. Les mutations non-synonymes peuvent s'ajouter au calcul en l'absence, ou pondérées par un indice de sélection. Il existe des différences de probabilité de survenue de mutations en fonction des bases devant être mutées et le mode de mutation en transition ou transversion. Le calcul du taux de mutation nécessite ainsi de prendre en compte le nombre de sites de mutation pour chaque type de mutation et le biais mutationnel. Les mutations n'ont donc pas la même valeur.

Le calcul du taux de mutation nécessite tout d'abord de déterminer si les isolats ont un lien généalogique ou non. Le premier cas se rencontre en général dans les observations *in natura* avec les prélèvements d'isolats évoluant dans le même milieu, alors que le deuxième cas se rencontre plutôt dans les expérimentations *in vivo* faisant évoluer des lignées indépendantes à

partir d'un ancêtre commun. L'ancêtre commun étant inconnu dans les expérimentations *in natura*, une approche Bayésienne permet d'inférer les points de divergence des isolats (apparition d'une mutation) prélevé dans le milieu et semble plus adapté pour déterminer un taux de mutation à partir d'isolats ayant un lien généalogique. Pour le calcul du taux de mutation des lignées indépendantes, deux approches sont possibles se différenciant par l'hypothèse de distribution des mutations dans le temps. On peut adopter une approche par régression linéaire si on considère que les mutations apparaissent suffisamment fréquemment ou considérer qu'il s'agit d'un phénomène rare suivant une loi de distribution de Poisson dont le paramètre devra être déterminé par un maximum de vraisemblance.

Le Ka/Ks, pouvant être noté ω, est un outil permettant de déterminer le régime d'évolution (sélection positive, purificatrice ou évolution neutre) d'une souche au sein de son milieu. Cet outil repose sur le rapport entre l'accumulation de mutations non-synonymes, responsable d'un changement d'expression de son acide aminé susceptible d'être sélectionné ou éliminé ou même s'accumuler de manière neutre en fonction de l'avantage ou désavantage fourni à la souche vis-à-vis de son environnement, et l'accumulation de mutations synonymes sensées théoriquement s'accumuler de manière neutre. Le calcul du Ka/Ks peut se faire par un rapport brut mais aussi comme précédemment par une approche par régression linéaire ou suivant la loi de Poisson dont le paramètre devra être déterminé par un maximum de vraisemblance.

DESCRIPTION DES QUESTIONS SCIENTIFIQUES

E. coli est un pathogène redoutable que chaque individu abrite dans son tube digestif. Le tube digestif est un réservoir de bactéries commensales potentiellement pathogènes. L'étude du commensalisme semble indispensable, la virulence étant un sous-produit de celui-ci. Il semble nécessaire de s'intéresser aux déterminants de l'adaptation de *E. coli* soumise à tant de pressions provenant du reste du microbiote, de l'hôte ou de facteurs extérieurs au tube digestif.

L'adaptation de *E. coli* dans le tube digestif a fait l'objet d'études d'évolution qui comportaient plusieurs limites dont l'utilisation d'isolats de K-12 de laboratoire évoluant dans des environnements relativement éloignés de ce qui peut s'observer dans la nature. Ces études ne prenaient pas en compte d'autres facteurs comme l'alimentation.

E. coli évolue dans un système ouvert exposé aux facteurs extérieurs comme les antibiotiques et surtout l'alimentation. L'effet de l'exposition prolongée des antibiotiques a été particulièrement bien étudié au vu du problème de santé publique majeur qu'il pose, ce n'est pas le cas de l'alimentation. L'alimentation engendre des modifications au sein du microbiote et change l'équilibre entre celui-ci et son hôte avec des conséquences sur la santé. La question se pose alors de l'adaptation de *E. coli* à ce nouvel environnement et du rôle joué par son fond génétique.

Nous avons mené trois études longitudinales une *in natura* et deux *in vivo* dans un modèle murin permettant de caractériser le mode d'adaptation de *E. coli* dans son milieu naturel, ainsi que l'impact combiné du fond génétique de la souche et de l'alimentation.

(i) Une première étude longitudinale observationnelle *in natura* suivant, pendant à peu près un an, *E. coli* ED1a dans le tube digestif d'un homme occidental.

(ii) Une deuxième étude expérimentale *in vivo* sur modèle murin suivant, pendant plus d'un an, *E. coli* 536 évoluant dans un environnement digestif que nous avons tenté de rapprocher le plus possible d'un environnement naturel et soumis à deux régimes alimentaires différents. Afin d'éviter l'emploi de souris axéniques ou de streptomycine permettant à notre souche d'intérêt de coloniser le tube digestif des souris, nous avons utilisé un modèle de colonisation du tube digestif mimant la transmission du microbiote de la mère à l'enfant.

(iii) Une troisième étude expérimentale *in vivo* sur modèle murin suivant, pendant plus d'un an, deux souches de *E. coli* 536 et HS évoluant dans un tube digestif de souris traitées à la streptomycine et soumises à deux régimes alimentaires différents.

PARTIE EXPERIMENTALE

CHAPITRE I : ETUDE DE L'ADAPTATION D'UN CLONE DOMINANT DE *E. COLI* ED1A DANS LE TUBE DIGESTIF D'UN HOMME OCCIDENTAL

I /Introduction

Plusieurs expérimentations évolutives *in vitro* nous ont permis de différencier les différentes stratégies d'adaptation de *E. coli* dans des conditions non naturelles voire extrêmes. Mais afin de comprendre le rythme et le mode d'adaptation de *E. coli* se produisant dans la nature, il s'est avéré nécessaire d'avoir recours aux expérimentations évolutive *in vivo* sur modèle murin. Cependant, l'étude de l'adaptation de *E. coli* dans le tube digestif est biaisée par l'utilisation de streptomycine ou de souris axéniques afin de garantir une niche de colonisation et donc le maintien de la souche d'intérêt. Par ailleurs, l'utilisation de souches de laboratoire non naturelles K-12 et d'un seul régime alimentaire ne permettaient pas d'extrapoler les résultats de ces études. Reeves *et al.* ont pu suivre pendant 3 ans une souche de *E. coli* proche de CFT073 au sein d'un foyer comptant 5 membres dont 1 chien mettant en évidence un faible taux de mutation, une faible diversité, peu de traces de sélection et un fort taux de transmission.

Le fait de retrouvé inconstamment le clone de *E. coli* chez un même individu à différent point de temps et le changement régulier d'environnement ne permettait pas d'étudier de manière rigoureuse l'impact du tube digestif *in natura* sur l'adaptation de *E. coli*. Afin de pallier ces limites, nous avons pu avoir l'opportunité de suivre un clone dominant de *E. coli* ED1a appartenant au groupe phylogénétique B2 retrouvé constamment dans le tube digestif d'un individu sain pendant à peu près un an. Le suivi d'un clone dominant de *E. coli* dans son environnement naturel nous a permis de déterminer le mode d'évolution naturel de cette souche constante chez un même individu, de déterminer un taux de mutation, d'estimer un taux de réplication et la taille de la population efficace.

II/Article n°1 (Publié dans Applied and Environnemental Microbiology en 2018)

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

GENETICS AND MOLECULAR BIOLOGY

Evolution of a Dominant Natural Isolate of *Escherichia coli* in the Human Gut over the Course of a Year Suggests a Neutral Evolution with Reduced Effective Population Size

Mohamed Ghalayini,[a,b,c] Adrien Launay,[a,d] Antoine Bridier-Nahmias,[a,d] Olivier Clermont,[a,d] Erick Denamur,[a,d,e] Mathilde Lescat,[a,b,c] Olivier Tenaillon[a,d]

[a]INSERM, IAME, UMR 1137, Paris, France
[b]Université Paris Nord, Sorbonne Paris Cité, Bobigny, France [c]AP-HP, Hôpitaux Universitaire Paris Seine Saint-Denis, Bondy, France [d]Université Paris Diderot, Sorbonne Paris Cité, Paris, France
[e]AP-HP, Hôpitaux Universitaire Paris Nord Val de Seine, Paris, France

ABSTRACT *In vitro* and *in vivo* evolution experiments on *Escherichia coli* revealed several principles of bacterial adaptation. However, few data are available in the literature describing the behavior of *E. coli* in its natural environment. We attempted here to study the evolution in the human gut of a commensal dominant *E. coli* clone, ED1a belonging to the B2 phylogroup, through a longitudinal genomic study. We sequenced 24 isolates sampled at three different time points within a healthy individual over almost a year. We computed a mutation rate of 6.90×10^{7} mutations per base per year of the chromosome for *E. coli* ED1a in healthy human gut. We observed very limited genomic diversity and could not detect any evidence of selection, in contrast to what is observed in experimental evolution over a similar length of time. We therefore suggest that ED1a, being well adapted to the healthy human gut, evolves mostly neutrally with a low effective population size (N_e of 500 to 1,700).

IMPORTANCE In this study, we follow the genomic fate of a dominant clone of *Escherichia coli* in the human gut of a healthy individual over about a year. We could compute a low annual mutation rate that supports low diversity, and we could not retrieve any clear signature of selection. These observations support a neutral evolution of *E. coli* in the human gut, compatible with a very limited effective population size that deviates drastically with the observations made previously in experimental evolution.

KEYWORDS ED1a, *Escherichia coli*, human gut, molecular evolution, mutation rate, neutral evolution, replication rate

Received 29 October 2017 **Accepted** 22 December 2017

Accepted manuscript posted online 5 January 2018

Citation Ghalayini M, Launay A, Bridier-Nahmias A, Clermont O, Denamur E, Lescat M, Tenaillon O. 2018. Evolution of a dominant natural isolate of *Escherichia coli* in the human gut over the course of a year suggests a neutral evolution with reduced effective population size. Appl Environ Microbiol 84:e02377-17. https://doi.org/10.1128/AEM.02377-17.

Editor Marie A. Elliot, McMaster University

Address correspondence to Mohamed Ghalayini, mohamed.ghalayini@inserm.fr.

M.G. and A.L. contributed equally to this work.

Resistance to antibiotics has sadly revealed that microbial evolution is an active process that can be witnessed in the short term (1–4). Antibiotic pressures represent an extreme form of selection. As a result, strong and fast selective responses are expected and documented. Much less is known about the pace of microbial evolution in the wild under milder selective pressures, even for bacterial species of medical interest such as *Escherichia coli*, a commensal of the gut and a versatile pathogen.

In vitro and *in vivo* evolution experiments on *E. coli* have revealed several principles of bacterial adaptation: (i) selection regularly favors an increased mutation rate that resulted in an improved adaptation rate (5–7), (ii) global regulators with pleiotropic effects are often recruited in the first stages of adaptation (8–11), (iii) lineages evolved under similar conditions evolve frequently through modification of the same genes albeit with different mutations (12–14), (iv) traces of selection overwhelm the genomic changes observed (6), and, finally, (v) the importance of selection is linked to maladaptation of the strain used in the experiment (14). Though globally consistent across experiments, these observations emerged nevertheless under artificial conditions. Even in the case of evolution in the mouse gut (11, 14), an environment much closer to *E. coli*'s habitat than test tubes, the caged life of the animals, the consistency of their diet, or the use of streptomycin to maintain *E. coli* are factors that are driving the process away from natural conditions. Hence, the relevance of these principles of adaptation under natural conditions, especially under human commensal conditions, remains questionable.

E. coli is a versatile species, known as both a widespread gut commensal of the vertebrates and a dangerous pathogen that can also be retrieved in the environment (15). As a pathogen *E. coli* is responsible for about 1 million human deaths yearly due to intraintestinal and now mostly extraintestinal diseases (16). Among the seven phylogenetic groups, A, B1, B2, C, D, E, and F (17), that compose the species, B2 and D are the ones most often associated with extraintestinal diseases. The great majority of *E. coli* strains belonging to genetic group B2 are highly virulent in a mouse model of extraintestinal infection (18). Yet one subgroup (subgroup VIII, corresponding to sequence type 452 [ST452] or ST149 using the Achtman or Pasteur Institute scheme, respectively, [19], and exhibiting the O81:H27 serotype) within this group does not exhibit any extraintestinal virulence in a mouse model of sepsis (20).

Moreover, this subgroup, which includes archetypal clone *E. coli* ED1a, happens to be specific to the human gut (21). When present, strains from that nonpathogenic B2 group are dominant compared to other strains of the phylogenetic group B2 and persist for long periods in the human gut. This suggests that these strains perform well in the human gut (21).

To circumvent the limitation of the *in vitro* and *in vivo* studies of experimental evolution, we attempt here to study the evolution of *E. coli* in its natural environment, the human gut, through a longitudinal genomic study. This approach was previously used in a 3-year-long study of the evolution and transmission of an *E. coli* clone in a household of six members, including a dog. The investigators followed a B2 clone (named clone D) that was not dominant but that could be recovered 14 times at different time points. They showed a high transmission rate of the clone within the household, a low mutation rate, limited diversity, and no clear traces of adaptation (22). In the present study, we focus on the evolution of a dominant clone retrieved over a 1-year period within a healthy individual.

To study the evolution and adaptation in the human gut of the commensal clone ED1a, we sequenced 24 isolates sampled at three different time points within an individual. The aim of our retrospective study was to identify the evolutionary forces shaping the evolution of its genome in the human gut over almost a year.

RESULTS

Dominance and persistence of *E. coli* ED1a. Eight isolates of *E. coli* were randomly selected after plating the feces on Drigalski medium at three time points: day 0 (D0), day 211 (D211), and day 315 (D315). All isolates belonged to the phylogenetic group B2, subgroup VIII, and exhibited the O81 type, confirming their belonging to the ED1a clone (21). The lack of alternative *E. coli* genotypes suggests that the *E. coli* niche was dominated, if not fully invaded, by the ED1a clone. This also suggests a persistence of this dominance over almost 1 year in healthy human gut although the number of clones sampled per time point (n = 8) was relatively small.

Description of *E. coli* ED1a evolution. *E. coli* ED1a, in addition to its chromosomic genome (5,209,548 bp), possesses a conjugative plasmid, pECOED, of 119,594 bp (23). On the 24 isolate genomes analyzed by the breseq pipeline, we identified 17 different mutations, among which 2 were located on the plasmid. We observed eight mutations among the eight isolates from D0 (including two deletions, four nonsynonymous single nucleotide polymorphisms [SNPs], and two synonymous SNPs),

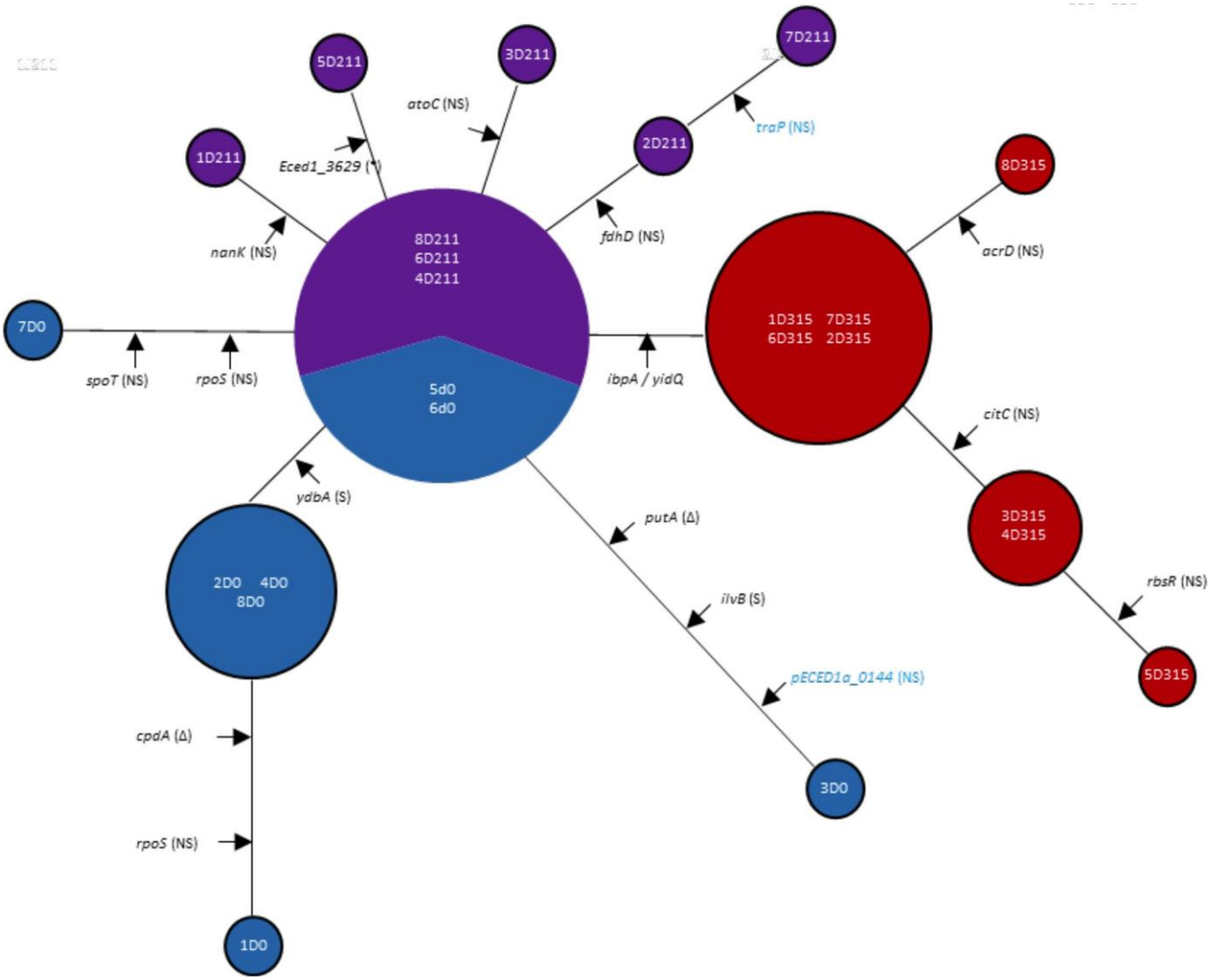


FIG 1 Genomic diversity of *E. coli* ED1a over almost a year in a healthy human gut. We represented a maximum likelihood unrooted tree relating all isolates sampled over a year reconstructed with the packages ape (55) and phangorn (56) in R software (34). The length of the branches is proportional to the number of mutations separating two samples. The diameters of circles are proportional to the numbers of isolates that are identical. Color codes are as follows: blue, isolates sampled at day 0 (D0); violet, isolates sampled at day 211 (D211); red, isolates sampled at day 315 (D315). The intergenic mutations are represented by two genes separated by a slash. The nonsynonymous and synonymous SNPs are marked by (NS) and (S), respectively. Deletion mutations (12 bp in the *putA* gene and 3 bp in a repeated region of the *cpdA* gene) are indicated with a Greek delta (Δ). A stop codon is indicated with an asterisk (*). The mutations occurring on the plasmid are shown in blue.

five mutations among the eight isolates from D211 (including four nonsynonymous SNPs and one SNP giving a stop codon), and four mutations (including three nonsynonymous SNPs and one SNP in intergenic region) among the eight isolates from D315. Several isolates had identical genomes (Fig. 1). While 13 out of the 17 mutations were specific to one isolate (singletons), four mutations were shared by several isolates. These mutations included an intergenic mutation between *ibpA* and *yidQ* that seems to have been fixed between D211 and D315 as all D315 isolates showed this mutation while none had it at D211 (Table 1).

In sum, SNPs represented 15 of the 17 mutations observed (88% of the mutations), with 11 nonsynonymous mutations, 2 synonymous mutations, 1 stop codon mutation, and 1 intergenic mutation. Among the SNPs, there were seven transitions and eight transversions. The two non-SNP mutations were deletions found at D0: a deletion of 3 bp within a repeat region of *cpdA* in isolate 1 obtained on D0 (1D0) and a 12-bp deletion in *putA* in sample 3D0. No large deletion, large insertion, or transposon insertions were detected (Table 1).

TABLE 1 Summary of mutations that were found in the 24 isolates of *E. coli* ED1a sampled over almost a year in a healthy human gut

Gene[a]	Position (nt)[b]	Mutation	Mutation effect	Type of mutation (no. of instances)	Isolate no.	Gene product(s)
acrD →	2852335	A → C	D164A (GAC → GCC)	Nonsynonymous SNP (1)	8D315	Aminoglycoside/multidrug efflux system
atoC →	2612636	T → G	I129S (ATC → AGC)	Nonsynonymous SNP (1)	3D211	Fused response regulator of *ato* operon in a two-component system with AtoS; response regulator; sigma54 interaction protein
citC ←	637957	T → G	T350P (ACC → CCC)	Nonsynonymous SNP (3)	3D315, 4D315, 5D315	Citrate lyase synthetase
cpdA ←	3623372	(GCA)4 → 3	Deletion of A	In-phase deletion (1)	1D0	3′–5′ Cyclic AMP phosphodiesterase
ECED1_3629 ←	3559530	C → A	E146* (GAG → TAG)[c]	Stop codon SNP (1)	5D211	Conserved hypothetical protein; putative membrane protein
fdhD →	4541860	C → T	A114V (GCG → GTG)	Nonsynonymous SNP (2)	2D211, 7D211	Formate dehydrogenase formation protein
ibpA ←/→ *yidQ*	4305206	C → T	NA	Intergenic SNP (8)	1D315, 2D315, 3D315, 4D315, 5D315, 6D315, 7D315, 8D315	Heat shock chaperone/conserved hypothetical protein; putative outer membrane protein
ilvB ←	4294479	C → A	A88A (GCG → GCT)	Synonymous SNP (1)	3D0	Acetolactate synthase I large subunit
nanK ←	3804472	T → C	T32A (ACG → GCG)	Nonsynonymous SNP (1)	1D211	Putative *N*-acetylmannosamine kinase
putA ←	1173242	Δ12 bp	Deletion of EEWQ	In-phase deletion (1)	3D0	Fused DNA-binding transcriptional regulator; proline dehydrogenase; pyrroline-5-carboxylate dehydrogenase
rbsR →	4375084	T → G	V18G (GTT → GGT)	Nonsynonymous SNP (1)	5D315	DNA-binding transcriptional repressor of ribose metabolism
rpoS ←	3132176	A → C	I198S (ATC → AGC)	Nonsynonymous SNP (1)	1D0	RNA polymerase, sigma S (sigma 38) factor
rpoS ←	3132465	C → T	V102 M (GTG → ATG)	Nonsynonymous SNP (1)	7D0	RNA polymerase, sigma S (sigma 38) factor
spoT →	4255773	G → A	E98K (GAG → AAG)	Nonsynonymous SNP (1)	7D0	Bifunctional (p)ppGpp synthetase II and guanosine-3′,5′-bis pyrophosphate 3′-pyrophosphohydrolase
ydbA →	1543455	G → T	G413G (GGG → GGT)	Synonymous SNP (4)	1D0, 2D0, 4D0, 8D0	Putative autotransported outer membrane protein involved in cell adhesion
pECED1a_0115 ←	85479	C → T	D115N (GAT → AAT)	Nonsynonymous SNP (1)	7D211	Conserved hypothetical protein
pECED1a_0144 →	112511	T → C	S49P (TCC → CCC)	Nonsynonymous SNP (1)	3D0	Hypothetical protein; putative membrane protein

[a]Arrows indicate the orientations of the genes; a slash indicates the intergenic region. The mutations present on the plasmid are preceded by the prefix *pECED1a*. *pECED1a_0115* corresponds to *traP*.
[b]nt, nucleotide.
[c]The asterisk (*) indicates a stop codon.

Analyses of prophage and supplementary plasmid sequences in *E. coli* ED1a. We observed several reads in all isolates that did not map to the ED1a genome reference. We searched for mobile elements that could explain these unmapped reads.

In the reads that did not map on the genome, the only phage we could detect was PhiX174. Given that Illumina recommends use of PhiX174 as a sequencing control in proportions varying between 1% and 40%, these reads are likely to result from a contamination of our fastq files by PhiX reads. This phenomenon is already documented in the literature and seems to be pervasive among sequenced genomes deposited in the Genome Online Database (https://gold.jgi.doe.gov/) (24).

To detect some plasmid DNA, we used plasmidSPAdes on the contigs assembled from the reads that did not map on the ED1a chromosome and large plasmid. The plasmidSPAdes algorithm could not reconstruct any full circular plasmid but, after the sequences were annotated with Prokka, the genes we identified were found to be common among mobilizable plasmids (25). The absence of a good reference for these potentially plasmidic sequences made it impossible to use breseq to detect genetic variations between the isolates. To overcome this hurdle, we used a reference-free SNP detection tool called discoSnp ++ and did not detect any polymorphic locus (26).

We performed a PCR assay with primers specific to the hypothetical plasmid matching an ORF coding a protein belonging to the *mobA-mobL* superfamily and obtained an amplicon in isolates 6D0, 6D211, and 6D315 showing the presence of this hypothetical plasmid from the first isolation of the strain to the end of the experiment almost 1 year later. Thus, we observed a small additional plasmid with no mutation.

***E. coli* ED1a mutation rate.** Mutation rates can be estimated simply by assuming that isolates evolve independently. To do so, we used the different mutations we observed in D211 isolates (four mutations, or five mutations including the plasmid) and in D315 isolates (four mutations), considering D0 and D211 isolates, respectively, to be their ancestors. We hypothesized consequently that mutations at D211 occurred during the interval between D0 and D211 and that mutations at D315 occurred during the interval between D211 and D315. We computed an average chromosomal mutation rate of 2.51×10^{7} mutations per base per year, assuming that we covered the whole chromosomal genome, and a rate of 2.66×10^{7} mutations per base per year taking into account the plasmid (Table 2).

This rough estimate of mutation rate that is commonly used neglects, however, that isolates are not independent but connected through a genealogy. Such genealogy can be inferred assuming the population is homogenous and can consequently be modeled according to coalescent theory. As our isolates are sampled from a single host, this hypothesis seems relevant. We therefore computed the mutation rate using a Bayesian approach implemented in BEAST (27). A mutation rate of 6.90×10^{7} mutations per base per year (95% confidence interval [CI], 3.14×10^{7} to 1.40×10^{6} mutations per base per year) was obtained on the chromosomic genome, and a rate of 7.18×10^{7} mutations per base per year (95% CI, 3.28×10^{7} to 1.44×10^{6} mutations per base per year) was obtained if the plasmid was included (Table 2). These estimates differ significantly from the estimates that rely on the independence of mutations, with the latter being outside the 95% confidence interval of the former.

We therefore applied the BEAST approach to another data set in which the authors followed diversification of a clone within a household over 3 years (22). They used a method that partially took into account the genealogy to compute a rate of 2.26×10^{7} mutations per base per year. However, by calculating the mutation rate based on the coalescent theory, we obtained a mutation rate of 1.40×10^{7} mutations per base per year (95% CI, 2.97×10^{8} to 2.77×10^{7} mutations per base per year), assuming a 5,038,386-bp-length clone D genome (Table 2). These two mutation rate estimations are quite close, but the Bayesian approach seems to be more suitable and accurate; indeed, there remained an uncertainty as to the period of time at which the mutations appeared.

Once the mutation rate per base per year (μ_{est}) was computed, we estimated the rate of replication (G) of *E. coli* ED1a in healthy human gut using the formula $G = \mu_{est}/(\mu_{g} \times 365)$ (see File S1 in the supplemental material), where $_{g}$ is the mutation rate per base per generation of ED1a in the healthy human gut. μ_{g} is unknown; however, several generational mutation rates have been previously reported using genome-wide approaches under *in vitro* conditions (28–30). Even an estimation of the ED1a mutation rate per generation was reported in Foster *et al.* (29). However, the environment seems to play a major role in the occurrence of mutations by generations (31). It is indeed difficult to establish experimentally a mutation rate per generation for *E. coli* because the rate depends on the strain (29) and the culture medium (31). However, we observe that the rates of mutations by generation determined with different methods, different strains, and different nonstressful media have very close values (of the order of a maximum factor of 3.5 between the extreme values). We

can therefore use the extreme estimates for our present study: 0.89×10^{10} (28) and 3.12×10^{10} (29) mutations per base per generation.

TABLE 2 Comparison of mutation rates per base per year and numbers of generations per day estimated by different models of evolution on different data sets from evolution studies[a]

Strain	Phylogenetic group	Independent evolution of lineage (no. of mutations/base/yr [95% CI])	Coalescent theory of evolution		Reference
			No. of mutations/base/yr (95% CI)	No. of generations/day (95% CI)[c]	
E. coli ED1a (chromosome only)	B2	2.51×10^{-7}	6.90×10^{-7} (3.14×10^{-7}–1.40×10^{-6})	6.06 (2.88–12.65)–21.24 (9.66–43.28)	This study
E. coli ED1a (chromosome + plasmid)	B2	2.66×10^{-7}	7.18×10^{-7} (3.28×10^{-7}–1.44×10^{-6})	6.31 (2.88–12.64)–22.10 (10.09–4 4.32)	This study
E. coli clone D	B2	2.26×10^{-7}[b]	1.40×10^{-7} (2.97×10^{-8}–2.77×10^{-7})	1.23 (0.26–2.43)–4.3 (0.91–8.52)	22
E. coli 536	B2	1.21×10^{-6}[b]	NA[d]	NA	14
E. coli REL606	A	2.16×10^{-7} (9.73×10^{-8}–3.4×10^{-7})[b]	NA	NA	28

[a]We calculated the mutation rate per genome per year and per base per year of each of the genomic elements according to two modalities: the coalescent theory (by Bayesian inference using BEAST) and the theory of the independence of the evolution of the lineages.

[b]Data published previously.

[c]The values were computed using a range of estimated mutation rates between 0.89×10^{-10} mutations per base per generation published in Wielgoss et al. (28) and 3.12×10^{-10} mutations per base per generation published in Foster et al. (29).

[d]NA, not available.

Assuming that the number of mutations per base per generation ranges from $0.89x10^{10}$ to $3.12x10^{10}$, we can compute, using the chromosome, a range from 6.06 (95% CI, 2.88 to 12.65) to 21.24 (95% CI, 9.66 to 43.28) generations per day (considering lineages with genealogical links) occurring for the *E. coli* ED1a chromosome in healthy human gut. We also estimated the number of generations per day for clone D followed in a household for 3 years from 1.23 (95% CI, 0.26 to 2.43) to 4.3 (95% CI, 0.91 to 8.52) (Table 2).

Looking for traces of selection during *E. coli* ED1a evolution. We then decided to determine whether the genome sequences could reveal some signal of selection. To compute how the population behaved, we used a population genetics approach. The first step was to compute an estimate of diversity based on the genome sequences. For that purpose, we focused on point mutations on the chromosome as indels or mutations on the plasmid may have different mutation rates. Using only point mutations on the chromosome, the measure of diversity as estimated by the theta of Watterson (θ_w) (32), using the pegas package (33) in R (34), was 1.93+/-1.13, 1.54+/-0.96, and 1.15+/-0.78 for sampling time points D0, D211, and D315, respectively. Using these later estimates, the genome length (L) and a range of mutation rates per base per generation from $0.89x10^{10}$ to $3.12x10^{10}$ (see above), as $\theta_w = 2N_e L\mu_g$, we estimated a range of effective population size (N_e) between 500 and 1,700. This means that the population studied has a standing genetic diversity that is equivalent to an idealized population of constant size N_e of 500 to 1,700 individuals. The observed diversity suggests therefore a small size.

We used Tajima's D to evaluate the deviation of the allele frequency distribution from that expected in a standard neutral evolution model (35). Taking into account point mutations on the chromosome, Tajima's D, using the pegas package (33) in R (34), was 0.84 (P=0.47) at D0, 1.03 (P=0.35) at D211, and 0.43 (P=0.74) at D315. Thus, the evolution of these isolates did not deviate significantly from the neutral evolution model at each time point. Yet the values are all negative and could support selection of bottleneck events.

We further looked at synonymous and nonsynonymous mutations to infer if selection acting on nonsynonymous mutations could lead to a differential evolution between the two categories. Taking all different mutations found in the whole data set, the ratio of the rate of nonsynonymous to that of synonymous changes, K_a/K_s, was 1.3 (95% CI estimated by bootstrapping mutations, 0.51 to infinity). Though the power is quite limited due to the small number of mutations, the value

of the ratio suggests that nonsynonymous mutations have accumulated as synonymous mutations in the different genomes and rejects an important contribution from selection. Furthermore, out of the four mutations that have been recovered in at least two isolates, one is synonymous, one is intergenic, and two are nonsynonymous, which suggests that among mutations that have survived drift to some "high" frequency, there is no excess of nonsynonymous mutations.

Among the four mutations that are not singletons, two mutations invaded the population over the year of the study. A synonymous mutation in the middle of gene *ybdA*, a putative autotransporter, was found in 50% (95% CI, 30% to 70%) of the isolates at D0 and fixed at D211. A mutation in the intergenic region between genes *ibpA* and *yidQ* is present in all isolates at D315, and none is present before. This mutation is in the promoter of *yidQ* that encodes an outer membrane protein whose function remains to be defined. While fixation is likely to reflect the contribution of selection, according to our estimates of the annual mutation rate and the mean diversity found at each time point, the average time to fix a neutral mutation should be, on average, 156 days. This corresponds to an average length of about 5.2 months for a neutral mutation to appear and invade the population, a time frame fully compatible with our observations. Moreover, given the annual mutation rate estimated at 3.32 mutations per genome per year, we can gauge that 1.66 neutral mutations that are destined to reach fixation should have occurred in the first 6 months of the study. Once again, this is a number that is close to the observation of one mutation fixing in the second half of the study. Hence, these rough estimations suggest that the fixation of the *yidQ* promoter mutation may result from genetic drift rather than from natural selection.

Signs of convergence, a signature of natural selection (12), were nevertheless observed in *rpoS*. Two different mutations on two different samples at D0 (I198S in 1D0 and V102M in 7D0) were recovered. These mutations had, however, no long-term selective advantage as they were just isolated from individual isolates from the early time points.

DISCUSSION

In the present work, we followed retrospectively the evolutionary history of an *E. coli* natural isolate: *E. coli* strain ED1a in its natural environment, the gut of a healthy human. The clone we studied exhibited the O81 type and belonged to a human-specific avirulent B2 subgroup, named subgroup VIII (ST452). It was the only *E. coli* we detected at the three time

points we sampled, D0, D211, and D315, which implies full dominance of the *E. coli* niche over almost a full year of sampling.

The high dominance of the clone is consistent with several observations. First, B2 strains are recovered at high frequency among humans of industrialized countries (15, 36). Changes in hygiene and/or diet are suspected to contribute to that high prevalence. Over the last 4 decades, a little more than a 3-fold (36) increase in B2 frequency was observed in the gut of French individuals from whom our sampled individual was chosen. Second, when B2 strains are present within an individual, they are usually more dominant than other strains (37, 38). Third, B2 subgroup VIII clones, encompassing *E. coli* ED1a, were found to be dominant in 42% of hosts when found, which was higher than the value for other B2 group strains, which were found to be dominant in only 17% of hosts (21). This suggests an overall high adaptation of *E. coli* ED1a to the human gut.

Most changes observed during this year of evolution were SNPs (88%). For instance, we did not detect any phage integration or any gain or loss of plasmids. Based on SNPs, we could estimate an annual mutation rate, assuming a homogenous population evolving according to coalescent theory. The rate we estimated was 6.90×10^{7} mutations per base per year on the chromosome and 7.18×10^{7} mutations per base per year if the plasmid is taken into account. This is 3.2 times higher than the annual rate of 2.16×10^{7} mutations per base per year found *in vitro* in the long-term evolution experiment (LTEE) of Richard Lenski (28). An annual rate 1.7 times higher, 1.21×10^{6} mutations per base per year, was found in strains evolving in streptomycin-treated mice (14), and a rate five times lower, 1.4×10^{7} mutations per base per year, was found using strains evolving within a household (22) (Table 2). The value we observed is hence intermediate between the estimates observed in the LTEE though the per-base per-generation mutation rate and number of generations per day may vary across conditions.

Assuming a constant mutation rate per generation that we estimated using a range of published mutation rates, we can compute that *E. coli* is replicating between 6.06 (95% CI, 2.88 to 12.65) and 21.24 (95% CI, 9.66 to 43.28) times per day in the human gut (Table 2). The range of values found encompasses the 8 (39) and 18 (40) generations per day found previously in the gut of mice treated with streptomycin. The estimate of eight generations per day relied on the dilution of preinduced fluorescence and reflects an average over the population (39).

The estimate of 18 generations per day relied on the quantification of ribosome content, a proxy for growth rate, using a 23S rRNA fluorescent oligonucleotide probe (40). These ranges of the numbers of divisions per day are compatible with *E. coli* keeping up with the mucus turnover rate (41, 42). We computed a lower number of generations per day, ranging from 1.23 (95% CI, 0.26 to 2.43) to 4.3 (95% CI, 0.91 to 8.52) in the case of *E. coli* clone D evolving in a household over 3 years (Table 2). Interestingly, in contrast to our clone that was dominant over almost a year, clone D evolved in the household but was neither dominant nor consistently recovered in any host. Transitions from one host to another require transitions of unspecified duration in secondary environments poor in nutrients and therefore supporting limited growth (43).

Experimental evolution over a year coupled to genomics, *in vitro* (6, 12) or *in vivo* (14), indicates that selection shapes genome evolution. Footprints of selection, such as overrepresentation of nonsynonymous mutations and traces of convergence, are overwhelming. We wondered whether we could find similar signals over almost 1 year of evolution in the human gut. To do so, we first looked at the number of synonymous and nonsynonymous mutations. Using the whole data set, we could not find a clear difference between the two types of mutations, suggesting that the mutations we observed were accumulated as neutral mutations. We then looked at convergence, and the only signal that emerged was from a couple of mutations found in *rpoS* at D0 which were not recovered later. Unfortunately, *rpoS* mutations have to be studied with caution as they emerge rapidly under laboratory conditions (44). Though our samples have experienced a minimum number of steps in the laboratory, it is still possible that laboratory evolution before DNA isolation could be responsible for these mutations. We have therefore no proof that selection is at play in the pattern of the mutations we observed.

Evolution without selection can be easily framed with population genetics. For instance, based on the number of polymorphic sites and mutation rate estimates, we computed the effective population size, N_e, as ranging from 500 to 1,700, the number of mutations appearing within a year in the population that has invaded the population in the long term, n_{fix}, as 3.32 (Table 2), and the mean time it takes for these mutations to reach fixation, T_{fix}, as 5.2 months (results independent of the per-generation mutation rate estimate). The observation of two fixations over almost a year, one starting from 50% and another one from a lower frequency, is fully compatible with neutral evolution. The low effective population size appeared

nevertheless to be strikingly low, knowing that *E. coli* can be found in from 10^6 to 10^8 cells per gram of feces (45). It is, however, possible that a small fraction of the cells within the mucus may be founders of the rest of the population. Accordingly, previous work in streptomycin-treated mice has suggested that *E. coli* is not dividing in the lumen (46).

Several factors could explain the lack of a clear signal of selection over almost a year. First, the study of the LTEE has revealed that a high level of adaptation results in a lower fraction of beneficial mutations (6). However, even after 50,000 generations of evolution in a constant environment, more nonsynonymous than synonymous mutations were recovered. One alternative explanation is that the environment is not constant but fluctuating through time. As ED1a is well adapted, there may not be any clear way to improve over the average environment. Finally, the low effective population size may limit the opportunity for beneficial mutations to appear. Along these lines, it is worth noting that the two selective sweeps observed involved a single mutation and not a cohort of multiple mutations, as observed in experimental evolution (47). This confirms that if selection is at play, its strength, as measured by the product of population size and beneficial mutation rate, is very weak.

Our observations suggest that there was no clear sign of selection over almost a year of evolution in the human gut. This contrasts with the genomic evolution observed in experimental evolution. The data are indeed compatible with a small effective population size evolving neutrally even though the clone was dominant in its host over almost a year. Analyzing more longitudinal data of *E. coli* in the human gut will be required to test how general is this pattern of evolution.

MATERIALS AND METHODS

Sampling strategy. A single feces sample was collected from one healthy 44-year-old male living in Paris, France, over a period of almost 1 year, between October 2001 and September 2002, with three sampling points, at day 0 (D0), day 211 (D211) and day 315 (D315). The study was approved by the ethics evaluation committee of Institut National de la Santé et de la Recherche Médicale (INSERM). This person had no medical issues or diarrheic episodes and did not take any drug during the studied period and the year before. Moreover, this person has a Western lifestyle in terms of diet and hygiene. An aliquot of the fresh feces was spread on Drigalski agar plates that select for Gram-negative aerobic bacilli. The plates were incubated at 37°C over night (O/N), and the feces were discarded. Eight colonies of *E. coli* identified by a yellow appearance were isolated from each plate at each time point. These samples were then grown on liquid lysogeny broth (LB) O/N, stored at 80°C in glycerol, and thereafter called "isolate." Finally, the 24 isolates obtained were identified by the numbers 1 to 8 and a suffix for the three time points of sampling, D0, D211, or D315. To confirm that the isolates belonged to the *E. coli* ED1a clone, we performed a PCR screening for the B2 phylogroup (48), followed by O81 and B2 subgroup VIII typing (21).

High-throughput sequencing. The glycerol stock of each of the 24 isolates was then plated on LB agar, and one colony was chosen to extract the total DNA by using a Wizard genomic DNA purification kit (Promega). The 24 genomes were then sequenced using an Illumina platform HiSeq instrument with an average of almost 22 million paired-end reads of 51 bp per read.

Read mapping and mutation identification. We used as a reference the genome of ED1a that was sequenced using Sanger technology, fully assembled, and annotated in a previous work (23). It is composed of a circular chromosome of 5,209,548 bp and a large conjugative plasmid of 119,594 bp. This sequenced clone is one of the clones sampled at day 0. We resequenced this clone with the same technology as that used with the other clones to identify errors in this reference using breseq, version 0.27.2 (49).

breseq uses Bowtie2 to map reads to a reference genome. It then identifies mutation evidence that can take the form of either read alignment (RA) evidence, corresponding to single nucleotide polymorphisms and short indels and missing coverage, and new junction evidence, corresponding to reads mapping to one part of the reference on one of its sides and to another part on the other side, indicating a possible rearrangement. The program then uses this evidence to make mutation predictions. The RAs are then transformed into predictions as SNPs or short indels when they are supported by at least 85% of the read. breseq can also identify a large deletion and chromosomal rearrangement when it is able to correlate a missing coverage in a region with new junction evidence on both sides of this region.

We then filtered the mutations identified by breseq. When filtering the outputs of breseq, we first looked at the predicted SNPs and short indels. We removed mutations that seemed to appear in each sample as those likely came from a sequencing error in the reference genome and were not informative on the dynamics of diversification during this experiment. We also removed mutations when they were too close to one another and discarded variations that were less than 51 bp apart. These clustered mutations are indeed usually caused by erroneously mapped reads, and previous analyses showed that they are typically found in prophagic regions. These mobile regions are repeated in the genome but do not have 100% identity, generating trouble in the mapping process as they are still close enough to one another to be erroneously mapped. Indeed, the phage-mediated exchange of DNA sequences among bacteria occurs with high frequency (50), resulting in constant modifications of specific regions of the genome. We also removed all mutations for which the frequency of the mutated reads was less than 0.95.

A total of 1,029 mutations with a mutant frequency higher than 0.95 were detected by breseq among the 24 isolates. On the chromosome, these correspond to 329 specific mutations. Among these, 192 were recovered in all strains, corresponding to sequencing errors in the Sanger-sequenced reference genome used for mapping the reads. A total of 122 other mutations were clustered in three genes, *ECED1_1442*, *ECED1_1710*, and *ECED1_2512*. All three genes are in prophagic regions that have multiple copies in the genome with some level of divergence. The mutations detected in these regions result most likely from assembly problems and/or the existence of partial gene conversion between the different copies. Furthermore, these mutations, while found in some genomes above the threshold read frequency of 0.95, are found at lower read frequencies in the other strains, confirming the idea of assembly issues in these regions. Finally, only 15 mutations were retained for analysis on the chromosome. On the plasmid, 700 specific mutations were recovered, 456 of which were sequencing errors in the reference. Out of the 244 remaining mutations, 242 were clustered in two regions, *pECED1a_0013/ pECED1a_0014* and *insC/pECED1a_0084*/*pECED1a_0085*. Ultimately only two mutations were retained for analysis on the plasmid.

Prophage and plasmid detection in isolate sequences. Using the sequences obtained for the 24 isolates, we started by filtering them to eliminate all the reads matching the reference genome (5,209,548 bp) and plasmid (119,594 bp) (23). To this end, Bowtie2 was used with default parameters to align the reads to the reference and to keep only the unaligned reads (51). Among the 24 isolates, the number of unaligned reads and their proportion are, respectively, around 1.5×10^6 and 7%. These unaligned reads were then assembled using SPAdes (52). To detect prophage sequences from these assembled unaligned reads, we used blastx with the PHASTER database (updated 3 August 2017) (53). Then, we assembled the remaining unaligned reads with plasmidSPAdes in order to capture plasmid sequences with greater accuracy (54). Prokka was used to annotate the plasmid sequences (25); we then used discoSnp to detect mutations between these potential plasmids (26).

In order to know if this hypothetical plasmid was merely the result of contamination or a bona fide genetic element present in the strain but undetected to this day, we designed a simple PCR assay with primers specific to the hypothetical plasmid matching an ORF coding a protein belonging to the *mobA-mobL* superfamily involved in bacterial conjugation (forward, 5=-GGTTCTGCTCACCGCGTTCT-3=; reverse, 5=-GCATGATTGCCGATGTGGCG-3=).

Calculation of mutation rates and phylogenetic and selection analyses. In our study, we calculated the mutation rate per base per year of each of the genomic elements according to two models: the coalescent theory (by Bayesian inference) and the independence of the evolution of the lineages.

The unrooted phylogenetic tree was reconstructed by the maximum likelihood method with the packages ape (55) and phangorn (56) in R software (34). Regarding the selection indices, the calculation of diversity by the theta of Watterson ($_w$) and the statistical test Tajima's *D* were done using the pegas package (33) in the R software (34), and the calculation of K_a/K_s was generated taking into account the codon bias and the proportion of transitions/transversions. Calculations for diversity, estimation of the number of generations per day, mean time for fixation in generations, and mean time for fixation in days are based

on population genetics assuming Kimura's neutral model, as described in File S1 in the supplemental material (57).

Data availability. The fastq files corresponding to each isolate and the genome reference used to analyze mutations and produce Fig. 1 and Tables 1 and 2 are accessible on the Dryad website https://doi.org/10.5061/dryad.v5374.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/AEM .02377-17.

SUPPLEMENTAL FILE 1, PDF file, 0.3 MB.

ACKNOWLEDGMENTS

This work was supported in part by grants from La Fondation pour la Recherche Médicale to Mohamed Ghalayini (thèse médico-scientifique, grant number FDM20150633803) and Erick Denamur (équipe FRM 2016, grant number DEQ20161136698) and from the European Research Council to Olivier Tenaillon under the European Union's Seventh Framework Program (FP7/2007–2013)/ERC grant 310944. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We gratefully acknowledge Francois Cohen and Hervé Le Nagard for their valuable and indispensable help with the CATIBioMed cluster.

REFERENCES

1. Bradford PA. 2001. Extended-spectrum beta-lactamases in the 21st century: characterization, epidemiology, and detection of this important resistance threat. Clin Microbiol Rev 14:933–951. https://doi.org/10.1128/CMR.14.4.933 -951.2001.
2. Friedman ND, Temkin E, Carmeli Y. 2016. The negative impact of antibiotic resistance. Clin Microbiol Infect 22:416–422. https://doi.org/10 .1016/j.cmi.2015.12.002.
3. Ciorba V, Odone A, Veronesi L, Pasquarella C, Signorelli C. 2015. Antibioticresistance as a major public health concern: epidemiology and economic impact. Ann Ig 27:562–579.
4. van Buul LW, van der Steen JT, Veenhuizen RB, Achterberg WP,Schellevis FG, Essink RT, van Benthem BH, Natsch S, Hertogh CM. 2012. Antibiotic use and resistance in long term care facilities. J Am Med Dir Assoc 13:568.e1–e13. https://doi.org/10.1016/j.jamda.2012 .04.004.
5. Giraud A, Matic I, Tenaillon O, Clara A, Radman M, Fons M, Taddei F.2001. Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. Science 291:2606–2608. https://doi.org/10 .1126/science.1056421.
6. Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, DasguptaA, Wu GC, Wielgoss S, Cruveiller S, Médigue C, Schneider D, Lenski RE. 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature 536:165–170. https://doi.org/10.1038/nature18959.
7. Wiser MJ, Ribeck N, Lenski RE. 2013. Long-term dynamics of adaptationin asexual populations. Science 342:1364–1367. https://doi.org/10.1126/ science.1243357.
8. Maddamsetti R, Lenski RE, Barrick JE. 2015. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with Escherichia coli. Genetics 200:619–631. https://doi.org/ 10.1534/genetics.115.176677.
9. Giraud A, Arous S, De Paepe M, Gaboriau-Routhiau V, Bambou J-C,Rakotobe S, Lindner AB, Taddei F, Cerf-Bensussan N. 2008. Dissecting the genetic components of adaptation of Escherichia coli to the mouse gut. PLoS Genet 4:e2. https://doi.org/10.1371/journal.pgen.0040002.
10. Hindré T, Knibbe C, Beslon G, Schneider D. 2012. New insights into bacterialadaptation through in vivo and in silico experimental evolution. Nat Rev Microbiol 10:352–365. https://doi.org/10.1038/nrmicro2750.
11. Barroso-Batista J, Sousa A, Lourenço M, Bergman M-L, Sobral D, Demengeot J, Xavier KB, Gordo I. 2014. The first steps of adaptation of Escherichia coli to the gut are dominated by soft sweeps. PLoS Genet 10: e1004182. https://doi.org/10.1371/journal.pgen.1004182.
12. 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. https://doi.org/10.1126/science.1212986.
13. Welling GW, Groen G, Tuinte JH, Koopman JP, Kennis HM. 1980. Biochemical effects on germ-free mice of association with several strains of anaerobic bacteria. J Gen Microbiol 117:57–63.
14. Lescat M, Launay A, Ghalayini M, Magnan M, Glodt J, Pintard C, Dion S,Denamur E, Tenaillon O. 2016. Using long-term experimental evolution to uncover the patterns and determinants of molecular evolution of an Escherichia coli natural isolate in the streptomycin-treated mouse gut. Mol Ecol 26:1802–1817. https://doi.org/10.1111/mec.13851.
15. Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The populationgenetics of commensal Escherichia coli. Nat Rev Microbiol 8:207–217. https://doi.org/10.1038/nrmicro2298.
16. 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. https://doi.org/ 10.1016/S1286-4579(03)00049-2.
17. Jaureguy 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 9:560. https://doi.org/10.1186/ 1471-2164-9-560.
18. 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. https://doi.org/10.1086/507305.
19. Clermont O, Gordon D, Denamur E. 2015. Guide to the various phylogenetic classification schemes for Escherichia coli and the correspondence among schemes. Microbiology 161:980–988. https://doi.org/10 .1099/mic.0.000063.
20. 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. https://doi.org/10.1093/molbev/msm172.
21. Clermont O, Lescat M, O'Brien CL, Gordon DM, Tenaillon O, Denamur E.2008. Evidence for a human-specific Escherichia coli clone. Environ Microbiol 10:1000–1006. https://doi.org/10.1111/j.1462-2920.2007.01520.x.
22. Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, Clabots C, Lan R, JohnsonJR, Wang L. 2011. Rates of mutation and host transmission for an Escherichia coli clone over 3 years. PLoS One 6:e26907. https://doi.org/ 10.1371/journal.pone.0026907.
23. 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 ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EPC, Denamur E. 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 5:e1000344. https://doi.org/10 .1371/journal.pgen.1000344.
24. Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. 2015. Largescale contamination of microbial isolate genomes by Illumina PhiX control. Stand Genomic Sci 10:18. https://doi.org/10.1186/1944-3277-10-18.
25. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.
26. Uricaru R, Rizk G, Lacroix V, Quillery E, Plantard O, Chikhi R, Lemaitre C,Peterlongo P. 2015. Reference-free detection of isolated SNPs. Nucleic Acids Res 43:e11. https://doi.org/10.1093/nar/gku1187.
27. Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysisby sampling trees. BMC Evol Biol 7:214. https://doi.org/10.1186/1471 -2148-7-214.
28. Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S,Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D. 2013. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc Natl Acad Sci U S A 110:222–227. https:// doi.org/10.1073/pnas.1219574110.
29. Foster PL, Lee H, Popodi E, Townes JP, Tang H. 2015. Determinants ofspontaneous mutation in the bacterium Escherichia coli as revealed by whole-genome sequencing. Proc Natl Acad Sci U S A 112:E5990–E5999. https://doi.org/10.1073/pnas.1512136112.
30. Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrumof spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. Proc Natl Acad Sci U S A 109: E2774–E2783. https://doi.org/10.1073/pnas.1210309109.
31. Maharjan RP, Ferenci T. 2017. A shifting mutational landscape in 6 nutritional states: stress-induced mutagenesis as a series of distinct stress inputmutation output relationships. PLoS Biol 15:e2001477. https://doi.org/10 .1371/journal.pbio.2001477.

32. Watterson GA. 1975. On the number of segregating sites in geneticalmodels without recombination. Theor Popul Biol 7:256–276. https://doi.org/10.1016/0040-5809(75)90020-9.
33. Paradis E. 2010. pegas: an R package for population genetics with anintegrated-modular approach. Bioinformatics 26:419–420. https://doi.org/10.1093/bioinformatics/btp696.
34. R Development Core Team. 2016. R: a language and environment forstatistical computing. R Foundation for Statistical Computing, Vienna, Austria.
35. Tajima F. 1989. Statistical method for testing the neutral mutationhypothesis by DNA polymorphism. Genetics 123:585–595.
36. Massot M, Daubié AS, 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 162:642–650. https://doi.org/ 10.1099/mic.0.000242.
37. Blyton MDJ, Cornall SJ, Kennedy K, Colligon P, Gordon DM. 2014. Sexdependent competitive dominance of phylogenetic group B2 Escherichia coli strains within human hosts. Environ Microbiol Rep 6:605–610. https://doi.org/10.1111/1758-2229.12168.
38. Smati M, Clermont O, Le Gal F, Schichmanoff O, Jauréguy F, Eddi A,Denamur E, Picard B, Coliville Group. 2013. Real-time PCR for quantitative analysis of human commensal *Escherichia coli* populations reveals a high frequency of subdominant phylogroups. Appl Environ Microbiol 79:5005–5012. https://doi.org/10.1128/AEM.01423-13.
39. Myhrvold C, Kotula JW, Hicks WM, Conway NJ, Silver PA. 2015. A distributedcell division counter reveals growth dynamics in the gut microbiota. Nat Commun 6:10039. https://doi.org/10.1038/ncomms10039.
40. Rang CU, Licht TR, Midtvedt T, Conway PL, Chao L, Krogfelt KA, CohenPS, Molin S. 1999. Estimation of growth rates of *Escherichia coli* BJ4 in streptomycin-treated and previously germfree mice by in situ rRNA hybridization. Clin Diagn Lab Immunol 6:434–436.
41. Conway T, Cohen PS. 2015. Commensal and pathogenic *Escherichia coli* metabolism in the gut. Microbiol Spectr 3:3. https://doi.org/10.1128/microbiolspec.MBP-0006-2014.
42. Johansson MEV. 2012. Fast renewal of the distal colonic mucus layers bythe surface goblet cells as measured by in vivo labeling of mucin glycoproteins. PLoS One 7:e41009. https://doi.org/10.1371/journal.pone .0041009.
43. Ihssen J, Egli T. 2005. Global physiological analysis of carbon- and energylimited growing Escherichia coli confirms a high degree of catabolic flexibility and preparedness for mixed substrate utilization. Environ Microbiol 7:1568–1581. https://doi.org/10.1111/j.1462-2920.2005.00846.x.
44. Bleibtreu A, Clermont O, Darlu P, Glodt J, Branger C, Picard B,Denamur E. 2014. The rpoS gene is predominantly inactivated during laboratory storage and undergoes source-sink evolution in Escherichia coli species. J Bacteriol 196:4276–4284. https://doi.org/10 .1128/JB.01972-14.
45. Ervin JS, Russell TL, Layton BA, Yamahara KM, Wang D, Sassoubre LM,Cao Y, Kelty CA, Sivaganesan M, Boehm AB, Holden PA, Weisberg SB, Shanks OC. 2013. Characterization of fecal concentrations in human and other animal sources by physical, culture-based, and quantitative realtime PCR methods. Water Res 47:6873–6882. https://doi.org/10.1016/j .watres.2013.02.060.
46. Poulsen LK, Lan F, Kristensen CS, Hobolth P, Molin S, Krogfelt KA. 1994. Spatial distribution of Escherichia coli in the mouse large intestine inferred from rRNA in situ hybridization. Infect Immun 62:5191–5194.
47. Lang GI, Desai MM. 2014. The spectrum of adaptive mutations in experimental evolution. Genomics 104:412–416. https://doi.org/10.1016/j .ygeno.2014.09.011.
48. Clermont O, Christenson JK, Denamur E, Gordon DM. 2013. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ Microbiol Rep 5:58–65. https://doi.org/10.1111/1758-2229.12019.
49. 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. https://doi.org/10.1007/978 -1-4939-0554-6_12.
50. Kenzaka T, Tani K, Sakotani A, Yamaguchi N, Nasu M. 2007. Highfrequency phage-mediated gene transfer among *Escherichia coli* cells, determined at the single-cell level. Appl Environ Microbiol 73: 3291–3299. https://doi.org/10.1128/AEM.02890-06.
51. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
52. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, LesinVM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
53. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44:W16–W21. https://doi.org/10.1093/nar/gkw387.
54. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. 2016. plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics 32:3380–3387. https://doi.org/10.1093/bioinformatics/ btv688.
55. Popescu A-A, Huber KT, Paradis E. 2012. ape 3.0: New tools for distancebased phylogenetics and evolutionary analysis in R. Bioinformatics 28: 1536–1537. https://doi.org/10.1093/bioinformatics/bts184.
56. Schliep KP. 2011. phangorn: phylogenetic analysis in R. Bioinformatics27:592–593. https://doi.org/10.1093/bioinformatics/btq706.
57. Nielsen R, Slatkin M. 2013. An introduction to population genetics: theory and applications. Sinauer Associates, Sunderland, Mass.

CHAPITRE II: ETUDE DE L'ADAPTATION DE *E. COLI* 536 DANS LE TUBE DIGESTIF DE SOURIS DANS DES CONDITIONS NATURELLES ET SOUMIS A DEUX REGIMES ALIMENTAIRES

I /Introduction

Dans notre précédente étude (Article n°1), nous avons observé que *E. coli* ED1a, appartenant au groupe phylogénétique B2, évoluait de façon neutre dans le tube digestif d'un individu occidental vivant en région parisienne et sans problème de santé avec une faible taille de population efficace. Ceci signifie que la souche ED1a dès le début de l'étude est très bien adaptée à son hôte. Cependant, le suivi d'une seule lignée évoluant chez un individu ne nous a pas permis de répondre à certaines questions sur l'impact de l'alimentation. Bien que le recours à une expérimentation évolutive *in vivo* sur modèle murin contrôlé paraissait nécessaire, les modèles d'évolution dans le tube digestif de souris ayant été fait jusqu'à maintenant ne nous paraissaient pas satisfaisantes. En effet, l'utilisation de souris axéniques (De Paepe *et al.* 2011) ou traitées à la streptomycine (Barroso-Batista *et al.* 2014; Lescat *et al.* 2017) éloigne les conditions expérimentales des conditions naturelles en introduisant des biais (comme l'absence de microbiote pour les souris axéniques ou l'adaptation à la streptomycine). Nous avons donc décidé d'utiliser un modèle murin tentant de récréer un environnement le plus naturel possible en colonisant le tube digestif de souris de manière la plus naturelle possible en mimant la transmission du microbiote de la mère à l'enfant. En effet, nous avons colonisé des mères initialement traitées à la streptomycine par *E. coli* 536. Puis à l'accouchement, nous avons arrêté la streptomycine et permis la transmission du microbiote de la mère vers les progénitures dans lequel se trouvait la souche d'*E. coli*. Après une période d'allaitement de 3 semaines, les progénitures ont été distribuées dans des cages indépendantes puis soumises à des régimes alimentaires différents : un régime standard et un régime riche en gras et en sucre. La souche *E. coli* 536 a été ensuite extraite des fèces régulièrement, décomptée puis analysée par séquençage haut débit pendant plus d'un an.

II/Article n°2 (Publié dans Molecular Ecology en 2019)

Long-term evolution of the natural isolate of *Escherichia coli* 536 in the mouse gut colonized after maternal transmission reveals convergence in the constitutive expression of the lactose operon.

Running title: *E. coli lacI* instability

Mohamed Ghalayini [1,2,3#], Melanie Magnan [1,3], Sara Dion [1,3], Ouassila Zatout [1], Lucie Bourguignon [3,4], Olivier Tenaillon [1,3]* and Mathilde Lescat [1,3,5]*

[1]Université Paris 13, IAME, INSERM, F-93017 Bobigny, France

[2]AP - HP, Hôpital Avicenne, Service de Réanimation Médico-Chirurgicale, F-93000 Bobigny, France

[3]Université de Paris, IAME, INSERM, F-75018 Paris, France

[4]École de l'Inserm Liliane Bettencourt, Paris, France

[5]AP - HP, Hôpital Avicenne, Service de Microbiologie, F-93000 Bobigny, France

*: The authors participated equally to the work

Mohamed Ghalayini : mohamed.ghalayini@inserm.fr ; Mélanie Magnan : melanie.magnan@inserm.fr ; Sara Dion : sara.dion@inserm.fr ; Ouassila Zatout : wassi_zat@hotmail.fr ; Lucie Bourguignon : luciebourguignon90@gmail.com ; Olivier Tenaillon: olivier.tenaillon@inserm.fr ; Mathilde Lescat: mathilde.lescat@inserm.fr

#: Corresponding author: Mohamed Ghalayini, IAME, INSERM UMR 1137, Faculté de Medicine Paris Diderot, Site Xavier Bichat, 16 rue Henri Huchard, 75018 Paris, France. email: mohamed.ghalayini@inserm.fr

ABSTRACT

In vitro experimental evolution has taught us many lessons on the molecular bases of adaptation. To move towards more natural settings, evolution in the mice gut has been successfully performed. Yet, these experiments suffered from the use of laboratory strains as well as the use of axenic or streptomycin treated mice to maintain the inoculated strains. To circumvent these limitations, we conducted a one-year experimental evolution *in vivo* using a natural isolate of *E. coli*, strain 536, in conditions mimicking as much as possible natural environment with mother to offspring microbiota transmission. Mice were then distributed in 24 independent cages and separated in two different diets: a regular one (Chow diet, CD) and high-fat high-sugar one (Western diet, WD). Genome sequences revealed an early and rapid selection during the breast-feeding period that selected the constitutive expression of the well-characterized lactose operon. *E. coli* was lost significantly more in CD than WD, however, we could not detect any genomic signature of selection, nor any diet specificities during the later part of the experiments. The apparently neutral evolution presumably due to low population size maintained nevertheless at high frequency the early selected mutations affecting lactose regulation. The rapid loss of lactose operon regulation challenges the idea that plastic gene expression is both optimal and stable in the wild.

KEYWORDS

Escherichia coli, lactose operon, *lacI*, adaptation, *in vivo* long-term evolution

INTRODUCTION

Genomes of bacteria have been shaped by billions of years of evolution to allow the colonization of all available ecological niches. In the last years, an explosive emergence of resistance of bacteria have been observed due to the massive utilization of antibiotics in human and veterinary medicine and brings us new evidence of the wonderful ability of bacteria to adapt to a challenging environment (Bradford 2001; Ciorba *et al.* 2015; Friedman *et al.* 2016; van Buul *et al.* 2012). Much less is known about the pace or determinants of microbial evolution in their natural environment even for a well-known bacterial species as *Escherichia coli*. While *E. coli* is a model for experimental evolution in the laboratory (Tenaillon *et al.* 2016), its evolution in its natural environment, the gut of vertebrate, is much less documented. Understanding this evolution is nevertheless important in particular at the medical level as beyond its life style as a commensal of the gut (Tenaillon *et al.* 2010), *E. coli* is a formidable pathogen, which is responsible of about a million deaths per year due to intestinal or extra-intestinal pathologies (Tenaillon *et al.* 2010).

To identify the molecular bases involved in *E. coli* adaptation, experimental evolutions in controlled conditions coupled with whole genome sequencing has been very successful (Barroso-Batista *et al.* 2014; Ghalayini *et al.* 2018; Lescat *et al.* 2017; Tenaillon *et al.* 2016). Over the last years, numerous *in vitro* experiments allowed us to decipher the main rules of bacterial adaptation. Among these, the repeated early-recruitment of mutations affecting global regulators suggests that the first steps of adaptation may involve global modifications rather than several specific modification of some pathways (Maddamsetti *et al.* 2015; Tenaillon *et al.* 2016; Tenaillon *et al.* 2012). However, the outcome of these experiments performed in environments very different from those in nature.

Several experimental studies, using a model of mice gut colonization have therefore been performed. Interestingly, genomic data revealed convergences in the inactivation (De

Paepe *et al.* 2011) or constitutive activation (Barroso-Batista *et al.* 2014; Lescat *et al.* 2017) of operons involved in metabolism of sugars. The first *in vivo* experiments were realized using gut colonization of germ-free mice. Among them, De Paepe *et al.* observed mutations of the *malT* gene encoding a transcriptional activator of genes responsible for the metabolism of maltose (De Paepe *et al.* 2011). Later, Barroso-Batista *et al.* using colonization of streptomycin treated mice by *E. coli* K-12 MG1655, observed mutations in *srlR* gene encoding the repressor of the *slr* operon and several genes in the *gat* operon (Barroso-Batista *et al.* 2014), whose operons are involved respectively in the metabolism of sorbitol and galactitol. Likewise, Lescat *et al.* performed the colonization of streptomycin treated mice using a natural isolate of *E. coli* and observed convergence of mutation in the *dgoRKAD* operon responsible for the metabolism of galactonate (Lescat *et al.* 2017). These nearly constant observations reveal that (i) metabolism adjustments are highly selected in the gut and that consequently diet should have an impact, (ii) these adjustments occur largely by change in the regulatory network, (iii) many of these changes involve gene inactivation that alter the regulatory structure of sugar metabolism.

The repeated modification of these regulatory structures, some of which being conserved, at the species level is however quite surprising. Several specificities of the experimental design could explain this apparently atypical evolution. First, these studies use streptomycin or germ-free mice to allow colonization and maintenance of *E. coli* in gut. These unnatural conditions could drive the pattern of molecular evolution. For instance, Lescat *et al.* observed repeated mutations in the *rluD* and the *gidB* genes involved in ribosome maturation and associated with adaptation to streptomycin (Lescat *et al.* 2017). Second the strain used in most evolutionary experiments are the emblematic *E. coli* laboratory strain K12 MG1655 whose initial maladaptation forces a strong selection in the gut (Couce & Tenaillon 2015). For instance, Barroso-Batista *et al.* followed a derived *E. coli* K-12 strain in streptomycin treated mice gut. The strain used could also contribute to the observed pattern of evolution. They

observed a recurrent inactivation of *gat* operon. To guarantee the constitutive expression of fluorescent proteins (used to follow the dynamic of adaptation through change in the frequency of these "neutral" markers) the lac promoter was used and a deletion of the *lac* operon was necessary (Barroso-Batista *et al.* 2014). The absence of a functional *lac* operon would not allow *E. coli* to degrade lactose to glucose and galactose and thus its reduced form, galactitol. Thus, the constitutive inactivation of *gat* operon could eventually be linked to the absence of its substrate, galactitol. Third, some of the adaptations could be specific of the mice diet used. The gut is an open system that is affected by the diet or the drugs consumed. These could impact *E. coli* direct environment through the nutrients availability, the composition of the microbiota and the local immune system (Giraud *et al.* 2001; Giraud *et al.* 2008). The impact of antibiotics as a drastic selective pressure has already been highlighted (Lescat *et al.* 2017; Wong *et al.* 2012). However, to our knowledge, no study has experienced diet as a determinant of molecular evolution in the gut.

To move forward in the understanding of the molecular determinants of adaptation in the mouse gut, and to test if changes in metabolism regulation occur broadly, we followed the long-term evolution of an unmodified natural strain of *E. coli* 536 in the mice gut under conditions mimicking natural transmission of microbiota between mother and their offspring (Payros *et al.* 2014) To study the impact of the diet, a high fat and high sugar diet (Western Diet, WD) and a regular diet (Chow Diet, CD) were compared.

MATERIALS AND METHODS

Bacterial strain

We used the *E. coli* 536 (phylogenetic group B2, serotype O6:K5:H31) reference strain. The 536 strain has been isolated from a patient suffering from urinary tract infection in the early 1980s (Berger *et al.* 1982) and subsequently shown to be a good gut colonizer in streptomycin-

treated mouse (Diard *et al.* 2010). This strain was acquired by our laboratory in 1998 from Jorg Hacker who isolated it and then stored at -80 °C with glycerol.

Mouse intestinal colonization assay and sampling strategy

All animal experiments were conducted using a protocol of colonization (N° APAFIS#4948-20160212162515484), approved by the French Ministry of Research and by the ethical committee for animal experiment. We have adopted a protocol of colonization of the mice gut allowing the natural transmission of the microbiota from mother to offspring (Payros *et al.* 2014). Fifteen 7-week-old primiparous pregnant mice were used for this study (Janvier® laboratory Swiss CD-1). All mice were on their 13^{th} day of gestation at the animal facility reception. Each pregnant mouse was placed in a single cage under pathogen-free conditions, with an access to standard breeding food and water supplemented with streptomycin (5g/l) *ad libitum*. On their 17^{th} day of gestation, the pregnant mice were inoculated by intra-gastric gavage with 100 µl of 10^7 Unit Forming Colony (UFC)/ml of *E. coli* 536 solution. Before inoculation, the feces of pregnant mice were collected and spread over the Drigalski agar media to ensure the absence of coliform flora in gut. Two pregnant mice with positive fecal cultures in Drigalski media were excluded from the experiment. All remaining mother were named Mo1 to Mo13. At the time of delivery, streptomycin was stopped and replaced by antibiotic-free water. The colonization of *E. coli* 536 from mother to offspring guts occurred naturally after birth (Payros *et al.* 2014) (Figure 1).

After a three-week breast-feeding period (day 24 after inoculation / 22-23 after offspring birth), 72 female offspring were housed in 24 independent pathogen-free cages (three mice per cage) during over one year. The three mice in the same cage came from different mothers to limit maternal effects. The mice were separated into two groups of 36 mice, within 12 cages, with two different diets: a high fat and sugar diet (Western Diet, WD, purified diet 260 HF,

U8978 version 19, Safe Diet®) and a regular diet (Chow Diet, CD, breeding diet, U8213G10R, Safe Diet®). In each cage, we considered that an independent lineage of *E. coli* 536 evolved.

During the experiment, the intestinal population of *E. coli* was estimated by counting colonies after plating dilutions of weighted fresh feces for each offspring (days after inoculation 28, 36, 43, 50, 57, 65, 78, 91, 106, 120, 151, 168, 182, 196, 210, 245, 273, 302, 322, 357, 385 and final) on Drigalski media. Feces were harvested for each mouse by placing the mouse in a sterile plastic cage until it defecated. The feces and those collected from mothers' at day 1 and 16 after inoculation were stored at -80°C with glycerol for further analysis.

High throughput sequencing

For sequencing, the frozen feces were plated on Drigalski media. An aliquot of the frozen feces was spread on Drigalski agar plates that select for Gram-negative aerobic bacilli. The plates were incubated at 37°C O/N (overnight). One colony of *E. coli* was isolated from each plate at three time points per lineage. At each time point, for each cage, we sampled a colony isolated from the feces of the mice with the highest *E. coli* density. Due to coprophagy, sampling one clone of one randomly selected mouse per cage is representative of the population of *E. coli* colonizing the three mice within a single cage (Lescat *et al.* 2017) (Figure 1). These samples were then grown on liquid lysogeny broth (LB) O/N and thereafter called "isolate". Finally, the 71 isolates obtained were labelled by the lineage (L) numbers from 1 to 24 and a suffix for the sampling day after inoculation (D). Total DNA was extracted using QIAamp® DNA Mini Kit (Qiagen). The 71 genomes of these isolates were then sequenced using the Illumina platform HiSeq® or MiSeq®. To confirm that the isolates evolved from the *E. coli* 536 isolate, we performed a PCR screening for B2 phylogroup (Clermont *et al.* 2013), followed by O6 and B2 subgroup VIII typing (Clermont *et al.* 2014). For the polymorphism study, we

picked a pool of about 100 colonies from feces spread on Drigalski plate for each of the 13 mothers at day 16 and for 4 mothers (Mo4, Mo5, Mo11, and Mo12) at day 1 after inoculation.

Read mapping and mutations identification

We used as a reference the genome of *E. coli* 536 that has been sequenced using Sanger technology, fully assembled and annotated in a previous work (Brzuszkiewicz *et al.* 2006) and then confirmed in a previous work (Lescat *et al.* 2017). It is composed of a circular chromosome of 4,938,920 bp without any plasmid.

We used breseq 0.23 (Deatherage & Barrick 2014) to identify differences between evolved genome sequenced and a reference genome. The pipeline breseq 0.23 uses bowtie2 (Langmead & Salzberg 2012) to map reads to a reference genome. It then identifies mutation evidence that can take the form of either Read Alignment (RA) evidence, corresponding to single nucleotide polymorphisms and short indels, missing coverage, and new junction evidence, corresponding to reads mapping to one part of the reference on one of its side and to another part on the other side, indicating a possible re-arrangement. The program then uses those evidences to make mutation predictions. The RA are then transformed into prediction as single-nucleotide polymorphisms (SNPs) or short indels when they are supported by at least 85% of the read. The pipeline breseq 0.23 can also identify a large deletion and chromosomal re-arrangement when it is able to correlate a missing coverage in a region with a new junction evidence on both sides of this region.

We filtered the mutations identified by breseq 0.23. When filtering the outputs of breseq 0.23, we first looked at the predicted SNPs and short indels. We removed mutations that seemed to appear in each sample as those likely came from a sequencing error in the reference genome and were not informative on the dynamics of the diversification during this experiment. We also removed mutations when they were too close from one another, by discarding variations

that were less than 51 bp apart. These clustered mutations are indeed usually caused by reads erroneously mapped, and previous analyses showed that they are typically found in proghagic regions (Ghalayini *et al.* 2018). Those mobile regions are repeated in the genome but do not have 100% identity generating trouble in the mapping process, as they are still close enough from one another to be erroneously mapped. Indeed, the phage-mediated exchange of DNA sequences among bacteria occurs with high frequency, resulting in constant modifications of specific regions of the genome. We also removed all mutations for which the frequency of the mutated reads was less than 0.95.

Mutation rates analysis

Mutation rates and non synonymous to synonymous selection bias ω was estimated through a maximum likelihood approach taking into account mutation biases. The probability to observe a strain with genotype {s1, s2, s3, s4, s5, s6, n1, n2, n3, n4, n5, n6} was sampled at day t after inoculation was derived as

$$P(s_1, s_2, s_3, s_4, s_5, s_6, n_1, n_2, n_3, n_4, n_5, n_6, t) = \prod_i Poisson(s_i, m\varepsilon_i S_i t) Poisson(n_i, \omega m \varepsilon_i N_i t)$$

with S_i (N_i) the number of sites were a mutation of type i was (non)synonymous and s_i (n_i) the number of observed (non)synonymous of type i, m the mutation rate of the most abundant transition GC to AT (i=5), and ε_i the observed bias in mutation rate of type i relative to the transition GC to AT ($\varepsilon_i = \frac{\sum_{all\ genotypes}(n_i+s_i)/(N_i+S_i)}{\sum_{all\ genotypes}(n_5+\varepsilon_5)/(N_5+S_5)}$). The annual mutation rate per base μ was then computed as $\mu = 365.25 \frac{\sum_i m\varepsilon_i(N_i+S_i)}{\sum_i(N_i+S_i)/3}$. The optim function of R ("R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL Https://Www.R-Project.Org/.," n.d.) was used to optimize the log likelihood, fitting for ε_i, m and ω. As optimizing for ε_i, did not improve the likelihood, we just report results in which ε_i are inferred from the data as presented ahead. The likelihood

of a model in which μ and ω are estimated was compared to a simpler model in which ω was set to one.

Calculation of K_a/K_s was generated taking into account the mutational biases. Basically, the expected of number of non synonymous of type i was computed as $n_{iexpected}=N_iS_i/S_i$. K_a/K_s was then defined as the ratio of observed to expected mutations:

$$\frac{Ka}{Ks} = \sum_i n_i \ / \sum_i n_{iexpected}$$ (Ghalayini *et al.* 2018; Lescat *et al.* 2017)

Growth curves

14 evolved lineages (L1D106 and L1D182 bringing a SNP mutation G→T in position 446,204 in *lacI* gene; isolates L3D106, L15D106 and L19D106 bringing a SNP mutation G→A in position 445,449 in intergenic region *lacZ/lacI*; isolates L12D106, L16D106 and L22D106 bringing a duplication $(CCTGCCAG)_{1\rightarrow2}$ in position 446,034 in *lacI* gene; isolates L14D106 and L14D182 bringing a one base deletion in position 446,035 in *lacI* gene; isolate L12D182 bringing a deletion $(AGACGC)_{2\rightarrow1}$ in position 446,039 in gene *lacI*; and isolates L5D106, L9D106 and L13D106 bringing a one base deletion in position 446,474 in gene *lacI*), a lactose minus control and the ancestor genotypes were distributed in two 96 well plates and stored in glycerol 40% at minus 20 degrees. Each of the lineages was present in 6 to 7 replicates in the two plates, while the ancestor was present in 92 replicates. An O/N culture was made from the 96 well storage plate and diluted about hundred fold using a pin replicator in 96 well plates containing M9 minimal media, with 2 mM $MgSO_4$, 0.1 mM $CaCl_2$ and either 0.4% P-Gal, 0.4% lactose, 0.1% lactose or 0.04% lactose plus 0.04% glucose. Each well contained 120 µl of media and were supplemented with 30 µl of mineral oil to limit evaporation. The growth curves were performed at 37 °C in a Tecan Infinite M200 Plate Reader® imbedded in a Tecan Evolution Robotic Platform. Growth curves were analyzed with R. For lag analysis, the minimal optical density (OD) of each culture was subtracted to all OD values. The median OD of the

whole plate after 10 hours of growth was measured, and the time to reach 25%, 50% or 75% of that value was measured using linear extrapolation between the two closest measures. Growth on P-Gal was evaluated as the OD at 25 hours. P-values were computed using a linear model (lm in R) linking the time to reach a given OD to the genotype of the lineage.

RESULTS

E. coli density evolution

To study *E. coli* adaptation to the mice gut, we inoculated 13 streptomycin-treated pregnant mice (one mouse per cage) with *E. coli* 536. Two mice gave birth at 18th gestation day and 11 mice gave birth at 19th gestation day. 160 offspring were born among the 13 pregnant mice with an average litter of 13.30+/-2.13 offspring, including 5.92+/-2.59 females per pregnant mouse. After the period of breast-feeding, during which the colonization of offspring's gut by *E. coli* 536 occurred, thirty six 3-week-old female mice were distributed in 12 independent cages (3 mice per cage) per diet (seventy-two mice in 24 independent cage in total) and followed during more than one year (figure 1). We considered that a lineage of *E. coli* 536 evolved in the three mice gut from the same cage as previously tested by Lescat *et al* (Lescat *et al.* 2017).

We observed an iterative loss and recolonization of *E. coli* 536 of the three mice guts from the same cage over the duration of the experiment. The presence of *E. coli* 536 was observed until the end-point for 15 lineages out of 24. In the other cages, loss of *E. coli* was observed in all mice (Figure 1). The full loss of *E. coli* was particularly relevant in mice fed with CD (8 lineages out of 12) compared with mice fed with WD (1 lineage out of 12) (Fisher Exact test, $p < 0.01$) (Figure 1).

The average *E. coli* 536 population density, when present, was $1.71x10^6$ (95% CI: $1.23x10^6 – 2.38x10^6$) UFC per gram (g) of feces in WD group and an average of $1.24x10^6$ (95%

CI: $7.96x10^5$ – $1.92x10^6$) UFC/g of feces in CD group. More precisely, density first decreased from $1.74x10^8$ UFC/g of feces to $1.41x10^6$ UFC/g of feces, in the two diets, until day 91 after inoculation and then reached a relative stability until the end of experiment (Figure 2A). There was no significant difference between the two diets relative to *E. coli* density when present. (Comparison of area under curve of density of *E. coli* 536 integrated to time between WD and CD, wilcoxson test, p = 0.37). However, if we include mice that have lost *E. coli* and count their density as 0, we then found a significant difference between the two diets (Comparison of area under curve of density of *E. coli* 536 integrated to time between WD and CD, wilcoxson test, $p < 10^{-4}$) with $9.64x10^4$ (95% CI: $5.72x10^4$ – $1.62x10^5$) UFC/g of feces for WD and $2.41x10^3$ (95% CI: $1.35x10^3$ – $4.31x10^3$) UFC/g of feces for CD (Figure 2B).

Mutations found in E. coli 536 evolved genomes

To study the molecular determinants of adaptation, one clone per cage and time point was sequenced. Using breseq 0.23 pipeline (Deatherage and Barrick 2014), we identified 137 different mutations that occurred during the experiment in the 71 genomes (Table S1).

The great majority of these (108/137) mutations were point mutations. Among these, 73 were non synonymous including 3 stop mutations, 26 were synonymous and 9 were intergenic. We observed an imbalance in favor of transitions for synonymous point mutations (19/26) as well as for non synonymous mutations (55/70) as previously reported (Lescat *et al.* 2017; Wielgoss *et al.* 2013) (Table S2). Two large deletions (>30bp) were observed as well as 18 small indels among which 10 lead to frameshifts, 3 were in frame and 5 were intergenic. In *E. coli* strains evolving in the mice gut subjected to CD we observed a total of 8 transposon *de novo* insertions, mostly from the IS 100 / IS21 family (Table S4). We have summarized the significant mutations at several occurrences in Table 1.

Convergence in the evolved genomes: inactivation of lac operon

The strongest signal of convergence was found in the *lac* operon (Table 1). The *lac* operon, the most characterized operon, is necessary for the transport and metabolism of lactose in *E. coli* through the expression of three genes: *lacZ*, *lacY* and *lacA*. The operon is regulated by *lacI* encoding a repressor protein which, in the absence of lactose or large amounts of glucose, binds to the *lacO* operator sites in the *lacI/lacZ* intergenic region, thereby creating a loop blocking the operon promoter (Krämer, Amouyal, Nordheim, & Müller-Hill, 1988; Krämer *et al.* 1987). Out of the 71 isolates analyzed, 44 carried a mutation in the lactose operon. These mutations were distributed in 19 lineages including 15 lineages where it was found at the final time point. Seven different mutations affected the operon, five in the *lacI* gene, one in the *lacO1* operator and one in *lacY*. The mutations found in *lacI/lacO1* resulted presumably in changes in the regulation of the operon. First four out of five of the *lacI* mutations were frameshift that inactivated *lacI* and therefore lead to constitutive expression of the operon. Second, the mutation found in *lacO1* was found at a site affecting the operon (Maity *et al.* 2012) regulation and in between two mutations described previously as leading to a very low threshold of induction of the operon (Oehler *et al.* 1990).

To test for a phenotypic effect of these mutations, we studied the growth of strains carrying some of these mutations in the presence of P-Gal. P-Gal is a substrate that can be degraded by *lacZ* but is unable to bind to *lacI* and therefore to induce the operon expression. Consequently, only isolates with constitutive expression of the operon can grow on P-Gal. While no growth was observed for the ancestral strain, all evolved genotypes showed a rapid growth on P-Gal (mean OD after 25 h of growth: 0.005 for the ancestor, 0.194 for *lacI* lineages and 0.198 for *lacO1* lineages, $p < 2x10^{-16}$ at the gene level and also for each individual mutation of *lacI*). This experiment confirms that all observed mutations lead to a constitutive expression of the operon confirming that convergence occurred at the phenotypic level.

We did not observe a difference in the distribution of this phenotypic change in LacI expression between the two diets (24 isolates for CD versus 20 isolates for WD, chi2 test, p-value = 0.25) (Figure 3). Moreover, the mutations being recovered in different cages multiple times, it is likely that rather than multiple occurrences of similar mutations, these were selected for before the assignment of the young mice to the different cages: as mice born from different mother were split in the different cages, so did the selected mutations. We therefore hypothesize that the emergence of these mutations that are recovered in different lineages predate the exposition of the young mice to the two diets.

To test for this hypothesis, we both tracked the origin of the mutations in different litters before the separation of the offspring in the 24 cages and also studied the diversity present in the different mother's gut before this separation. For the later part, we sequenced a pool of about 100 colonies from all mothers at day 16 after inoculation. We found that mutations in the *lacI* gene and *lacO1* operator were very frequent in the mothers and were distributed among 9 or 3 different alleles respectively (Table 2 and figure 3). The *lacI* mutations were either frameshifts, non-sense mutation or non synonymous among which some were found previsouly as inactivating *lacI* (Sousa *et al.* 2016). Similarly, one of the mutations found in *lacO1* was previously described as leading to a lower threshold of activation of the operon (Quan *et al.* 2012). While the *E. coli* population of 6 mothers carried a single polymorphism in one or the other gene, 5 others had two to five different polymorphisms at these loci. This reveals that several alleles compete with one another in these populations, a phenomenon of clonal interference (Gerrish & Lenski 1998).

The fraction of the population with a modified expression of the *lac* operon was in most cases dominant revealing an important selection. The overall frequency of the different alleles affecting the *lac* operon expression, was greater than 90% in 7 mothers (Mo4, Mo7, Mo9, Mo10, Mo11, Mo12 and Mo13) and between 40 and 70% for four (Mo1, Mo2, Mo5 and Mo8).

Only Mo6 carried no variant in the *lacI* gene. Thus, there is a real convergence in the modulation of the lactose operon regulation by mutations on *lacI* and *lacO1* in all but one of the mother mice analyzed (Table 2). The analysis of *E. coli* 536 populations extracted from the feces of 4 mothers on day 1 after inoculation does not show any mutations on *lacI* or *lacO1*. This confirms our hypothesis of an early selection during the breast-feeding period that is independent of the diet used later on. Alternatively, lac mutations may have surfaced within a period of few days of streptomycin treatment inside the mothers' guts but it seems unlikely given past studies (Barroso-Batista *et al.* 2014; Lescat *et al.* 2017).

Beyond the selection at the phenotypic level, we found some consistency in the analysis of the genotype of the strains isolated in the evolved mice and the polymorphism found in the mothers. Most mutations could be traced back to a polymorphism found in a mother with two exceptions. The first one was a point mutation found in cage 1. However, one mice of that cage came from a mother that we could not test for polymorphism as we could not retrieve *E. coli* from the frozen feces. More importantly one mutation in *lacO1* was found in many evolved strains but was not detected in any of the mother.

Finally, an insertion of 4 bp (CGAG) in *lacY* was found in 3 lineages (L6, L16 and L22) without affiliation with a common mother. This mutation appears early in the different lineages but was not fixed. Contrary to *lacI* and *lacO1* mutations that were never found simultaneously in the same genome, the *lacY* mutation was always associated to a *lacI* inactivation.

To test for potential benefit of constitutive expression of *lac* operon, we performed some growth curves of several evolved lineages in the presence of two concentrations of lactose, 0.4% and 0.1%. Using the ancestor (92 replicates) and 14 different evolved lineages, with 6 to 7 replicates for each, covering the 6 different mutations affecting the regulation of *lac* operon, we could show that the lag was shorter for *lacI/O1* mutants. Excluding two lineages that showed very poor growth on lactose, the *lacI* lineages reached half max OD (see methods) 40 minutes

before the ancestor ($p < 2x10^{-13}$) and *lacO1* lineages 24 minutes ($p = 7x10^{-5}$) for a lactose concentration of 0.4%. The gain was even bigger for a concentration of 0.1% lactose: 58 minutes ($p < 2x10^{-16}$) gain for *lacI* and 36 minutes for *lacO1* lineages ($p < 5x10^{-6}$). To prove that the effect was due to a faster transition in the usage of lactose, we performed growth curves in a media containing glucose 0.04% and lactose 0.04%. At 25% of the final OD, when glucose is used, no difference was found ($p = 0.867$ for *lacI* and $p = 0.57$ for *lacO1)* while a significant difference was already found at that OD in the pure lactose media (at least $p < 1x10^{-5}$). Nevertheless, 75% of the final OD was reached 34 and 40 minutes before ($p < 5x10^{-6}$ and $p < 2.1x10^{-4}$) for *lacI* and *lacO1* lineages respectively.

Other signature of convergence

We also found a signature of convergence in the regulation of maltose and to a lesser extent in methionine transporter gene *metN* and a regulator for the genes encoding proteins produced in response to oxidative stress *oxyR.* For these three genes a single mutation was found in several cages which shared the same mother. These observations are compatible with an early occurrence of the mutation. However only in the case of *malT* could we found a signature of selection among the mothers. Similarly to the *lacI* most frequent mutation, the frameshift found in *malT*, was selected between day 1 and day 16 post-inoculation during the breast-feeding period in 7 mothers with frequencies ranging from 7% to 83% (Table 2). In addition, other segregating polymorphisms were found in other genes of the operon.

Genomics rates of mutation and signature of selection

Beyond genetic convergence, we investigated the signature of selection through the analysis of the rates of synonymous and non synonymous mutations. We first focused only on synonymous mutations that are mostly neutral and consequently supposed to accumulate in a

clock like fashion according to mutation rate (Figure 4). With a regression approach we found a mutation rate of $1.28x10^{-7}$ (95% CI: $1.1x10^{-9}$ - $2.55x10^{-7}$) per base per year in CD and of $1.39x10^{-7}$ (95% CI: $1.07x10^{-8}$ - $2.67x10^{-7}$) in WD. As the numbers were not different (ANOVA test between the diets, $p = 0.32$), we could combine them and found a rate of $1.27x10^{-7}$ (95% CI: $2.13x10^{-8}$ - $2.34x10^{-7}$). We then focused on non synonymous mutations that are more likely to be affected by selection. We computed a rate of $2.34x10^{-7}$ (95% CI: $6.36x10^{-8}$ - $4.05x10^{-7}$) mutations per base per year for CD and $3.51x10^{-7}$ (95% CI: $1.55x10^{-7}$ - $5.47x10^{-7}$) for WD. Similarly, the rate did not differ between treatments (ANOVA test between the diets, $p = 0.16$), and was computed as $3.15x10^{-7}$ (95% CI: $2x10^{-7}$ - $4.31x10^{-7}$).

We then combined the mutations observed in the two different diets to estimate if non synonymous mutations accumulate at rate equivalent to the one of synonymous. We took into account the observed mutation bias and used a maximum likelihood approach to estimate the genomic mutation rate and the bias ω between non synonymous and synonymous mutations accumulations. ω of one reveals that non synonymous mutations accumulated randomly through genetic drift, ω greater than one suggests a contribution of natural selection in the accumulation of non synonymous mutations, and ω lesser than one a purifying selection filtering out non synonymous mutations. With that approach, we found a mutation rate, μ, of $4.9x10^{-7}$ (95% CI: $3.3x10^{-7}$ - $6.8x10^{-7}$) and $\omega = 1.40$ (95% CI: 0.93 - 2.13). ω is therefore marginally different from 1. Accordingly, model comparison was not conclusive with Akaike Information Criterion (AIC) test favoring slightly a model with $\omega > 1$ ($AIC_{\mu} - AIC_{\mu,\omega} = 0.48$) but Bayesian Information Criterion (BIC) favoring a model with just ω fixed to 1 ($BIC_{\mu} - BIC_{\mu,\omega} = -1.78$). When ω was fixed to one, we found $\mu = 6.2x10^{-7}$ (95% CI: $5.1x10^{-7}$ - $7.4x10^{-7}$). Alternatively, we computed a ratio of the rates of non synonymous mutations (Ghalayini *et al.* 2018), using the set of all mutations that were found only once in the data set. Taking into

account synonymous mutation biases, we found Ka/Ks= 1.31 (95% CI estimated by bootstrap: 0.80 - 1.84).

Because we know mutations in *lacI* were selected for, we recomputed the numbers without the only SNP observed in that gene, other mutations were indels or intergenic not taken into account in the synonymous to non synonymous ratios. We found numbers very similar to the ones mentioned previously with $\mu = 4.9 \times 10^{-7}$ (95% CI: $3.3 \times 10^{-7} – 6.8 \times 10^{-7}$) and $\omega = 1.38$ (95% CI: 0.91 - 2.11), ω is fixed to one $\mu = 6.1 \times 10^{-7}$ (95% CI: 5.0×10^{-7} - 7.3×10^{-7}) and the comparison of the models $AIC_{\mu} - AIC_{\mu,\omega} = 0.23$ and $BIC_{\mu} - BIC_{\mu,\omega} = -2.03$. Finally we found in that case Ka/Ks= 1.15 (95% CI estimated by bootstrap: 0.78 - 1.83). Hence, no convincing sign of positive selection for non synonymous mutation was found with the synonymous to non synonymous ratios.

DISCUSSION

We prospectively conducted a long-term evolution experiment of a natural isolate of *E. coli* in dominant situation in conditions that we tried to make as close as possible to natural conditions.

Characterization of the colonization model

Previous experiments performed in the mice gut suffered from several limitations such as (i) the use of laboratory isolates such as *E. coli* K12 MG1655 (Lawrence & Ochman 1998; Welch *et al.* 2002), that may undergo a more rapid adaptation because of its initial maladaptation (Couce & Tenaillon 2015) and/or (ii) the use of streptomycin (Barroso-Batista *et al.* 2014; Lescat *et al.* 2017) or germ-free mice (Giraud *et al.* 2001; Giraud *et al.* 2008) to facilitate *E. coli* colonization. In the present study we used a natural isolate of *E. coli* and to be closer to natural conditions, streptomycin was only used to inoculate the strain in pregnant mice.

They subsequently naturally transmitted *E. coli* along with the rest of their microbiota to their offspring (Payros *et al.* 2014). The combination of the massive initial intake of *E. coli* during inoculation and the early streptomycin driven dysbiosis (Garner *et al.* 2009) perturbed the mice gut environment making it less "natural" than hoped. However, the rapid initial decrease of *E. coli* density previously observed (Poulsen *et al.* 1994) and the rapid microbiota resilience after the end of streptomycin administration (Stecher *et al.* 2007) suggest that these initial perturbations are short lived.

In that setting, *E. coli* 536 could be recovered after more than one year of evolution in the majority of the lineages, especially in high fat high sugar diet group. Mice could lose *E. coli* 536 definitely or recover it by eating the feces of mice still colonized in the same cage. Mice are known to be coprophages. The resistance to gut colonization of *E. coli* (van der Waaij *et al.* 1971) was probably secondary to competition with the rest of the microbiota (Kamada *et al.* 2013). In previous studies, colonization resistance was diverted by the absence of microbiota, using axenic mice (De Paepe *et al.* 2011; Giraud *et al.* 2008) or its modification, using streptomycin-treated mice (Barroso-Batista *et al.* 2014; Lescat *et al.* 2017). The loss of *E. coli* seems also to be associated with diet. The analysis of the digestive microbiota of children in Burkina-Faso whose diet is high in fiber and children in Europe whose diet is rich in fat and sugar have reported an under-representation of *E. coli* in the microbiota of Burkinabe children (De Filippo *et al.* 2010). Furthermore, the more consistant maintenance of *E. coli* 536 in the WD diet is probably due to its genetic background rather than to *de novo* adaptation as we failed to see a clear signature of adaptation. The maintenance of a B2 background with WD echoes the spread of the phylogenetic group B2 (to which *E. coli* 536 belongs), in a commensal situation, in the digestive tract of individuals living in industrialized countries (Tenaillon *et al.* 2010) for the past 30 years (Massot *et al.* 2016). This model seems therefore consistent with some epidemiological patterns of commensal *E. coli* in the human gut, suggesting its relevance.

Estimation of mutation rates and biases in the mice gut

The sequencing of the evolved genomes could be used to estimate the mutation rate based on the SNPs that represented 78.2% of mutations. The neutral mutation rate we estimated was about $5x10^{-7}$ mutations per base per year with no statistically significant difference between the two applied diets. This rate is in the ranged of the ones estimated previously (i) in Richard Lenski's long-term evolution experiment (~$2x10^{-7}$) (Wielgoss *et al.* 2013), (ii) the mutation rate of a strain evolving in a household (Reeves *et al.* 2011) or the mutation rate of a dominant strain evolving in human gut (6.9×10^{-7}) (Ghalayini *et al.* 2018) or the synonymous mutation rate of the same strain evolving in the streptomycin-treated mice gut ($6.4 x10^{-7}$, 95% CI: $5.0x10^{-7}$ - $7.3x10^{-7}$ estimated assuming the same mutational biases as the one observed here) (Lescat *et al.* 2017).

The mutational bias was largely in favor of the transition GC to AT which represented 69% (73%) of synonymous (all) mutations while the other transition AT to GC represented only 3% (5%) of synonymous (all). This bias was consistent with the bias observed for the laboratory strain REL606 evolved *in vitro* (Wielgoss *et al.* 2013) (Chi^2 test using synonymous mutations $p = 0.14$) comforting the idea of a strong mutational pressure towards AT rich sequences (Rocha and Feil 2010b) with 87% of all mutations going towards AT. Yet the bias was different from the one observed in a mutation accumulation experiment (MAE) using another laboratory strain (Chi^2 test using all mutations from the MAE and synonymous from our experiment $p = 0.0025$), opening the possibility of environment by strains specificities for the mutational biases.

Detecting selection with the synonymous to non synonymous ratios

Experimental evolution over a year coupled to genomics, *in vitro* (Tenaillon *et al.* 2016; Tenaillon *et al.* 2012) or *in vivo* (Lescat *et al.* 2017), indicates that selection shapes genome

evolution. Footprints of selection such as overrepresentation of non synonymous mutations and traces of convergence are overwhelming. We wondered whether we could find similar signal over almost one year of evolution in the mice gut. To do so, we looked at the number of synonymous and non synonymous mutations. Using the whole dataset, we could not retrieve a clear difference between the two types of mutations. Moreover, we did not observe any difference between the number of synonymous and non synonymous mutations depending on the diet or the sampling period. All these observations suggest that the mutations accumulated mostly neutrally once the diets were imposed. Two factors may explain the lack of selection in that experiment. First, *E. coli* 536 is a commensal gut bacteria already adapted to its environment. Second, the population size decreased to low density in the course of the experiment (Figure 2). The resulting low diversity will limit the efficiency of natural selection, a situation observed in human gut (Ghalayini *et al.* 2018).

Detecting selection with convergence

Despite the lack of a consistent signal of selection using synonymous and non synonymous, a clear pattern of selection was found using genetic convergences. The difference between the two approaches is largely due to the fact that the observed convergence involved mostly indels and intergenic mutations that are not part of synonymous to non synonymous approaches. The convergences involved changes in the regulation of metabolic operons. The metabolism of sugar is a crucial issue for *E. coli* in the gut. Indeed it has been shown with mutant analysis that only mutations in sugar pathways affected colonization. Mutations in phospholipid and amino acid catabolism, gluconeogenesis, the tricarboxylic acid cycle, and the pentose phosphate pathway had not such an effect (Chang *et al.* 2004). Moreover, selection experiment in the mice gut recruited mainly mutations affecting sugar pathways (Barroso-Batista *et al.* 2014; De Paepe *et al.* 2011; Giraud *et al.* 2008; Lescat *et al.* 2017). Among these

sugars, galactose, a degradation product of lactose, seems to occupy a major place by being one of the sugars consumed in priority by *E. coli in vitro* (Fabich *et al.* 2008). Mutation of the pathway for galactose catabolism caused a colonization defect in *E. coli* EDL 93 in mice gut fed with regular diet (Fabich *et al.* 2008).

We observed a strong convergence targeting the regulation of the lactose operon that degrades lactose into one molecule of glucose and galactose, two main energy substrates of *E. coli* (Fabich *et al.* 2008). Five different mutations on *lacI*, and a mutation on the *lacO1* operator were recovered in 18 lineages out of 24. All sequenced genotypes had only one of these mutations suggesting that they have a similar effect. Indeed, we tested phenotypically these mutations and showed that they all resulted in a constitutive expression of the *lac* operon. The pattern of convergence observed involved the occurrence of the same point mutation in multiple cages, a quite rare phenomenon that suggested a very early selection of the mutations that were then transferred to the different cages with the young mice. We confirmed this hypothesis by looking at the polymorphism present in the mothers before mice were spread in the different cages. In a few weeks, in all but one mice mother inactivation of the *lac* operon regulation reached 40 to 100% frequencies. Yet, at the end of the experiments, not all lineages had fixed this constitutive expression.

A strong polymorphism of the *lacI* gene has been demonstrated in spontaneous or mutagen-associated mutations carried out on *E. coli* K-12. The main mutation found in the lacI- strains isolated on P-gal plates, present in about 67% to 82% of the mutants, was an indel (duplication or deletion) of a nucleotide segment TGGC that is repeated three times between position 623 and 624 of *lacI* (Murata-Kamiya *et al.* 1997; Schaaper & Dunn 1991; Sedwick *et al.* 1986; Swerdlow & Schaaper 2014). This repeat sequence is absent in *E. coli* 536, which, as we observed in our study, but it has another hotspot at position 592 due to duplication of a CCTGCCAG nucleotide segment. Whereas the indels at a specific hotspot constitutes the

overwhelming majority of the mutations leading to a constitutive activation of the lactose operon, the SNPs represent around 10% of the mutations on the *lacI* gene (Murata-Kamiya *et al.* 1997; Schaaper & Dunn 1991). Mutations, exclusively SNPs, were also found to affect the *lacO1* site, under the same conditions, with no mutation found on the *lacO3* or *lacO2* operators, nor any other site in the *lacI/lacZ* intergenic region (Quan *et al.* 2012; Swerdlow & Schaaper 2014). This suggests a second mutational hotspot and/or a narrower spectrum of mutational targets that would lead to a lacI- phenotype in that region.

The early strong selection for this constitutive expression of the lactose operon can only be justified by the presence of lactose in the environment (Andrews & Hegeman 1976; Koch 1983; Novick & Weiner 1957). Lactose intake seems obvious at the time of breast-feeding, but less clear afterwards. Nevertheless the constitutive expression of the operon could was still present after a year. A possible explanation for this maintenance is genetic drift. Indeed, mutations leading to a constitutive expression of the operon are highly present at the end of the breast-feeding period. If they are no more selected for, their high frequency assure their maintenance or fixation in the population with high probability (Kimura & Ohta 1978). As besides convergence in *lacI* and *malT*, we did not detect a global signal of selection, is it possible that after the breast-feeding period selection for these mutations stopped and their frequencies fluctuated later on according to genetic drift. This scenario is consistent with the mutations in *lacI*, as the most frequent mutation found in the mothers is also the most frequently found in the evolved lineages. However, in the *lacO1* operator, the mutation recovered largely at the end of the evolution position was not detected in the mothers, suggesting some selection after the breast-feeding period.

After the breast-feeding period, there is much less lactose in the diet. However, the profound decrease of the mice lactase expression (an enterocyte brush border membrane beta-glycosidase that splits lactose) (Sebastio *et al.* 1989; Smith & James 1987) make this fraction

(up to 0.26% in WD) accessible to the microbiota. Lactose resource is therefore redistributed from the mouse to the microbiota (Szilagyi 2015). As a result, the constitutive activation of lactose activity could still constitute a selective advantage and could explain the maintenance of the *lacI/O1* alleles and the selection of some specific alleles post breast-feeding period.

In our experiment, for both *malT* and *lacI/O1* mutations, selection favored a change in regulation by either inactivating the regulator or modifying its binding site. Similarly, in a previous experiment, we found that the regulator of the galactonate operon, *dgoR*, was the main target of adaptation (Lescat *et al.* 2017). Functional regulation of these operons is supposed to give the most cost efficient pattern of expression: inactivated when the sugar is absent and activated otherwise. It seems therefore surprising that a constitutive expression can be selected for. Yet, as observed in galactonate, we confirmed that constitutive expression could be beneficial in the presence of lactose: the growh curves showed that the time to reach the mid OD was shorter for the *lacI/O1* lineages and a faster transition from glucose to lactose usage.

The operon mostly affected in our experiment, the lactose operon, is the most characterized operon in the history of genetics and was the very one used to define the concept of regulation upon demand. Our results, associated with observations of convergence in maltose operon regulator *malT* and those reported in the literature, suggest that the genetic regulation of operon may be less stable than previously thought and may depend on a complex balance between the presence and absence of substrate.

ACKNOWLEDGMENTS

This work was supported in part by grants from La Fondation pour la Recherche Médicale to Mohamed Ghalayini (thèse médico-scientifique, grant number FDM20150633803) and from European Research Council to Olivier Tenaillon under the European Union's Seventh Framework Program (FP7/2007–2013)/ERC Grant 310944. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We

gratefully acknowledge Hervé Le Nagard for his valuable and indispensable help with the CATIBioMed cluster.

REFERENCES

Andrews, K. J., & Hegeman, G. D. (1976). Selective disadvantage of non-functional protein synthesis in Escherichia coli. *Journal of Molecular Evolution*, *8*(4), 317–328.

Barroso-Batista, J., Sousa, A., Lourenço, M., Bergman, M.-L., Sobral, D., Demengeot, J., … Gordo, I. (2014). The first steps of adaptation of Escherichia coli to the gut are dominated by soft sweeps. *PLoS Genetics*, *10*(3), e1004182. https://doi.org/10.1371/journal.pgen.1004182

Berger, H., Hacker, J., Juarez, A., Hughes, C., & Goebel, W. (1982). Cloning of the chromosomal determinants encoding hemolysin production and mannose-resistant hemagglutination in Escherichia coli. *Journal of Bacteriology*, *152*(3), 1241–1247.

Bradford, P. A. (2001). Extended-spectrum beta-lactamases in the 21st century: characterization, epidemiology, and detection of this important resistance threat. *Clinical Microbiology Reviews*, *14*(4), 933–951, table of contents. https://doi.org/10.1128/CMR.14.4.933-951.2001

Brzuszkiewicz, E., Brüggemann, H., Liesegang, H., Emmerth, M., Olschläger, T., Nagy, G., … Dobrindt, U. (2006). How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic Escherichia coli strains. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(34), 12879–12884. https://doi.org/10.1073/pnas.0603038103

Chang, D.-E., Smalley, D. J., Tucker, D. L., Leatham, M. P., Norris, W. E., Stevenson, S. J., … Conway, T. (2004). Carbon nutrition of Escherichia coli in the mouse intestine. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(19), 7427–7432. https://doi.org/10.1073/pnas.0307888101

Ciorba, V., Odone, A., Veronesi, L., Pasquarella, C., & Signorelli, C. (2015). Antibiotic resistance as a major public health concern: epidemiology and economic impact. *Annali Di Igiene: Medicina Preventiva E Di Comunita*, *27*(3), 562–579.

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. *Environmental Microbiology Reports*, *5*(1), 58–65. https://doi.org/10.1111/1758-2229.12019

Couce, A., & Tenaillon, O. A. (2015). The rule of declining adaptability in microbial evolution experiments. *Frontiers in Genetics*, *6*, 99. https://doi.org/10.3389/fgene.2015.00099

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., … Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(33), 14691–14696. https://doi.org/10.1073/pnas.1005963107

De Paepe, M., Gaboriau-Routhiau, V., Rainteau, D., Rakotobe, S., Taddei, F., & Cerf-Bensussan, N. (2011). Trade-off between bile resistance and nutritional competence drives Escherichia coli diversification in the mouse gut. *PLoS Genetics*, *7*(6), e1002107. https://doi.org/10.1371/journal.pgen.1002107

Deatherage, D. E., & Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods in Molecular Biology (Clifton, N.J.)*, *1151*, 165–188. https://doi.org/10.1007/978-1-4939-0554-6_12

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. *Journal of Bacteriology*, *192*(19), 4885–4893. https://doi.org/10.1128/JB.00804-10

Fabich, A. J., Jones, S. A., Chowdhury, F. Z., Cernosek, A., Anderson, A., Smalley, D., … Conway, T. (2008). Comparison of carbon nutrition for pathogenic and commensal Escherichia coli strains in the mouse intestine. *Infection and Immunity*, *76*(3), 1143–1152. https://doi.org/10.1128/IAI.01386-07

Friedman, N. D., Temkin, E., & Carmeli, Y. (2016). The negative impact of antibiotic resistance. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *22*(5), 416–422. https://doi.org/10.1016/j.cmi.2015.12.002

Garner, C. D., Antonopoulos, D. A., Wagner, B., Duhamel, G. E., Keresztes, I., Ross, D. A., … Altier, C. (2009). Perturbation of the small intestine microbial ecology by streptomycin alters pathology in a Salmonella enterica serovar typhimurium murine model of infection. *Infection and Immunity*, *77*(7), 2691–2702. https://doi.org/10.1128/IAI.01570-08

Gerrish, P. J., & Lenski, R. E. (1998). The fate of competing beneficial mutations in an asexual population. *Genetica*, *102–103*(1–6), 127–144.

Ghalayini, M., Launay, A., Bridier-Nahmias, A., Clermont, O., Denamur, E., Lescat, M., & Tenaillon, O. (2018). Evolution of a dominant natural isolate of Escherichia coli in the human gut over a year suggests a neutral evolution with reduced effective population size. *Applied and Environmental Microbiology*. https://doi.org/10.1128/AEM.02377-17

Giraud, A., Arous, S., De Paepe, M., Gaboriau-Routhiau, V., Bambou, J.-C., Rakotobe, S., … Cerf-Bensussan, N. (2008). Dissecting the genetic components of adaptation of Escherichia coli to the mouse gut. *PLoS Genetics*, *4*(1), e2. https://doi.org/10.1371/journal.pgen.0040002

Giraud, A., Matic, I., Tenaillon, O., Clara, A., Radman, M., Fons, M., & Taddei, F. (2001). Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science (New York, N.Y.)*, *291*(5513), 2606–2608. https://doi.org/10.1126/science.1056421

Kamada, N., Chen, G. Y., Inohara, N., & Núñez, G. (2013). Control of pathogens and pathobionts by the gut microbiota. *Nature Immunology*, *14*(7), 685–690. https://doi.org/10.1038/ni.2608

Kimura, M., & Ohta, T. (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences of the United States of America*, *75*(6), 2868–2872.

Koch, A. L. (1983). The protein burden of lac operon products. *Journal of Molecular Evolution*, *19*(6), 455–462.

Krämer, H., Amouyal, M., Nordheim, A., & Müller-Hill, B. (1988). DNA supercoiling changes the spacing requirement of two lac operators for DNA loop formation with lac repressor. *The EMBO Journal*, *7*(2), 547–556.

Krämer, H., Niemöller, M., Amouyal, M., Revet, B., von Wilcken-Bergmann, B., & Müller-Hill, B. (1987). lac repressor forms loops with linear DNA carrying two suitably spaced lac operators. *The EMBO Journal*, *6*(5), 1481–1491.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lawrence, J. G., & Ochman, H. (1998). Molecular archaeology of the Escherichia coli genome. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(16), 9413–9417.

Lescat, M., Launay, A., Ghalayini, M., Magnan, M., Glodt, J., Pintard, C., … Tenaillon, O. (2017). Using long-term experimental evolution to uncover the patterns and determinants of molecular evolution of an Escherichia coli natural isolate in the streptomycin-treated mouse gut. *Molecular Ecology*, *26*(7), 1802–1817. https://doi.org/10.1111/mec.13851

Maddamsetti, R., Lenski, R. E., & Barrick, J. E. (2015). Adaptation, Clonal Interference, and Frequency-Dependent Interactions in a Long-Term Evolution Experiment with Escherichia coli. *Genetics*, *200*(2), 619–631. https://doi.org/10.1534/genetics.115.176677

Maity, T. S., Jha, R. K., Strauss, C. E. M., & Dunbar, J. (2012). Exploring the sequence-function relationship in transcriptional regulation by the lac O1 operator. *The FEBS Journal*, *279*(14), 2534–2543. https://doi.org/10.1111/j.1742-4658.2012.08635.x

Massot, M., Daubié, A.-S., Clermont, O., Jauréguy, F., Couffignal, C., Dahbi, G., … The Coliville Group, null. (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, England)*, *162*(4), 642–650. https://doi.org/10.1099/mic.0.000242

Murata-Kamiya, N., Kamiya, H., Kaji, H., & Kasai, H. (1997). Mutational specificity of glyoxal, a product of DNA oxidation, in the lacI gene of wild-type Escherichia coli W3110. *Mutation Research*, *377*(2), 255–262.

Novick, A., & Weiner, M. (1957). ENZYME INDUCTION AS AN ALL-OR-NONE PHENOMENON. *Proceedings of the National Academy of Sciences of the United States of America*, *43*(7), 553–566.

Oehler, S., Eismann, E. R., Krämer, H., & Müller-Hill, B. (1990). The three operators of the lac operon cooperate in repression. *The EMBO Journal*, *9*(4), 973–979.

Payros, D., Secher, T., Boury, M., Brehin, C., Ménard, S., Salvador-Cartier, C., … Oswald, E. (2014). Maternally acquired genotoxic Escherichia coli alters offspring's intestinal homeostasis. *Gut Microbes*, *5*(3), 313–325. https://doi.org/10.4161/gmic.28932

Poulsen, L. K., Lan, F., Kristensen, C. S., Hobolth, P., Molin, S., & Krogfelt, K. A. (1994). Spatial distribution of Escherichia coli in the mouse large intestine inferred from rRNA in situ hybridization. *Infection and Immunity*, *62*(11), 5191–5194.

Quan, S., Ray, J. C. J., Kwota, Z., Duong, T., Balázsi, G., Cooper, T. F., & Monds, R. D. (2012). Adaptive evolution of the lactose utilization network in experimentally evolved populations of Escherichia coli. *PLoS Genetics*, *8*(1), e1002444. https://doi.org/10.1371/journal.pgen.1002444

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. (n.d.).

Reeves, P. R., Liu, B., Zhou, Z., Li, D., Guo, D., Ren, Y., … Wang, L. (2011). Rates of mutation and host transmission for an Escherichia coli clone over 3 years. *PloS One*, *6*(10), e26907. https://doi.org/10.1371/journal.pone.0026907

Rocha, E. P. C., & Feil, E. J. (2010). Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? *PLoS Genetics*, *6*(9), e1001104. https://doi.org/10.1371/journal.pgen.1001104

Schaaper, R. M., & Dunn, R. L. (1991). Spontaneous mutation in the Escherichia coli lacI gene. *Genetics*, *129*(2), 317–326.

Sebastio, G., Villa, M., Sartorio, R., Guzzetta, V., Poggi, V., Auricchio, S., … Semenza, G. (1989). Control of lactase in human adult-type hypolactasia and in weaning rabbits and rats. *American Journal of Human Genetics*, *45*(4), 489–497.

Sedwick, W. D., Brown, O. E., & Glickman, B. W. (1986). Deoxyuridine misincorporation causes site-specific mutational lesions in the lacI gene of Escherichia coli. *Mutation Research*, *162*(1), 7–20.

Smith, M. W., & James, P. S. (1987). Cellular origin of lactase decline in postweaned rats. *Biochimica Et Biophysica Acta*, *905*(2), 503–506.

Sousa, F. L., Parente, D. J., Shis, D. L., Hessman, J. A., Chazelle, A., Bennett, M. R., … Swint-Kruse, L. (2016). AlloRep: A Repository of Sequence, Structural and Mutagenesis Data for the LacI/GalR Transcription Regulators. *Journal of Molecular Biology*, *428*(4), 671–678. https://doi.org/10.1016/j.jmb.2015.09.015

Stecher, B., Robbiani, R., Walker, A. W., Westendorf, A. M., Barthel, M., Kremer, M., … Hardt, W.-D. (2007). Salmonella enterica serovar typhimurium exploits inflammation to compete with the intestinal microbiota. *PLoS Biology*, *5*(10), 2177–2189. https://doi.org/10.1371/journal.pbio.0050244

Swerdlow, S. J., & Schaaper, R. M. (2014). Mutagenesis in the lacI gene target of *E. coli*: improved analysis for lacI(d) and lacO mutants. *Mutation Research*, *770*, 79–84. https://doi.org/10.1016/j.mrfmmm.2014.09.004

Szilagyi, A. (2015). Adaptation to Lactose in Lactase Non Persistent People: Effects on Intolerance and the Relationship between Dairy Food Consumption and Evalution of Diseases. *Nutrients*, *7*(8), 6751–6779. https://doi.org/10.3390/nu7085309

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., … Lenski, R. E. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, *536*(7615), 165–170. https://doi.org/10.1038/nature18959

Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D., & Gaut, B. S. (2012). The molecular diversity of adaptive convergence. *Science (New York, N.Y.)*, *335*(6067), 457–461. https://doi.org/10.1126/science.1212986

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

van Buul, L. W., van der Steen, J. T., Veenhuizen, R. B., Achterberg, W. P., Schellevis, F. G., Essink, R. T. G. M., … Hertogh, C. M. P. M. (2012). Antibiotic use and resistance in long term care facilities. *Journal of the American Medical Directors Association*, *13*(6), 568.e1-13. https://doi.org/10.1016/j.jamda.2012.04.004

van der Waaij, D., Berghuis-de Vries, J. M., & Lekkerkerk Lekkerkerk-v, null. (1971). Colonization resistance of the digestive tract in conventional and antibiotic-treated mice. *The Journal of Hygiene*, *69*(3), 405–411.

Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., … Blattner, F. R. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(26), 17020–17024. https://doi.org/10.1073/pnas.252529799

Wielgoss, S., Barrick, J. E., Tenaillon, O., Wiser, M. J., Dittmar, W. J., Cruveiller, S., … Schneider, D. (2013). Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 222–227. https://doi.org/10.1073/pnas.1219574110

Wong, A., Rodrigue, N., & Kassen, R. (2012). Genomics of adaptation during experimental evolution of the opportunistic pathogen Pseudomonas aeruginosa. *PLoS Genetics*, *8*(9), e1002928. https://doi.org/10.1371/journal.pgen.1002928

DATA ACCESSIBILITY

The fastq files corresponding to each isolate used to analyze mutations and produce Table 1, 2, S2, S3 and S4 and Fig 1, 3 et 4 are accessible on the Dryad website: (http://datadryad.org) with the number DOI: https://doi.org/10.5061/dryad.8t15p7t.

AUTHORS CONTRIBUTIONS

ML and OT designed research; MG, OT, MM, SD, OZ and LB performed research; MG and OT analyzed data; MG, ML and OT wrote the paper.

TABLES AND FIGURES

<u>Figure 1.</u> Experimental scheme, summary of the distribution and appearance of mutations over time

We represented in parallel the experimental scheme and the distribution of the mutations observed on strains extracted in the feces of mothers and offspring at different time points. *E. coli* 536 was inoculated in streptomycin-treated pregnant mice. Streptomycin was only used to release a colonization nest for *E. coli* in the gut of pregnant mice and was stopped at delivery to avoid exposing the offspring. Then, the mothers naturally transmitted *E. coli* along with the rest of their microbiota to their offspring. After a 22 to 23 days breast-feeding period, the offspring were distributed in 24 independent cages, with 3 mice per cage from different mothers. The contribution of mothers to a lineage is represented by grey box in the left table. Two diets were used: a high fat high sugar diet (Western diet, WD) for lineages 1 to 12 and regular diet (Chow diet, CD) for lineages 13 to 24. We represented the mutations found in *E. coli* 536 lineages during the 22-23 days breast-feeding period and the rest of experiment. The length of each arrow is proportional to the duration of the follow-up, Cages which lost their lineage before the final time point were marked with a plain black cross followed by the last day when *E. coli* was found in, at least, one mouse in the cage. Each lineage was sampled at three times points (except for the lineage 18 because of its quick disappearance) represented by the tip of arrow and two vertical lines across it. The observed mutations in the sequenced genomes are represented by arrows or triangles. The nonsynonymous SNPs are represented by blue arrows, synonymous SNPs by red arrows, intergenic SNPs by black arrows and SNPs resulting in a stop codon by a crossed black arrow. The deletion mutations are represented by an orange triangle, the mutations by insertions are represented by a purple triangle, the mutations by duplication are represented by a green triangle.

Figure 2. Density of *E. coli* in mouse feces as a function of time

A) The mean log10 density for WD (blue) or CD (red) group is presented for the independent cages in which *E. coli* 536 was detected. B) The mean log10 density is presented for the independent cages according to diet taking account the loss of *E. coli*. Error bars represent standard deviation.

Figure 3. Time distribution of mutations in the *lac* operon and their characteristics

The reference sequence of the *lac* operon and mutations positions is indicated at the top of the figure. SNP mutations are represented by a nucleotide change, insertions by a (+) sign and deletions by a (-) sign followed by the nucleotide introduced or lost. The contribution of mothers to a lineage is represented by grey box in a table placed at the intersection of mother column and lineage row. The red boxes represents how the intergenic SNP in *lacO1* from Mo9 was propagated to the different cages, the blue boxes the *lacI* one base deletion from Mo13.

Figure 4. Accumulation of SNP mutations per genome per day after inoculation

A) Accumulation of synonymous (red) and non synonymous (blue) mutations according to diet. CD is in plain, WD in dashed. An ANOVA test did not reveal any significant difference between the two diets for the accumulation of synonymous ($p = 0.58$) or non synonymous ($p = 0.25$) mutations. B) Per site accumulation of synonymous (red) and non synonymous (blue) mutations. When normalizing mutation accumulation by the number of available sites of each type and the observed overall mutation bias, mutation per genome accumulated similarly for synonymous and non synonymous mutations. 95 confidence intervals of the regression slope are presented in grey.

Table 1. Summary of several mutations that were found in the evolved lineages

The name of the genes as well as the orientation (defined by the arrow), the position along the genome, and the flanking genes in case of an intergenic mutation are displayed, with the mutations, their type, the number of occurrences, and a description of the gene products. Stop codons are indicated with the symbol *.

Table 2. Polymorphism of *E. coli 536* isolates extracted from the feces of mother mice at day 16 after inoculation

The proportion of the different alleles found in the mother's feces at day 16 after inoculation for *lac* and *mal* operons. Arrows indicate the orientations of the genes; a slash indicates the intergenic region. The asterisk (*) indicates a stop codon.

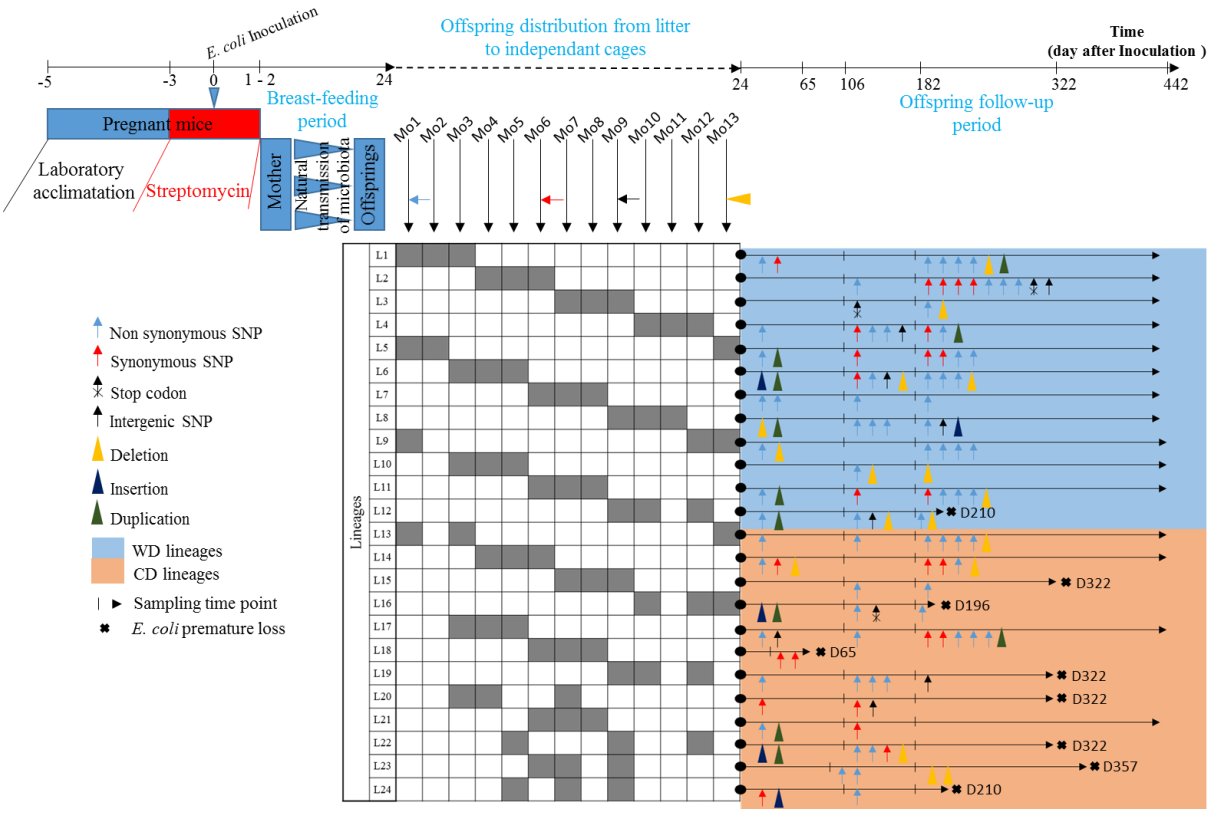
Table S1. Mice characteristics of the 71 sequenced strains

For each of the 71 isolates sequenced, we reported the sampling time, the sampled mice ID (lineage and mouse number), the diet, the time of death (spontaneous or sacrifice), the *E. coli* density per gram of feces and the number of mutations found in the genome.

Table S2. List of the SNPs that were found in the 71 sequenced isolates of *E. coli* 536

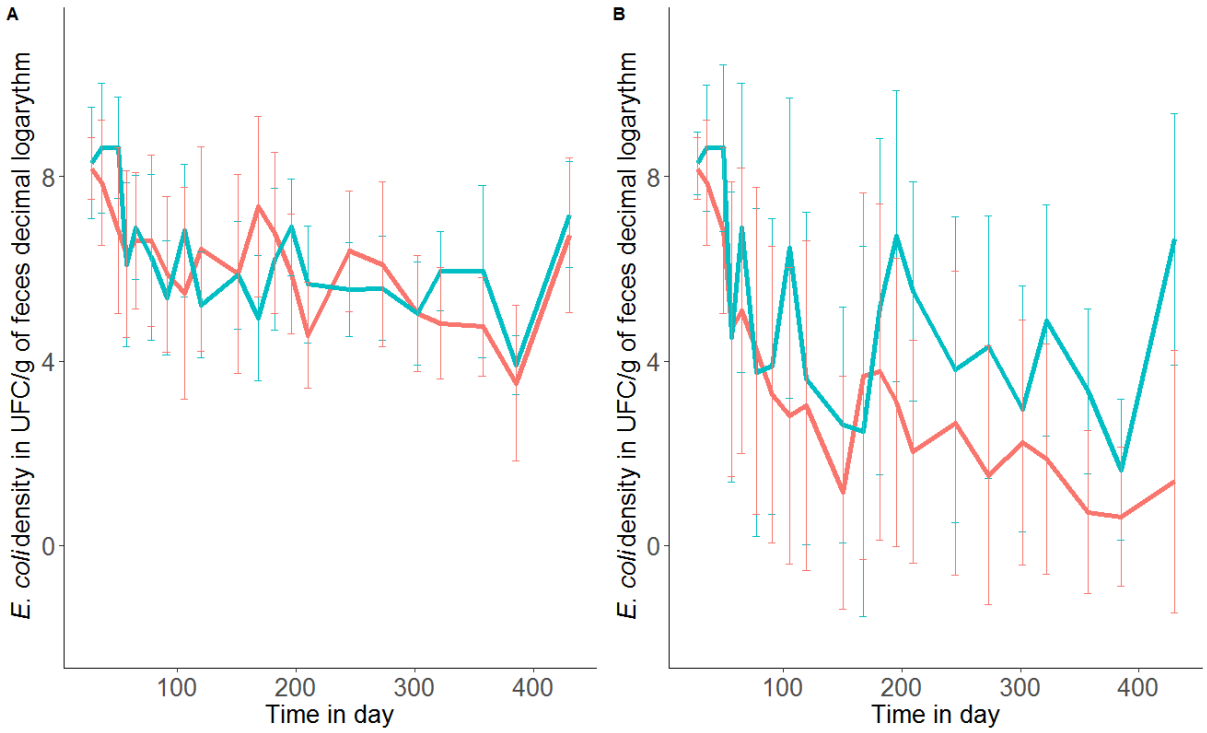
Table S3. List of the deletion or insertion mutations that were found in the 71 sequenced isolates of *E. coli* 536

Table S4. List of the transposition mutations that were found in the 71 sequenced isolates of *E. coli* 536



<u>Figure 1.</u> Experimental scheme, summary of the distribution and appearance of mutations over time

We represented in parallel the experimental scheme and the distribution of the mutations observed on strains extracted in the feces of mothers and offspring at different time points. *E. coli* 536 was inoculated in streptomycin-treated pregnant mice. Streptomycin was only used to release a colonization nest for *E. coli* in the gut of pregnant mice and was stopped at delivery to avoid exposing the offspring. Then, the mothers naturally transmitted *E. coli* along with the rest of their microbiota to their offspring. After a 22 to 23 days breast-feeding period, the offspring were distributed in 24 independent cages, with 3 mice per cage from different mothers. The contribution of mothers to a lineage is represented by grey box in the left table. Two diets were used: a high fat high sugar diet (Western diet, WD) for lineages 1 to 12 and regular diet (Chow diet, CD) for lineages 13 to 24. We represented the mutations found in *E. coli* lineages during the 22-23 days breast-feeding period and the rest of experiment. The length of each arrow is proportional to the duration of the follow-up, Cages which lost their lineage before the final time point were marked with a plain black cross followed by the last day when *E. coli* was found in, at least, one mouse in the cage. Each lineage was sampled at three times points (except for the lineage 18 because of its quick disappearance) represented by the tip of arrow and two vertical lines across it. The observed mutations in the sequenced genomes are represented by arrows or triangles. The nonsynonymous SNPs are represented by blue arrows, synonymous SNPs by red arrows, intergenic SNPs by black arrows and SNPs resulting in a stop codon by a crossed black arrow. The deletion mutations are represented by an orange triangle, the mutations by insertions are represented by a purple triangle, the mutations by duplication are represented by a green triangle.



<u>Figure 2.</u> Density of *E. coli* in mouse feces as a function of time.

A) The mean log10 density for WD (blue) or CD (red) group is presented for the independent cages in which *E. coli* 536 was detected. B) The mean log10 density is presented for the independent cages according to diet taking account the loss of *E. coli*. Error bars represent standard deviation.

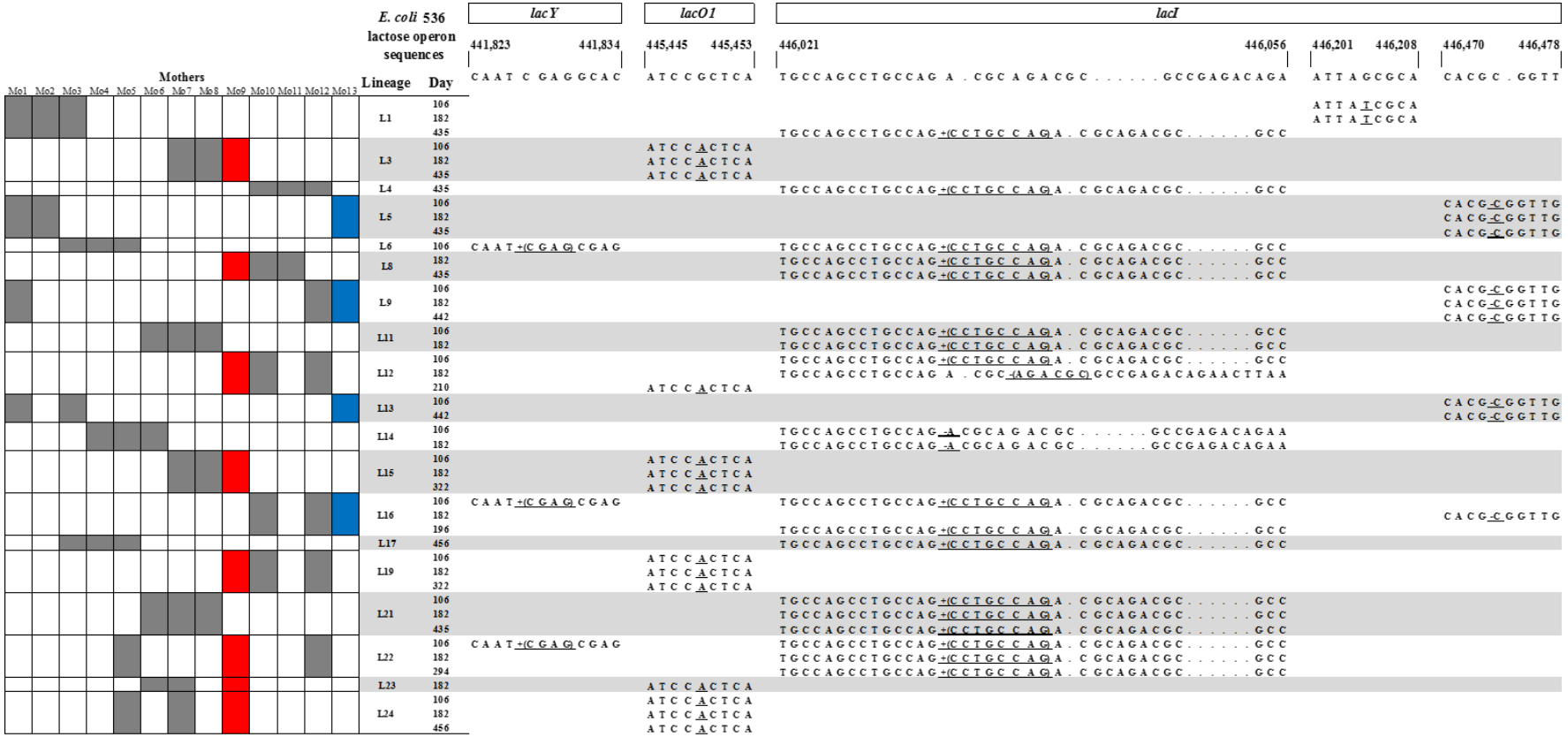


Figure 3. Time distribution of mutations in the *lac* operon and their characteristics

The reference sequence of the *lac* operon and mutations positions is indicated at the top of the figure. SNP mutations are represented by a nucleotide change, insertions by a (+) sign and deletions by a (-) sign followed by the nucleotide introduced or lost. The contribution of mothers to a lineage is represented by grey box in a table placed at the intersection of mother column and lineage row. The red boxes represents how the intergenic SNP in *lacO1* from Mo9 was propagated to the different cages, the blue boxes the *lacI* one base deletion from Mo13.

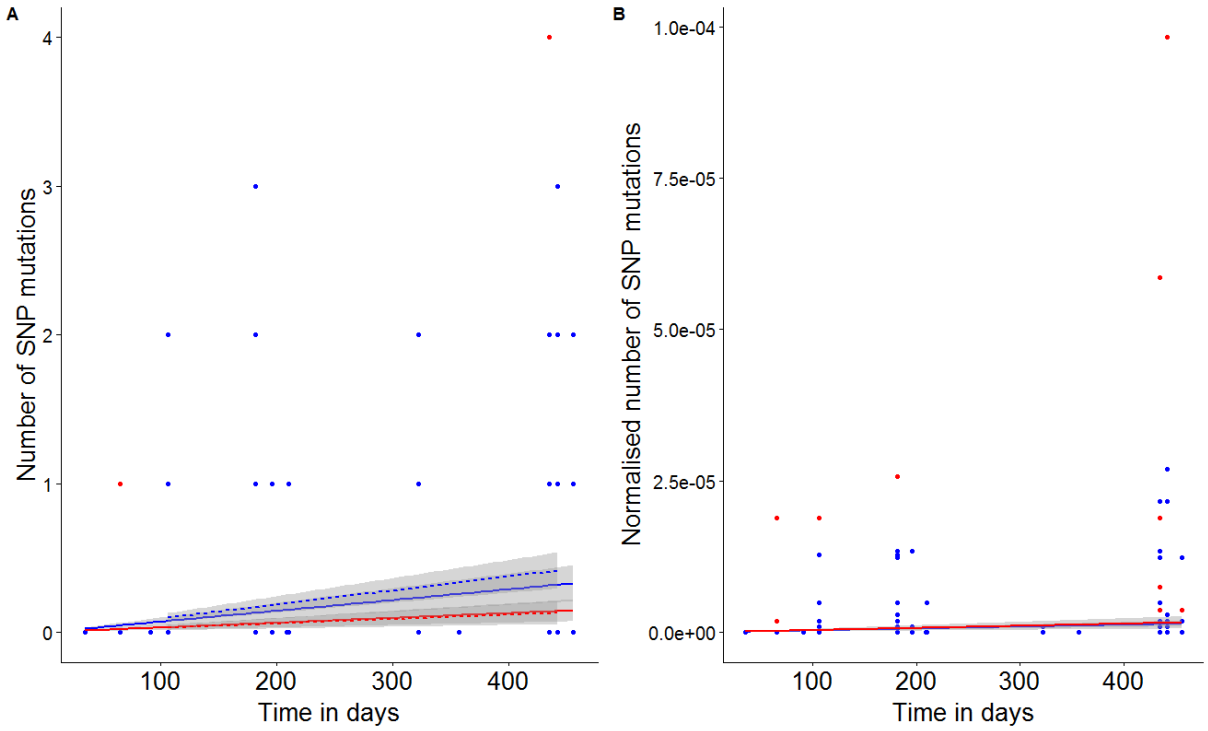


Figure 4. Accumulation of SNP mutations per genome per day after inoculation

A) Accumulation of synonymous (red) and non synonymous (blue) mutations according to diet. CD is in plain, WD in dashed. An ANOVA test did not reveal any significant difference between the two diets for the accumulation of synonymous ($p = 0.58$) or non synonymous ($p = 0.25$) mutations. B) Per site accumulation of synonymous (red) and non synonymous (blue) mutations. When normalizing mutation accumulation by the number of available sites of each type and the observed overall mutation bias, mutation per genome accumulated similarly for synonymous and non synonymous mutations. 95 confidence intervals of the regression slope are presented in grey.

Gene	Position	Mutation	Mutation type	Occurrence	Mice ID	Diet	Gene information
bcsG →	3,813,987	G→A	S493N (AGC→AAC)	2	L22D182, L22D322	CD	Putative endoglucanase involved in cellulose biosynthesis
ECP_1501 ←	1,551,577	C→T	G211D (GGC→GAC)	3	L7D106, L7D182, L7D435	WD	Hypothetical protein
ECP_3029 ←	3,189,472	C→T	G193D (GGC→GAC)	2	L5D106, L5D182	WD	Conserved hypothetical protein
ECP_4610 ←	4,817,157	A→T	Y102N (TAT→AAT)	3	L9D106, L9D182, L9D442	WD	Conserved hypothetical protein, putative Transcriptional regulator
ecpA ←	377,733	C→T	G77D (GGT→GAT)	2	L14D106, L14D182	CD	*E. coli* common pilus (ECP)
glpC →	2,393,197	G→T	V324V (GTG→GTT)	3	L20D106, L20D182, L20D322	CD	Sn-glycerol-3-phosphate dehydrogenase (anaerobic), small subunit
hscA ←	2,667,858	C→T	S110N (AGC→AAC)	2	L22D182, L22D322	CD	DnaK-like molecular chaperone specific for IscU
int – ECP_0342	294,395	Δ68,078 bp	NA	5	L11D442, L3D435, L10D182, L9D106, L6D182	WD	76 genes: int, xis, ECP_0275, ECP_0276, ECP_0277, ECP_0278, ECP_0279, ECOLI0280, ECP_0280, ECP_0281, ECP_0282, insF, insE, ECP_0285, ECP_0286, ECOLI0287, ECP_0287, ECP_0288, ECP_0289, ECP_0290, papI.2, papB, ECP_0293, ECP_0294, ECP_0295, ECP_0296, ECP_0297, ECP_0298, ECP_0299, ECP_0300, ECP_0301, ECOLI0303, ECP_0302, ECP_0303, ECP_0304, ECP_0305, ECP_0306, ECP_0307, insG, ECOLI0311, ECP_0309, ECP_0310, ECP_0311, ECP_0312, ECP_0313, ECP_0314, ECP_0315, ECP_0316, ECOLI0318, ECP_0317, ECP_0318, ECP_0319, ECP_0320, ECP_0321, ECP_0322, ECP_0323, ECP_0324, ECP_0325, ECP_0326, ECP_0327, ECP_0328, ECP_0329, ECP_0330, ECP_0331, flu, ECP_0333, ECP_0334, yeeS, yeeT, ECOLI0337, yeeU, yeeV, yeeW.2, ECP_0340, ECP_0341, ECP_0342
lacI ←	446,034	(CCTGCCAG)1→2	coding (592/1083 nt)	17	L22D106, L22D182, L22D456, L21D182, L21D106, L21D435, L17D456, L16D106, L16D196, L12D106, L11D106, L11D182, L8D182, L8D435, L6D106, L4D435, L1D435	WD (8) / CD (9)	DNA-binding transcriptional repressor
lacI ←	446,035	Δ1 bp	coding (591/1083 nt)	2	L14D182, L14D106	CD	DNA-binding transcriptional repressor
lacI ←	446,474	Δ1 bp	coding (152/1083 nt)	9	L16D182, L13D442, L9D106, L9D182, L9D442, L5D106, L5D182, L5D435, L13D106	WD (6) / CD (3)	DNA-binding transcriptional repressor
lacI ←	446,204	G→T	A141D (GCT→GAT)	2	L1D106, L1D182	WD	DNA-binding transcriptional repressor
lacI ←	446,039	(AGACGC)2→1	coding (582-587/1083 nt)	1	L12D182	WD	DNA-binding transcriptional repressor
lacY ←	441,830	+CGAG	coding (465/1254 nt)	3	L22D106, L6D106, L16D106	WD (1) / CD (2)	Lactose/galactose transporter
lacZ ← / ← *lacI*	445,449	G→A	intergenic (-29/+94)	14	L24D106, L24D182, L24D209, L23D182, L19D106, L19D182, L19D322, L15D106, L15D182, L15D322, L12D210, L3D106, L3D182, L3D435	CD (10) / WD (4)	Beta-D-galactosidase/DNA-binding transcriptional repressor
leuB ←	83,137	G→A	A48V (GCA→GTA)	2	L8D182, L8D435	WD	3-isopropylmalate dehydrogenase
malT →	3,660,530	(GCTGCCAGTGGGAACTGGCGGCGGA)1->2	coding (992/2706 nt)	5	L10D182, L17D456, L22D182, L22D456, L2D182	WD (2) / CD (3)	Fused conserved hypothetical protein ; DNA-binding transcriptional activator, maltotriose-ATP-binding
manX →	1,814,451	(TCTCCC)1→2	coding (568/972 nt)	2	L5D106, L5D182	WD	Fused mannose-specific PTS enzymes: IIA component ; IIB component
metN ←	226,259	A→T	L182Q (CTG→CAG)	4	L5D106, L5D182, L5D435, L13D106	WD (3) / CD (1)	DL-methionine transporter subunit ; ATP-binding component of ABC superfamily
oxyR →	4,368,673	G→A	L270L (CTG→CTA)	7	L23D91, L23D357, L21D182, L21D106, L21D435, L11D106, L11D182	WD (2) / CD (5)	DNA-binding transcriptional dual regulator
rihA ←	709,873	C→T	G155R (GGG→AGG)	2	L19D182, L19D322	CD	Ribonucleoside hydrolase 1
ubiD →	4,232,244	Δ1 bp	pseudogene (1407/1494 nt)	2	L22D182, L22D456	CD	3-octaprenyl-4-hydroxybenzoate decarboxylase
ydiO →	1,692,914	C→T	S315S (AGC→AGT)	2	L14D182, L14D106	CD	Putative acyl-CoA dehydrogenase
yfgB ←	2,653,765	C→T	L71L (CTG→CTA)	2	L22D182, L22D322	CD	Putative Fe-S containing enzyme
yhfK →	3,599,400	C→T	R345W (CGG→TGG)	2	L5D106, L5D182	WD	Conserved hypothetical protein; putative inner membrane protein

Table 1. Summary of several mutations that were found in the evolved lineages

The name of the genes as well as the orientation (defined by the arrow), the position along the genome, and the flanking genes in case of an intergenic mutation are displayed, with the mutations, their type, the number of occurrences, and a description of the gene products.

Theorical phenotype	Gene	Position	Mutation	Mutation type	Mother mice ID											
					Mo1	Mo2	Mo4	Mo5	Mo6	Mo7	Mo8	Mo9	Mo10	Mo11	Mo12	Mo13
lactose operon - constitutive activation	*lacI* ←	445,641	T→G	L332V (TTG→GTG)		9.4%										
		446,034	(CCTGCCAG)1→2	coding (592/1083 nt)	44.8%	68.1%	96.1%			76.4%	71.7%	93.7%	94.9%	27.3%	12%	7.1%
		446,039	(AGACGC)2→1	coding (582-587/1083 nt)											8.4%	
		446,432	C→T	G65D (GGC→GAC)						25.6%						
		446,471	A→T	V52E (GTG→GAG)												6.9%
		446,474	Δ1 bp	coding (152/1083 nt)												52.1%
		446,550	G→A	Q26* (CAG→TAG)												4.7%
		446,559	C→T	V23M (GTG→ATG)											11.9%	
		446,562	G→A	R22C (CGC→TGC)								5.4%				
		446,573	T→A	Q18L (CAG→CTG)											5.3%	
		446,595	C→A	E11* (GAG→TAG)											48.8%	
	lacZ ← / ← *lacI*	445,441	T→C	intergenic (-21/+102)				40.0%								
		445,448	C→T	intergenic (-28/+95)										66.7%		47.1%
maltose operon - inactivation (except for *malI*)	*malT* →	3,660,554	+25 bp	coding (1016/2706 nt)	9.9%		83.9%		45.4%		8.7%		28.6%	7.7%	10.8%	
	malQ ←	3,654,993	24 bp→24 bp	coding (1498-1521/2085 nt)						9.6%	12.2%					
	malK →	4,461,316	Δ3 bp	coding (1044-1046/1116 nt)		10.6%										
	malE ←	4,459,358	C→T	W184* (TGG→TAG)												8.3%
	malF ←	4,458,426	Δ1 bp	coding (139/1545 nt)	22.9%	5.2%										
	malI ←	1,617,723	G→A	T235M (ACG→ATG)		5.7%										

Table 2. Polymorphism of *E. coli 536* isolates extracted from the feces of mother mice at day 16 post inoculation

The proportion of the different alleles found in the mother's feces at day 16 after inoculation for *lac* and *mal* operons. Arrows indicate the orientations of the genes; a slash indicates the intergenic region. The asterisk (*) indicates a stop codon.

CHAPITRE III : ETUDE DE L'ADAPTATION DE DEUX ISOLATS NATURELS DE *E. COLI* 536 ET HS DANS LE TUBE DIGESTIF DE SOURIS TRAITEES A LA STREPTOMYCINE SOUMISES A DEUX REGIMES ALIMENTAIRES

I /Introduction

Dans notre précédente étude (Article n°2), nous avons pu mettre en évidence une sélection avec une forte convergence vers une activation constitutive de l'opéron lactose dans *E. coli* 536 apparaissant rapidement lors de la période d'allaitement de 3 semaines. Cette expression était secondaire à des mutations, notamment au niveau d'un point chaud, au niveau du répresseur transcritpionnel *lacI*. Des mutations dans les régulateurs des opérons et/ou centrées sur le métabolisme des sucres ont déjà été observées dans d'autres études d'évolution de *E. coli* dans le tube digestif de souris traitées à la streptomycine avec des mutations dans l'opéron *gat* et *srlR* (De Paepe *et al.* 2011), ou axéniques avec des mutations dans *malT* (Barroso-Batista *et al.* 2014). Dans notre modèle, la densité de *E. coli* dans les fèces était faible mais avec une différence significative entre les deux régimes. Le maintien et la densité de *E. coli* était plus forte dans les fèces de souris nourries avec un régime riche en gras et en sucre faisant écho à la forte proportion de *E. coli* du groupe phylogénétique B2 dans les pays occidentaux. Nous n'avons observé aucune sélection particulière durant la période de suivi d'un an. Deux hypothèses à cette observation: (i) *E. coli* 536 évolue de façon neutre et est bien adapté au tube digestif de souris ; ou (ii) la taille de population est trop faible, résultant du modèle, pour pouvoir détecter une sélection. Pour éviter cette limite et répondre à la question de l'adaptation de *E. coli* en fonction de son fond génétique et du régime alimentaire de l'hôte, nous avons mis en place une évolution expérimentale à long terme sur une période d'un an en utilisant des souris traitées à la streptomycine. Le but de cette expérience était de comparer l'évolution de 2 isolats naturels: un virulent, à nouveau *E. coli* 536, et *E. coli* HS purement commensal, appartenant au groupe A. Les souris ont également été soumises à deux régimes alimentaires différents.

II/Article n°3

Genomic patterns of *Escherichia coli* long-term adaptation to the mouse gut are largely shaped by the natural isolated used and the mice diet

Running title: Commensal *E. coli* adaptation depending on strain and diet

Mohamed Ghalayini [1,2,3#], Lucie Bourguignon [3,4], Audrey Chapron [3], Mélanie Magnan [1,3], Sara Dion [1,3], Mathilde Lescat [1,3,5] and Olivier Tenaillon [1,3]

[1]Université Paris 13, IAME, INSERM, F-93017 Bobigny, France

[2]AP - HP, Hôpital Avicenne, Service de Réanimation Médico-Chirurgicale, F-93000 Bobigny, France

[3]Université de Paris, IAME, INSERM, F-75018 Paris, France

[4]École de l'Inserm Liliane Bettencourt, Paris, France

[5]AP - HP, Hôpital Avicenne, Service de Microbiologie, F-93000 Bobigny, France

Mohamed Ghalayini : mohamed.ghalayini@inserm.fr ; Mélanie Magnan : melanie.magnan@inserm.fr ; Sara Dion : sara.dion@inserm.fr ; Lucie Bourguignon : luciebourguignon90@gmail.com ; Mathilde Lescat: mathilde.lescat@inserm.fr ; Olivier Tenaillon: olivier.tenaillon@inserm.fr

#: Corresponding author: Mohamed Ghalayini, IAME, INSERM UMR 1137, Faculté de Medicine Paris Diderot, Site Xavier Bichat, 16 rue Henri Huchard, 75018 Paris, France.
email: mohamed.ghalayini@inserm.fr

ABSTRACT

Escherichia coli has a clonal structure that makes it possible to differentiate several phylogenetic groups predictive of the virulence and commensalism of *E. coli*. It also has a partly variable structure subject to flows of genes from the environment. It is also subject to mutations capable of expressing a phenotypic trait that may persist depending on the selection imposed by the medium or the genetic drift. *E. coli* is partly shaped by its environment, which remains an open environment subject to external factors such as food. The objective of this study was to determine the factors influencing the adaptation of *E. coli* in its natural environment: the digestive tract. We were particularly interested in the influence of genetic background and diet on the adaptation of *E. coli* in streptomycin treated mice.

For more than one year, we studied the adaptation of *E. coli* 536 (belonging to phylogenetic group B2) extracted from feces of mice fed with a high fat and high sugar diet (Western Diet, WD) and *E. coli* HS coli extracted from feces of mice fed with WD diet or fed with a standard diet with fiber (Chow diet, CD). Strains were extracted at three time points, sequenced and mutations were analyzed.

We found a stronger signal of selection in *E. coli* HS than *E. coli* 536 mainly related to mutational convergence in genes of sugar metabolism. In particular, convergent mutations on Entner-Doudoroff pathway substrates (gluconate and 2-kéto-3-deoxy-gluconate) transporter gene *gntT*, especially in WD, and *kdgT* exclusively in CD. We found other convergent mutations on gene encoding for mannose, ribose and maltose metabolism in *E. coli* HS and on maltose and lactose genes in *E. coli* 536 in WD. We also observed other convergent mutations conferring theorically a constitutive activation of cpx system, on amino-acids metabolism genes in anaerobic condition and genes involved in production of energy in low O_2 condition.

The very preliminary results of this study allowed us to distinguish two adaptation profiles according to the strain combined with the diet. It seems that the intensity of adaptation depends on the strain and probably its genetic background, and that the direction of adaptation depends on the diet.

KEYWORDS

Escherichia coli, adaptation, selection, *in vivo* long-term evolution, sugar metabolism, diet

INTRODUCTION

Escherichia coli is a commensal bacterium of the gut, which, depending on the host and the strain genetic specificities can turn into a terrible pathogen causing intra or extra-intestinal diseases. With the rise of antibiotic resistance and the propagation of more virulent and resistant clones not only in the hospitals but also in the community, *E. coli* constitutes now a major public health problem. Interestingly, the commensal niche is both a reservoir for some *E. coli* strains that may produce later on infections if they colonise other sites such as the bladder or the blood, and a place targeted recurrently by antibiotics in which resistance is most likely to emerge. It is therefore crucial to determine the factors that can influence the behavior of *E. coli* in this reservoir.

Several *in vivo* experiments have been conducted to determine the adaptation mode of *E. coli* in its natural environment. Genomic data revealed convergences in the inactivation of genes or constitutive activation of operons involved in sugar metabolism. De Paepe *et al.* followed the adaptation of *E. coli* K-12 in axenic mouse gut. They reported a convergence of mutations in the *ompBSG1* gene, in the *flhDC* operon, and the *malT* gene with an impact on the growth of these strains in the presence of all or some sugars (De Paepe *et al.* 2011). Later, Barroso-Batista *et al.* followed the adaptation of *E. coli* K-12 MG1655 in streptomycin-treated mice and observed mutations in the *srlR* gene coding for the repressor of the *slr* operon and/or in several genes of the *gat* operon, these operons being respectively involved in the metabolism of sorbitol and galactitol. Evolutive experiments following a natural isolate of *E. coli,* strain 536 in streptomycin-treated mice gut or during a 24 days breastfeeding period have demonstrated an inactivation of the *dgo* operon, coding for galactonate metabolism, or a constitutive activation of the lactose operon respectively (Lescat *et al.* 2017; Ghalayini *et al.* 2019). Mostly mutations in sugar pathways seemed to be recruited during colonization; neither Phospholipid and amino acid catabolism, nor gluconeogenesis, tricarboxylic acid cycle, the pentose phosphate pathway were affected. Therefore these *in vivo* evolutive experiments suggest that sugar metabolism is a central determinant of growth and a prime target for the adaptation of *E. coli* to the mouse gut.

The gut is an open system; as such it is affected by outside factors among which the diet plays an important role being the main provider of sugar. The diet has been shown to impact the composition of the microbiota (Murphy, Velazquez, et Herbert 2015) and could therefore also

affect the behavior of *E. coli* (Tenaillon *et al.* 2010). Accordingly, the increasing proportion of *E. coli* from phylogroup B2 in the gut of inhabitants of industrialized countries over the last 30 years could reflect an impact of western dietary habits (Massot *et al.* 2016). Moreover, experiments with *E. coli* 536, a strain from that phylogroup, revealed a better maintenance in mice fed with a western diet rich in fat and sugar compared to a standard fiber rich diet (Ghalayini *et al.* 2019).

E. coli 536 has a genetic background characterized by the expression of extra-intestinal virulence factors favoring the colonization of *E. coli* in gut, virulence appearing as a byproduct of commensalism (Diard *et al.* 2010). *E. coli* 536 seems to evolve neutrally in mouse gut colonization model as close as possible to natural conditions (Ghalayini *et al.* 2019) or with a weak selection signal in streptomycin treated mice gut (Lescat *et al.* 2017). Moreover, the observation, for about a year, of a neutral evolution of *E. coli* ED1a in a western man gut suggests a link between this genetic background and an adaptation mode common to B2 phylogenetic group in gut (Ghalayini *et al.* 2018). *In vivo* evolutionary experiments have been conducted with modified laboratory strains, or a natural strain which, by the clonal nature of *E. coli*, cannot be representative of the entire species. The question arises about the adaptation mode on different genetic background, for example a strain belonging to phylogenetic group A, devoid of factor of extra-intestinal virulence. Moreover, the question arises of the link between diet, adaptation and genetic background.

To answer these questions, we have evolved for more than one year two strains of *E. coli* 536, belonging to the group phylogenetic B2 reputedly virulent extra-intestinal and HS, belonging to the group phylogenetic A known to be commensal without pathogenicity, in the mice gut treated with streptomycin. We subjected *E. coli* HS colonized mice to a high fat and high sugar diet (Western diet, WD) and a standard diet with fiber (Chow diet, CD). While mice colonized with *E. coli* 536 were submitted to a WD diet, the study of the evolution of *E. coli* 536 in streptomycin-treated mice with CD had already been conducted (Lescat *et al.* 2017).

MATERIAL AND METHODS

Bacterial strain

In this study, we used strain *E. coli* 536 from phylogenetic group B2 that has a serotype O6:K5:H31. The 536 strain has been isolated from a patient suffering from urinary tract

infection in the early 1980s (Berger *et al.* 1982) and subsequently shown to be a good gut colonizer in streptomycin-treated mouse (Diard *et al.* 2010). This strain was acquired by our laboratory in 1998 from Jorg Hacker who isolated it and then stored at -80 °C with glycerol. This strain is naturally resistant to streptomycin and we have tested it in many mouse gut colonization protocols (Lescat *et al.* 2017; Ghalayini *et al.* 2019).

We also used *E. coli* HS (phylogenetic group A0, serotype O9:H4) obtained from the feces of a healthy man without any sign of acute or chronic gastrointestinal pathology. As such it appear to be a good model for gut commensalism. This isolate was originally collected in 1978 and stored at the Center for Vaccine Development at the University of Maryland (Levine et Rennels 1978). It is a strain that can colonize the streptomycin-treated (non-consanguine CD-1) mouse gut with a density of approximately 10^9 CFU/g of feces at 2 weeks (Maltby *et al.* 2013). Moreover, the complete sequencing of this strain was carried out (Rasko *et al.* 2008) and available on MAGE genoscope (http://www.genoscope.cns.fr/agc/microscope/mage/viewer.php?O_id=296). This strain was not naturally resistant to streptomycin. The natural resistance of *E. coli* 536 to streptomycin is conferred by a G to A point mutation at the 3,586,077 position in *rpsL* gene coding for 30S ribosomal subunit protein S12 (replacing the AGA codon coding for arginine by codon AAA coding for lysine). We introduced that mutation in HS genome through the electroporation of a 70 base oligonucleotide containing the mutation (5'-CGGTGTGGTAACGAACACCCGGGAGGTCTCTAACACGACCACCACGGATCAGGATCACGGAGTGCTCCTG-3') in a *E. coli* HS strain containing a recombinase on a previously introduced thermosensitive plasmid inspired by the technique of B. Wanner (Datsenko et Wanner 2000). The presence of the mutation was confirmed by Sanger sequencing.

Mouse intestinal colonization assay and sampling strategy

All animal experiments were performed in compliance with the recommendations of the French Ministry of Agriculture and approved by the French Veterinary Services (accreditation A 75-18-05). The mouse gut colonization model was conducted after approval by the Debré-Bichat Ethics Committee for Animal Experimentation (Protocol Number 2012-17/722-0076) in accordance to the European Decret and French law on the protection of animals. Intestinal colonization by 536 and HS *E. coli* strains was performed using a mouse model as described elsewhere (Lescat *et al.* 2017; Diard *et al.* 2010). Sixty-nine 7-week-old female outbred mice numbered from S1 to S69 (Janvier lab Swiss CD-1) were housed in 21 cages (numbered from C1 to C21) during 351 to 353 days.

In order to study the impact of diet on the evolution of the two different strains, the experimental scheme included three groups of mice: 24 mice colonized with *E. coli* 536 subjected to the high sugar and high fat diet, Western Diet (WD), (B2WD), 18 *E. coli* HS colonized mice fed the the same WD diet (A0WD) and 21 *E. coli* HS colonized mice under a fiber rich diet, Chow Diet (CD), (A0CD). Here, we have worked on only three groups, the study of the long-term evolution of a strain of *E. coli* 536 in the mouse gut subjected to a standard diet had already been the subject of a previous study (6). The food provided to the mice throughout the experiment was sterile.

Five days before inoculation of the strains, and throughout the experiment, streptomycin was added to the sterile drinking water at a final concentration of 5 g/L. Mice free of coliform flora (controlled by plating the faeces on Drigalski plates) were inoculated per os with 10^6 bacteria in 200 µl of physiological serum. For practical reasons, at each sampling time, it took three days to sample all mice. The day after the inoculation and during the experiment (Days 5-6-7, 26-27-28, 54-55-56, 82-83-84, 110-111-112, 138-139-140, 166-167-168, 201-202-203, 230-231-232, 264-265-266, 292-293-294, 320-321-322 and, at least, days 350-351-352-353), the density of gut population of *E. coli* was estimated by plating dilutions of weighted fresh faeces for each mouse on Drigalski agar plates and supplementary samples were performed for sequencing assays (Days 110-111-112, 230-231-232 and days 350-351-352-353). Faeces were harvested for each mouse by putting the mouse in a sterile plastic cage until it defecated. The faeces were stored at -80°C with glycerol for further analysis.

High throughput sequencing

For sequencing, the frozen faeces were diluted and plated on Drigalski. We observed differences in sizes between colonies on some plates, as reported in previous evolution experiments (Barroso-Batista *et al.* 2014). Two colonies per Drigalski plate were then randomly picked with a sterile tip, grown in LB with streptomycin, stored at -80°C with glycerol and thereafter called "isolate". Mice being coprophages, it is considered that studying an isolate extracted from the feces of a single mouse per cage is representative of the population and the dominant strain of *E. coli* present in the cage (Lescat *et al.* 2017). Thus, the mouse with the highest densities of *E. coli* was favored in order to increase the probability of obtaining isolates from its feces. Finally, the 126 isolates obtained were labelled by the mice and cage ID into lineages (lineages 49, 57, 60, 63 66 and 69 for *E. coli* HS in WD; lineages 72, 74, 78, 79, 84, 87 and 90 for *E. coli* HS in CD; lineages 3, 6, 9, 11, 13, 17, 19 and 23 for *E. coli* 536 in WD),

and a suffix was added to illustrate colony size (A bigger than B) and sampling day after inoculation (J).

Total DNA was extracted using QIAamp® DNA Mini Kit (Qiagen). The 126 genomes of these isolates were then sequenced using the Illumina platform HiSeq® or MiSeq®. To confirm that the isolates evolved from the *E. coli* 536 or HS isolate, we performed a PCR screening for B2 or A phylogroup (Clermont *et al.* 2013).

Mutation rates analysis

Mutation rates and non-synonymous to synonymous selection bias ω was estimated through a maximum likelihood approach taking into account mutation biases. The probability to observe a strain with genotype {s1, s2, s3, s4, s5, s6, n1, n2, n3, n4, n5, n6} was sampled at day t after inoculation was derived as

$$P(s_1, s_2, s_3, s_4, s_5, s_6, n_1, n_2, n_3, n_4, n_5, n_6, t) = \prod_i Poisson(s_i, m\varepsilon_i S_i t) Poisson(n_i, \omega m \varepsilon_i N_i t)$$

with S_i (N_i) the number of sites were a mutation of type i was (non)synonymous and s_i (n_i) the number of observed (non)synonymous of type i, m the mutation rate of the most abundant transition GC to AT (i=5), and ε_i the observed bias in mutation rate of type i relative to the transition GC to AT ($\varepsilon_i = \frac{\sum_{all\ genotypes}(n_i+s_i)/(N_i+S_i)}{\sum_{all\ genotypes}(n_5+\varepsilon_5)/(N_5+S_5)}$). The annual mutation rate per base μ was then computed as $\mu = 365.25 \frac{\sum_i m\varepsilon_i(N_i+S_i)}{\sum_i(N_i+S_i)/3}$. The optim function of R (« R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. », s. d.) was used to optimize the log likelihood, fitting for ε_i, m and ω. As optimizing for ε_i, did not improve the likelihood, we just report results in which ε_i are inferred from the data as presented ahead. The likelihood of a model in which μ and ω are estimated was compared to a simpler model in which ω was set to one.

RESULTS

Mouse colonization and E. coli density

To study the adaptation of *E. coli* in streptomycin treated mice gut, we inoculated 48 streptomycin-treated 7 weeks old female mice with *E. coli* 536 and 78 with *E. coli* HS. All mice inoculated *with E. coli* 536 were fed with WD diet, while 36 mice inoculated with *E. coli* HS

were fed with WD and 42 with CD. Mice were housed by 3 in independent cages, we considered that the *E. coli* lineages evolved in the gut of the three mice of the same cage as previously tested (Lescat *et al.* 2017). Thus, we followed 8 lineages in the *E. coli* 536 / WD group, 6 lineages in *the E. coli* HS / WD group and 7 lineages in the *E. coli* HS / CD group.

The mean density of *E. coli* 536 was $4.84x10^7$ UFC per gram (g) of feces (CI95%: $3.98x10^7$ – $5.9x10^7$) and HS strain was $4.44x10^7$ UFC/g of feces (CI95%: $2.73x10^7$ – $5.27x10^7$) in WD and $2.73x10^7$ UFC/g of feces (CI95%: $2.23x10^7$ – $3.34x10^7$) in CD. Except for the first 82 days, it appears that the average densities of *E. coli* in the three groups follow a parallel trajectory. Specifically, the density of *E. coli* in the three groups converge to a mean density of $5.07x10^7$ UFC/g of feces (CI95%: $2.64x10^7$ – $9.76x10^7$) at day 82 post inoculation before decreasing to a mean density of $8.94x10^6$ UFC per g of feces (CI95%: $4.9x10^6$ – $1.63x10^7$) at day 230, before increasing in all three groups until reaching a mean density of $2.98x10^8$ UFC/g of feces (CI95%: $1.93x10^8$ – $4.59x10^8$) (Figure 1).

We did not find any significant difference in density over time between the three groups (AUC, Kruskall-wallis test, p=0.6), the potential difference between the groups is probably masked by the use of streptomycin. There is probably no difference in population size between groups.

Comparative diversity in E. coli 536 and E. coli HS

According to genetic population, genetic diversity depends on the effective population size and mutation rate. *E. coli* density is probably not a good representation of the effective population size (Ghalayini *et al.* 2018), but the relative equivalence of *E. coli* density in all groups can allow us to hypothesize a comparable effective population size.

We first studied genomic differences among the *E. coli* 536 isolates sampled. Overall we observed 162 independent mutations on 48 isolates. Among these 123 SNP mutations have been identified of which 79 non-synonymous SNP with 11 stop SNP, 27 synonymous SNP, 15 intergenic SNP and 2 SNP in pseudogenes. Of the 39 remaining mutations, 30 were indels among these 18 deletions of less than 1000 bp, 4 duplication and 8 insertions. Using a maximum likelihood estimation of SNP accumulations through time mutation, we computed a Ka/Ks ratio (noted ω) between the accumulation of non-synonymous and synonymous SNP of 2.02 (CI95%: 1.39 – 3.01). The mutation rate, weighted by mutation bias and synonymous and non-synonymous ratio ω, was found to be $1.22x10^{-6}$ (CI95%: $8.52x10^{-7}$ – $1.68x10^{-6}$) mutations per base par year.

We observed a higher number of mutations in *E. coli* HS with 798 independent mutations on 568 genes. Of these 798 independent mutations 475 were SNP mutations (241 independent mutations in CD, 91 independent mutations in WD and 7 mutations shared in the two diets). We observed 83 synonymous mutations (63 independent mutations in CD, 20 in WD and none shared between the two diets), 326 non-synonymous (208 independent mutations in CD, 113 in WD and 5 shared between the two diets) among which 46 stop SNP (16 independent mutations in CD, 27 in WD and 3 shared between the two diets), 49 intergenic SNP (35 independent mutations in CD, 12 in WD and two shared between the two diets) and finally 17 SNP non-coding pseudogene region (10 independent mutations in CD, 7 in WD and none shared between the two diets). Of 323 remaining mutations, 238 were indels with 136 deletions of less than 350 bp (34 independent mutations in WD, 93 in CD and 9 shared between the two diets), 88 duplications (25 independent mutations in WD, 56 in CD and 7 shared between the two diets) and 14 insertions (5 independent mutations in WD, 7 in CD and 2 shared between the two diets). Interestingly, the great majority of indels in *E. coli* HS (66%) were the product of the deletion or insertion of one base pair in mononucleotide repeats. 59 of them were located in intergenic region possibly in DNA binding site for a regulator or a gene promoter and 86 of them were located in coding gene. So, we found a higher number of changes in *E. coli* HS than *E. coli* 536 especially in CD. However, a closer look at the data revealed some heterogeneity among HS isolates suggesting the presence of hypermutator strains.

We found that 22 *E. coli* HS isolates carried mutations in gene involved in mismatch repair system that result in hypermutator strains. In three lineages (lineages 57, 72 and 79) these genotypes had presumably fully invaded the populations at day 203, in one (lineage 74) it was in all genomes (2 isolates) at the final time point at day 351. Several other mutator isolates were also identified in lineage 74, 79 84, but did not fix, being outcompeted in some cases by other hypermutator alleles. So it seems that hypermutator invaded 3 lineages in CD and one in WD lineage. *mutL* was the main gene in the mismatch repair system to be mutated with three different alleles recovered, all indels, occurring between positions 4,422,361 and 4,422,367 in a 3-fold repeated region of TGGCGC sequence, a known mutation hotspot in that gene (Li *et al.* 2017). Others mutations are Insertion Sequence transpositions at position 2,988,577 in *mutH* and at position 2,875,090 in *mutS*. Another *mutS* mutation (a stop SNP at position 2,875,090) was also found. Isolates carrying these mutations have clearly a hypermutator phenotype trait compared with isolates which did not carry these mutations (comparison of linear regression by ANOVA, p-value $< 10^{-16}$).

In order to compute a mutation rate, we first need to get rid of SNP mutations from hypermutator isolates that may have some confounding effects (Couce *et al.* 2017). Eliminating the 22 hypermutator isolates, we observed 252 non-synonymous (155 non-synonymous SNP in WD on 29 isolates and 97 in CD on 26 isolates and only 13 synonymous SNP (10 synonymous SNP in WD and 3 in CD). Using a maximum likelihood approach, we computed a Ka/Ks ratio (ω) between the accumulation of non-synonymous and synonymous SNP of 9.90 (CI95%: 5.88 – 14.12) and a mutation rate of $4.6x10^{-7}$ (CI95%: $2.95x10^{-7}$ – $7.5x10^{-7}$) mutations per base par year. We computed an ω of 8.15 (CI95%: 4.57 – 14.12) with a mutation rate of $6.23x10^{-7}$ (CI95%: $3.39x10^{-7}$ – $1.08x10^{-6}$) in WD and an ω of 15.55 (CI95%: ? - ?) with a mutation rate of $1.53x10^{-7}$ (CI95%: ? - ?) in CD.

Including the data of Lescat *et al.* to have both diets for the two strains, and eliminating hypermutator strains, we observed no differences in the number of synonymous and non-synonymous mutations across the two diets. We found however a marked strain effect, HS having simultaneously a lower rate of synonymous mutations (Poisson regression adjusted for time sampling and diet, p=0.005) regardless of diet (Poisson regression adjusted for time sampling and strain, p=0.2), and more nonsynonymous mutations (Poisson regression adjusted for time sampling and diet, p=0.01) regardless of diet (Poisson regression adjusted for time sampling and strain, p=0.74). Furthermore, discarding hypermutator strains that are known to generate a large number of indels, we also observed a strain effect on indels accumulation, HS havng more than 536 (Poisson regression adjusted for time sampling and diet, $p<10^{-6}$) regardless the diet (Poisson regression adjusted for time sampling and strain, p=0.76). We observed 46 indels (22 of all isolates in WD, 20 in CD and 4 shared between the two diets) including 30 deletions (14 of all isolates in WD, 14 in CD and 2 shared in the two diets), 11 duplications (5 isolates in WD, 5 in CD and one shared in the two diet) and 5 insertions (one isolate in WD, 3 in CD and one shared in the two diet). Of note, only 8 indels (17%) instead of 59 (66%) when hypermutator strain were included occurred in mononucleotide repeats.

We observed 22 large deletions (> 1,000 bp) in *E. coli* HS in 39 isolates (14 larges deletions in CD, 8 in WD and none shared between the two diets). We distinguished 6 hotspot regions for large deletion in *E. coli* HS. Two large deletions of 3704 bp and 6940 bp beginning around position 800,000 and comprising, among others, *cydA* or/and *cydB* gene were found in 1 or 2 isolates (respectively 60BJ202 and 63BJ84/63BJ203) in WD. We noticed a region between position 1,927,210 and 1,941,766 susceptible to be deleted (a region of around 14,556 bp) with the loss of 14,556 bp in isolate 90AJ351 and 3,696 bp in isolates 66AJ322 and 66BJ322. We

found a large genomic region susceptible to be deleted between position 2,109,390 and 2,182,933 (a region of around 74,000 bp) with a median loss of 28,679+/-19,386 bp in 19 isolates (18 of all isolates in CD, 1 in WD and none shared between the two diets). This large region was surrounded by transposases. Another large genomic region between position 2,800,568 and 2,829,548 (a region of around 29,000 bp) was susceptible to be deleted with a minimum of 1,337 bp and a maximum of 28,980 bp. Finally, we found a large deletion of 12,822 bp at position 3,591,863 in isolates 90BJ203 and 90BJ351 which deleted notably genes involved in gluconate transport and maltose metabolism; and another large deletion of 1,281 bp at position 4,274,718 in isolates 84AJ351, 84BJ84, 87AJ203 and 87BJ203 which affected genes involved in maltose metabolism. We also observed 4 large deletions in *E. coli* 536 in 5 isolates at position 294,387 with a defect of 68,039 bp for isolate 9AJ353, at position 2,139,900 with of defect of 19,522 bp for isolate 11BJ353, at position 4,021,738 with a defect of 1,629 bp for isolate 6AJ353 and at position 294,395 with a defect of 68,030 bp for isolate 3AJ353 and 3BJ353 (Table 2). Including datas of Lescat *et al.* and not taking into account hypermutator strains, we found once again a higher rate of large deletion in *E. coli* HS (Poisson regression adjusted for time sampling and diet, $p<10^{-4}$) compared to strain 536, with a higher prevalence of deletions in CD (Poisson regression adjusted for time sampling and strain, p=0.005).

Finally, we noticed a strong activity of replicative transposition exclusively in *E. coli* HS. 64 transpositions were found in that genetic background against 5 in strain 536. Among the observed 64 IS transpositions, 52 were from IS1 elements (14 in WD, 35 in CD and 3 shared between the two diets), 7 were from IS110 (none in WD, 7 in CD and one shared between the two diets), one were from IS3 in CD and 4 were from putative transposase (one in WD and 3 in CD). We found a higher rate of IS transposition, including datas of Lescat *et al.* and excluding hypermutator strains, in *E. coli* HS than *E. coli* 536 (Poisson regression adjusted for time sampling and diet, $p<10^{-15}$) regardless of diet (Poisson regression adjusted for time sampling and diet, p=0.28).

Overall, we found much more mutations in *E. coli* HS than in *E. coli* 536, even including data on the evolution of *E. coli* 536 in streptomycin treated mice in CD (Lescat *et al.* 2017). This is partially due to the occurrence of hypermutator isolates, but is also driven by a higher transposition activity in HS, a higher rate of large deletions and a larger rate of non-synonymous mutation despite a lower synonymous rate of mutation, the later suggesting a larger contribution of selection.

These results, notably the selection of hypermutators and a higher rate of nonsynonymous evolutions associated with a much larger Ka/Ks suggest that selection is largely contributing to genome evolution in strain HS and to a much lower extent in 536. Beyond mutations counts and the estimation of Ka/Ks, we can use the signature of genetic convergence to detect signals of selection: finding repeatedly mutations in one gene suggest the filtering action of natural selection. We observed several convergences on *E. coli* HS in both diets and in a lesser extent on *E. coli* 536 in WD. Before going into the details of the nature of convergence, we first did a simple global analysis in which we used only the genes involved in convergence (genes found at least 3 times in independent lineages) to cluster strains with a Principal component analysis. Such an analysis reavealed several important differences between the two diets and two type of strains (Figure 2). First, as just mentioned, 536 stayed much closer to the ancestral genotype than HS. Second 536 and HS occupied different regions of the space suggesting different adaptations. Third, while 536 showed a poor discrimination between the two diets, the effect was much more contrasted for HS in which both diets used clearly different directions.

Weak convergence between strains 536 and HS

We have not been able to highlight many convergences between the strains, except mutations in the *gidB* gene, in the intergenic region between *yjiH* and *kptA* and in the maltose operon.

We first found ubiquitous convergent mutations in *gidB* gene in both strain and both diet (Lescat *et al.* 2017). These alleles seems to associated with a higher growth in streptomycin (Lescat *et al.* 2017). We identified 20 independent mutations in *E. coli* HS with 15 non-synonymous SNPs including 7 stop SNPs, 1 IS transposition, and 4 indels. While we found 13 alleles in *E. coli* 536 with 8 non-synonymous SNPs including 2 stop SNPs, 3 insertions with frameshift and 2 deletions. Convergent mutations to inactivation of *gidB* invade significantly more *E. coli* HS than *E. coli* 536 (80% of isolates in *E. coli* HS and 41% in *E. coli* 536 including Lescat *et al.* data, chi2 test, $p<10^{-5}$) (Lescat *et al.* 2017). In fact, in 536, we observed another convergence conferring an advantage in media with streptomycin, in *rluD*. We observed 8 alleles with 3 non-synonymous SNPs including 2 stop SNPs and 5 indels. Convergent mutations in *rluD* were exclusively found in *E. coli* 536 with none in *E. coli* HS despite the presence of the gene in this strain. As found previously, mutations in *rluD* gene and *gidB* gene were exclusive of one another, they were not present together in the same isolate (Lescat *et al.* 2017).

The inactivation of maltose operon was observed in the two strain and in the two diets. (Lescat *et al.* 2017). The analysis of the mutations present in the operon reveals a certain polymorphism with 17 alleles present on 53 isolates of *E. coli* HS distributed in 3 lineages in WD and the 7 lineages in CD. Two other alleles were detected in 9 isolates of *E. coli* 536 distributed in 2 lineages. We mainly observed mutations in *malT* regulatory regions and its target regions. 7 alleles present on *E. coli* HS and the 2 alleles of *E. coli* 536 target the transcriptional regulator *malT* resulting in the activation of the expression of the genes of the maltose operon. These alleles are the result of 4 non-synonymous mutations, 2 frameshifts, a transposition or a large deletion suggesting a strong selection for the inactivation of the maltose operon. Two mutations by IS transposition affected two intergenic regions malE/malK in lineage 78 and malF/malE in 13 isolates distributed in 7 lineages (60, 63, 72, 78, 84, 87 et 90). The first intergenic region is a DNA binding site of *malT* activator (Danot et Raibaud 1994; Boos et Shuman 1998). The second is a repetitive extragenic palindrome element that is a short base pair sequence capable of producing a stem-loop structure. They may act as mRNA stabilizers within an operon (Messing *et al.* 2012) and are a hotspots for transposition and recombination events (Stern *et al.* 1984). Gene *malF* is an integral membrane component of the maltose ABC transporter (Shuman, Silhavy, et Beckwith 1980) counting 5 alleles resulting from two IS transpositions, a frameshift, a non-synonymous mutation and a large deletion integrating also affecting *malG* possibly inhibiting the transport of maltose (Merino et Shuman 1998; 1997; Ehrle *et al.* 1996). Two mutations on the gene *malE* (encoding the periplasmic substrate-binding component of the maltose ABC transporter) by a replicative transposition and *malP* (encoding for maltodextrin phosphorylase) by a non synonymous SNP with one isolate each.

We found convergent mutations, exclusively SNP, in the intergenic region between genes *yjiH* and *kptA* concentrated between position 4,565,756 and 4,565,781 distributed in 9 isolates all occurring at the last time point at day 351 (except one at day 203) in 5 lineages (lineages 74, 79, 84, 87 and 90) in CD. Interestingly, these convergent mutations were also found in a previous *E. coli* experimental evolution in streptomycin treated mice fed with CD (Lescat *et al.* 2017). This convergence is therefore the only convergence specific to CD diet regardless of the strain. Unfortunately, the specific function of gene *yjiH* and *kpta* and moreover of their intergenic region remain unclear.

Weak differences between diets in E. coli 536

The signature of convergence in *E. coli* 536 across diet was weak. As mentioned, mutational convergence regardless of the diet was found in the genes involved in ribosome maturation and

is most likely related to streptomycin treatment given to mice (*gidB* and *rluD*). To a lower extant, Lescat *et al.* found 2 mutations on the *malF* gene that can confer an inactivation of the use of maltose and no mutation on *malT*. Mutations in the intergenic region between the *yijH* gene and *kptA* were found only in the CD and appear more specific to the diet (Lescat *et al.* 2017).

Finally, we observed a small convergence in the constitutive activation of the lactose operon by mutations on the *lacI* transcriptionnal repressor and an allele in the *lacO* intergenic region already observed on the 536 strain evolving in more natural conditions (Ghalayini *et al.* 2019). These 4 alleles are present only on 9 isolates in 4 lineages (lineages 9, 13, 19 and 23) and never reached high frequencies. Moreover, while Lescat *et al.* had shown a strong convergence in the constitutive activation of the operon galactonate with *dgoR* mutations (Lescat *et al.* 2017), we did not observe it in our study on *E. coli* 536 or HS for both diets. This weak signal of convergence in *E. coli* 536 contrasts with the one found in *E. coli* HS.

Convergent mutations centered on the sugar metabolism in E. coli HS

Most of the convergences observed in that strain concern the metabolism of sugar and notably ribose, maltose, mannose, gluconate and its derivate 2-keto-3-deoxy-D-gluconate. We mainly found mutations in regulatory genes (*malT, gntR, kdgR*) or DNA binding site for regulators inside the operon suggesting a global adaptation depending on the disponibility of the energetic substrate (Ghalayini *et al.* 2019). We also observed mutations in gene encoding for sugar transporter sub-unit (*rbsA, rbsB, rbsC, malE, malF, malG, gntT, kdgT* and *manZ*) suggesting an adaptation depending on the presence of sugar in the environment. The rest of mutated genes concern the utilization of these substrates (*rbsD, cpsG, cpsB* and *manA*) (Figure 3 and 4).

The analysis of mutations in the *rbsDACBKR* operon show in gene coding for subunit transporter of ribose (*rbsA, rbsC* and *rbsB*) and for *RbsD* encoding for a ribose pyranase; it binds specifically to the β-furan and β-pyran forms of ribose and increases the exchange rate between the two forms. An *rbsD* defect abolishes pyranase activity is impaired in the utilization of D-ribose as a source of carbon and energy (Ryu *et al.* 2004). We observed 17 mutations on this operon on 47 isolates spread in the 6 lineages in WD and 6 on 7 lineages in CD with significant difference between diet (69% of all isolates in WD and 52% in CD, chi2 test, p=0,1). 6 alleles were observed in *rbsD* and *rbsA* with 3 stop SNPs and one deletion with frameshift each and for the rest, one duplication and one synonymous mutation for *rbsD* and two

replicative transposition for *rbsA*. 4 and one indels with frameshift were observed respectively on gene *rbsC* and gene *rbsB*.

Mutations in genes encoding the metabolism of mannose were observed in our study. The mannose entering, via a mannose PTS permease whose sub units are encoded by the tree genes *manXYZ* (Erni et Zanolari 1985), then the bacteria undergoes phosphorylation into mannose-6-phosphate (M6P). M6P can follow the path of glycolysis after a transformation to fructose-6-phosphate by the mannose-6-phosphate isomerase enzyme encoded by *manA* (Sampaio, Santos, et Boos 2003). M6P can follow, via phosphomannomutase encoded by *cpsG* and Mannose-1-phosphateguanylyl transferase encoded by *cpsB*, the GDP-mannose synthesis pathway (Sampaio, Santos, et Boos 2003). GDP-mannose is a precursor for the biosynthesis of GDP-L-fucose, a nucleotide-activated form of L-fucose, a monosaccharide constituent of colanic acid, an extracellular polysaccharide found in Enterobacteriaceae. Mutations on *cpsG* or/and *cpsB* are found in 41 isolates, these mutations are theorically responsible for diversion of the mannose use from GDP-mannose synthesis pathway to glycolysis. These isolates are spread in only 3 lineages in WD and all 7 lineages in CD (22% of all isolates in WD and 80% in CD, chi2 test, $p<10^{-5}$). The great majority of mutations are indels with only one stop SNP mutation present in 2 isolates and IS transposition present in one isolate in gene *cpsB*. 15 isolates had a partial or total deletion of gene *cpsG* and *cpsB* integrated in larger deletions greater than 20,000 bp. Finally, 24 isolates among 7 lineages had an indel mutation with frameshift in a repeated 8 times nucleotide C region at position 2,153,455 on the gene *cpsG*. This convergence seems to favor the utilization of mannose as carbon source, especially since we found only 4 isolates in 4 different lineages, without mutations in genes *manA* or *manZ*.

Gluconate is one of most important energetic substrate of *E. coli* in the mice gut. Surprisingly, we found a great convergence in potential inactivation of gluconate transporter *gntT*. *gntT* encodes for a high-affinity gluconate transporter, one of four known transporters for gluconate in *E. coli* (Porco *et al.* 1997), induced at low levels of gluconate and partially repressed by glucose (Porco *et al.* 1997). Mutations on gene *gntT* were found in 53 isolates spread in all lineages whatever the diet, with relative higher number of isolates carrying these mutations in WD (91 % of all isolates in WD and 66% in CD, chi2 test, p=0.006). *gntT* had a high polymorphism with 21 alleles including 13 SNP with 8 non synonymous mutation and 5 stop SNP, 7 indels with 4 duplications, 2 short deletions of 1 and 14 bp and a large deletion of 12,822 bp (taking a part of *gntT* gene). We noticed a high presence of stop SNP and indels suggesting a high selection for *gntT* inactivation. Paradoxically, we observed 11 isolates

carrying mutations in gene *gntR* equally distributed in the two diets (11% of all isolates in WD and 16% in CD, chi2 test, p=0.53). *gntR* encodes for *gntT* repressor, so its mutations result in theory in a constitutive activation of *gntT*. This mutation seems to be fixed in only one lineage in CD (lineage 84) and two isolates evolved in WD share mutations in *gntR* and *gntT*. These two elements suggest a relative selective tolerance for the constitutive activation of gluconate metabolism in CD and less so in WD.

We found convergent mutations in *kdgT* gene encoding 2-keto-3-deoxy-D-gluconate (KDG) uptake system. The cloned *kdgT* gene increased uptake of KDG and to a lesser extent glucoronate (Mandrand-Berthelot, Ritzenthaler, et Mata-Gilsinger 1984). Mutations in gene *kdgT* were found in 38 isolates exclusively in CD in all lineages. The great majority of the isolates (29 isolates) carried indels at position 4,129,737 in a 7 times repeated G region. The rest of the mutations are replicative transposition (one isolate), non-synonymous SNP (2 isolates), a duplication (one isolate) and an insertion of a T nucleotide at position 4,130,364 (4 isolates) and 4,130,460 (one isolate). Interestingly, 19 of these CD isolates have mutations in both *kdgT* and *gntT*, transporters of the two substrates of the Entner-Doudoroff pathway (Peekhaus et Conway 1998). 5 isolates in lineages 84 and 87 share mutations in *kdgT* and *gntR* and one isolate (74BJ351) carryed mutations in *kdgR*, *kdgT* and *gntT*.

So, we observed a large predominance of mutations in gene involved in sugar metabolism. However, the signature of convergence in *E. coli* HS was not restricted to metabolism.

Others convergences in E. coli HS

Sugar metabolism was the main target of the selection but we identified others convergences (Figure 5 and 6).

CpxAR is a two-component system which functions to protect the inner membrane against stress; it is one of several envelope stress response systems by which the cell monitors the envelope for defects and damage and mediates a response to restore homeostasis (Vogt et Raivio 2012). We found mutations in the gene involved in sensor domain of the membrane associated histidine kinase *cpxA* (T. L. Raivio et Silhavy 1997) and in gene involved in the periplasmic accessory protein CpxP that acts as a negative feedback regulator for the pathway (Fleischer *et al.* 2007). Mutations in *cpxA* are all non-synonymous SNP concentrated in the first part of the gene corresponding to the sensor domain of the membrane associated histidine kinase that theoretically should make the strain insensitive to activating signal by preventing

interaction between cpxA and cpxP (T. L. Raivio et Silhavy 1997). The four alleles in *cpxP* involve a nonsense mutation occurring in the first half of the gene suggesting strongly an inactivation. There is no redundancy between these two mutations; no isolate carried these two mutations simultaneously in WD (Correlation test, R=-0.56, p=0.0004). Interestingly, all isolates carrying one of these mutations were exclusively found in WD lineages. The mutations invaded early on lineages 49, 57, 60 and 66, more lately lineage 63 and were found in only one isolate at day 352 in lineage 69. All these elements suggest an important selection for constitutive activation of the response stress CpxAR two-component system (Figure 7).

We found mutations in gene *cydA* and *cydB* encoding for the sub-unit of cytochrome bd-I oxidase involved as a major terminal oxidases in aerobic respiratory chain of *E. coli* under oxygen limited conditions (Cotter *et al.* 1990). There are two others enzymes which act as a terminal oxidases in aerobic respiratory: (i) cytochrome oxidase encoded by *CyoABCDE* appears to be produced only under oxygen rich growth conditions (Cotter *et al.* 1990), (ii) while the role of cytochrome bd-II oxidase encoded by *AppCD* remain unclear (Bekker *et al.* 2009). We observed no mutation in *CyoABCDE* or *AppCD* in our study. We recovered 10 alleles resulting presumably in the inactivation of cydA, cydB or both. An IS transposition in the intergenic region between *EcHS_A0784* and *cydA* at position 798,113, was found in 35 isolates. It targets a rich and complex regulators binding sites region where Cra, Fnr and ArcA interact to modulate the expression of *cyd* operon (Ramseier, Chien, et Saier 1996). This regulation results in the expression of the operon in high level oxygen condition and strict anaerobiosis or activation in limited oxygen condition (Cotter *et al.* 1997). Interestingly, we found deletion of all the regulation region in one isolate 60BJ202 that theoretically inactivates the expression of *cyd* operon that required functional induction (Cotter *et al.* 1997). We observed 5 alleles in *cydA* gene, 3 non-synonymous SNP present in 2, 1 and 5 isolates, 1 stop SNP early invading lineage 79 and one pair base deletion inducing a frameshift in isolate 60AJ202. We also observed 2 alleles in *cydB* involving a non-synonymous mutation each in one isolate. We found a complete loss of both genes *cydA* and *cydB* due to a lager deletion. This massive convergence towards a nonfunctional operon suggest an adaptation to the level of oxygen found in the mouse gut.

We found exclusively in WD mutations in gene involved in cysteine detoxication pathway notably in gene *ybaO* and *yhaM* and. *ybaO* is DNA-binding transcriptional regulator capable to induce genes involved in cysteine detoxication (*yhaO* and *yhaM*), in aromatic amino acid metabolism (two genes of the *tyrA-aroF* operon, *aroP*, and *aroL)*; and several genes for proteins involved in carbohydrate metabolism (*lacI*, two genes of the *srlA* operon, and all the

genes of the *uidABC* and *dhaKLM* operons). The transcriptional regulator *yhaO* is capable of repressing genes involved in degradation of glucarate and galacturate (*gudP* and the adjacent *garPLRK* and *garD* operons) and in the maltose degradation (Loddeke *et al.* 2017). We first observed mutations occurring early in gene *yhaM* in 17 isolates invading early tree lineages (49, 57 and 60) and only present in one isolate at day 84 in lineage 63. Mutations occurring in *yhaM* gene were mainly in phase indels for tree alleles, the others were a non-synonymous SNP and an IS transposition. There were diverse mutations on *ybaO* transcriptional regulator, with 6 alleles (mainly non-synonymous SNP), fixed lately at day 352 in lineage 49, 57 and 60 that occurred in the same isolate in which genes *yhaM* had been mutated in WD (Correlation test, R=0.41, p=0.01).

tdcA is a gene encoding for a transcriptionnal activator of *tdc* operon involved in transport and metabolism of threonine and serine during anaerobiosis (Ganduri *et al.* 1993). We found 6 alleles, differing exclusively by non-synonymous mutations, in 11 isolates spread in 4 lineages (lineages 49, 63, 66 and 69) exclusively in WD. We also noticed mutations in another transcriptional activator *ytfA* exclusively in WD whose function remain unclear. 4 alleles of that gene were observed including a frameshift and two nonsense mutations suggesting a strong selection for inactivation (Burke *et al.* 2013).

We finally observed mutations in two genes potentially involved in antibiotic resistance with two different mechanisms, impermeability and multidrug efflux. There are mutations in *mprA* transcriptional repressor that negatively regulates the transcription of genes that code for multidrug resistance pumps that extrude structurally unrelated antimicrobial agents from the cell (Lomovskaya, Lewis, et Matin 1995). We observed 7 alleles with two larges deletions, one 5 pair base deletion, tree IS transposition and a non-synonymous SNP in 9 isolates spread in 5 lineages exclusively in CD. There were maybe very few mutations but almost concentrated in end time point. These indels and large deletions suggest a strong selection for inactivation of this gene and then a constitutive activation of multidrug resistance pumps. We also found mutations in *sbmA* encoding for a transport protein responsible for microcin uptake (Laviña, Pugsley, et Moreno 1986). Its mutation could confer theoretically a resistance to microcin. Gene *sbmA* mutations have produced 5 alleles distributed in 6 isolates and in 3 lineages exclusively in WD (lineages 57, 63 and 66).

DISCUSSION

In this work, we observed that two different natural isolates of *E. coli* facing similar environments showed very contrasted adaptation, both in mechanistic strategies of adaptation and in the targeted genes an functions.

Diversity of E. coli in mice gut treated with streptomycin

Much more mutations were recovered *in E. coli* HS than in *E. coli* 536 with a comparable density in feces through time and paradoxically a lower basal mutation rate of $4.6x10^{-7}$ (CI95%: $2.95x10^{-7}$ – $7.5x10^{-7}$), ($6.23x10^{-7}$ (CI95%: $3.39x10^{-7}$ – $1.08x10^{-6}$) in WD and $1.53x10^{-7}$ (CI95%: ? - ?) in CD) in comparison with a mutation rate of $1.22x10^{-6}$ (CI95%: $8.52x10^{-7}$ – $1.68x10^{-6}$) in *E. coli* 536 in WD. The mutation rate of *E. coli* 536 was therefore estimated to be almost twice higher than the basal global mutation rate of *E .coli* HS and concordant with the mutation rate of $1.25x10^{-6}$ found for *E. coli* 536 evolving in streptomycin treated mice fed with CD (Lescat *et al.* 2017). However this mutation rate is almost 2 times higher than the one found in the same strain evolving in mice the mice gut without streptomycin $5x10^{-7}$ (Ghalayini *et al.* 2019) or in another phylogenetic group B2 *E. coli* ED1a evolving in healthy human gut at 6.9×10^{-7}. The large number of mutations found in of *E. coli* HS was not due its basal mutation rate, but due on the one hand to the occurrence of hypermutator isolates invading 3 lineages in CD and one in WD, and on the other hand to a higher rate of IS transposition, of indels, of large deletions and of nonsynonymous mutations.

The computation of Ka/Ks revealed a weak but significant excess of nonsynonymous mutations (Ka/Ks =2.02) in *E. coli* 536 in streptomycin treated mice gut for more than one year in WD notably associated with a convergence in *rluD* and *gidB*. This convergence had already been detected in an experiment in the evolution of *E. coli* 536 in the streptomycin treated mice (Lescat *et al.* 2017) but not in mice gut reproducing some more natural conditions (Ghalayini *et al.* 2019). Although *E. coli* 536 is naturally resistant to streptomycin by a G to A mutation mutation at the 3,586,077 position, the presence of a mutation on *gidB* or *rluD* appears to be associated with a selective advantage in the presence of streptomycin (Lescat *et al.* 2017). *E. coli* 536 presents a small signal for convergent mutations on lactose and maltose operon, which we discuss about below. *E. coli* HS, made resistant by integration of a transposon carrying the point mutation of the *rpsL* gene, had mutations on *gidB* in both diets suggesting a mutation quite specific to the presence of streptomycin treatment and not to the strain or the diet.

E. coli HS underwent intense adaptation within the streptomycin treated mice gut (Figure) with a Ka / Ks of 8.15 in WD and 15.55 in CD. These extreme values suggest that more than 90%

of the observed nonsynonymous mutations result from the action of natural selection. Several genetic mechanisms contributed to this adaptation. First, we osberved the appearance of hypermutator strains that increase the mutation rate and result in a faster adaptation of the population. This phenotypic trait was in many experimental evolution in the laboratroy (Tenaillon *et al.* 2016) and also found in nature in bacteria such as *Burkholderia dolosa* (Lieberman *et al.* 2014), *Pseudomonas aeruginosa* (Marvig *et al.* 2013) and also in *E. coli* (Levert *et al.* 2010) during acute infection. Theory and simulations have suggested that when facing a challenging environment, hypermutators may be seleted for thanks to the beneficial mutation they generate. Second, the high proportion of indels (duplication / insertion / deletion) occurring often on short repetitive seqeunces that are error prone for the DNA polymerase (Maki 2002; Chattopadhyay *et al.* 2009). These regions are located, on coding regions but also in an intergenic region likely to touch a regulatory zone or a promoter upstream of a gene. Third, an intense transposition activity in comparison to *E. coli* 536 was observed and involved the inactivation of genes of interest. Fourth, large deletions responsible of gene losses, including genes involved in convergence were also osberved. These genetic mechanisms contributed to the adaptation of HS strain to the gut.

Sugar metabolism: a predominant target of E. coli adaptation in streptomycin treated mice

Sugar are the main energetic substrate in bacteria and depend to tree central degradation pathway to product energy: the glycolysis (Embden-Meyerhof-Parnas pathway), the Entner-Doudoroff pathway (ED) and the pentose phosphate pathway (Peekhaus et Conway 1998). The ED and pentose phosphate pathway products are intermediate substrate joining the second part of the glycolysis (Wood 1986; Conway 1992). These three pathways can be followed by glucose to be degraded into pyruvate, a substrate for aerobic or anaerobic energy production pathways.

Sugars follow one of the three metabolic pathways depending on the structural proximity of their intermediate substrates. Gluconate and 2-keto-3-deoxy-gluconate are integrated in the ED (Peekhaus et Conway 1998) and ribose in pentose phosphate pathway (Diether *et al.* 2019). We found no mutations in genes involved in ED, pentose phosphate or glycolysis pathway. The integrity of ED pathway is crucial in colonization of streptomycin treated mice (Sweeney, Laux, et Cohen 1996). We noticed that mutations targeted essentially genes involved in Gluctonate and 2-keto-3-deoxy-gluconate (2KDG) transporter *gntT* and *kdgT* with a probable lower

possibility to obtain these substrates from outside the cell (Marvig *et al.* 2013; Tong *et al.* 1996). Gluconate and 2KDG can be some intermediate substrates of some sugar degradation pathway before they enter in the ED pathway (Peekhaus et Conway 1998). L-iodonate can be transported in bacteria by transporter encoding by *idnT* and successively degraded by enzymes coded by *idnD, idnO* and *idnK* into D-gluconate before entering the ED pathway. On the other side, D-glucuronate and D-Glucuronate can be transported in bacteria by transporters encoding by *kdgT* and especially *exuT* and then be successively degraded by enzymes coded by *uxaC* in common, *uxaB* and *uxuB* respectively into 2KDG before entering the ED pathway (Peekhaus et Conway 1998). Since the ED degradation pathway is essential for the colonization of streptomycin treated mice gut, the convergent mutations theoretically inactivating its two main substrates suggest the use of other substrates whose transporters and metabolism remained intact.

Interestingly, we observed differences between the two diets relative to Gluconate or/and 2KDG transporter inactivation. In WD, we observed mutations in *gntT* gene, which should decrease gluconate extraction from the gut and not in *kdgT*. This suggests a diversion to others sugar susceptible of producing gluconate such as L-iodonate or to the other substrate of the ED pathway, 2KDG, whose genes, involved in its transport and metabolism, remained intact in WD. We noticed very few mutations in *exuT* which theoretically inactivate the transport of D-galacturonate and to a lesser extend D-glucuronate, confirming the hypothesis that *E. coli* HS uses 2KDG in WD. In CD, we found a high prevalence of mutations in both *gntT* and *kdgT*, the later coding for a protein involved in 2KDG transport and to a lesser extent in D-Glucuronate transport. Moreover, a little less than a half of *E. coli* HS isolates in CD carried mutations in both genes suggesting a complete deviation to sugars whose degradation products include gluconate or 2KDG such as L-idonate, D-galacturonate, D-fructuronate or D-tugaturonate.

Surprisingly, we found mutations in the repressor *gntR* in the lineage evolved in the two diets. This gene represses genes involved in transport and metabolism of gluconate including *gntT* and *kdgT*. It therefore seems paradoxical to alleviate the repression of genes selected to be inactivated. But *gntR* is also involved in repression of *idn* operon (Rodionov *et al.* 2000). This suggests that: (i) first, in presence of gluconate, the bacteria favors utilization of outside gluconate over L-iodonate and, (ii) second, even with *gntT* defect, *gntR* defect allows the constitutive usage of L-idonate. Mutations in *gntT* and *kdgT can* essentially suggest a lack of gluconate and/or 2KDG in gut environment forcing bacteria to redirect its sugar supply to other sugars having gluconate or 2KDG as byproduct that feed the ED pathway. To our knowledge, no other sugar whose one the degradation intermediary product can enter in ED pathway have

been tested for colonization of streptomycin treated mice. The question of key sugars for colonization of gut in WD and CD remain to be answer.

We found mutations in genes involved in ribose metabolism, in its transport with *rbs* complex and its degradation pathway with *rbsD* that presumably prevents the use of ribose. These convergent mutations were found in almost all lineage except 90 in the two diets and fixed in 9 lineages *in E. coli* HS with high polymorphism of 16 alleles spread in gene *rbsD*, *rbsA*, *rbsC* and *rbsB*. Ribose is one of the sugar that needs to pass through pentose phosphate pathway after its degradation. Inactivation of the uptake and degradation of ribose is not that surprising as the pentose phosphate pathway does not seem to be a indispensable for *E. coli* colonization in the streptomycin treated mice gut and ribose appears to be the least preferred substrate for the growth of *E. coli* in mucus (Chang *et al.* 2004).

D-mannose can be a precursor for colonic acid biosynthesis or an energetic substrate. Mannose can be degradated in coloic acid precursor through pathway using enzyme coding by *cpsB* and *cpsG* or be degradeted into D-frutctose-6-phosphate by an enzyme encoding by *manA* before entering glycolysis. Mutations, predominantly found in CD, lead to an inactivation of the colanic acid biosynthesis pathway and a diversion of mannose towards the energy pathway. These convergent mutations invaded all lineages in CD. The inactivation of the colanic acid pathway could also be an adaptation as colanic acid capsule in *E. coli* is made primarily at low temperatures, suggesting a role when *E. coli* is outside mammalian hosts (Navasa *et al.* 2013). Accordingly, the colanic acid capsule does not have a role in pathogenesis (Russo *et al.* 1995), and high-level colanic acid expression appears to interfere with adhesion of uropathogenic *E. coli* (Hanna *et al.* 2003).

We observed in our study an important convergence for the inactivation of the maltose operon in both diets and more unusual in our data set for both strains. Most mutations affected the transcriptional activator *malT*, one of its intergenic DNA binding site and genes *malF* and *malE*, two part of maltoseABC transporter. Maltose metabolism is the only sugar metabolism that we found affected in both strains and diets. An inactivation of maltose was previously found in *E. coli* strains evolving in axenic mice gut (De Paepe *et al.* 2011), or in more natural conditions lacking streptomycin (Ghalayini *et al.* 2019). But until now, very few mutations in maltose operon were found in streptomycin treated mice (Lescat *et al.* 2017; Barroso-Batista *et al.* 2014). This convergence for inactivation of maltose operon in our study and others suggests a common advantage for adaptation in mice gut. Yet, this observation is in opposition with the finding of Jones *et al.* that showed that catabolism of the disaccharide maltose provided a

competitive advantage *in vivo* to pathogenic *E. coli* O157:H7 and commensal *E. coli* K-12 in streptomycin treated mice (Jones *et al.* 2008). The suggestion of a benefit of maltose metabolism limited to early colonization (Chang *et al.* 2004) may help reconcile the two observations.

In addition to mutations in *malT*, we observed few convergent mutations leading to constitutive activation of lactose use in *E. coli* 536. These mutations were found in only 4 lineages with no fixation and a low polymorphism of 4 alleles. We already observed this phenotypic trait in *E. coli* 536 evolving during a breast-feeding period probably due to high amount of lactose in the media (Ghalayini *et al.* 2019). We did not observe these mutation in *E. coli* HS in our study or in *E. coli* K12 in previous studies (De Paepe *et al.* 2011; Barroso-Batista *et al.* 2014).

Instead of *E. coli* 536 have a weak adaptation pattern confirming the result of former study studying its evolution in more natural condition (Ghalayini *et al.* 2019), *E. coli* HS allowed us to determine a strategy for adaptation to the use of sugar: (i) global inhibition or activation of sugar utilization operons and (ii) deviation of substrate in the pathway. Surprisingly, except for maltose and a little for lactose degradation metabolism, *E. coli* 536 does not adapt that much while *E. coli* HS reorganize completely its sugar supply metabolic pathways.

Others convergences

We detected several convergent mutations that presumably activates constitutively with a probable gain of function (T. L. Raivio et Silhavy 1997) the cpx two component system implied in membrane stress response (Vogt et Raivio 2012). Interestingly, mutations occurred both in the gene of effector of the inhibition *cpxP* and its target in the first part of *cpxA*. This system allows the activation or repression of several genes involved in the homeostasis of the stress-related cell membrane (Vogt et Raivio 2012). However, we have detected no significant mutation in the known target genes (Tracy L. Raivio, Leblanc, et Price 2013) except the *sbmA* gene involved in the expression of a putative transporter whose specific role is unknown but whose inactivation confers resistance to the microcin (Krizsan, Knappe, et Hoffmann 2015). The constitutive activity of this system with a probable gain of function and the relative integrity of the downstream metabolic pathways allowing the repression or the activation of the genes responsible for the maintenance of the bacterial membrane homeostasis suggest the presence of permanent disruption of the membrane of *E. coli* HS in streptomycin-treated mice gut in WD

diet and absent in CD. A broad range of stimuli can activate Cpx two component system such pH, osmolality copper, zinc, indole, lipid composisiton of the inner membrane and surface adhesion (Hunke, Keller, et Müller 2012). We need further investigation to determine the element in streptomycin treated mice needing a constitutive activation of cpx system.

We observed a mutational convergence leading to theoretical inactivation in the genes involved in the metabolism of three amino acids, cysteine (*ybaO* and *yhaM*), threonine and serine (*tdcA*). The presence of cysteine improves the interaction between ybaO and the promoter region of the operon *yhaM*, thereby causing activation (Shimada, Tanaka, et Ishihama 2016). Interestingly, it is the anaerobic metabolism of the uptake of these amino acids that is mutated (Loddeke *et al.* 2017; Ganduri *et al.* 1993). This observation joins another convergence in the inactivation of the *cydA* and *cydB* genes coding for the two subunits of the cytochrome bd-I responsible for the production of energy in an oxygen-poor medium, whereas the genes coding for the activated cytochrome bo for a high level of oxygen remain intact. This suggests an adaptation to a high level of oxygen in the streptomycin treated mice in WD consistent with the difference in digestive microbiota tending towards a richer composition of aero-tolerant bacterium belonging to the phylum Firmicutes and a decrease in anaerobic bacteria belonging to the phylum Bacteroidetes in WD (Murphy, Velazquez, et Herbert 2015).

We observed two other mutations that theorically confer a resistance to antibiotic, especially to microcin, in the two diet with different mechanism. Mutations on gene *mprA* confer resistance by efflux in CD (Lomovskaya, Lewis, et Matin 1995) and *sbmA* confer resistance by impermeability (Laviña, Pugsley, et Moreno 1986). We finally found interesting mutations in the intergenic region between *yjiH* and *kptA*. these convergent mutations were also found in a previous *E. coli* experimental evolution in streptomycin treated mice fed with CD (Lescat *et al.* 2017), so that make the only convergence specific to CD regardless of strain. Unfortunaly, the specific function of gene *yjiH* and *kptA* and of their intergenic region remain unclear.

CONCLUSION

This very preliminary analyses of natural isolates evolving in streptomycin treated mice gut during more than year allow us to distinguish some interesting results. First, adaptation profile could be very different between two natural *E. coli* isolates. They are genetically

different one belonging to phylogenetic group B2 and the other belonging to phylogenetic group A. This past history could have an influence on the intensity of adaptation strategy on the same media. First, *E. coli* belonging to phylogenetic group B2 are known to carry genes encoding for virulent factors that are involved in commensal colonization (Diard *et al.* 2010) while *E. coli* belonging to phylogenetic group A are not. Second, adaptation seems to depend on the combination of strain and diet in streptomycin treated mice targeting in priority genes involved in sugar metabolism.

ACKNOWLEDGMENTS

This work was supported in part by grants from La Fondation pour la Recherche Médicale to Mohamed Ghalayini (thèse médico-scientifique, grant number FDM20150633803) and from European Research Council to Olivier Tenaillon under the European Union's Seventh Framework Program (FP7/2007–2013)/ERC Grant 310944. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We gratefully acknowledge Hervé Le Nagard for his valuable and indispensable help with the CATIBioMed cluster.

REFERENCES

Barroso-Batista, João, Ana Sousa, Marta Lourenço, Marie-Louise Bergman, Daniel Sobral, Jocelyne Demengeot, Karina B. Xavier, et Isabel Gordo. 2014. « The First Steps of Adaptation of Escherichia Coli to the Gut Are Dominated by Soft Sweeps ». *PLoS Genetics* 10 (3): e1004182. https://doi.org/10.1371/journal.pgen.1004182.

Bekker, M., S. de Vries, A. Ter Beek, K. J. Hellingwerf, et M. J. Teixeira de Mattos. 2009. « Respiration of Escherichia Coli Can Be Fully Uncoupled via the Nonelectrogenic Terminal Cytochrome Bd-II Oxidase ». *Journal of Bacteriology* 191 (17): 5510‑17. https://doi.org/10.1128/JB.00562-09.

Berger, H., J. Hacker, A. Juarez, C. Hughes, et W. Goebel. 1982. « Cloning of the Chromosomal Determinants Encoding Hemolysin Production and Mannose-Resistant Hemagglutination in Escherichia Coli ». *Journal of Bacteriology* 152 (3): 1241‑47.

Boos, W., et H. Shuman. 1998. « Maltose/Maltodextrin System of Escherichia Coli: Transport, Metabolism, and Regulation ». *Microbiology and Molecular Biology Reviews: MMBR* 62 (1): 204‑29.

Burke, Catherine, Michael Liu, Warwick Britton, James A. Triccas, Torsten Thomas, Adrian L. Smith, Steven Allen, Robert Salomon, et Elizabeth Harry. 2013. « Harnessing Single Cell Sorting to Identify Cell Division Genes and Regulators in Bacteria ». *PloS One* 8 (4): e60964. https://doi.org/10.1371/journal.pone.0060964.

Chang, Dong-Eun, Darren J. Smalley, Don L. Tucker, Mary P. Leatham, Wendy E. Norris, Sarah J. Stevenson, April B. Anderson, et al. 2004. « Carbon Nutrition of Escherichia Coli in the Mouse

Intestine ». *Proceedings of the National Academy of Sciences of the United States of America* 101 (19): 7427‑32. https://doi.org/10.1073/pnas.0307888101.

Chattopadhyay, Sujay, Scott J. Weissman, Vladimir N. Minin, Thomas A. Russo, Daniel E. Dykhuizen, et Evgeni V. Sokurenko. 2009. « High Frequency of Hotspot Mutations in Core Genes of Escherichia Coli Due to Short-Term Positive Selection ». *Proceedings of the National Academy of Sciences of the United States of America* 106 (30): 12412 ‑ 17. https://doi.org/10.1073/pnas.0906217106.

Clermont, Olivier, Julia K. Christenson, Erick Denamur, et David M. Gordon. 2013. « The Clermont Escherichia Coli Phylo-Typing Method Revisited: Improvement of Specificity and Detection of New Phylo-Groups ». *Environmental Microbiology Reports* 5 (1): 58 ‑ 65. https://doi.org/10.1111/1758-2229.12019.

Conway, T. 1992. « The Entner-Doudoroff Pathway: History, Physiology and Molecular Biology ». *FEMS Microbiology Reviews* 9 (1): 1‑27. https://doi.org/10.1111/j.1574-6968.1992.tb05822.x.

Cotter, P. A., V. Chepuri, R. B. Gennis, et R. P. Gunsalus. 1990. « Cytochrome o (CyoABCDE) and d (CydAB) Oxidase Gene Expression in Escherichia Coli Is Regulated by Oxygen, PH, and the Fnr Gene Product ». *Journal of Bacteriology* 172 (11): 6333 ‑ 38. https://doi.org/10.1128/jb.172.11.6333-6338.1990.

Cotter, P. A., S. B. Melville, J. A. Albrecht, et R. P. Gunsalus. 1997. « Aerobic Regulation of Cytochrome d Oxidase (CydAB) Operon Expression in Escherichia Coli: Roles of Fnr and ArcA in Repression and Activation ». *Molecular Microbiology* 25 (3): 605‑15. https://doi.org/10.1046/j.1365-2958.1997.5031860.x.

Danot, O., et O. Raibaud. 1994. « Multiple Protein-DNA and Protein-Protein Interactions Are Involved in Transcriptional Activation by MalT ». *Molecular Microbiology* 14 (2): 335 ‑ 46. https://doi.org/10.1111/j.1365-2958.1994.tb01294.x.

Datsenko, K. A., et B. L. Wanner. 2000. « One-Step Inactivation of Chromosomal Genes in Escherichia Coli K-12 Using PCR Products ». *Proceedings of the National Academy of Sciences of the United States of America* 97 (12): 6640‑45. https://doi.org/10.1073/pnas.120163297.

De Paepe, Marianne, Valérie Gaboriau-Routhiau, Dominique Rainteau, Sabine Rakotobe, François Taddei, et Nadine Cerf-Bensussan. 2011. « Trade-off between Bile Resistance and Nutritional Competence Drives Escherichia Coli Diversification in the Mouse Gut ». *PLoS Genetics* 7 (6): e1002107. https://doi.org/10.1371/journal.pgen.1002107.

Diard, Médéric, Louis Garry, Marjorie Selva, Thomas Mosser, Erick Denamur, et Ivan Matic. 2010. « Pathogenicity-Associated Islands in Extraintestinal Pathogenic Escherichia Coli Are Fitness Elements Involved in Intestinal Colonization ». *Journal of Bacteriology* 192 (19): 4885‑93. https://doi.org/10.1128/JB.00804-10.

Diether, Maren, Yaroslav Nikolaev, Frédéric Ht Allain, et Uwe Sauer. 2019. « Systematic Mapping of Protein-Metabolite Interactions in Central Metabolism of Escherichia Coli ». *Molecular Systems Biology* 15 (8): e9008. https://doi.org/10.15252/msb.20199008.

Ehrle, R., C. Pick, R. Ulrich, E. Hofmann, et M. Ehrmann. 1996. « Characterization of Transmembrane Domains 6, 7, and 8 of MalF by Mutational Analysis ». *Journal of Bacteriology* 178 (8): 2255‑62. https://doi.org/10.1128/jb.178.8.2255-2262.1996.

Erni, B., et B. Zanolari. 1985. « The Mannose-Permease of the Bacterial Phosphotransferase System. Gene Cloning and Purification of the Enzyme IIMan/IIIMan Complex of Escherichia Coli ». *The Journal of Biological Chemistry* 260 (29): 15495‑503.

Fleischer, Rebecca, Ralf Heermann, Kirsten Jung, et Sabine Hunke. 2007. « Purification, Reconstitution, and Characterization of the CpxRAP Envelope Stress System of Escherichia Coli ». *The Journal of Biological Chemistry* 282 (12): 8583‑93. https://doi.org/10.1074/jbc.M605785200.

Ganduri, Y. L., S. R. Sadda, M. W. Datta, R. K. Jambukeswaran, et P. Datta. 1993. « TdcA, a Transcriptional Activator of the TdcABC Operon of Escherichia Coli, Is a Member of the LysR Family of Proteins ». *Molecular & General Genetics: MGG* 240 (3): 395 ‑ 402. https://doi.org/10.1007/bf00280391.

Ghalayini, Mohamed, Adrien Launay, Antoine Bridier-Nahmias, Olivier Clermont, Erick Denamur, Mathilde Lescat, et Olivier Tenaillon. 2018. « Evolution of a Dominant Natural Isolate of Escherichia Coli in the Human Gut over the Course of a Year Suggests a Neutral Evolution with Reduced Effective Population Size ». *Applied and Environmental Microbiology* 84 (6). https://doi.org/10.1128/AEM.02377-17.

Ghalayini, Mohamed, Melanie Magnan, Sara Dion, Ouassila Zatout, Lucie Bourguignon, Olivier Tenaillon, et Mathilde Lescat. 2019. « Long-Term Evolution of the Natural Isolate of Escherichia Coli 536 in the Mouse Gut Colonized after Maternal Transmission Reveals Convergence in the Constitutive Expression of the Lactose Operon ». *Molecular Ecology*, septembre. https://doi.org/10.1111/mec.15232.

Hanna, Andrea, Michael Berg, Valerie Stout, et Anneta Razatos. 2003. « Role of Capsular Colanic Acid in Adhesion of Uropathogenic Escherichia Coli ». *Applied and Environmental Microbiology* 69 (8): 4474‑81. https://doi.org/10.1128/aem.69.8.4474-4481.2003.

Hunke, Sabine, Rebecca Keller, et Volker S. Müller. 2012. « Signal Integration by the Cpx-Envelope Stress System ». *FEMS Microbiology Letters* 326 (1): 12‑22. https://doi.org/10.1111/j.1574-6968.2011.02436.x.

Jones, Shari A., Mathias Jorgensen, Fatema Z. Chowdhury, Rosalie Rodgers, James Hartline, Mary P. Leatham, Carsten Struve, Karen A. Krogfelt, Paul S. Cohen, et Tyrrell Conway. 2008. « Glycogen and Maltose Utilization by Escherichia Coli O157:H7 in the Mouse Intestine ». *Infection and Immunity* 76 (6): 2531‑40. https://doi.org/10.1128/IAI.00096-08.

Krizsan, Andor, Daniel Knappe, et Ralf Hoffmann. 2015. « Influence of the YjiL-MdtM Gene Cluster on the Antibacterial Activity of Proline-Rich Antimicrobial Peptides Overcoming Escherichia Coli Resistance Induced by the Missing SbmA Transporter System ». *Antimicrobial Agents and Chemotherapy* 59 (10): 5992‑98. https://doi.org/10.1128/AAC.01307-15.

Laviña, M., A. P. Pugsley, et F. Moreno. 1986. « Identification, Mapping, Cloning and Characterization of a Gene (SbmA) Required for Microcin B17 Action on Escherichia Coli K12 ». *Journal of General Microbiology* 132 (6): 1685‑93. https://doi.org/10.1099/00221287-132-6-1685.

Lescat, Mathilde, Adrien Launay, Mohamed Ghalayini, Mélanie Magnan, Jérémy Glodt, Coralie Pintard, Sara Dion, Erick Denamur, et Olivier Tenaillon. 2017. « Using Long-Term Experimental Evolution to Uncover the Patterns and Determinants of Molecular Evolution of an Escherichia Coli Natural Isolate in the Streptomycin-Treated Mouse Gut ». *Molecular Ecology* 26 (7): 1802‑17. https://doi.org/10.1111/mec.13851.

Levert, Maxime, Oana Zamfir, Olivier Clermont, Odile Bouvet, Sylvain Lespinats, Marie Claire Hipeaux, Catherine Branger, et al. 2010. « Molecular and Evolutionary Bases of Within-Patient Genotypic and Phenotypic Diversity in Escherichia Coli Extraintestinal Infections ». *PLoS Pathogens* 6 (9): e1001125. https://doi.org/10.1371/journal.ppat.1001125.

Levine, M. M., et M. B. Rennels. 1978. « E. Coli Colonisation Factor Antigen in Diarrhoea ». *Lancet (London, England)* 2 (8088): 534.

Li, Qinrui, Ying Han, Angel Belle C. Dy, et Randi J. Hagerman. 2017. « The Gut Microbiota and Autism Spectrum Disorders ». *Frontiers in Cellular Neuroscience* 11: 120. https://doi.org/10.3389/fncel.2017.00120.

Lieberman, Tami D., Kelly B. Flett, Idan Yelin, Thomas R. Martin, Alexander J. McAdam, Gregory P. Priebe, et Roy Kishony. 2014. « Genetic Variation of a Bacterial Pathogen within Individuals with Cystic Fibrosis Provides a Record of Selective Pressures ». *Nature Genetics* 46 (1): 82‑87. https://doi.org/10.1038/ng.2848.

Loddeke, Melissa, Barbara Schneider, Tamiko Oguri, Iti Mehta, Zhenyu Xuan, et Larry Reitzer. 2017. « Anaerobic Cysteine Degradation and Potential Metabolic Coordination in Salmonella Enterica and Escherichia Coli ». *Journal of Bacteriology* 199 (16). https://doi.org/10.1128/JB.00117-17.

Lomovskaya, O., K. Lewis, et A. Matin. 1995. « EmrR Is a Negative Regulator of the Escherichia Coli Multidrug Resistance Pump EmrAB ». *Journal of Bacteriology* 177 (9): 2328‑34. https://doi.org/10.1128/jb.177.9.2328-2334.1995.

Maki, Hisaji. 2002. « Origins of Spontaneous Mutations: Specificity and Directionality of Base-Substitution, Frameshift, and Sequence-Substitution Mutageneses ». *Annual Review of Genetics* 36: 279‑303. https://doi.org/10.1146/annurev.genet.36.042602.094806.

Maltby, Rosalie, Mary P. Leatham-Jensen, Terri Gibson, Paul S. Cohen, et Tyrrell Conway. 2013. « Nutritional Basis for Colonization Resistance by Human Commensal Escherichia Coli Strains HS and Nissle 1917 against E. Coli O157:H7 in the Mouse Intestine ». *PloS One* 8 (1): e53957. https://doi.org/10.1371/journal.pone.0053957.

Mandrand-Berthelot, M. A., P. Ritzenthaler, et M. Mata-Gilsinger. 1984. « Construction and Expression of Hybrid Plasmids Containing the Structural Gene of the Escherichia Coli K-12 3-Deoxy-2-Oxo-D-Gluconate Transport System ». *Journal of Bacteriology* 160 (2): 600‑606.

Marvig, Rasmus Lykke, Helle Krogh Johansen, Søren Molin, et Lars Jelsbak. 2013. « Genome Analysis of a Transmissible Lineage of Pseudomonas Aeruginosa Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators ». *PLoS Genetics* 9 (9): e1003741. https://doi.org/10.1371/journal.pgen.1003741.

Massot, Méril, Anne-Sophie Daubié, Olivier Clermont, Françoise Jauréguy, Camille Couffignal, Ghizlane Dahbi, Azucena Mora, 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, England)* 162 (4): 642‑50. https://doi.org/10.1099/mic.0.000242.

Merino, G., et H. A. Shuman. 1997. « Unliganded Maltose-Binding Protein Triggers Lactose Transport in an Escherichia Coli Mutant with an Alteration in the Maltose Transport System ». *Journal of Bacteriology* 179 (24): 7687‑94. https://doi.org/10.1128/jb.179.24.7687-7694.1997.

———. 1998. « Truncation of MalF Results in Lactose Transport via the Maltose Transport System of Escherichia Coli ». *The Journal of Biological Chemistry* 273 (4): 2435‑44. https://doi.org/10.1074/jbc.273.4.2435.

Messing, Simon A. J., Bao Ton-Hoang, Alison B. Hickman, Andrew J. McCubbin, Graham F. Peaslee, Rodolfo Ghirlando, Michael Chandler, et Fred Dyda. 2012. « The Processing of Repetitive Extragenic Palindromes: The Structure of a Repetitive Extragenic Palindrome Bound to Its Associated Nuclease ». *Nucleic Acids Research* 40 (19): 9964‑79. https://doi.org/10.1093/nar/gks741.

Murphy, E. Angela, Kandy T. Velazquez, et Kyle M. Herbert. 2015. « Influence of High-Fat Diet on Gut Microbiota: A Driving Force for Chronic Disease Risk ». *Current Opinion in Clinical Nutrition and Metabolic Care* 18 (5): 515‑20. https://doi.org/10.1097/MCO.0000000000000209.

Navasa, Nicolás, Leandro Rodríguez-Aparicio, Miguel Ángel Ferrero, Andrea Monteagudo-Mera, et Honorina Martínez-Blanco. 2013. « Polysialic and Colanic Acids Metabolism in Escherichia Coli K92 Is Regulated by RcsA and RcsB ». *Bioscience Reports* 33 (3). https://doi.org/10.1042/BSR20130018.

Peekhaus, N., et T. Conway. 1998. « What's for Dinner?: Entner-Doudoroff Metabolism in Escherichia Coli ». *Journal of Bacteriology* 180 (14): 3495‑3502.

Porco, A., N. Peekhaus, C. Bausch, S. Tong, T. Isturiz, et T. Conway. 1997. « Molecular Genetic Characterization of the Escherichia Coli GntT Gene of GntI, the Main System for Gluconate Metabolism ». *Journal of Bacteriology* 179 (5): 1584‑90. https://doi.org/10.1128/jb.179.5.1584-1590.1997.

« R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. » s. d.

Raivio, T. L., et T. J. Silhavy. 1997. « Transduction of Envelope Stress in Escherichia Coli by the Cpx Two-Component System ». *Journal of Bacteriology* 179 (24): 7724‑33. https://doi.org/10.1128/jb.179.24.7724-7733.1997.

Raivio, Tracy L., Shannon K. D. Leblanc, et Nancy L. Price. 2013. « The Escherichia Coli Cpx Envelope Stress Response Regulates Genes of Diverse Function That Impact Antibiotic Resistance and Membrane Integrity ». *Journal of Bacteriology* 195 (12): 2755 - 67. https://doi.org/10.1128/JB.00105-13.

Ramseier, T. M., S. Y. Chien, et M. H. Saier. 1996. « Cooperative Interaction between Cra and Fnr in the Regulation of the CydAB Operon of Escherichia Coli ». *Current Microbiology* 33 (4): 270‑74.

Rasko, David A., M. J. Rosovitz, Garry S. A. Myers, Emmanuel F. Mongodin, W. Florian Fricke, Pawel Gajer, Jonathan Crabtree, et al. 2008. « The Pangenome Structure of Escherichia Coli: Comparative Genomic Analysis of E. Coli Commensal and Pathogenic Isolates ». *Journal of Bacteriology* 190 (20): 6881‑93. https://doi.org/10.1128/JB.00619-08.

Rodionov, D. A., A. A. Mironov, A. B. Rakhmaninova, et M. S. Gelfand. 2000. « Transcriptional Regulation of Transport and Utilization Systems for Hexuronides, Hexuronates and Hexonates in Gamma Purple Bacteria ». *Molecular Microbiology* 38 (4): 673 - 83. https://doi.org/10.1046/j.1365-2958.2000.02115.x.

Russo, T. A., G. Sharma, J. Weiss, et C. Brown. 1995. « The Construction and Characterization of Colanic Acid Deficient Mutants in an Extraintestinal Isolate of Escherichia Coli (O4/K54/H5) ». *Microbial Pathogenesis* 18 (4): 269‑78.

Ryu, Kyoung-Seok, Changhoon Kim, Insook Kim, Seokho Yoo, Byong-Seok Choi, et Chankyu Park. 2004. « NMR Application Probes a Novel and Ubiquitous Family of Enzymes That Alter Monosaccharide Configuration ». *The Journal of Biological Chemistry* 279 (24): 25544‑48. https://doi.org/10.1074/jbc.M402016200.

Sampaio, Maria-Manuel, Helena Santos, et Winfried Boos. 2003. « Synthesis of GDP-Mannose and Mannosylglycerate from Labeled Mannose by Genetically Engineered Escherichia Coli without Loss of Specific Isotopic Enrichment ». *Applied and Environmental Microbiology* 69 (1): 233‑40. https://doi.org/10.1128/aem.69.1.233-240.2003.

Shimada, Tomohiro, Kan Tanaka, et Akira Ishihama. 2016. « Transcription Factor DecR (YbaO) Controls Detoxification of L-Cysteine in Escherichia Coli ». *Microbiology (Reading, England)* 162 (9): 1698‑1707. https://doi.org/10.1099/mic.0.000337.

Shuman, H. A., T. J. Silhavy, et J. R. Beckwith. 1980. « Labeling of Proteins with Beta-Galactosidase by Gene Fusion. Identification of a Cytoplasmic Membrane Component of the Escherichia Coli Maltose Transport System ». *The Journal of Biological Chemistry* 255 (1): 168‑74.

Stern, M. J., G. F. Ames, N. H. Smith, E. C. Robinson, et C. F. Higgins. 1984. « Repetitive Extragenic Palindromic Sequences: A Major Component of the Bacterial Genome ». *Cell* 37 (3): 1015‑26. https://doi.org/10.1016/0092-8674(84)90436-7.

Sweeney, N. J., D. C. Laux, et P. S. Cohen. 1996. « Escherichia Coli F-18 and E. Coli K-12 Eda Mutants Do Not Colonize the Streptomycin-Treated Mouse Large Intestine ». *Infection and Immunity* 64 (9): 3504‑11.

Tenaillon, Olivier, Jeffrey E. Barrick, Noah Ribeck, Daniel E. Deatherage, Jeffrey L. Blanchard, Aurko Dasgupta, Gabriel C. Wu, et al. 2016. « Tempo and Mode of Genome Evolution in a 50,000-Generation Experiment ». *Nature* 536 (7615): 165‑70. https://doi.org/10.1038/nature18959.

Tenaillon, Olivier, David Skurnik, Bertrand Picard, et Erick Denamur. 2010. « The Population Genetics of Commensal Escherichia Coli ». *Nature Reviews. Microbiology* 8 (3): 207 - 17. https://doi.org/10.1038/nrmicro2298.

Tong, S., A. Porco, T. Isturiz, et T. Conway. 1996. « Cloning and Molecular Genetic Characterization of the Escherichia Coli GntR, GntK, and GntU Genes of GntI, the Main System for Gluconate Metabolism ». *Journal of Bacteriology* 178 (11): 3260 - 69. https://doi.org/10.1128/jb.178.11.3260-3269.1996.

Touchon, Marie, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, et al. 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.

Vogt, Stefanie L., et Tracy L. Raivio. 2012. « Just Scratching the Surface: An Expanding View of the Cpx Envelope Stress Response ». *FEMS Microbiology Letters* 326 (1): 2‑11. https://doi.org/10.1111/j.1574-6968.2011.02406.x.

Wood, T. 1986. « Physiological Functions of the Pentose Phosphate Pathway ». *Cell Biochemistry and Function* 4 (4): 241‑47. https://doi.org/10.1002/cbf.290040403.

AUTHORS CONTRIBUTIONS

ML and OT designed research; MG, LB, AC, MM and SD performed research; MG and OT analyzed data; MG and OT wrote the paper.

FIGURES AND TABLES

Figure 1. Density of *E. coli* in mouse feces as a function of time

The mean log10 density for *E. coli* 536 in WD (red), *E. coli* HS in WD (blue) and *E. coli* HS in CD (green) are presented. Error bars represent standard deviation.

Figure 2. Principal component analyses of all isolates depending on their mutations with and without hypermutator strain in *E. coli* HS

We represented the isolates from our experiment evolving during more than one year distributed in WD group inoculated with *E. coli* 536 (48 isolates), in WD group inoculated with *E. coli* HS (34 isolates) and in CD group inoculated in *E. coli* HS (42 isolates). We added the database from a previous study (Lescat *et al.* 2017) that followed an evolution of *E. coli* 536 strain during one year in streptomycin treated mice fed with CD. Hypermutator isolates are surrounded by a purple circle in the left figure.

Figure 3. The genetic basis of adaptive mutations and convergences between lineages in *E. coli* 536 in WD

We characterized and located convergent mutations identified in *E. coli* 536. For simplicity, the genomes are represented linearly and vertically drawn. We represented genes involved in ribosome maturation and conferring growth advantage in streptomycine media (*gidA* and *rluD*), in constitutive utilization of lactose (*lacI* and *lacO*) and in maltose use inhibition (*malT*). The type of mutation were shown by different color: pink for synonymous mutation, black for stop

SNPs, red for non-synonymous mutation, yellow for duplication, green for deletion, violet for insertion.

Figure 4. The genetic basis of adaptive mutation on sugar or energetic substrate metabolism

We characterized and located convergent identified mutations on sugar or energetic substrate metabolism. We represented ribose (*rbsD, rbsA, rbsC* and *rbsB*), maltose (*malF*, *malF/malE, malE, malE/malK, malP* and *malT*), gluconate (*gntT* and *gntR*), 2-keto-3-deoxy-D-gluconate (*kdgT* and *kdgR*) and mannose (*cpsG, cpsB, manA* and *manZ*) metabolism genetic support. The type of mutation were shown by different color: pink for synonymous mutation, black for stop SNPs, red for non-synonymous mutation, yellow for duplication, green for deletion, violet for insertion and white for transposition. The short deletion were represented only with a triangle while large deletion were represented with a green band below the region deleted.

Figure 5. Mutations convergence on sugar metabolism between lineages

Identified mutations in the two different isolates (A and B) from the 13 lineages at each time point were represented along the *E. coli* chromosome. For simplicity, the genomes are represented linearly and vertically drawn. Lineages from WD group are in orange while CD group are in blue. We represented ribose (*rbsD, rbsA, rbsC* and *rbsB*), maltose (*malF, malF/malE, malE, malE/malK, malP* and *malT*), gluconate (*gntT* and *gntR*), 2-keto-3-deoxy-D-gluconate (*kdgT* and *kdgR*) and mannose (*cpsG, cpsB, manA* and *manZ).* The type of mutation were shown by different color: pink for synonymous mutation, black for stop SNPs, red for non-synonymous mutation, yellow for duplication, green for deletion, violet for insertion and white for transposition. The short deletion were represented only with a triangle while large deletion were represented with a green band below the region deleted.

Figure 6. The genetic basis of adaptive mutation on non-sugar metabolism in *E. coli* HS

We characterized and located convergent identified mutations on non-sugar metabolism. We represented mutations types in genes involved in hypermutator phenotype (*mutS, mutL* and *mutH*), in amino-acid metabolism in anaerobiosis condition (*ybaO, yhaM* and *tdcA*), in

energetic production in low oxygen condition (*cydA* and *cydB*), in antibiotic resistance (*mprA* and *sbmA*), in ribosome maturation and conferring growth advantage in streptomycine media (*gidB*) and gene or region whose function remain unclear (intergenic region between *yijH* and *kptA*, and *ytfA*). The type of mutation were shown by different color: pink for synonymous mutation, black for stop SNPs, red for non-synonymous mutation, yellow for duplication, green for deletion, violet for insertion and white for transposition. The short deletion were represented only with a triangle while large deletion were represented with a green band below the region deleted.

Figure 7. Mutations convergence on non-sugar metabolism between lineages

Identified mutations in the two different isolates (A and B) from the 13 lineages at each time point were represented along the *E. coli* chromosome. For simplicity, the genomes are represented linearly and vertically drawn. Lineages from WD group are in orange while CD group are in blue. We represented mutations types in genes involved in hypermutator phenotype (*mutS*, *mutL* and *mutH*), in amino-acid metabolism in anaerobiosis condition (*ybaO*, *yhaM* and *tdcA*), in energetic production in low oxygen condition (*cydA* and *cydB*), in antibiotic resistance (*mprA* and *sbmA*), in ribosome maturation and conferring growth advantage in streptomycine media (*gidB*) and gene or region whose function remain unclear (intergenic region between *yijH* and *kptA*, and *ytfA*). The type of mutation were shown by different color: pink for synonymous mutation, black for stop SNPs, red for non-synonymous mutation, yellow for duplication, green for deletion, violet for insertion and white for transposition.

Figure 8. Functional and genetic support of constitutive activation of cpx pathway in WD.

We represented a schematic structure and fonction of cpx pathway. We represented the mutations occuring on the sensor domain of *cpxA* , on the cpx inhibitor *cpxP* and compensation mutations on *sbmA*. The type of mutation were shown by different color: black for stop SNPs, red for non-synonymous mutation, yellow for duplication and green for deletion.

Table 1. Characteristics and distributions of hypermutators strains.

The name of the genes as well as the orientation (defined by the arrow), the position along the genome, and the flanking genes in case of an intergenic mutation are displayed, with the mutations, their type, the number of occurrences, isolates which carry the mutations and diet group. Stop codons are indicated with the symbol *.

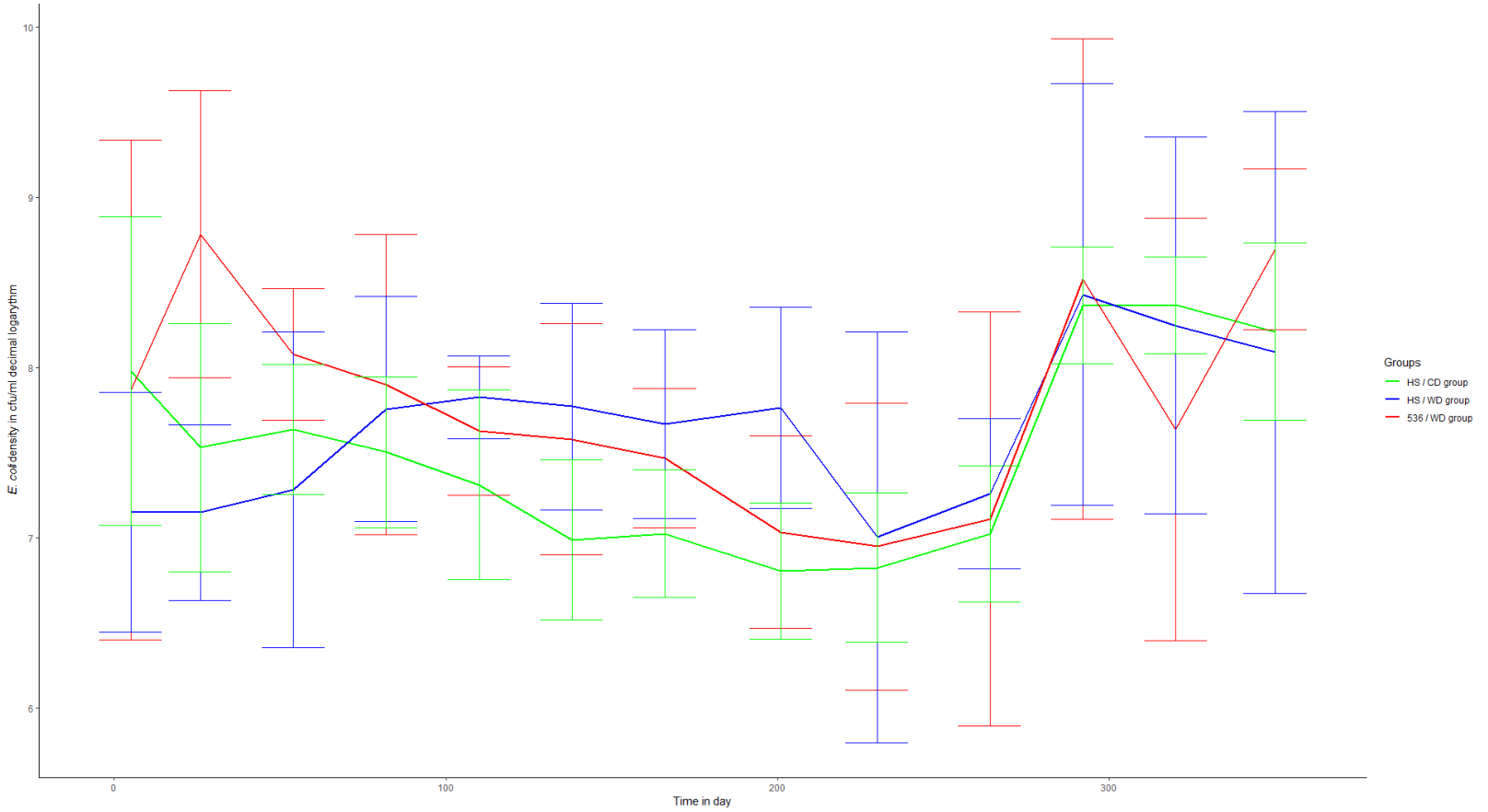
Table 2. Summary of large deletion (>1000 bp) that we found in evolved lineages

Table S1. Mice characteristics of the 126 sequenced strains

For each of the 126 isolates sequenced, we reported the sampling time, the sampled mice ID (lineage and mouse number), the diet, the relative size of the colonie (small, middle or big) and the *E. coli* density per gram of feces.

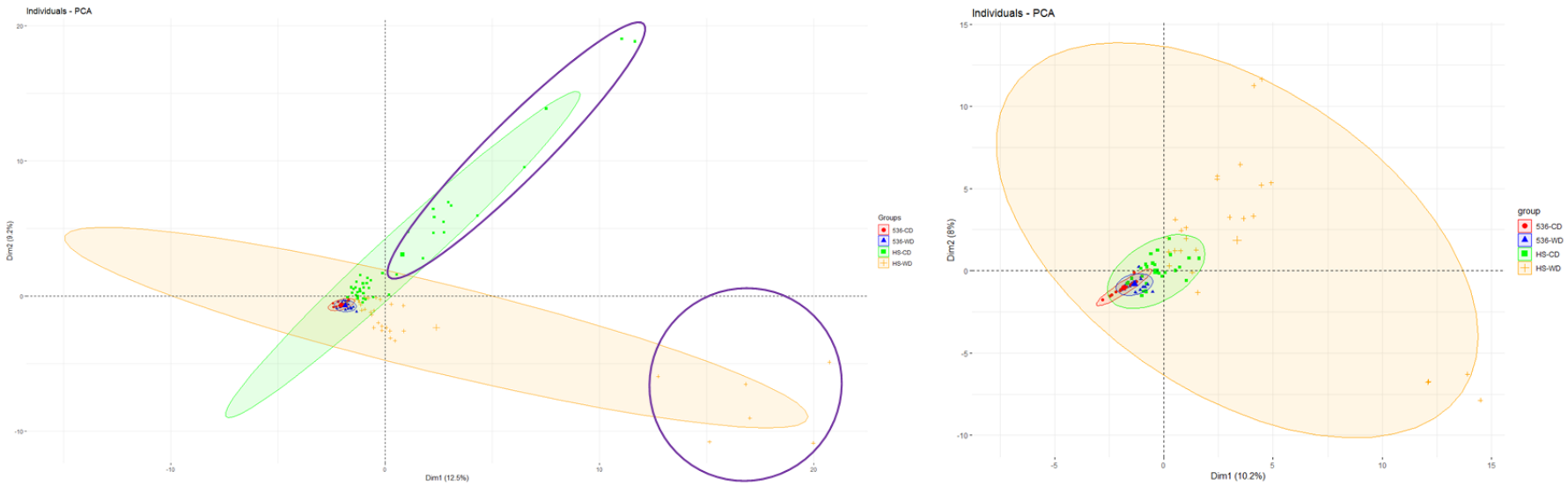
Table S2. List of the mutations that were found in the 48 sequenced isolates of *E. coli* 536

Table S3. List of the mutations that were found in the 78 sequenced isolates of *E. coli* HS



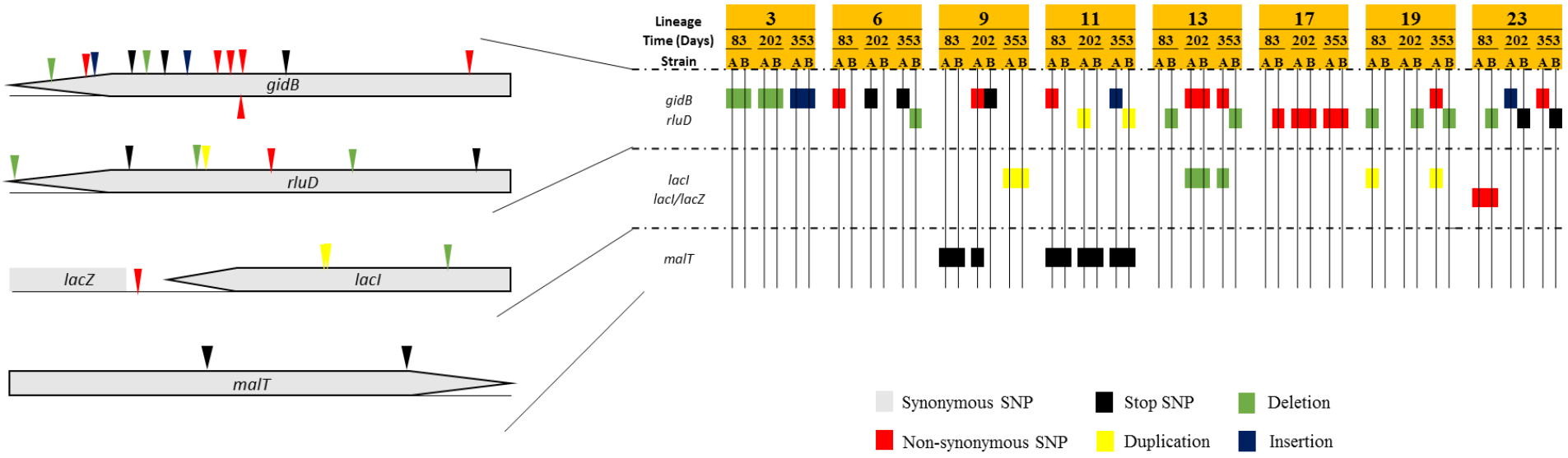
<u>Figure 1.</u> Density of *E. coli* in mouse feces as a function of time

The mean log10 density for *E. coli* 536 in WD (red), *E. coli* HS in WD (blue) and *E. coli* HS in CD (green) are presented. Error bars represent standard deviation.



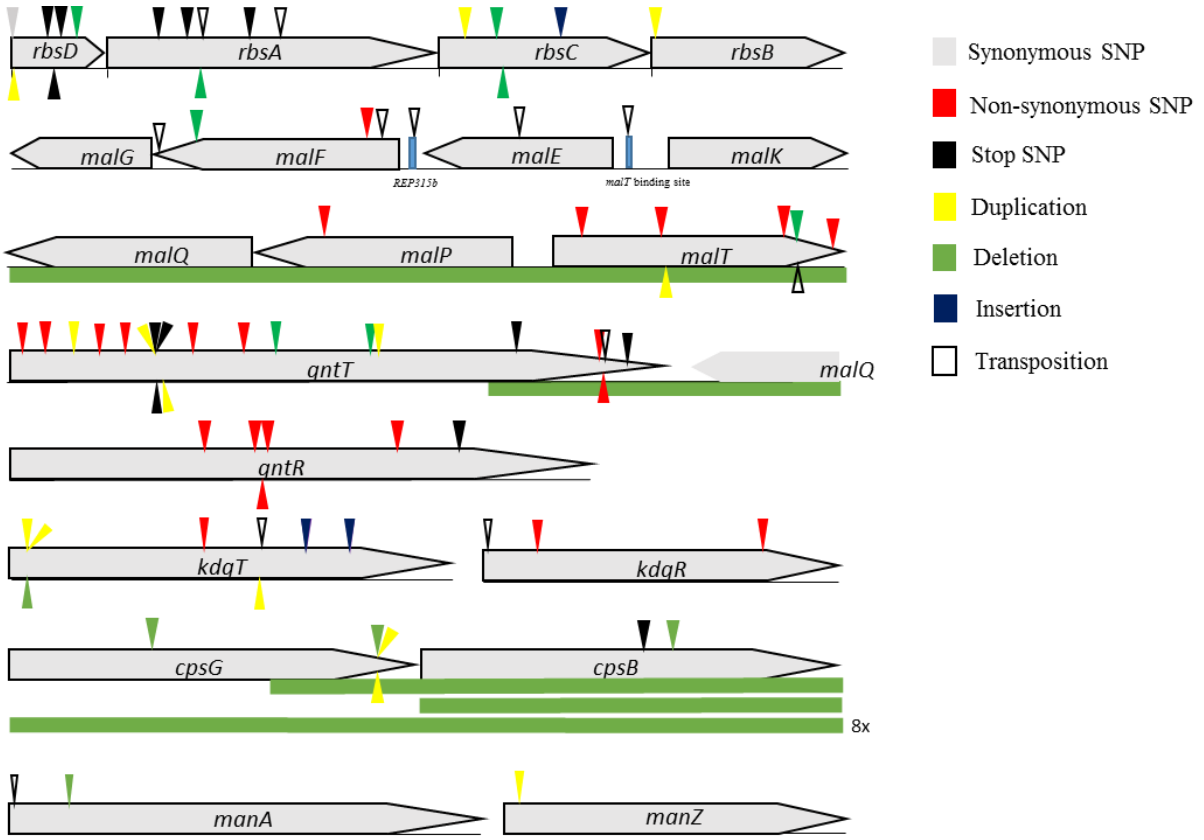
<u>Figure 2.</u> Principal component analyses of all isolates depending on their mutations with and without hypermutator strain in *E. coli* HS

We represented the isolates from our experiment evolving during more than one year distributed in WD group inoculated with *E. coli* 536 (48 isolates), in WD group inoculated with *E. coli* HS (34 isolates) and in CD group inoculated in *E. coli* HS (42 isolates). We added the database from a previous study (Lescat *et al.* 2017) that followed an evolution of *E. coli* 536 strain during one year in streptomycin treated mice fed with CD. Hypermutator isolates are surrounded by a purple circle in the left figure.



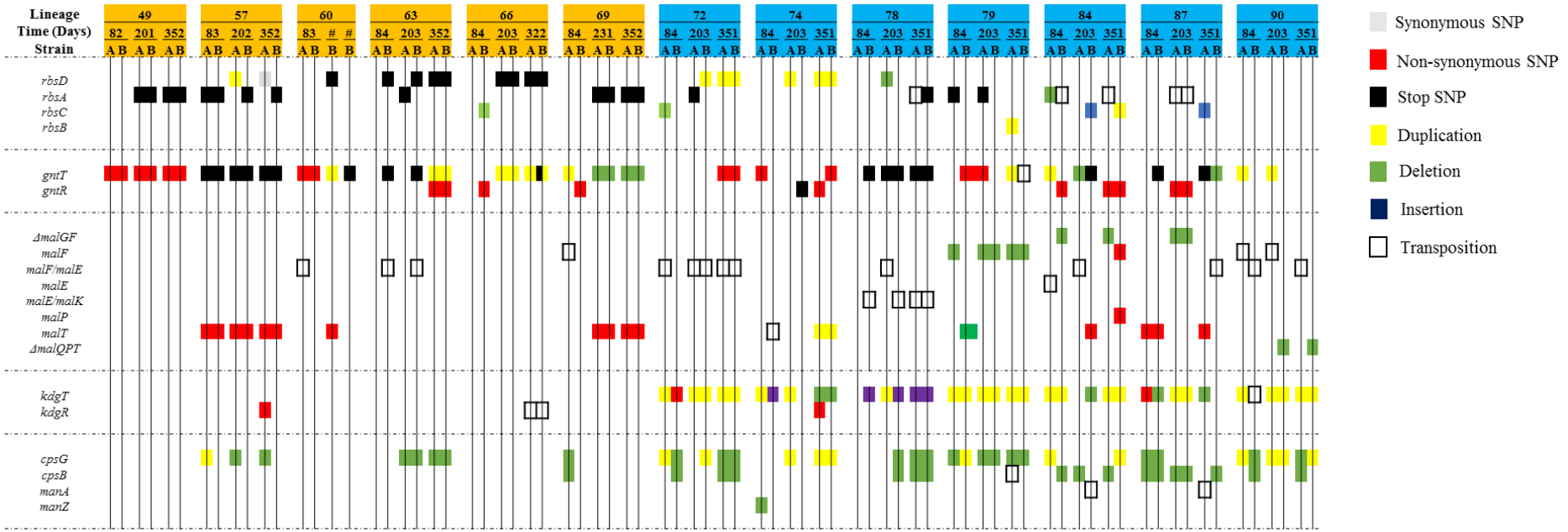
<u>Figure 3.</u> The genetic basis of adaptive mutations and convergences between lineages in *E. coli* 536 in WD

We characterized and located convergent mutations identified in *E. coli* 536. For simplicity, the genomes are represented linearly and vertically drawn. We represented genes involved in ribosome maturation and conferring growth advantage in streptomycine media (*gidA* and *rluD*), in constitutive utilization of lactose (*lacI* and *lacO*) and in maltose use inhibition (*malT*). The type of mutation were shown by different color: pink for synonymous mutation, black for stop SNPs, red for non-synonymous mutation, yellow for duplication, green for deletion, violet for insertion.



<u>Figure 4.</u> The genetic basis of adaptive mutation on sugar or energetic substrate metabolism

We characterized and located convergent identified mutations on sugar or energetic substrate metabolism. We represented ribose (*rbsD*, *rbsA*, *rbsC* and *rbsB*), maltose (*malF*, *malF/malE*, *malE*, *malE/malK*, *malP* and *malT*), gluconate (*gntT* and *gntR*), 2-keto-3-deoxy-D-gluconate (*kdgT* and *kdgR*) and mannose (*cpsG*, *cpsB*, *manA* and *manZ*) metabolism genetic support. The type of mutation were shown by different color: pink for synonymous mutation, black for stop SNPs, red for non-synonymous mutation, yellow for duplication, green for deletion, violet for insertion and white for transposition. The short deletion were represented only with a triangle while large deletion were represented with a green band below the region deleted.



<u>Figure 5.</u> Mutations convergence on sugar metabolism between lineages

Identified mutations in the two different isolates (A and B) from the 13 lineages at each time point were represented along the *E. coli* chromosome. For simplicity, the genomes are represented linearly and vertically drawn. Lineages from WD group are in orange while CD group are in blue. We represented ribose (*rbsD*, *rbsA*, *rbsC* and *rbsB*), maltose (*malF*, *malF/malE*, *malE*, *malE/malK*, *malP* and *malT*), gluconate (*gntT* and *gntR*), 2-keto-3-deoxy-D-gluconate (*kdgT* and *kdgR*) and mannose (*cpsG*, *cpsB*, *manA* and *manZ)*. The type of mutation were shown by different color: pink for synonymous mutation, black for stop SNPs, red for non-synonymous mutation, yellow for duplication, green for deletion, violet for insertion and white for transposition. The short deletion were represented only with a triangle while large deletion were represented with a green band below the region deleted.

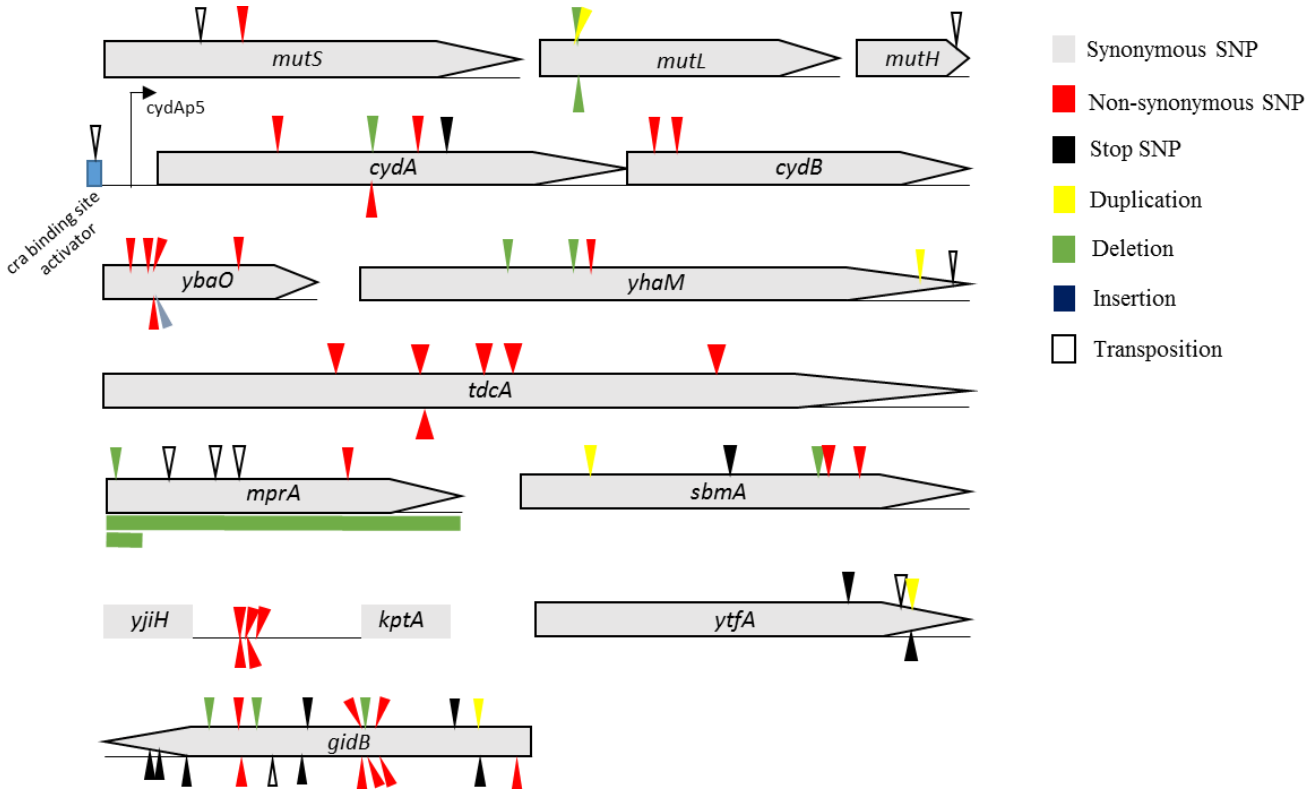


Figure 6. The genetic basis of adaptive mutation on non-sugar metabolism in *E. coli* HS

We characterized and located convergent identified mutations on non-sugar metabolism. We represented mutations types in genes involved in hypermutator phenotype (*mutS*, *mutL* and *mutH*), in amino-acid metabolism in anaerobiosis condition (*ybaO*, *yhaM* and *tdcA*), in energetic production in low oxygen condition (*cydA* and *cydB*), in antibiotic resistance (*mprA* and *sbmA*), in ribosome maturation and conferring growth advantage in streptomycine media (*gidB*) and gene or region whose function remain unclear (intergenic region between *yijH* and *kptA*, and *ytfA*). The type of mutation were shown by different color: pink for synonymous mutation, black for stop SNPs, red for non-synonymous mutation, yellow for duplication, green for deletion, violet for insertion and white for transposition. The short deletion were represented only with a triangle while large deletion were represented with a green band below the region deleted.

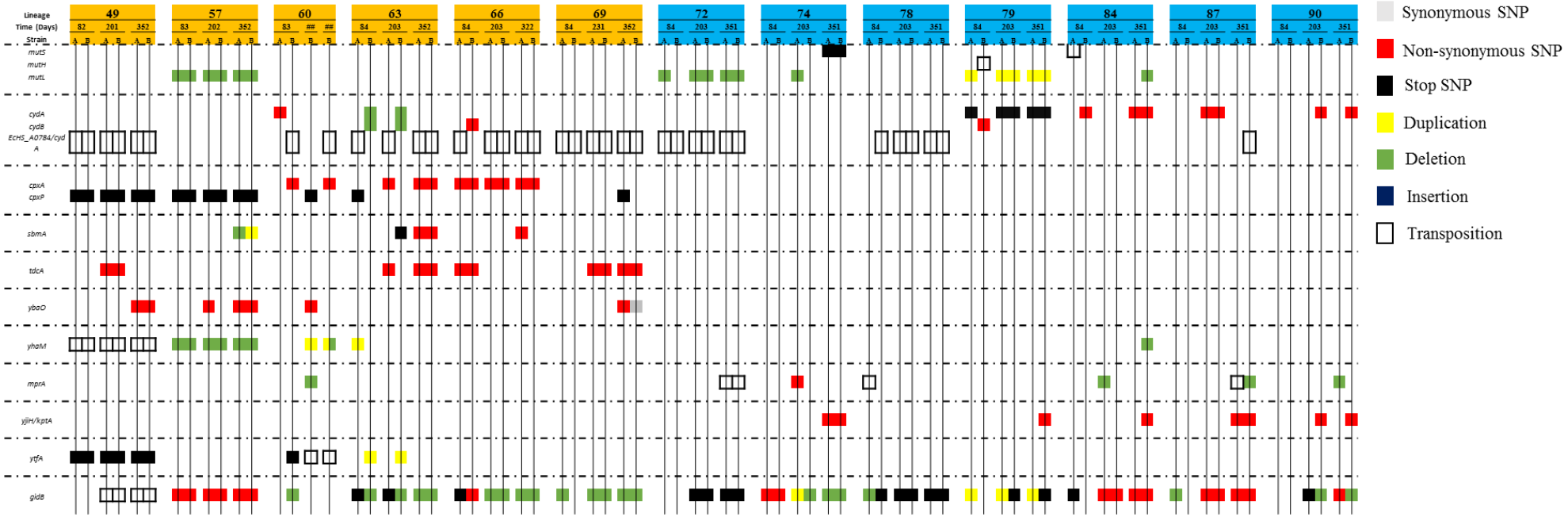


Figure 7. Mutations convergence on non-sugar metabolism between lineages

Identified mutations in the two different isolates (A and B) from the 13 lineages at each time point were represented along the *E. coli* chromosome. For simplicity, the genomes are represented linearly and vertically drawn. Lineages from WD group are in orange while CD group are in blue. We represented mutations types in genes involved in hypermutator phenotype (*mutS*, *mutL* and *mutH*), in amino-acid metabolism in anaerobiosis condition (*ybaO*, *yhaM* and *tdcA*), in energetic production in low oxygen condition (*cydA* and *cydB*), in antibiotic resistance (*mprA* and *sbmA*), in ribosome maturation and conferring growth advantage in streptomycine media (*gidB*) and gene or region whose function remain unclear (intergenic region between *yijH* and *kptA*, and *ytfA*). The type of mutation were shown by different color: pink for synonymous mutation, black for stop SNPs, red for non-synonymous mutation, yellow for duplication, green for deletion, violet for insertion and white for transposition.

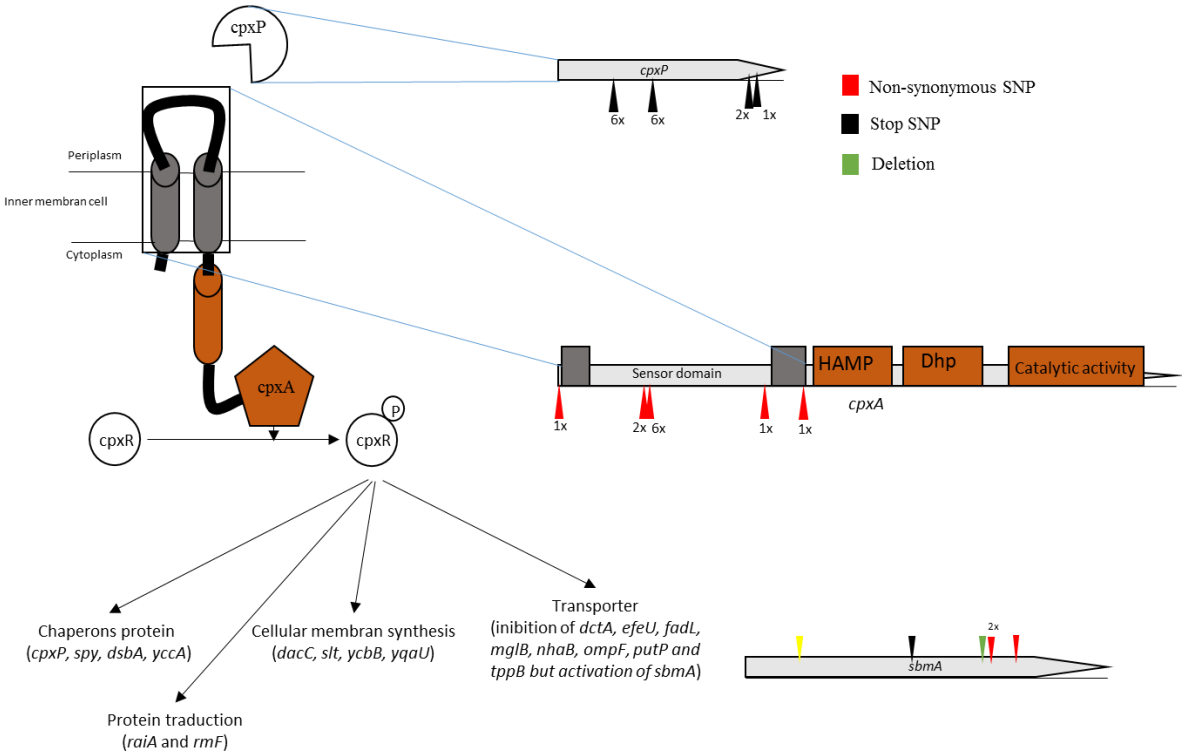


Figure 8. Functional and genetic support of constitutive activation of cpx pathway in WD.

We represented a schematic structure and fonction of cpx pathway. We represented the mutations occuring on the sensor domain of *cpxA* , on the cpx inhibitor *cpxP* and compensation mutations on *sbmA*. The type of mutation were shown by different color: black for stop SNPs, red for non-synonymous mutation, yellow for duplication and green for deletion.

Gene	Position	Mutation	Mutation effect	Mutation type	Occurrence	Mice ID	Diet
mutH →	2,988,577	IS1	coding (596/690 nt)	replication	1	79BJ84	CD
	4,422,361	Δ87 bp	coding (214-300/1848 nt)	deletion	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD
mutL →	4,422,362	Δ6 bp	coding (215-220/1848 nt)	deletion	7	72AJ84; 72AJ203; 72AJ351; 72BJ203; 72BJ351; 74AJ203; 84BJ351	CD
	4,422,367	(TGGCGC)3→4	coding (220/1848 nt)	duplication	5	79AJ84; 79AJ203; 79AJ351; 79BJ203; 79BJ351	CD
mutS →	2,875,090	IS1	coding (580/2562 nt)	conservative	1	84AJ84	CD
	2,875,351	C→T	Q281* (CAG→TAG)	stop SNP	2	74AJ351; 74BJ351	CD

Table 1. Characteristics and distributions of hypermutators strains.

The name of the genes as well as the orientation (defined by the arrow), the position along the genome, and the flanking genes in case of an intergenic mutation are displayed, with the mutations, their type, the number of occurrences, isolates which carry the mutations and diet group. Stop codons are indicated with the symbol *.

Gene	Position	Mutation	Occurrence	Mice ID	Strain	Diet	Gene information
[ECP_0273]–ECP_0342	294,387	Δ68,039	1	9AJ353	536	WD	77 genes - [ECP_0273], int, xis, ECP_0275, ECP_0276, ECP_0277, ECP_0278, ECP_0279, ECOLI0280, ECP_0280, ECP_0281, ECP_0282, insF, insE, ECP_0285, ECP_0286, ECOLI0287, ECP_0287, ECP_0288, ECP_0289, ECP_0290, papI.2, papB, ECP_0293, ECP_0294, ECP_0295, ECP_0296, ECP_0297, ECP_0298, ECP_0299, ECP_0300, ECP_0301, ECOLI0303, ECP_0302, ECP_0303, ECP_0304, ECP_0305, ECP_0306, ECP_0307, insG, ECOLI0311, ECP_0309, ECP_0310, ECP_0311, ECP_0312, ECP_0313, ECP_0314, ECP_0315, ECP_0316, ECOLI0318, ECP_0317, ECP_0318, ECP_0319, ECP_0320, ECP_0321, ECP_0322, ECP_0323, ECP_0324, ECP_0325, ECP_0326, ECP_0327, ECP_0328, ECP_0329, ECP_0330, ECP_0331, flu, ECP_0333, ECP_0334, yeeS, yeeT, ECOLI0337, yeeU, yeeV, yeeW.2, ECP_0340, ECP_0341, ECP_0342
[cpsG]–[cpsG]	2,139,900	Δ19,522	1	11BJ353	536	WD	16 genes - [cpsG], ECP_2074, ECP_2075, ECP_2076, ECP_2077, ECP_2078, ECP_2079, ECP_2080, ECP_2081, galF, wcaM, wcaL, wcaK, wzxC, wcaJ, [cpsG]
[yeeT]–[ECP_3861]	4,021,738	Δ1,629	1	6AJ353	536	WD	[yeeT], yeeU, yeeV, yeeW.2, [ECP_3861]
[mngB]–EcHS_A0784	795,175	Δ3,704 bp	1	60BJ202	HS	WD	[mngB], [insB], insA, mngB, EcHS_A0783, EcHS_A0784/cydA
[mngB]–[ybgE]	801,406	Δ6,940 bp	2	63BJ84, 63BJ203	HS	WD	[mngB], [insB], insA, mngB, EcHS_A0783, EcHS_A0784, cydA, cydB, ybgT, [ybgE]
[insA]–yobF	1,927,210	Δ14,556 bp	1	90AJ351	HS	CD	18 genes - [insA], insB, EcHS_A1901, yoaH, pabB, yeaB, sdaA, adrB, yoaE, manX, manY, manZ, yobD, ESCCO1871, yebN, rrmA, cspC, yobF
[insA]–[kdgR]	1,928,293	Δ3,696 bp	2	66AJ322, 66BJ322	HS	WD	[mngB], [insB], insA, mngB, EcHS_A0783, EcHS_A0784
[EcHS_A2124]–[yegH]	2,109,390	Δ73,543 bp	1	69AJ84	HS	WD	92 genes - [EcHS_A2124], EcHS_A2127, EcHS_A2127, EcHS_A2127, EcHS_A2127, EcHS_A2128, EcHS_A2129, EcHS_A2129, EcHS_A2131, ESCCO2087, EcHS_A2132, EcHS_A2133, ESCCO2089, EcHS_A2134, ESCCO2091, EcHS_A2135, EcHS_A2136, ESCCO2094, EcHS_A2137, EcHS_A2138, EcHS_A2139, yeeU, yeeV, yeeW, EcHS_A2142, ESCCO2101, EcHS_A2143, ESCCO2103, ESCCO2104, EcHS_A2144, yeeX, yeeA, sbmC, dacD, sbcB, yeeD, yeeE, yeeF, yeeY, yeeZ, yoeB, yefM, hisL, hisG, hisD, hisC, hisB, hisH, hisA, hisF, hisI, EcHS_A2166, EcHS_A2167, EcHS_A2168, EcHS_A2169, EcHS_A2170, EcHS_A2171, cpsG, cpsB, EcHS_A2174, ESCCO2133, EcHS_A2175, insB, insA, ugd, EcHS_A2179, EcHS_A2180, EcHS_A2181, EcHS_A2182, gnd, EcHS_A2184, EcHS_A2185, EcHS_A2186, EcHS_A2187, EcHS_A2188, EcHS_A2189, EcHS_A2190, EcHS_A2191, EcHS_A2192, EcHS_A2193, EcHS_A2194, EcHS_A2195, EcHS_A2196, insA, insB, EcHS_A2199, ESCCO2156, EcHS_A2200, EcHS_A2200, galF, ESCCO2160, [yegH]
[yeeA]–[EcHS_A2171]	2,125,896	Δ26,098 bp	1	74BJ203	HS	CD	26 genes - [yeeA], sbmC, dacD, sbcB, yeeD, yeeE, yeeF, yeeY, yeeZ, yoeB, yefM, hisL, hisG, hisD, hisC, hisB, hisH, hisA, hisF, hisI, EcHS_A2166, EcHS_A2167, EcHS_A2168, EcHS_A2169, EcHS_A2170, EcHS_A2171, [EcHS_A2171]
[dacD]–[galF]	2,127,546	Δ53,668 bp	1	78BJ351	HS	CD	57 genes - [dacD], sbcB, yeeD, yeeE, yeeF, yeeY, yeeZ, yoeB, yefM, hisL, hisG, hisD, hisC, hisB, hisH, hisA, hisF, hisI, EcHS_A2166, EcHS_A2167, EcHS_A2168, EcHS_A2169, EcHS_A2170, EcHS_A2171, cpsG, cpsB, EcHS_A2174, ESCCO2133, EcHS_A2175, insB, insA, ugd, EcHS_A2179, EcHS_A2180, EcHS_A2181, EcHS_A2182, gnd, EcHS_A2184, EcHS_A2185, EcHS_A2186, EcHS_A2187, EcHS_A2188, EcHS_A2189, EcHS_A2190, EcHS_A2191, EcHS_A2192, EcHS_A2193, EcHS_A2194, EcHS_A2195, EcHS_A2196, insA, insB, EcHS_A2199, ESCCO2156, EcHS_A2200, EcHS_A2200, [galF]
[yeeE]–[EcHS_A2200]	2,131,128	Δ48,622 bp	1	87AJ84	HS	CD	52 genes - [yeeE], yeeF, yeeY, yeeZ, yoeB, yefM, hisL, hisG, hisD, hisC, hisB, hisH, hisA, hisF, hisI, EcHS_A2166, EcHS_A2167, EcHS_A2168, EcHS_A2169, EcHS_A2170, EcHS_A2171, cpsG, cpsB, EcHS_A2174, ESCCO2133, EcHS_A2175, insB, insA, ugd, EcHS_A2179, EcHS_A2180, EcHS_A2181, EcHS_A2182, gnd, EcHS_A2184, EcHS_A2185, EcHS_A2186, EcHS_A2187, EcHS_A2188, EcHS_A2189, EcHS_A2190, EcHS_A2191, EcHS_A2192, EcHS_A2193, EcHS_A2194, EcHS_A2195, EcHS_A2196, insA, insB, EcHS_A2199, [ESCCO2156], [EcHS_A2200]
[yeeF]–[gnd]	2,131,575	Δ30,294 bp	2	72AJ351, 72BJ351	HS	CD	33 genes - [yeeF], yeeY, yeeZ, yoeB, yefM, hisL, hisG, hisD, hisC, hisB, hisH, hisA, hisF, hisI, EcHS_A2166, EcHS_A2167, EcHS_A2168, EcHS_A2169, EcHS_A2170, EcHS_A2171, cpsG, cpsB, EcHS_A2174, ESCCO2133, EcHS_A2175, insB, insA, ugd, EcHS_A2179, EcHS_A2180, EcHS_A2181, EcHS_A2182, [gnd]
[ykgN]–[EcHS_A0326]	2,131,576	Δ32,877 bp	3	72BJ351,72AJ351, 78BJ351	HS	CD	53 genes - [ykgN], EcHS_A0281, EcHS_A0282, EcHS_A0283, ESCCO0275, EcHS_A0284, ESCCO0277, EcHS_A0285, EcHS_A0286, EcHS_A0287, EcHS_A0288, cII, cro, EcHS_A0291, ESCCO0285, EcHS_A0292, P, EcHS_A0294, ESCCO0289, EcHS_A0295, EcHS_A0296, ninF, EcHS_A0298, EcHS_A0299, rusA, EcHS_A0301, EcHS_A0302, ESCCO0296, EcHS_A0303, r, EcHS_A0305, EcHS_A0306, EcHS_A0307, EcHS_A0308, EcHS_A0309, EcHS_A0310, EcHS_A0311, EcHS_A0312, EcHS_A0313, EcHS_A0314, EcHS_A0315, EcHS_A0316, ESCCO0311, EcHS_A0317, EcHS_A0318, EcHS_A0319, EcHS_A0320, EcHS_A0321, EcHS_A0322, EcHS_A0323, EcHS_A0324, ykgN, [EcHS_A0326]
[EcHS_A2169]–[EcHS_A2187]	2,148,795	Δ19,038 bp	2	78AJ351, 78BJ203	HS	CD	20 genes - [EcHS_A2169], EcHS_A2170, EcHS_A2171, cpsG, cpsB, EcHS_A2174, ESCCO2133, EcHS_A2175, insB, insA, ugd, EcHS_A2179, EcHS_A2180, EcHS_A2181, EcHS_A2182, gnd, EcHS_A2184, EcHS_A2185, EcHS_A2186, [EcHS_A2187]
[cpsG]–[EcHS_A2200]	2,152,639	Δ27,065 bp	2	90AJ351, 90BJ84	HS	CD	31 genes - [cpsG], cpsB, EcHS_A2174, ESCCO2133, EcHS_A2175, insB, insA, ugd, EcHS_A2179, EcHS_A2180, EcHS_A2181, EcHS_A2182, gnd, EcHS_A2184, EcHS_A2185, EcHS_A2186, EcHS_A2187, EcHS_A2188, EcHS_A2189, EcHS_A2190, EcHS_A2191, EcHS_A2192, EcHS_A2193, EcHS_A2194, EcHS_A2195, EcHS_A2196, insA, insB, EcHS_A2199, [ESCCO2156], [EcHS_A2200]
[cpsB]–[EcHS_A2182]	2,154,435	Δ7,103 bp	4	84AJ351, 84BJ84, 87AJ203, 87BJ203	HS	CD	cpsB], EcHS_A2174, ESCCO2133, EcHS_A2175, insB, insA, ugd, EcHS_A2179, EcHS_A2180, EcHS_A2181, [EcHS_A2182]
[ugd]–[insB]	2,157,053	Δ21,948 bp	2	90BJ203, 90BJ351	HS	CD	21 genes - [ugd], EcHS_A2179, EcHS_A2180, EcHS_A2181, EcHS_A2182, gnd, EcHS_A2184, EcHS_A2185, EcHS_A2186, EcHS_A2187, EcHS_A2188, EcHS_A2189, EcHS_A2190, EcHS_A2191, EcHS_A2192, EcHS_A2193, EcHS_A2194, EcHS_A2195, EcHS_A2196, insA, [insB]
[EcHS_A2785]–[mprA]	2,800,568	Δ28,980 bp	1	60BJ202	HS	WD	37 genes - [EcHS_A2785], EcHS_A2786, ypjC, EcHS_A2788, ykgN, EcHS_A2790, EcHS_A2791, ESCCO2738, EcHS_A4687, EcHS_A2793, ygaT, ygaF, gabD, gabT, gabP, csiR, ygaU, yqaE, ygaV, ygaP, stpA, EcHS_A2805, ygaW, ygaC, ygaM, nrdH, nrdI, nrdE, nrdF, proV, proW, proX, EcHS_A2816, EcHS_A2817, ygaZ, ygaH, [mprA]
[EcHS_A2785]–[ESCCO2738]	2,802,988	Δ3,120 bp	2	74AJ351, 74BJ351	HS	CD	[EcHS_A2785], EcHS_A2786, ypjC, EcHS_A2788, ykgN, EcHS_A2790, [EcHS_A2791], [ESCCO2738]
[EcHS_A2817]–[mprA]	2,827,037	Δ1,337 bp	2	84AJ203, 87BJ351	HS	CD	[EcHS_A2817], ygaZ, ygaH, [mprA]
[gntT]–[glpR]	3,591,863	Δ12,822 bp	2	90BJ203, 90BJ351	HS	CD	[gntT], malQ, malP, malT, rtcA, rtcB, rtcR, [glpR]
[malG]–[malF]	4,274,718	Δ1,281 bp	4	84AJ351, 84BJ84, 87AJ203, 87BJ203	HS	CD	[malG], [malF]

Table 2. Summary of large deletion (>1000 bp) that we found in evolved lineages

SYNTHESE ET PERSPECTIVES

E. coli est une bactérie versatile commensale du tube digestif qui dans certaines circonstances dépendantes de la souche (Picard *et al.* 1999) et de l'hôte (Lefort *et al.* 2011) peut devenir un pathogène redoutable. Elle entraine une mortalité et une morbidité significative en provoquant des pathologies intestinales ou extra-intestinales telles que des infections urinaires, des atteintes digestives (péritonites, cholécystite, …) ou les méningites néo-natales. L'ensemble de ces infections peuvent s'accompagner de bactériémies. La situation est d'autant plus préoccupante que *E. coli*, et plus généralement les entérobactéries, semblent être devenues les bactéries le plus souvent impliquées dans des infections communautaires ou nosocomiales chez des patients de réanimation (Javaloyas, Garcia-Somoza, and Gudiol 2002; Diekema *et al.* 2003; Laupland *et al.* 2007; Lyytikäinen *et al.* 2002). Elle est aussi associée à une forte résistance aux antibiotiques avec notamment l'émergence et la large diffusion du clone ST131 producteur de la BLSE CTX-M-15 (Chong, Shimoda, and Shimono 2018). *E. coli*, et plus généralement les entérobactéries, seraient même impliquées dans des pathologies métaboliques comme l'insulinorésistance ou l'obésité (Cani *et al.* 2008) ou des pathologies carcinologiques (Raisch *et al.* 2014).

E. coli possède une structure à la fois clonale avec un « core » génome susceptible de d'être le site de recombinaisons. Les groupes phylogénétiques de l'espèce *E. coli*, sont prédictifs de sa virulence. Notamment, les souches du groupe phylogénétique A sont connues pour être très majoritairement commensales alors que les souches *E. coli* du groupe phylogénétique B2 sont connues non seulement pour être de bonnes commensales mais aussi susceptibles de provoquer des pathologies extra-intestinales. Les souches de *E. coli* du groupe phylogénétique B2 sont porteuses de gènes de virulences dans des îlots de pathogénicité leur permettant de coloniser le tube digestif (Diard *et al.* 2010), faisant de la virulence un sous-produit du commensalisme. *E. coli* possède donc un fond génétique pouvant lui permettre d'influencer sa capacité à coloniser le tube digestif. Sa grande adaptabilité est expliquée d'une part par le fait

que l'espèce *E. coli* possède un génome variable modulable soumis à des flux entrant et sortant de gènes provenant de son environnement. Par ailleurs, tout le génome est susceptible de subir des mutations ponctuelles, indels ou par transposition pouvant exprimer un trait phénotypique pouvant persister en fonction de la sélection ou de la dérive génétique. Ainsi, l'espèce *E. coli* est capable de coloniser un environnement ouvert complexe associant le reste du microbiote, le système immunitaire et l'épithélium digestif de l'hôte, ainsi que des facteurs extérieurs comme les antibiotiques et l'alimentation. Autant d'éléments auxquels cette espèce doit potentiellement s'adapter.

La distribution de la proportion des différents groupes phylogénétiques dans le tube digestif des individus n'est pas la même dans les différentes régions du monde. En effet, il semblerait qu'il y ait une plus forte proportion de souches d' *E. coli* du groupe phylogénétique B2 dans les pays occidentaux ou industrialisées depuis 30 ans (Massot *et al.* 2016). Différents facteurs susceptibles de favoriser un groupe phylogénétique par rapport à un autre peuvent être évoqués. Ils sont probablement liés au mode de vie telles que l'alimentation ou encore l'hygiène.

Connaissant l'importance de l'alimentation dans la diversité du microbiote (Ley *et al.* 2005), je me suis focalisé sur son impact en utilisant une approche génomique basée sur la théorie de l'évolution pour répondre à cette question. En utilisant le séquençage, nous pouvons détecter des mutations et identifier celles qui sont associées à de l'adaptation. Ces modifications au sein du génome de *E. coli* permettent de comprendre comment l'environnement façonne et sélectionne les populations de *E. coli* s'adaptant au tube digestif. Bien sûr, d'autres études, utilisant cette approche, ont déjà été menées permettant d'identifier certaines cibles de l'adaptation de *E. coli* dans le tube digestif (Barroso-Batista *et al.* 2014; De Paepe *et al.* 2011; Lescat *et al.* 2017). Cependant, certaines études souffraient de limitations telles que l'utilisation d'une souche de laboratoire maladaptée au tube digestif (Couce and Tenaillon 2015) qui

évoluait, dans toutes les études, dans un milieu non naturel avec un seul régime alimentaire. Mon propos durant ce travail de thèse a été de comprendre l'adaptation de *E. coli* dans son milieu naturel ainsi que l'impact de l'alimentation de l'hôte et du fond génétique de la souche colonisatrice.

Dans **notre première étude,** nous avons eu l'opportunité de suivre le clone *E. coli* ED1a, appartenant au groupe phylogénétique B2, persistant et dominant dans le tube digestif d'un individu vivant à Paris pendant à peu près un an. Au cours de cette année de suivi, l'individu n'a pas présenté de pathologie infectieuse ou autre et n'a donc pas pris d'antibiotique. De plus il n'a pas modifié ses habitudes alimentaires. L'analyse des séquences ne nous ont pas permis de mettre en évidence de sélection suggérant une évolution neutre de *E. coli* ED1a dans le tube digestif d'un individu parisien en bonne santé.

La persistance de la souche ED1a, nous a permis d'étudier rigoureusement le taux de mutation d'une même souche *E. coli* au fil du temps dans des conditions naturelles en utilisant une approche Bayésienne. L'approche Bayésienne infère l'âge des nœuds de la généalogie, inconnu *a priori*, à partir d'une hypothèse de distribution inférée par la théorie de la coalescence auquel on ajoute une hypothèse d'horloge moléculaire pour l'accumulation des mutations. Nous avons donc calculé le taux de mutation en utilisant une approche bayésienne mise en œuvre dans BEAST (Drummond and Rambaut 2007). Un taux de mutation de $6{,}90 \times 10^{-7}$ mutations par base et par an (intervalle de confiance à 95%, $3{,}14 \times 10^{-7}$ à $1{,}40 \times 10^{-6}$ mutations par base par an) a été obtenu sur le génome chromosomique, et un taux de de $7{,}18 \times 10^{-7}$ mutations par base et par an (IC à 95%, $3{,}28 \times 10^{-7}$ à $1{,}44 \times 10^{-6}$ mutations par base par an) était obtenu si le plasmide était inclus.

Une fois que le taux de mutation par base et par an (μ_{est}) a été calculé, nous avons estimé le taux de réplication (G) de *E. coli* ED1a dans un tube digestif humain sain en utilisant la formule $G = \mu_{est} / (\mu_g \times 365)$. μ_g, le taux de mutation par base et par génération, a été estimé à

partir d'une fourchette de valeurs extrêmes retrouvées dans la littérature : $0,89x10^{-10}$ (Wielgoss *et al.* 2013) et $3,12x10^{-10}$ (Foster *et al.* 2015) mutations par base par génération. Ainsi, nous avons obtenu un taux de génération compris entre 6,06 (95% CI, 2,88 to 12,65) et 21,24 (95% CI, 9,66 to 43,28) générations par jour pour *E. coli* ED1a dans le tube digestif d'un individu en bonne santé. En utilisant seulement des mutations ponctuelles sur le chromosome, la mesure de la diversité estimée par le thêta de Watterson (θ_w) était de 1,93 ± 1,13, 1,54 ± 0,96 et 1,15 ± 0,78 pour les trois points temporels. En utilisant ces dernières estimations, la longueur du génome (L) et une fourchette de taux de mutation par base et par génération allant de $0,89x10^{-10}$ à $3,12x10^{-10}$, $\theta_w = 2N_eL\mu_g$, nous avons estimé la taille de la population efficace (N_e) entre 500 et 1700. Ce qui veut dire qu'une petite fraction des cellules dans le mucus se reproduisent et sont à l'origine du reste de la population. Les données sont en effet compatibles avec une petite taille de population efficace évoluant de manière neutre même si le clone a été dominant dans son hôte pendant presque un an.

Cette étude, nous a donc permis de déterminer le mode d'adaptation d'une souche de *E. coli*, probablement déjà bien adaptée, dans le tube digestif d'un individu occidental. Mais l'étude de l'impact du facteur alimentation sur l'adaptation d'un isolat naturel de *E. coli* nécessite le recours à une expérimentation évolutive contrôlée reproduisant au maximum un milieu naturel.

L'objectif de **notre seconde étude** était d'étudier l'adaptation d'un isolat naturel de *E. coli* 536 appartenant au groupe phylogénétique B2, dans un milieu aussi proche que possible des conditions naturelles. Ce milieu a été le tube digestif de souris soumises à deux régimes différents : un régime standard et un régime riche en gras et en sucre. Le propos de cette expérience était de se rapprocher encore plus des conditions naturelles en éliminant le plus possible la streptomycine. Pour ce faire, la streptomycine n'a été utilisée que pour inoculer la souche chez des souris gravides et arrêtée au moment de l'accouchement. Les mères ont ensuite

transmis *E. coli* avec le reste de leur microbiote à leur progéniture. Après 3 semaines d'allaitement, X souris femelles ont été séparées dans des cages indépendantes et *E. coli* a été extrait de leurs fèces de manière itérative pendant un an pour des analyses génomiques.

Nous avons d'abord observé une difficulté de *E. coli* à coloniser le tube digestif de souris avec des phénomènes de clairance complète associée à des phénomènes de recolonisation (Reeves *et al.* 2011) d'un point de temps de prélèvements à un autre. Le régime alimentaire riche en gras et en sucre permettait un maintien significativement supérieur de *E. coli* dans le tube digestif de souris par rapport à CD. Cela fait écho à la constatation d'une proportion de *E. coli*, appartenant au groupe phylogénétique B2, plus importante en commensal digestif chez les individus des pays industrialisés (Tenaillon *et al.* 2010).

Nous avons pu déterminer un taux de mutation et la sélection de *E. coli* 536 dans le tube digestif de souris reproduisant le plus possible les conditions naturelles par 2 approches différentes. Le choix dans l'utilisation des différentes approches dépend de l'hypothèse de fréquence d'apparition des mutations. Posant l'hypothèse que les mutations sont un phénomène suffisamment fréquent dans le temps pour s'accumuler sur un modèle linéaire, nous avons donc adopté dans un premier temps une approche par régression linéaire. Nous avons pu calculer un taux de mutation de 1.27×10^{-7} (95% CI: 2.13×10^{-8} - 2.34×10^{-7}). Posant maintenant l'hypothèse que les mutations sont un phénomène rare dont la distribution dans le temps suit une loi de Poisson, nous pouvons estimer un taux de mutation et un différentiel de sélection entre mutations synonymes et non synonymes ω par maximum de vraisemblance. Nous avons ainsi pu calculer un taux de mutation à 4.9×10^{-7} (95% CI: 3.3×10^{-7} - 6.8×10^{-7}) pondéré par ω et le biais mutationnel.

Nous avons mis en évidence une adaptation de *E. coli* 536 en fonction du régime alimentaire durant la période d'allaitement. Nous avons mis en évidence des mutations sur le répresseur transcriptionnel de l'opéron lactose *lacI* et un de ses sites de liaison à l'ADN *lacO1*

entrainant une activation constitutive de l'opéron lactose. Ce trait phénotypique a principalement émergé lors de la période d'allaitement de 3 semaines probablement adapté au fort apport de lactose dans le lait. Par contre, nous n'avons pas détecté de sélection lors de la période de suivi de *E. coli* 536 dans les fèces des progénitures quelquesoit le régime alimentaire. Ce dernier résultat peut témoigner de l'évolution neutre de *E. coli* 536 dans un tube digestif de souris se rapprochant de son milieu naturel, confirmant ainsi le résultat de l'étude précédente. L'une des limites de notre étude était une certaine difficulté à maintenir une population de *E. coli* dans le tube digestif de souris et une densité bactérienne plutôt basse par rapport à ceux constatés dans le modèle de colonisation de souris traitées à la streptomycine (Lescat *et al.* 2017) ou axénique (De Paepe *et al.* 2011). La diversité génétique dépend du taux de mutation générationnel et de la taille de population efficace. Nous ne pouvons ainsi pas écarter l'absence de détection de sélection sur notre impossibilité à produire suffisamment de diversité dans les conditions naturelles sans streptomycine.

Dans **notre troisième étude**, nous avons voulu étudier l'impact du fond génétique de la souche colonisatrice combiné au facteur du régime alimentaire de l'hôte dans l'adaptation de deux souches différentes dans le tube digestif de souris traitées à la streptomycine, afin d'éviter la potentielle limite d'une faible taille de population. Nous avons donc étudié l'évolution de 2 isolats naturels: l'un virulent, à nouveau la souche *E. coli* 536, et l'autre la souche *E. coli* HS purement commensale, appartenant au groupe A. Les souris colonisées avec la souche HS ont également été soumises à deux régimes différents comme précédemment. Les souris colonisées avec la souche 536 ont été soumises à un régime riche en gras et sucre. Nous avons ensuite comparé ces résultats avec ceux de l'étude que nous avons mené précédemment où nous avions colonisé des souris soumises à un régime standard avec la souche 536 (Lescat *et al.* 2017).

Cette troisième étude est à un stade tout à fait préliminaire, mais nous pouvons en dégager un certain nombre de résultats. Nous avons pu observer une différence de profils

d'adaptation entre les souches *E. coli* 536 et HS dans le tube digestif de souris traitées à la streptomycine.

E. coli 536 semble ne pas subir une forte sélection face à son environnement. En effet, chez les souches issues des souris soumises à un régime alimentaire riche en gras et en sucre, un faible signal pour une inactivation théorique du métabolisme du maltose et une activation constitutive du lactose, ainsi qu'une convergence mutationnelle sur les gènes *rluD* et *gidB* déjà décrite dans l'adaptation de *E. coli* 536 dans un tube digestif de souris traitées à la streptomycine soumises à un régime alimentaire standard (Lescat *et al.* 2017). Nous n'avons pas mis en évidence de mutation dans le répresseur *dgoR* dans le groupe des souches des souris soumises à régime alimentaire riche en gras et en sucre menant vers l'activation constitutive de l'utilisation du galactonate décrite dans Lescat *et al.* où un régime standard était utilisé suggérant une adaptation spécifique au régime alimentaire. Ces données sont assez proches de l'observation que les souches appartenant au groupe phylogénétique B2, *E. coli* 536 ou ED1a évoluent de façon neutre respectivement dans un tube digestif de souris reproduisant des conditions naturelles (Article n°2) ou dans le tube digestif d'un individu occidental en bonne santé (Article n°1). Ce faisceau d'observations semble concourir à une évolution naturellement neutre de *E. coli* appartenant au groupe phylogénétique B2 dans son milieu naturel. Cela suggère que les souches *E. coli* du groupe phylogénétique B2, déjà bien adaptées au tube digestif dans les hôtes étudiés, limite l'opportunité de mutations bénéfiques de persister (Couce and Tenaillon 2015).

Au contraire, *E. coli* HS, appartenant au groupe phylogénétique A, semble subir une intense sélection dans le tube digestif. Premièrement, *E. coli* HS possède une forte diversité génétique, dans les deux régimes alimentaires, rendue possible par (i) la sélection de souches hypermutatrices envahissant 4 lignées (dont 3 précocement), (ii) une intense activité de transposition et (iii) plusieurs larges délétions de plus de 1,000 bp. Deuxièmement, on observe

plusieurs convergences mutationnelles dont les cibles principales sont les gènes impliqués dans le métabolisme de sucres comme le maltose, le ribose, le mannose et les deux substrats de la voie d'Entner-Doudoroff (ED): gluconate et le 2-kéto-3-deoxy-gluconate (2KDG). Nous avons observé des mutations principalement dans les gènes codant pour le transport (*rbsA, rbsB, rbsC, malE, malF, malG, gntT, kdgT* and *manZ*), les gènes régulateurs (*malT, gntR, kdgR*) ou leurs sites de liaison à l'ADN et dans les gènes impliqués dans leur utilisation (*rbsD, cpsG, cpsB* and *manA*). La voie ED est indispensable à la colonisation du tube digestif de souris traitées à la streptomycine suggérant une consommation préférentielle de gluconate ou de 2KDG par *E. coli* (Peekhaus and Conway 1998). Or la présence de mutations, potentiellement inhibitrices, sur les gènes *gntT* codant pour un transporteur du gluconate (préférentiellement dans le régime riche en gras et en sucre) et *kdgT* codant pour une protéine de transport du 2KDG (exclusivement dans le régime alimentaire standard) suggèrent l'utilisation de l'un ou l'autre sucre en fonction du régime alimentaire. Par ailleurs, la présence de mutations à la fois sur *kdgT* et *gntT* dans près de la moitié des isolats de *E. coli* HS dans le régime alimentaire standard suggère l'utilisation d'autres sucres présents dans l'environnement susceptibles de passer par la voie ED et ayant comme produit de dégradation intermédiaire le gluconate ou le 2KDG. Les mutations des gènes de l'opéron maltose conduisant à une inhibition théorique de sa voie de métabolisation sont les seuls que nous retrouvons dans les deux souches de *E. coli* 536 dans le groupe nourri par un régime riche en gras et en sucre et dans les souches *E. coli* HS quelquesoit le régime alimentaire. La présence du même type de mutation sur le gène *malT* dans une étude dévolution de *E. coli* K12 dans le tube digestif de souris axénique suggère une adaptation commune au tube digestif de souris. Le mannose peut emprunter deux voies métaboliques menant soit à une dégradation en substrat énergétique ou à la biosynthèse d'un précurseur de l'acide colonique après transformation par des enzymes codés par *cpsG* et *cpsB.* Les mutations sur les gènes *cpsG* et *cpsB* prédominantes chez les souches des souris nourries avec le régime alimentaire standard

suggèrent théoriquement un détournement de la ressource mannose vers la dégradation en substrat énergétique dans le régime alimentaire standard.

Nous avons mis en évidence d'autres convergences chez *E. coli* HS différentes en fonction du régime alimentaires des souris. Des mutations observées spécifiquement chez les souches issues des souris nourries par un régime alimentaire riche en gras et en sucre, convergent vers l'activation constitutive du système à deux composantes *cpx* garantissant l'homéostasie de la membrane bactérienne et répondant au stress environnemental (Vogt and Raivio 2012). Cette observation dans notre étude suggère qu'un élément de ce milieu riche en gras et en sucre semble être délétère à la membrane bactérienne (Raivio and Silhavy 1997). Nous avons aussi observé chez les souches des souris soumises à un régime riche en gras et en sucre, une convergence mutationnelle menant à l'inactivation théorique des gènes impliqués dans le métabolisme de trois acides aminés, la cystéine, la thréonine et la sérine. Fait intéressant, les gènes mutés sont impliqués dans l'absorption de ces acides aminés dans des conditions anaérobies (Loddeke *et al.* 2017; Ganduri *et al.* 1993). Cette observation rejoint une autre convergence chez les souches issues des mêmes souris, dans l'inactivation des gènes *cydA* et *cydB* codant pour les deux sous-unités du cytochrome *bd*-I, responsables de la production d'énergie dans un milieu pauvre en oxygène, alors que les gènes codant pour le cytochrome *bo* responsable de la production d'énergie à un haut niveau d'oxygène reste intact. Cela peut suggérer une tendance à l'adaptation du métabolisme de *E. coli* vers des conditions plus aérobies surtout dans le régime riche en gras et en sucre.

Les résultats très préliminaires de cette étude nous ont permis de distinguer deux profils d'adaptation en fonction de la souche combinée à l'alimentation. Il semblerait que l'intensité de l'adaptation dépend de la souche et probablement de son fond génétique, et que les cibles de l'adaptation dépendent du régime alimentaire.

Les **perspectives** concernent, en premier lieu, les suites à donner à notre troisième étude. L'analyse génétique nous a permis de poser des hypothèses sur les forces à l'œuvre dans le tube digestif des souris traitées à la streptomycine et sur les stratégies adoptées par *E. coli* pour s'adapter. L'expression de ces mutations nécessite dans un premier temps d'être testées phénotypiquement *in vitro*. Nous avons prévu de tester la valeur sélective des différents mutants sur milieu minimum avec différentes gammes de sucres en condition aérobie ou anaérobie.

E. coli s'intègre dans un système complexe avec le reste du microbiote digestif et l'épithélium digestif. Afin de caractériser au maximum le milieu d'adaptation de *E. coli*, nous devons réaliser des études du microbiote digestif par l'étude de l'ARN16S et des études de l'épithélium digestif par immunohistochimie.

Alors que le profil adaptatif de *E. coli* 536 et plus généralement du groupe phylogénétique B2 commence à être bien caractérisé dans le tube digestif de souris traitées à la streptomycine et dans des conditions plus naturelles, nous pourrions réaliser d'autres expérimentations *in vivo* pour caractériser le profil adaptatif de *E. coli* HS dans son milieu naturel en l'absence de streptomycine et peut être différencier un profil adaptatif spécifique du groupe phylogénétique.

BIBLIOGRAPHIE

Al Nabhani, Ziad, Sophie Dulauroy, Rute Marques, Clara Cousu, Shahed Al Bounny, François Déjardin, Tim Sparwasser, Marion Bérard, Nadine Cerf-Bensussan, et Gérard Eberl. 2019a. « A Weaning Reaction to Microbiota Is Required for Resistance to Immunopathologies in the Adult ». *Immunity* 50 (5): 1276-1288.e5. https://doi.org/10.1016/j.immuni.2019.02.014.

Al Nabhani. 2019b. « Excess calorie intake early in life increases susceptibility to colitis in the adult ».

Armstrong, Julie, John J. Reilly, et Child Health Information Team. 2002. « Breastfeeding and Lowering the Risk of Childhood Obesity ». *Lancet (London, England)* 359 (9322): 2003‑4. https://doi.org/10.1016/S0140-6736(02)08837-2.

Arthur, Janelle C., Ernesto Perez-Chanona, Marcus Mühlbauer, Sarah Tomkovich, Joshua M. Uronis, Ting-Jia Fan, Barry J. Campbell, *et al.* 2012. « Intestinal Inflammation Targets Cancer-Inducing Activity of the Microbiota ». *Science (New York, N.Y.)* 338 (6103): 120‑23. https://doi.org/10.1126/science.1224820.

Bäckhed, Fredrik, Jill K. Manchester, Clay F. Semenkovich, et Jeffrey I. Gordon. 2007. « Mechanisms Underlying the Resistance to Diet-Induced Obesity in Germ-Free Mice ». *Proceedings of the National Academy of Sciences of the United States of America* 104 (3): 979‑84. https://doi.org/10.1073/pnas.0605374104.

Bager, P., J. Wohlfahrt, et T. Westergaard. 2008. « Caesarean Delivery and Risk of Atopy and Allergic Disease: Meta-Analyses ». *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology* 38 (4): 634‑42. https://doi.org/10.1111/j.1365-2222.2008.02939.x.

Barrick, Jeffrey E., et Richard E. Lenski. 2013. « Genome Dynamics during Experimental Evolution ». *Nature Reviews. Genetics* 14 (12): 827‑39. https://doi.org/10.1038/nrg3564.

Barrick, Jeffrey E., Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E. Lenski, et Jihyun F. Kim. 2009. « Genome Evolution and Adaptation in a Long-Term Experiment with Escherichia Coli ». *Nature* 461 (7268): 1243‑47. https://doi.org/10.1038/nature08480.

Barroso-Batista, João, Ana Sousa, Marta Lourenço, Marie-Louise Bergman, Daniel Sobral, Jocelyne Demengeot, Karina B. Xavier, et Isabel Gordo. 2014. « The First Steps of Adaptation of Escherichia Coli to the Gut Are Dominated by Soft Sweeps ». *PLoS Genetics* 10 (3): e1004182. https://doi.org/10.1371/journal.pgen.1004182.

Bentley, R., et R. Meganathan. 1982. « Biosynthesis of Vitamin K (Menaquinone) in Bacteria ». *Microbiological Reviews* 46 (3): 241‑80.

Bergthorsson, U., et H. Ochman. 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.

Bertin, Yolande, J. P. Girardeau, F. Chaucheyras-Durand, Bernard Lyan, Estelle Pujos-Guillot, Josée Harel, et Christine Martin. 2011. « Enterohaemorrhagic Escherichia Coli Gains a Competitive Advantage by Using Ethanolamine as a Nitrogen Source in the Bovine Intestinal Content ». *Environmental Microbiology* 13 (2): 365‑77. https://doi.org/10.1111/j.1462-2920.2010.02334.x.

Bibbò, Stefano, Gianluca Ianiro, Antonio Gasbarrini, et Giovanni Cammarota. 2017. « Fecal Microbiota Transplantation: Past, Present and Future Perspectives ». *Minerva Gastroenterologica E Dietologica* 63 (4): 420‑30. https://doi.org/10.23736/S1121-421X.17.02374-1.

Bingen, E. 1998. « [Escherichia coli virulence factors in pediatric urinary tract infections] ». *Archives De Pediatrie: Organe Officiel De La Societe Francaise De Pediatrie* 5 Suppl 3: 279S-281S. https://doi.org/10.1016/s0929-693x(98)80149-8.

Bingen-Bidois, Martine, Olivier Clermont, Stéphane Bonacorsi, Mustapha Terki, Naïma Brahimi, Chawki Loukil, Dominique Barraud, et Edouard Bingen. 2002. « Phylogenetic Analysis and Prevalence of Urosepsis Strains of Escherichia Coli Bearing Pathogenicity Island-like Domains ». *Infection and Immunity* 70 (6): 3216‑26. https://doi.org/10.1128/iai.70.6.3216-3226.2002.

Bloomquist, Erik W., Philippe Lemey, et Marc A. Suchard. 2010. « Three Roads Diverged? Routes to Phylogeographic Inference ». *Trends in Ecology & Evolution* 25 (11): 626 - 32. https://doi.org/10.1016/j.tree.2010.08.010.

Bonacorsi, Stéphane, Olivier Clermont, Véronique Houdouin, Christophe Cordevant, Naima Brahimi, Armelle Marecat, Colin Tinsley, Xavier Nassif, Marc Lange, et Edouard Bingen. 2003. « Molecular Analysis and Experimental Virulence of French and North American Escherichia Coli Neonatal Meningitis Isolates: Identification of a New Virulent Clone ». *The Journal of Infectious Diseases* 187 (12): 1895‑1906. https://doi.org/10.1086/375347.

Bonnet, R. 2004. « Growing Group of Extended-Spectrum Beta-Lactamases: The CTX-M Enzymes ». *Antimicrobial Agents and Chemotherapy* 48 (1): 1‑14. https://doi.org/10.1128/aac.48.1.1-14.2004.

Bush, Karen, et George A. Jacoby. 2010. « Updated Functional Classification of Beta-Lactamases ». *Antimicrobial Agents and Chemotherapy* 54 (3): 969‑76. https://doi.org/10.1128/AAC.01009-09.

Cani, P. D., S. Possemiers, T. Van de Wiele, Y. Guiot, A. Everard, O. Rottier, L. Geurts, *et al.* 2009. « Changes in Gut Microbiota Control Inflammation in Obese Mice through a Mechanism Involving GLP-2-Driven Improvement of Gut Permeability ». *Gut* 58 (8): 1091 - 1103. https://doi.org/10.1136/gut.2008.165886.

Cani, Patrice D., Rodrigo Bibiloni, Claude Knauf, Aurélie Waget, Audrey M. Neyrinck, Nathalie M. Delzenne, et Rémy Burcelin. 2008. « Changes in Gut Microbiota Control Metabolic Endotoxemia-Induced Inflammation in High-Fat Diet-Induced Obesity and Diabetes in Mice ». *Diabetes* 57 (6): 1470‑81. https://doi.org/10.2337/db07-1403.

Cannatelli, Antonio, Vincenzo Di Pilato, Tommaso Giani, Fabio Arena, Simone Ambretti, Paolo Gaibani, Marco Maria D'Andrea, et Gian Maria Rossolini. 2014. « In Vivo Evolution to Colistin Resistance by PmrB Sensor Kinase Mutation in KPC-Producing Klebsiella Pneumoniae Is Associated with Low-Dosage Colistin Treatment ». *Antimicrobial Agents and Chemotherapy* 58 (8): 4399‑4403. https://doi.org/10.1128/AAC.02555-14.

Cantón, Rafael, José María González-Alba, et Juan Carlos Galán. 2012. « CTX-M Enzymes: Origin and Diffusion ». *Frontiers in Microbiology* 3: 110. https://doi.org/10.3389/fmicb.2012.00110.

Cario, E. 2005. « Bacterial Interactions with Cells of the Intestinal Mucosa: Toll-like Receptors and NOD2 ». *Gut* 54 (8): 1182‑93. https://doi.org/10.1136/gut.2004.062794.

Charlesworth, Brian. 2009. « Fundamental Concepts in Genetics: Effective Population Size and Patterns of Molecular Evolution and Variation ». *Nature Reviews. Genetics* 10 (3): 195 - 205. https://doi.org/10.1038/nrg2526.

Chattopadhyay, Sujay, Scott J. Weissman, Vladimir N. Minin, Thomas A. Russo, Daniel E. Dykhuizen, et Evgeni V. Sokurenko. 2009. « High Frequency of Hotspot Mutations in Core Genes of Escherichia Coli Due to Short-Term Positive Selection ». *Proceedings of the National Academy of Sciences of the United States of America* 106 (30): 12412 - 17. https://doi.org/10.1073/pnas.0906217106.

Chong, Yong, Shinji Shimoda, et Nobuyuki Shimono. 2018. « Current Epidemiology, Genetic Evolution and Clinical Impact of Extended-Spectrum β-Lactamase-Producing Escherichia Coli and Klebsiella Pneumoniae ». *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 61: 185 - 88. https://doi.org/10.1016/j.meegid.2018.04.005.

Clermont, Olivier, Julia K. Christenson, Erick Denamur, et David M. Gordon. 2013. « The Clermont Escherichia Coli Phylo-Typing Method Revisited: Improvement of Specificity and Detection of New Phylo-Groups ». *Environmental Microbiology Reports* 5 (1): 58 - 65. https://doi.org/10.1111/1758-2229.12019.

Clermont, Olivier, David Gordon, et Erick Denamur. 2015. « Guide to the Various Phylogenetic Classification Schemes for Escherichia Coli and the Correspondence among Schemes ». *Microbiology (Reading, England)* 161 (Pt 5): 980‑88. https://doi.org/10.1099/mic.0.000063.

Couce, Alejandro, et Olivier A. Tenaillon. 2015. « The Rule of Declining Adaptability in Microbial Evolution Experiments ». *Frontiers in Genetics* 6: 99. https://doi.org/10.3389/fgene.2015.00099.

Cunliffe, R. N., et Y. R. Mahida. 2004. « Expression and Regulation of Antimicrobial Peptides in the Gastrointestinal Tract ». *Journal of Leukocyte Biology* 75 (1): 49 - 58. https://doi.org/10.1189/jlb.0503249.

Datsenko, K. A., et B. L. Wanner. 2000. « One-Step Inactivation of Chromosomal Genes in Escherichia Coli K-12 Using PCR Products ». *Proceedings of the National Academy of Sciences of the United States of America* 97 (12): 6640‑45. https://doi.org/10.1073/pnas.120163297.

De Paepe, Marianne, Valérie Gaboriau-Routhiau, Dominique Rainteau, Sabine Rakotobe, François Taddei, et Nadine Cerf-Bensussan. 2011a. « Trade-off between Bile Resistance and Nutritional Competence Drives Escherichia Coli Diversification in the Mouse Gut ». *PLoS Genetics* 7 (6): e1002107. https://doi.org/10.1371/journal.pgen.1002107.

Diard, Médéric, Louis Garry, Marjorie Selva, Thomas Mosser, Erick Denamur, et Ivan Matic. 2010a. « Pathogenicity-Associated Islands in Extraintestinal Pathogenic Escherichia Coli Are Fitness Elements Involved in Intestinal Colonization ». *Journal of Bacteriology* 192 (19): 4885‑93. https://doi.org/10.1128/JB.00804-10.

Didelot, Xavier, A. Sarah Walker, Tim E. Peto, Derrick W. Crook, et Daniel J. Wilson. 2016a. « Within-Host Evolution of Bacterial Pathogens ». *Nature Reviews. Microbiology* 14 (3): 150 - 62. https://doi.org/10.1038/nrmicro.2015.13.

Diekema, D. J., S. E. Beekmann, K. C. Chapin, K. A. Morel, E. Munson, et G. V. Doern. 2003. « Epidemiology and Outcome of Nosocomial and Community-Onset Bloodstream Infection ». *Journal of Clinical Microbiology* 41 (8): 3655 - 60. https://doi.org/10.1128/jcm.41.8.3655-3660.2003.

Drummond, Alexei J., et Andrew Rambaut. 2007. « BEAST: Bayesian Evolutionary Analysis by Sampling Trees ». *BMC Evolutionary Biology* 7 (novembre): 214. https://doi.org/10.1186/1471-2148-7-214.

Ducluzeau, R. 1993. « [Development, equilibrium and role of microbial flora in the newborn] ». *Annales De Pediatrie* 40 (1): 13‑22.

Dunne, Karl A., Roy R. Chaudhuri, Amanda E. Rossiter, Irene Beriotto, Douglas F. Browning, Derrick Squire, Adam F. Cunningham, Jeffrey A. Cole, Nicholas Loman, et Ian R. Henderson. 2017. « Sequencing a Piece of History: Complete Genome Sequence of the Original Escherichia Coli Strain ». *Microbial Genomics* 3 (3): mgen000106. https://doi.org/10.1099/mgen.0.000106.

Eckburg, Paul B., Elisabeth M. Bik, Charles N. Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael Sargent, Steven R. Gill, Karen E. Nelson, et David A. Relman. 2005. « Diversity of the Human Intestinal Microbial Flora ». *Science (New York, N.Y.)* 308 (5728): 1635 - 38. https://doi.org/10.1126/science.1110591.

Elkharrat, D., L. Arrouy, F. Benhamou, A. Dray, J. Grenet, et A. Le Corre. 2007. « Épidémiologie de l'infection urinaire communautaire de l'adulte en France ». In *Les infections urinaires*, édité par Bernard Lobel et Claude-James Soussy, 1 - 20. Paris: Springer Paris. https://doi.org/10.1007/978-2-287-48617-3_1.

Ellegren, Hans, et Nicolas Galtier. 2016. « Determinants of Genetic Diversity ». *Nature Reviews. Genetics* 17 (7): 422‑33. https://doi.org/10.1038/nrg.2016.58.

Enright, M. C., et B. G. Spratt. 1999. « Multilocus Sequence Typing ». *Trends in Microbiology* 7 (12): 482 ‑87.

Escobar-Páramo, Patricia, Karine Grenet, Arnaud Le Menac'h, Luc Rode, Emmanuelle Salgado, Christine Amorin, Stéphanie Gouriou, *et al.* 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.

Faria, Nuno R., Andrew Rambaut, Marc A. Suchard, Guy Baele, Trevor Bedford, Melissa J. Ward, Andrew J. Tatem, *et al.* 2014. « HIV Epidemiology. The Early Spread and Epidemic Ignition of

HIV-1 in Human Populations ». *Science (New York, N.Y.)* 346 (6205): 56 - 61. https://doi.org/10.1126/science.1256739.

Ferrarese, R., E. R. Ceresola, A. Preti, et F. Canducci. 2018. « Probiotics, Prebiotics and Synbiotics for Weight Loss and Metabolic Syndrome in the Microbiome Era ». *European Review for Medical and Pharmacological Sciences* 22 (21): 7588 - 7605. https://doi.org/10.26355/eurrev_201811_16301.

Fihman, Vincent, Jonathan Messika, David Hajage, Véronique Tournier, Stéphane Gaudry, Fatma Magdoud, Guilène Barnaud, *et al.* 2015. « Five-Year Trends for Ventilator-Associated Pneumonia: Correlation between Microbiological Findings and Antimicrobial Drug Consumption ». *International Journal of Antimicrobial Agents* 46 (5): 518 - 25. https://doi.org/10.1016/j.ijantimicag.2015.07.010.

Fong, Stephen S., Andrew R. Joyce, et Bernhard Ø Palsson. 2005. « Parallel Adaptive Evolution Cultures of Escherichia Coli Lead to Convergent Growth Phenotypes with Different Gene Expression States ». *Genome Research* 15 (10): 1365‑72. https://doi.org/10.1101/gr.3832305.

Foster, Patricia L., Heewook Lee, Ellen Popodi, Jesse P. Townes, et Haixu Tang. 2015a. « Determinants of Spontaneous Mutation in the Bacterium Escherichia Coli as Revealed by Whole-Genome Sequencing ». *Proceedings of the National Academy of Sciences of the United States of America* 112 (44): E5990-5999. https://doi.org/10.1073/pnas.1512136112.

Foxman, Betsy. 2003. « Epidemiology of Urinary Tract Infections: Incidence, Morbidity, and Economic Costs ». *Disease-a-Month: DM* 49 (2): 53‑70. https://doi.org/10.1067/mda.2003.7.

Fraher, Marianne H., Paul W. O'Toole, et Eamonn M. M. Quigley. 2012. « Techniques Used to Characterize the Gut Microbiota: A Guide for the Clinician ». *Nature Reviews. Gastroenterology & Hepatology* 9 (6): 312‑22. https://doi.org/10.1038/nrgastro.2012.44.

Friedman, Noah J., et Robert S. Zeiger. 2005. « The Role of Breast-Feeding in the Development of Allergies and Asthma ». *The Journal of Allergy and Clinical Immunology* 115 (6): 1238‑48. https://doi.org/10.1016/j.jaci.2005.01.069.

Ganduri, Y. L., S. R. Sadda, M. W. Datta, R. K. Jambukeswaran, et P. Datta. 1993. « TdcA, a Transcriptional Activator of the TdcABC Operon of Escherichia Coli, Is a Member of the LysR Family of Proteins ». *Molecular & General Genetics: MGG* 240 (3): 395 - 402. https://doi.org/10.1007/bf00280391.

Gao, Wei, Kyra Chua, John K. Davies, Hayley J. Newton, Torsten Seemann, Paul F. Harrison, Natasha E. Holmes, *et al.* 2010. « Two Novel Point Mutations in Clinical Staphylococcus Aureus Reduce Linezolid Susceptibility and Switch on the Stringent Response to Promote Persistent Infection ». *PLoS Pathogens* 6 (6): e1000944. https://doi.org/10.1371/journal.ppat.1000944.

Gauger, Eric J., Mary P. Leatham, Regino Mercado-Lubo, David C. Laux, Tyrrell Conway, et Paul S. Cohen. 2007. « Role of Motility and the FlhDC Operon in Escherichia Coli MG1655 Colonization of the Mouse Intestine ». *Infection and Immunity* 75 (7): 3315 - 24. https://doi.org/10.1128/IAI.00052-07.

Gauzit, Rémy, Yves Péan, Xavier Barth, Frédéric Mistretta, Olivier Lalaude, et Top Study Team. 2009. « Epidemiology, Management, and Prognosis of Secondary Non-Postoperative Peritonitis: A French Prospective Observational Multicenter Study ». *Surgical Infections* 10 (2): 119‑27. https://doi.org/10.1089/sur.2007.092.

Gerrish, P. J., et R. E. Lenski. 1998. « The Fate of Competing Beneficial Mutations in an Asexual Population ». *Genetica* 102‑103 (1‑6): 127‑44.

Golubchik, Tanya, Elizabeth M. Batty, Ruth R. Miller, Helen Farr, Bernadette C. Young, Hanna Larner-Svensson, Rowena Fung, *et al.* 2013. « Within-Host Evolution of Staphylococcus Aureus during Asymptomatic Carriage ». *PloS One* 8 (5): e61319. https://doi.org/10.1371/journal.pone.0061319.

Good, Benjamin H., et Michael M. Desai. 2014. « Deleterious Passengers in Adapting Populations ». *Genetics* 198 (3): 1183‑1208. https://doi.org/10.1534/genetics.114.170233.

Gordon, David M., et Ann Cowling. 2003. « The Distribution and Genetic Structure of Escherichia Coli in Australian Vertebrates: Host and Geographic Effects ». *Microbiology (Reading, England)* 149 (Pt 12): 3575‑86. https://doi.org/10.1099/mic.0.26486-0.

Gregory, Katherine E., Buck S. Samuel, Pearl Houghteling, Guru Shan, Frederick M. Ausubel, Ruslan I. Sadreyev, et W. Allan Walker. 2016. « Influence of Maternal Breast Milk Ingestion on Acquisition of the Intestinal Microbiome in Preterm Infants ». *Microbiome* 4 (1): 68. https://doi.org/10.1186/s40168-016-0214-x.

Guo, Cong, Ian C. McDowell, Michael Nodzenski, Denise M. Scholtens, Andrew S. Allen, William L. Lowe, et Timothy E. Reddy. 2017. « Transversions Have Larger Regulatory Effects than Transitions ». *BMC Genomics* 18 (1): 394. https://doi.org/10.1186/s12864-017-3785-4.

Hamet, Maël, Arnaud Pavon, Frédéric Dalle, André Pechinot, Sébastien Prin, Jean-Pierre Quenot, et Pierre-Emmanuel Charles. 2012. « Candida Spp. Airway Colonization Could Promote Antibiotic-Resistant Bacteria Selection in Patients with Suspected Ventilator-Associated Pneumonia ». *Intensive Care Medicine* 38 (8): 1272‑79. https://doi.org/10.1007/s00134-012-2584-2.

Hatfield, G. Wesley, et Craig J. Benham. 2002. « DNA Topology-Mediated Control of Global Gene Expression in Escherichia Coli ». *Annual Review of Genetics* 36: 175 ‑ 203. https://doi.org/10.1146/annurev.genet.36.032902.111815.

Hershberg, Ruth, et Dmitri A. Petrov. 2008. « Selection on Codon Bias ». *Annual Review of Genetics* 42: 287‑99. https://doi.org/10.1146/annurev.genet.42.110807.091442.

Hill, Cian J., Denise B. Lynch, Kiera Murphy, Marynka Ulaszewska, Ian B. Jeffery, Carol Anne O'Shea, Claire Watkins, *et al.* 2017. « Evolution of Gut Microbiota Composition from Birth to 24 Weeks in the INFANTMET Cohort ». *Microbiome* 5 (1): 4. https://doi.org/10.1186/s40168-016-0213-y.

Hooper, Lora V., Dan R. Littman, et Andrew J. Macpherson. 2012. « Interactions between the Microbiota and the Immune System ». *Science (New York, N.Y.)* 336 (6086): 1268 ‑ 73. https://doi.org/10.1126/science.1223490.

Hoorn, C., F. P. Wesselingh, H. ter Steege, M. A. Bermudez, A. Mora, J. Sevink, I. Sanmartín, *et al.* 2010. « Amazonia through Time: Andean Uplift, Climate Change, Landscape Evolution, and Biodiversity ». *Science (New York, N.Y.)* 330 (6006): 927 ‑ 31. https://doi.org/10.1126/science.1194585.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, et J. P. Bollback. 2001. « Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology ». *Science (New York, N.Y.)* 294 (5550): 2310‑14. https://doi.org/10.1126/science.1065889.

Jakobsson, Hedvig E., Ana M. Rodríguez-Piñeiro, André Schütte, Anna Ermund, Preben Boysen, Mats Bemark, Felix Sommer, Fredrik Bäckhed, Gunnar C. Hansson, et Malin E. V. Johansson. 2015. « The Composition of the Gut Microbiota Shapes the Colon Mucus Barrier ». *EMBO Reports* 16 (2): 164‑77. https://doi.org/10.15252/embr.201439263.

Javaloyas, M., D. Garcia-Somoza, et F. Gudiol. 2002. « Epidemiology and Prognosis of Bacteremia: A 10-y Study in a Community Hospital ». *Scandinavian Journal of Infectious Diseases* 34 (6): 436 ‑41. https://doi.org/10.1080/00365540110080629.

Jiang, Cizhong, et Zhongming Zhao. 2006. « Mutational Spectrum in the Recent Human Genome Inferred by Single Nucleotide Polymorphisms ». *Genomics* 88 (5): 527 ‑ 34. https://doi.org/10.1016/j.ygeno.2006.06.003.

Johnson, James R., Krista L. Owens, Connie R. Clabots, Scott J. Weissman, et Steven B. Cannon. 2006. « Phylogenetic Relationships among Clonal Groups of Extraintestinal Pathogenic Escherichia Coli as Assessed by Multi-Locus Sequence Analysis ». *Microbes and Infection* 8 (7): 1702‑13. https://doi.org/10.1016/j.micinf.2006.02.007.

Johnson, James R., et Thomas A. Russo. 2018. « Molecular Epidemiology of Extraintestinal Pathogenic Escherichia Coli ». *EcoSal Plus* 8 (1). https://doi.org/10.1128/ecosalplus.ESP-0004-2017.

Johnston, Calum, Bernard Martin, Gwennaele Fichant, Patrice Polard, et Jean-Pierre Claverys. 2014. « Bacterial Transformation: Distribution, Shared Mechanisms and Divergent Control ». *Nature Reviews. Microbiology* 12 (3): 181‑96. https://doi.org/10.1038/nrmicro3199.

Kamada, Nobuhiko, Sang-Uk Seo, Grace Y. Chen, et Gabriel Núñez. 2013. « Role of the Gut Microbiota in Immunity and Inflammatory Disease ». *Nature Reviews. Immunology* 13 (5): 321‑35. https://doi.org/10.1038/nri3430.

Kasai, Chika, Kazushi Sugimoto, Isao Moritani, Junichiro Tanaka, Yumi Oya, Hidekazu Inoue, Masahiko Tameda, *et al.* 2016. « Comparison of Human Gut Microbiota in Control Subjects and Patients with Colorectal Carcinoma in Adenoma: Terminal Restriction Fragment Length Polymorphism and next-Generation Sequencing Analyses ». *Oncology Reports* 35 (1): 325‑33. https://doi.org/10.3892/or.2015.4398.

Kennemann, Lynn, Xavier Didelot, Toni Aebischer, Stefanie Kuhn, Bernd Drescher, Marcus Droege, Richard Reinhardt, *et al.* 2011. « Helicobacter Pylori Genome Evolution during Human Infection ». *Proceedings of the National Academy of Sciences of the United States of America* 108 (12): 5033‑38. https://doi.org/10.1073/pnas.1018444108.

Kimura, M. 1979. « The Neutral Theory of Molecular Evolution ». *Scientific American* 241 (5): 98‑100, 102, 108 passim.

Kimura, M., et T. Ota. 1969. « The Average Number of Generations until Extinction of an Individual Mutant Gene in a Finite Population ». *Genetics* 63 (3): 701‑9.

Kisiela, Dagmara I., Matthew Radey, Sandip Paul, Stephen Porter, Kseniya Polukhina, Veronika Tchesnokova, Sofiya Shevchenko, *et al.* 2017. « Inactivation of Transcriptional Regulators during Within-Household Evolution of Escherichia Coli ». *Journal of Bacteriology* 199 (13). https://doi.org/10.1128/JB.00036-17.

Kull, I., M. Wickman, G. Lilja, S. L. Nordvall, et G. Pershagen. 2002. « Breast Feeding and Allergic Diseases in Infants-a Prospective Birth Cohort Study ». *Archives of Disease in Childhood* 87 (6): 478‑81. https://doi.org/10.1136/adc.87.6.478.

Lamrabet, Otmane, Jacqueline Plumbridge, Mikaël Martin, Richard E. Lenski, Dominique Schneider, et Thomas Hindré. 2019. « Plasticity of Promoter-Core Sequences Allows Bacteria to Compensate for the Loss of a Key Global Regulatory Gene ». *Molecular Biology and Evolution*, mars. https://doi.org/10.1093/molbev/msz042.

Laupland, K. B., D. B. Gregson, W. W. Flemons, D. Hawkins, T. Ross, et D. L. Church. 2007. « Burden of Community-Onset Bloodstream Infection: A Population-Based Assessment ». *Epidemiology and Infection* 135 (6): 1037‑42. https://doi.org/10.1017/S0950268806007631.

Leatham, Mary P., Sarah J. Stevenson, Eric J. Gauger, Karen A. Krogfelt, Jeremy J. Lins, Traci L. Haddock, Steven M. Autieri, Tyrrell Conway, et Paul S. Cohen. 2005. « Mouse Intestine Selects Nonmotile FlhDC Mutants of Escherichia Coli MG1655 with Increased Colonizing Ability and Better Utilization of Carbon Sources ». *Infection and Immunity* 73 (12): 8039‑49. https://doi.org/10.1128/IAI.73.12.8039-8049.2005.

Lefort A, Panhard X, Clermont O, Woerther PL, Branger C, Mentré F, Fantin B, Wolff M, Denamur E; COLIBAFI Group. 2011. « Host factors and portal of entry outweigh bacterial determinants to predict the severity of Escherichia coli bacteremia ». *J Clin Microbiol.*, mars 2011.

Lefort, Agnès, Xavière Panhard, Olivier Clermont, Paul-Louis Woerther, Catherine Branger, France Mentré, Bruno Fantin, Michel Wolff, Erick Denamur, et 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): 777‑83. https://doi.org/10.1128/JCM.01902-10.

Lescat, Mathilde, Adrien Launay, Mohamed Ghalayini, Mélanie Magnan, Jérémy Glodt, Coralie Pintard, Sara Dion, Erick Denamur, et Olivier Tenaillon. 2017. « Using Long-Term Experimental Evolution to Uncover the Patterns and Determinants of Molecular Evolution of an Escherichia Coli Natural Isolate in the Streptomycin-Treated Mouse Gut ». *Molecular Ecology* 26 (7): 1802‑17. https://doi.org/10.1111/mec.13851.

Levert, Maxime, Oana Zamfir, Olivier Clermont, Odile Bouvet, Sylvain Lespinats, Marie Claire Hipeaux, Catherine Branger, *et al.* 2010. « Molecular and Evolutionary Bases of Within-Patient Genotypic and Phenotypic Diversity in Escherichia Coli Extraintestinal Infections ». *PLoS Pathogens* 6 (9): e1001125. https://doi.org/10.1371/journal.ppat.1001125.

Ley, Ruth E., Fredrik Bäckhed, Peter Turnbaugh, Catherine A. Lozupone, Robin D. Knight, et Jeffrey I. Gordon. 2005. « Obesity Alters Gut Microbial Ecology ». *Proceedings of the National Academy of Sciences of the United States of America* 102 (31): 11070‑75. https://doi.org/10.1073/pnas.0504978102.

Li, Qinrui, Ying Han, Angel Belle C. Dy, et Randi J. Hagerman. 2017. « The Gut Microbiota and Autism Spectrum Disorders ». *Frontiers in Cellular Neuroscience* 11: 120. https://doi.org/10.3389/fncel.2017.00120.

Lieberman, Tami D., Kelly B. Flett, Idan Yelin, Thomas R. Martin, Alexander J. McAdam, Gregory P. Priebe, et Roy Kishony. 2014. « Genetic Variation of a Bacterial Pathogen within Individuals with Cystic Fibrosis Provides a Record of Selective Pressures ». *Nature Genetics* 46 (1): 82‑87. https://doi.org/10.1038/ng.2848.

Loddeke, Melissa, Barbara Schneider, Tamiko Oguri, Iti Mehta, Zhenyu Xuan, et Larry Reitzer. 2017. « Anaerobic Cysteine Degradation and Potential Metabolic Coordination in Salmonella Enterica and Escherichia Coli ». *Journal of Bacteriology* 199 (16). https://doi.org/10.1128/JB.00117-17.

Loewe, Laurence, et William G. Hill. 2010. « The Population Genetics of Mutations: Good, Bad and Indifferent ». *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365 (1544): 1153‑67. https://doi.org/10.1098/rstb.2009.0317.

Lourenço, Marta, Ricardo S. Ramiro, Daniela Güleresi, João Barroso-Batista, Karina B. Xavier, Isabel Gordo, et Ana Sousa. 2016. « A Mutational Hotspot and Strong Selection Contribute to the Order of Mutations Selected for during Escherichia Coli Adaptation to the Gut ». *PLoS Genetics* 12 (11): e1006420. https://doi.org/10.1371/journal.pgen.1006420.

Lozupone, Catherine A., Jesse I. Stombaugh, Jeffrey I. Gordon, Janet K. Jansson, et Rob Knight. 2012. « Diversity, Stability and Resilience of the Human Gut Microbiota ». *Nature* 489 (7415): 220‑30. https://doi.org/10.1038/nature11550.

Lukjancenko, Oksana, Trudy M. Wassenaar, et David W. Ussery. 2010. « Comparison of 61 Sequenced Escherichia Coli Genomes ». *Microbial Ecology* 60 (4): 708‑20. https://doi.org/10.1007/s00248-010-9717-3.

Lyons, Daniel M., et Adam S. Lauring. 2017. « Evidence for the Selective Basis of Transition-to-Transversion Substitution Bias in Two RNA Viruses ». *Molecular Biology and Evolution* 34 (12): 3205‑15. https://doi.org/10.1093/molbev/msx251.

Lyytikäinen, O., J. Lumio, H. Sarkkinen, E. Kolho, A. Kostiala, P. Ruutu, et Hospital Infection Surveillance Team. 2002. « Nosocomial Bloodstream Infections in Finnish Hospitals during 1999-2000 ». *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 35 (2): e14-19. https://doi.org/10.1086/340981.

Maddocks, Oliver D. K., Abigail J. Short, Michael S. Donnenberg, Scott Bader, et David J. Harrison. 2009. « Attaching and Effacing Escherichia Coli Downregulate DNA Mismatch Repair Protein in Vitro and Are Associated with Colorectal Adenocarcinomas in Humans ». *PloS One* 4 (5): e5517. https://doi.org/10.1371/journal.pone.0005517.

Maki, Hisaji. 2002. « Origins of Spontaneous Mutations: Specificity and Directionality of Base-Substitution, Frameshift, and Sequence-Substitution Mutageneses ». *Annual Review of Genetics* 36: 279‑303. https://doi.org/10.1146/annurev.genet.36.042602.094806.

Marvig, Rasmus Lykke, Helle Krogh Johansen, Søren Molin, et Lars Jelsbak. 2013a. « Genome Analysis of a Transmissible Lineage of Pseudomonas Aeruginosa Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators ». *PLoS Genetics* 9 (9): e1003741. https://doi.org/10.1371/journal.pgen.1003741.

Massot, Méril, Anne-Sophie Daubié, Olivier Clermont, Françoise Jauréguy, Camille Couffignal, Ghizlane Dahbi, Azucena Mora, *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, England)* 162 (4): 642‑50. https://doi.org/10.1099/mic.0.000242.

Mathers, Amy J., Nicole Stoesser, Anna E. Sheppard, Louise Pankhurst, Adam Giess, Anthony J. Yeh, Xavier Didelot, *et al.* 2015. « Klebsiella Pneumoniae Carbapenemase (KPC)-Producing K. Pneumoniae at a Single Institution: Insights into Endemicity from Whole-Genome Sequencing ». *Antimicrobial Agents and Chemotherapy* 59 (3): 1656 ‑ 63. https://doi.org/10.1128/AAC.04292-14.

Matute, D. R. 2013. « The Role of Founder Effects on the Evolution of Reproductive Isolation ». *Journal of Evolutionary Biology* 26 (11): 2299‑2311. https://doi.org/10.1111/jeb.12246.

McAdams, Harley H., Balaji Srinivasan, et Adam P. Arkin. 2004. « The Evolution of Genetic Regulatory Systems in Bacteria ». *Nature Reviews. Genetics* 5 (3): 169 ‑ 78. https://doi.org/10.1038/nrg1292.

Miura, Kouichi, Ekihiro Seki, Hirohide Ohnishi, et David A. Brenner. 2010. « Role of Toll-like Receptors and Their Downstream Molecules in the Development of Nonalcoholic Fatty Liver Disease ». *Gastroenterology Research and Practice* 2010: 362847. https://doi.org/10.1155/2010/362847.

Miyata, T., S. Miyazawa, et T. Yasunaga. 1979. « Two Types of Amino Acid Substitutions in Protein Evolution ». *Journal of Molecular Evolution* 12 (3): 219‑36.

Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, *et al.* 2002. « Initial Sequencing and Comparative Analysis of the Mouse Genome ». *Nature* 420 (6915): 520 ‑ 62. https://doi.org/10.1038/nature01262.

Mueller, N. T., R. Whyatt, L. Hoepner, S. Oberfield, M. G. Dominguez-Bello, E. M. Widen, A. Hassoun, F. Perera, et A. Rundle. 2015. « Prenatal Exposure to Antibiotics, Cesarean Section and Risk of Childhood Obesity ». *International Journal of Obesity (2005)* 39 (4): 665 ‑ 70. https://doi.org/10.1038/ijo.2014.180.

Nascimento, Fabrícia F., Ana Lazar, Albert N. Menezes, Andressa da Matta Durans, Jânio C. Moreira, Jorge Salazar-Bravo, Paulo S. D'Andrea, et Cibele R. Bonvicino. 2013. « The Role of Historical Barriers in the Diversification Processes in Open Vegetation Formations during the Miocene/Pliocene Using an Ancient Rodent Lineage as a Model ». *PloS One* 8 (4): e61924. https://doi.org/10.1371/journal.pone.0061924.

Nascimento, Fabrícia F., Mario Dos Reis, et Ziheng Yang. 2017. « A Biologist's Guide to Bayesian Phylogenetic Analysis ». *Nature Ecology & Evolution* 1 (10): 1446 ‑ 54. https://doi.org/10.1038/s41559-017-0280-x.

Nataro, J. P., et J. B. Kaper. 1998. « Diarrheagenic Escherichia Coli ». *Clinical Microbiology Reviews* 11 (1): 142‑201.

Natividad, Jane M. M., et Elena F. Verdu. 2013. « Modulation of Intestinal Barrier by Intestinal Microbiota: Pathological and Therapeutic Implications ». *Pharmacological Research* 69 (1): 42 ‑51. https://doi.org/10.1016/j.phrs.2012.10.007.

Nei, Masatoshi, Takeo Maruyama, et Ranajit Chakraborty. 1975. « THE BOTTLENECK EFFECT AND GENETIC VARIABILITY IN POPULATIONS ». *Evolution; International Journal of Organic Evolution* 29 (1): 1‑10. https://doi.org/10.1111/j.1558-5646.1975.tb00807.x.

Niba, A. T., J. D. Beal, A. C. Kudi, et P. H. Brooks. 2009. « Bacterial Fermentation in the Gastrointestinal Tract of Non-Ruminants: Influence of Fermented Feeds and Fermentable Carbohydrates ». *Tropical Animal Health and Production* 41 (7): 1393‑1407. https://doi.org/10.1007/s11250-009-9327-6.

Nicolas-Chanoine, Marie-Hélène, Xavier Bertrand, et Jean-Yves Madec. 2014. « Escherichia Coli ST131, an Intriguing Clonal Group ». *Clinical Microbiology Reviews* 27 (3): 543‑74. https://doi.org/10.1128/CMR.00125-13.

Nougayrède, Jean-Philippe, Stefan Homburg, Frédéric Taieb, Michèle Boury, Elzbieta Brzuszkiewicz, Gerhard Gottschalk, Carmen Buchrieser, Jörg Hacker, Ulrich Dobrindt, et Eric Oswald. 2006. « Escherichia Coli Induces DNA Double-Strand Breaks in Eukaryotic Cells ». *Science (New York, N.Y.)* 313 (5788): 848‑51. https://doi.org/10.1126/science.1127059.

Ochman, H., J. G. Lawrence, et E. A. Groisman. 2000. « Lateral Gene Transfer and the Nature of Bacterial Innovation ». *Nature* 405 (6784): 299‑304. https://doi.org/10.1038/35012500.

Ochman, H., et R. K. Selander. 1984. « Evidence for Clonal Population Structure in Escherichia Coli ». *Proceedings of the National Academy of Sciences of the United States of America* 81 (1): 198‑201. https://doi.org/10.1073/pnas.81.1.198.

Olson-Manning, Carrie F., Maggie R. Wagner, et Thomas Mitchell-Olds. 2012. « Adaptive Evolution: Evaluating Empirical Support for Theoretical Predictions ». *Nature Reviews. Genetics* 13 (12): 867‑77. https://doi.org/10.1038/nrg3322.

Orskov, F., I. Orskov, D. J. Evans, R. B. Sack, D. A. Sack, et T. Wadström. 1976. « Special Escherichia Coli Serotypes among Enterotoxigenic Strains from Diarrhoea in Adults and Children ». *Medical Microbiology and Immunology* 162 (2): 73‑80.

Osawa, S., T. H. Jukes, K. Watanabe, et A. Muto. 1992. « Recent Evidence for Evolution of the Genetic Code ». *Microbiological Reviews* 56 (1): 229‑64.

Otto, S. P., et M. C. Whitlock. 1997. « The Probability of Fixation in Populations of Changing Size ». *Genetics* 146 (2): 723‑33.

Owen, Christopher G., Richard M. Martin, Peter H. Whincup, George Davey Smith, et Derek G. Cook. 2006. « Does Breastfeeding Influence Risk of Type 2 Diabetes in Later Life? A Quantitative Analysis of Published Evidence ». *The American Journal of Clinical Nutrition* 84 (5): 1043‑54. https://doi.org/10.1093/ajcn/84.5.1043.

Paterson, David L., et Robert A. Bonomo. 2005. « Extended-Spectrum Beta-Lactamases: A Clinical Update ». *Clinical Microbiology Reviews* 18 (4): 657‑86. https://doi.org/10.1128/CMR.18.4.657-686.2005.

Pédron, Thierry, Céline Mulet, Catherine Dauga, Lionel Frangeul, Christian Chervaux, Gianfranco Grompone, et Philippe J. Sansonetti. 2012. « A Crypt-Specific Core Microbiota Resides in the Mouse Colon ». *MBio* 3 (3). https://doi.org/10.1128/mBio.00116-12.

Peekhaus, N., et T. Conway. 1998. « What's for Dinner?: Entner-Doudoroff Metabolism in Escherichia Coli ». *Journal of Bacteriology* 180 (14): 3495‑3502.

Perlman, Robert L. 2016. « Mouse Models of Human Disease: An Evolutionary Perspective ». *Evolution, Medicine, and Public Health* 2016 (1): 170‑76. https://doi.org/10.1093/emph/eow014.

Picard, B., J. S. Garcia, S. Gouriou, P. Duriez, N. Brahimi, E. Bingen, J. Elion, et E. Denamur. 1999. « The Link between Phylogeny and Virulence in Escherichia Coli Extraintestinal Infection ». *Infection and Immunity* 67 (2): 546‑53.

Poirel, L., T. Naas, et P. Nordmann. 2008. « Genetic Support of Extended-Spectrum Beta-Lactamases ». *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 14 Suppl 1 (janvier): 75‑81. https://doi.org/10.1111/j.1469-0691.2007.01865.x.

Poulsen, L. K., T. R. Licht, C. Rang, K. A. Krogfelt, et S. Molin. 1995. « Physiological State of Escherichia Coli BJ4 Growing in the Large Intestines of Streptomycin-Treated Mice ». *Journal of Bacteriology* 177 (20): 5840‑45. https://doi.org/10.1128/jb.177.20.5840-5845.1995.

Powell, D. W. 1981. « Barrier Function of Epithelia ». *The American Journal of Physiology* 241 (4): G275-288. https://doi.org/10.1152/ajpgi.1981.241.4.G275.

Price, Erin P., Derek S. Sarovich, Mark Mayo, Apichai Tuanyok, Kevin P. Drees, Mirjam Kaestli, Stephen M. Beckstrom-Sternberg, *et al.* 2013. « Within-Host Evolution of Burkholderia Pseudomallei

over a Twelve-Year Chronic Carriage Infection ». *MBio* 4 (4). https://doi.org/10.1128/mBio.00388-13.

Prum, Richard O., Jacob S. Berv, Alex Dornburg, Daniel J. Field, Jeffrey P. Townsend, Emily Moriarty Lemmon, et Alan R. Lemmon. 2015. « A Comprehensive Phylogeny of Birds (Aves) Using Targeted next-Generation DNA Sequencing ». *Nature* 526 (7574): 569 - 73. https://doi.org/10.1038/nature15697.

Putsep, K., L. G. Axelsson, A. Boman, T. Midtvedt, S. Normark, H. G. Boman, et M. Andersson. 2000. « Germ-Free and Colonized Mice Generate the Same Products from Enteric Prodefensins ». *The Journal of Biological Chemistry* 275 (51): 40478 - 82. https://doi.org/10.1074/jbc.M007816200.

Pybus, Oliver G., Marc A. Suchard, Philippe Lemey, Flavien J. Bernardin, Andrew Rambaut, Forrest W. Crawford, Rebecca R. Gray, *et al.* 2012. « Unifying the Spatial Epidemiology and Molecular Evolution of Emerging Epidemics ». *Proceedings of the National Academy of Sciences of the United States of America* 109 (37): 15066‑71. https://doi.org/10.1073/pnas.1206598109.

Rabot, Sylvie, Mathieu Membrez, Aurélia Bruneau, Philippe Gérard, Taoufiq Harach, Mireille Moser, Frederic Raymond, Robert Mansourian, et Chieh J. Chou. 2010. « Germ-Free C57BL/6J Mice Are Resistant to High-Fat-Diet-Induced Insulin Resistance and Have Altered Cholesterol Metabolism ». *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 24 (12): 4948‑59. https://doi.org/10.1096/fj.10-164921.

Raisch, Jennifer, Emmanuel Buc, Mathilde Bonnet, Pierre Sauvanet, Emilie Vazeille, Amélie de Vallée, Pierre Déchelotte, *et al.* 2014. « Colon Cancer-Associated B2 Escherichia Coli Colonize Gut Mucosa and Promote Cell Proliferation ». *World Journal of Gastroenterology* 20 (21): 6560‑72. https://doi.org/10.3748/wjg.v20.i21.6560.

Raivio, T. L., et T. J. Silhavy. 1997. « Transduction of Envelope Stress in Escherichia Coli by the Cpx Two-Component System ». *Journal of Bacteriology* 179 (24): 7724 - 33. https://doi.org/10.1128/jb.179.24.7724-7733.1997.

Rannala, B., et Z. Yang. 1996. « Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference ». *Journal of Molecular Evolution* 43 (3): 304‑11.

Rau, Martin H., Rasmus Lykke Marvig, Garth D. Ehrlich, Søren Molin, et Lars Jelsbak. 2012. « Deletion and Acquisition of Genomic Content during Early Stage Adaptation of Pseudomonas Aeruginosa to a Human Host Environment ». *Environmental Microbiology* 14 (8): 2200‑2211. https://doi.org/10.1111/j.1462-2920.2012.02795.x.

Reeves, Peter R., Bin Liu, Zhemin Zhou, Dan Li, Dan Guo, Yan Ren, Connie Clabots, Ruiting Lan, James R. Johnson, et Lei Wang. 2011a. « Rates of Mutation and Host Transmission for an Escherichia Coli Clone over 3 Years ». *PloS One* 6 (10): e26907. https://doi.org/10.1371/journal.pone.0026907.

Reis, Mario dos, Yuttapong Thawornwattana, Konstantinos Angelis, Maximilian J. Telford, Philip C. J. Donoghue, et Ziheng Yang. 2015. « Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales ». *Current Biology: CB* 25 (22): 2939 - 50. https://doi.org/10.1016/j.cub.2015.09.066.

Rice, M. C., et S. J. O'Brien. 1980. « Genetic Variance of Laboratory Outbred Swiss Mice ». *Nature* 283 (5743): 157‑61.

Robinson, T. P., D. P. Bu, J. Carrique-Mas, E. M. Fèvre, M. Gilbert, D. Grace, S. I. Hay, *et al.* 2016. « Antibiotic Resistance Is the Quintessential One Health Issue ». *Transactions of the Royal Society of Tropical Medicine and Hygiene* 110 (7): 377 - 80. https://doi.org/10.1093/trstmh/trw048.

Rocha, Eduardo P. C., et Edward J. Feil. 2010. « Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria? » *PLoS Genetics* 6 (9): e1001104. https://doi.org/10.1371/journal.pgen.1001104.

Rodríguez-Beltrán, Jerónimo, Jérôme Tourret, Olivier Tenaillon, Elena López, Emmanuelle Bourdelier, Coloma Costas, Ivan Matic, Erick Denamur, et Jesús Blázquez. 2015. « High Recombinant

Frequency in Extraintestinal Pathogenic Escherichia Coli Strains ». *Molecular Biology and Evolution* 32 (7): 1708‑16. https://doi.org/10.1093/molbev/msv072.

Ruiter, J. de, J. Weel, E. Manusama, W. P. Kingma, et P. H. J. van der Voort. 2009. « The Epidemiology of Intra-Abdominal Flora in Critically Ill Patients with Secondary and Tertiary Abdominal Sepsis ». *Infection* 37 (6): 522‑27. https://doi.org/10.1007/s15010-009-8249-6.

Rutayisire, Erigene, Kun Huang, Yehao Liu, et Fangbiao Tao. 2016. « The Mode of Delivery Affects the Diversity and Colonization Pattern of the Gut Microbiota during the First Year of Infants' Life: A Systematic Review ». *BMC Gastroenterology* 16 (1): 86. https://doi.org/10.1186/s12876-016-0498-0.

Salomón, R. A., et R. N. Farías. 1992. « Microcin 25, a Novel Antimicrobial Peptide Produced by Escherichia Coli ». *Journal of Bacteriology* 174 (22): 7428 ‑ 35. https://doi.org/10.1128/jb.174.22.7428-7435.1992.

Seki, Ekihiro, Samuele De Minicis, Christoph H. Osterreicher, Johannes Kluwe, Yosuke Osawa, David A. Brenner, et Robert F. Schwabe. 2007. « TLR4 Enhances TGF-Beta Signaling and Hepatic Fibrosis ». *Nature Medicine* 13 (11): 1324‑32. https://doi.org/10.1038/nm1663.

Sekirov, Inna, Shannon L. Russell, L. Caetano M. Antunes, et B. Brett Finlay. 2010. « Gut Microbiota in Health and Disease ». *Physiological Reviews* 90 (3): 859 ‑ 904. https://doi.org/10.1152/physrev.00045.2009.

Selander, R. K., D. A. Caugant, H. Ochman, J. M. Musser, M. N. Gilmour, et T. S. Whittam. 1986. « Methods of Multilocus Enzyme Electrophoresis for Bacterial Population Genetics and Systematics ». *Applied and Environmental Microbiology* 51 (5): 873‑84.

Sender, Ron, Shai Fuchs, et Ron Milo. 2016. « Revised Estimates for the Number of Human and Bacteria Cells in the Body ». *PLoS Biology* 14 (8): e1002533. https://doi.org/10.1371/journal.pbio.1002533.

Shin, Na-Ri, Tae Woong Whon, et Jin-Woo Bae. 2015. « Proteobacteria: Microbial Signature of Dysbiosis in Gut Microbiota ». *Trends in Biotechnology* 33 (9): 496 ‑ 503. https://doi.org/10.1016/j.tibtech.2015.06.011.

Sinsheimer, J. S., J. A. Lake, et R. J. Little. 1997. « Inference for Phylogenies under a Hybrid Parsimony Method: Evolutionary-Symmetric Transversion Parsimony ». *Biometrics* 53 (1): 23‑38.

Sniegowski, P. D., P. J. Gerrish, et R. E. Lenski. 1997. « Evolution of High Mutation Rates in Experimental Populations of *E. coli* ». *Nature* 387 (6634): 703‑5. https://doi.org/10.1038/42701.

Stevens, J. P., M. Eames, A. Kent, S. Halket, D. Holt, et D. Harvey. 2003. « Long Term Outcome of Neonatal Meningitis ». *Archives of Disease in Childhood. Fetal and Neonatal Edition* 88 (3): F179-184. https://doi.org/10.1136/fn.88.3.f179.

Swanson, Kelly S., Scot E. Dowd, Jan S. Suchodolski, Ingmar S. Middelbos, Brittany M. Vester, Kathleen A. Barry, Karen E. Nelson, *et al.* 2011. « Phylogenetic and Gene-Centric Metagenomics of the Canine Intestinal Microbiome Reveals Similarities with Humans and Mice ». *The ISME Journal* 5 (4): 639‑49. https://doi.org/10.1038/ismej.2010.162.

Swerdlow, Sarah J., et Roel M. Schaaper. 2014. « Mutagenesis in the LacI Gene Target of *E. coli*: Improved Analysis for LacI(d) and LacO Mutants ». *Mutation Research* 770 (décembre): 79‑84. https://doi.org/10.1016/j.mrfmmm.2014.09.004.

Takao, Keizo, et Tsuyoshi Miyakawa. 2015. « Genomic Responses in Mouse Models Greatly Mimic Human Inflammatory Diseases ». *Proceedings of the National Academy of Sciences of the United States of America* 112 (4): 1167‑72. https://doi.org/10.1073/pnas.1401965111.

Tannock, G. W., R. Fuller, S. L. Smith, et M. A. Hall. 1990. « Plasmid Profiling of Members of the Family Enterobacteriaceae, Lactobacilli, and Bifidobacteria to Study the Transmission of Bacteria from Mother to Infant ». *Journal of Clinical Microbiology* 28 (6): 1225‑28.

Tenaillon, O., F. Taddei, M. Radmian, et I. Matic. 2001. « Second-Order Selection in Bacterial Evolution: Selection Acting on Mutation and Recombination Rates in the Course of Adaptation ». *Research in Microbiology* 152 (1): 11‑16.

Tenaillon, Olivier, Jeffrey E. Barrick, Noah Ribeck, Daniel E. Deatherage, Jeffrey L. Blanchard, Aurko Dasgupta, Gabriel C. Wu, *et al.* 2016a. « Tempo and Mode of Genome Evolution in a 50,000-Generation Experiment ». *Nature* 536 (7615): 165‑70. https://doi.org/10.1038/nature18959.

Tenaillon, Olivier, Alejandra Rodríguez-Verdugo, Rebecca L. Gaut, Pamela McDonald, Albert F. Bennett, Anthony D. Long, et Brandon S. Gaut. 2012. « The Molecular Diversity of Adaptive Convergence ». *Science (New York, N.Y.)* 335 (6067): 457 ‑ 61. https://doi.org/10.1126/science.1212986.

Tenaillon, Olivier, David Skurnik, Bertrand Picard, et Erick Denamur. 2010. « The Population Genetics of Commensal Escherichia Coli ». *Nature Reviews. Microbiology* 8 (3): 207 ‑ 17. https://doi.org/10.1038/nrmicro2298.

Thavagnanam, S., J. Fleming, A. Bromley, M. D. Shields, et C. R. Cardwell. 2008. « A Meta-Analysis of the Association between Caesarean Section and Childhood Asthma ». *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology* 38 (4): 629‑33. https://doi.org/10.1111/j.1365-2222.2007.02780.x.

Touchon, Marie, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, *et al.* 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.

Turrientes, María-Carmen, Fernando Baquero, Bruce R. Levin, José-Luis Martínez, Aida Ripoll, José-María González-Alba, Raquel Tobes, *et al.* 2013. « Normal Mutation Rate Variants Arise in a Mutator (Mut S) Escherichia Coli Population ». *PloS One* 8 (9): e72963. https://doi.org/10.1371/journal.pone.0072963.

Vangay, Pajau, Tonya Ward, Jeffrey S. Gerber, et Dan Knights. 2015. « Antibiotics, Pediatric Dysbiosis, and Disease ». *Cell Host & Microbe* 17 (5): 553 ‑ 64. https://doi.org/10.1016/j.chom.2015.04.006.

Vanhooren, Valerie, et Claude Libert. 2013. « The Mouse as a Model Organism in Aging Research: Usefulness, Pitfalls and Possibilities ». *Ageing Research Reviews* 12 (1): 8 ‑ 21. https://doi.org/10.1016/j.arr.2012.03.010.

Vieira, S. M., O. E. Pagovich, et M. A. Kriegel. 2014. « Diet, Microbiota and Autoimmune Diseases ». *Lupus* 23 (6): 518‑26. https://doi.org/10.1177/0961203313501401.

Vogel, F., et M. Kopun. 1977. « Higher Frequencies of Transitions among Point Mutations ». *Journal of Molecular Evolution* 9 (2): 159‑80.

Vogt, Stefanie L., et Tracy L. Raivio. 2012. « Just Scratching the Surface: An Expanding View of the Cpx Envelope Stress Response ». *FEMS Microbiology Letters* 326 (1): 2 ‑ 11. https://doi.org/10.1111/j.1574-6968.2011.02406.x.

Wakeley, J. 1996. « The Excess of Transitions among Nucleotide Substitutions: New Methods of Estimating Transition Bias Underscore Its Significance ». *Trends in Ecology & Evolution* 11 (4): 158‑62.

Watterson, G. A. 1975. « On the Number of Segregating Sites in Genetical Models without Recombination ». *Theoretical Population Biology* 7 (2): 256‑76.

Weisberg, Stuart P., Daniel McCann, Manisha Desai, Michael Rosenbaum, Rudolph L. Leibel, et Anthony W. Ferrante. 2003. « Obesity Is Associated with Macrophage Accumulation in Adipose Tissue ». *The Journal of Clinical Investigation* 112 (12): 1796 ‑ 1808. https://doi.org/10.1172/JCI19246.

Wielgoss, Sébastien, Jeffrey E. Barrick, Olivier Tenaillon, Stéphane Cruveiller, Béatrice Chane-Woon-Ming, Claudine Médigue, Richard E. Lenski, et Dominique Schneider. 2011a. « Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With Escherichia Coli ». *G3 (Bethesda, Md.)* 1 (3): 183‑86. https://doi.org/10.1534/g3.111.000406.

Wielgoss, Sébastien, Jeffrey E. Barrick, Olivier Tenaillon, Michael J. Wiser, W. James Dittmar, Stéphane Cruveiller, Béatrice Chane-Woon-Ming, Claudine Médigue, Richard E. Lenski, et Dominique Schneider. 2013a. « Mutation Rate Dynamics in a Bacterial Population Reflect Tension

between Adaptation and Genetic Load ». *Proceedings of the National Academy of Sciences of the United States of America* 110 (1): 222‑27. https://doi.org/10.1073/pnas.1219574110.

Wilfert, L., G. Long, H. C. Leggett, P. Schmid-Hempel, R. Butlin, S. J. M. Martin, et M. Boots. 2016. « Deformed Wing Virus Is a Recent Global Epidemic in Honeybees Driven by Varroa Mites ». *Science (New York, N.Y.)* 351 (6273): 594‑97. https://doi.org/10.1126/science.aac9976.

Wiser, Michael J., Noah Ribeck, et Richard E. Lenski. 2013. « Long-Term Dynamics of Adaptation in Asexual Populations ». *Science (New York, N.Y.)* 342 (6164): 1364 ‑ 67. https://doi.org/10.1126/science.1243357.

World Health Organization. 2017. « Global Priority List of Antibiotic-resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics ». http://apps.who.int/medicinedocs/documents/s23171en/s23171en.pdf.

Wu, Gary D., Frederic D. Bushmanc, et James D. Lewis. 2013. « Diet, the Human Gut Microbiota, and IBD ». *Anaerobe* 24 (décembre): 117‑20. https://doi.org/10.1016/j.anaerobe.2013.03.011.

Yang, Zhi-Kai, Hao Luo, Yanming Zhang, Baijing Wang, et Feng Gao. 2019. « Pan-Genomic Analysis Provides Novel Insights into the Association of E.Coli with Human Host and Its Minimal Genome ». *Bioinformatics (Oxford, England)* 35 (12): 1987 ‑ 91. https://doi.org/10.1093/bioinformatics/bty938.

Young, Bernadette C., et Daniel J. Wilson. 2012. « On the Evolution of Virulence during Staphylococcus Aureus Nasal Carriage ». *Virulence* 3 (5): 454‑56. https://doi.org/10.4161/viru.21189.

Zdziarski, Jaroslaw, Elzbieta Brzuszkiewicz, Björn Wullt, Heiko Liesegang, Dvora Biran, Birgit Voigt, Jenny Grönberg-Hernandez, *et al.* 2010. « Host Imprints on Bacterial Genomes--Rapid, Divergent Evolution in Individual Patients ». *PLoS Pathogens* 6 (8): e1001078. https://doi.org/10.1371/journal.ppat.1001078.

Zhang, J. 2000. « Rates of Conservative and Radical Nonsynonymous Nucleotide Substitutions in Mammalian Nuclear Genes ». *Journal of Molecular Evolution* 50 (1): 56‑68.

Zhang, Ming, Kaiji Sun, Yujun Wu, Ying Yang, Patrick Tso, et Zhenlong Wu. 2017. « Interactions between Intestinal Microbiota and Host Immune Response in Inflammatory Bowel Disease ». *Frontiers in Immunology* 8: 942. https://doi.org/10.3389/fimmu.2017.00942.

Zhang, Qiucen, Guillaume Lambert, David Liao, Hyunsung Kim, Kristelle Robin, Chih-kuan Tung, Nader Pourmand, et Robert H. Austin. 2011. « Acceleration of Emergence of Bacterial Antibiotic Resistance in Connected Microenvironments ». *Science (New York, N.Y.)* 333 (6050): 1764‑67. https://doi.org/10.1126/science.1208747.

Zhang, Zhaolei, et Mark Gerstein. 2003. « Patterns of Nucleotide Substitution, Insertion and Deletion in the Human Genome Inferred from Pseudogenes ». *Nucleic Acids Research* 31 (18): 5338‑48. https://doi.org/10.1093/nar/gkg745.

ANNEXE N°1: FORMULES UTILISEES POUR ESTIMER LE COMPORTEMENT DE *E. COLI* ED1A EN ADOPTANT UNE APPROCHE DE GENETIQUE DES POPULATIONS DANS L'ARTICLE 1 DE LA THESE (NIELSEN R, SLATKIN M. 2013. AN INTRODUCTION TO POPULATION GENETICS: THEORY AND APPLICATIONS. SINAUER ASSOCIATES, SUNDERLAND, MASS, CH.3).

Estimation de la diversité par le Thêta de Watterson ($\boldsymbol{\theta_w}$) :

$\boldsymbol{\theta_w} = S/a_n$

S nombre de sites polymorphiques dans le génome

$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$; n étant le nombre de génomes comparés

$\boldsymbol{\theta_w} = 2 . N_e . L . \boldsymbol{\mu_g}$

$\mathbf{N_e}$ Taille de la population efficace

$\boldsymbol{\mu_g}$ Estimation du taux de mutation par base et par génération

L Longueur du génome

Estimation de la taille de la population efficace ($\mathbf{N_e}$):

$\mathbf{N_e} = S/(2 . a_n . L . \boldsymbol{\mu_g})$

$\boldsymbol{\mu_{est}}$ Le taux de mutation annuel est égal au nombre de fixation par an

Estimation du nombre de générations par jour (**G**):

$\mathbf{G} = \boldsymbol{\mu_{est}} / (\boldsymbol{\mu_g} . 365)$

Temps moyen de fixation en générations:

$2\mathbf{N_e} = \boldsymbol{\theta_w}/(L . \boldsymbol{\mu_g}) = S/(a_n . L . \boldsymbol{\mu_g})$

Temps moyen de fixation en jours:

$2\mathbf{N_e}/\mathbf{G} = \theta_w/(\mathbf{G} . L . \boldsymbol{\mu_g}) = S/(\mathbf{G} . a_n . L . \boldsymbol{\mu_g}) = S /((a_n . \boldsymbol{\mu_{est}}) /365)$

ANNEXE N°2: CARACTERISTIQUES DES SOURIS DES 71 SOUCHES SEQUENCEES DANS L'ARTICLE N°2 DE LA THESE (TABLE S1)

Pour chacun des 71 isolats séquencés, nous avons rapporté la période d'échantillonnage, l'identifiant de souris échantillonné (lignée et nombre de souris), le régime alimentaire, l'heure du décès (spontané ou sacrifice), la densité de *E. coli* par gramme de feces et le nombre de mutation trouvées dans le génome.

Isolate ID	Lineage/Cage	Mouse	Time of sampling for sequencing in day	Diet	Log (UFC of *E. coli* /g of feces) at last day of sampling for quantification	Number of mutations	Time of death or sacrifice in day
L1D106	1	2	106	WD	5.3665315444241	3	435
L1D182	1	2	182	WD	5.816731563172	3	435
L1D435	1	2	435	WD	5.43062609038496	7	435
L2D106	2	4	106	WD	5.542118132661	2	435
L2D435	2	4	435	WD	5.17392519729917	9	435
L2D182	2	6	182	WD	4.8794266879415	2	435
L3D106	3	9	106	WD	5.649161949569	2	435
L3D182	3	9	182	WD	6.12349349573412	2	435
L3D435	3	9	435	WD	7.16749108729376	5	435
L4D106	4	10	106	WD	5.38486859982428	1	435
L4D182	4	10	182	WD	6.48172622429	4	435
L4D435	4	10	435	WD	4.94949135384832	5	435
L5D106	5	13	106	WD	5.67985371388894	6	435
L5D182	5	13	182	WD	7.64975198166584	6	435
L5D435	5	13	435	WD	6.6458915608526	8	435
L6D106	6	16	106	WD	5.2687214643	3	435
L6D182	6	16	182	WD	4.59773861754532	5	435
L6D435	6	16	435	WD	4.43415218132648	4	435
L7D106	7	21	106	WD	7.33724216831842	2	435
L7D182	7	21	182	WD	6.88842391262	3	435
L7D435	7	21	435	WD	7.67026322520084	3	435
L8D106	8	22	106	WD	8.571655622991	0	435
L8D182	8	22	182	WD	7.125363436555	5	435
L8D435	8	22	435	WD	6.73361148999123	6	435
L9D106	9	25	106	WD	6.8757437921719	3	442
L9D182	9	25	182	WD	6.377885688939	4	442
L9D442	9	26	442	WD	6.88605664769316	5	442
L10D106	10	28	106	WD	6.8413475977	2	442
L10D442	10	28	442	WD	7.51144928349956	1	442
L10D182	10	29	182	WD	3.6197887582884	4	442
L11D442	11	31	442	WD	3.97197127639976	5	442
L11D106	11	32	106	WD	6.2915799986529	4	442
L11D182	11	32	182	WD	7.481486612212	4	442
L12D182	12	34	182	WD	6.3496924768687	4	235
L12D210	12	34	210	WD	3.7716852813441	3	235
L12D106	12	36	106	WD	4.34294414714291	2	442
L13D106	13	37	106	CD	7.8916652521117	4	442
L13D442	13	37	442	CD	5.69164905141327	7	442
L13D182	13	38	182	CD	3.33859300760939	1	385
L14D106	14	40	106	CD	2.6733467433	3	442
L14D182	14	40	182	CD	6.2971483596676	4	442
L14D442	14	40	442	CD	3.61367942610595	6	442
L15D106	15	45	106	CD	2.31695296176115	2	442
L15D182	15	45	182	CD	6.2881927795881	2	442
L15D322	15	45	322	CD	5.96772669414448	3	442
L16D106	16	46	106	CD	8.246611539454	3	442
L16D196	16	46	196	CD	1.997419476868	4	442
L16D182	16	47	182	CD	5.35753547975788	5	442
L17D106	17	49	106	CD	4.2134935455473	3	456
L17D182	17	50	182	CD	5.433562578493	1	456
L17D456	17	50	456	CD	5.11294562194904	8	456
L18D28	18	54	28	CD	7.54668165995296	1	456
L18D65	18	54	65	CD	3.9858695895587	4	456
L19D106	19	55	106	CD	5.383489723264	3	456
L19D322	19	55	322	CD	4.7468798535	3	456
L19D182	19	56	182	CD	3.37365963262496	5	350
L20D106	20	58	106	CD	3.2133776293425	3	456
L20D182	20	58	182	CD	3.32513885926218	3	456
L20D322	20	59	322	CD	4.451615818678	1	456
L21D182	21	62	182	CD	3.896683182973	3	435
L21D435	21	62	435	CD	4.07228753809972	3	435
L21D106	21	63	106	CD	3.5718652597121	3	435
L22D106	22	65	106	CD	2.144488443322	2	456
L22D182	22	65	182	CD	4.566415464966	7	456
L22D322	22	65	322	CD	6.5436339668796	6	456
L23D91	23	67	91	CD	4.125625155596	2	456
L23D357	23	67	357	CD	3.648329959473	2	456
L23D182	23	69	182	CD	6.1291118623942	3	456
L24D106	24	71	106	CD	4.38854223433166	3	456
L24D182	24	71	182	CD	4.7774392864353	2	456
L24D210	24	71	210	CD	5.34486156518862	2	456

ANNEXE N°3: LISTE DES SNP TROUVES DANS LES 71 ISOLATS D'*E. COLI* 536 SEQUENCES DANS L'ARTICLE 2 DE LA THESE (TABLE S2)

Gene	Position	Base change ment	Mutation effect	Kind of SNP	Occurrence	Mice ID	Diet	Gene information
atoS →	2,358,237	C→T	A397V (GCG→GTG)	non synonymous	1	L10D182	WD	Sensory histidine kinase in two-component regulatory system with AtoC
clpA →	939,086	G→A	A701T (GCA→ACA)	non synonymous	1	L11D442	WD	ATPase and specificity subunit of ClpA-ClpP ATP-dependent serine protease, chaperone activity
epsE ←	3,203,757	G→A	P321L (CCG→CTG)	non synonymous	1	L11D442	WD	General secretion pathway protein E
irp3 →	1,996,572	C→T	A206V (GCC→GTC)	non synonymous	1	L11D442	WD	Thiazolinyl-S-HMWP1 reductase YbtU (Yersiniabactin siderophore biosynthetic protein)
ubiE →	4,225,199	C→T	R165R (CGC→CGT)	synonymous	1	L11D442	WD	Bifunctional 2-octaprenyl-6-methoxy-1,4-benzoquinone methylase and S-adenosylmethionine:2-DMK methyltransferase
nuoG ←	2,430,886	C→T	A177T (GCG→ACG)	non synonymous	1	L11D106	WD	NADH:ubiquinone oxidoreductase, chain G
uvrD →	4,182,865	G→A	Q705Q (CAG→CAA)	synonymous	1	L11D182	WD	DNA-dependent ATPase I and helicase II
narY ←	1,509,534	C→T	R278H (CGC→CAC)	non synonymous	1	L12D182	WD	Nitrate reductase 2 (NRZ), beta subunit
tolA → / → *tolB*	792,997	G→C	intergenic (+24/-109)	intergenic	1	L12D182	WD	Membrane anchored protein in TolA-TolQ-TolR complex/periplasmic protein
ECP_1141 ←	1,192,779	G→T	A254D (GCC→GAC)	non synonymous	1	L12D210	WD	Recombination protein bet from phage origin
bipA →	4,264,753	C→T	T291M (ACG→ATG)	non synonymous	1	L12D106	WD	GTP-binding protein
glpR ←	3,666,639	C→T	G141D (GGT→GAT)	non synonymous	1	L13D106	CD	DNA-binding transcriptional repressor

ECP_1179 →	1,217,899	T→C	V141A (GTC→GCC)	non synonymous	1	L13D442	CD	Minor tail protein T
ECP_4571 ←	4,777,699	A→G	V483A (GTT→GCT)	non synonymous	1	L13D442	CD	Conserved hypothetical protein
hcr ←	925,232	G→A	A272V (GCC→GTC)	non synonymous	1	L13D442	CD	HCP oxidoreductase, NADH-dependent
pldB →	4,194,355	G→A	D257N (GAT→AAT)	non synonymous	1	L13D442	CD	Lysophospholipase L(2)
kefF →	49,060	G→A	G91S (GGC→AGC)	non synonymous	1	L13D182	CD	Glutathione-regulated potassium-efflux system ancillary protein KefF
ecpA ←	377,733	C→T	G77D (GGT→GAT)	non synonymous	2	L14D106 , L14D182	CD	*E. coli* common pilus (ECP)
ydiO →	1,692,914	C→T	S315S (AGC→AGT)	synonymous	2	L14D182 , L14D106	CD	Putative acyl-CoA dehydrogenase
agp →	1,060,908	C→T	A249V (GCC→GTC)	non synonymous	1	L14D442	CD	Glucose-1-phosphatase/inositol phosphatase
glgB ←	3,688,552	G→A	R255R (CGC→CGT)	synonymous	1	L14D442	CD	1,4-alpha-glucan branching enzyme
ydjR ←	1,737,328	T→G	A88A (GCA→GCC)	synonymous	1	L14D442	CD	Conserved hypothetical protein
tyrP →	1,899,402	T→A	F295I (TTT→ATT)	non synonymous	1	L15D182	CD	Tyrosine transporter
ydhV ←	1,666,775	C→T	D669N (GAC→AAC)	non synonymous	1	L15D322	CD	Putative ferredoxin:oxidoreductase subunit
ECP_0699 →	740,587	C→T	Q46* (CAG→TAG)	stop codon	1	L16D196	CD	Putative transcriptional regulator, DeoR-family
ECP_4610 ←	4,816,916	A→G	V182A (GTT→GCT)	non synonymous	1	L16D196	CD	Conserved hypothetical protein, putative Transcriptional regulator
malX →	1,619,358	G→A	G253D (GGC→GAC)	non synonymous	1	L16D182	CD	Fused maltose and glucose-specific PTS enzymes: IIB and IIC components
fliR →	1,928,497	C→T	A231V (GCA→GTA)	non synonymous	1	L17D106	CD	Flagellar export pore protein
yjjN → / ← *mdoB*	4,895,229	T→C	intergenic (+98/+42)	intergenic	1	L17D106	CD	Putative oxidoreductase, Zn-dependent and NAD(P)-binding/phosphoglycerol transferase I
irp7 ←	1,975,577	G→A	R531W (CGG→TGG)	non synonymous	1	L17D182	CD	Permease and ATP-binding protein of yersiniabactin-iron ABC transporter YbtQ

atoC →	2,359,900	G→A	D344N (GAC→AAC)	non synonymous	1	L17D456	CD	Fused response regulator of ato opeon, in two-component system with AtoS: response regulator ; sigma54 interaction protein
fimA →	4,848,451	C→G	T71S (ACC→AGC)	non synonymous	1	L17D456	CD	Major type 1 subunit fimbrin (pilin)
fucA ←	2,926,472	C→T	G210R (GGG→AGG)	non synonymous	1	L17D456	CD	L-fuculose-1-phosphate aldolase
sixA ←	2,488,401	C→T	P121P (CCG→CCA)	synonymous	1	L17D456	CD	Phosphohistidine phosphatase
yddG ←	1,516,096	G→A	V215V (GTC→GTT)	synonymous	1	L17D456	CD	Metabolite and drug efflux pump
dadA →	1,259,229	G→T	P363P (CCG→CCT)	synonymous	1	L18D65	CD	D-amino acid dehydrogenase
ECP_0287 ←	307,221	G→A	F35F (TTC→TTT)	synonymous	1	L18D65	CD	Conserved hypothetical protein
ECP_4629 →	4,833,406	C→T	R137W (CGG→TGG)	non synonymous	1	L19D106	CD	Conserved hypothetical protein
ECP_0787 →	823,630	C→T	L294F (CTC→TTC)	non synonymous	1	L19D182	CD	Conserved hypothetical protein
hsdM ←	4,877,187	G→A	A202V (GCC→GTC)	non synonymous	1	L19D182	CD	Type I restriction enzyme EcoEI M protein (M.EcoEI)
rihA ←	709,873	C→T	G155R (GGG→AGG)	non synonymous	2	L19D182 , L19D322	CD	Ribonucleoside hydrolase 1
yeeZ ← / ← *yoeB*	2,127,384	C→T	intergenic (-8/+75)	intergenic	1	L19D322	CD	Putative epimerase, with NAD(P)-binding Rossmann-fold domain/toxin of the YoeB-YefM toxin-antitoxin system
ECP_0227 ←	248,978	C→A	P231P (CCG→CCT)	synonymous	1	L1D106	WD	Conserved hypothetical protein Aec29, contains a TPR-like_helical domain
lacI ←	446,204	G→T	A141D (GCT→GAT)	non synonymous	2	L1D106, L1D182	WD	DNA-binding transcriptional repressor
murF →	96,469	G→T	A168S (GCC→TCC)	non synonymous	1	L1D435	WD	UDP-N-acetylmuramoyl-tripeptide:D-alanyl-D-alanine ligase
tatD →	4,229,941	G→A	E184K (GAA→AAA)	non synonymous	1	L1D435	WD	DNase, magnesium-dependent
ushA →	573,205	C→T	A392V (GCA→GTA)	non synonymous	1	L1D435	WD	Bifunctional UDP-sugar hydrolase and 5'-nucleotidase
ytfM →	4,680,249	C→G	Q207E (CAA→GAA)	non synonymous	1	L1D435	WD	Conserved hypothetical protein; putative outer membrane protein and surface antigen

glpC →	2,393,197	G→T	V324V (GTG→GTT)	synonymous	3	L20D106, L20D182, L20D322	CD	Sn-glycerol-3-phosphate dehydrogenase (anaerobic), small subunit
ECP_0306 ← / → *ECP_0307*	329,774	A→G	intergenic (-145/-393)	intergenic	1	L20D182	CD	IroB protein/conserved hypothetical protein
flhB ←	1,876,940	G→A	L78L (CTC→CTT)	synonymous	1	L20D182	CD	Flagellar export pore protein
truB ←	3,422,792	G→A	I122I (ATC→ATT)	synonymous	1	L21D182	CD	tRNA pseudouridine synthase
ydiF →	1,690,619	G→A	G86S (GGT→AGT)	non synonymous	1	L21D106	CD	Short chain acyl-CoA transferase: fused alpha subunit ; beta subunit
hscA ←	2,667,858	C→T	S110N (AGC→AAC)	non synonymous	2	L22D182, L22D322	CD	DnaK-like molecular chaperone specific for IscU
yfgB ←	2,653,765	C→T	L71L (CTG→CTA)	synonymous	2	L22D182, L22D322	CD	Putative Fe-S containing enzyme
bcsG →	3,813,987	G→A	S493N (AGC→AAC)	non synonymous	2	L22D182, L22D322	CD	Putative endoglucanase involved in cellulose biosynthesis
oxyR →	4,368,673	G→A	L270L (CTG→CTA)	synonymous	7	L23D91, L23D357, L21D182, L21D106, L21D435, L11D106, L11D182	WD (2) / CD (5)	DNA-binding transcriptional dual regulator
ycaQ →	980,053	C→T	R40C (CGC→TGC)	non synonymous	1	L23D182	CD	Conserved hypothetical protein

ynjC →	1,751,354	C→A	A497E (GCG→GAG)	non synonymous	1	L23D182	CD	Fused putative transporter subunits of ABC superfamily: membrane components
thiP ←	75,540	C→A	A508A (GCG→GCT)	non synonymous	1	L24D106	CD	Fused thiamin transporter subunits of ABC superfamily: membrane components
lacZ ← / ← *lacI*	445,449	G→A	intergenic (-29/+94)	intergenic	14	L24D106, L24D182, L24D209, L23D182, L19D106, L19D182, L19D322, L15D106, L15D182, L15D322, L12D210, L3D106, L3D182, L3D435	CD (10) / WD (4)	Beta-D-galactosidase/DNA-binding transcriptional repressor
ECP_1501 ←	1,551,482	T→A	S243C (AGT→TGT)	non synonymous	1	L24D182	CD	Hypothetical protein
alsK ←	4,540,279	C→T	L45L (CTG→CTA)	synonymous	1	L2D435	WD	D-allose kinase
amiC ← / → *argA*	2,979,756	T→C	intergenic (-161/-71)	intergenic	1	L2D435	WD	N-acetylmuramoyl-L-alanine amidase/fused acetylglutamate kinase homolog (inactive) ; amino acid N-acetyltransferase

nagB ←	729,735	C→T	T44T (ACG→ACA)	synonymous	1	L2D435	WD	Glucosamine-6-phosphate deaminase
putP →	1,074,058	C→T	R117R (CGC→CGT)	synonymous	1	L2D435	WD	Proline:sodium symporter
slyB →	1,639,412	G→A	G63S (GGC→AGC)	non synonymous	1	L2D435	WD	Putative outer membrane lipoprotein
yeaO →	1,794,317	A→C	Q4P (CAG→CCG)	non synonymous	1	L2D435	WD	Conserved hypothetical protein
yebS →	1,828,579	G→A	W162* (TGG→TGA)	stop codon	1	L2D435	WD	Conserved hypothetical protein; putative inner membrane protein
yeeV →	3,174,801	G→A	T37T (ACG→ACA)	synonymous	1	L2D435	WD	Toxin of the YeeV-YeeU toxin-antitoxin system; CP4-44 prophage
ynbC →	1,456,418	C→G	H401D (CAT→GAT)	non synonymous	1	L2D435	WD	Putative hydrolase
yegQ →	2,208,206	C→T	P90S (CCG→TCG)	non synonymous	1	L2D182	WD	Putative peptidase
glgP ←	3,681,807	G→A	Q197* (CAG→TAG)	stop codon	1	L3D182	WD	Glycogen phosphorylase
entF →	654,108	C→T	P1277L (CCG→CTG)	non synonymous	1	L3D435	WD	Enterobactin synthase multienzyme complex component, ATP-dependent
csrA ←	2,795,949	G→A	L12F (CTC→TTC)	non synonymous	1	L4D182	WD	Pleiotropic regulatory protein for carbon source metabolism
yeeJ → / ← *ECP_1952*	2,005,187	C→T	intergenic (+110/+18)	intergenic	1	L4D182	WD	Fragment of adhesin (part 2)/conserved hypothetical protein
yihX →	4,277,958	C→T	Y71Y (TAC→TAT)	synonymous	1	L4D182	WD	Putative hydrolase
yneE ←	1554,115	C→T	R115Q (CGG→CAG)	non synonymous	1	L4D182	WD	Conserved hypothetical protein; putative inner membrane protein
ECP_0231 ←	254,184	C→T	L351L (CTG→CTA)	synonymous	1	L4D435	WD	Conserved hypothetical protein Aec25
ygiF ←	3,311,591	G→A	R364R (CGC→CGT)	synonymous	1	L4D435	WD	Putative adenylate cyclase
ECP_3029 ←	3,189,472	C→T	G193D (GGC→GAC)	non synonymous	2	L5D106, L5D182	WD	Conserved hypothetical protein
yhfK →	3,599,400	C→T	R345W (CGG→TGG)	non synonymous	2	L5D106, L5D182	WD	Conserved hypothetical protein; putative inner membrane protein
metN ←	226,259	A→T	L182Q (CTG→CAG)	non synonymous	4	L5D106, L5D182, L5D435, L13D106	WD (3) / CD (1)	DL-methionine transporter subunit ; ATP-binding component of ABC superfamily
ECP_4092 →	4,276,239	G→A	T287T (ACG→ACA)	synonymous	1	L5D182	WD	Putative transport protein (shikimate homolog)

dacB →	3,440,331	C→T	A331V (GCC→GTC)	non synonymous	1	L5D435	WD	D-alanyl-D-alanine carboxypeptidase
ECP_0334 →	357,101	G→C	V51V (GTG→GTC)	synonymous	1	L5D435	WD	Conserved hypothetical protein; Putative antirestriction protein
slt →	4,929,003	G→A	A353T (GCG→ACG)	non synonymous	1	L5D435	WD	Lytic murein transglycosylase, soluble
yeeS →	357,596	G→T	L49L (CTG→CTT)	synonymous	1	L5D435	WD	Putative DNA repair protein; CP4-44 prophage
fmt →	3,544,317	A→G	Q178Q (CAA→CAG)	synonymous	1	L6D182	WD	10-formyltetrahydrofolate:L-methionyl-tRNA(fMet) N-formyltransferase
glpF ← / → *yiiU*	4,317,747	C→T	intergenic (-365/-66)	intergenic	1	L6D182	WD	Glycerol facilitator/conserved hypothetical protein
ydjJ ←	1,775,772	G→A	P140S (CCT→TCT)	non synonymous	1	L6D182	WD	Putative iditol dehydrogenase
dacC →	895,820	G→A	G318D (GGC→GAC)	non synonymous	1	L6D435	WD	D-alanyl-D-alanine carboxypeptidase (penicillin-binding protein 6a)
fabD →	1,141,152	C→T	A186V (GCG→GTG)	non synonymous	1	L6D435	WD	Malonyl-CoA-[acyl-carrier-protein] transacylase
pta →	2,446,433	A→G	K149E (AAA→GAA)	non synonymous	1	L6D435	WD	Phosphate acetyltransferase
znuA ←	1,854,014	G→A	H154Y (CAT→TAT)	non synonymous	1	L7D106	WD	Zinc transporter subunit: periplasmic-binding component of ABC superfamily
ECP_1501 ←	1,551,577	C→T	G211D (GGC→GAC)	non synonymous	3	L7D106, L7D182, L7D435	WD	ypothetical protein
ECP_1970 ←	2,035,828	G→C	A72G (GCA→GGA)	non synonymous	1	L7D182	WD	Putative peptide synthetase
manA →	1,608,784	G→A	E264K (GAA→AAA)	non synonymous	1	L7D435	WD	Mannose-6-phosphate isomerase
mdtG ←	1,105,597	T→C	N255D (AAC→GAC)	non synonymous	1	L8D182	WD	Putative metabolite efflux transporter
papH ←	4,752,992	C→T	R111Q (CGG→CAG)	non synonymous	1	L8D182	WD	Minor pilin protein PapH
leuB ←	83,137	G→A	A48V (GCA→GTA)	non synonymous	2	L8D182, L8D435	WD	3-isopropylmalate dehydrogenase

cyoA ← / ← *ampG*	519,253	A→T	intergenic (-419/+41)	intergenic	1	L8D435	WD	Cytochrome o ubiquinol oxidase subunit II/muropeptide transporter
ECP_3029 ←	3,189,842	C→T	D70N (GAT→AAT)	non synonymous	1	L8D435	WD	Conserved hypothetical protein
ECP_4610 ←	4,817,157	A→T	Y102N (TAT→AAT)	non synonymous	3	L9D106, L9D182, L9D442	WD	Conserved hypothetical protein, putative Transcriptional regulator
fldB →	3,059,194	C→A	A169E (GCA→GAA)	non synonymous	1	L9D182	WD	Flavodoxin 2
ECP_0241 →	266,134	C→T	H82Y (CAC→TAC)	non synonymous	1	L9D442	WD	Conserved hypothetical protein Aec14
rcsD →	2,351,207	A→C	I239L (ATC→CTC)	non synonymous	1	L9D442	WD	Phosphotransfer intermediate protein in two-component regulatory system with RcsBC
ydcF →	1,463,442	C→T	A43V (GCA→GTA)	non synonymous	1	L9D442	WD	Conserved hypothetical protein

ANNEXE N°4: LISTE DES MUTATIONS DE DELETION OU D'INSERTION TROUVEES DANS LES 71 ISOLATS D'*E. COLI* 536 SEQUENCES DANS L'ARTICLE N°2 DE LA THESE (TABLE S3)

Gene	Position	Mutation	Mutation effect	Type of mutation	Occurrence	Mice ID	Diet	Gene information
cadA ←	4,588,081	Δ7 bp	coding (892-898/2148 nt)	Deletion with frameshift	1	L10D442	WD	Lysine decarboxylase 1
ECP_0018 →	18,393	(A)7→6	coding (1381/1494 nt)	Deletion with frameshift	1	L1D435	WD	Putative Cerebroside-sulfatase
ECP_1980 → / → *ECP_1981*	2,066,746	(ACAGATAC)10→9	intergenic (+179/-21)	Intergenic indel	1	L14D442	CD	Hypothetical protein/hypothetical protein
ECP_2077 ←	2,145,824	(A)8→7	coding (218/1146 nt)	Deletion with frameshift	1	L23D456	CD	Conserved hypothetical protein
ECP_2966 ←	3,132,987	(T)7→6	coding (13/993 nt)	Deletion with frameshift	1	L23D456	CD	PixG protein
ECP_4735	231,939	+ATCTCGGTGGTCGCCGTATCATTAAAAAAAAAA	noncoding (2115/2904 nt)	Insertion with no frameshift	1	L24D106	CD	23S ribosomal RNA
fruB ← / → *set B*	2,304,373	(A)7→6	intergenic (-51/-317)	Intergenic indel	1	L12D210	WD	Fused fructose-specific PTS enzymes: IIA component ; HPr component/lactose/glucose efflux system
int – ECP_0342	294,395	Δ68,078 bp	NA	Large deletion	5	L11D442, L3D435, L10D182, L9D106, L6D182	WD	76 genes: int, xis, ECP_0275, ECP_0276, ECP_0277, ECP_0278, ECP_0279, ECOLI0280, ECP_0280, ECP_0281, ECP_0282, insF, insE, ECP_0285, ECP_0286, ECOLI0287, ECP_0287, ECP_0288, ECP_0289, ECP_0290, papI.2, papB,

								ECP_0293, ECP_0294, ECP_0295, ECP_0296, ECP_0297, ECP_0298, ECP_0299, ECP_0300, ECP_0301, ECOLI0303, ECP_0302, ECP_0303, ECP_0304, ECP_0305, ECP_0306, ECP_0307, insG, ECOLI0311, ECP_0309, ECP_0310, ECP_0311, ECP_0312, ECP_0313, ECP_0314, ECP_0315, ECP_0316, ECOLI0318, ECP_0317, ECP_0318, ECP_0319, ECP_0320, ECP_0321, ECP_0322, ECP_0323, ECP_0324, ECP_0325, ECP_0326, ECP_0327, ECP_0328, ECP_0329, ECP_0330, ECP_0331, flu, ECP_0333, ECP_0334, yeeS, yeeT, ECOLI0337, yeeU, yeeV, yeeW.2, ECP_0340, ECP_0341, ECP_0342
lacI ←	446,034	(CCTGCCAG)1→2	coding (592/1083 nt)	Duplication with frameshift	17	L22D106, L22D182, L22D456, L21D182, L21D106, L21D435, L17D456, L16D106, L16D196, L12D106,	WD (8) / CD (9)	DNA-binding transcriptional repressor

						L11D106, L11D182, L8D182, L8D435, L6D106, L4D435, L1D435		
lacI ←	446,035	Δ1 bp	coding (591/1083 nt)	Deletion with frameshift	2	L14D182, L14D106	CD	DNA-binding transcriptional repressor
lacI ←	446,039	(AGACGC)2→1	coding (582-587/1083 nt)	Deletion with no frameshift	1	L12D182	WD	DNA-binding transcriptional repressor
lacI ←	446,474	Δ1 bp	coding (152/1083 nt)	Deletion with frameshift	9	L16D182, L13D442, L9D106, L9D182, L9D442, L5D106, L5D182, L5D435, L13D106	WD (6) / CD (3)	DNA-binding transcriptional repressor
lacY ←	441,830	+CGAG	coding (465/1254 nt)	Insertion with frameshift	3	L22D106, L6D106, L16D106	WD (1) / CD (2)	Lactose/galactose transporter
malT →	3,660,530	(GCTGCCAGTGGGAAC TGGCGGCGGA)1->2	coding (992/2706 nt)	Duplication with frameshift	5	L10D182, L17D456, L22D182, L22D456, L2D182	WD (2) / CD (3)	Fused conserved hypothetical protein ; DNA-binding transcriptional activator, maltotriose-ATP-binding
manX →	1,814,451	(TCTCCC)1→2	coding (568/972 nt)	Duplication with no frameshift	2	L5D106, L5D182	WD	Fused mannose-specific PTS enzymes: IIA component ; IIB component

ptsA ← / → *frwC*	4,349,140	+T	intergenic (-308/-1)	Intergenic indel	1	L8D435	WD	Fused putative PTS enzymes: Hpr component ; enzyme I component ; enzyme IIA component/enzyme IIC component of PTS
ubiD →	4,232,244	Δ1 bp	pseudogene (1407/1494 nt)	Deletion with frameshift	2	L22D182, L22D456	CD	3-octaprenyl-4-hydroxybenzoate decarboxylase
ydhQ ← / → *ECP_1612*	1,661,294	Δ19 bp	intergenic (-220/-71)	Intergenic indel	1	L6D435	WD	Conserved hypothetical protein/tRNA-Val
ytfJ ← / → *ytfK*	4,676,714	(A)7→6	intergenic (-203/-122)	Intergenic indel	1	L13D442	CD	Putative transcriptional regulator/conserved hypothetical protein
ECP_0758 – *ECP_0759/ECP_0760*	796,391	Δ124 bp	noncoding (76/76 nt) – intergenic (+1/-33)	Large deletion with frameshift	1	L8D106	WD	tRNA-Lys – tRNA-Lys/tRNA-Lys

ANNEXE N°5: LISTE DES MUTATIONS DE TRANSPOSITION DECOUVERTES DANS LES 71 ISOLATS SEQUENCES D'*E. COLI* 536 DANS L'ARTICLE N°2 (TABLE S4)

Target gene	Position	IS	Transposition type	Mutation type	Mice ID	Diet	Target gene information
ECP_0381/ECP_0382	406,296	recombinaison	NA	intergenic (-40/+290)	L1D182, L4D435, L5D435, L10D106	WD	putative transcription regulator; LuxR-type HTH domain/Putative DNA recombinase similar to Type 1 fimbriae Regulatory proteins
ECP_2995	3,157,822	IS 3	replicative	coding (459/1023 nt)	L10D106	CD	transposase ORF A, IS100, IS21 family
ydhT	1,664,213	IS 100 / IS 21	replicative	coding (248/765 nt)	L14D442	CD	conserved hypothetical protein
ECP_0275/ECP_0276	296,501	IS 100 / IS 21	replicative	intergenic (-276/-22)	L15D322	CD	conserved hypothetical protein/transposase ORF A, IS100, IS21 family
ydhW	1,666,381	IS 100 / IS 21	replicative	coding (284/639 nt)	L16D182	CD	conserved hypothetical protein
ECP_4523	4,738,995	IS 100 / IS 21	replicative	coding (1978/3516 nt)	L16D182	CD	Putative superfamily I DNA helicase
ECP_4524	4,742,095	IS 100 / IS 21	replicative	coding (47/1500 nt)	L17D456	CD	Putative membrane-associated, metal-dependent hydrolase
ECP_1911	1,952,296	IS 3	replicative	coding (90/1464 nt)	L18D28, L18D65	CD	PilV-like protein
ECP_2592	2,732,131	IS 110	replicative	coding (115/405 nt)	L13D442	CD	Putative transposase (fragment), IS110 family

ANNEXE N°6: CARACTERISTIQUES DES SOURIS DES 126 SOUCHES SEQUENCEES DANS L'ARTICLE N°3 (TABLE S1)

Pour chacun des 126 isolats séquencés, nous avons rapporté la période d'échantillonnage, l'identifiant de souris échantillonné (lignée et nombre de souris), le régime alimentaire, la taille relative de la colonie (petite, moyenne ou grande) et la densité de *E. coli* par gramme de feces.

isolate ID	strain	mouse ID	Time of sampling for sequencing in day	colonies	Diet	*E. coli* density
49AJ201	HS	49	201	Middle	WD	14179104.4776119
49AJ352	HS	49	352	Middle	WD	642857142.857142
49AJ82	HS	49	82	Small	WD	239425379.090185
49BJ201	HS	49	201	Middle	WD	14179104.4776119
49BJ352	HS	49	352	Small	WD	642857142.857142
49BJ82	HS	49	82	Small	WD	239425379.090185
57AJ202	HS	57	202	Big	WD	13450292.3976607
57AJ352	HS	57	352	Middle	WD	3053763440.86019
57AJ83	HS	57	83	Big	WD	202839756.592292
57BJ202	HS	57	202	Small	WD	13450292.3976607
57BJ352	HS	57	352	Small	WD	3053763440.86019
57BJ83	HS	57	83	Middle	WD	202839756.592292
60AJ202	HS	60	202	Small	WD	27210.8843537415
60AJ352	HS	60	352	Middle	WD	3527336.86067018
60AJ83	HS	60	83	Middle	WD	45941807.0444106
60BJ202	HS	60	202	Small	WD	27210.8843537415
60BJ352	HS	60	352	Small	WD	3527336.86067018
60BJ83	HS	60	83	Small	WD	45941807.0444106
63AJ203	HS	63	203	Middle	WD	20671834.625323
63AJ352	HS	63	352	Small	WD	339350180.505414
63AJ84	HS	63	84	Middle	WD	13763896.2413976
63BJ203	HS	63	203	Small	WD	20671834.625323
63BJ352	HS	63	352	Small	WD	339350180.505414
63BJ84	HS	63	84	Small	WD	13763896.2413976
66AJ203	HS	66	203	Small	WD	46783625.7309938
66AJ322	HS	66	322	Middle	WD	1250000.00000001
66AJ84	HS	66	84	Middle	WD	31620553.3596831
66BJ203	HS	66	203	Small	WD	46783625.7309938
66BJ322	HS	66	322	Middle	WD	1250000.00000001
66BJ84	HS	66	84	Small	WD	31620553.3596831
69AJ231	HS	69	231	Small	WD	2770083.10249306
69AJ352	HS	69	352	Middle	WD	44609665.4275094
69AJ84	HS	69	84	Big	WD	8872458.4103512
69BJ231	HS	69	231	Small	WD	2770083.10249306
69BJ352	HS	69	352	Small	WD	44609665.4275094
69BJ84	HS	69	84	Small	WD	8872458.4103512
72AJ203	HS	72	203	Middle	CD	32997250.2291476
72AJ351	HS	72	351	Middle	CD	184032476.31935

72AJ84	HS	72	84	Big	CD	55363321.7993082
72BJ203	HS	72	203	Small	CD	32997250.2291476
72BJ351	HS	72	351	Small	CD	184032476.31935
72BJ84	HS	72	84	Middle	CD	55363321.7993082
74AJ203	HS	74	203	Big	CD	20434227.330779
74AJ351	HS	74	351	Big	CD	71914480.0777454
74AJ84	HS	74	84	Big	CD	27411702.6884554
74BJ203	HS	74	203	Middle	CD	20434227.330779
74BJ351	HS	74	351	Middle	CD	71914480.0777454
74BJ84	HS	74	84	Big	CD	27411702.6884554
78AJ203	HS	78	203	Middle	CD	3515332.83470457
78AJ351	HS	78	351	Middle	CD	18641810.918775
78AJ84	HS	78	84	Big	CD	30273437.5
78BJ203	HS	78	203	Small	CD	3515332.83470457
78BJ351	HS	78	351	Middle	CD	18641810.918775
78BJ84	HS	78	84	Middle	CD	30273437.5
79AJ203	HS	79	203	Middle	CD	19164955.5099247
79AJ351	HS	79	351	Small	CD	131195335.276967
79AJ84	HS	79	84	Small	CD	2369668.24644551
79BJ203	HS	79	203	Small	CD	19164955.5099247
79BJ351	HS	79	351	Small	CD	131195335.276967
79BJ84	HS	79	84	Small	CD	2369668.24644551
84AJ203	HS	84	203	Big	CD	14475271.411339
84AJ351	HS	84	351	Big	CD	248226950.35461
84AJ84	HS	84	84	Big	CD	79638009.0497737
84BJ203	HS	84	203	Big	CD	14475271.411339
84BJ351	HS	84	351	Middle	CD	248226950.35461
84BJ84	HS	84	84	Middle	CD	79638009.0497737
87AJ203	HS	87	203	Middle	CD	1645278.05199078
87AJ351	HS	87	351	Big	CD	321839080.459769
87AJ84	HS	87	84	Big	CD	51122194.5137157
87BJ203	HS	87	203	Middle	CD	1645278.05199078
87BJ351	HS	87	351	Middle	CD	321839080.459769
87BJ84	HS	87	84	Middle	CD	51122194.5137157
90AJ203	HS	90	203	Big	CD	2569593.1477516
90AJ351	HS	90	351	Middle	CD	136029411.764706
90AJ84	HS	90	84	Big	CD	20202020.2020202
90BJ203	HS	90	203	Middle	CD	2569593.1477516
90BJ351	HS	90	351	Middle	CD	136029411.764706
90BJ84	HS	90	84	Middle	CD	20202020.2020202
3AJ202	536	3	202	Big	WD	85385878.4893265
3AJ353	536	3	353	Big	WD	1388888888.88889
3AJ83	536	3	83	Big	WD	574162679.425833
3BJ202	536	3	202	Big	WD	85385878.4893265
3BJ353	536	3	353	Middle	WD	1388888888.88889
3BJ83	536	3	83	Big	WD	574162679.425833
6AJ202	536	6	202	Big	WD	17857142.8571428
6AJ353	536	6	353	Big	WD	1999999999.99978

6AJ83	536	6	83	Big	WD	14371257.4850301
6BJ202	536	6	202	Middle	WD	17857142.8571428
6BJ353	536	6	353	Middle	WD	1999999999.99978
6BJ83	536	6	83	Middle	WD	14371257.4850301
9AJ202	536	9	202	Big	WD	30816640.9861324
9AJ353	536	9	353	Big	WD	521739130.43478
9AJ83	536	9	83	Big	WD	545454.545454569
9BJ202	536	9	202	Big	WD	30816640.9861324
9BJ353	536	9	353	Middle	WD	521739130.43478
9BJ83	536	9	83	Middle	WD	545454.545454569
11AJ202	536	11	202	Middle	WD	4223227.75263952
11AJ353	536	11	353	Big	WD	139860139.86014
11AJ83	536	11	83	Big	WD	352112676.056344
11BJ202	536	11	202	Middle	WD	4223227.75263952
11BJ353	536	11	353	Middle	WD	139860139.86014
11BJ83	536	11	83	Big	WD	352112676.056344
13AJ202	536	13	202	Big	WD	17283950.617284
13AJ353	536	13	353	Big	WD	282208588.957059
13AJ83	536	13	83	Big	WD	105263157.894736
13BJ202	536	13	202	Big	WD	17283950.617284
13BJ353	536	13	353	Middle	WD	282208588.957059
13BJ83	536	13	83	Middle	WD	105263157.894736
17AJ202	536	17	202	Small	WD	49180327.8688525
17AJ353	536	17	353	Middle	WD	1408450704.22535
17AJ83	536	17	83	Big	WD	241379310.344817
17BJ202	536	17	202	Middle	WD	49180327.8688525
17BJ353	536	17	353	Middle	WD	1408450704.22535
17BJ83	536	17	83	Middle	WD	241379310.344817
19AJ202	536	19	202	Big	WD	8762886.59793811
19AJ353	536	19	353	Big	WD	604026845.637578
19AJ83	536	19	83	Big	WD	27586206.8965517
19BJ202	536	19	202	Middle	WD	8762886.59793811
19BJ353	536	19	353	Middle	WD	604026845.637578
19BJ83	536	19	83	Small	WD	27586206.8965517
23AJ202	536	23	202	Big	WD	860986.547085206
23AJ353	536	23	353	Big	WD	471698113.207545
23AJ83	536	23	83	Big	WD	18867924.5283019
23BJ202	536	23	202	Middle	WD	860986.547085206
23BJ353	536	23	353	Middle	WD	471698113.207545
23BJ83	536	23	83	Middle	WD	18867924.5283019

ANNEXE N°7: LISTE DES MUTATIONS TROUVEES DANS LES 48 ISOLATS SEQUENCES DE *E. COLI* 536 DANS L'ARTICLE N°3 DE LA THESE (TABLE S2)

Gene	Position	Mutation	Mutation effect	Mutation type	Occurrence	Mice ID	Diet	Gene information
[cpsG]–[cpsG]	2139900	Δ19,522	NA	deletion	1	11BJ353	WD	16 genes
[ECP_0273]–ECP_0342	294387	Δ68,039	NA	deletion	1	9AJ353	WD	77 genes
[yeeT]–[ECP_3861]	4021738	Δ1,629	NA	deletion	1	6AJ353	WD	[yeeT], yeeU, yeeV, yeeW.2, [ECP_3861]
A →	1208550	C→T	C158C (TGC→TGT)	synonymous SNP	1	19BJ353	WD	Terminase large subunit from bacteriophage origin
acrB ←	548907	C→T	A912T (GCG→ACG)	non synonymous SNP	1	11AJ83	WD	multidrug efflux system protein
adhE ←	1318677	G→A	N741N (AAC→AAT)	synonymous SNP	1	17BJ353	WD	fused acetaldehyde-CoA dehydrogenase ; iron-dependent alcohol dehydrogenase ; pyruvate-formate lyase deactivase
	1320093	G→A	H269H (CAC→CAT)	synonymous SNP	1	9AJ353	WD	
aldA →	1465242	G→A	A311T (GCG→ACG)	non synonymous SNP	1	3AJ353	WD	aldehyde dehydrogenase A, NAD-linked
allC ←	606706	C→T	G396R (GGG→AGG)	non synonymous SNP	2	13AJ202;13AJ353	WD	allantoate amidohydrolase
amtB →	541449	G→A	L416L (CTG→CTA)	synonymous SNP	1	9BJ353	WD	ammonium transporter
arcA ←	4936786	G→A	H226Y (CAC→TAC)	non synonymous SNP	3	13AJ83;13BJ83;13BJ353	WD	DNA-binding response regulator in two-component regulatory system with ArcB or CpxA
arcC ← / ← *ECP_4500*	4713424	C→T	intergenic (-10/+1)	intergenic SNP	1	11BJ202	WD	Carbamate kinase/Putative arginine deiminase (Arginine dihydrolase)
atoC →	2359256	T→G	I129S (ATC→AGC)	non synonymous SNP	1	13BJ353	WD	fused response regulator of ato opeon, in two-component system with AtoS: response regulator ; sigma54 interaction protein
atoS →	2357877	C→G	A277G (GCC→GGC)	non synonymous SNP	1	19AJ202	WD	sensory histidine kinase in two-component regulatory system with AtoC
	2358102	G→A	R352H (CGT→CAT)	non synonymous SNP	1	17AJ202	WD	
atpI/gidB-[gidB]	4106438	Δ779 bp	intergenic (-300/+305)	deletion	2	9AJ353;9BJ353	WD	methyltransferase, SAM-dependent methyltransferase, glucose-inhibited cell-division protein
bglH ←	4084412	C→T	A532T (GCC→ACC)	non synonymous SNP	1	3BJ353	WD	carbohydrate-specific outer membrane porin, cryptic
clcB	1585431	IS21	coding (735/1293 nt)	replicative	1	11AJ353	WD	putative voltage-gated chloride channel
cpsG ←	2139875	G→A	N420N (AAC→AAT)	synonymous SNP	1	11BJ353	WD	phosphomannomutase
csrA ←	2795891	C→T	R31H (CGT→CAT)	non synonymous SNP	1	13BJ353	WD	pleiotropic regulatory protein for carbon source metabolism
cstA →	667412	C→T	A528V (GCG→GTG)	non synonymous SNP	1	3AJ353	WD	carbon starvation protein
cusB →	635231	C→T	A122V (GCC→GTC)	non synonymous SNP	5	17AJ202;17AJ353;17BJ83;17BJ202;17BJ353	WD	copper/silver efflux system, membrane fusion protein
cvpA ←	2461908	G→A	T33I (ACA→ATA)	non synonymous SNP	2	3AJ202;3BJ202	WD	membrane protein required for colicin V production

cvpA ← / ← *dedD*	2462181	G→A	intergenic (-176/+629)	intergenic SNP	1	23AJ353	WD	membrane protein required for colicin V production/conserved hypothetical protein
cydA →	787548	Δ1 bp	coding (725/1569 nt)	deletion	1	19BJ83	WD	cytochrome d terminal oxidase, subunit I
cysS →	618162	G→A	G199D (GGC→GAC)	non synonymous SNP	1	19AJ353	WD	cysteinyl-tRNA synthetase
dcp ←	1573480	(TTG)2→1	coding (165-167/2046 nt)	deletion	1	11BJ353	WD	dipeptidyl carboxypeptidase II
dnaK →	12923	C→T	D233D (GAC→GAT)	synonymous SNP	1	19BJ353	WD	chaperone Hsp70, co-chaperone with DnaJ
dsdX ←	4801282	G→A	L327F (CTC→TTC)	non synonymous SNP	1	19BJ202	WD	putative transporter
ECOLI2729	2844472	IS21	coding (117/261 nt)	replicative	1	19BJ353	WD	hypothetical protein
ECOLI2999/ECP_2973	3139370	Δ111	NA	deletion	1	13BJ83	WD	PixB protein/conserved hypothetical protein
ECOLI3635 →	3758957	C→T	A17V (GCT→GTT)	non synonymous SNP	2	3AJ353;3BJ353	WD	hypothetical protein
ECP_0297 →	316282	C→T	P133L (CCG→CTG)	non synonymous SNP	1	23BJ202	WD	S-fimbrial adhesin protein sfaG precursor
ECP_0742 → / → *cydA*	786180	C→A	intergenic (+460/-644)	intergenic SNP	1	11AJ83	WD	conserved hypothetical protein/cytochrome d terminal oxidase, subunit I
ECP_0758/ECP_0759	796392	+GT	intergenic (+2/-47)	insertion	1	3AJ202	WD	tRNA-Lys/tRNA-Lys
ECP_0759–[ECP_0760]	796393	Δ172	NA	deletion	2	3AJ202;3BJ202	WD	ECP_0759, [ECP_0760]
ECP_1244 ←	1271344	G→A	T119M (ACG→ATG)	non synonymous SNP	2	9AJ353;9BJ353	WD	putative ABC transporter periplasmic binding protein
ECP_1345 →	1378205	G→A	D592N (GAC→AAC)	non synonymous SNP	1	6BJ202	WD	Putative efflux pump (Putative multidrug-efflux transport protein)
ECP_1347 →	1381241	G→A	G106E (GGG→GAG)	non synonymous SNP	1	11BJ353	WD	putative membrane transport protein of an efflux system
ECP_1500 →	1550528	G→A	pseudogene (158/672 nt)	SNP	1	6BJ353	WD	fragment of putative transcriptional regulator, lysR family (partial)
ECP_1897 →	1938511	C→T	G8G (GGC→GGT)	synonymous SNP	1	3AJ353	WD	hypothetical protein
ECP_2007 →	2087710	C→T	S100F (TCT→TTT)	non synonymous SNP	1	19AJ353	WD	putative phosphotriesterase-related protein
ECP_2074 ←	2142613	C→T	G19D (GGT→GAT)	non synonymous SNP	1	19AJ202	WD	Mannose-1-phosphate guanylyltransferase [GDP] (GDP- mannose pyrophosphorylase) (GMP)
ECP_2077 ←	2145824	(A)8→7	coding (218/1146 nt)	deletion	1	11AJ202	WD	conserved hypothetical protein
ECP_2709/ECP_2710	2848693	IS21	intergenic (-31/+238)	replicative	1	13AJ353	WD	conserved hypothetical protein/hypothetical protein
ECP_2754 →	2894570	C→T	R43C (CGT→TGT)	non synonymous SNP	1	17AJ353	WD	Sucrose operon repressor
ECP_3347 ←	3514752	G→A	R71C (CGC→TGC)	non synonymous SNP	1	9BJ202	WD	Putative fructose/tagatose biphosphate aldolase (EC 4.1.2.-)
ECP_3765/ECOLI3810	3949287	IS3	intergenic (+37/-208)	replicative	1	23AJ83	WD	Putative prophage integrase/Prophage P4 integrase (Int(P4)) (fragment)

ECP_3784 ←	3964543	C→G	A136P (GCC→CCC)	non synonymous SNP	1	19AJ83	WD	Chaperone protein hifB precursor
ECP_3789 ←	3968091	C→T	D94N (GAC→AAC)	non synonymous SNP	1	13AJ353	WD	conserved hypothetical protein
ECP_3849 ←	4016584	G→A	L35L (CTG→TTG)	synonymous SNP	2	3AJ353;3BJ353	WD	conserved hypothetical protein, putative transposase ORF A, IS3 family
ECP_3861 →	4023412	A→G	K48E (AAG→GAG)	non synonymous SNP	1	6AJ353	WD	conserved hypothetical protein
	4023415	G→T	D49Y (GAT→TAT)	non synonymous SNP	1	6AJ353	WD	
	4023419	A→G	K50R (AAA→AGA)	non synonymous SNP	1	6AJ353	WD	
ECP_4106 ←	4287399	+T	coding (961/2415 nt)	insertion	1	17BJ353	WD	formate dehydrogenase-O, large subunit, selenoprotein, molybdoenzyme
ECP_4523	4739538	IS3	coding (2521/3516 nt)	replicative	2	9AJ353;9BJ353	WD	Putative superfamily I DNA helicase
ECP_4610 ←	4816928	C→T	R178H (CGT→CAT)	non synonymous SNP	5	6AJ83;6AJ202;6AJ353;6BJ83;6BJ202	WD	conserved hypothetical protein, putative Transcriptional regulator
	4817259	C→T	A68T (GCC→ACC)	non synonymous SNP	1	3AJ353	WD	
ECP_4630 → / → *yeeS*	4834398	+G	intergenic (+6/-9)	insertion	1	11AJ83	WD	fragment of conserved hypothetical protein; Putative antirestriction protein (partial)/putative DNA repair protein; CP4-44 prophage
	4834405	T→C	intergenic (+13/-2)	intergenic SNP	1	11AJ83	WD	
ECP_4717 ← / ← *ECP_4716*	2540176	G→A	intergenic (-25/+15)	intergenic SNP	1	11BJ353	WD	tRNA-Ala/tRNA-Ala
ECP_4733	2734934	+TTTTTTTTTTAATGATACGGCGACCACCGAGAT	noncoding (2149/2904 nt)	insertion	1	9AJ202	WD	23S ribosomal RNA
ECP_4733 ← / ← *ECP_4715*	2737105	T→G	intergenic (-23/+171)	intergenic SNP	1	13BJ202	WD	23S ribosomal RNA/tRNA-Glu
	2737110	T→C	intergenic (-28/+166)	intergenic SNP	1	13BJ202	WD	
	2737123	A→G	intergenic (-41/+153)	intergenic SNP	1	13BJ202	WD	
ECP_4735	231939	+ATCTCGGTGGTCGCCGTATCATTAAAAAAAAAA	noncoding (2115/2904 nt)	insertion	1	3BJ83	WD	23S ribosomal RNA
entD ← / ← *fepA*	646209	G→A	intergenic (-7/+159)	intergenic SNP	1	11AJ353	WD	phosphopantetheinyltransferase component of enterobactin synthase multienzyme complex/iron-enterobactin outer membrane transporter
fliD →	1906952	C→T	A87V (GCG→GTG)	non synonymous SNP	1	19AJ353	WD	flagellar filament capping protein
fucP →	2928048	A→G	K135E (AAA→GAA)	non synonymous SNP	1	9BJ202	WD	L-fucose transporter
gatY ← / ← *fbaB*	2217937	C→T	intergenic (-174/+53)	intergenic SNP	1	23AJ353	WD	D-tagatose 1,6-bisphosphate aldolase 2, catalytic subunit/fructose-bisphosphate aldolase class I

	4106956	+ATTACGCC ATCAAATGG CGGCTCTGA CGGAAACTC TTCTACCCTA CTCTG	coding (410/624 nt)	insertion	1	11AJ353	WD	
	4106777	G→A	R197C (CGT→TGT)	non synonymous SNP	1	19AJ353	WD	
	4106902	Δ25 bp	coding (440-464/624 nt)	deletion	2	3AJ83;3BJ83	WD	
	4106948	C→T	A140T (GCT→ACT)	non synonymous SNP	1	11AJ83	WD	
	4107005	G→A	Q121* (CAG→TAG)	stop SNP	1	9BJ202	WD	
gidB	4107022	Δ2 bp	coding (343-344/624 nt)	deletion	2	3AJ202;3BJ2 02	WD	methyltransferase, SAM-dependent methyltransferase, glucose-inhibited cell-division protein
	4107047	G→A	Q107* (CAG→TAG)	stop SNP	2	6AJ202;6AJ3 53	WD	
	4107075	+G	coding (291/624 nt)	insertion	2	3AJ353;3BJ3 53	WD	
	4107116	A→G	S84P (TCT→CCT)	non synonymous SNP	1	6AJ83	WD	
	4107130	G→A	P79L (CCT→CTT)	non synonymous SNP	1	23AJ353	WD	
	4107145	G→A	T74I (ACC→ATC)	non synonymous SNP	1	9AJ202	WD	
	4107148	C→T	G73D (GGC→GAC)	non synonymous SNP	3	13AJ202;13A J353;13BJ20 2	WD	
	4107204	+T	coding (162/624 nt)	insertion	1	23AJ202	WD	
glnP ←	862717	C→T	M1M (ATG→ATA) †	synonymous SNP	1	23AJ202	WD	glutamine transporter subunit ; membrane component of ABC superfamily
hrpB →	169554	G→A	P667P (CCG→CCA)	synonymous SNP	1	9BJ202	WD	ATP-dependent (RNA) helicase
hyaD →	1039634	C→T	R167R (CGC→CGT)	synonymous SNP	1	19BJ353	WD	protein involved in processing of HyaA and HyaB proteins
hydN ←	2814848	G→A	A64V (GCA→GTA)	non synonymous SNP	2	23BJ202;23B J353	WD	formate dehydrogenase-H, [4Fe-4S] ferredoxin subunit
hypF ←	2812656	G→A	A568V (GCA→GTA)	non synonymous SNP	1	19BJ353	WD	carbamoyl phosphate phosphatase and maturation protein for [NiFe] hydrogenases
ibpB ←	4051064	G→A	Q78* (CAG→TAG)	stop SNP	1	13BJ353	WD	heat shock chaperone
icd →	1186387	C→T	C127C (TGC→TGT)	synonymous SNP	1	17AJ353	WD	isocitrate dehydrogenase, specific for NADP+; e14 prophage
int–ECP_0342	294395	Δ68,030	NA	deletion	2	3AJ353;3BJ3 53	WD	76 genes
irp2 →	1983431	C→T	R1053R (CGC→CGT)	synonymous SNP	1	9BJ202	WD	High-molecular-weight nonribosomal peptide/polyketide synthetase 2 (HMWP2) (Yersiniabactin siderophore biosynthetic protein)
kdpB ←	758862	G→A	A594V (GCG→GTG)	non synonymous SNP	4	11AJ202;11A J353;11BJ20 2;11BJ353	WD	potassium translocating ATPase, subunit B

	758897	G→A	N582N (AAC→AAT)	synonymous SNP	1	17AJ83	WD	
	760363	C→T	A94T (GCC→ACC)	non synonymous SNP	1	3BJ353	WD	
lacI ←	446034	(CCTGCCAG) 1→2	coding (592/1083 nt)	duplication	2	9AJ353;9BJ353	WD	DNA-binding transcriptional repressor
	446044	(AGACGC)2 →3	coding (582/1083 nt)	duplication	2	19AJ83;19AJ353	WD	
	446426	Δ1 bp	coding (200/1083 nt)	deletion	3	13AJ202;13AJ353;13BJ202	WD	
lacI ← / → *yaiL*	446663	C→A	intergenic (-38/-164)	intergenic SNP	1	23AJ202	WD	DNA-binding transcriptional repressor/nucleoprotein/polynucleotide-associated enzyme
lacY ←	441061	G→A	Q412* (CAG→TAG)	stop SNP	2	9AJ353;9BJ353	WD	lactose/galactose transporter
lacZ ←	442914	Δ957 bp	coding (1551-2507/3075 nt)	deletion	1	9BJ202	WD	beta-D-galactosidase
lacZ ← / ← *lacI*	445452	C→A	intergenic (-32/+91)	intergenic SNP	2	23AJ83;23BJ83	WD	beta-D-galactosidase/DNA-binding transcriptional repressor
leuB ←	82477	G→A	P268L (CCT→CTT)	non synonymous SNP	1	19BJ353	WD	3-isopropylmalate dehydrogenase
maa ← / ← *hha*	547265	G→A	intergenic (-111/+63)	intergenic SNP	1	19BJ353	WD	maltose O-acetyltransferase/modulator of gene expression, with H-NS
malT →	3660683	G→A	W382* (TGG→TAG)	stop SNP	6	11AJ83;11AJ202;11AJ353;11BJ83;11BJ202;11BJ353	WD	fused conserved hypothetical protein ; DNA-binding transcriptional activator, maltotriose-ATP-binding
	3661756	C→T	Q740* (CAG→TAG)	stop SNP	3	9AJ83;9AJ202;9BJ83	WD	
melR ←	4572957	C→T	M225I (ATG→ATA)	non synonymous SNP	2	9AJ353;9BJ353	WD	DNA-binding transcriptional dual regulator
	4573091	G→A	H181Y (CAC→TAC)	non synonymous SNP	1	19AJ353	WD	
metH →	4433317	C→T	S386S (AGC→AGT)	synonymous SNP	1	3BJ83	WD	homocysteine-N5-methyltetrahydrofolate transmethylase, B12-dependent
mfd ←	1161527	G→A	Q1125* (CAG→TAG)	stop SNP	1	19BJ353	WD	transcription-repair coupling factor
mogA → / ← *yaaH*	9887	G→A	intergenic (+97/+108)	intergenic SNP	1	6AJ353	WD	molybdochelatase MogA, involved in Moco biosynthesis/conserved hypothetical protein; putative inner membrane protein associated with acetate transport
murE →	94932	G→A	G150E (GGG→GAG)	non synonymous SNP	1	13AJ353	WD	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate:meso-diaminopimelate ligase
narH →	1305578	C→T	P48S (CCG→TCG)	non synonymous SNP	1	3AJ353	WD	nitrate reductase 1, beta (Fe-S) subunit
nrfA →	4514101	C→T	T269T (ACC→ACT)	synonymous SNP	1	13BJ202	WD	nitrite reductase, formate-dependent, cytochrome
nrfE →	4517596	C→T	A185V (GCC→GTC)	non synonymous SNP	1	23AJ353	WD	heme lyase (NrfEFG) for insertion of heme into c552, subunit NrfE

nudC →	4408324	C→A	A206E (GCG→GAG)	non synonymous SNP	1	6BJ353	WD	NADH pyrophosphatase
nuoG ←	2429868	G→A	A516V (GCT→GTT)	non synonymous SNP	1	11BJ202	WD	NADH:ubiquinone oxidoreductase, chain G
ompA ← / → *ECP_0963*	1023615	(A)8→7	intergenic (-198/-4)	deletion	1	19BJ202	WD	outer membrane protein A (3a;II*;G;d)/fragment of conserved hypothetical protein (partial)
pepB ←	2664517	C→A	A404A (GCG→GCT)	synonymous SNP	1	6BJ353	WD	aminopeptidase B
pgm →	749038	C→T	P93S (CCT→TCT)	non synonymous SNP	2	11AJ202;11BJ353	WD	phosphoglucomutase
plsB ←	4467720	C→T	pseudogene (523/1273 nt)	SNP	2	9AJ353;9BJ353	WD	fragment of glycerol-3-phosphate O-acyltransferase (part 1)
prmC →	1289292	T→A	L58Q (CTA→CAA)	non synonymous SNP	2	11AJ202;11BJ353	WD	N5-glutamine methyltransferase, modifies release factors RF-1 and RF-2
ptrB ←	1840662	Δ1 bp	coding (14/2061 nt)	deletion	1	17BJ353	WD	protease II
rho →	4150124	C→T	R323R (CGC→CGT)	synonymous SNP	1	19AJ353	WD	transcription termination factor
rluD ←	2742901	Δ5 bp	coding (927-931/981 nt)	deletion	1	19BJ202	WD	23S rRNA pseudouridine synthase
	2743129	G→A	Q235* (CAG→TAG)	stop SNP	2	23BJ202;23BJ353	WD	
	2743262	Δ1 bp	coding (570/981 nt)	deletion	2	13BJ83;13BJ353	WD	
	2743277	(TGGTTCATCT)1→2	coding (555/981 nt)	duplication	2	11AJ202;11BJ353	WD	
	2743407	G→A	T142I (ACC→ATC)	non synonymous SNP	5	17AJ202;17AJ353;17BJ83;17BJ202;17BJ353	WD	
	2743564	Δ6 bp	coding (263-268/981 nt)	deletion	1	6BJ353	WD	
	2743564	Δ6 bp	coding (263-268/981 nt)	deletion	2	19BJ353;23BJ83	WD	
	2743789	G→A	Q15* (CAA→TAA)	stop SNP	1	19AJ83	WD	
rpoZ →	3926631	G→A	E53K (GAA→AAA)	non synonymous SNP	3	13AJ202;13AJ353;13BJ202	WD	RNA polymerase, omega subunit
rtcB ←	3663618	C→T	A306T (GCT→ACT)	non synonymous SNP	1	19AJ202	WD	conserved hypothetical protein
sbmA →	464122	(T)6→5	coding (17/1221 nt)	deletion	1	6AJ353	WD	transporter involved in cell envelope modification
scrA →	2892821	G→A	G395S (GGC→AGC)	non synonymous SNP	1	17AJ353	WD	PTS system sucrose-specific EIIBC component (EIIBC-Scr) (EII-Scr)
tdcC ←	3380093	C→T	A150T (GCT→ACT)	non synonymous SNP	1	13AJ353	WD	L-threonine/L-serine transporter
trmA ←	4374330	C→T	A315T (GCG→ACG)	non synonymous SNP	1	23AJ202	WD	tRNA (uracil-5-)-methyltransferase
ugpA ←	3709198	G→A	L13L (CTG→TTG)	synonymous SNP	2	9AJ353;9BJ353	WD	glycerol-3-phosphate transporter subunit ; membrane component of ABC superfamily

valS ←	4721635	C→T	D244N (GAT→AAT)	non synonymous SNP	1	17AJ202	WD	valyl-tRNA synthetase
waaL ←	3904184	G→A	H144Y (CAT→TAT)	non synonymous SNP	1	23BJ353	WD	lipid A-core surface O-antigen ligase
wcaD ←	2167748	C→T	A318A (GCG→GCA)	synonymous SNP	1	11AJ353	WD	putative colanic acid polymerase
yaeT →	203161	C→T	A343V (GCG→GTG)	non synonymous SNP	1	9AJ202	WD	outer membrane protein assembly factor
ybaO →	535661	C→T	Q43* (CAG→TAG)	stop SNP	1	11BJ353	WD	putative DNA-binding transcriptional regulator with homology to Lrp
ybbJ ←	584518	G→A	A67A (GCC→GCT)	synonymous SNP	1	17BJ202	WD	conserved hypothetical protein; putative inner membrane protein
ybiN →	858483	C→G	S136R (AGC→AGG)	non synonymous SNP	1	23AJ353	WD	putative AdoMet-dependent methyltransferase
ycaL → / → *cmk*	971615	G→A	intergenic (+38/-135)	intergenic SNP	1	3AJ353	WD	putative lipoprotein with metallohydrolase domain/cytidylate kinase
ydcS →	1485767	A→T	M79L (ATG→TTG)	non synonymous SNP	1	13BJ83	WD	putative spermidine/putrescine transporter subunit ; periplasmic-binding component of ABC superfamily transporter
ydhS →	1662427	G→A	Q116Q (CAG→CAA)	synonymous SNP	3	19AJ202;19BJ202;19BJ353	WD	conserved hypothetical protein; putative FAD/NAD(P)-binding domain
ydjX →	1746739	C→T	L229F (CTT→TTT)	non synonymous SNP	1	9BJ83	WD	conserved hypothetical protein; putative inner membrane protein
yeeT →	358046	G→A	A18T (GCA→ACA)	non synonymous SNP	1	23BJ83	WD	conserved hypothetical protein; CP4-44 prophage
	358060	C→T	F22F (TTC→TTT)	synonymous SNP	1	23BJ83	WD	
	358162	T→C	R56R (CGT→CGC)	synonymous SNP	1	23BJ83	WD	
	3174012	C→T	C25C (TGC→TGT)	synonymous SNP	1	13BJ202	WD	

ANNEXE N°8: LISTE DES MUTATIONS DECOUVERTES DANS LES 78 ISOLATS SEQUENCES DE *E. COLI* HS DANS L'ARTICLE N°3 DE LA THESE (TABLE S3)

Gene	Position	Mutation	Mutation effect	Mutation type	Occurrence	Mice ID	Diet	Gene information
[cpsB]–[EcHS_A2182]	2,154,435	Δ7103bp	coding (588/1416 nt)	large deletion	4	84AJ351, 84BJ84, 87AJ203, 87BJ203	CD	Mannose-1-phosphate guanylyltransferase (GDP)
[cpsG]–[EcHS_A2200]	2,152,639	Δ27,065 bp	coding (946/1383 nt)	large deletion	2	90AJ351, 90BJ84	CD	31 genes
[dacD]–[EcHS_A2191]	2,127,766	IS2	coding (636/1167 nt)	IS transposition	1	72BJ84	CD	45 genes
[dacD]–[galF]	2,127,546	Δ53668bp	coding (856/1167 nt)	large deletion	1	78BJ351	CD	*57 genes*
[deoR]–[EcHS_A0921]	911,03	IS1	coding (404/732 nt)	IS transposition	2	69AJ231, 69AJ352, 69BJ352	WD	24 genes
[EcHS_A2124]–[yegH]	2,109,390	Δ73,543 bp	coding (136/162 nt)	large deletion	1	69AJ84	WD	92 genes
[EcHS_A2169]–[EcHS_A2187]	2,148,795	Δ19,038 bp	coding (1299/2127 nt)	large deletion	2	78AJ351, 78BJ203	CD	20 genes
[EcHS_A2785]–[ESCCO2738]	2,802,988	Δ3120bp	coding (128/153 nt)	large deletion	2	74AJ351, 74BJ351	CD	7 genes
[EcHS_A2785]–[mprA]	2,800,568	Δ1337bp	intergenic (+1/-126)	large deletion	1	60BJ202	WD	DNA-binding transcriptional repressor of microcin B17 synthesis and multidrug efflux/multidrug efflux system
[EcHS_A2817]–[mprA]	2,827,037	Δ1,337 bp	coding (1178/1185 nt)	large deletion	2	84AJ203, 87BJ351	CD	[EcHS_A2817], ygaZ, ygaH, [mprA]
[gntT]–[glpR]	3,591,863	Δ12,822 bp	coding (963/1317 nt)	large deletion	2	90BJ203, 90BJ351	CD	[gntT], malQ, malP, malT, rtcA, rtcB, rtcR, [glpR]
[insA]–[kdgR]	1,928,293	Δ3696bp	coding (784/792 nt)	large deletion	2	66AJ322, 66BJ322	WD	DNA-binding transcriptional repressor KdgR
[insA]–yobF	1,927,210	Δ14556bp	intergenic (+15/-188)	large deletion	1	90AJ351	CD	18 genes
[malG]–[malF]	4,274,718	Δ1,281 bp	coding (544/891 nt)	large deletion	4	84AJ351, 84BJ84, 87AJ203, 87BJ203	CD	[malG], [malF]
[mngB]–EcHS_A0784	795,175	Δ3704bp	intergenic (-55/-10)	large deletion	1	60BJ202	WD	6 genes
[mngB]–[ybgE]	801,406	Δ6940bp	coding (245/294 nt)	large deletion	2	63BJ84, 63BJ203	WD	11 genes
[ugd]–[insB]	2,157,053	Δ21,948 bp	coding (955/1167 nt)	large deletion	2	90BJ203, 90BJ351	CD	21 genes
[yeeA]–[EcHS_A2171]	2,125,896	Δ26,098 bp	coding (551/1059 nt)	large deletion	1	74BJ203	CD	26 genes
[yeeE]–[EcHS_A2200]	2,131,128	Δ48,622 bp	coding (253/1059 nt)	large deletion	1	87AJ84	CD	52 genes
[yeeF]–[gnd]	2,131,575	Δ30,294 bp	coding (1342/1359 nt)	large deletion	2	72AJ351, 72BJ351	CD	33 genes
[ykgN]–[EcHS_A0326]	2,131,576	Δ32,877 bp	large deletion of 53 genes	large deletion	3	72BJ351,72AJ351, 78BJ351	CD	53 genes
aapJ → / → aapQ	3,464,462	G→A	intergenic (+52/-16)	intergenic SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	amino-acid transporter subunit ; periplasmic-binding component of ABC superfamily/amino-acid transporter subunit ; membrane component of ABC superfamily

aas ← / → *galR*	2,994,632	(A)8→9	intergenic (-298/-288)	duplication	2	72AJ351; 72BJ351	CD	fused 2-acylglycerophospho-ethanolamine acyl transferase ; acyl-acyl carrier protein synthetase/DNA-binding transcriptional repressor
acnA →	1,384,856	A→G	Y488C (TAT→TGT)	non synonymous SNP	2	79AJ84; 79AJ203	CD	aconitate hydratase 1
acnB →	133,733	G→A	P160P (CCG→CCA)	synonymous SNP	1	72BJ203	CD	bifunctional aconitate hydratase 2 and 2-methylisocitrate dehydratase
acrF →	3,460,158	T→C	R261R (CGT→CGC)	synonymous SNP	1	79BJ203	CD	multidrug efflux system protein
adiA ←	4,363,797	A→G	V250A (GTC→GCC)	non synonymous SNP	1	72BJ351	CD	biodegradative arginine decarboxylase
aes ← / → *gsk*	566,090	T→C	intergenic (-52/-100)	intergenic SNP	1	69BJ352	WD	acetyl esterase/inosine/guanosine kinase
aidB →	4,440,213	C→T	G401G (GGC→GGT)	synonymous SNP	2	72AJ351; 72BJ351	CD	isovaleryl CoA dehydrogenase
allD ← / → *fdrA*	606,031	(A)7→6	intergenic (-91/-226)	deletion	1	72BJ203	CD	ureidoglycolate dehydrogenase/putative acyl-CoA synthetase with NAD(P)-binding Rossmann-fold domain
aphA →	4,301,436	(A)7→6	pseudogene (259/333 nt)	deletion	1	72BJ203	CD	fragment of acid phosphatase/phosphotransferase, class B, non-specific (part 1)
appA →	1,103,259	C→T	P343P (CCC→CCT)	synonymous SNP	1	72BJ203	CD	phosphoanhydride phosphorylase
arcA ←	4,641,595	T→C	T200A (ACT→GCT)	non synonymous SNP	2	79BJ203; 79BJ351	CD	DNA-binding response regulator in two-component regulatory system with ArcB or CpxA
	4,936,786	G→A	H226Y (CAC→TAC)	non synonymous SNP	1	60AJ352	WD	
aroB ←	3,562,194	T→C	Y217C (TAC→TGC)	non synonymous SNP	1	79BJ351	CD	3-dehydroquinate synthase
aroF ← / → *yfiL*	2,780,811	G→A	intergenic (-48/-162)	intergenic SNP	2	57AJ202; 57AJ352	WD	3-deoxy-D-arabino-heptulosonate-7-phospha te synthase, tyrosine-repressible/conserved hypothetical protein
arsC → / → *EcHS_A3706*	3,693,191	(C)7→6	intergenic (+18/-37)	deletion	2	79BJ203; 79BJ351	CD	arsenate reductase/conserved hypothetical protein
artQ ←	962,100	G→A	A129V (GCG→GTG)	non synonymous SNP	1	79AJ203	CD	arginine transporter subunit ;membrane component of ABC superfamily
ascB →	2,859,232	G→A	P275P (CCG→CCA)	synonymous SNP	2	72AJ351; 72BJ351	CD	cryptic 6-phospho-beta-glucosidase
ascF →	2,857,785	G→A	V282I (GTT→ATT)	non synonymous SNP	1	72BJ203	CD	fused cellobiose/arbutin/salicin-specific PTS enzymes: IIB component ; IIC component
asnB ←	727,030	(GC)5→4	coding (1157-1158/1665 nt)	deletion	2	57BJ202; 57BJ352	WD	asparagine synthetase B
asnS/pncB	1,050,673	SNP	pseudogene (1358/1362 nt)	SNP	2	66AJ322, 66BJ322	WD	asparaginyl tRNA synthetase/nicotinate phosphoribosyltransferase
aspA ← / → *fxsA*	4,392,436	T→C	intergenic (-86/-251)	intergenic SNP	2	72AJ351; 72BJ351	CD	aspartate ammonia-lyase/inner membrane protein
aspC ← / ← *ompF*	1,047,463	(A)8→9	intergenic (-137/+48)	duplication	1	74AJ351	CD	aspartate aminotransferase, PLP-dependent/outer membrane porin 1a (Ia;b;F)
	1,047,465	A→T	intergenic (-139/+46)	intergenic SNP	1	78BJ351	CD	

astD ←	1,847,272	C→T	R182H (CGT→CAT)	non synonymous SNP	1	84BJ351	CD	succinylglutamic semialdehyde dehydrogenase
atoC →	2,363,230	A→G	E128G (GAG→GGG)	non synonymous SNP	1	79AJ351	CD	fused response regulator of ato opeon, in two-component system with AtoS: response regulator ; sigma54 interaction protein
	2,363,233	T→G	I129S (ATC→AGC)	non synonymous SNP	3	78AJ203; 84BJ203; 87AJ351	CD	
atoC → / → *atoD*	2,364,381	A→G	intergenic (+148/-48)	intergenic SNP	17	66AJ84; 72AJ84; 72AJ203; 72AJ351; 72BJ203; 72BJ351; 74AJ203; 78AJ203; 79AJ84; 79AJ203; 79AJ351; 84BJ203; 87AJ84; 87AJ351; 87BJ84; 90AJ84; 90AJ203	WD (1)/CD (16)	fused response regulator of ato opeon, in two-component system with AtoS: response regulator ; sigma54 interaction protein/acetyl-CoA:acetoacetyl-CoA transferase, alpha subunit
atpE ←	3,948,184	G→A	A62V (GCT→GTT)	non synonymous SNP	2	63BJ84; 63BJ203	WD	F0 sector of membrane-bound ATP synthase, subunit c
barA →	2,934,121	G→A	A286T (GCC→ACC)	non synonymous SNP	2	79BJ203; 79BJ351	CD	hybrid sensory histidine kinase, in two-component regulatory system with UvrY
bglF ← / ← *bglG*	3,932,819	(A)9→10	intergenic (-42/+92)	duplication	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	fused beta-glucoside-specific PTS enzymes: IIA component ; IIB component ; IIC component/transcriptional antiterminator of the bgl operon
bglH ←	1,640,012	G→A	pseudogene (210/1401 nt)	SNP	1	74AJ203	CD	fragment of carbohydrate-specific outer membrane porin, cryptic (part 1)
	3,929,327	G→A	A25V (GCT→GTT)	non synonymous SNP	1	74BJ351	CD	carbohydrate-specific outer membrane porin, cryptic
bioD → / → *uvrB*	839,893	T→C	intergenic (+343/-236)	intergenic SNP	1	72AJ203	CD	dethiobiotin synthetase/excinulease of nucleotide excision repair, DNA damage recognition component
bipA → / → *yihL*	4,090,419	(T)9→8	intergenic (+50/-167)	deletion	1	79AJ203	CD	GTP-binding protein/putative DNA-binding transcriptional regulator
cadC ←	4,385,471	(A)8→7	coding (487/1539 nt)	deletion	2	57AJ202; 57AJ352	WD	DNA-binding transcriptional activator
ccmA	2,338,357	putative transposase	coding (261/618 nt)	IS transposition	1	78BJ351	CD	heme exporter subunit ; ATP-binding component of ABC superfamily
ccmH ←	2,332,596	G→A	A265V (GCG→GTG)	non synonymous SNP	1	90BJ351	CD	heme lyase, CcmH subunit
cdaR →	181,784	G→A	E151K (GAA→AAA)	non synonymous SNP	1	78AJ84	CD	DNA-binding transcriptional activator
	181,917	C→T	P195L (CCC→CTC)	non synonymous SNP	2	84AJ203; 87BJ351	CD	
chbF ← / ← *chbR*	1,836,632	(C)8→7	intergenic (-51/+55)	deletion	2	72AJ351; 72BJ351	CD	cryptic phospho-beta-glucosidase, NAD(P)-binding/DNA-binding transcriptional dual regulator

citA →	688,756	C→T	R147R (CGC→CGT)	synonymous SNP	1	90AJ203	CD	sensory histidine kinase in two-component regulatory system with citB
cls ←	1,354,401	A→G	T434T (ACT→ACC)	synonymous SNP	1	79BJ351	CD	cardiolipin synthase 1
cobB →	1,243,846	C→T	A207A (GCC→GCT)	synonymous SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	deacetylase of acetyl-CoA synthetase, NAD-dependent
cobS-[EcHS_A2191]	2,103,705	Δ69885	coding (120/744 nt)	large deletion	1	87BJ84	CD	88 genes
cobU ←	2,104,349	(C)6→5	coding (18/546 nt)	deletion	1	74AJ351	CD	bifunctional cobinamide kinase and cobinamide phosphate guanylyltransferase
cpsB	2,154,560	NA	coding (463/1416 nt)	deletion	1	79AJ351	CD	Mannose-1-phosphate guanylyltransferase (GDP)
	2,154,359	G→A	Q222* (CAG→TAG)	stop SNP	2	84AJ203; 87BJ351	CD	
cpsG ←	2,152,682	Δ2 bp	coding (901-902/1383 nt)	deletion	2	63AJ352; 63BJ352	WD	Phosphomannomutase (EC 5.4.2.8) (PMM)
	2,153,455	(C)8→9	coding (129/1383 nt)	duplication	10	57AJ83; 72AJ84; 72BJ203; 74AJ203; 79BJ84; 84BJ351; 90AJ84; 90AJ203; 90BJ203; 90BJ351	WD (1)/CD (9)	
	2,153,455	(C)8→7	coding (129/1383 nt)	deletion	9	57AJ202; 57AJ352; 63AJ203; 63BJ203; 79AJ84; 79AJ203; 79BJ203; 79AJ351; 79BJ351	WD (4)/CD (5)	
	2,153,455	(C)8→11	coding (129/1383 nt)	duplication	3	74AJ351; 74BJ351; 84AJ84	CD	
cpxA ←	4,132,426	C→T	A187T (GCA→ACA)	non synonymous SNP	1	66AJ84	WD	sensory histidine kinase in two-component regulatory system with CpxR
	4,132,511	G→T	N158K (AAC→AAA)	non synonymous SNP	1	60BJ83	WD	
	4,132,765	C→T	A74T (GCG→ACG)	non synonymous SNP	6	63AJ352; 63BJ352; 66AJ203; 66BJ203; 66AJ322; 66BJ322	WD	
	4,132,780	G→A	R69W (CGG→TGG)	non synonymous SNP	2	63AJ203; 66BJ84	WD	
	4,132,968	G→T	T6N (ACC→AAC)	non synonymous SNP	1	60BJ352	WD	
cpxP →	4,133,952	C→T	Q42* (CAG→TAG)	stop SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	periplasmic protein combats stress
	4,134,036	C→T	Q70* (CAG→TAG)	stop SNP	6	49AJ82; 49AJ201; 49AJ352; 49BJ82; 49BJ201; 49BJ352	WD	

	4,134,252	C→T	Q142* (CAG→TAG)	stop SNP	1	69AJ352	WD	
	4,134,270	C→T	Q148* (CAA→TAA)	stop SNP	2	60BJ202; 63AJ84	WD	
cpxP → / → *fieF*	4,134,396	(T)8→7	intergenic (+67/-82)	deletion	1	74AJ203	CD	periplasmic protein combats stress/zinc transporter
cpxR ← / → *cpxP*	4,133,755	(C)5→4	intergenic (-76/-74)	deletion	1	69BJ84	WD	DNA-binding response regulator in two-component regulatory system with CpxA/periplasmic protein combats stress
creC →	4,639,694	G→A	A371A (GCG→GCA)	synonymous SNP	1	57BJ83	WD	sensory histidine kinase in two-component regulatory system with CreB or PhoB, regulator of the CreBC regulon
cspE → / ← *crcB*	693,631	(T)9→8	intergenic (+49/+6)	deletion	4	72AJ84; 72AJ351; 72BJ351; 79AJ351	CD	DNA-binding transcriptional repressor/conserved hypothetical protein; putative inner membrane protein associated with chromosome condensation
cstA → / → *ybdD*	668,098	(G)7→8	intergenic (+18/-165)	duplication	2	79AJ84; 79AJ203	CD	carbon starvation protein/conserved hypothetical protein
cueO →	140,217	A→G	H512R (CAC→CGC)	non synonymous SNP	1	79BJ351	CD	multicopper oxidase (laccase)
cusA →	635,432	G→A	A209T (GCG→ACG)	non synonymous SNP	1	84BJ351	CD	copper/silver efflux system, membrane component
cvpA/dedD	2,472,550	IS110	intergenic (-216/+43)	IS transposition	4	66BJ203, 78BJ351, 90BJ203, 90BJ351	WD (1)/ CD (3)	membrane protein required for colicin V production/conserved hypothetical protein
cvrA ←	1,289,569	A→G	R314R (CGT→CGC)	synonymous SNP	1	84BJ351	CD	putative cation/proton antiporter
cydA →	799,036	Δ1 bp	coding (725/1569 nt)	deletion	1	60AJ202	WD	cytochrome d terminal oxidase, subunit I
	798,72	C→T	S137L (TCA→TTA)	non synonymous SNP	2	90BJ203; 90BJ351	CD	
	799,032	C→T	S241F (TCC→TTC)	non synonymous SNP	1	84BJ351	CD	
	799,188	G→A	G293D (GGC→GAC)	non synonymous SNP	5	60AJ83; 84AJ351; 84BJ84; 87AJ203; 87BJ203	WD (1)/CD (4)	
	799,285	C→A	Y325* (TAC→TAA)	stop SNP	5	79AJ84; 79AJ203; 79AJ351; 79BJ203; 79BJ351	CD	
cydB →	799,992	G→A	G33D (GGC→GAC)	non synonymous SNP	1	66BJ84	WD	cytochrome d terminal oxidase, subunit II
	800,067	A→G	D58G (GAC→GGC)	non synonymous SNP	1	79BJ84	CD	
cynX →	425,333	G→A	V45V (GTG→GTA)	synonymous SNP	1	57BJ352	WD	putative cyanate transporter
cyoB ←	516,434	G→A	A91V (GCG→GTG)	non synonymous SNP	1	63BJ352	WD	cytochrome o ubiquinol oxidase subunit I
cyoE ←	513,716	G→A	N18N (AAC→AAT)	synonymous SNP	1	84BJ351	CD	protoheme IX farnesyltransferase
cysJ ←	2,908,202	G→T	P310H (CCT→CAT)	non synonymous SNP	1	84BJ351	CD	sulfite reductase, alpha subunit, flavoprotein

cysS →	614,197	A→G	E22E (GAA→GAG)	synonymous SNP	1	79AJ203	CD	cysteinyl-tRNA synthetase
cysU ←	2,569,576	G→A	G167G (GGC→GGT)	synonymous SNP	1	57AJ352	WD	sulfate/thiosulfate transporter subunit ; membrane component of ABC superfamily
D →	947,870	A→G	D361G (GAT→GGT)	non synonymous SNP	1	74AJ203	CD	Late control gene D protein (GpD)
dcuB ← / ← *dcuR*	4,372,770	T→C	intergenic (-4/+567)	intergenic SNP	1	79BJ351	CD	C4-dicarboxylate antiporter/DNA-binding response regulator in two-component regulatory system with DcuS
dcuS ←	4,374,975	G→A	A237V (GCC→GTC)	non synonymous SNP	1	79AJ351	CD	sensory histidine kinase in two-component regulatory system with DcuR, regulator of anaerobic fumarate respiration
ddpB ←	1,576,930	C→T	G271D (GGC→GAC)	non synonymous SNP	1	79BJ203	CD	D-ala-D-ala transporter subunit ; membrane component of ABC superfamily
deoR	911,022	IS1	coding (549/759 nt)	IS transposition	2	69BJ231, 69BJ352	WD	DNA-binding transcriptional repressor
	911,114	Δ1 bp	coding (457/759 nt)	deletion	1	57BJ352	WD	
deoR ← / ← *ybjG*	911,617	+A	intergenic (-47/+11)	insertion	1	57AJ352	WD	DNA-binding transcriptional repressor/undecaprenyl pyrophosphate phosphatase
dfp →	3,843,194	A→G	T295A (ACG→GCG)	non synonymous SNP	2	72AJ351; 72BJ351	CD	fused 4'-phosphopantothenoylcysteine decarboxylase ; phosphopantothenoylcysteine synthetase, FMN-binding
dhaR →	1,300,735	C→T	R411W (CGG→TGG)	non synonymous SNP	2	72AJ351; 72BJ351	CD	DNA-binding transcriptional regulator, dihydroxyacetone metabolism
dnaK →	12,685	C→T	A175V (GCT→GTT)	non synonymous SNP	1	69BJ352	WD	chaperone Hsp70, co-chaperone with DnaJ
dos ←	1,582,501	Δ1 bp	coding (69/2424 nt)	deletion	1	72BJ351	CD	cAMP phosphodiesterase, heme-regulated
dps ← / ← *EcHS_A0869*	875,639	(A)9→8	intergenic (-124/+45)	deletion	1	79AJ351	CD	Fe-binding and storage protein/[Protein-PII] uridylyltransferase (PII uridylyl- transferase) (Uridylyl-removing enzyme) (UTase) (fragment)
dsbG ←	674,540	C→T	D39N (GAT→AAT)	non synonymous SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	periplasmic disulfide isomerase/thiol-disulphide oxidase
dsdC ←	2,508,642	A→G	S131S (TCT→TCC)	synonymous SNP	3	79AJ84; 79AJ203; 79AJ351	CD	DNA-binding transcriptional dual regulator
dusB →	3,455,403	G→A	A261T (GCG→ACG)	non synonymous SNP	2	79BJ203; 79BJ351	CD	tRNA-dihydrouridine synthase B
eamA ←	1,643,969	G→A	A79V (GCC→GTC)	non synonymous SNP	1	72BJ203	CD	cysteine and O-acetyl-L-serine efflux system
EcHS_A0229 ←	246,947	A→G	V899A (GTG→GCG)	non synonymous SNP	1	79AJ84	CD	putative ATP-dependent Clp proteinase Aec27 ATP-binding chain, with chaperone activity
EcHS_A0235 ←	255,885	C→T	D197N (GAC→AAC)	non synonymous SNP	2	57BJ202; 57BJ352	WD	conserved hypothetical protein

EcHS_A0243 →	260,815	G→A	R70H (CGC→CAC)	non synonymous SNP	1	57AJ202	WD	conserved hypothetical protein, putative VgrG protein, Encoded within repeats that are hotspots for chromosomal duplication formation; Function of protein is unknown
EcHS_A0327 →	330,887	C→T	G216G (GGC→GGT)	synonymous SNP	1	74AJ203	CD	hypothetical protein
EcHS_A0340 →	351,836	C→T	R749C (CGC→TGC)	non synonymous SNP	1	66BJ84	WD	putative cytoplasmic protein
EcHS_A0351 →	365,812	C→T	N1333N (AAC→AAT)	synonymous SNP	1	57AJ352	WD	putative adhesin
EcHS_A0366 ←	380,519	(T)8→7	coding (505/522 nt)	deletion	1	84BJ351	CD	Putative DNA recombinase similar to Type 1 fimbriae Regulatory proteins
EcHS_A0434 →	454,836	T→C	pseudogene (642/1527 nt)	SNP	2	72AJ351; 72BJ351	CD	fragment of Putative flagellin-like structural protein similar to yaiT; putative exported protein (part 1)
EcHS_A0702 →	716,469	C→T	pseudogene (120/798 nt)	SNP	1	74AJ203	CD	fragment of conserved hypothetical protein; putative membrane protein (part 1)
EcHS_A0724 ←	736,326	(T)9→8	coding (4/144 nt)	deletion	1	84BJ351	CD	hypothetical protein
EcHS_A0756	767,105	IS110	coding (67/120 nt)	IS transposition	1	74AJ84	CD	hypothetical protein
EcHS_A0784/cydA	798,113	IS1	intergenic (-216/-198)	IS transposition	35	49AJ82, 49AJ201, 49AJ352, 49BJ82, 49BJ201, 49BJ352, 60BJ83, 60BJ352, 63AJ84, 63AJ203, 63AJ352, 63BJ352, 66AJ84, 66AJ203, 66AJ322, 66BJ203, 66BJ322, 69AJ84, 69AJ231, 69AJ352, 69BJ84, 69BJ231, 69BJ352, 72AJ84, 72AJ203, 72AJ351, 72BJ84, 72BJ203, 72BJ351, 78AJ203, 78AJ351, 78BJ84, 78BJ203, 78BJ351, 87BJ351	WD (23) / CD (12)	hypothetical protein/cytochrome d terminal oxidase, subunit I
EcHS_A0923 ←	927,617	(T)9→8	coding (956/1026 nt)	deletion	1	84BJ351	CD	putative phage portal protein

EcHS_A1402 → / ← EcHS_A1403	1,398,756	(G)7→8	intergenic (+44/+12)	duplication	3	57AJ352; 72BJ351; 79BJ351	WD (1)/CD (2)	hypothetical protein/conserved hypothetical protein
	1,398,756	(G)7→6	intergenic (+44/+12)	deletion	2	57BJ202; 57BJ352	WD	
EcHS_A1557 →	1,566,100	G→A	G418S (GGT→AGT)	non synonymous SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	Formate dehydrogenase, nitrate-inducible, major subunit (Formate dehydrogenase-N subunit alpha) (FDH-N subunit alpha) (Anaerobic formate dehydrogenase major subunit) (fragment)
EcHS_A1560	1,568,953	IS110	coding (66/153 nt)	IS transposition	1	78BJ203	CD	hypothetical protein
EcHS_A1592 ←	1,610,954	(C)7→9	coding (3955/5421 nt)	duplication	3	72AJ351; 72BJ351; 74BJ351	CD	conserved hypothetical protein; putative exported protein
	1,611,078	(C)8→9	coding (3831/5421 nt)	duplication	7	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ352; 79AJ351; 79BJ351	WD (5)/CD (2)	
	1,613,595	(C)6→7	coding (1314/5421 nt)	duplication	1	57BJ352	WD	
EcHS_A1615 ←	1,637,980	T→A	N182I (AAT→ATT)	non synonymous SNP	2	57AJ202; 57AJ352	WD	putative 6-phospho-beta-glucosidase (EC 3.2.1.86)
EcHS_A1624 → / ← ydeH	1,645,841	Δ1 bp	intergenic (+33/+186)	deletion	1	90BJ351	CD	conserved hypothetical protein/conserved hypothetical protein
EcHS_A1648/blc	1,672,507	IS1	intergenic (-27/-166)	IS transposition	1	74BJ351	CD	Conserved Hypothetical protein from bacteriophage origin/Outer membrane lipoprotein blc precursor (LIPOCALIN) from bacteriophage origin
EcHS_A1801 →	1,820,715	T→C	pseudogene (337/1245 nt)	SNP	1	84BJ351	CD	fragment of putative ankyrin repeat regulatory protein (part 2)
EcHS_A1868 → / → yeaG	1,884,685	T→C	intergenic (+15/-200)	intergenic SNP	1	79BJ351	CD	hypothetical protein/conserved hypothetical protein; putative nucleoside triphosphate hydrolase domain
EcHS_A1874 ←	1,892,067	T→C	E47E (GAA→GAG)	synonymous SNP	2	49AJ352; 49BJ352	WD	conserved hypothetical protein
EcHS_A2081	2,068,917	IS1	coding (937/1092 nt)	IS transposition	2	90BJ203, 90BJ351	CD	conserved hypothetical protein
EcHS_A2087 ←	2,073,336	(C)6→7	coding (446/1086 nt)	duplication	1	57BJ202	WD	putative bacteriophage tail protein
EcHS_A2089 ←	2,074,566	G→A	P11L (CCT→CTT)	non synonymous SNP	1	84BJ351	CD	Putative phage protein (modular protein)
EcHS_A2093 ←	2,077,901	G→A	Q17* (CAG→TAG)	stop SNP	1	79AJ351	CD	Putative phage protein (fragment)
EcHS_A2127 ←	2,110,118	(C)9→7	pseudogene (22-23/276 nt)	deletion	1	79AJ351	CD	fragment of conserved hypothetical protein, putative transposase (fragment) (partial)
	2,110,119	(C)9→8	pseudogene (22/276 nt)	deletion	5	57AJ83; 72AJ351; 72BJ351; 74AJ203; 74BJ351	WD (1)/CD (4)	
	2,110,119	(C)9→10	pseudogene (22/276 nt)	duplication	1	79BJ203	CD	

EcHS_A2127 ← / ← *EcHS_A2127*	2,110,927	G→A	intergenic (-29/+122)	intergenic SNP	1	84BJ351	CD	fragment of putative transposase, ISL3 family (part 2)/fragment of putative transposase, ISL3 family (part 1)
EcHS_A2128 ←	2,111,568	(C)7→6	coding (372/411 nt)	deletion	2	57AJ202; 57AJ352	WD	conserved hypothetical protein
	2,111,568	(C)7→8	coding (372/411 nt)	duplication	1	74BJ351	CD	
EcHS_A2132 ← / → *EcHS_A2133*	2,114,291	T→C	intergenic (-61/-174)	intergenic SNP	1	74BJ351	CD	conserved hypothetical protein/Rhomboid domain-containing protein 2 (fragment)
EcHS_A2138 →	2,119,672	A→G	D251G (GAC→GGC)	non synonymous SNP	2	79AJ84; 79AJ203	CD	conserved hypothetical protein with GTPase domain
EcHS_A2166	2,144,004	IS1	coding (228/1116 nt)	IS transposition	2	66AJ203, 66BJ203	WD	Mannosyl transferase (modular protein), wbdC
	2,143,216	C→T	G339D (GGC→GAC)	non synonymous SNP	1	57AJ83	WD	
	2,143,280	Δ11 bp	coding (942-952/1116 nt)	deletion	2	66AJ322; 66BJ322	WD	
EcHS_A2167 ←	2,145,240	(G)5→4	coding (147/1146 nt)	deletion	1	84BJ351	CD	WbdB (mannosyltransferase B)
EcHS_A2168 ←	2,146,970	C→T	C304Y (TGC→TAC)	non synonymous SNP	2	57AJ202; 57AJ352	WD	Mannosyltransferase A
EcHS_A2170 ←	2,150,544	(C)6→5	coding (864/1296 nt)	deletion	3	79AJ203; 79BJ203; 79BJ351	CD	ATP binding component of ABC-transporter
EcHS_A2179 ←	2,158,622	Δ1 bp	coding (120/555 nt)	deletion	1	84BJ351	CD	dTDP-4-dehydrorhamnose 3,5-epimerase (dTDP-4-keto-6- deoxyglucose 3,5-epimerase) (dTDP-L-rhamnose synthetase)
EcHS_A2181 ←	2,160,387	+G	coding (161/870 nt)	insertion	1	79AJ351	CD	Glucose-1-phosphate thymidylyltransferase (dTDP-glucose synthase) (dTDP-glucose pyrophosphorylase)
EcHS_A2182 ←	2,161,394	G→A	A82V (GCA→GTA)	non synonymous SNP	55	49AJ82; 49AJ201; 49AJ352; 49BJ82; 49BJ201; 49BJ352; 57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352; 60AJ83; 60BJ83; 60BJ202; 60BJ352; 63AJ84; 63AJ203; 63AJ352; 63BJ84; 63BJ203; 63BJ352; 66AJ84; 66AJ203; 66AJ322; 66BJ84; 66BJ203; 66BJ322; 69BJ84; 72AJ84; 72AJ203; 72BJ203; 74AJ84; 74AJ203 ;	WD (29)/ CD (26)	dTDP-glucose 4,6-dehydratase

						74AJ351; 74BJ84; 74BJ203; 74BJ351; 78AJ84; 78AJ203; 78BJ84; 79AJ84; 79AJ203; 79AJ351; 79BJ84; 79BJ203 ; 79BJ351; 84AJ84; 84AJ203; 84BJ203; 84BJ351; 87AJ351; 87BJ351; 90AJ84; 90AJ203		
	2,166,929	IS1	coding (148/1251 nt)	IS transposition	1	74BJ203	CD	
EcHS_A2186	2,166,193	(C)6→5	coding (884/1251 nt)	deletion	9	57BJ202; 57BJ352; 72AJ351; 72BJ351; 74AJ351; 74BJ351; 79AJ351; 79BJ203; 79BJ351	WD (2)/CD (7)	putative transmembrane protein
	2,166,193	(C)6→7	coding (884/1251 nt)	duplication	2	79AJ84; 79AJ203	CD	
	2,166,760	(A)9→8	coding (317/1251 nt)	deletion	4	57AJ202; 57AJ352; 72AJ351; 72BJ351	WD (2)/CD(2)	
EcHS_A2188 ← / ← EcHS_A2189	2,170,022	(A)8→9	intergenic (-349/+98)	duplication	1	66AJ203	WD	hypothetical protein/O antigen biosynthesis rhamnosyltransferase rfbN
EcHS_A2190 ←	2,171,581	(T)6→7	coding (916/1428 nt)	duplication	1	72AJ84	CD	UDP-galactose-lipid carrier transferase
EcHS_A2191	2,172,919	IS1	coding (1849/2169 nt)	IS transposition	24	63AJ352, 63BJ352, 66AJ203, 66AJ322, 66BJ203, 69BJ84, 57AJ202, 57AJ352, 66BJ322, 69AJ231,	WD (11) / CD (13)	conserved hypothetical protein

						69AJ352, 69BJ231, 69BJ352, 72AJ351, 72BJ84, 72BJ351, 74BJ84, 74BJ203, 74BJ351, 78BJ203, 78BJ351, 79BJ203, 79BJ351, 87BJ84		
EcHS_A2191 ←	2,173,894	G→A	R292W (CGG→TGG)	non synonymous SNP	1	57AJ83	WD	conserved hypothetical protein
	2,174,359	Δ1 bp	coding (409/2169 nt)	deletion	1	79AJ351	CD	
	2,174,399	Δ1 bp	coding (369/2169 nt)	deletion	2	57BJ202; 57BJ352	WD	
	2,174,415	(A)5→6	coding (353/2169 nt)	duplication	2	72AJ351; 72BJ351	CD	
EcHS_A2193	2,175,598	IS1	coding (752/1137 nt)	IS transposition	2	57AJ202, 57AJ352	WD	conserved hypothetical protein
	2,175,642	IS1	coding (709/1137 nt)	IS transposition	1	78BJ203	CD	
	2,175,695	IS1	coding (663/1137 nt)	IS transposition	1	74BJ351	CD	
	2,175,288	A→G	L357P (CTT→CCT)	non synonymous SNP	1	57BJ83	WD	
EcHS_A2194 ←	2,177,517	(C)7→8	coding (499/1515 nt)	duplication	1	84BJ351	CD	conserved hypothetical protein
EcHS_A2196 ←	2,178,308	G→A	A19V (GCG→GTG)	non synonymous SNP	2	79AJ84; 79AJ203	CD	hypothetical protein
	2,178,314	G→A	P17L (CCA→CTA)	non synonymous SNP	1	74AJ351	CD	
EcHS_A2307 → / → *set B*	2,306,004	T→C	intergenic (+34/-37)	intergenic SNP	1	79BJ351	CD	conserved hypothetical protein/lactose/glucose efflux system
EcHS_A2381 →	2,393,563	C→T	R31* (CGA→TGA)	stop SNP	1	74BJ351	CD	hypothetical protein
EcHS_A2644 ←	2,656,130	(G)8→7	coding (983/4098 nt)	deletion	3	57BJ352; 79BJ203; 79BJ351	WD	conserved hypothetical protein
	2,656,130	(G)8→9	coding (983/4098 nt)	duplication	1	72BJ203	CD	
	2,656,468	C→T	T215T (ACG→ACA)	synonymous SNP	1	72BJ203	CD	
	2,656,999	Δ1 bp	coding (114/4098 nt)	deletion	1	79BJ351	CD	
EcHS_A2648	2,661,465	IS1	coding (580/645 nt)	IS transposition	1	84BJ203	CD	Protein traJ (Relaxosome protein) (modular protein)
EcHS_A2651 ← / ← *EcHS_A2652*	2,664,328	(T)8→7	intergenic (-15/+55)	deletion	1	79AJ203	CD	Plasmid encoded RepA protein precursor/Phage transcriptional regulator, AlpA
EcHS_A2793 →	2,804,169	(A)5→6	coding (328/2253 nt)	duplication	2	57AJ202; 57AJ352	WD	putative Alpha-amylase
	2,806,084	T→C	L748P (CTC→CCC)	non synonymous SNP	1	74AJ203	CD	

EcHS_A2893	2,895,389	putative transposase	pseudogene (114/261 nt)	IS transposition	1	87BJ84	CD	fragment of conserved hypothetical protein (part 1)
EcHS_A3018 ←	3,021,810	Δ3 bp	coding (538-540/666 nt)	deletion	1	72AJ351	CD	putative type III secretion system protein EpaP
EcHS_A3158 →	3,171,113	Δ1 bp	coding (6/1173 nt)	deletion	1	72BJ203	CD	Putative 8-amino-7-oxononanoate synthase
EcHS_A3278 ← / ← *EcHS_A3279*	3,293,996	+T	intergenic (-46/+27)	insertion	1	57AJ83	WD	conserved hypothetical protein; putative exported protein/putative adhesin major subunit pilin
EcHS_A3612 → / → *gntT*	3,590,888	G→A	intergenic (+4/-12)	intergenic SNP	1	66BJ322	WD	hypothetical protein/gluconate transporter, high-affinity GNT I system
EcHS_A3893 ←	3,888,784	G→A	G125G (GGC→GGT)	synonymous SNP	1	74AJ351	WD	6-phospho-glucosidase
EcHS_A4139/kdgT	4,129,604	IS1	intergenic (-8/-95)	IS transposition	2	84AJ203, 87BJ351	CD	conserved hypothetical protein/2-keto-3-deoxy-D-gluconate transporter
EcHS_A4454 ← / → *ytfA*	4,460,374	C→T	intergenic (-47/-166)	intergenic SNP	2	49AJ201; 49BJ201	WD	putative dehydrogenase/putative transcriptional regulator
EcHS_A4459 →	4,464,782	C→T	pseudogene (156/435 nt)	SNP	3	57BJ83; 57BJ202; 57BJ352	WD	fragment of conserved hypothetical protein (partial)
EcHS_A4570 →	4,578,931	(G)8→7	pseudogene (1716/1767 nt)	deletion	4	57AJ202; 57AJ352; 74AJ351; 84AJ84	WD(2)/CD(2)	fragment of Penicillin G acylase precursor (part 1)
	4,578,931	(G)8→9	pseudogene (1716/1767 nt)	duplication	3	72AJ203; 72AJ351; 72BJ351	CD	
EcHS_A4643 → / → *dkgB*	228,422	(T)8→9	intergenic (+39/-123)	intergenic SNP	3	72AJ351; 72BJ203; 72BJ351	CD	tRNA-Asp/2,5-diketo-D-gluconate reductase B
EcHS_A4644	236,606	NA	noncoding (78/79 nt)	insertion	1	79AJ203	CD	tRNA-Asp
EcHS_A4659	807,978	NA	noncoding (76/76 nt)	insertion	1	72AJ351	CD	tRNA-Lys
EcHS_A4712 →	4,011,097	(C)8→9	noncoding (80/89 nt)	duplication	1	57AJ83	WD	tRNA-Leu
EcHS_A4715 → / → *rrl*	4,067,707	A→G	intergenic (+113/-70)	intergenic SNP	1	72AJ203	CD	tRNA-Ala/23S ribosomal RNA
EcHS_A4727 ←	4,609,409	(G)8→9	noncoding (73/89 nt)	duplication	1	72AJ203	CD	tRNA-Leu
	4,609,410	A→G	noncoding (72/89 nt)	SNP	1	74BJ351	CD	
EcHS_A4728 ←	4,609,530	(G)8→9	noncoding (72/87 nt)	duplication	1	72AJ351	CD	tRNA-Leu
EcHS_A4729 ←	4,609,645	(G)8→9	noncoding (73/89 nt)	duplication	3	57BJ83; 57BJ202; 57BJ352	WD	tRNA-Leu
ECP_4610 ←	4,816,995	G→A	R156C (CGC→TGC)	non synonymous SNP	1	60AJ202	WD	conserved hypothetical protein, putative Transcriptional regulator
eda ←	1,951,633	T→C	K34R (AAA→AGA)	non synonymous SNP	1	72AJ84	CD	multifunctional 2-keto-3-deoxygluconate 6-phosphate aldolase and 2-keto-4-hydroxyglutarate aldolase and oxaloacetate decarboxylase

efeB →	1,145,786	(A)6→7	coding (112/1272 nt)	duplication	1	74BJ351	CD	heme-binding periplasmic protein involved in iron transport; Tat-dependent exported
ego →	1,618,653	(A)7→6	coding (342/1536 nt)	deletion	1	84BJ351	CD	fused AI2 transporter subunits of ABC superfamily: ATP-binding components
elbA ← / ← *ycgX*	1,259,304	(C)11→8	intergenic (-86/+615)	deletion	5	72AJ203; 72BJ203; 74AJ351; 79AJ351; 79BJ351	CD	conserved hypothetical protein associated with lycopene formation/conserved hypothetical protein
	1,259,305	(C)11→9	intergenic (-87/+615)	deletion	1	79BJ203	CD	
	1,259,306	(C)11→10	intergenic (-88/+615)	deletion	7	57AJ83; 57AJ202; 57BJ352; 72AJ84; 72AJ351; 72BJ351; 79BJ84	WD (3) / CD (4)	
	1,259,306	(C)11→12	intergenic (-88/+615)	duplication	3	72AJ203; 79AJ351; 79BJ351	CD	
	1,259,295	NA	intergenic (-77/+626)	insertion	3	57AJ352, 74AJ203, 84BJ351	WD (1)/ CD (2)	
elbB ←	3,401,104	1 bp→CC	coding (283/654 nt)	insertion	1	79AJ203	CD	isoprenoid biosynthesis protein with amidotransferase-like domain
	3,401,111	(C)7→8	coding (276/654 nt)	deletion	7	72AJ351; 72BJ203; 72BJ351; 74AJ351; 74BJ351; 79BJ351; 84BJ351	CD	
	3,401,111	(C)7→6	coding (276/654 nt)	deletion	1	74AJ203	CD	
eno ←	2,924,862	C→T	G430D (GGC→GAC)	non synonymous SNP	1	84BJ351	CD	enolase
	2,925,021	T→C	D377G (GAC→GGC)	non synonymous SNP	2	72AJ351; 72BJ351	CD	
eprH ← / → *EcHS_A3012*	3,018,414	(C)7→6	intergenic (-12/-246)	deletion	1	84BJ351	CD	Putative Type III secretion EprH protein/conserved hypothetical protein, putative response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain
eprK ←	3,015,482	T→C	pseudogene (10/384 nt)	SNP	1	57BJ352	WD	fragment of Putative Type III secretion system lipoprotein EprK (part 2)
eptB ←	3,738,864	(C)6→7	coding (915/1692 nt)	duplication	5	57AJ352; 74AJ351; 79AJ351; 79BJ203; 79BJ351	WD (1) / CD (4)	putative metal dependent hydrolase
	3,738,864	(C)6→5	coding (915/1692 nt)	deletion	3	57BJ352, 72AJ351, 72BJ351	WD (1)/	

							CD (2)	
ESCCO0743 ← / → sdhC	781,027	C→T	intergenic (-110/-220)	intergenic SNP	1	74BJ351	CD	conserved hypothetical protein/succinate dehydrogenase, membrane subunit, binds cytochrome b556
ESCCO1408/ycjY	1,438,345	IS1	intergenic (-32/+39)	IS transposition	2	79BJ203, 79BJ351	CD	conserved hypothetical protein/putative hydrolase
ESCCO2449 →	2,504,509	(C)7→8	coding (105/153 nt)	duplication	1	79AJ203	CD	hypothetical protein
ESCCO2727 → / → ESCCO2728	2,796,070	G→A	intergenic (+115/-452)	intergenic SNP	1	57BJ83	WD	hypothetical protein/hypothetical protein
ESCCO2854/EcHS_A2916	2,921,047	Δ122 bp	intergenic (-104/-13)	deletion	1	87AJ84	CD	hypothetical protein/conserved hypothetical protein
ESCCO3118 → / → metC	3,196,888	T→C	intergenic (+19/-179)	intergenic SNP	1	74BJ351	CD	hypothetical protein/cystathionine beta-lyase, PLP-dependent
ESCCO3166 ←	3,244,419	(T)6→7	coding (159/207 nt)	duplication	1	57AJ352	WD	conserved hypothetical protein
ESCCO3623 → / → arsR	3,690,886	(T)7→8	intergenic (+386/-153)	duplication	2	74AJ351; 74BJ351	CD	conserved hypothetical protein/DNA-binding transcriptional repressor
ESCCO4090 ←	4,167,786	T→C	pseudogene (11/153 nt)	SNP	3	84BJ203; 87AJ351; 87BJ84	CD	fragment of conserved hypothetical protein (partial)
ESCCO4453 ←	4,545,529	T→C	pseudogene (171/222 nt)	SNP	1	72BJ351	CD	fragment of conserved hypothetical protein (partial)
exuR →	3,296,000	T→C	F146L (TTC→CTC)	non synonymous SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	DNA-binding transcriptional repressor
exuT →	3,288,960	C→T	Q62* (CAA→TAA)	stop SNP	1	60BJ202	WD	hexuronate transporter
	3,289,504	G→A	W243* (TGG→TAG)	stop SNP	2	49AJ352; 49BJ352	WD	
	3,289,585	G→A	W270* (TGG→TAG)	stop SNP	2	49AJ201; 49BJ201	WD	
fabD →	1,213,494	C→T	V29V (GTC→GTT)	synonymous SNP	1	84BJ351	CD	malonyl-CoA-[acyl-carrier-protein] transacylase
fadA ←	4,058,181	C→T	P284P (CCG→CCA)	synonymous SNP	1	74AJ203	CD	3-ketoacyl-CoA thiolase (thiolase I)
fdnH →	1,567,685	A→G	Y138C (TAC→TGC)	non synonymous SNP	1	57AJ352	WD	formate dehydrogenase-N, Fe-S (beta) subunit, nitrate-inducible
fdoH ←	4,110,688	G→A	Q87* (CAG→TAG)	stop SNP	1	74BJ351	CD	formate dehydrogenase-O, Fe-S subunit
fhlA →	2,872,794	G→A	A347T (GCC→ACC)	non synonymous SNP	2	74AJ351; 84AJ203; 87BJ351	CD	DNA-binding transcriptional activator
fhuF ←	4,608,832	G→A	P49P (CCC→CCT)	synonymous SNP	1	84BJ351	CD	ferric iron reductase involved in ferric hydroximate transport
fieF →	4,134,648	A→G	L57L (TTA→TTG)	synonymous SNP	1	79BJ351	CD	zinc transporter
fixC →	48,070	C→T	L211L (CTG→TTG)	synonymous SNP	1	66AJ203	WD	putative oxidoreductase with FAD/NAD(P)-binding domain
fliA ←	2,020,872	C→T	V182M (GTG→ATG)	non synonymous SNP	1	84AJ351	CD	RNA polymerase, sigma 28 (sigma F) factor

fliF →	2,033,620	A→G	T218A (ACC→GCC)	non synonymous SNP	1	74BJ351	CD	flagellar basal-body MS-ring and collar protein
fliH →	2,035,900	T→C	M98T (ATG→ACG)	non synonymous SNP	1	72AJ203	CD	flagellar biosynthesis protein
fnr ←	1,445,686	C→T	P31P (CCG→CCA)	synonymous SNP	1	72AJ203	CD	DNA-binding transcriptional dual regulator, global regulator of anaerobic growth
frlB →	3,546,079	G→A	C163Y (TGC→TAC)	non synonymous SNP	2	57AJ202; 57AJ352	WD	fructoselysine-6-P-deglycase
frlC →	3,546,877	G→A	G65S (GGC→AGC)	non synonymous SNP	1	79AJ203	CD	fructoselysine 3-epimerase
frlD →	3,547,765	A→G	H85R (CAC→CGC)	non synonymous SNP	1	79AJ351	CD	fructoselysine 6-kinase
frlD → / → *frlR*	3,548,335	T→C	intergenic (+38/-62)	intergenic SNP	1	57AJ352	WD	fructoselysine 6-kinase/putative DNA-binding transcriptional regulator
	3,548,384	IS1	intergenic (+87/-13)	IS transposition	1	60BJ202	WD	
frlR →	3,548,595	C→T	R67C (CGC→TGC)	non synonymous SNP	2	66AJ203; 66BJ203	WD	putative DNA-binding transcriptional regulator
	3,548,878	A→G	Y161C (TAC→TGC)	non synonymous SNP	1	79AJ203	CD	
frmR ← / ← *yaiO*	445,303	(G)8→7	intergenic (-139/+47)	deletion	1	84BJ351	CD	regulator protein that represses frmRAB operon/conserved hypothetical protein
fruB ←	2,304,852	C→T	L274L (CTG→CTA)	synonymous SNP	1	84BJ351	CD	fused fructose-specific PTS enzymes: IIA component ; HPr component
frwC →	4,172,715	G→A	G92R (GGG→AGG)	non synonymous SNP	1	72AJ203	CD	enzyme IIC component of PTS
fsr ←	570,085	A→G	V226A (GTA→GCA)	non synonymous SNP	1	79AJ351	CD	fosmidomycin efflux system, member of the major facilitator superfamily
fucA ←	2,951,341	G→A	A186V (GCG→GTG)	non synonymous SNP	3	57AJ83; 57AJ202; 57AJ352	WD	L-fuculose-1-phosphate aldolase
fucI →	2,954,392	G→A	W200* (TGG→TGA)	stop SNP	1	79AJ84	CD	L-fucose isomerase
fumA/manA	1,710,662	IS1	intergenic (-151/-48)	IS transposition	1	60BJ83	WD	fumarate hydratase (fumarase A), aerobic Class I/mannose-6-phosphate isomerase
fur ←	740,741	T→C	H87R (CAC→CGC)	non synonymous SNP	1	84BJ351	CD	DNA-binding transcriptional dual regulator of siderophore biosynthesis and transport
	740,844	C→T	A53T (GCT→ACT)	non synonymous SNP	1	72BJ203	CD	
gabP →	2,812,278	C→T	G194G (GGC→GGT)	synonymous SNP	1	74BJ351	CD	gamma-aminobutyrate transporter
gadB ← / ← *pqqL*	1,589,009	(T)7→8	intergenic (-139/+224)	duplication	1	57BJ83	WD	glutamate decarboxylase B, PLP-dependent/putative membrane-associated peptidase
galF ←	2,180,325	A→G	L298P (CTG→CCG)	non synonymous SNP	1	72BJ351	CD	putative subunit with GalU
garK ←	3,319,723	T→C	D223G (GAC→GGC)	non synonymous SNP	2	90BJ203; 90BJ351	CD	glycerate kinase I
	3,319,780	G→A	A204V (GCA→GTA)	non synonymous SNP	2	90AJ351; 90BJ84	CD	
	3,319,988	C→T	G135S (GGC→AGC)	non synonymous SNP	2	79BJ203; 79BJ351	CD	
	3,320,365	G→T	S9Y (TCT→TAT)	non synonymous SNP	1	74BJ203	CD	
garP ← / → *EcHS_A3317*	3,323,665	C→A	intergenic (-138/-71)	intergenic SNP	4	78AJ351; 78BJ84; 78BJ203; 78BJ351	CD	putative (D)-galactarate transporter/hypothetical protein

garR ←	3,320,585	T→C	M263V (ATG→GTG)	non synonymous SNP	2	79BJ203; 79BJ351	CD	tartronate semialdehyde reductase
gcvA ← / ← *ygdI*	2,960,903	Δ1 bp	intergenic (-127/+224)	deletion	2	57BJ202; 57BJ352	WD	DNA-binding transcriptional dual regulator/conserved hypothetical protein
	3,950,490	IS1	coding (370/624 nt)	IS transposition	4	49AJ201, 49AJ352, 49BJ201, 49BJ352	WD	
	3,950,313	C→A	E183* (GAA→TAA)	stop SNP	1	63AJ84	WD	
	3,950,319	G→A	Q181* (CAG→TAG)	stop SNP	3	72AJ351; 72BJ203; 72BJ351	CD	
	3,950,361	G→A	Q167* (CAA→TAA)	stop SNP	6	72AJ203; 78AJ351; 78BJ84; 78BJ203; 78BJ351; 84AJ84	CD	
	3,950,396	Δ25 bp	coding (440-464/624 nt)	deletion	10	60BJ83; 63BJ84; 63BJ203; 69AJ84; 74AJ351; 74BJ203; 74BJ351; 78AJ84; 90BJ203; 90BJ351	WD (4)/ CD (6)	
	3,950,441	G→A	A140V (GCT→GTT)	non synonymous SNP	1	84BJ351	CD	
	3,950,444	C→T	R139H (CGC→CAC)	non synonymous SNP	2	84AJ203; 87BJ351	CD	
gidB	3,950,469	Δ1 bp	coding (391/624 nt)	deletion	10	63AJ352; 63BJ352; 66AJ203; 66BJ203; 66AJ322; 66BJ322; 69AJ231; 69AJ352; 69BJ231; 69BJ352	WD	methyltransferase, SAM-dependent methyltransferase, glucose-inhibited cell-division protein
	3,950,535	G→A	Q109* (CAA→TAA)	stop SNP	3	78AJ203; 79BJ203; 79BJ351	CD	
	3,950,541	G→A	Q107* (CAG→TAG)	stop SNP	1	66AJ84	WD	
	3,950,624	G→A	P79L (CCA→CTA)	non synonymous SNP	5	74AJ84; 84BJ203; 87AJ203; 87AJ351; 90AJ351	CD	
	3,950,625	G→A	P79S (CCA→TCA)	non synonymous SNP	1	66BJ84	WD	
	3,950,632	Δ9 bp	coding (220-228/624 nt)	deletion	1	87AJ84	CD	
	3,950,633	G→A	P76L (CCA→CTA)	non synonymous SNP	1	84AJ351	CD	
	3,950,643	C→T	G73S (GGC→AGC)	non synonymous SNP	1	87BJ203	CD	
	3,950,649	C→T	D71N (GAT→AAT)	non synonymous SNP	1	74BJ84	CD	

	3,950,763	T→A	K33* (AAA→TAA)	stop SNP	1	63AJ203	WD	
	3,950,799	(T)5→6	coding (61/624 nt)	duplication	4	74AJ203; 79AJ84; 79AJ203; 79AJ351	CD	
	3,950,802	G→A	Q20* (CAG→TAG)	stop SNP	1	90AJ203	CD	
	3,950,858	A→G	M1A (GTG→GCG) †	non synonymous SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	
	4,106,867	G→A	Q167* (CAA→TAA)	stop SNP	1	60AJ202	WD	
gldA	4,168,333	IS1	coding (615/1104 nt)	IS transposition	1	78AJ351	CD	glycerol dehydrogenase, NAD
	4,168,500	IS1	coding (448/1104 nt)	IS transposition	1	78BJ351	CD	
	4,167,900	C→T	A350T (GCC→ACC)	non synonymous SNP	2	72AJ351; 72BJ351	CD	
	4,168,130	(T)5→4	coding (818/1104 nt)	deletion	1	57BJ202	WD	
glnE ←	3,240,237	G→A	R932C (CGT→TGT)	non synonymous SNP	1	57BJ202	WD	fused deadenylyltransferase ; adenylyltransferase for glutamine synthetase
glpF ← / → *EcHS_A4159*	4,146,185	G→A	intergenic (-86/-151)	intergenic SNP	10	57AJ83; 60AJ83; 69AJ231; 69AJ352; 69BJ84; 69BJ231; 69BJ352; 74AJ351; 79BJ84; 90AJ203	WD (7)/ CD (3)	glycerol facilitator/fragment of conserved hypothetical protein (partial)
glpR	3,604,809	IS1	coding (67/759 nt)	IS transposition	1	84AJ351	CD	DNA-binding transcriptional repressor
	3,604,176	G→A	P234S (CCG→TCG)	non synonymous SNP	1	90AJ84	CD	
	3,604,370	A→C	I169S (ATT→AGT)	non synonymous SNP	1	72BJ203	CD	
	3,604,386	G→A	R164C (CGC→TGC)	non synonymous SNP	1	87AJ203	CD	
	3,604,778	G→A	P33L (CCG→CTG)	non synonymous SNP	2	74AJ351; 74BJ351	CD	
glpT ←	2,392,618	G→A	Q247* (CAG→TAG)	stop SNP	1	60AJ83	WD	sn-glycerol-3-phosphate transporter
	2,392,865	Δ1 bp	coding (492/1359 nt)	deletion	4	69AJ231; 69AJ352; 69BJ231; 69BJ352	WD	
gltP →	4,327,282	A→G	T346A (ACG→GCG)	non synonymous SNP	2	79BJ203; 79BJ351	CD	glutamate/aspartate:proton symporter
glxK →	602,089	G→A	G209S (GGC→AGC)	non synonymous SNP	1	74AJ351	CD	glycerate kinase II
gmk →	3,851,585	G→A	R192H (CGC→CAC)	non synonymous SNP	1	57BJ83	WD	guanylate kinase
gnd ← / ← *EcHS_A2184*	2,163,419	(A)9→8	intergenic (-151/+50)	deletion	2	74AJ203; 79BJ203	CD	gluconate-6-phosphate dehydrogenase, decarboxylating/Capsule repeat unit export
gntR ←	3,625,102	A→G	S221P (TCT→CCT)	non synonymous SNP	1	84BJ351	CD	DNA-binding transcriptional repressor
	3,625,186	G→A	R193C (CGT→TGT)	non synonymous SNP	1	66BJ84	WD	

	3,625,200	G→A	A188V (GCA→GTA)	non synonymous SNP	3	63AJ352; 63BJ352; 69BJ84	WD	
	3,625,210	A→G	Y185H (TAT→CAT)	non synonymous SNP	1	74AJ351	CD	
	3,625,440	T→A	E108V (GAG→GTG)	non synonymous SNP	4	84AJ351; 84BJ84; 87AJ203; 87BJ203	CD	
	3,625,537	G→A	Q76* (CAG→TAG)	stop SNP	1	74BJ203	CD	
	3,592,099	IS1	coding (1200/1317 nt)	IS transposition	1	79BJ351	CD	
	3,590,924	G→A	G9S (GGT→AGT)	non synonymous SNP	2	72AJ351; 72BJ351	CD	
	3,590,970	G→A	G24D (GGC→GAC)	non synonymous SNP	1	79AJ203	CD	
	3,591,028	(AATGCCGCTG)1→2	coding (129/1317 nt)	duplication	2	90AJ84; 90AJ203	CD	
	3,591,080	C→T	L61F (CTT→TTT)	non synonymous SNP	1	60BJ83	WD	
	3,591,132	G→A	C78Y (TGC→TAC)	non synonymous SNP	1	74AJ84	CD	
	3,591,184	(A)6→7	coding (285/1317 nt)	duplication	2	79AJ351, 84AJ84	CD	
	3,591,191	C→T	Q98* (CAG→TAG)	stop SNP	2	63BJ84; 63BJ203	WD	
	3,591,195	G→A	W99* (TGG→TAG)	stop SNP	1	60BJ352	WD	
	3,591,196	G→A	W99* (TGG→TGA)	stop SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	
gntT	3,591,208	(GTACTG)1→2	coding (309/1317 nt)	duplication	6	63AJ352;63BJ352; 66AJ203; 66AJ322; 66BJ203; 66BJ322	WD	gluconate transporter, high-affinity GNT I system
	3,591,267	C→T	P123L (CCG→CTG)	non synonymous SNP	1	74BJ351	CD	
	3,591,370	T→A	H157Q (CAT→CAA)	non synonymous SNP	1	60AJ83	WD	
	3,591,435	Δ1 bp	coding (536/1317 nt)	deletion	2	84AJ203, 87BJ351	CD	
	3,591,630	Δ14 bp	coding (731-744/1317 nt)	deletion	4	69AJ231; 69AJ352; 69BJ231; 69BJ352	WD	
	3,591,643	(TGATCCTGCCGAAA)1→2	coding (744/1317 nt)	duplication	2	60BJ202, 69AJ84	WD	
	3,591,921	G→A	W341* (TGG→TAG)	stop SNP	4	66AJ322; 84BJ203; 87AJ351; 87BJ84	CD(3)/WD (1)	
	3,592,089	G→A	G397D (GGC→GAC)	non synonymous SNP	1	79BJ84	CD	
	3,592,098	T→G	L400R (CTG→CGG)	non synonymous SNP	6	49AJ82; 49AJ201; 49AJ352; 49BJ82; 49BJ201; 49BJ352	WD	

	3,592,146	G→A	W416* (TGG→TAG)	stop SNP	5	78AJ203; 78AJ351; 78BJ84; 78BJ203; 78BJ351	CD	
gspD →	3,501,146	(C)6→5	coding (439/1965 nt)	deletion	1	74BJ351	CD	general secretory pathway component, cryptic
guaD →	3,054,369	A→G	T363A (ACG→GCG)	non synonymous SNP	1	72AJ351	CD	guanine deaminase
gudP ← / ← *yqcA*	2,940,435	T→C	intergenic (-126/+309)	intergenic SNP	1	72AJ203	CD	putative D-glucarate transporter/putative flavoprotein
hcaD →	2,710,016	G→A	R138Q (CGA→CAA)	non synonymous SNP	1	72AJ203	CD	phenylpropionate dioxygenase, ferredoxin reductase subunit
hcaF →	2,708,058	G→A	E38K (GAA→AAA)	non synonymous SNP	1	72BJ351	CD	3-phenylpropionate dioxygenase, small (beta) subunit
hcaT ← / ← *hcaR*	2,705,465	(A)5→4	intergenic (-62/+98)	deletion	2	57AJ202; 57AJ352	WD	putative 3-phenylpropionic transporter/DNA-binding transcriptional activator of 3-phenylpropionic acid catabolism
hcr ←	971,650	G→A	P246S (CCG→TCG)	non synonymous SNP	1	60BJ202	WD	HCP oxidoreductase, NADH-dependent
hisI →	2,142,675	C→T	G69G (GGC→GGT)	synonymous SNP	1	79AJ351	CD	fused phosphoribosyl-AMP cyclohydrolase ; phosphoribosyl-ATP pyrophosphatase
hokA/insJ - insK/glyS	3,749,936	IS1	intergenic (-33/-47)	IS transposition	1	74AJ351	CD	IS150
htpG →	562,595	T→C	Y471H (TAT→CAT)	non synonymous SNP	1	57BJ83	WD	molecular chaperone HSP90 family
hybO ← / ← *yghW*	3,191,129	Δ1 bp	intergenic (-38/+151)	deletion	1	57BJ352	WD	hydrogenase 2, small subunit/conserved hypothetical protein
hyfB →	2,624,588	C→T	pseudogene (250/1404 nt)	SNP	1	74AJ203	CD	fragment of hydrogenase 4, membrane subunit (part 2)
ibpA ←	3,894,570	C→T	P8P (CCG→CCA)	synonymous SNP	1	72AJ203	CD	heat shock chaperone
idnR	4,526,898	IS1	coding (368/999 nt)	IS transposition	1	72BJ84	CD	DNA-binding transcriptional repressor, 5-gluconate-binding
	4,527,222	G→A	A15V (GCT→GTT)	non synonymous SNP	2	72AJ351; 72BJ351	CD	
idnT ←	4,528,042	G→A	P204S (CCA→TCA)	synonymous SNP	1	63BJ352	WD	L-idonate and D-gluconate transporter
infB ←	3,364,571	C→T	L790L (CTG→CTA)	synonymous SNP	1	72AJ84	CD	fused protein chain initiation factor 2, IF2: membrane protein ; conserved hypothetical protein
kdgR ←	1,928,403	C→T	S225N (AGC→AAC)	non synonymous SNP	1	74AJ351	CD	DNA-binding transcriptional repressor KdgR
	1,928,908	T→C	T57A (ACC→GCC)	non synonymous SNP	1	57AJ352	WD	
kdgT	4,130,261	IS1	coding (563/984 nt)	IS transposition	1	90BJ84	CD	2-keto-3-deoxy-D-gluconate transporter

	4,129,737	(G)7→8	coding (39/984 nt)	duplication	23	72AJ84; 72AJ203; 72AJ351; 72BJ203; 72BJ351; 74AJ84; 74AJ203; 78AJ203; 79AJ84; 79AJ203; 79AJ351; 79BJ84; 79BJ203; 79BJ351; 84AJ84; 84AJ351; 84BJ84; 87AJ203; 87BJ203; 90AJ84; 90AJ203; 90BJ203; 90BJ351	CD	
	4,129,737	(G)7→6	coding (39/984 nt)	deletion	5	74AJ351; 74BJ351; 84BJ203; 87AJ351; 87BJ84	CD	
	4,129,737	(G)7→9	coding (39/984 nt)	duplication	1	84BJ351	CD	
	4,130,135	C→T	S146L (TCG→TTG)	non synonymous SNP	2	72BJ84; 87AJ84	CD	
	4,130,257	(CTGGGGAACC)1→2	coding (559/984 nt)	duplication	1	90AJ351	CD	
	4,130,364	+T	coding (666/984 nt)	insertion	4	78AJ351; 78BJ84; 78BJ203; 78BJ351	CD	
	4,130,460	+T	coding (762/984 nt)	insertion	1	74BJ84	CD	
kdpA ←	753,88	T→C	A311A (GCA→GCG)	synonymous SNP	1	84BJ351	CD	potassium translocating ATPase, subunit A
kduI ←	3,001,852	G→A	H198H (CAC→CAT)	synonymous SNP	1	79AJ351	CD	5-keto 4-deoxyuronate isomerase
kefB ←	3,524,295	G→A	A219V (GCC→GTC)	non synonymous SNP	1	84AJ351	CD	potassium:proton antiporter
ldrB ← / ← *ldrB*	1,318,546	T→G	intergenic (-272/+129)	intergenic SNP	3	79AJ84; 79AJ203; 79AJ351	CD	fragment of small toxic polypeptide (partial)/fragment of small toxic polypeptide (partial)
leuA ←	84,796	G→A	A163A (GCC→GCT)	synonymous SNP	1	57AJ83	WD	2-isopropylmalate synthase
lhr →	1,752,857	T→C	V594A (GTG→GCG)	non synonymous SNP	1	57BJ352	WD	putative ATP-dependent helicase
	1,755,367	G→A	A1431T (GCA→ACA)	non synonymous SNP	1	72AJ84	CD	
livF ←	3,639,375	A→G	F103L (TTT→CTT)	non synonymous SNP	1	79AJ351	CD	leucine/isoleucine/valine transporter subunit ; ATP-binding component of ABC superfamily
lpxC → / → *secM*	109,308	(T)9→10	intergenic (+79/-77)	duplication	1	84AJ84	CD	UDP-3-O-acyl N-acetylglucosamine deacetylase/regulator of secA translation

lrhA ←	2,447,878	G→A	T112M (ACG→ATG)	non synonymous SNP	8	49AJ352; 49BJ352; 63AJ352; 63BJ352; 66AJ203; 66AJ322; 66BJ203; 66BJ322	WD	DNA-binding transcriptional repressor of flagellar, motility and chemotaxis genes
mak →	474,795	(G)6→7	coding (755/909 nt)	duplication	2	79BJ203; 79BJ351	CD	manno(fructo)kinase
malE	4,277,565	IS1	coding (599/1191 nt)	IS transposition	1	84AJ84	CD	maltose transporter subunit ; periplasmic-binding component of ABC superfamily
malE/malK	4,278,259	IS1	intergenic (-96/-269)	IS transposition	4	78AJ351, 78BJ84, 78BJ203, 78BJ351	CD	maltose transporter subunit ; periplasmic-binding component of ABC superfamily/fused maltose transport subunit, ATP-binding component of ABC superfamily ; regulatory protein
malF	4,275,291	IS1	coding (1529/1545 nt)	IS transposition	1	69AJ84	WD	maltose transporter subunit ; membrane component of ABC superfamily
	4,276,705	IS1	coding (115/1545 nt)	IS transposition	2	90AJ84, 90AJ203	CD	
	4,275,539	Δ24 bp	coding (1258-1281/1545 nt)	deletion	5	79AJ84; 79AJ203; 79AJ351; 79BJ203; 79BJ351	CD	
	4,276,609	C→T	G71R (GGA→AGA)	non synonymous SNP	1	84BJ351	CD	
malF/malE	4,276,909	IS1	intergenic (-90/+64)	IS transposition	13	60AJ83, 63BJ84, 63BJ203, 72AJ84, 72AJ203, 72AJ351, 72BJ203, 72BJ351, 78AJ203, 84AJ203, 87BJ351, 90AJ351, 90BJ84	WD (3) / CD (10)	maltose transporter subunit ; membrane component of ABC superfamily/maltose transporter subunit ; periplasmic-binding component of ABC superfamily
malP ←	3,595,011	A→G	Y580H (TAC→CAC)	non synonymous SNP	1	84BJ351	CD	maltodextrin phosphorylase
malT	3,599,623	IS1	coding (2264/2706 nt)	IS transposition	1	74BJ84	CD	
	3,597,604	T→A	V82E (GTG→GAG)	non synonymous SNP	4	84BJ203; 87AJ84; 87AJ351; 87BJ84	CD	
	3,598,353	T→C	C332R (TGC→CGC)	non synonymous SNP	1	60BJ202	WD	
	3,598,381	(AGTGGGAACTGGC GGCGGAGCTGCC)1 →2	coding (1022/2706 nt)	duplication	1	74AJ351;74BJ351	CD	

	3,599,488	T→C	V710A (GTT→GCT)	non synonymous SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	
	3,599,601	Δ1 bp	coding (2242/2706 nt)	deletion	1	79BJ84	CD	
	3,599,967	C→T	H870Y (CAT→TAT)	non synonymous SNP	4	69AJ231; 69AJ352; 69BJ231; 69BJ352	WD	
manA	1,710,721	IS1	coding (12/1176 nt)	IS transposition	2	84BJ203, 87AJ351	CD	mannose-6-phosphate isomerase
	1,710,857	Δ210 bp	coding (148-357/1176 nt)	deletion	1	74AJ84	CD	
manZ →	1,922,909	(A)6→7	coding (36/852 nt)	duplication	1	79AJ351	CD	mannose-specific enzyme IID component of PTS
mdaB →	3,217,381	C→T	I7I (ATC→ATT)	synonymous SNP	1	74AJ203	CD	NADPH quinone reductase
mdh ←	3,427,660	G→A	A66V (GCG→GTG)	non synonymous SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	malate dehydrogenase, NAD(P)-binding
mdh ← / → *argR*	3,427,892	G→A	intergenic (-36/-399)	intergenic SNP	1	74AJ351	CD	malate dehydrogenase, NAD(P)-binding/DNA-binding transcriptional dual regulator, L-arginine-binding
mdtB →	2,199,627	C→T	P42S (CCC→TCC)	non synonymous SNP	1	57AJ202	WD	multidrug efflux system, subunit B
mdtF →	3,704,219	G→A	G832E (GGG→GAG)	non synonymous SNP	2	72AJ351, 72BJ351	CD	multidrug transporter, RpoS-dependent
mdtG ←	1,178,541	C→T	R209Q (CGG→CAG)	non synonymous SNP	3	57BJ83; 57BJ202; 57BJ352	WD	putative metabolite efflux transporter
melB →	4,367,644	G→A	L85L (TTG→TTA)	synonymous SNP	1	74BJ351	CD	melibiose:sodium symporter
	4,367,842	(T)7→6	coding (453/1422 nt)	deletion	1	72AJ203	CD	
menC	2,416,611	putative transposase	coding (139/963 nt)	IS transposition	1	60BJ202	WD	o-succinylbenzoyl-CoA synthase
metE →	4,045,207	C→T	D599D (GAC→GAT)	synonymous SNP	1	72AJ203	CD	5-methyltetrahydropteroyltriglutamate-homocystein e S-methyltransferase
metQ ←	219,968	A→G	S118P (TCC→CCC)	non synonymous SNP	2	72AJ351; 72BJ351	CD	DL-methionine transporter subunit ; periplasmic-binding component of ABC superfamily
mglA ←	2,280,280	G→A	Q413* (CAA→TAA)	stop SNP	1	74AJ351	CD	fused methyl-galactoside transporter subunits of ABC superfamily: ATP-binding components
	2,281,377	Δ1 bp	coding (140/1521 nt)	deletion	4	72AJ351; 72BJ351; 74BJ351; 79BJ351	CD	
mngB →	796,083	G→A	pseudogene (899/2280 nt)	SNP	2	57BJ202; 57BJ352	WD	fragment of alpha-mannosidase (part 2)
mprA	2,828,410	IS1	coding (91/531 nt)	IS transposition	1	78AJ84	CD	DNA-binding transcriptional repressor of microcin B17 synthesis and multidrug efflux
	2,828,484	IS1	coding (165/531 nt)	IS transposition	2	72AJ351, 72BJ351	CD	
	2,828,518	IS1	coding (199/531 nt)	IS transposition	1	87AJ351	CD	

	2,828,331	Δ5 bp	coding (12-16/531 nt)	deletion	1	90AJ351	CD	
	2,828,681	G→A	G121D (GGT→GAT)	non synonymous SNP	1	74AJ203	CD	
mtr ←	3,355,739	(C)7→9	coding (1005/1245 nt)	duplication	1	79AJ351	CD	tryptophan transporter of high affinity
mutH	2,988,577	IS1	coding (596/690 nt)	IS transposition	1	79BJ84	CD	methyl-directed mismatch repair protein
mutL →	4,422,361	Δ87 bp	coding (214-300/1848 nt)	deletion	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	methyl-directed mismatch repair protein
	4,422,362	Δ6 bp	coding (215-220/1848 nt)	deletion	7	72AJ84; 72AJ203; 72AJ351; 72BJ203; 72BJ351; 74AJ203; 84BJ351	CD	
	4,422,367	(TGGCGC)3→4	coding (220/1848 nt)	duplication	5	79AJ84; 79AJ203; 79AJ351; 79BJ203; 79BJ351	CD	
mutS	2,875,090	IS1	coding (580/2562 nt)	IS transposition	1	84AJ84	CD	methyl-directed mismatch repair protein
	2,875,351	C→T	Q281* (CAG→TAG)	stop SNP	2	74AJ351; 74BJ351	CD	
nadB →	2,750,163	C→T	L470F (CTC→TTC)	non synonymous SNP	1	72AJ84	CD	quinolinate synthase, L-aspartate oxidase (B protein) subunit
napF ← / → *ESCCO2305*	2,344,491	A→G	intergenic (-15/-85)	intergenic SNP	1	87BJ351	CD	ferredoxin-type protein, putative role in electron transfer to periplasmic nitrate reductase (NapA)/conserved hypothetical protein
ndh →	1,229,982	C→T	A70A (GCC→GCT)	synonymous SNP	1	79AJ351	CD	respiratory NADH dehydrogenase 2/cupric reductase
nikA →	3,661,202	(GC)3→2	coding (1266-1267/1575 nt)	deletion	1	57AJ352	WD	nickel transporter subunit ; periplasmic-binding component of ABC superfamily
nlpE → / → *EcHS_A0196*	214,180	G→A	intergenic (+44/-206)	intergenic SNP	3	84BJ203; 87AJ351; 87BJ84	CD	lipoprotein involved with copper homeostasis and adhesion/transposase, IS110 family
nrdD ←	4,495,823	T→C	S685G (AGC→GGC)	non synonymous SNP	1	74AJ203	CD	anaerobic ribonucleoside-triphosphate reductase
nrdF → / → *proV*	2,822,049	(T)6→7	intergenic (+145/-209)	duplication	1	74AJ203	CD	ribonucleoside-diphosphate reductase 2, beta subunit, ferritin-like/glycine betaine transporter subunit ; ATP-binding compoent of ABC superfamily
	2,821,960	NA	intergenic (+56/-298)	insertion	1	66AJ322	WD	
nrdR →	496,972	C→T	P60L (CCG→CTG)	non synonymous SNP	2	79AJ84; 79AJ203	CD	transcriptional repressor of nrd genes
nrfB →	4,321,263	C→T	P85S (CCA→TCA)	non synonymous SNP	1	74AJ351	CD	nitrite reductase, formate-dependent, penta-heme cytochrome c
nuoA ← / ← *lrhA*	2,446,988	(C)6→7	intergenic (-345/+286)	duplication	1	57AJ352	WD	NADH:ubiquinone oxidoreductase, membrane subunit A/DNA-binding transcriptional repressor of flagellar, motility and chemotaxis genes
	2,447,198	(A)8→9	intergenic (-555/+76)	duplication	2	72BJ203; 79AJ351	CD	

nupG → / ← *speC*	3,135,933	(T)8→7	intergenic (+45/+5)	deletion	1	74BJ351	CD	nucleoside transporter/ornithine decarboxylase, constitutive
ompA ← / ← *sulA*	1,081,798	G→A	intergenic (-131/+226)	intergenic SNP	1	74BJ351	CD	outer membrane protein A (3a;II*;G;d)/SOS cell division inhibitor
	1,081,866	(A)8→7	intergenic (-199/+158)	deletion	1	79AJ351	CD	
	1,081,962	A→G	intergenic (-295/+62)	intergenic SNP	1	74AJ203	CD	
ompC ←	2,352,803	Δ3 bp	coding (929-931/1107 nt)	deletion	1	90BJ351	CD	outer membrane porin protein C
	2,352,812	C→T	G308S (GGT→AGT)	non synonymous SNP	1	72BJ84	CD	
ompF	1,048,350	IS1	coding (250/1089 nt)	IS transposition	1	78AJ351	CD	outer membrane porin 1a (Ia;b;F)
	1,048,395	Δ1 bp	coding (205/1089 nt)	deletion	1	78BJ203	CD	
ompF ← / ← *asnS*	1,048,625	(T)8→7	intergenic (-26/+576)	deletion	1	84BJ351	CD	outer membrane porin 1a (Ia;b;F)/asparaginyl tRNA synthetase
ompR ←	3,580,337	C→T	V203M (GTG→ATG)	non synonymous SNP	2	90AJ351; 90BJ84	CD	DNA-binding response regulator in two-component regulatory system with EnvZ
	3,580,855	T→C	Q30R (CAG→CGG)	non synonymous SNP	1	79AJ351	CD	
oppD →	1,352,403	C→T	A198V (GCG→GTG)	non synonymous SNP	1	72AJ203	CD	oligopeptide transporter subunit ; ATP-binding component of ABC superfamily
paaG →	1,481,818	(G)6→7	coding (362/789 nt)	duplication	1	57AJ352	WD	acyl-CoA hydratase
paaH →	1,482,788	T→C	V181A (GTG→GCG)	non synonymous SNP	1	84BJ351	CD	3-hydroxybutyryl-CoA dehydrogenase
pdxA ←	56,323	C→T	G119D (GGC→GAC)	non synonymous SNP	1	74BJ351	CD	4-hydroxy-L-threonine phosphate dehydrogenase, NAD-dependent
pflC →	4,353,306	T→G	G146G (GGT→GGG)	synonymous SNP	1	60AJ352	WD	pyruvate formate lyase II activase
pheT ←	1,814,313	T→C	T145A (ACC→GCC)	non synonymous SNP	1	79BJ84	CD	phenylalanine tRNA synthetase, beta subunit
pinR →	1,656,992	(G)6→5	coding (305/591 nt)	deletion	1	84BJ351	CD	putative site-specific recombinase; DNA invertase; Rac prophage
pntA ←	1,699,278	G→A	T254I (ACC→ATC)	non synonymous SNP	1	87BJ351	CD	pyridine nucleotide transhydrogenase, alpha subunit
polA →	4,077,721	C→T	G192G (GGC→GGT)	synonymous SNP	1	74AJ203	CD	fused DNA polymerase I 5'->3' exonuclease ; 3'->5' polymerase ; 3'->5' exonuclease
polA → / ← *yihA*	4,080,161	Δ1 bp	intergenic (+229/+153)	deletion	1	84BJ351	CD	fused DNA polymerase I 5'->3' exonuclease ; 3'->5' polymerase ; 3'->5' exonuclease/GTP-binding protein
ppiD →	528,944	C→T	T322I (ACT→ATT)	non synonymous SNP	2	72AJ351; 72BJ351	CD	peptidyl-prolyl cis-trans isomerase (rotamase D)
pps ←	1,803,127	(T)7→6	coding (787/2379 nt)	deletion	1	84AJ84	CD	phosphoenolpyruvate synthase
prc ←	1,931,945	T→C	N617D (AAT→GAT)	non synonymous SNP	1	74AJ203	CD	carboxy-terminal protease for penicillin-binding protein 3
proP →	4,354,697	(G)5→6	coding (175/1503 nt)	duplication	1	84BJ351	CD	proline/glycine betaine transporter
proQ ←	1,934,077	Δ5 bp	coding (431-435/699 nt)	deletion	1	63BJ203	WD	ProP transporter amplifier
	1,934,507	T→A	E2V (GAA→GTA)	non synonymous SNP	1	84BJ203	CD	

proQ/yebR	1,934,595	IS1	intergenic (-84/+13)	IS transposition	1	78AJ351	CD	ProP transporter amplifier/conserved hypothetical protein
proW →	2,823,485	G→A	T11T (ACG→ACA)	synonymous SNP	1	57AJ352	WD	glycine betaine transporter subunit ; membrane component of ABC superfamily
prpD →	417,143	C→T	G111G (GGC→GGT)	synonymous SNP	1	84AJ84	CD	2-methylcitrate dehydratase
ptsP	2,986,024	IS1	coding (731/2247 nt)	IS transposition	1	60AJ83	WD	fused PTS enzyme: PEP-protein phosphotransferase (enzyme I) ; GAF domain containing protein
purK	611,589	IS1	coding (527/1068 nt)	IS transposition	2	49AJ352, 49BJ352	WD	N5-carboxyaminoimidazole ribonucleotide synthase
	611,824	G→A	Q98* (CAG→TAG)	stop SNP	1	66BJ322	WD	
	611,863	G→A	Q85* (CAG→TAG)	stop SNP	2	66AJ203; 66BJ203	WD	
purL ←	2,733,341	G→A	G179G (GGC→GGT)	synonymous SNP	1	79AJ84	CD	phosphoribosylformyl-glycineamide synthetase
purL ← / → *yfhD*	2,733,920	(G)8→9	intergenic (-43/-216)	duplication	4	57BJ202; 72AJ351; 79BJ351; 84AJ84	WD (1)/ CD (3)	phosphoribosylformyl-glycineamide synthetase/putative transglycosylase
	2,733,920	(G)8→7	intergenic (-43/-216)	deletion	1	79AJ351	CD	
putP →	1,142,457	C→T	A285V (GCG→GTG)	non synonymous SNP	1	57BJ202	WD	proline:sodium symporter
puuR →	1,409,393	A→G	K10K (AAA→AAG)	synonymous SNP	1	79AJ203	CD	DNA-binding transcriptional repressor
	1,409,819	G→A	A152A (GCG→GCA)	synonymous SNP	2	72AJ351; 72BJ351	CD	
pykA →	1,957,340	A→G	I239V (ATC→GTC)	non synonymous SNP	1	79AJ203	CD	pyruvate kinase II
pyrC ←	1,185,311	T→C	I326V (ATC→GTC)	non synonymous SNP	1	72BJ203	CD	dihydro-orotase
qor ←	4,295,860	C→T	G39D (GGC→GAC)	non synonymous SNP	1	79BJ351	CD	quinone oxidoreductase, NADPH-dependent
ravA ← / → *trkD*	3,958,362	(G)7→8	intergenic (-66/-133)	duplication	2	72AJ351; 72BJ351	CD	fused regulatory ATPase variant A ; conserved hypothetical protein/potassium transporter
	3,958,362	(G)7→6	intergenic (-66/-133)	deletion	1	79AJ351	CD	
rbfA ←	3,363,914	G→A	A64V (GCG→GTG)	non synonymous SNP	1	79BJ351	CD	fused D-ribose transporter subunits of ABC superfamily: ATP-binding components
rbsA	3,961,394	NA	coding (438/1506 nt)	duplication	1	78AJ351	CD	fused D-ribose transporter subunits of ABC superfamily: ATP-binding components
	3,961,741	IS1	coding (785/1506 nt)	IS transposition	4	84AJ351, 84BJ84, 87AJ203, 87BJ203	CD	
	3,961,185	C→T	Q77* (CAG→TAG)	stop SNP	10	49AJ201; 49AJ352; 49BJ201; 49BJ352; 57BJ83; 69AJ231; 69AJ352; 69BJ231; 69BJ352; 78BJ351	WD (9) / CD (1)	

	3,961,317	G→T	E121* (GAA→TAA)	stop SNP	1	63AJ203	WD	
	3,961,382	Δ2 bp	coding (426-427/1506 nt)	deletion	1	84AJ84	CD	
	3,961,608	C→T	Q218* (CAA→TAA)	stop SNP	6	57AJ83; 57BJ202; 57BJ352; 72AJ203; 79AJ84; 79AJ203	WD (3) / CD (3)	
rbsB →	3,963,471	(A)6→7	coding (9/885 nt)	duplication	1	79AJ351	CD	D-ribose transporter subunit ; periplasmic-binding compoent of ABC superfamily
rbsC →	3,962,588	(T)5→6	coding (122/966 nt)	duplication	1	84BJ351	CD	D-ribose transporter subunit ; membrane component of ABC superfamily
	3,962,733	Δ4 bp	coding (267-270/966 nt)	deletion	1	66BJ84	WD	
	3,962,749	Δ1 bp	coding (283/966 nt)	deletion	1	72AJ84	CD	
	3,963,021	+CGCG	coding (555/966 nt)	insertion	2	84BJ203; 87AJ351	CD	
rbsD →	3,960,532	G→A	M1M (ATG→ATA) †	synonymous SNP	1	57AJ352	WD	putative cytoplasmic sugar-binding protein
	3,960,538	(A)6→7	coding (9/420 nt)	duplication	7	57AJ202; 72AJ351; 72BJ203; 72BJ351; 74AJ203; 74AJ351; 74BJ351	CD(6)/WD (1)	
	3,960,686	C→T	Q53* (CAG→TAG)	stop SNP	6	63AJ352; 63BJ352; 66AJ203; 66AJ322; 66BJ203; 66BJ322	WD	
	3,960,716	C→T	Q63* (CAG→TAG)	stop SNP	2	63BJ84; 63BJ203	WD	
	3,960,752	C→T	Q75* (CAA→TAA)	stop SNP	1	60BJ202	WD	
	3,960,819	Δ1 bp	coding (290/420 nt)	deletion	1	78AJ203	CD	
rcsA →	2,043,907	A→G	H67R (CAC→CGC)	non synonymous SNP	1	79AJ203	CD	DNA-binding transcriptional activator, co-regulator with RcsB
recB ←	2,971,172	G→A	P1051S (CCG→TCG)	non synonymous SNP	1	74AJ203	CD	exonuclease V (RecBCD complex), beta subunit
recC ←	2,979,399	A→G	V450A (GTA→GCA)	non synonymous SNP	2	79AJ84; 79AJ203	CD	exonuclease V (RecBCD complex), gamma chain
rfe →	3,996,484	(G)6→7	coding (156/1104 nt)	duplication	3	57BJ83; 57BJ202; 57BJ352	WD	UDP-GlcNAc:undecaprenylphosphate GlcNAc-1-phosphate transferase
	3,997,291	G→A	W321* (TGG→TGA)	stop SNP	1	69BJ84	WD	
rffE →	3,999,555	C→T	T337M (ACG→ATG)	non synonymous SNP	1	57AJ83	WD	UDP-N-acetyl glucosamine-2-epimerase
rhaT ←	4,128,286	A→C	C83W (TGT→TGG)	non synonymous SNP	1	74BJ84	CD	L-rhamnose:proton symporter

rho →	3,995,585	C→T	R252R (CGC→CGT)	synonymous SNP	1	79AJ203	CD	transcription termination factor
rihA ←	711,793	A→T	T256T (ACT→ACA)	synonymous SNP	1	74BJ203	CD	ribonucleoside hydrolase 1
rlmB → / → *yjfI*	4,434,789	(T)9→8	intergenic (+47/-80)	deletion	2	57AJ202; 74AJ351	WD	23S rRNA (Gm2251)-methyltransferase/conserved hypothetical protein
rlpB ←	708,163	T→A	I37F (ATC→TTC)	non synonymous SNP	1	79BJ84	CD	minor lipoprotein
rluC → / ← *yceF*	1,209,625	T→G	intergenic (+46/+66)	intergenic SNP	1	90AJ351	CD	23S rRNA pseudouridylate synthase/conserved hypothetical protein
rluD ←	2,743,262	Δ1 bp	coding (570/981 nt)	deletion	1	60AJ352	WD	23S rRNA pseudouridine synthase
	2,775,209	A→T	D139E (GAT→GAA)	non synonymous SNP	1	60BJ202	WD	
rnc ← / ← *lepB*	2,742,667	G→A	intergenic (-268/+4)	intergenic SNP	3	57BJ83; 57BJ202; 57BJ352	WD	RNase III/leader peptidase (signal peptidase I)
rncA →	2,228,324	T→C	V20A (GTC→GCC)	non synonymous SNP	1	72AJ203	CD	nickel and cobalt resistance
rne ←	1,206,706	C→T	A448T (GCT→ACT)	non synonymous SNP	2	57AJ202; 57AJ352	WD	fused ribonucleaseE: endoribonuclease ; RNA-binding protein ;RNA degradosome binding protein
rpoB →	4,212,163	Δ9 bp	coding (1337-1345/4029 nt)	deletion	1	69BJ352	WD	RNA polymerase, beta subunit
	4,212,226	G→A	G467D (GGC→GAC)	non synonymous SNP	1	69AJ352	WD	
rpoC →	4,215,698	A→G	D256G (GAT→GGT)	non synonymous SNP	2	57BJ202; 57BJ352	WD	RNA polymerase, beta prime subunit
rpoS ←	2,884,128	Δ1 bp	coding (843/993 nt)	deletion	1	79BJ351	CD	RNA polymerase, sigma S (sigma 38) factor
rpoZ →	3,851,829	(CCACT)1→2	coding (141/276 nt)	duplication	1	66BJ84	WD	RNA polymerase, omega subunit
rpsA →	1,023,781	G→A	E482K (GAA→AAA)	non synonymous SNP	1	84BJ351	CD	30S ribosomal subunit protein S1
rrf →	4,201,332	C→T	noncoding (112/116 nt)	SNP	1	72AJ84	CD	5S ribosomal RNA
rrl	2,766,683	NA	noncoding (2115/2904 nt)	insertion	8	49BJ201, 49BJ352, 57AJ83, 57BJ83, 60BJ202, 69AJ84, 78BJ203, 90BJ351	WD (6) / CD (2)	23S ribosomal RNA
	3,974,229	NA	noncoding (2115/2904 nt)	insertion	1	69AJ352	WD	
rrmA ←	1,925,480	A→G	L187P (CTG→CCG)	non synonymous SNP	1	74AJ351	CD	23S rRNA m1G745 methyltransferase
	1,925,996	(T)7→8	coding (44/810 nt)	duplication	1	79AJ351	CD	
rrs ←	3,471,568	G→A	noncoding (1538/1542 nt)	SNP	1	49BJ352	WD	16S ribosomal RNA
rrs ← / → *yrdA*	3,473,186	G→T	intergenic (-81/-177)	intergenic SNP	2	79BJ203; 79BJ351	CD	16S ribosomal RNA/conserved hypothetical protein
rspB ←	1,675,837	C→T	V72M (GTG→ATG)	non synonymous SNP	1	74BJ84	CD	putative oxidoreductase, Zn-dependent and NAD(P)-binding

rsxC →	1,729,647	A→G	Q259R (CAG→CGG)	non synonymous SNP	1	79BJ351	CD	putative 4Fe-4S ferredoxin-type protein fused with unknown protein
rsxD →	1,731,361	T→C	I89I (ATT→ATC)	synonymous SNP	1	74BJ351	CD	putative inner membrane oxidoreductase
rutD ←	1,133,695	G→A	R123R (CGC→CGT)	synonymous SNP	4	69AJ231; 69AJ352; 69BJ231; 69BJ352	WD	enzyme of the alternative pyrimidine degradation pathway
sbmA →	460,716	(T)6→7	coding (181/1221 nt)	duplication	1	57BJ352	WD	transporter involved in cell envelope modification
	461,109	C→T	Q192* (CAG→TAG)	stop SNP	1	63BJ203	WD	
	461,336	Δ1 bp	coding (801/1221 nt)	deletion	1	57AJ352	WD	
	461,365	C→T	A277V (GCC→GTC)	non synonymous SNP	2	63AJ352; 63BJ352	WD	
	461,451	C→T	R306W (CGG→TGG)	non synonymous SNP	1	66AJ322	WD	
sdaA →	1,917,057	C→T	P383S (CCG→TCG)	non synonymous SNP	1	74BJ351	CD	L-serine deaminase I
sdaC →	2,946,655	G→A	R73H (CGC→CAC)	non synonymous SNP	1	74AJ203	CD	putative serine transporter
sdiA ←	2,016,124	T→C	A112A (GCA→GCG)	synonymous SNP	2	72AJ351; 72BJ351	CD	DNA-binding transcriptional activator
selB ←	3,788,415	Δ16 bp	coding (735-750/1845 nt)	deletion	1	90AJ203	CD	selenocysteinyl-tRNA-specific translation factor
	3,789,136	T→C	D10G (GAC→GGC)	non synonymous SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	
setC	3,866,618	IS1	coding (83/1185 nt)	IS transposition	1	87AJ84	CD	putative sugar efflux system
sgrR ←	78,149	C→G	W302C (TGG→TGC)	non synonymous SNP	1	72BJ84	CD	DNA-binding transcriptional regulator
shiA →	2,093,469	C→T	A408A (GCC→GCT)	synonymous SNP	1	74AJ203	CD	shikimate transporter
sodB/ydhP	1,757,959	IS110	intergenic (+49/+186)	IS transposition	2	87BJ351, 90AJ351	CD	superoxide dismutase, Fe/putative efflux protein; MFS family
speA ←	3,112,371	G→A	P390S (CCT→TCT)	non synonymous SNP	1	63BJ203	WD	biosynthetic arginine decarboxylase, PLP-binding
	3,112,994	A→G	L182P (CTG→CCG)	non synonymous SNP	1	74BJ351	CD	
speB ← / ← *speA*	3,111,459	(A)7→6	intergenic (-35/+103)	deletion	2	72AJ351; 72BJ351	CD	agmatinase/biosynthetic arginine decarboxylase, PLP-binding
ssb →	4,306,058	C→T	R57C (CGC→TGC)	non synonymous SNP	1	84BJ351	CD	Single-stranded DNA-binding protein
ssnA →	3,047,521	C→T	L283L (CTG→TTG)	synonymous SNP	1	84BJ351	CD	putative chlorohydrolase/aminohydrolase
sstT →	3,284,537	(G)7→6	coding (1041/1245 nt)	deletion	1	74BJ351	CD	sodium:serine/threonine symporter
stpA ←	2,815,814	(TTC)4→3	coding (124-126/405 nt)	deletion	2	79BJ203; 79BJ351	CD	DNA binding protein, nucleoid-associated
sucA →	784,717	G→A	W9* (TGG→TGA)	stop SNP	1	79AJ351	CD	2-oxoglutarate decarboxylase, thiamin-requiring
sufI ←	3,207,161	C→T	G114S (GGC→AGC)	non synonymous SNP	2	72AJ351; 72BJ351	CD	repressor protein for FtsI
tatA →	4,052,380	C→T	G26G (GGC→GGT)	synonymous SNP	2	72AJ351; 72BJ351	CD	TatABCE protein translocation system subunit

tatB →	4,052,745	A→G	Q57R (CAG→CGG)	non synonymous SNP	1	74AJ351	CD	TatABCE protein translocation system subunit
tauA →	450,075	A→G	Q300R (CAG→CGG)	non synonymous SNP	1	79BJ84	CD	taurine transporter subunit ; periplasmic-binding component of ABC superfamily
tauC →	451,310	C→T	P132L (CCG→CTG)	non synonymous SNP	1	74AJ351	CD	taurine transporter subunit ; membrane component of ABC superfamily
tdcA ←	3,315,215	A→G	V230A (GTC→GCC)	non synonymous SNP	1	66AJ84	WD	DNA-binding transcriptional activator
	3,315,307	C→G	L199F (TTG→TTC)	non synonymous SNP	1	63AJ203	WD	
	3,315,311	A→G	V198A (GTG→GCG)	non synonymous SNP	2	63AJ352; 63BJ352	WD	
	3,315,378	C→A	V176L (GTA→TTA)	non synonymous SNP	2	49AJ201; 49BJ201	WD	
	3,315,408	C→T	E166K (GAG→AAG)	non synonymous SNP	4	69AJ231; 69AJ352; 69BJ231; 69BJ352	WD	
	3,315,629	G→A	S92F (TCT→TTT)	non synonymous SNP	1	66BJ84	WD	
thiI →	506,433	G→A	E365K (GAG→AAG)	non synonymous SNP	1	87BJ203	CD	sulfurtransferase required for thiamine and 4-thiouridine biosynthesis
thrL → / → *thrA*	309	(T)8→7	intergenic (+54/-27)	deletion	1	57AJ352	WD	thr operon leader peptide/fused aspartokinase I ; homoserine dehydrogenase I
tnaC →	3,915,673	G→A	R23H (CGC→CAC)	non synonymous SNP	2	72AJ203; 72BJ203	CD	tryptophanase leader peptide
tnaC → / → *tnaA*	3,915,683	T→G	intergenic (+3/-218)	intergenic SNP	2	49AJ352; 49BJ352	WD	tryptophanase leader peptide/tryptophanase/L-cysteine desulfhydrase, PLP-dependent
torY ←	1,976,716	G→A	G196G (GGC→GGT)	synonymous SNP	2	79BJ203; 79BJ351	CD	TMAO reductase III (TorYZ), cytochrome c-type subunit
torZ ←	1,974,801	A→G	W466R (TGG→CGG)	non synonymous SNP	1	84BJ351	CD	trimethylamine N-oxide reductase system III, catalytic subunit
trkD → / → *rbsD*	3,960,443	C→T	intergenic (+80/-87)	intergenic SNP	1	79BJ84	CD	potassium transporter/putative cytoplasmic sugar-binding protein
trkH →	4,063,686	G→A	P94P (CCG→CCA)	synonymous SNP	1	57AJ83	WD	potassium transporter
trmA ← / → *btuB*	4,193,055	G→T	intergenic (-119/-250)	intergenic SNP	1	84AJ351	CD	tRNA (uracil-5-)-methyltransferase/vitamin B12/cobalamin outer membrane transporter
	4,193,298	C→T	intergenic (-362/-7)	intergenic SNP	1	74AJ203	CD	
trmE →	3,915,089	G→A	G364D (GGC→GAC)	non synonymous SNP	1	72BJ203	CD	GTPase
trpC ←	1,365,701	C→T	G382D (GGC→GAC)	non synonymous SNP	2	72AJ351; 72BJ351	CD	fused indole-3-glycerolphosphate synthetase ; N-(5-phosphoribosyl)anthranilate isomerase
trpE ← / ← *trpL*	1,370,032	(A)8→9	intergenic (-29/+63)	duplication	1	57AJ83	WD	component I of anthranilate synthase/trp operon leader peptide
trxA →	3,994,442	C→T	T90I (ACC→ATC)	non synonymous SNP	1	57AJ352	WD	thioredoxin 1
tynA ←	1,473,905	Δ1 bp	coding (638/2274 nt)	deletion	1	84BJ351	CD	tyramine oxidase, copper-requiring

tyrS ←	1,738,561	G→A	I205I (ATC→ATT)	synonymous SNP	1	87BJ351	CD	tyrosyl-tRNA synthetase
ubiB →	4,051,301	A→C	M240L (ATG→CTG)	non synonymous SNP	1	84BJ351	CD	2-octaprenylphenol hydroxylase
ubiG →	2,380,618	G→A	A24T (GCC→ACC)	non synonymous SNP	3	57BJ83; 57BJ202; 57BJ352	WD	bifunctional 3-demethylubiquinone-9 3-methyltransferase and 2-octaprenyl-6-hydroxy phenol methylase
uhpA ← / ← *ilvN*	3,880,335	(A)9→8	intergenic (-27/+48)	deletion	1	79BJ203	CD	DNA-binding response regulator in two-component regulatory system wtih UhpB/acetolactate synthase I, small subunit
ulaG ← / → *ulaA*	4,444,595	T→C	intergenic (-234/-121)	intergenic SNP	2	72AJ351; 72BJ351	CD	putative L-ascorbate 6-phosphate lactonase/L-ascorbate-specific enzyme IIC component of PTS
upp ← / → *purM*	2,642,844	(A)8→7	intergenic (-35/-290)	deletion	1	84BJ351	CD	uracil phosphoribosyltransferase/phosphoribosylaminoimidazole synthetase
uraA ← / ← *upp*	2,642,145	(A)8→7	intergenic (-48/+38)	deletion	1	79BJ84	CD	uracil transporter/uracil phosphoribosyltransferase
uspA → / → *yhiP*	3,683,261	A→G	intergenic (+206/-112)	intergenic SNP	1	57BJ352	WD	universal stress global response regulator/transporter
	3,683,352	G→A	intergenic (+297/-21)	intergenic SNP	2	57BJ202; 57BJ352	WD	
uxaB ←	1,627,381	T→C	H41R (CAC→CGC)	non synonymous SNP	1	72AJ351	CD	altronate oxidoreductase, NAD-dependent
uxuB →	4,558,756	A→G	D283G (GAC→GGC)	non synonymous SNP	1	84BJ351	CD	D-mannonate oxidoreductase, NAD-binding
waaL ←	3,828,855	T→G	Y325S (TAT→TCT)	non synonymous SNP	1	90BJ351	CD	lipid A-core surface O-antigen ligase
waaV	3,830,233	IS1	coding (625/984 nt)	IS transposition	1	60BJ202	WD	putative beta1,3-glucosyltransferase WaaV
	3,830,487	C→T	G124D (GGC→GAC)	non synonymous SNP	1	60AJ83	WD	
wzyE →	4,007,825	C→T	P271L (CCG→CTG)	non synonymous SNP	1	79BJ351	CD	putative Wzy protein involved in ECA polysaccharide chain elongation
wzzE →	3,997,562	Δ21 bp	coding (119-139/1047 nt)	deletion	1	87BJ203	CD	Entobacterial Common Antigen (ECA) polysaccharide chain length modulation protein
	3,998,445	(G)6→7	coding (1002/1047 nt)	duplication	2	72AJ351; 84BJ351	CD	
xdhD →	3,051,200	G→A	G791S (GGC→AGC)	non synonymous SNP	1	72BJ351	CD	fused putative xanthine/hypoxanthine oxidase: molybdopterin-binding subunit ; Fe-S binding subunit
xylB ←	3,757,421	(C)8→7	coding (1254/1455 nt)	deletion	1	79AJ84	CD	xylulokinase
yaaX →	5,311	A→G	T27A (ACG→GCG)	non synonymous SNP	1	79BJ84	CD	conserved hypothetical protein; putative exported protein
yaeQ →	212,613	C→T	Q56* (CAA→TAA)	stop SNP	1	63AJ203	WD	conserved hypothetical protein
yafD →	230,959	A→G	N147D (AAT→GAT)	non synonymous SNP	2	57AJ202; 57AJ352	WD	conserved hypothetical protein
yafV ←	269,165	C→T	G195S (GGC→AGC)	non synonymous SNP	2	72AJ351; 72BJ351	CD	putative C-N hydrolase family amidase
yahE →	397,447	G→A	L120L (CTG→CTA)	synonymous SNP	1	72BJ203	CD	conserved hypothetical protein

yahG/yahI	400,973	IS110	intergenic (+67/-76)	IS transposition	1	90AJ351	CD	conserved hypothetical protein/putative carbamate kinase
yahM →	411,589	(G)5→6	pseudogene (93/141 nt)	duplication	1	79AJ351	CD	fragment of conserved hypothetical protein (partial)
yaiI → / → aroL	470,155	T→A	intergenic (+36/-147)	intergenic SNP	1	72BJ203	CD	conserved hypothetical protein/shikimate kinase II
yaiP ←	447,543	(A)7→6	coding (338/1197 nt)	deletion	1	79AJ351	CD	putative membrane-associated glycosyltransferase
yajQ →	508,583	G→C	A37P (GCC→CCC)	non synonymous SNP	1	57AJ352	WD	putative nucleotide binding protein (UPF0234 protein)
ybaO →	534,506	C→T	S20F (TCT→TTT)	non synonymous SNP	1	69AJ352	WD	putative DNA-binding transcriptional regulator with homology to Lrp
	534,541	A→G	T32A (ACC→GCC)	non synonymous SNP	2	57AJ202; 57AJ352	WD	
	534,553	T→C	W36R (TGG→CGG)	non synonymous SNP	1	57BJ352	WD	
	534,559	C→T	R38C (CGC→TGC)	non synonymous SNP	1	60BJ202	WD	
	534,562	C→T	L39L (CTG→TTG)	synonymous SNP	1	69BJ352	WD	
	534,739	C→T	R98C (CGC→TGC)	non synonymous SNP	2	49AJ352; 49BJ352	WD	
ybdA →	659,249	T→C	L290S (TTG→TCG)	non synonymous SNP	1	79AJ203	CD	enterobactin exporter
ybdK ←	642,687	T→C	T260A (ACC→GCC)	non synonymous SNP	1	57BJ352	WD	gamma-glutamyl:cysteine ligase
ybfO →	761,179	A→C	R294R (AGG→CGG)	synonymous SNP	1	79BJ351	CD	conserved hypothetical protein; putative rhs-like
	761,181	G→A	R294R (AGG→AGA)	synonymous SNP	1	79BJ351	CD	
	760,600	C→G	Q101E (CAA→GAA)	non synonymous SNP	1	57AJ83	WD	
ybgQ ←	777,753	C→T	R185H (CGC→CAC)	non synonymous SNP	1	57AJ83	WD	conserved hypothetical protein; putative outer membrane protein
ybhK ←	842,717	G→A	H178H (CAC→CAT)	synonymous SNP	1	57BJ352	WD	putative transferase with NAD(P)-binding Rossmann-fold domain; UPF0052 family
ybiP ←	878,165	C→T	A346T (GCT→ACT)	non synonymous SNP	2	74AJ351; 74BJ351	CD	putative enzyme, UPF0141 family, inner membrane
ybiT →	883,425	Δ2 bp	coding (860-861/1593 nt)	deletion	1	79AJ351	CD	putative transporter fused subunits of ABC superfamily: ATP-binding components
ybjJ ←	915,568	T→C	Q203R (CAG→CGG)	non synonymous SNP	1	72AJ203	CD	putative transporter
	915,704	Δ1 bp	coding (472/1209 nt)	deletion	2	66AJ322; 66BJ322	WD	
ybjM →	950,638	(T)7→6	coding (250/378 nt)	deletion	1	57BJ352	WD	conserved hypothetical protein; putative inner membrane protein
ycaN	1,009,890	SNP	intergenic (-39/+15)	SNP	2	66AJ322, 66BJ322	WD	putative DNA-binding transcriptional regulator
ycbB →	1,042,188	(G)6→5	coding (804/1848 nt)	deletion	2	72AJ351; 72BJ351	CD	putative exported enzyme
ycbB → / → ycbK	1,043,301	(C)5→4	intergenic (+69/-112)	deletion	2	74AJ351; 74BJ351	CD	putative exported enzyme/conserved hypothetical protein
ycbF →	1,065,775	(G)7→8	coding (241/738 nt)	duplication	1	84AJ84	CD	putative periplasmic pilin chaperone

ycbQ	1,059,501	putative transposase	coding (27/513 nt)	IS transposition	1	78AJ84	CD	putative fimbrial-like adhesin protein
yccW ←	1,091,353	(A)8→7	coding (231/1191 nt)	deletion	3	57BJ202; 72BJ351; 84BJ351	WD (1)/ CD(2)	putative AdoMet-dependent methyltransferase, UPF0064 family
yccZ ←	1,106,815	(C)7→10	coding (624/1161 nt)	duplication	2	72AJ351; 72BJ351	CD	putative exopolysaccharide export protein
ycdU → / ← *EcHS_A4664*	1,160,034	(G)5→4	intergenic (+129/+426)	deletion	1	84BJ351	CD	conserved hypothetical protein; putative inner membrane protein/tRNA-Ser
	1,160,113	(G)7→8	intergenic (+208/+347)	duplication	1	72AJ84	CD	
	1,160,294	(G)7→6	intergenic (+389/+166)	deletion	1	74AJ203	CD	
yceB ←	1,186,904	C→A	V17F (GTT→TTT)	non synonymous SNP	1	79AJ203	CD	putative lipoprotein
ycgE ←	1,260,836	C→T	E148K (GAA→AAA)	non synonymous SNP	1	79AJ351	CD	putative DNA-binding transcriptional regulator
ycgV ←	1,302,598	(C)6→5	coding (1794/2868 nt)	deletion	1	79BJ351	CD	putative adhesin; putative autotransporter
ycgZ →	1,263,040	G→A	G12E (GGA→GAA)	non synonymous SNP	1	78BJ351	CD	conserved hypothetical protein
ychE →	1,347,243	G→A	G130G (GGG→GGA)	synonymous SNP	1	74AJ351	CD	conserved hypothetical protein; putative inner membrane protein
ycjS →	1,425,270	A→G	Y330C (TAC→TGC)	non synonymous SNP	1	72BJ203	CD	putative oxidoreductase, NADH-binding
ydbH →	1,468,368	A→G	Q709R (CAG→CGG)	non synonymous SNP	1	84BJ351	CD	conserved hypothetical protein
ydbK ←	1,463,778	C→T	V67M (GTG→ATG)	non synonymous SNP	1	79BJ351	CD	putative 2-oxoacid-flavodoxin fused oxidoreductase:conserved protein; 4Fe-4S cluster binding protein
ydcK ←	1,520,409	C→T	R157H (CGC→CAC)	non synonymous SNP	1	57AJ352	WD	putative enzyme
ydcL →	1,523,354	C→T	G156G (GGC→GGT)	synonymous SNP	2	57AJ202; 57AJ352	WD	putative lipoprotein
ydcP →	1,527,330	C→T	A54A (GCC→GCT)	synonymous SNP	2	79BJ203; 79BJ351	CD	putative peptidase
	1,528,176	C→T	G336G (GGC→GGT)	synonymous SNP	1	74AJ203	CD	
yddB ←	1,592,247	C→T	L732L (CTG→CTA)	synonymous SNP	1	74AJ203	CD	putative porin protein
yddE ←	1,552,133	C→T	D180N (GAC→AAC)	non synonymous SNP	1	79AJ203	CD	conserved hypothetical protein
ydfI ←	1,653,396	A→G	V196A (GTC→GCC)	non synonymous SNP	1	72BJ351	CD	putative mannonate dehydrogenase
ydfK →	1,656,153	A→G	T17A (ACC→GCC)	non synonymous SNP	1	74BJ351	CD	conserved hypothetical protein; putative DNA-binding transcriptional regulator; Qin prophage
ydgR → / → *gst*	1,736,279	(T)9→8	intergenic (+55/-51)	deletion	2	72AJ351; 72BJ351	CD	putative oligopeptide transporter/glutathionine S-transferase
ydhF ←	1,747,453	C→T	G45S (GGC→AGC)	non synonymous SNP	2	72AJ351; 72BJ351	CD	putative oxidoreductase
ydhJ →	1,743,419	C→T	A68V (GCG→GTG)	non synonymous SNP	5	72AJ84; 72AJ203; 72AJ351; 72BJ203; 72BJ351	CD	undecaprenyl pyrophosphate phosphatase

ydhK →	1,744,577	Δ3 bp	coding (504-506/2013 nt)	deletion	1	79BJ203	CD	conserved hypothetical protein; putative inner membrane protein
ydhY ← / ← *ydhZ*	1,776,944	C→T	intergenic (-443/+12)	intergenic SNP	1	72AJ203	CD	putative 4Fe-4S ferridoxin-type subunit of oxidoreductase/conserved hypothetical protein
ydiK →	1,791,502	(G)7→8	coding (405/1113 nt)	duplication	1	57AJ202	WD	conserved hypothetical protein; putative inner membrane protein
ydjK ←	1,877,211	A→C	G332G (GGT→GGG)	synonymous SNP	1	74AJ203	CD	putative transporter; MFS superfamily
yeaR	1,897,804	putative transposase	coding (122/360 nt)	IS transposition	1	90AJ351	CD	conserved hypothetical protein
yeaS	1,898,353	IS1	coding (384/639 nt)	IS transposition	1	78BJ351	CD	putative neutral amino-acid efflux system
yebK →	1,956,019	(A)8→9	coding (391/870 nt)	duplication	2	57BJ202; 72BJ203	WD (1)/CD (1)	putative DNA-binding transcriptional regulator
yebS →	1,936,375	G→A	D381N (GAT→AAT)	non synonymous SNP	1	87AJ351	CD	conserved hypothetical protein; putative inner membrane protein
yebV →	1,940,799	G→A	G15S (GGC→AGC)	non synonymous SNP	1	60AJ83	WD	conserved hypothetical protein
yebZ ←	1,942,923	(G)6→7	coding (646/873 nt)	duplication	2	57AJ202; 57AJ352	WD	conserved hypothetical protein; putative inner membrane protein
yedA →	2,049,536	(C)6→7	coding (191/921 nt)	duplication	1	79BJ351	CD	conserved hypothetical protein; putative inner membrane protein
yeeF–EcHS_A2182	2,131,537	NA	intergenic (-158/+21)	large deletion	5	69AJ231, 69AJ352, 69BJ231, 69BJ352	WD	32 genes
yeeO ←	2,097,525	A→G	L308L (TTA→CTA)	synonymous SNP	1	84BJ351	CD	putative Na+-driven multidrug efflux system
yegL ←	2,197,153	C→T	V73M (GTG→ATG)	non synonymous SNP	1	72AJ203	CD	conserved hypothetical protein
yehI →	2,242,802	(A)8→7	coding (176/3633 nt)	duplication	2	79BJ203; 84BJ351	CD	conserved hypothetical protein
	2,244,220	G→A	V532I (GTT→ATT)	non synonymous SNP	2	79AJ84; 79AJ203	CD	
yehR →	2,253,794	T→C	A74A (GCT→GCC)	synonymous SNP	1	74BJ351	CD	conserved hypothetical protein
yehU ←	2,255,896	C→T	M366I (ATG→ATA)	non synonymous SNP	2	66AJ322; 66BJ322	WD	sensory kinase in two-component system with YehT
	2,255,928	G→A	L356F (CTT→TTT)	non synonymous SNP	1	57AJ352	WD	
yeiC ← / ← *fruA*	2,301,581	+ATG	intergenic (-108/+316)	insertion	1	57BJ202	WD	putative kinase/fused fructose-specific PTS enzymes: IIBcomponent ; IIC components
yejO ←	2,329,283	G→A	T626M (ACG→ATG)	non synonymous SNP	1	57AJ83	WD	putative autotransporter outer membrane protein; type V secretion
yfaL ←	2,381,820	G→A	A1111V (GCC→GTC)	non synonymous SNP	1	84BJ351	CD	adhesin
yfaV ←	2,400,163	G→A	F259F (TTC→TTT)	synonymous SNP	1	84BJ351	CD	putative carboxylate transporter; phthalate permease family
yfbF →	2,408,616	G→T	L253L (CTG→CTT)	synonymous SNP	1	84BJ351	CD	undecaprenyl phosphate-L-Ara4FN transferase

yfbN	2,429,308	IS1	coding (690/717 nt)	IS transposition	4	69AJ231, 69AJ352, 69BJ231, 69BJ352	WD	conserved hypothetical protein
yfcJ ←	2,480,559	G→A	A378V (GCG→GTG)	non synonymous SNP	1	79AJ351	CD	putative transporter; UPF0226 family
yfdZ ←	2,528,962	A→G	Y246H (TAC→CAC)	non synonymous SNP	1	57AJ352	WD	putative aminotransferase, PLP-dependent
yfeA ←	2,547,591	G→A	S548S (AGC→AGT)	synonymous SNP	1	72BJ203	CD	putative diguanylate cyclase
yfeG ←	2,582,897	G→T	T328K (ACG→AAG)	non synonymous SNP	1	84AJ351	CD	putative DNA-binding transcriptional regulator
	2,582,987	C→T	R298H (CGC→CAC)	non synonymous SNP	1	72BJ203	CD	
	2,583,251	C→A	G210V (GGG→GTG)	non synonymous SNP	1	90AJ351	CD	
	2,583,527	A→G	V118A (GTG→GCG)	non synonymous SNP	3	72AJ203; 78BJ351; 84BJ351	CD	
	2,583,528	C→T	V118M (GTG→ATG)	non synonymous SNP	4	72BJ203; 74AJ203; 74BJ351; 84AJ351	CD	
yfeO →	2,542,274	(C)7→9	coding (1243/1257 nt)	duplication	1	74BJ351	CD	putative ion channel protein
yfeS →	2,565,401	G→A	G30D (GGC→GAC)	non synonymous SNP	1	79AJ351	CD	conserved hypothetical protein
yfhH →	2,737,370	G→A	D93N (GAT→AAT)	non synonymous SNP	1	57AJ202	WD	fused putative DNA-binding transcriptional regulator; putative isomerase
yfhK ←	2,728,733	G→A	Q234* (CAA→TAA)	stop SNP	3	79AJ84; 79AJ203; 79AJ351	CD	putative sensory kinase in two-component system
	2,729,411	G→A	P8S (CCC→TCC)	non synonymous SNP	1	79AJ351	CD	
yfhM ←	2,689,676	A→G	V84A (GTG→GCG)	non synonymous SNP	1	72AJ351	CD	conserved hypothetical protein
ygaD ← / ← *mltB*	2,841,895	(C)6→7	intergenic (-132/+13)	duplication	1	72AJ84	CD	conserved hypothetical protein/membrane-bound lytic murein transglycosylase B
ygaV →	2,814,941	T→C	V96A (GTC→GCC)	non synonymous SNP	1	57AJ352	WD	putative DNA-binding transcriptional regulator
ygbJ →	2,879,485	T→C	V213A (GTG→GCG)	non synonymous SNP	1	74BJ351	CD	putative dehydrogenase, with NAD(P)-binding Rossmann-fold domain
ygbN	2,883,516	IS1	coding (997/1365 nt)	IS transposition	1	78BJ351	CD	putative transporter
ygcB ←	2,903,028	(C)7→6	coding (1425/2667 nt)	deletion	1	74BJ351	CD	conserved hypothetical protein; putative member of DEAD box family
ygcN →	2,909,916	(ATT)4→3	coding (28-30/1272 nt)	deletion	1	84BJ351	CD	putative oxidoreductase with FAD/NAD(P)-binding domain
ygcW ← / → *yqcE*	2,917,538	(A)9→8	intergenic (-33/-286)	deletion	1	74AJ203	CD	putative deoxygluconate dehydrogenase/putative transporter
ygfQ →	3,055,466	G→A	A278T (GCT→ACT)	non synonymous SNP	4	72AJ84; 74AJ351; 74BJ351; 84AJ84	CD	putative transporter
yggH ←	3,131,258	(C)8→7	coding (513/720 nt)	deletion	1	79BJ351	CD	tRNA (m7G46) methyltransferase, SAM-dependent

yggN ←	3,129,957	(C)7→8	coding (582/717 nt)	duplication	1	84BJ351	CD	conserved hypothetical protein; periplasmic protein
	3,130,262	G→A	R93C (CGC→TGC)	non synonymous SNP	1	79BJ351	CD	
yggW →	3,127,224	C→T	A347V (GCC→GTC)	non synonymous SNP	1	57BJ202	WD	putative oxidoreductase
yghD ←	3,139,794	G→A	A84A (GCC→GCT)	synonymous SNP	1	74AJ203	CD	putative secretion pathway M-type protein, membrane anchored
yghK ←	3,158,659	G→A	V85V (GTC→GTT)	synonymous SNP	1	79AJ351	CD	glycolate transporter
yghU/hybG	3,184,530	IS110	intergenic (+107/+16)	IS transposition	1	84AJ351	CD	S-transferase/hydrogenase 2 accessory protein
yhaM	3,305,455	IS1	coding (35/1311 nt)	IS transposition	6	49AJ82, 49AJ201, 49AJ352, 49BJ82, 49BJ201, 49BJ352	WD	conserved hypothetical protein
	3,304,494	Δ3 bp	coding (994-996/1311 nt)	deletion	1	60BJ352	WD	
	3,304,638	Δ3 bp	coding (850-852/1311 nt)	deletion	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	
	3,304,675	C→T	G272D (GGC→GAC)	non synonymous SNP	1	84BJ351	CD	
	3,305,385	(GCC)3→4	coding (105/1311 nt)	duplication	3	60BJ83; 60BJ202; 63AJ84	WD	
yhbE ← / ← *rpmA*	3,383,739	C→T	intergenic (-78/+49)	intergenic SNP	1	74BJ203	CD	conserved hypothetical protein; putative inner membrane protein/50S ribosomal subunit protein L27
yhbT ←	3,352,140	Δ1 bp	coding (63/525 nt)	deletion	1	57BJ352	WD	putative lipid carrier protein
yheT →	3,528,431	(C)6→5	coding (887/1023 nt)	deletion	1	79BJ351	CD	putative hydrolase
yhfL → / → *frlA*	3,544,165	G→T	intergenic (+205/-39)	intergenic SNP	4	49AJ201; 49AJ352; 49BJ201; 49BJ352	WD	conserved hypothetical protein; putative secreted peptide/putative fructoselysine transporter
yhfW ←	3,553,494	T→C	Q222R (CAG→CGG)	non synonymous SNP	1	74AJ351	CD	putative mutase
yhfZ ←	3,556,207	Δ1 bp	coding (483/906 nt)	deletion	1	74AJ203	CD	conserved hypothetical protein
yhgG →	3,587,257	T→C	C64R (TGC→CGC)	non synonymous SNP	1	74AJ351	CD	putative DNA-binding transcriptional regulator
yhgN →	3,622,477	T→C	L124P (CTG→CCG)	non synonymous SNP	2	57AJ202; 57AJ352	WD	putative antibiotic transporter
yhhI →	762,695	T→C	pseudogene (171/255 nt)	SNP	1	66BJ322	WD	fragment of putative transposase (partial)
yhhL →	3,651,515	(A)6→7	coding (267/270 nt)	duplication	1	79AJ351	CD	conserved hypothetical protein; putative inner membrane protein
yhhX ← / → *yhhY*	3,627,965	(A)9→8	intergenic (-124/-208)	deletion	1	79AJ351	CD	putative oxidoreductase with NAD(P)-binding Rossmann-fold domain/putative acetyltransferase
yhhZ →	3,629,850	C→T	P318L (CCT→CTT)	non synonymous SNP	1	84AJ351	CD	conserved hypothetical protein
yhjA ←	3,709,294	G→A	P402P (CCC→CCT)	synonymous SNP	1	57BJ352	WD	cytochrome C peroxidase (cpp-like)

yhjB ←	3,713,000	C→T	V69V (GTG→GTA)	synonymous SNP	1	79BJ351	CD	putative DNA-binding response regulator in two-component regulatory system
yhjK ←	3,726,337	G→T	A187A (GCC→GCA)	synonymous SNP	1	57BJ83	WD	putative diguanylate cyclase
yiaI ←	3,770,503	G→A	P128L (CCG→CTG)	non synonymous SNP	1	79AJ203	CD	putative hydrogenase, 4Fe-4S ferredoxin-type component
yiaY ←	3,786,336	C→T	M265I (ATG→ATA)	non synonymous SNP	5	72AJ84; 72AJ203; 72AJ351; 72BJ203; 72BJ351	CD	Fe-containing alcohol dehydrogenase
yicC →	3,846,722	G→A	L155L (TTG→TTA)	synonymous SNP	1	72AJ203	CD	conserved hypothetical protein
	3,846,792	C→T	Q179* (CAA→TAA)	stop SNP	4	63AJ352; 63BJ352; 66AJ322; 66BJ322	WD	
yidC →	3,912,507	C→T	Y87Y (TAC→TAT)	synonymous SNP	1	72BJ203	CD	cytoplasmic insertase into membrane protein, Sec system
yidQ →	3,894,988	(G)6→7	coding (165/408 nt)	duplication	1	57AJ352	WD	conserved hypothetical protein; putative outer membrane protein
yidZ →	3,920,564	T→C	W210R (TGG→CGG)	non synonymous SNP	1	74AJ351	CD	putative DNA-binding transcriptional regulator
yifK →	4,010,510	C→T	P404L (CCG→CTG)	non synonymous SNP	1	74BJ351	CD	putative amino acid transporter
yiiM/cpxA	4,131,518	IS110	intergenic (+13/+93)	IS transposition	1	78AJ203	CD	conserved hypothetical protein/sensory histidine kinase in two-component regulatory system with CpxR
yjbB →	4,256,727	(T)7→8	coding (356/1632 nt)	duplication	2	84BJ203; 87AJ351	CD	putative transporter
yjbC →	4,259,223	G→A	G77S (GGT→AGT)	non synonymous SNP	1	79BJ351	CD	23S rRNA pseudouridine synthase
yjbM →	4,292,432	(T)5→4	coding (90/708 nt)	deletion	1	79BJ351	CD	conserved hypothetical protein
yjcQ ←	4,333,658	A→G	L208P (CTA→CCA)	non synonymous SNP	1	74AJ203	CD	multidrug efflux system component
yjeF →	4,419,815	(A)6→5	coding (1014/1548 nt)	deletion	1	57BJ352	WD	putative carbohydrate kinase
yjeH ←	4,393,671	A→G	L255L (CTT→CTC)	synonymous SNP	1	74AJ351	CD	transporter
yjgR ←	4,525,793	C→T	V133M (GTG→ATG)	non synonymous SNP	1	84BJ351	CD	putative ATPase
yjiH ← / → *kptA*	4,565,756	C→T	intergenic (-68/-182)	intergenic SNP	1	84BJ351	CD	conserved hypothetical protein; putative inner membrane protein/2'-phosphotransferase
	4,565,757	C→T	intergenic (-69/-181)	intergenic SNP	2	87AJ351; 87BJ351	CD	
	4,565,765	C→A	intergenic (-77/-173)	intergenic SNP	2	74AJ351; 74BJ351	CD	
	4,565,766	C→A	intergenic (-78/-172)	intergenic SNP	2	90BJ203; 90BJ351	CD	
	4,565,781	A→G	intergenic (-93/-157)	intergenic SNP	1	79BJ351	CD	
yjiR ←	4,575,795	(GA)5→6	coding (798/1413 nt)	duplication	1	74AJ351	CD	fused putative DNA-binding transcriptional regulator ; putative aminotransferase
yjiZ ← / ← *yjjM*	4,598,049	Δ3 bp	intergenic (-15/+198)	deletion	1	87BJ203	CD	putative transporter/putative DNA-binding transcriptional regulator

yjjM	4,598,259	IS1	coding (905/915 nt)	IS transposition	1	87BJ203	CD	putative DNA-binding transcriptional regulator
	4,598,323	G→A	L281F (CTC→TTC)	non synonymous SNP	1	87BJ203	CD	
	4,598,765	T→A	L133F (TTA→TTT)	non synonymous SNP	1	87BJ203	CD	
yjjY → / *→ yjtD*	4,642,665	(T)6→8	intergenic (+237/-163)	duplication	1	57BJ352	WD	conserved hypothetical protein/rRNA methyltransferase
	4,642,711	(T)6→7	intergenic (+283/-117)	duplication	4	69AJ231; 69AJ352; 69BJ231; 69BJ352	WD	
	4,642,588	IS1	intergenic (+160/-240)	IS transposition	1	74BJ203	CD	
ymcA ←	1,107,779	A→G	pseudogene (1007/1323 nt)	SNP	1	79AJ84	CD	fragment of conserved hypothetical protein (part 2)
	1,108,096	A→G	pseudogene (690/1323 nt)	SNP	6	57AJ83; 57AJ202; 57AJ352; 57BJ83; 57BJ202; 57BJ352	WD	
ymfC ← / *→ icd*	1,257,399	A→G	intergenic (-29/-143)	intergenic SNP	2	74AJ351; 74BJ351	CD	23S rRNA pseudouridine synthase/isocitrate dehydrogenase, specific for NADP+; e14 prophage
yncA ←	1,539,542	T→C	Q113R (CAG→CGG)	non synonymous SNP	1	74BJ351	CD	putative acyltransferase with acyl-CoA N-acyltransferase domain
ynfE →	1,680,639	(A)7→8	coding (436/2427 nt)	duplication	1	74BJ351	CD	oxidoreductase subunit
ynfL ←	1,691,153	T→C	E192G (GAA→GGA)	non synonymous SNP	1	84BJ351	CD	putative DNA-binding transcriptional regulator
ynfN	1,667,102	IS3	coding (66/156 nt)	IS transposition	1	90BJ84	CD	conserved hypothetical protein; Qin prophage
ynfN ← / *← cspI*	1,667,226	(A)8→7	intergenic (-59/+288)	deletion	1	72AJ84	CD	conserved hypothetical protein; Qin prophage/cold shock protein; Qin prophage
yodB → / *← EcHS_A2076*	2,062,205	(A)6→7	intergenic (+207/+695)	duplication	1	79BJ351	CD	putative cytochrome/conserved hypothetical protein
yohL ← / *→ rncA*	2,228,224	(G)7→6	intergenic (-79/-42)	deletion	1	79AJ351	CD	conserved hypothetical protein/nickel and cobalt resistance
ypdG ←	2,538,625	G→A	N224N (AAC→AAT)	synonymous SNP	1	79AJ351	CD	putative enzyme IIC component of PTS
ypfE ← / *← maeB*	2,597,911	(T)8→7	intergenic (-127/+94)	deletion	4	74AJ203; 79BJ203; 79BJ351; 84BJ351	CD	putative carboxysome-like ethanolaminosome structural protein with putative role in ethanol utilization/putative fused malic enzyme oxidoreductase ; putative phosphotransacetylase
ypfJ ←	2,618,257	G→A	P130S (CCA→TCA)	non synonymous SNP	1	84BJ351	CD	conserved hypothetical protein
ypjA → / *← minE*	1,272,433	(A)7→6	intergenic (+77/+295)	deletion	1	74AJ203	CD	fragment of adhesin-like autotransporter (partial)/cell division topological specificity factor
yqcA	2,940,765	SNP	coding (429/450 nt)	SNP	2	78AJ351, 78BJ203	CD	putative flavoprotein
yqcE →	2,918,126	G→A	M101I (ATG→ATA)	non synonymous SNP	1	57BJ83	WD	putative transporter

yqeG →	3,004,864	(A)7→6	coding (697/1230 nt)	deletion	1	72AJ203	CD	putative transporter
yqhC ←	3,199,869	G→A	R61C (CGT→TGT)	non synonymous SNP	1	74BJ203	CD	putative DNA-binding transcriptional regulator
yqiC →	3,229,689	(A)7→8	coding (79/351 nt)	duplication	1	79AJ351	CD	conserved hypothetical protein
yqiG →	3,233,306	(C)9→10	coding (2451/2502 nt)	duplication	2	57AJ352; 74BJ351	WD	outer membrane usher protein
	3,233,306	(C)9→8	coding (2451/2502 nt)	deletion	1	72AJ203	CD	
	3,233,306	(C)9→12	coding (2451/2502 nt)	duplication	2	72AJ351; 72BJ351	CD	
yqiK →	3,237,330	T→C	pseudogene (20/606 nt)	SNP	1	57AJ352	WD	fragment of conserved hypothetical protein (part 2)
ytfA	4,461,085	IS1	coding (546/648 nt)	IS transposition	2	60BJ202, 60BJ352	WD	putative transcriptional regulator
	4,461,008	C→T	Q157* (CAG→TAG)	stop SNP	1	49AJ82	WD	
	4,461,101	(C)6→7	coding (562/648 nt)	duplication	2	63BJ84; 63BJ203	WD	
	4,461,101	C→T	Q188* (CAG→TAG)	stop SNP	6	49AJ201; 49AJ352; 49BJ82; 49BJ201; 49BJ352; 60BJ83	WD	
zraR →	4,234,004	A→G	E375G (GAA→GGA)	non synonymous SNP	1	84BJ351	CD	fused DNA-binding response regulator in two-component regulatory system with ZraS: response regulator ; sigma54 interaction protein
zraS →	4,232,211	T→C	L235P (CTG→CCG)	non synonymous SNP	2	79BJ203; 79BJ351	CD	sensory histidine kinase in two-component regulatory system with ZraR
	4,232,321	G→A	E272K (GAA→AAA)	non synonymous SNP	1	79AJ351	CD	
	4,232,349	C→T	A281V (GCG→GTG)	non synonymous SNP	1	84AJ351	CD	
	4,232,382	G→A	S292N (AGC→AAC)	non synonymous SNP	1	78AJ351	CD	

<u>ANNEXE N°9:</u> ARTICLE "USING LONG-TERM EXPERIMENTAL EVOLUTION TO UNCOVER THE PATTERNS AND DETERMINANTS OF MOLECULAR EVOLUTION OF AN ESCHERICHIA COLI NATURAL ISOLATE IN THE STREPTOMYCIN-TREATED MOUSE GUT" DE LESCAT *ET AL*. PUBLIÉ DANS MOLECULAR ECOLOGY EN 2017

Published in final edited form as:
Mol Ecol. 2017 April ; 26(7): 1802–1817. doi:10.1111/mec.13851.

Using long-term experimental evolution to uncover the patterns and determinants of molecular evolution of an *Escherichia coli* natural isolate in the streptomycin treated mouse gut

Mathilde Lescat[#1,2,3], **Adrien Launay**[#1,4], **Mohamed Ghalayini**[1,2], **Mélanie Magnan**[1,4], **Jérémy Glodt**[1,4], **Coralie Pintard**[1,4], **Sara Dion**[1,4], **Erick Denamur**[1,4,5], and **Olivier Tenaillon**[1,4,#]

[1]INSERM, IAME, UMR 1137, Paris, France

[2]Univ Paris Nord, Sorbonne Paris Cité, Bobigny, France

[3]APHP, Hôpitaux Universitaires Paris Seine Saint-Denis, Bondy, France

[4]Univ Paris Diderot, Sorbonne Paris Cité, Paris, France

[5]APHP, Hôpitaux Universitaires Paris Nord Val de Seine, Paris, France

[#] These authors contributed equally to this work.

Abstract

Though microbial ecology of the gut is now a major focus of interest, little is known about the molecular determinants of microbial adaptation in the gut. Experimental evolution coupled with whole genome sequencing can provide insights of the adaptive process. *In vitro* experiments have revealed some conserved patterns: intermediate convergence, epistatic interactions between beneficial mutations and mutations in global regulators. To test the relevance of these patterns and to identify the selective pressures acting *in vivo*, we have performed a long-term adaptation of an *E. coli* natural isolate, the streptomycin resistant strain 536, in the digestive tract of streptomycin treated mice. After a year of evolution, a clone from 15 replicates was sequenced. Consistently with *in vitro* observations, the identified mutations revealed a strong pattern of convergence at the mutation, gene, operon and functional levels. Yet, the rate of molecular evolution was lower than in *in vitro* and no mutations in global regulators were recovered. More specific targets were observed: the *dgo* operon, involved in the galactonate pathway that improved growth on D-galactonate, and *rluD* and *gidB*, implicated in the maturation of the ribosomes, which mutations improved growth only in the presence of streptomycin. As *in vitro*, the non-random associations of mutations within the same pathways suggested a role of epistasis in shaping the adaptive landscape. Overall, we show that “evolve and sequence” approach coupled to an analysis of

[#]**Corresponding author:** Olivier Tenaillon, IAME, INSERM UMR 1137, Faculté de Médecine Paris Diderot, Site Xavier Bichat, 16 rue Henri Huchard, 75018 Paris, France. olivier.tenaillon@inserm.fr.

Data accessibility

The date used to produce the figure 1, 4, 5 and 6 are accessible on the Dryad website : (http://datadryad.org) with the number doi: 10.5061/dryad.4g503 and in the table S1

The date used to produce the figures 2 and 3 are presented in the table S2 and in the MicroScope website (http://www.genoscope.cns.fr/agc/microscope/home).

Europe PMC Funders Author Manuscripts

convergence, when applied to a natural isolate, can be used to study adaptation *in vivo* and uncover the specific selective pressures of that environment.

Keywords

Escherichia coli; experimental evolution; gut colonization mouse model; 536; streptomycin resistance; D-galactonate

Europe PMC Funders Author Manuscripts

Background

Bacteria have evolved for billions of years and colonized all available ecological niches. Many studies have been performed to understand how evolution has shaped the genome of these complex organisms. Comparative genomics as unravelled the dynamics of gene acquisition and losses in bacterial genomes (Touchon *et al.* 2009), and the general constraints acting on genome composition (Hershberg & Petrov 2010). Yet, much less is known about bacterial genomes evolution in the short term in their natural habitat. While we know through the emergence of resistance to antibiotics that bacterial evolution can be caught in the act, most of our evidence of short-term evolution in the wild is linked to drastic selective pressures such as antibiotics (Wong *et al.* 2012) or escape of the immune system (Flint *et al.* 2016). Can other selective pressures be strong enough to drive significant changes in bacterial genomes in the sort term? To investigate that possibility, we have decided to study the evolution of a natural isolate of *Escherichia* coli in the gut of mice.

E. coli is one of the most characterised bacterial species; it has been used as a model organism in the laboratory to unravel the multiple facets of bacterial biology from biochemistry to genetics. Yet, the natural environment of *E. coli* is not the laboratory. In the wild, *E. coli* is a versatile species, known as both a widespread gut commensal of the vertebrates and a dangerous pathogen that can also be retrieved in the environment (Tenaillon *et al.* 2010). As a pathogen *E. coli* is responsible for about 1 million human deaths yearly due to intra and now mostly extra intestinal diseases (Russo & Johnson 2003). The species is a structured species subdivided in seven phylogenetic groups A, B1, B2, C, D, E and F (Jaureguy *et al.* 2008). These groups are not randomly distributed. Previous studies have shown an association between the phylogenetic groups and virulence, with most extra intestinal *E. coli* pathogens (including urinary tract infection and meningitis strains) belonging to phylogenetic group B2 (Bingen *et al.* 1998; Picard *et al.* 1999). The prevalence of the different phylogenetic groups as commensals varies largely across host species and even within host populations (Tenaillon *et al.* 2010). For instance, the B2 phylogenetic group can be frequently retrieved in wild animal's guts (Gordon & Cowling 2003; Lescat *et al.* 2013), rarely found (<10%) as a commensal among human tropical populations while it is dominant (>30%) among the populations of industrialized countries (Tenaillon *et al.* 2010). Indeed, over the last decades, the prevalence of the B2 phylogroup has significantly increased in frequency in some western countries such as France. B2 phylogroup strains being both highly efficient extra intestinal pathogens, and more and more prevalent as commensals in industrialized countries, they are becoming an increasing public health

concern. Yet, the molecular determinants explaining the differences in prevalence of the different subgroups of *E. coli* among different hosts or populations remain largely unknown.

To identify the molecular bases that may lead *E. coli* to specialize to certain environmental conditions, two main approaches can be used: comparative genomics of natural isolates and experimental evolution. Thanks to the sequencing revolution, more than two hundred complete genomes of *E. coli* are now accessible and thousands more are underway. We can subsequently compare these genomes in order to identify some genetic characteristics likely to contribute to the phenotypic differentiation of the subgroups. For instance, we compared 130 sequenced genomes to identify the core region that was the most different between B2 to non-B2 strains: the *nhaAR* operon involved in pH control. Though we could find a contribution of the operon in extra intestinal virulence, we were not able to observe its implication in colonization of the mouse gut (Lescat *et al.* 2014) nor in any pH linked phenotypes we could measure in the laboratory. This example illustrates the limitation of comparative genomics: though some specific genomic signatures may be found, we have not access to the environmental conditions that shaped them and it is therefore difficult to assess their evolutionary relevance.

An alternative approach relies in evolving bacteria in controlled conditions. *In vitro* experimental evolution has been used for decades to study bacterial adaptation. Over the last years, the coupling of experimental evolution with whole genome sequencing allowed progresses in the understanding of the genetic bases of adaptation. The most charismatic experiment in the field is the one initiated more than 25 years ago by Richard Lenski. His 12 lineages long-term evolution experiment (LTEE) has now reached more than 60,000 generations. The genomic data as well as targeted sequencing approaches have revealed convergence among the first targets of adaptation involving several global regulators (Maddamsetti *et al.* 2015). Another large scale study confirmed these results using 115 lineages adapting for 2,000 generations to high temperature: the first target of adaptation was the RNA polymerase one of the most pleiotropic gene in the genome, and a large convergence was observed between replicates (Tenaillon *et al.* 2012). These studies and others (Herring *et al.* 2006; Hindré *et al.* 2012; Le Gac *et al.* 2013) suggest that the first steps of adaptation may involve global modifications of gene expression rather than a specific small-scale optimization of some specific pathways. Interestingly, global regulators such as *rpoS*, have also been found to be involved in the diversification that seems to occur during infections (Levert *et al.* 2010). However, the relevance of these *in vitro* or *in infectio* observations in the primary habitat of *E. coli,* the gut of vertebrates, remains to be tested.

In vivo assays have also been performed to study adaptation to a more relevant environment for *E. coli*. In the early 2000's, Giraud *et al.* tested the implication of mutator alleles of *E. coli* in the adaptation to the gut of germ free mice and showed that mutator bacteria possessed a short-term advantage and long-term disadvantage in colonizing new environments (Giraud *et al.* 2001). Later, Giraud *et al.* also used germ-free mice inoculated with *E. coli* MG1655 (Giraud *et al.* 2008) and observed a rapid adaptation by a single point mutation in EnvZ/OmpR, a two component signal transduction system, which controls more than 100 genes. These observations suggest that global regulators could also be the targets of adaptation in the mouse gut. However, these works are entailed by an important limitation:

the gut of adult germ free mice is an artificial environment. Mice that have developed without flora have abnormal immune system and gut physiology (Khoury *et al.* 1969; Welling *et al.* 1980; Umekasi 2014). Furthermore, in these experiments, *E. coli* was alone in the gut. In the wild, *E. coli* is a minority species of the gut microbiota, which is largely dominated by numerous anaerobes (Arumugam *et al.* 2011). As a consequence, *E. coli* is facing intense competition for nutrients, and may also have access to other nutrients processed by the rest of the microbiota. More recently, Barroso-Batista *et al.* performed an evolution experiment using the laboratory strain *E. coli* K-12 in streptomycin treated mice. Streptomycin treatment allows the colonization of the digestive tract by a streptomycin resistant *E. coli*, without altering theoretically the anaerobic microbiota (Barroso-Batista *et al.* 2014). They observed the rapid apparition of mutations in the different lineages leading to convergence in inactivation of the *gat* operon conferring advantage to the strains bearing the mutations and unable to utilize the galactitol. Here the first target of adaptation was not a global regulator. However, one limitation of these studies is the use of a laboratory strain that has been subject to many genetic alterations and that has not seen a gut for close to a century. *E. coli* K-12 is a strain evolved in *in vitro* conditions since 1922, date of its isolation in Palo Alto, from a diphtheria convalescent patient. Genetic modifications, adaptation to oxygen rich media from the laboratory as well as accumulation of mutations favoured by UV irradiations or strong bottlenecks may modify its adaptation to the mouse gut.

In this context, the aims of this study are (i) to identify the basis of molecular adaptation in *in vivo* commensal conditions using long-term evolution during one year of a natural isolate of *E. coli* belonging to the phylogenetic group B2 in the digestive tract of streptomycin treated mice and (ii) to test the phenotypic significances of the molecular modifications observed and their possible advantages in this natural environment.

Materials and Methods

Bacterial strains

We used *E. coli* K-12 BW25113 (phylogenetic group A) from the KEIO collection (Baba *et al.* 2006) and 536 (phylogenetic group B2, serotype O6:K5:H31) reference strains. The 536 strain has been isolated from a patient suffering from urinary tract infection in the early 1980's (Berger *et al.* 1982) and subsequently shown to be a good gut colonizer in streptomycin treated mouse (Diard *et al.* 2010). This strain was acquired by our laboratory in 1998 from Jörg Hacker who isolated it and then stored at -80°C with glycerol.

Mutant strains of K-12, obtained from the KEIO collection (K-12Δ*dgoR*:kan, K-12Δ*rluD*:kan, K-12Δ*gidB*:kan) and of 536, obtained during the course of this study using the lambda-red disruption method (Datsenko & Wanner 2000) (536Δ*dgoR*:cm, 536Δ*rluD*:cm, 536Δ*gidB*:cm, 536Δ*gidB*Δ*rluD*:cm), were also used. For the flux cytometry assays, we used K-12 p*dgoR* issued from the Zaslaver collection (Zaslaver *et al.* 2006) which beared the plasmid p*dgoR* which contained the *gfp* gene placed under the control of the promoter of the *dgoRKADT* operon. After extraction of this plasmid, we electroporated it in the K-12Δ*dgoR*:kan mutant strain. We thus obtained K-12Δ*dgoR* p*dgoR* strain.

Europe PMC Funders Author Manuscripts

Mouse intestinal colonization assay and sampling strategy

All animal experiments were performed in compliance with the recommendations of the French Ministry of Agriculture and approved by the French Veterinary Services (accreditation A 75-18-05). The mouse gut colonization model was conducted after approval by the Debré-Bichat Ethics Committee for Animal Experimentation (Protocol Number 2012-17/722-0076) in accordance to the European Decret and French law on the protection of animals. Intestinal colonization by 536 *E. coli* strain was performed using a mouse model as described elsewhere (Diard *et al.* 2010). Thirty 6-week-old female outbred mice (Charles River CD-1) were housed in 15 cages (numbered from C1 to C15) during one year (two mice per cage named Ma and Mb). Five days before inoculation of the 536 strain, and throughout the experiment, streptomycin was added to the sterile drinking water at a final concentration of 5 g/L. Likewise the mice were fed with sterile food all along the experiment. Mice free of coliform flora (controlled by plating the faeces on Drigalski plates) were inoculated *per os* with 10^6 bacteria in 200 µl of physiological water. The day after the inoculation and during the experiment (Days 1, 5, 33, 47, 155, 188, 217, 294, 336 and 363, the intestinal population of *E. coli* was estimated by plating dilutions of weighted fresh faeces for each mouse on lysogeny broth (LB) agar with streptomycin (5 g/L) and supplementary samples were performed for sequencing assays (Days 61, 124, 155 and 250). Faeces were harvested for each mouse by putting the mouse in a sterile plastic cage until it defecated. The faeces were stored at -80°C with glycerol for further analysis. For sequencing, the frozen faeces were diluted and plated on LB agar with streptomycin. One colony per LB plate was then randomly picked, grown in LB, stored at -80°C with glycerol and thereafter called “clone”. In addition, to study the reproducibility between the sample of the two mice from each cage, 50 colonies were picked from the faeces on the LB plate for each mouse (100 colonies in total) in cage 1 at Day 363. For the polymorphism study, we picked a pool of about 50 colonies from 4 mice (C2Ma, C3Ma, C8Ma and C11Ma) at both the last time point (Days 250, 336, 363 and 363, respectively) and an intermediate time point (Days 124, 155, 155 and 155, respectively).

High throughput sequencing

From the glycerol stocks, the clones were plated on LB plates, and one colony was randomly picked, and DNA was extracted using Wizard® genomic DNA purification kit (Promega). The genomes of the evolved clones, as well as that of the ancestor 536 strain used for inoculation were sequenced on the Illumina HiSeq 2000 (IntegraGen), with one line of single-end 51-bp reads per genome. We also sequenced after DNA extraction the pool of colonies for the reproducibility and polymorphism assays.

Sequence analyses

Point mutations were identified by comparing sequence reads to the genome of the ancestral strain 536, considered as the reference, using breseq 0.23, a pipeline for analyzing resequenced microbial genomes (Deatherage & Barrick 2014).

Reference reconstruction—In a first step, we sequenced the 536 reference genome and compared the reads to the original strain genome (Brzuszkiewicz *et al.* 2006) from the

MicroScope website (https://www.genoscope.cns.fr/agc/microscope/mage/viewer.php) to identify differences using breseq 0.23 (Deatherage & Barrick 2014). This pipeline maps the reads on an annotated reference genome in gbk format to precisely identify the mutations that occurred: the type of the mutation, the genes it fell into, or the closest ones, and in the case of point mutation the changes in amino acid associated. We then used gdtools utility program from breseq 0.27.1 (Deatherage & Barrick 2014) to integrate the identified mutations inside the reference genome. We repeated this operation of comparison and modification of the reference five times until there was no more mutation identified. The iterations are required to correct mutations in close neighbourhood of one another. Using this method, we identified a total of 351 mutations (230 being point mutations, and 121 indels) between our ancestral 536 strain and the one from the database we used, a number compatible with errors due to Sanger sequencing method at the genome level.

Detection of mutations—Once the reference genome was obtained, we compared this reference to the sequences of the different evolved clones using the same approach. The identified mutations were then manually checked. We did the same for the pooled colonies and also obtained the frequencies for each mutation from breseq 0.27.1(Deatherage & Barrick 2014).

Quantification of convergence—To quantify how significant convergence at the gene level is, for each gene of the genome, we computed the expected number of mutations E_i, that would alter this gene if mutations were randomly spread in the chromosome. We compared that expectation to the observed results O_i through the computation of a G statistics for nonsynonymous mutations: in which the sum is made on all genes with some observed mutations. The number of degrees of freedom of G statistics being difficult to compute when many classes are null (genes with no mutations) we compared the observed G value to that of simulated dataset with mutations randomly spread (Tenaillon *et al.* 2016).

Phenotypic assays

Flux cytometry—The strain K-12 p*dgoR* and the mutant strain K-12Δ*dgoR* p*dgoR* were compared using flux cytometry as described elsewhere (Bleibtreu *et al.* 2013).

Growth curves—For the comparative growth assays, strains were grown at 37°C in four media, all supplemented with streptomycin at 5 g/L: LB and minimum Davies medium (DM) with glucose, gluconate or galactonate (20 mmol/L for each sugar). Each medium was adjusted at pH 7 and pH 5.7 with MOPS® and TAPS® products, respectively (Sigma). LB is a complex media, whereas DM is a minimal media with only one source of carbon. All the strains studied were grown overnight (O/N) in LB media in deep-weel plates at 37°C with constant stirring at 280 rpm. O/N cultures were prediluted at 1/100 in physiological serum and strains were inoculated each at 1/100 in a Costar® 96 flat-bottomed well plate and recovered by 50µl of mineral oil. Growth was recorded by an Infinite 200 Tecan®, which measured the OD600 in each well every 5 minutes at 37°C, while stirring for 24 hours. Growth assays were repeated 3 times. The maximum growth rate (MGR) was calculated from growth curves obtained by Tecan®. OD600 in nm were collected and Log-transformed. Curves were calculated from a smoothed spline function. The MGR was defined as the time

point at which the maximum value of the derivative of the smoothed function was observed. The doubling times (DTs) in mn have then been computed as followed: DT=Log2/(MGR*60).

Statistical analysis

The DTs were compared by strain and by medium using a Welch test. All statistics were computed using R (R Development Core Team, 2009, Vienna, Austria) and statistical significance was determined at a p-value of less than 0.05.

Results

Mouse colonization assay and clone sampling

Thirty mice in 15 independent cages (2 mice per cage) were colonized by the *E. coli* 536 strain. As it is known that mice are coprophages, we postulated that sampling one clone of one randomly selected mouse per cage should be representative of the population of *E. coli* colonizing both mice within a single cage. To validate this approach, we sampled in cage 1 at Days 61, 155, 250 and 363,, one clone from each mouse as well as a mix of 50 clones at Day 363 for each mouse and performed sequencing of the individual clones and the mix. We observed a single mutation difference between the two individual clones per cage and this site was the only polymorphic site recovered in the pooled sequencing. These results validate our strategy, sampling a single mouse per cage being relevant.

The 536 strain was maintained in all alive mice throughout the experiment that lasted 363 days. However, ten mice died in the course of the experiment (figure 1), most of these were issued from different cages except for 2 cages (cages 7 and 14) in which the two mice died before day 363 (supplementary table 1). The average population density was $1.4.10^9$ unity forming colony (UFC)/g of faeces along the experiment. More precisely, density first decreased from 3.10^9 UFC/g of faeces to 2.10^4 UFC/g of faeces at day 188 and stabilized at 6.10^7 until the end of experiment at day 363 (figure 1). One clone per cage has been used for sequencing. The sampling points corresponds (i) to the end of the experiment when at least one of the two mice of the cage was still alive (Day 363 for 10 cages: C1, C4, C5, C6, C8, C9, C10, C11, C13, C15), (ii) the time at which the last mouse of the cage died (Day 250 for cage C2, C7 and C14), and for three cages we could not recover *E. coli* from the frozen faeces and we sampled the last point that produced an *E. coli* clone (Day 124 for cage C12 and Day 336 for C3) (supplementary table 1).

Description of mutations in the 536 evolved clones

Using breseq pipeline (Deatherage & Barrick 2014), we identified 95 independent mutations that occurred during the experiment of mice gut evolution in the 15 genomes as shown in the supplementary table 2. Amongst those mutations, 70 were point mutations, 17 were small indels (less than 50bp), 6 were large indel events (>25bp) and two were IS insertion. For the point mutations, 44 were non-synonymous including 10 stop mutations; 13 were synonymous and 13 intergenic. The synonymous mutations recovered were only transitions and 11 out of 13 were of C-G to T-A type, a bias reported before (Wielgoss *et al.* 2011). Assuming 18 generations a day in the mice gut (Rang *et al.* 1999), we could compute a

mutation rate of 1.85 10^{-10} per base-pair per generation using synonymous mutation. This is consistent with previous estimates of mutation rate of 8.9 10^{-11} found in the LTEE (Wielgoss *et al.* 2011). The ratio of rate of non-synonymous to the rate of synonymous, Ka/Ks, was 1.90 (Confidence Interval (CI) estimated by bootstrapping mutations: 1.09-4.01). Such a value suggests that about 50% of non-synonymous mutations (including nonsense mutations) are the product of natural selection. Stop codons were even more overrepresented in the dataset compared to the expectation based on synonymous (Kstop/Ks=6.72, CI: 2.78-15.86) or on all non-stop mutations (Kstop/Ka_nonstop=4.51, CI: 1.96-8.39). This suggests a strong selection for gene inactivation.

Among the 17 small indels (<50bp), 13 lead to frameshifts, and none were intergenic. For the 10 lineages that survived over the 363 days, an average of 7 mutations were found per sequenced genome (4 to 11).

Convergence among lineages

Convergence was observed at several levels. First, while 95 mutations were recovered, they corresponded indeed to 76 different mutations. Six mutations were recovered in 2 to 8 lineages. For instance 8 lineages shared the same C to T mutation in an intergenic region between *yjiH* and *kptA*, and 5 shared a G to A non-sense mutation in *dgoR*. While these shared mutations could be a signal of cross contamination, the star like phylogeny of the samples and the mapping of mutations on that phylogeny suggest that these common mutations must have occurred multiple times (figure 2). In other words, using the previous two examples, the lineages having the *yjiH*/*kptA* mutations were not fully included in, fully encompassing or mutually exclusive with the lineages having the afore mentioned *dgoR* mutations: some lineages have both mutations, one, the other or none. This suggests that if the population evolved asexually, in at least one locus the same mutations occurred several times independently. These convergent mutations account for 26% (25/95) of the mutations found.

Convergence could be integrated further at the gene level. Nine genes were found to be mutated repeatedly in different cages. Adding a large deletion recovered 3 times, the mutations showing some sign of convergence among lineages at the gene level accounted for 56% (53/95) of the mutations found. With the limited number of mutations found in each genome, this is a clear sign of convergence (see below). This signal is further reinforced by the fact that all genes affected multiple times presented at least 2 different mutations at the origin of the convergence (table 1). The fact that different alleles are found in the different lineages, excludes the possibility that convergence is only due to a single mutation that was by chance at high frequency in the initial stock used to inoculate the different mice. For instance, ECP_4610, a putative transporter is affected in 5 lineages with 5 different mutations.

To quantify how significant this convergence at the gene level is, for each gene of the genome, we computed a statistical test for convergence (see method) in the form of a G statistics (Tenaillon *et al.* 2016). We found G=482.3 that has to be compared to a mean value of 399.2 for simulations, with a Z-score of 8.3. This suggests that the observed value is more than 8 standard deviations away from the simulations mean and results in a p-value lower

than 2.10^{-16}. This overwhelming signal of convergence suggests the filtering action of natural selection.

Among this list of genes showing signs of convergence, two clusters of linked genes can be further identified. *dgoR, dgoD* and *dgoT* are implicated in the same metabolic pathways. This observation suggests further that convergence may occur at a higher level of integration than the gene. In the same manner, 6 mutations in the *rluD* gene were found, and 7 in *gidB*. Both genes are implicated in the maturation of the ribosomes, and both are linked to streptomycin resistance.

Polymorhism within lineages

To have a better insight on the dynamics of adaptation, we sequenced a pool of about 50 colonies from 4 mice (C2Ma, C3Ma, C8Ma and C11Ma) at both the last time point (Days 250, 336, 363 and 363, respectively) and an intermediate time point (Days 124, 155, 155 and 155, respectively). Using an analysis of polymorphism in these data, we can detect mutations with a frequency higher than 5% in most samples. The analysis of the mutations found at the population level confirmed the presence of the mutations of interest we detected in the clone sequenced. Polymorphism data revealed however if mutations were fully fixed (frequency of 100%) in the population or if they had reached intermediate frequencies. Moreover, several mutations affecting the same gene were found simultaneously in some populations. This suggests some competition between these various alleles. Interestingly, all cases of multiple allele affecting the same gene were restricted to the genes we identified as target of selection with the clone sequencing approach (supplementary table 2). This confirms that we have identified the biggest targets of selection.

The loci under selection showed traces of polymorphism in two ways (supplementary table 3). Either a single allele was present in the population but not fixed or several alleles were found simultaneously in the population. *gidB* showed the strongest sign of polymorphism. In mouse C11Ma, at day 155, the whole population seems to have mutated *gidB*, but 4 alleles are involved: a 25bp deletion, a frameshift, P79L and Q23* at respective frequencies of 8%, 60.2%, 8.7% and 31.3% (note that the frequencies are based on local coverage statistics so the sum is not supposed to be precisely 1). At day 363, the 25bp deletion overtook the other alleles and reached a frequency of 100%. In mice C2Ma, two alleles of *gidB*, Q167* and the Δ25bp deletion, are found polymorphic at both sampling time. *dgoR* also had traces of polymorphism, with C8Ma having two competing alleles (Δ360bp at 75.3% and Q212* at 20.9%) at one time point (Day 155). Out of the 3 mice studied for polymorphism that had mutations in *dgoR* at the last sampling time point, only in one case fixation was reached, while in the others the frequency of the alleles was of 55.8% and 74.1%. The study of polymorphism revealed also new alleles in some of the locus of interest. A mutation in the promoter of *dgoR* reached a frequency of 7.4% in C2Ma at day 124 and another allele between *yjiH* and *kptA* reached a frequency of 24% in C8Ma at Day 155 before being out competed by another allele. Finally, in mouse C11Ma, while we did not find any mutation in *dgoR* in the clone sequenced at the last time point, a mutation was recovered at high frequency in that gene at an earlier time point, but disappeared later on.

The comparison of the alleles found polymorphic in the different populations studied revealed another gene with evidence of convergence. Clone sequencing revealed a frameshift in *uxuR* in mouse C11Ma. The polymorphism analysis revealed that this mutation was fixed in that population. In 2 other mice we found mutations affecting *uxuR*. These mutations were at low frequencies (Q116* at 16.7% in mouse C2Ma, A213V at 6.6% in mouse C3Ma) which explains why we have not sampled them with the clone approach. UxuR is a transcriptional factor that regulates a cluster of operons involved in transport and degradation of the sugar acids β-D-glucuronides, glucuronate, gluconate and L-galactonate.

We further focused on both *dgoRKADT* operon and *rluD* and *gidB* complex as the other important convergence targets, the intergenic region between *yijH* and *kptA* (hit 11 times but with only 3 alleles) or ECP_4610 (hit 5 times) affect proteins of unknown functions. To limit the use of animals, functional tests were made only *in vitro*.

Galactonate operon

Role of *dgoR* and *dgoRKADT* operon—Among the most notable mutated genes, the genes related to the galactonate operon, *dgoRKADT,* seemed particularly interesting. Seventeen mutations comprising 6 stop codons, 2 frameshifts, and 3 missenses were recovered in *dgoR*, 4 missenses in *dgoD* and 2 missenses in *dgoT* such that in 11 out of the 15 cages the operon was mutated (figure 3). The operon is composed of *dgoR* (coding for the repressor), *dgoK* (coding for a kinase), *dgoA (*coding for a aldolase), *dgoD* (coding for a dehydratase) and *dgoT* (coding for the D-galactonate transporter). This operon is responsible for the utilization of the D-galactonate sugar in a pathway that is somehow parallel to the Entner Doudouroff pathway (EDp): contrary to this pathway the input is D-galactonate rather than gluconate but likewise to the EDp the outputs are pyruvate and D-glyceraldehyde 3P (Lamble *et al.* 2004). We confirmed that *dgoR* is the repressor of the pathway, which is expressed at low level in the absence of D-galactonate and allows the inhibition of the expression of the operon. We observed this regulator activity using reporter strains carrying the *gfp* gene under control of the *dgoRKADT* operon promoter in K-12 strain. In LB, which does not contain galactonate, no fluorescence was observed. The presence of D-galactonate, induced fluorescence that was revealed using flux cytometry. Moreover the strain *E. coli* K-12Δ*dgoR* (Baba *et al.* 2006) using the same construct presented a constitutive induction of fluorescence in both media (figure 4). These results confirmed the repressor activity of DgoR, which had not yet been observed. To test the effects of the *dgoR* mutations selected for in our experiments, one evolved lineages of 536 obtained at the end of the experiment from the mouse C8Ma and the ancestor 536 were transformed with the reporter construct plasmid *pdgoR*. It showed a constitutive expression of fluorescence similar to the one found in *E. coli* K-12Δ*dgoR*, confirming that inactivation of the repressor has been selected for (data not shown).

Interestingly, mutations in *dgoD* always occurred in lineages having mutation in *dgoR* and *dgoT* mutations always occurred in lineages having also *dgoR* and *dgoD* mutations (p=0.009). For instance lineages 5 and 13 both had the 3 genes modified, using 6 different mutations.

To test for a potential benefit linked to mutations in the *dgoRKADT* operon, we cultured the 15 evolved lineages in different conditions. Growth curves were assayed in different media (LB and minimum medium with glucose or gluconate as the sole carbon source) at different pH (5.7 and 7) and in minimum medium with D-galactonate at pH 5.7 because the sugar is not soluble at neutral pH (figure 5). We then compared the DTs of the seven evolved strains with mutations in *dgoR*, the two evolved strains with mutations in *dgoR* and *dgoD* and the two evolved strains with mutations in *dgoR, dgoD* and *dgoT* and the four evolved lineages unaffected in *dgoR* as well that of the ancestor strains. A significant difference in growth was found in minimum medium with D-galactonate between *dgoR* mutant strains and the others using a Welch test (p=0.000114). In more details, the data suggest that *dgoR* and *dgoR-dgoD* (p=0.001) mutants have a higher growth rate than strains non mutated in *dgo* operon, and that strains with the three mutations *dgoR*, *dgoD* and *dgoT* grow even better than the ones with only two (p=0.0014). Of note, truncating (non sense and frame shift) mutations were observed only in the regulator coding gene *dgoR*. This suggest that despite an induction of the operon of *dgoRKADT* operon by D-galactonate, inactivation of the repressor of the operon promotes growth in D-galactonate, and that growth is further improved when genes implicated in the pathway of degradation of the D-galactonate are mutated. No difference was observed in LB, minimum medium with glucose and minimum medium with gluconate between the wild type and the mutant strains at both pH 5.7 and 7 (data not shown).

Focus on *rluD* and *gidB*

rluD and *gidB*, are other genes that attracted our attention. Specifically, 6 mutations, including frame shifts were found in *rluD*, and 7 mutations, including 25bp deletions were recovered in *gidB*. *rluD* encodes for a pseudo uridin synthetase that catalyses the formation of pseudo uridine on the 23S rRNA, in other words *rluD* is involved in 23S rRNA maturation. Similarly *gidB* is implicated in the maturation of the 16S subunit of the ribosomes as it encodes for a S-adenoside methionine dependant methyltransferase. In other bacterial species, both genes have been shown to be involved in streptomycin resistance, as streptomycin antibiotic targets the ribosome (Okamoto *et al.* 2007; Sato & Iino 2010). It is therefore very likely that these mutations have been selected for the adaptation to the antibiotic selective pressure that was maintained during our experiment. Moreover the selected functions of both genes seem to be associated as no evolved lineage presented mutations in both genes. Given the frequency of these mutations (7 and 6 mutations), we could find a significant repulsion between the mutations (test of non random association using a correlation: $p<0.001$), suggesting that once one is found in one of the genes the acquisition of a mutation in the other is no more selected for.

We first wanted to test if these mutations increased resistance to streptomycin in our strain. First, we tested if *E. coli* K-12 mutants of those two genes, K-12Δ*rluD* and K-12Δ*gidB* had a higher resistance level by determining the minimum inhibition concentrations (MIC) of streptomycin. K-12Δ*rluD* and K-12Δ*gidB* had an MIC of 20 and 80 mg/L, compared to 16 mg/L for K-12, confirming that the genes could play a role in resistance. We then measured the MIC of streptomycin of strain 536 and the 15 lineages. The mutation K87R responsible for the resistance to streptomycin has been identified previously by others and is responsible

for high resistance to streptomycin (Pelchovich *et al.* 2013). Indeed, MIC was over 150 g/L (more than 30 times higher than the concentration used in the experiment in the water of mice) and fairly similar across evolved lineages and the ancestor. We decided not to go further with MIC as such an incredibly high level of resistance prevents any accurate and reproducible measurement.

To go further, we then inactivated both genes independently or in combination in strain 536 and performed growth curves in LB supplemented or not with streptomycin at 5 g/L. The underlying idea was that the mutations could compensate the cost of resistance to streptomycin that has been shown to be costly. Wild type strain 536 and 536Δ*gidB* grew faster than 536Δ*rluD* and 536Δ*rluD*Δ*gidB* ($p = 0.001$) in LB rejecting the idea that the mutations were compensating a general cost of resistance. Yet, the mutant strains grew faster than the wild type strain 536 in LB with streptomycin (figure 6), revealing that these mutations could readily be selected for under the antibiotic selective pressure used. *rluD* mutants were dominant over *gidB* mutations in the sense that *rluD* was sufficient to promote the reducing of the DT; adding *gidB* to it did not bring any further improvement. Hence, once *rluD* is inactivated, we do not expect to find any further mutations in *gidB*, however, under the conditions we tested, if *gidB* is mutated first, *rluD* could still be selected for. Hence our experimental test in LB could only partially explain the observed full repulsion between the two mutated genes.

Discussion

Experimental evolution coupled to whole genome sequencing is a powerful tool to study bacterial adaptation (Long *et al.* 2015). Over the last decades, many *E. coli* populations have been adapted to different media using that approach. Genomics was then used to uncover the molecular determinants of these adaptations and to follow how reproducible adaptation is. Most studies have been done *in vitro* and have shown some similarities. First, the response to selection is proportional to the initial maladaptation: clones with initial low fitness tend to improve faster (Couce & Tenaillon 2015). Second, intermediate convergence is observed among replicate lineages. Moreover this convergence is involving different mutations that affect the same genes or functional units. Though integration at the gene level seems the most relevant to study convergence, mutations in different genes involved in the same function can also be identified as a source of convergence (Tenaillon *et al.* 2012). Third, some epistatic interactions may be found between mutations within genes or among genes (Tenaillon *et al.* 2012). Fourth, global regulators tend to be frequently recruited during the first steps of adaptations (Hindré *et al.* 2012). As most studies, especially long-term ones have been done *in vitro* or *in vivo* with laboratory strains, the relevance of these observations remained to be questioned. We therefore wanted to know what fraction of these results would still hold if we were to adapt a natural isolate of *E. coli* to a more natural environment, the mouse gut. We therefore performed a 363-day assay of colonization of the digestive tract of 30 streptomycin treated mice in 15 independent cages with the B2 strain *E. coli* 536, and sequenced one evolved clone from each cage. Assuming 18 generations per day for *E. coli* in streptomycin treated mice (Poulsen *et al.* 1995), this represents about 6,534 generations.

Similarities and differences with *in vitro* lessons of experimental evolution

We found some similarities and differences with the previous experiments done.

- We recovered an important signal of convergence between lineages at the mutation, gene/operon and function levels. Convergence at the genome level was highly significant; it was variable among lineages. For instance, 3 out of 10 affected genes of in lineage 9 were targeted in other lineages, while the 6 genes affected in lineage 6 were at least found in another lineage. On average, 56% of the mutations affected genes that were also found mutated in at least another lineage. For instance, lineages 5 and 13 shared mutations in *dgoR*, *dgoD*, *dgoT*, *rluD* and in between *yjiH-kptA*. They have therefore 5 common targets (hit by different mutations for 3 out of 5) out of 9 or 7 affected genes respectively. They have hence a more than 50% convergence at the gene level.

-As found *in vitro*, this convergence was based on different mutations in the affected genes, and could be integrated into higher functional levels: *rluD-gidB* modifications of the ribosomal RNA for instance.

-Also some interactions among mutations within genes or among genes could be found. For instance, though 11 mutations were found in *dgoR*, no lineage was found with a double mutation in *dgoR* ($p<0.004$), which suggest that once the gene is inactivated, no further mutations in the gene are selected for. Among genes, mutations in *gidB* and *rluD* were mutually exclusive, while mutations in *dgoT* only occurred in lineages having a mutation in both *dgoR* and *dgoT*. These nonrandom associations of mutations suggest an important role of epistasis in shaping the adaptive landscape.

One important difference with previous work was that no global regulator was recovered among the mutations found. *In vitro*, for instance *rpoB* is a very frequent target of adaptation (Conrad *et al.* 2010; Tenaillon *et al.* 2012). Despite a extremely high level of conservation within the species, with just 4 non-synonymous mutations in *rpoB* among 128 *E. coli* isolated genomes for a 4 Kb gene (Lescat *et al.* 2014), the mutations is one of the most frequent target of adaptation in many systems (Long *et al.* 2015)). The lack of mutations in global regulators in our system, suggests that the conditions encountered may be less artificial than *in vitro* conditions. Yet, if no global regulators are hit, we may still question the targeted functions to the mouse gut to uncover how specific they are from our experimental conditions or from our strain?

Functional targets of adaptation

One of the dominant targets of adaptation is the inactivation of *dgoR* that leads to an enhanced ability to grow on D-galactonate. Other studies performed in chemostats with specific sugar as nutrients allowed the observation of mutation in genes enhancing the increase of utilization of these sugars (Ferenci 2007; Lee & Palsson 2010). However, the specific convergence in the *dgoR* gene has never been reported so far. Particularly, it was not recovered by Barroso-Batista *et al* (Barroso-Batista *et al.* 2014) who used also streptomycin treated mice to study the evolution of the K-12 strain. Many hypotheses could be proposed. These observations could be linked to the strain studied. We used the strain *E. coli* 536 which was isolated after a urinary tract infection in a human patient (Brzuszkiewicz *et al.*

2006), whereas in other studies the strain used for the experiment of colonization was *E. coli* K-12 which is the typical laboratory strain (Blattner *et al.* 1997). One possibility is that strain 536 could harbor an abnormal genetic background relative to other strains. For instance, Barroso-Batista *et al* found that the first response *of E. coli* K-12 to the gut was mutations inactivating *gat* operon, but this operon is abnormal in *E. coli* K-12: it is constitutively expressed due to an insertion sequence integration in its repressor (Barroso-Batista *et al.* 2014). Hence the observed response could be more a specificity of the strain than a signature of the environment. When we looked at *dgoR*, in *E. coli* 536, we could not find any clear specificity. Our functional assays showed that *dgoR* does repress efficiently the operon in the presence of glucose and release the repression in the presence of D-galactonate. We performed further analysis of *dgoR* gene history using fully sequenced genome of *E. coli* isolates. We observed that the entire operon *dgoRKADT* was missing in the strains *E. coli* from the serotype O157:H7 belonging to the E group. This group is usually responsible for enterohemorragic diarrhea in patients but it is also retrieved as commensal in digestive tracts of cattle (Cray & Moon 1995). This particular observation could be linked to the specific digestive tract of cattle. We further used more than 110 *E. coli* genome sequences (Lescat et al 2014) to retrieve the mean pairwise rate of synonymous (Ks) and nonsynonymous substitutions (Ka) of *dgoR*, *dgoD* and *dgoT*. For each of these genes, the ratio Ka/Ks was much less than 1 (0.0055, 0.013 and 0.019 respectively) which comforts the idea that the operon is under strong purifying selection.

We then tried to identify the source of D-galactonate in the digestive tract of animals or human individuals, but nothing has been described about this sugar except the big amount produced in specific tissues of patients or mice with a specific genetic disease (Yager *et al.* 2004; Ficicioglu *et al.* 2005). However, this sugar can be produced from galactose using the De Ley Doudoroff pathway present in specific aerobic bacteria such as *Stenotrophomonas maltophilia* (Brechtel *et al.* 2002). Saffarian *et al.* recently identified that these bacteria especially *S. maltophilia* can be retrieved in the crypts of mice (Saffarian *et al.* 2015). Besides, we know that *S. maltophilia* can be retrieved in the mucous membranes of patients under antibiotic selective pressure especially with imipenem treatment, this bacteria being naturally resistant to this antibiotic as well as to all aminoglycosides, including streptomycin. We therefore retrospectively searched for the presence of this species in mouse feces stored at -80°C in glycerol, but failed to find it. Surprisingly, in other experiments with CD1 mice used for colonization assay, we identified the presence of this species in the feces but were not able to find if the species could produce D-galactonate (data not shown). We can also hypothesize that *dgoRKADT* operon may be used to catabolize some alternative sugars found in the digestive tract of mice, and that this sugar is unable to relieve the repression by *dgoR* and favors therefore its inactivation. Overall, from the data we have, we can say that selection for *dgoR* inactivation seem not to be a specificity of the strain we used, and must be due to the environmental conditions we evolved the strains in. However, it is hard to know if the selection for *dgoR* loss is generally selected in the mouse gut or if it is selected here as an indirect consequence of the use of streptomycin that alters the composition of the microbiota as Leatham-Jensen *et al* observed missenses in other genes of an *E. coli* K-12 strain evolved model of colonization (Leatham-Jensen *et al.* 2012). Nevertheless, these convergences illustrate the ability of *E. coli* to adapt to the

specificity of its gut environment, whether they are linked to the host, its diet or *E. coli* interactions with the gut microbiota (Maltby *et al.* 2013).

Our second target of adaptation was directly linked to the use of antibiotic in our experimental setting. Thirteen out of the 15 lineages had mutations either in *gidB* or in *rluD,* two genes involved in maturation of ribosomal RNAs and linked to the *rpsL* gene in which mutations may be responsible of resistance to streptomycin. Experiments in which bacteria are evolved in presence of antibiotics represents one of the typical cases of selection that favors the fast emergence of mutations. It is also a case of medical importance due to the drastic increase of resistance among pathogenic bacteria (Andersson & Hughes 2010; Toprak *et al.* 2011). Two types of mutations are generally selected for in such a regime: mutations that increase resistance to the antibiotic on the one hand and compensatory mutations on the other. The latter are mutations that are selected as they compensate the cost of some previously acquired costly resistance mutations (Andersson & Hughes 2010). In our case, as the concentration of antibiotic used was much lower than the MIC of the strain (5 g/L compared to an MIC of more than 150 g/L), we were therefore not expecting resistance mutations to be selected for. Indeed, in the case of streptomycin, subinhibitory concentrations higher than 1/4 of the MIC can select for resistance mutations (Gullberg *et al.* 2011), while in the present setting, strains were exposed to concentrations lower than 1/30 of the MIC. We found however that *gidB* and *rluD* inactivations were not compensatory mutations, but mutations that improved growth only in the presence of the antibiotic. Interestingly *gidB* was found to be inactivated in several streptomycin resistant laboratory strains (Jeong *et al.* 2009) suggesting that the selection of its inactivation may occur quite broadly under streptomycin selective pressure.

Finally, the pattern of convergence suggests that two other sets of genes of unknown functions are involved in adaptation to the streptomycin treated mice gut. ECP4610, a putative transcriptional regulator is affected 8 times and the intergenic region between *yjiH* and *kptA* is modified 11 times. It would be interesting to look in more details at the phenotypic effects of these mutations to uncover the specificity of their contribution to adaptation.

Rates of adaptation

Besides the targets of adaptation, our data gave also the rate at which mutations are selected for in the mice gut. In experimental evolution, fast rate of substitutions can be found, with about 11 mutations in 2,000 generations in a thermal adaptation experiment using a very maladapted strain (Tenaillon *et al.* 2012), 15 mutations after 5,000 generations in a less challenging environment (Barrick *et al.* 2009) *In vivo*, Barroso-Batista *et al.* found a mean of 2.3 mutations per lineage over 400 generations (Barroso-Batista *et al.* 2014). Here we found 95 mutations over 4,838 days of evolution, a rate 4.4, 2.4 and 4.7 times lower than the one found in the afore mentioned experiments if we assume 18 generations per day (Rang *et al.* 1999) and no mutation in the global regulators. Hence, the selective pressures may not be as strong as in other setting, reflecting that despite the antibiotic selective pressure, the strain was more adapted to its environment than the ones of the previous experiments. The polymorphism data were suggesting low selective pressure as in mice C3Ma and C2Ma

Europe PMC Funders Author Manuscripts

some polymorphic alleles shifted their frequency by less than 40% over more than 125 days of evolution (table S3) and over 4 populations sequenced after more than 125 days of evolution, a single mutation was found fixed (data not shown). The value of Ka/Ks was substantially higher than one (Ka/Ks=1.90, 95% CI 1.09-4.02) but less than the one found in other studies. For instance, a Ka/Ks close to 5 was found in thermal adaptation suggesting that 80% of non-synonymous mutations were the product of selection in that setting. Such high values of Ka/Ks are not rare, for instance in the long-term experimental evolution, the population sequenced (Barrick *et al.* 2009) had no synonymous mutations fixed after 20,000 generations. The lower value found here is coherent with a much weaker selective pressure in the mouse gut compared to *in vitro* evolution.

Overall, our data show that an “evolve and sequence” approach (Long et al 2015) coupled to an analysis of convergence can be used to uncover the molecular bases of short-term adaptation in an *in vivo* study. With that approach, we identified several genes as targets of selection in the gut of streptomycin treated mice and showed that adaptation could be caught in the act in a somehow natural setting with a natural isolate. This suggests that short term adaptation may contribute to some of the patterns of prevalence we observed in the wild (Tenaillon et al 2010). While some of the dominant targets, *gidB*/*rluD* and *dgo* operon do not seem to be specific of the strain we used, others such as ECP4610 and mutations between *yjiH* and *kptA* are more enigmatic and could be specific of the strain, especially ECP4610 that is only found in few strains. Further studies with various strains will be required to investigate the fraction of that adaptation that is dependent on the genetic background of the host. Finally, the presence of genetic responses linked to the antibiotic used to maintain the strain, call for the development of some colonization assays without such a selective pressure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding information

This work was supported by the European Research Council under the European Union’s Seventh Framework Program (FP7/2007-2013)/ERC Grant 310944. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. MG was supported by the grant FDM20150633803 obtained from the « Fondation pour la Recherche Médicale ».

Bibliographic references

Andersson DI, Hughes D. Antibiotic resistance and its cost: is it possible to reverse resistance? Nature Reviews Microbiology. 2010; 8:260–271. [PubMed: 20208551]

Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. Nature. 2011; 473:174–180. [PubMed: 21508958]

Baba T, Ara T, Hasegawa M, et al. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular systems biology. 2006; 2 2006.0008.

Barrick JE, Yu DS, Yoon SH, et al. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature. 2009; 461:1243–1247. [PubMed: 19838166]

Barroso-Batista J, Sousa A, Lourenço M, et al. The First Steps of Adaptation of *Escherichia coli* to the Gut Are Dominated by Soft Sweeps. PLoS Genetics. 2014; 10(3):4182. 1371/jourbal.pgen.1004182.

Berger H, Hacker J, Juarez A, Hughes C, Goebel W. Cloning of the chromosomal determinants encoding hemolysin production and mannose-resistant hemagglutination in *Escherichia coli*. Journal of Bacteriology. 1982; 152:1241–1247. [PubMed: 6754701]

Bingen E, Picard B, Brahimi N, et al. Phylogenetic Analysis of *Escherichia coli* Strains Causing Neonatal Meningitis Suggests Horizontal Gene Transfer from a Predominant Pool of Highly Virulent B2 Group Strains. Journal of Infectious Diseases. 1998; 177:642–650. [PubMed: 9498443]

Blattner FR, Plunkett G 3rd, Bloch CA, et al. The complete genome sequence of *Escherichia coli* K-12. Science (New York, N.Y.). 1997; 277:1453–1462.

Bleibtreu A, Gros P-A, Laouénan C, et al. Fitness, Stress Resistance, and Extraintestinal Virulence in *Escherichia coli*. Infection and immunity. 2013; 81:2733–2742. [PubMed: 23690401]

Brechtel E, Huwig A, Giffhorn F. l-Glucitol Catabolism in *Stenotrophomonas maltophilia* Ac. Applied and Environmental Microbiology. 2002; 68:582–587. [PubMed: 11823194]

Brzuszkiewicz E, Brüggemann H, Liesegang H, et al. How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103:12879–12884. [PubMed: 16912116]

Conrad TM, Frazier M, Joyce AR, et al. RNA polymerase mutants found through adaptive evolution reprogram *Escherichia coli* for optimal growth in minimal media. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107:20500–20505. [PubMed: 21057108]

Couce A, Tenaillon OA. The rule of declining adaptability in microbial evolution experiments. Frontiers in Genetics. 2015; 6(99)doi: 10.3389/fgene.2015.00099

Cray WC, Moon HW. Experimental infection of calves and adult cattle with *Escherichia coli* O157:H7. Applied and Environmental Microbiology. 1995; 61:1586–1590. [PubMed: 7747972]

Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proceedings of the National Academy of Sciences of the United States of America. 2000; 97:6640–6645. [PubMed: 10829079]

Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods in Molecular Biology (Clifton, N.J.). 2014; 1151:165–188.

Diard M, Garry L, Selva M, et al. Pathogenicity-associated islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal colonization. Journal of Bacteriology. 2010; 192:4885–4893. [PubMed: 20656906]

Ferenci, T. Bacterial Physiology, Regulation and Mutational Adaptation in a Chemostat Environment. Advances in Microbial Physiology. Poole, RK., editor. Academic Press; 2007. p. 169-315.

Ficicioglu C, Yager C, Segal S. Galactitol and galactonate in red blood cells of children with the Duarte/galactosemia genotype. Molecular Genetics and Metabolism. 2005; 84:152–159. [PubMed: 15670721]

Flint A, Butcher J, Stintzi A. Stress Responses, Adaptation, and Virulence of Bacterial Pathogens During Host Gastrointestinal Colonization. Microbiology Spectrum. 2016; 4(2)doi: 10.1128/microbiolspec.VMBF-0007-2015

Giraud A, Arous S, Paepe MD, et al. Dissecting the Genetic Components of Adaptation of *Escherichia coli* to the Mouse Gut. PLoS Genet. 2008; 4:e2. [PubMed: 18193944]

Giraud A, Matic I, Tenaillon O, et al. Costs and Benefits of High Mutation Rates: Adaptive Evolution of Bacteria in the Mouse Gut. Science. 2001; 291:2606–2608. [PubMed: 11283373]

Gordon DM, Cowling A. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. Microbiology (Reading, England). 2003; 149:3575–3586.

Gullberg E, Cao S, Berg OG, et al. Selection of Resistant Bacteria at Very Low Antibiotic Concentrations. PLoS Pathogens. 2011; 7(7):e1002158. Epub 2011 Jul 21. doi: 10.1371/journal.ppat.1002158 [PubMed: 21811410]

Herring CD, Raghunathan A, Honisch C, et al. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. Nature Genetics. 2006; 38:1406–1412. [PubMed: 17086184]

Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. PLoS genetics. 2010; 6:e1001115. [PubMed: 20838599]

Hindré T, Knibbe C, Beslon G, Schneider D. New insights into bacterial adaptation through *in vivo* and *in silico* experimental evolution. Nature Reviews Microbiology. 2012; 10:352–365. [PubMed: 22450379]

Jaureguy F, Landraud L, Passet V, et al. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. BMC Genomics. 2008; 9:560. [PubMed: 19036134]

Jeong H, Barbe V, Lee CH, et al. Genome sequences of *Escherichia coli* strains REL606 and BL21(DE3). Journal of Molecular Biology. 2009; 394:644–652. [PubMed: 19786035]

Khoury KA, Floch MH, Hersh T. Small intestinal mucosal cell proliferation and bacterial flora in the conventionalization of the germfree mouse. The Journal of Experimental Medicine. 1969; 130:659–670. [PubMed: 4896909]

Lamble HJ, Milburn CC, Taylor GL, Hough DW, Danson MJ. Gluconate dehydratase from the promiscuous Entner–Doudoroff pathway in *Sulfolobus solfataricus*. FEBS Letters. 2004; 576:133–136. [PubMed: 15474024]

Leatham-Jensen MP, Frimodt-Møller J, Adediran J, et al. The Streptomycin-Treated Mouse Intestine Selects *Escherichia coli envZ* Missense Mutants That Interact with Dense and Diverse Intestinal Microbiota. Infection and Immunity. 2012; 80:1716–1727. [PubMed: 22392928]

Lee D-H, Palsson BØ. Adaptive Evolution of *Escherichia coli* K-12 MG1655 during Growth on a Nonnative Carbon Source, l-1,2-Propanediol. Applied and Environmental Microbiology. 2010; 76:4158–4168. [PubMed: 20435762]

Le Gac M, Cooper TF, Cruveiller S, Médigue C, Schneider D. Evolutionary history and genetic parallelism affect correlated responses to evolution. Molecular Ecology. 2013; 22:3292–3303. [PubMed: 24624420]

Lescat M, Clermont O, Woerther PL, et al. Commensal *Escherichia coli* strains in Guiana reveal a high genetic diversity with host-dependant population structure. Environmental microbiology reports. 2013; 5:49–57. [PubMed: 23757130]

Lescat M, Reibel F, Pintard C, et al. The Conserved nhaAR Operon Is Drastically Divergent between B2 and Non-B2 *Escherichia coli* and Is Involved in Extra-Intestinal Virulence. PLoS ONE. 2014; 9(9):e108738. eCollection 2014. doi: 10.1371/journal.pone.0108738 [PubMed: 25268639]

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 Pathogens. 2010; 6(9):e1001125.doi: 10.1371/journal.ppat.1001125 [PubMed: 20941353]

Long A, Liti G, Luptak A, Tenaillon O. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. Nature Reviews Genetics. 2015; 16:567–582.

Maddamsetti R, Lenski RE, Barrick JE. Adaptation, Clonal Interference, and Frequency-Dependent Interactions in a Long-Term Evolution Experiment with *Escherichia coli*. Genetics. 2015; 200:619–631. [PubMed: 25911659]

Maltby R, Leatham-Jensen MP, Gibson T, Cohen PS, Conway T. Nutritional Basis for Colonization Resistance by Human Commensal *Escherichia coli* Strains HS and Nissle 1917 against E. coli O157:H7 in the Mouse Intestine. PLoS ONE. 2013; 8:e53957. [PubMed: 23349773]

Okamoto S, Tamaru A, Nakajima C, et al. Loss of a conserved 7-methylguanosine modification in 16S rRNA confers low-level streptomycin resistance in bacteria. Molecular Microbiology. 2007; 63:1096–1106. [PubMed: 17238915]

Pelchovich G, Schreiber R, Zhuravlev A, Gophna U. The contribution of common rpsL mutations in *Escherichia coli* to sensitivity to ribosome targeting antibiotics. International journal of medical microbiology: IJMM. 2013; 303:558–562. [PubMed: 23972615]

Picard B, Garcia JS, Gouriou S, et al. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. Infection and Immunity. 1999; 67:546–553. [PubMed: 9916057]

Poulsen LK, Licht TR, Rang C, Krogfelt KA, Molin S. Physiological state of *Escherichia coli* BJ4 growing in the large intestines of streptomycin-treated mice. Journal of Bacteriology. 1995; 177:5840–5845. [PubMed: 7592332]

Rang CU, Licht TR, Midtvedt T, et al. Estimation of Growth Rates of *Escherichia coli* BJ4 in Streptomycin-Treated and Previously Germfree Mice by In Situ rRNA Hybridization. Clinical and Diagnostic Laboratory Immunology. 1999; 6:434–436. [PubMed: 10225851]

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

Saffarian A, Mulet C, Naito T, et al. Draft Genome Sequences of *Acinetobacter parvus* CM11, *Acinetobacter radioresistens* CM38, and *Stenotrophomonas maltophilia* BR12, Isolated from Murine Proximal Colonic Tissue. Genome Announcements. 2015; 3(5) pii: e0108915. doi: 10.1128/genomeA.01089-15

Sato T, Iino T. Genetic analyses of the antibiotic resistance of *Bifidobacterium bifidum* strain Yakult YIT 4007. International Journal of Food Microbiology. 2010; 137:254–258. [PubMed: 20051305]

Tenaillon O, Barrick JE, Ribeck N, et al. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature. 2016; 536(7615):165–70. [PubMed: 27479321]

Tenaillon O, Rodríguez-Verdugo A, Gaut RL, et al. The molecular diversity of adaptive convergence. Science (New York, N.Y.). 2012; 335:457–461.

Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*. Nature Reviews Microbiology. 2010; 8:207–217. [PubMed: 20157339]

Toprak E, Veres A, Michel J-B, et al. Evolutionary paths to antibiotic resistance under dynamically sustained drug stress. Nature genetics. 2011; 44:101–105. [PubMed: 22179135]

Touchon M, Hoede C, Tenaillon O, et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genetics. 2009; 5:e1000344. [PubMed: 19165319]

Umekasi Y. Use of gnotobiotic mice to identify and characterize key microbes responsible for the development of the intestinal immune system. Proceedings of the Japan Academy Series B, Physical and Biological Sciences. 2014; 90:313–332.

Welling GW, Groen G, Tuinte JH, Koopman JP, Kennis HM. Biochemical effects on germ-free mice of association with several strains of anaerobic bacteria. Journal of General Microbiology. 1980; 117:57–63. [PubMed: 7391821]

Wielgoss S, Barrick JE, Tenaillon O, et al. Mutation Rate Inferred From Synonymous Substitutions in a Long-Term Evolution Experiment With *Escherichia coli*. G3 (Bethesda, Md.). 2011; 1:183–186.

Wong A, Rodrigue N, Kassen R. Genomics of adaptation during experimental evolution of the opportunistic pathogen *Pseudomonas aeruginosa*. PLoS genetics. 2012; 8:e1002928. [PubMed: 23028345]

Yager C, Ning C, Reynolds R, Leslie N, Segal S. Galactitol and galactonate accumulation in heart and skeletal muscle of mice with deficiency of galactose-1-phosphate uridyltransferase. Molecular Genetics and Metabolism. 2004; 81:105–111. [PubMed: 14741191]

Zaslaver A, Bren A, Ronen M, et al. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. Nature methods. 2006; 3:623–628. [PubMed: 16862137]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

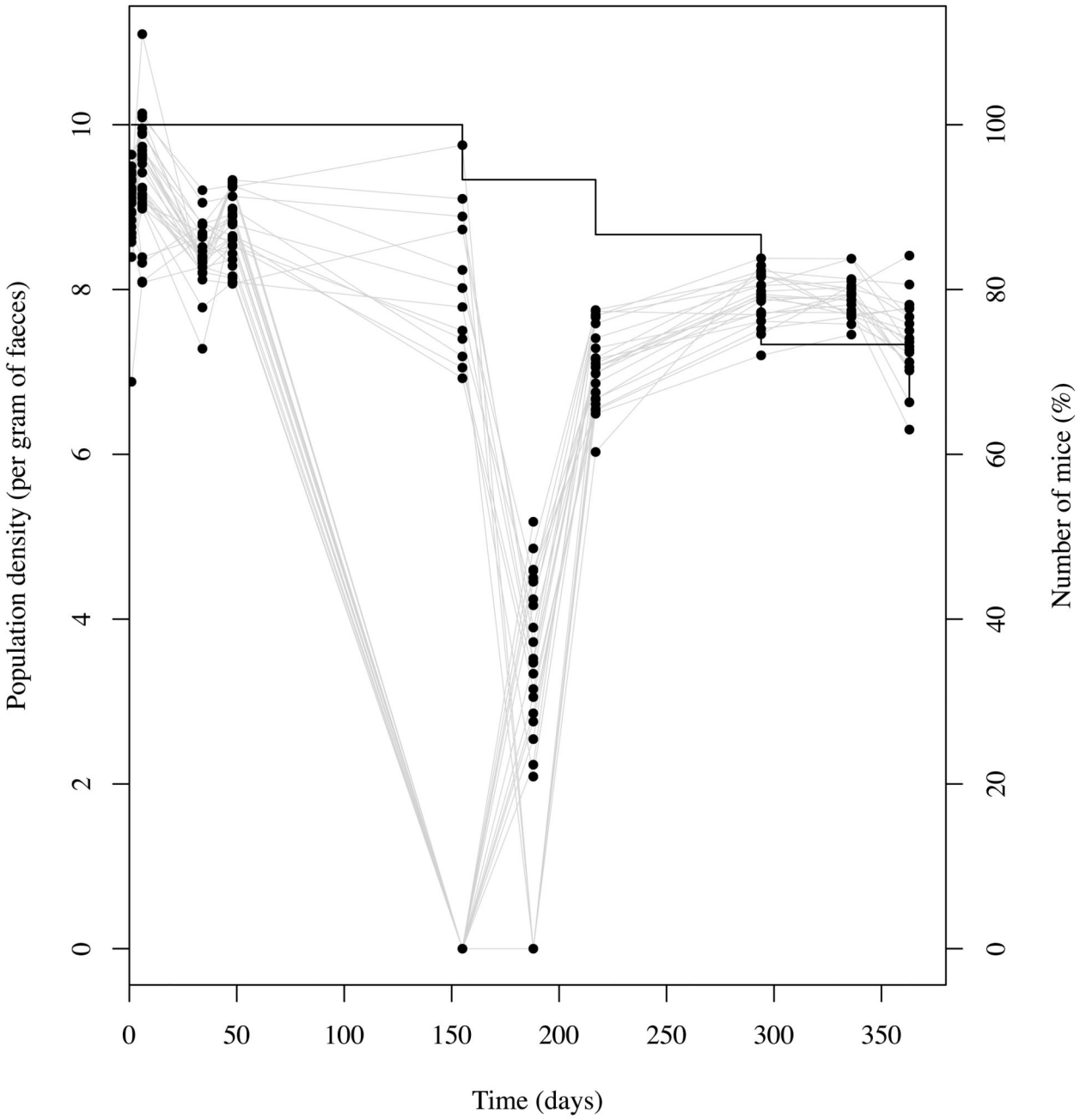


Figure 1. Population density of *Escherichia coli* 536 per gram of feces of the streptomycin treated mice in 15 independent cages after oral colonization of the strain during 363 days.

The population density is represented in logarithmic scale. All density populations from the same mice at the different points in the time are relied by a grey curve. The percentage of alive mice is also represented with the black curve.

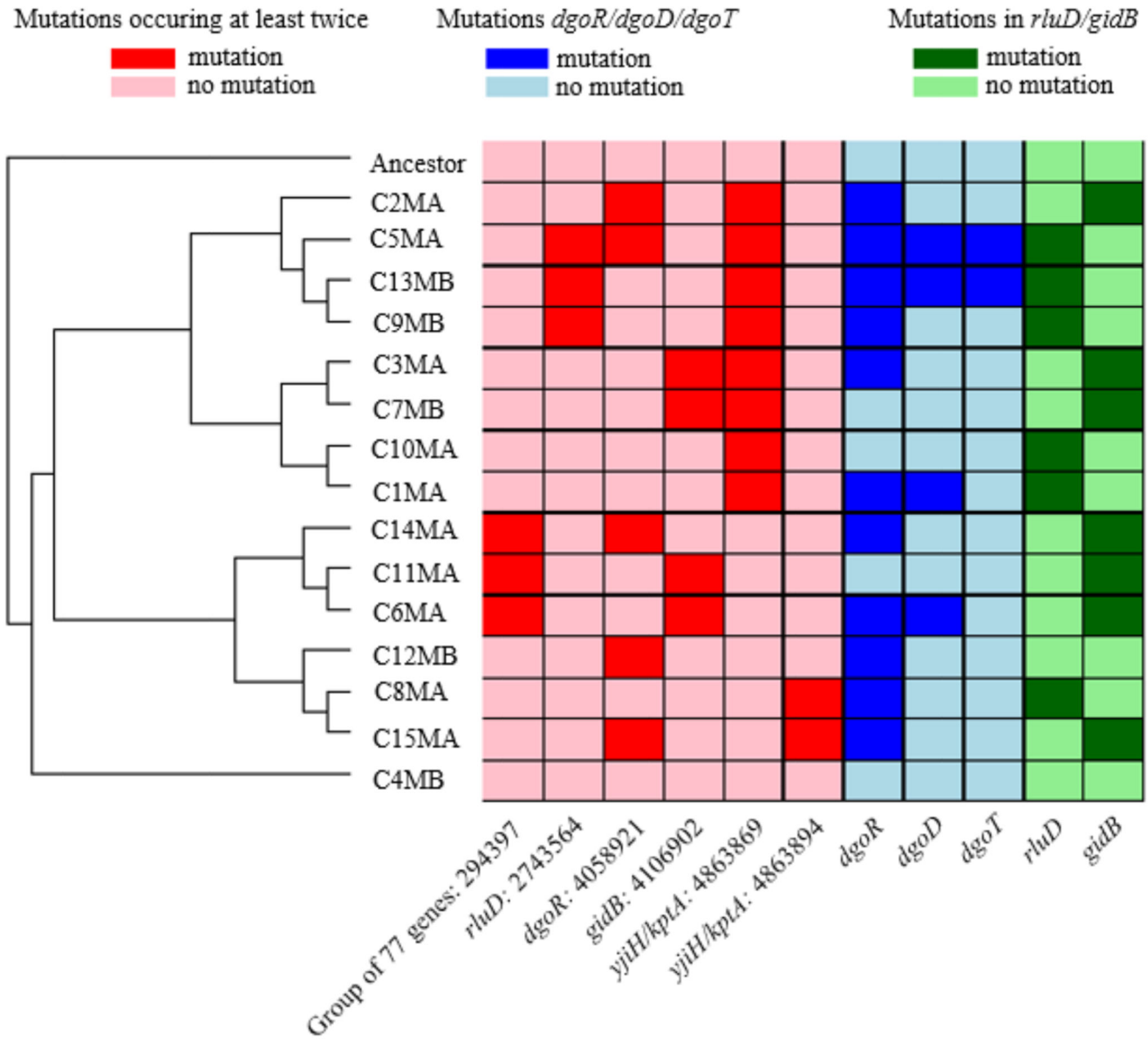


Figure 2. Mutation pattern among the evolved lineages.

The dendrogram of the evolved lineages (on the left) was computed using parsimony method and the presence or absence of mutations among the lineages. The convergence at the mutation level is showed in red (light red absence of mutation, bright red: presence of mutation) and illustrates similar mutations that appeared in different lineages with the name of the gene or regions it falls into, and the position on the *Escherichia coli* 536 genome. The specificity at the gene level is shown in blue (light blue: absence of mutations, bright blue: presence of mutation) for the *dgoR*, *dgoT*, and *dgoD* genes, and in green (light green: absence of mutations, bright green: presence of mutation) for the *gidB* and *rluD* genes illustrating mutations falling in the same genes, but with possible differences.

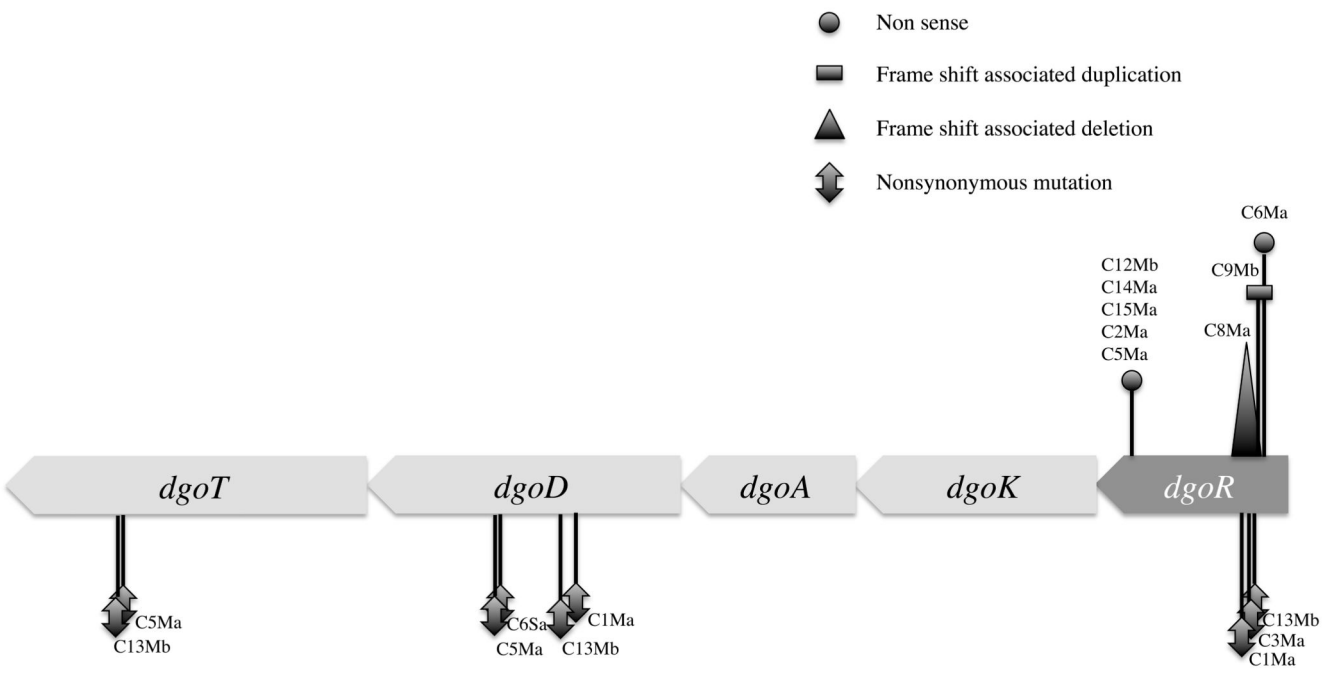


Figure 3. Diagrammatic representation of the *dgoRKADT* operon of the 15 evolved lineages compared to the *Escherichia coli* 536 ancestor strain.

The mutations are shown at scale all along the operon. Each evolved lineage mutation is represented with a symbol defined in the legend.

Europe PMC Funders Author Manuscripts

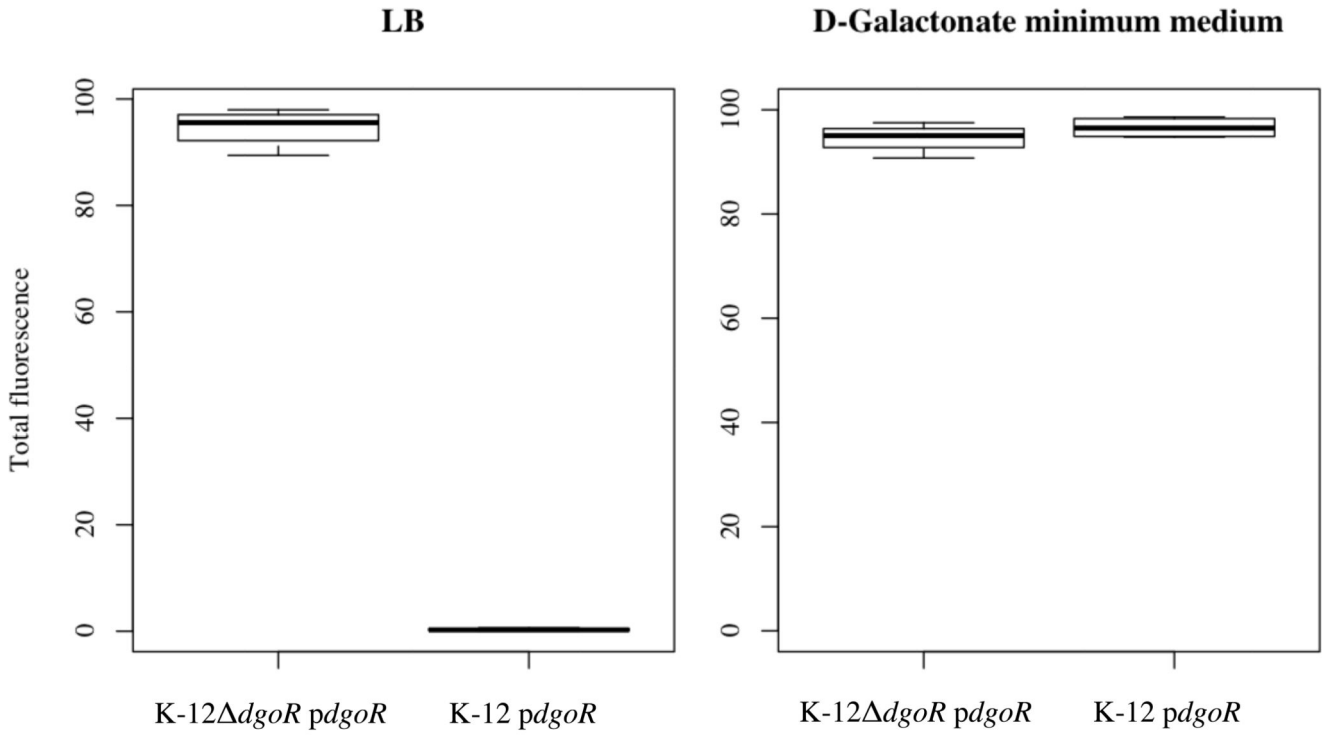


Figure 4. Comparison of total fluorescence of a reporter plasmid of the promoter activity of the *dgoRKADT* operon in *Escherichia coli* K-12 derived strains.

Boxplots of the total fluorescence of the strain K-12 p*dgoR* and the mutated strain K-12Δ*dgoR* p*dgoR* have been compared in LB and in minimum medium D-galactonate.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

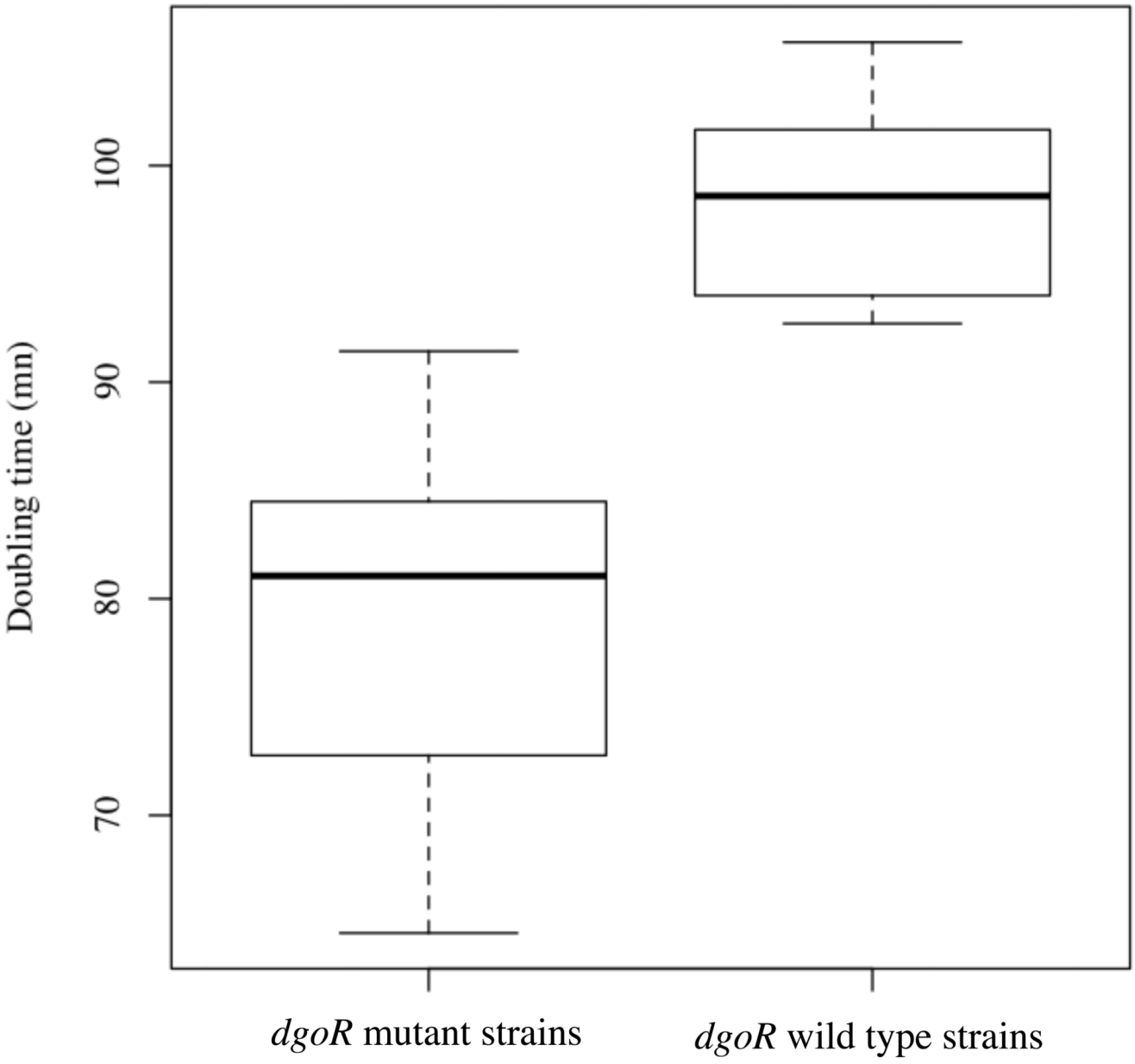


Figure 5. Mutated strains of *Escherichia coli* 536 in *dgoR* grew faster than wild type strains in minimum medium D-galactonate.

Boxplots of the doubling times in minutes of 11 mutated strains of *E. coli* 536 in *dgoR* and 5 non mutated strains in *dgoR* with the ancestor, in minimum medium D-galactonate at pH 5.7. We found a significant difference between the *dgoR* mutated strains and the wild type strains using a Welch test (p =0.001).

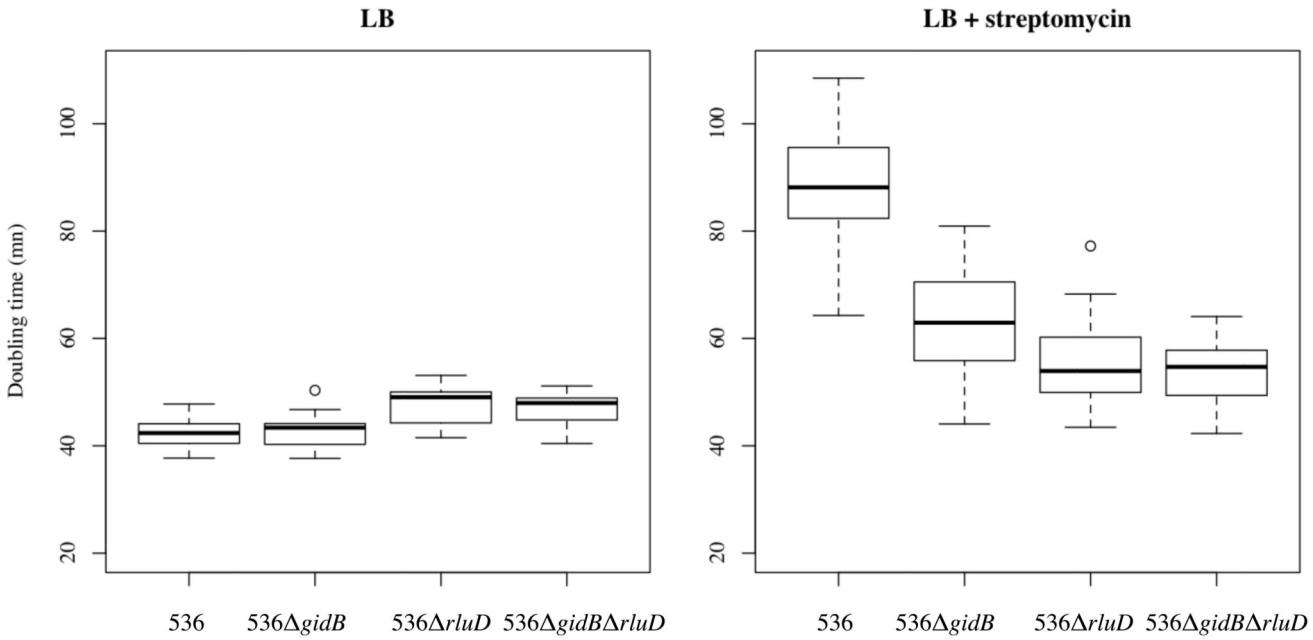


Figure 6. Inactivation of *rluD* and *gidB* genes confers an advantage to the *Escherichia coli* strains in LB with streptomycin.

Boxplots of the doubling times in minutes of 536, 536Δ*gidB,* 536Δ*rluD* and 536Δ*rluD*Δ*gidB* in LB, and in LB with streptomycin at 5 g/L. We observed in one side that wild type strain 536 and 536Δ*gidB* grew faster than 536Δ*rluD* and 536Δ*rluD*Δ*gidB* in LB and that the mutant strains grew faster than the wild type strains 536 in LB with streptomycin ($p < 0.001$).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Summary of the mutations that were found in the evolved lineages.

The name of the genes as well as the orientation (defined by the arrow), the position along the genome, and the flanking genes in case of an intergenic mutation are furthermore displayed, with the mutations, its type, the number of occurrence, and a description of the gene product. Stop codons are indicated with the symbol *.

Gene	Position	Mutation	Mutation type	Number of occurrence of the mutation	Mouse	Gene info
rluD ←	2742937	IS100	IS insertion	1	C8Ma	23S rRNA pseudouridine synthase
	2743335	Δ151 bp	deletion	1	C10Ma	
	2743427	Δ7 bp	deletion	1	C1Ma	
	2743564	Δ6 bp	deletion	3	C5Ma ; C9Mb ; C13Mb	
dgoT ←	4055032	C→T	A380T (GCA→ACA)	1	C13Mb	D-galactonate transporter
	4055282	C→T	M296I (ATG→ATA)	1	C5Ma	
dgoD ←	4056705	G→A	P230S (CCA→TCA)	1	C5Ma	Galactonate dehydratase
	4056714	T→C	T227A (ACG→GCG)	1	C6Ma	
	4056917	G→A	A159V (GCG→GTG)	1	C13Mb	
	4057008	T→C	I129V (ATC→GTC)	1	C1Ma	
dgoR ←	4058921	G→A	Q212* (CAG→TAG)	5	C2Ma; C5Ma; C12Mb; C14Ma; C15Ma	DNA binding transcriptional regulator
	4059047	Δ360 bp	deletion	1	C8Ma	
	4059269	8 bp x 2	duplication	1	C9Mb	
	4059403	G→A	S51L (TCG→TTG)	1	C1Ma	
	4059431	G→A	R42C (CGC→TGC)	1	C3Ma	
	4059436	G→A	T40I (ACC→ATC)	1	C13Mb	
	4059478	G→T	S26* (TCG→TAG)	1	C6Ma	
gidB ←	4106867	G→A	Q167* (CAA→TAA)	1	C2Ma	S-adenosine methionine dependent methyltransferase-
	4106902	Δ25 bp	deletion	4	C3Ma; C6Ma; C7Mb; C11Ma	
	4107005	G→A	Q121* (CAG→TAG)	1	C15Ma	
	4107347	Δ80 bp	deletion	1	C14Ma	

Europe PMC Funders Author Manuscripts

Gene	Position	Mutation	Mutation type	Number of occurrence of the mutation	Mouse	Gene info
ECP_4610 ←	4816907	A→T	L185Q (CTG→CAG)	1	C3Ma	Putative transcriptional regulator
	4816998	10 bp→10 bp	substitution	1	C8Ma	
	4817034	C→T	G143S (GGC→AGC)	1	C5Ma	
	4817174	G→A	T96I (ACT→ATT)	1	C14Ma	
	4817222	G→A	A80V (GCC→GTC)	1	C1Ma	
yjiH ← / → *kptA*	4863869	C→T	intergenic (-68/-182)	8	C1Ma; C2Ma; C3Ma; C5Ma; C7Mb; C9Mb; C10Ma; C13Mb	Putative inner membrane protein/2'-phosphotransferase
	4863869	C→G	intergenic (-68/-182)	1	C6Ma	
	4863894	A→G	intergenic (-93/-157)	2	C8Ma; C15Ma	
Group of 77 genes	294397	Δ68069 bp	deletion	3	C6Ma; C11Ma; C14Ma	Prophagic region