Impact des processus de mutation et de recombinaison sur la diversité génomique au sein de l'espèce Escherichia coli

Claire Hoede

UNIVERSITE PARIS.DIDEROT (Paris 7)

ECOLE DOCTORALE : Gc2ID

DOCTORAT EN SCIENCES DE LA VIE ET DE LA SANTE

Discipline : Microbiologie

CLAIRE HOEDE

Impact des processus de mutation et de recombinaison sur la diversité génomique au sein de l'espèce *Escherichia coli.*

Mutation and recombination impact on the genomic diversity in *Escherichia coli* species

---

Soutenue le 22 septembre 2010

Jury

Pr Catherine Etchebest	Président
Dr Laurent Duret	Rapporteur
Dr Céline Brochier	Rapporteur
Dr Catherine Schouler	Examinateur
Pr Erick Denamur	Directeur
Dr Olivier Tenaillon	Co-Directeur, Encadrant

Remerciements

Je tiens à remercier en tout premier lieu, Olivier Tenaillon qui a accepté que j'effectue mon stage de DESS sous sa direction, puis qui m'a permise de travailler trois ans de plus à ses cotés. Merci de la confiance que tu m'as si rapidement accordée.

Un grand merci à toute l'équipe de l'unité 722 de l'INSERM pour m'avoir si gentiment accueillie, spécialement à tous ceux ayant fait un séjour dans le bureau du fond (vous savez celui des étudiants ....). Une pensée spéciale à Jérôme Tourret, Agnès Lefort, Mathilde Lescat, Victor Sabarly, Alix Michel, Maximes Levert et tous les autres.... Je suis vraiment très heureuse d'avoir passé ces années en votre compagnie. Il y avait alors une vraie cohésion, une vraie solidarité, une vraie émulation scientifique.

Merci à Erick Denamur d'avoir accepté de diriger ma thèse, et pour la relecture attentive que tu as fait de mon manuscrit. Merci aussi à Bertrand Picard d'avoir également participé à sa relecture. Un grand merci à Marie-Agnès Petit pour son aide quant à l'écriture de la partie concernant la recombinaison homologue.

Je tiens également à remercier Joelle Amselem de m'avoir fais confiance, de m'avoir acceptée en CDD et de m'avoir inconditionnellement encouragée dans la poursuite de ma thèse. Merci à Hadi Quesneville de m'avoir permise d'apporter ma contribution à REPET pendant deux ans. Plus généralement, merci à toute l'équipe de l'URGI pour son accueil. Je voudrais plus spécifiquement remercier l'équipe de pair, dont les membres ne se sont jamais plaints de ma fatigue du lundi matin, lorsque j'avais travaillé tout le week-end ... Parmi eux, je tiens à citer : Sandie Arnoux, Olivier Inizan, Françoise Alfama, Laetitia Brigitte, Jonathan Kreplak.

Merci à mes plus proches amies : les deux Julia et Emilie d'avoir su vous adaptez à mes rares disponibilités et d'être si fières de moi.

Merci à mes parents de m'avoir appris l'ambition, la persévérance et l'exigence.

Et enfin, merci à Sébastien de m'avoir poussée à travailler les jours où j'étais découragée.

Table des matières

Liste des figures

Liste des abbréviations

ADN : acide désoxyribonucléique

ARN : acide ribonucléique

ARNt : acide ribonucléique de transfert

BER : « base excision repair » ou réparation par excision de base

BIR : « Break-induced replication » ou réplication induite par une cassure

CGA : « clonal group A»

DAEC : « Diffusely Adherent *E. coli* » ou *E. coli* entéroadhérent

DSBR : « Double Strand Break Repair » ou réparation de coupure double brins

ECOR: « *E. coli* reference collection »

EPEC : « Enteropathogenic *E. coli* » ou *E. coli* entéropathogène

EAEC : « Enteroaggregative *E. coli* » ou *E. coli* entéroaggrégatif

EHEC : « Enterohaemorrhagic *E. coli* » ou *E. coli* entérohémorragique

EIEC : « Enteroinvasive *E. coli* » ou *E. coli* entéroinvasif

ETEC : « Enterotoxigenic *E. coli* » ou *E. coli* entérotoxinogène

ExPEC : « extraintestinal pathogenic *E. coli* » ou *E. coli* pathogène extraintestinal

GRM : « genomic resistance module »

LT : « heat-labile toxin » ou toxine thermolabile

Hfr : « high frequency of recombination »

HPI : « high pathogenicity island »

InPEC : « intraintestinal pathogenic *E. coli* » ou *E. coli* pathogène intraintestinal

ITU : Infection du tractus urinaire

PAI : « pathogenicity islands » ou îlot de pathogénicité

MCU : « major codon usage » ou usage du codon majoritaire

MLST : « multilocus sequence typing »

MLEE : « multilocus enzyme electrophoresis»

NER : « nucleotide excision repair » ou réparation par excision de nucléotides

NMEC : « New Born Meningitis *E. coli* » ou *E. coli* engendrant des méningites du nouveau-né

RAPD : « random amplified polymorphic DNA » ou polymorphisme d'amplification aléatoire

RFLP : « restriction fragment length polymorphism » ou polymorphisme de longueur de fragments de restriction

SDSA : « synthesis-dependent strand annealing » ou synthèse dépendante du brin apparié

SRM : système de réparation des mésappariements

SSA : « single strand annealing » ou hybridation simple brin

ST : « heat-stable toxin » ou toxine thermostable

UPEC : « uropathogenic *E. coli* » ou *E. coli* uropathogène

UV : ultra-violet

VIH : virus de l'immunodéficience humaine

VMP : « variable major protein »

INTRODUCTION

Les bactéries de l'espèce *Escherichia coli* sont présentes aussi bien chez l'homme que chez de nombreux animaux. Elles constituent la majeure partie de la flore microbienne commensale aéro-anaérobie du tube digestif de l'hôte. Pourtant *E. coli* est aussi une des espèces les plus fréquemment rencontrées en pathologie humaine et animale intestinale et extra-intestinale.

En plus de son intérêt physiopathologique évident, *E. coli* est une espèce modèle. C'est l'une des espèces bactériennes les plus étudiées et les plus connues (Neidhart, Curtiss et al. 1996). Actuellement, 45 souches d'*E. coli* sont entièrement séquencées et 32 sont en cours d'assemblage.

L'évolution des génomes au sein de l'espèce repose sur deux mécanismes distincts : la mutation et la recombinaison, qui génèrent une diversité génétique sur laquelle la sélection naturelle peut opérer.

Dans la partie bibliographique de ce travail, nous présenterons l'espèce *E. coli* et ses modes de vie. Je détaillerai ensuite les principaux mécanismes de mutation mis en évidence chez cette bactérie. Puis, nous nous intéresserons aux mécanismes de la recombinaison, très importante chez *E. coli* et à son impact sur la phylogénie de l'espèce.

Nous présenterons ensuite la partie expérimentale dont l'objectif a été d'étudier les processus de mutation et de recombinaison et les traces qu'ils laissent dans les génomes. Ceci nous permet à la fois d'apprécier leur importance relative et en même temps de mieux caractériser l'histoire évolutive des souches révélée par la reconstruction de la phylogénie de l'espèce.

La structuration de la population que nous mettrons en évidence a de multiples applications, nous en présenterons trois : la localisation d'un îlot de résistance, la mise en évidence d'un marqueur de phylogénie (l'estérase B) et l'identification de sous-groupes phylogénétiques comprenant des souches pathogènes humaines et animales.

PARTIE BIBLIOGRAPHIQUE

Chapitre I : *Escherichia coli* : une espèce à la fois commensale et pathogène.

E. coli est une bactérie à Gram négatif, aérobie, anaérobie facultative et non sporulée. Elle a en général une forme de bâtonnet de 2 µm de long et de 0,5 µm de large. Certaines souches sont dotées de flagelles et sont donc mobiles (Fig. 1). Son habitat primaire est le tube digestif des vertébrés, son habitat secondaire, l'eau et les sédiments. Quand *E. coli* se trouve dans son habitat primaire, ce microorganisme peut devenir pathogène.

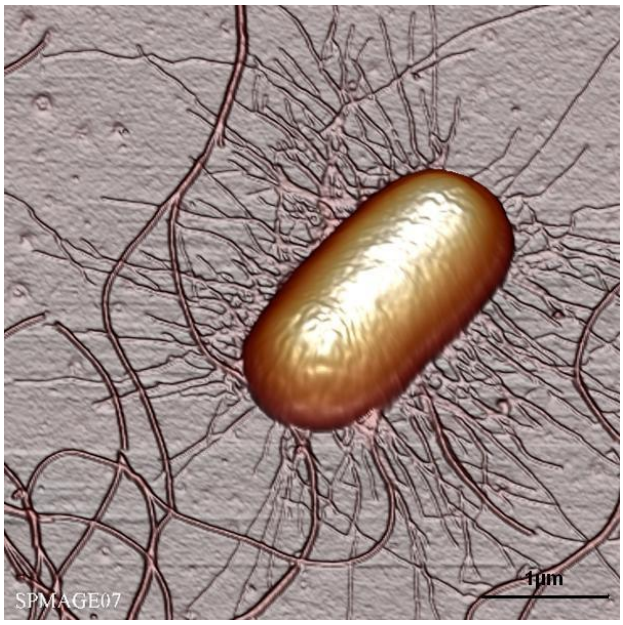

Fig. 1 : *E. coli* avec pili et flagelles. Cette image a été produite par microscopie à force atomique en mode contact intermittent en condition sèche. On peut clairement observer les pili, qui peuvent être impliqués dans la conjugaison bactérienne ou dans l'adhérence aux cellules épithéliales (dans ce cas on les appellera plutôt des fimbriae). On remarque également des flagelles, nettement plus grands, permettant à la bactérie de se déplacer. Le diamètre des pili est d'environ 20 nm, celui des flagelles de 30 nm. (SPMage Prize : http://www.icmm.csic.es/spmage/, Mr Ang Li. National University of Singapore (Singapore))

1. *E. coli* : une bactérie commensale de la flore intestinale

La principale niche écologique des souches d'*E. coli* est le mucus tapissant les cellules épithéliales du colon des vertébrés (mammifères et oiseaux). *E. coli* y est un compétiteur très performant et constitue la majeure partie de la flore microbienne aérobie facultative du tube digestif de l'hôte. Pourtant, les γ-proteobactéries dont le principal représentant est *E. coli* représentent moins de 1% de la flore intestinale totale (Berg 1996). La prévalence d'*E. coli* chez les oiseaux est de 23%, de 56% chez les mammifères, et de plus de 90% dans le cas des humains (Tenaillon, Skurnik et al. 2010).

La niche écologique secondaire d'*E. coli* est l'eau et les sédiments. Cette bactérie est souvent utilisée comme indicateur de pollution fécale de l'eau. Il a été estimé que la moitié des populations d'*E. coli* peuplait ces habitats secondaires (Savageau 1983). Certaines études récentes ont montré que certaines souches étaient capables de saprophytisme (Solo-Gabriele, Wolfert et al. 2000; Power, Littlefield-Wyer et al. 2005).

Dans notre organisme, on trouve en moyenne de 1 à 2 kg de bactéries (soit environ 10^{14} bactéries), ce qui représente dix fois plus de cellules bactériennes que de cellules humaines. Dans le colon, on observe 10^{10-11} bactéries par gramme de fèces (Berg 1996). C'est *E. coli* qui colonise le tractus gastro-intestinal des individus quelques heures après leur naissance. *E. coli* sera ensuite partiellement remplacé par de nombreuses espèces pour appartenir à un panel estimé de 800 espèces dont la grande majorité est anaérobie (Bambou, Giraud et al. 2006). Par exemple, la quantité d'*E. coli* par gramme de fèces varie chez l'homme de 10^7 à 10^9 individus. Cette flore intestinale dépend des ressources disponibles chez l'hôte, qu'il soit humain ou animal, c'est à dire indirectement de son alimentation (Slanetz and Bartley 1957; Mitsuoka and Hayakawa 1973; Penders, Thijs et al. 2006). L'implantation de la flore dépend aussi de sa capacité à s'adapter à ces ressources. En effet, les bactéries de la flore intestinale utilisent à leur profit le métabolisme de l'hôte, qui leur fournit des nutriments. Leur hôte leur procure aussi, un environnement stable, et un moyen de transport et de dissémination (Rastegar Lari, Gold et al. 1990; Vollaard and Clasener 1994; Hudault, Guignot et al. 2001; Conway, Krogfelt et al. 2004; Hudault, Spiller et al. 2004). Il apparaît maintenant clairement qu'elles peuvent aussi procurer certains bénéfices à l'hôte. Un premier effet de la flore établi de longue date est son rôle de barrière

contre les germes pathogènes. Cet effet a longtemps été expliqué par la compétition pour les ressources disponibles chez l'hôte, mais il apparaît désormais comme étant la conséquence d'une série de mécanismes complexes induits par la colonisation qui mettent en œuvre les défenses immunitaires de l'hôte (Vollaard and Clasener 1994). De plus, Hooper et collaborateurs ont montré chez la souris que la colonisation du tube digestif par *E. coli* induit l'expression de gènes épithéliaux impliqués dans l'absorption et la digestion des nutriments, la formation des jonctions serrées, la détoxification de certains métabolites et des fonctions de défense (Hooper, Wong et al. 2001). D'autre part *E. coli* synthétise et excrète des vitamines (Bentley and Meganathan 1982). En fait, la relation entre *E. coli* et son hôte tient davantage du mutualisme que du commensalisme, puisque chacun procure un bénéfice à l'autre.

E. coli est également l'une des bactéries les plus fréquemment rencontrées en pathologie humaine. Elle fait partie des 5 agents infectieux causant le plus de perte humaine dans le monde (Denamur, Picard et al. 2010), chaque année elle cause près de 2 millions de morts dues à des diarrhées ou à des infections extraintestinales (septicémies dérivées d'une infection urinaire principalement). Il existe des souches pathogènes de l'homme, mais également d'autres mammifères, ou même d'oiseaux. La virulence des souches d'*E. coli* pathogènes peut s'exercer à différents niveaux et de diverses manières.

2. *E. coli* : une bactérie pathogène

Les souches pathogènes ont été classées selon la localisation (intestinale ou extra-intestinale) des infections qu'elles produisent (Tableau I). Les souches d'*E. coli* pathogènes extra-intestinales sont nommées les ExPEC (« Extraintestinal Pathogenic *E. coli* ») (Russo and Johnson 2000). On distingue 2 pathovars au sein des ExPEC :

- Les *E. coli* uro-pathogènes : UPEC (« Uro-Pathogenic *E. coli* »*)*
- Les *E. coli* responsables de méningites, le plus souvent chez le nouveau né : NMEC (« New Born Meningitis *E. coli* »*)*

Les souches d'*E. coli* pathogènes intestinales sont nommées les InPEC (« Intestinal Pathogenic *E. coli* »). On distingue six pathovars au sein des InPEC (Nataro and Kaper 1998) :

- Les *E. coli* entéro-toxinogènes : ETEC (« Enterotoxigenic *E. coli* »)

- Les *E. coli* entéro-hémorragiques : EHEC (« Enterohaemorrhagic *E. coli* »)
- Les *E. coli* entéro-invasifs : EIEC (« Enteroinvasive *E. coli* ») et les *Shigella*
- Les *E. coli* entéro-pathogènes : EPEC (« Enteropathogenic *E. coli* »)
- Les *E. coli* à adhérence diffuse : DAEC (« Diffusely Adherent *E. coli* »)
- Les *E. coli* entéro-aggrégatifs : EAEC (« Enteroaggregative *E. coli* »)

	Pathovars	**Hôtes**	**Principales caractéristiques**
ExPEC (« Extraintestinal Pathogenic *E. coli* »)	UPEC (« Uro-Pathogenic *E. coli* »	Humains Animaux	Responsables d'infections urinaires, adhésions par les fimbriae, nécessité d'un système de captation du fer, libèrent des toxines, forment une sorte de biofilm à l'intérieur des cellules superficielles de la vessie, et se déversent dans la lumière de la vessie, souvent en adoptant une forme filamenteuse.
	NMEC (« New Born Meningitis *E. coli* »	Humains	Responsables de méningites chez les nouveaux nés, phase de multiplication dans les vaisseaux sanguins, puis, si la bactériémie requise est atteinte elles traversent la barrière hémato-méningée.
InPEC (« Intestinal Pathogenic *E. coli* »)	ETEC (« Enterotoxigenic *E. coli* »)	Humains Porcs Moutons Chèvres Bovins Chats Chevaux	Responsables de diarrhées sans fièvre pédiatriques et chez le voyageur, utilisent des adhésines fimbriales pour se lier aux entérocytes, produisent deux entérotoxines protéiques.
	EHEC (« Enterohaemorrhagic *E. coli* »)	Humains Bovins Chèvres	Responsables de diarrhées sanglantes, sans fièvre, partagent certains facteurs de virulence avec les EPEC, possèdent des facteurs de virulence supplémentaires, par exemple une entérohémolysine et des Shiga-toxines, modérément invasives.
	EIEC (« Enteroinvasive *E. coli* ») *Shigella*	Humains	Responsables d'abondantes diarrhées mêlées de sang et de mucus, forte fièvre, invasion des cellules intestinales et dissémination de bactéries de cellules en cellules, induisent l'apoptose des macrophages infectés, produisent des toxines.

	EPEC (« Enteropathogenic *E. coli* »)	Humains Lapins Chiens Chats Chevaux	Responsables de diarrhées, utilisent l'adhésine intimine pour se lier aux cellules intestinales qui forment une structure en piédestal, sont modérément invasifs.
	DAEC (« Diffusely Adherent *E. coli* »)	Humains Animaux	Agrégation diffuse sur les cellules hôtes, Certains sont responsables d'UTI et sont donc des ExPEC et d'autres provoquent des diarrhées aqueuses sans sang. Groupe hétérogène selon la nature des facteurs de virulence actifs
	EAEC (« Enteroaggregative *E. coli* »)	Humains	Responsables de diarrhées aqueuses sans fièvre, pédiatriques, chez l'adulte séropositif au VIH ou chez le voyageur. Possèdent des fimbriae spécifiques qui permettent une auto-agglomération des cellules bactériennes entre elles (biofilm dense), non invasifs.

Tableau I : Les différents pathovars. Tableau récapitulatif des différents pathovars d'*E. coli*, de leurs hôtes connus ainsi que de leurs principales caractéristiques.

2.1 Les ExPEC (« Extraintestinal Pathogenic *E. coli* »)

Les infections extra-intestinales à *E. coli* se rencontrent dans toutes les classes d'âge et peuvent affecter de nombreux organes ou sites anatomiques humains. Elles regroupent des infections du tractus urinaire (ITU) qui peuvent être des cystites (vessie) ou des pyélonéphrites (reins). Les ExPEC peuvent aussi être à l'origine de méningites (le plus souvent chez le nouveau né), de diverses infections extra-abdominales, de pneumonies (chez les patients hospitalisés), d'infections sur dispositifs intra-vasculaires, d'ostéomyélites (moelle osseuse et tissus osseux adjacents) et d'infections des tissus mous (muscles ou organes) et de septicémies (sang). Les ExPEC sont incapables de produire des infections intestinales, par contre, elles peuvent coloniser le tractus intestinal.

Les ITU sont la forme d'infection extra-intestinale à *E. coli* la plus courante, et *E. coli* est la bactérie la plus souvent responsable d'ITU. Au cours de leur vie, 12% des hommes et jusqu'à 20% des femmes contracteront une ITU (Johnson 1991).

Les méningites néo-natales sont une cause majeure de mortalité des nourrissons. Elles induisent dans près de la moitié des cas des séquelles neurologiques. *E. coli* est la seconde cause de méningites néo-natales, la première quand il s'agit de prématurés.

L'incidence des méningites à *E. coli* dans les pays industrialisés est de 0.1‰ (Bonacorsi and Bingen 2005). Les méningites néo-natales à *E. coli* présentent un taux de létalité de 14% (Houdouin, Bonacorsi et al. 2008). La plupart des cas de méningites à *E. coli* concernent des nourrissons de moins de 1 mois. Mais il arrive exceptionnellement, qu'elles touchent des adultes dans un contexte neurochirurgical ou traumatique.

Actuellement, les souches impliquées dans les infections urinaires (UPEC) et dans les méningites néo-natales (NMEC) sont les pathogènes ExPEC les mieux caractérisés.

2.1.1 Les UPEC (« Uro-Pathogenic E. coli *»)*

Il existe des souches uro-pathogènes responsables d'infections urinaires chez l'homme mais aussi chez les animaux.

L'adhésion des UPEC aux cellules hôtes se fait par les fimbriae (P, type 1 ou S). La désignation « P » correspond aux pyélonéphrites, dans ce cas, l'adhésine PapG à l'extrémité du pilus reconnait et se lie à l'épithélium rénal (Anderson, Martin et al. 2004). Les pili de type 1 sont indispensables lors de l'initiation d'une infection de la vessie. Comme les pili P, la partie distale des pili de type 1 contient une adhésine : FimH. Les pili S sont associés aux pyélonéphrites et aux cystites. Cette fois c'est l'adhésine SfaS qui permet la liaison aux cellules épithéliales. Un système de captation du fer leur est nécessaire pour coloniser l'appareil urinaire, en effet, ce nutriment indispensable au métabolisme de la cellule est présent en trop faible concentration dans le milieu extra-cellulaire. Ces bactéries libèrent des toxines (Johnson 1991), par exemple l'α-hemolysine qui lyse les érythrocytes, mais qui a également un rôle dans l'inflammation, la détérioration des tissus et du potentiel de défense de l'hôte. Elles forment une sorte de biofilm à l'intérieur des cellules superficielles de la vessie (Anderson, Martin et al. 2004; Rosen, Hooton et al. 2007). Elles sortent ensuite des cellules et se déversent dans la lumière de la vessie, souvent en adoptant une forme filamenteuse qui les rend plus résistantes aux polynucléaires neutrophiles.

2.1.2 Les NMEC (« New Born Meningitis E. coli *»)*

Les souches responsables de méningites ont, jusqu'à présent, été isolées uniquement chez l'homme.

Il a été montré chez le rat nouveau-né, que pour atteindre le sang, ces bactéries passent le plus souvent par la barrière intestinale, mais elles peuvent provenir également du tractus urinaire (20% des cas) ou d'une contamination maternelle *in utero* (Bonacorsi and Bingen 2005). Puis pour induire une méningite, elles doivent survivre dans les vaisseaux sanguins et s'y multiplier. Les facteurs de virulence impliqués dans cette phase de multiplication dans les vaisseaux sanguins seraient l'antigène de capsule K1 et la salmocheline IroN. Il faut ensuite qu'elles traversent la barrière hémato-méningée, pour cela une bactériémie importante est requise. Cette phase impliquerait l'antigène de capsule K1, l'adhésine S, l'invasine IbeA et la cytotoxine Cnf1, mais le mécanisme exact d'invasion n'est pas encore élucidé (Bonacorsi and Bingen 2005).

2.2 Les InPEC (« Intestinal Pathogenic *E. coli* »)

On distingue six pathovars InPEC répartis en trois groupes selon le type de processus physiopathologique dont ils sont responsables :

- L'adhérence aux cellules épithéliales intestinales (EPEC, DAEC, EAEC)
- La production de toxines (ETEC, EHEC)
- L'invasion des cellules épithéliales intestinales (*Shigella*, EIEC qui sont les seuls pathogènes obligatoires)

2.2.1 Les EPEC (« Enteropathogenic E. coli *»)*

Ce pathotype est responsable de diarrhées chez l'homme, plus particulièrement chez le très jeune enfant, mais aussi chez le lapin, le chien, le chat et le cheval.

Les souches EPEC ne possèdent pas de fimbriae ni de toxines, mais utilisent l'adhésine intimine (codée par le gène *eae*) pour se lier aux cellules intestinales. L'adhérence à la muqueuse intestinale déclenche un réarrangement d'actine dans la cellule hôte, ce qui la déforme significativement (formant une structure en piédestal (Nataro and Kaper 1998)). Les changements de structure des microvillosités des cellules intestinales par « attachement et effacement » entrainent une malabsorption et sont probablement la première cause de diarrhées chez les personnes infectées par ces bactéries (Nataro and Kaper 1998). Après la liaison de la bactérie à la cellule eucaryote, une cascade de phosphorylation est déclenchée ayant pour conséquences une altération du transport des ions ainsi qu'une augmentation de

la perméabilité des jonctions serrées, ce qui peut constituer une autre cause de diarrhées. Les EPEC provoquent une réponse inflammatoire (caractérisée par une migration des granulocytes) malgré le fait qu'ils ne soient que modérément invasifs (Nataro and Kaper 1998). Les EPEC ont en effet été observés à l'intérieur de certaines cellules épithéliales, mais par contre ils ne s'y multiplient pas.

2.2.2 Les EAEC (« Enteroaggregative E. coli *»)*

Les EAEC ont été mis en évidence uniquement chez l'homme.

Bien que les EAEC soient le plus souvent associés à des diarrhées pédiatriques dans les pays en voie de développement, ils sont aussi mis en évidence lors de diarrhées chez l'adulte séropositif au VIH (virus de l'immunodéficience humaine) ou chez le voyageur dans les pays industrialisés.

Ils sont nommés ainsi car ils possèdent des fimbriae spécifiques qui permettent une auto-agglomération des cellules bactériennes entre elles. Les EAEC se lient à la muqueuse intestinale, forment un biofilm dense et provoquent des diarrhées aqueuses sans fièvre (Nataro and Kaper 1998). Ils sont non invasifs, et certains produisent une hémolysine et une entérotoxine ST (thermostable) semblable à celle des ETEC. De plus, un gène *aggR* a été identifié comme régulant de nombreux facteurs de virulence chez les souches de ce pathotype. Par exemple, il régule certains gènes plasmidiques contribuant à la biogénèse des fimbriae et d'autres situés sur un îlot chromosomique (Harrington, Dudley et al. 2006).

2.2.3 Les DAEC (« Diffusely Adherent E. coli *»)*

Les DAEC peuvent être rencontrés chez l'homme mais également chez certains animaux.

Les DAEC se reconnaissent par une agrégation diffuse sur les cellules contrairement aux EAEC qui présentent une agrégation dense. Certains sont responsables d'UTI et sont donc des ExPEC et d'autres provoquent des diarrhées aqueuses sans sang. Nous ne présenterons ici que ces derniers.

Leur interaction avec les cellules épithéliales qui se fait également par l'intermédiaire de fimbriae, active différentes cascades de transduction de signaux cellulaires qui conduisent à une altération des enzymes de la bordure en brosse.

Les DAEC forment un groupe hétérogène selon la nature des facteurs de virulence actifs (Servin 2005). La première catégorie de DAEC utilise les adhésines Afa/Dr pour se lier aux cellules épithéliales humaines. La seconde catégorie exprime une adhésine impliquée dans l'adhérence diffuse : AIDA-I. Elle semble causer des diarrhées chez l'enfant.

2.2.4 Les ETEC (« Enterotoxigenic E. coli *»)*

Les ETEC sont responsables de diarrhées (sans fièvre) chez l'homme, le porc, le mouton, la chèvre, les bovins, le chat et le cheval.

Les ETEC utilisent des adhésines fimbriales pour se lier aux entérocytes de l'intestin grêle. Elles produisent deux entérotoxines protéiques :

- la plus grande, l'entérotoxine LT (thermolabile), est semblable à la toxine cholérique structuralement et fonctionnellement (Nataro and Kaper 1998). Elle initialise une cascade d'activation (passant par une augmentation de la quantité de d'AMPc intracellulaire) menant à la phosphorylation des canaux chlorure et donc à une sécrétion des ions Cl^- . Ceci entraine une diarrhée osmotique par appel d'eau dans la lumière intestinale.
- la plus petite, l'entérotoxine ST existe sous deux formes qui diffèrent dans leur structure et leur mécanisme d'action (Nataro and Kaper 1998). La STa cause l'accumulation de GMPc dans la cellule cible, ce qui stimule la sécrétion des ions chlorure et inhibe l'absorption des ions sodium. Comme précédemment, il y aura par conséquent appel d'eau dans la lumière intestinale. La STb cause des dommages aux cellules de l'épithélium intestinal en entraînant une perte de villosité de la membrane. Elle stimule également la sécrétion de bicarbonates.

Les ETEC sont la première cause de diarrhées chez l'enfant dans les pays en développement ainsi qu'une cause courante de diarrhée du voyageur (Turista). Chaque année, les ETEC causent près de 200 millions de cas de diarrhées et 170000 morts, pour la plupart des enfants dans les pays en développement (Niyogi 2005).

2.2.5 Les EHEC (« Enterohaemorrhagic E. coli »*)*

Les EHEC ont été observés chez l'homme, les bovins et les chèvres.

Ce pathotype provoque des diarrhées sanglantes, sans fièvre. Les EHEC peuvent provoquer le syndrome hémolytique et urémique et une brusque défaillance rénale. Bien que les souches de sérotype O157:H7 sont les plus prévalentes, d'autres sérotypes présentent un potentiel pathogénique similaire. On peut citer par exemple les sérotypes O26, O111, O103 (Ogura, Ooka et al. 2009). Les EHEC partagent certains facteurs de virulence avec les EPEC (dont le système de sécrétion de type III et les protéines qui y sont liées telles que l'adhésine intimine codée par le locus *eae*). C'est pourquoi ils présentent certaines caractéristiques communes quand à la pathogénèse (dont le mode d'attachement à la membrane de la cellule épithéliale ainsi que la condensation des filaments d'actine menant à l'effacement des microvillosités de la paroi intestinale). Pourtant les EHEC possèdent des facteurs de virulence supplémentaires, par exemple une entérohémolysine dont le gène se trouve sur un plasmide, et Stx1 et Stx2, des Shiga-toxines codées par un prophage qui entraine la mort de certains patients. Ces bactéries utilisent les fimbriae pour la liaison, sont modérément invasives et la Shiga-toxine libérée peut déclencher une réponse inflammatoire intense.

Ce pathotype continue à avoir une certaine incidence (surtout chez les enfants de moins de 5 ans) par ingestion de viande de bœuf ou de lait contaminés, ceci même dans les pays les plus développés. Par exemple, en 2008, en France, cinquante neuf cas ont été confirmés (http://www.invs.sante.fr/surveillance/shu/).

2.2.6 Les Shigella et les EIEC (« Enteroinvasive E. coli *»)*

Les *Shigella* sont aujourd'hui considérées comme faisant partie à part entière de l'espèce *E. coli*. En effet, de nombreuses études ont démontrées que *Shigella* et *E. coli* sont d'un point de vue taxonomique impossibles à distinguer au niveau de l'espèce (Pupo, Lan et al. 2000; Escobar-Paramo, Giudicelli et al. 2003; Wei, Goldberg et al. 2003). La distinction taxonomique officielle a toutefois été maintenue essentiellement parce que la pathologie induite est sévère et spécifique. Cette distinction reste donc utile à des fins de diagnostic, de traitement, et d'épidémiologie. Les EIEC sont, d'un point de vue biochimique, génétique et physio-pathologique, très proche des *Shigella* (Nataro and Kaper 1998). Quatre groupes de

Shigelles sont classiquement distingués : *S. dysenteriae*, *S. flexneri*, *S. boydii*, *S. sonnei*, chacun regroupant un ou plusieurs sérotypes, un seul dans le cas de *S. sonnei* qui est clonal, plusieurs pour les autres. Il n'y a pas de lien entre ces différents groupes et la phylogénie des souches (Pupo, Lan et al. 2000). Il apparait trois clusters phylogénétiques principaux, le premier contient la majorité des *S. boydii* et des *S. dysenteriae* et quelques *S. flexneri*. Le second cluster contient des *S. boydii* et au moins une *S. dysenteriae*. Le troisième cluster contient le reste des *S. flexneri* et au moins une *S. boydii*. A l'extérieur de ces trois clusters principaux se trouvent *S. sonnei* et certaines *S. dysenteriae*.

Les *Shigella* et les EIEC sont des pathogènes strictement humains ; ce sont les agents de la dysenterie bacillaire ou shigellose (caractérisée par d'abondantes diarrhées mêlées de sang et de mucus et une forte fièvre). Des millions de cas sont déclarés chaque année, essentiellement dans les pays en voie de développement et lorsque les conditions d'hygiène sont insuffisantes. Elles tuent plusieurs centaines de milliers d'individus dans le monde, pour l'essentiel des enfants de moins de 5 ans. Mais les pays développés ne sont pas complètement épargnés. Dans ce contexte, c'est *S. sonnei* qui est le plus souvent impliquée. En 1996, plusieurs centaines de shigelloses à *S. sonnei* dont certaines étaient résistantes à l'amoxicilline et au cotrimoxazole sont apparues en Ile de France, surtout chez des enfants de 1 à 4 ans. Moins de 20% des personnes touchées étaient des adultes (http://ile-de-france.sante.gouv.fr). En 2007, dans le Val de Marne, d'autres cas ont été signalés : au total 53 cas de gastro-entérite aigüe à *S. sonnei* entre le 21 octobre 2007 et le 21 janvier 2008 dans 3 écoles privées dans le Val de Marne et à Paris.

Le pouvoir pathogène des *Shigella* et des EIEC est essentiellement lié à l'invasion des cellules intestinales et à la dissémination de bactéries de cellules en cellules produisant une nécrose de la muqueuse du colon (Nataro and Kaper 1998). En plus de l'invasion et de la dissémination à l'intérieur des cellules épithéliales, les *Shigella* et probablement aussi les EIEC induisent l'apoptose des macrophages qu'ils infectent. Les *Shigella* et les EIEC produisent également des toxines qui participent à l'amplification des dommages, on peut citer par exemple la Shiga-toxine produite exclusivement par *S. dysenteriae* sérotype 1 qui est la plus connue et la plus pathogène. La dose infectieuse des Shigelles est très faible : de 10 à 100 Shigelles (Crockett, Haas et al. 1996). Pour les EIEC elle s'avère supérieure (Nataro and Kaper 1998).

Le modèle de pathogénèse des Shigelles et des EIEC comprend 5 phases : (i) la pénétration de la cellule épithéliale, (ii) la lyse de la vacuole endocytique (qui lui a permis de rentrer dans la cellule), (iii) la multiplication intracellulaire, (iv) le déplacement à travers le cytoplasme, (v) la pénétration dans une cellule épithéliale adjacente (Nataro and Kaper 1998). Les gènes impliqués dans l'invasion sont portés par un plasmide nommé pInv, que ce soit chez *Shigella* ou dans le cas des EIEC (Parsot 2005).

CONCLUSION :

E. coli est un organisme modèle. Mais c'est aussi une bactérie dont les souches présente des modes de vie très variés. En effet, certaines d'entre elles sont dites commensales, nous devrions même dire symbiotiques avec les mammifères et avec l'homme en particulier. Alors que d'autres sont pathogènes. Dans ce cas, elles peuvent produire des infections intestinales impliquant des diarrhées plus ou moins sévères, ou extra-intestinales (cystites, pyélonéphrites, méningites, septicémie ....), ce sont alors des pathogènes opportunistes à l'exception des *Shigella* et EIEC qui sont des pathogènes obligatoires. *E. coli* cause chaque année près de deux millions de mort. Chaque pathotype est caractérisé par son mode d'action. Parce que les souches d'*E. coli* sont facile à cultiver et donc à étudier, mais aussi parce qu'elles présentent une diversité physiopathologique importante, une masse impressionnante de connaissance a été amassée par les scientifiques depuis sa découverte en 1885 par Thomas Escherich. C'est l'étude d'*E. coli* qui a posé les fondements de la génétique bactérienne et de la biologie moléculaire. La diversité génétique sur laquelle agissent les processus évolutifs responsables des capacités d'adaptation de l'organisme est le fruit de deux mécanismes que nous détaillerons dans les chapitres suivants : la mutation et la recombinaison.

Chapitre II : *E. coli* : sa mutagénèse.

E. coli est un bon modèle pour étudier la mutagénèse car il est génétiquement haploïde, facile à cultiver et a un temps de génération en moyenne de 20 minutes. Le chromosome bactérien peut être modifié par des altérations physico-chimiques ou des erreurs de réplications. Quand de tels changements sont transmis à la génération suivante, on parle de mutations. Les mutations peuvent être de différentes natures : des substitutions lorsqu'une paire de base est modifiée, des délétions ou des insertions lorsque des bases sont ajoutées ou supprimées. Elles peuvent aussi consister en des inversions : un fragment d'ADN est alors coupé et réinséré dans l'autre sens, ou encore en des duplications : une forme d'insertion qui met en œuvre une séquence présente ailleurs dans le génome, généralement immédiatement adjacente à l'insert. Elles peuvent enfin être le résultat de la combinaison de plusieurs des types de mutations précédentes, elles sont alors dites complexes.

1. Les systèmes de réparation en jeu, les types de mutation et leur cause

L'ADN support de l'information génétique indispensable à la reproduction, et à la vie de la cellule, est une molécule particulièrement stable, mais sa grande longueur et la précision de l'information qui y est codée font qu'une mutation sur 5 millions de paires de bases peut suffire à rendre une bactérie non viable (mutation dite létale). En conséquence malgré sa robustesse, la molécule d'ADN reste une structure fragile qui nécessite un fort investissement de la part de la bactérie pour rester le plus intègre possible. On peut considérer deux sources principales de mutations : les erreurs associées à des lésions physico-chimiques et les erreurs de réplication. On distingue les mutations spontanées, liées à la vie normale de la cellule et les mutations provoquées, liées à l'environnement. Par exemple les rayons X, les ultra-violets, ou certains agents chimiques augmentent le taux de mutation. La résultante des mutations proprement dites et des mécanismes de réparation mis en œuvre n'est pas nulle et correspond au taux de mutations observé.

1.1. Les systèmes de réparation

Les dommages subits par l'ADN peuvent résulter de plusieurs processus que je présenterai juste après, par exemple : l'hydrolyse, la désamination, l'alkylation et l'oxydation. En conséquence, une part importante de l'activité enzymatique est dédiée à la

protection et à la réparation de la molécule d'ADN. Les mécanismes de protection éliminent les nucléotides altérés avant leur intégration dans l'ADN au cours de la réplication. Les mécanismes de réparation enlèvent les lésions physico-chimiques présentes dans la molécule d'ADN. L'ensemble de ces mécanismes, couplé à la forte fidélité de la polymérase assure un taux de mutation par base et par réplication de l'ordre de 5×10^{-8} maximum pour *E. coli* avant la mise en œuvre du système de réparation post réplicatif (appelé système de réparation des mésappariements ou SRM). En effet, quand le SRM est inactivé, le taux de mutation est augmenté de 100 à 1000 fois par rapport au taux de mutation standard de *E. coli* qui est de 5×10^{-10} (Miller 1996).

1.1.1. Principaux mécanismes de réparation pré-réplicatif

Il existe plusieurs mécanismes de réparation pré-réplicatif (Friedberg, Walker et al. 1995).

L'un d'entre eux est la réparation directe des dommages. Dans ce cas, une enzyme catalyse la réaction inverse de celle qui a causé la lésion. Par exemple, la méthyl-transférase (codée par le gène *ada*) débarrasse les bases des alkylations qu'elles ont subies.

Un autre mécanisme de réparation s'effectue par excision de base (BER). La base portant une lésion est d'abord excisée par une ADN glycosylase. Le site abasique résultant est ensuite éliminé par une endonucléase, le nucléotide manquant est alors remplacé grâce à l'action conjuguée d'une polymérase et d'une ligase. De nombreuses lésions ne sont pas reconnues par le BER, notamment celles affectant deux bases voisines comme les dimères de pyrimidine engendrés par les rayons ultra-violets.

La réparation de l'ADN peut également se faire par excision de nucléotides (NER). Son activité principale est de réparer les dimères de pyrimidine (engendré par les UV). Au lieu d'enlever uniquement une base, un fragment d'une dizaine de nucléotides comportant la lésion est excisé. Le complexe UvrABC coupe le brin portant la lésion à 4 nucléotides en 3' et à 7 en 5' du site problématique. L'action de PolI et de la ligase assure alors la synthèse de la partie excisée.

Il existe un type de réparation qui fait appel à la recombinaison. Il est utile lorsque l'ADN a subit une cassure double brins. C'est dans ce cas le processus de recombinaison

homologue qui permet de restituer la molécule d'ADN intacte. Pour que cela soit possible, il faut qu'il existe une autre molécule d'ADN identique n'ayant pas subie de dommage dans la région homologue de la région à réparer.

Les nucléotides, avant même leur insertion dans la matrice d'ADN, peuvent être sujets à des altérations chimiques. Certaines enzymes dégradent ces nucléotides altérés. Il s'agit par exemple de MutT (Maki and Sekiguchi 1992) qui dégrade la 8-oxo-guanine dans le cytoplasme.

L'ensemble des systèmes décrits précédemment assure une stabilité chimique de l'ADN et donc la conservation de l'information qui y est contenue. Cette information, pour être transmise à la génération suivante doit être répliquée avec une forte fidélité. Cette fidélité est le fait de deux caractéristiques de l'ADN polymérase : sa sélectivité ce qui signifie qu'elle associe préférentiellement les bases complémentaires et son activité exonucléase de 3' vers 5' qui lui permet de corriger une base mésappariée avant d'insérer la suivante (Hutchinson, 1996).

Pourtant, la réplication peut produire, elle aussi, des erreurs qui peuvent être corrigées soit par des enzymes très spécifiques reconnaissant un type de mésappariement particulier dans une séquence particulière, soit par un système généralisé de réparation des mésappariements.

1.1.2. Principaux mécanismes de réparation post-réplicatif

La réparation spécifique de mésappariements ressemble au NER et peut entraîner l'excision de quelques nucléotides seulement (Friedberg, Walker et al. 1995).

Le système généralisé de réparation des mésappariements (SRM), au contraire, entraîne l'excision d'environ un millier de paires de bases. Chez *E. coli*, il repose sur l'action de MutS, L, H et U, ainsi que de dam-méthylases. Juste après la réplication, les mésappariements sont reconnus par MutS qui forme alors un complexe avec MutL et MutH. Ce dernier parcourt le brin d'ADN de part et d'autre du mésappariement. Lorsqu'un site de méthylation de l'adénine, GATC est rencontré, MutH coupe le brin dont le site n'est pas méthylé : il s'agit du brin néo-synthétisé (en effet, la dam-méthylase n'a pas encore eu le temps de méthyler l'adénine de ce site). Une fois le brin néo-synthétisé coupé, l'hélicase II

(MutU) élimine la portion du brin néosynthétisé entre le mésappariement et la coupure. Le complexe PolIII permet alors de resynthétiser le fragment (Fig 2).

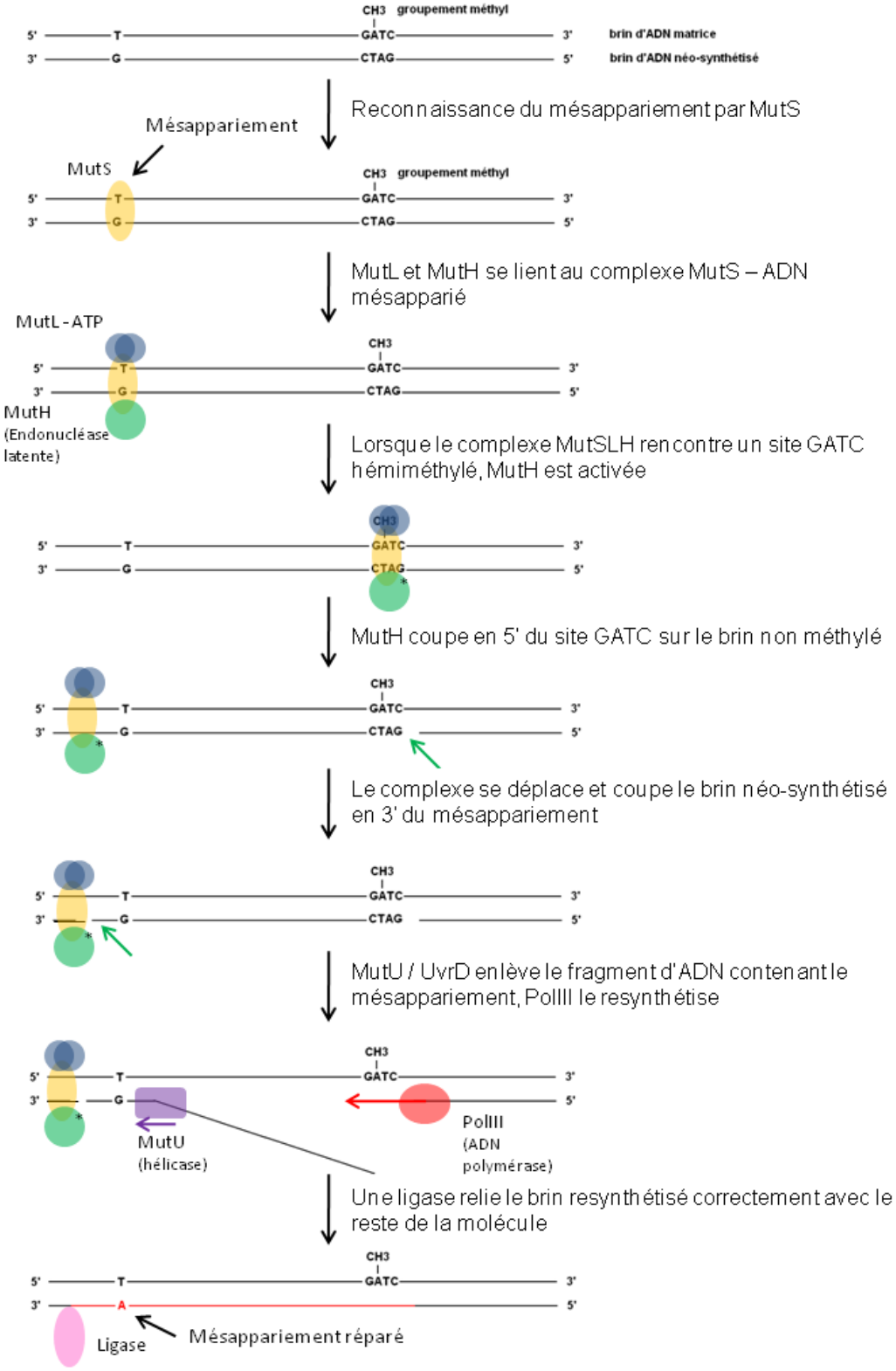

Fig. 2 : Le système de réparation des mésappariements. Ce mécanisme met en œuvre les protéines MutS, MutL, MutH, MutU et PolIII, ainsi qu'une ligase. L'ADN resynthétisé par PolIII est figuré en rouge.

Certains de ces systèmes de réparation sont fortement conservés au cours de l'évolution. C'est le cas par exemple du BER, du NER, de la recombinaison homologue et du SRM. Chez l'homme, certains gènes du NER, s'ils sont inactivés prédisposent à certains cancers de la peau et des yeux, en augmentant la sensibilité au soleil (Hoeijmakers 2001). Les syndromes dans lesquels un gène du NER défectueux a été mis en évidence sont : le *xeroderma pigmentosum* (cette maladie multiplie le risque de cancer de la peau par 1000), le syndrome de Cockayne, la trichothiodystrophie. Il a été également découvert que des gènes du SRM défectueux sont impliqués dans le cancer colorectal héréditaire sans polypose (HNPCC : « hereditary non polyposis colorectal cancer », ou syndrome de Lynch). Il s'agit d'une prédisposition héréditaire au cancer du colon, mais aussi de bien d'autres cancers. Ce syndrome se traduit par une instabilité des microsatellites.

1.2. Les altérations chimiques et les systèmes de réparation pré-réplicatif mis en jeu

La mutation, résultante d'altérations chimiques, peut avoir de nombreuses origines (Fig 3):

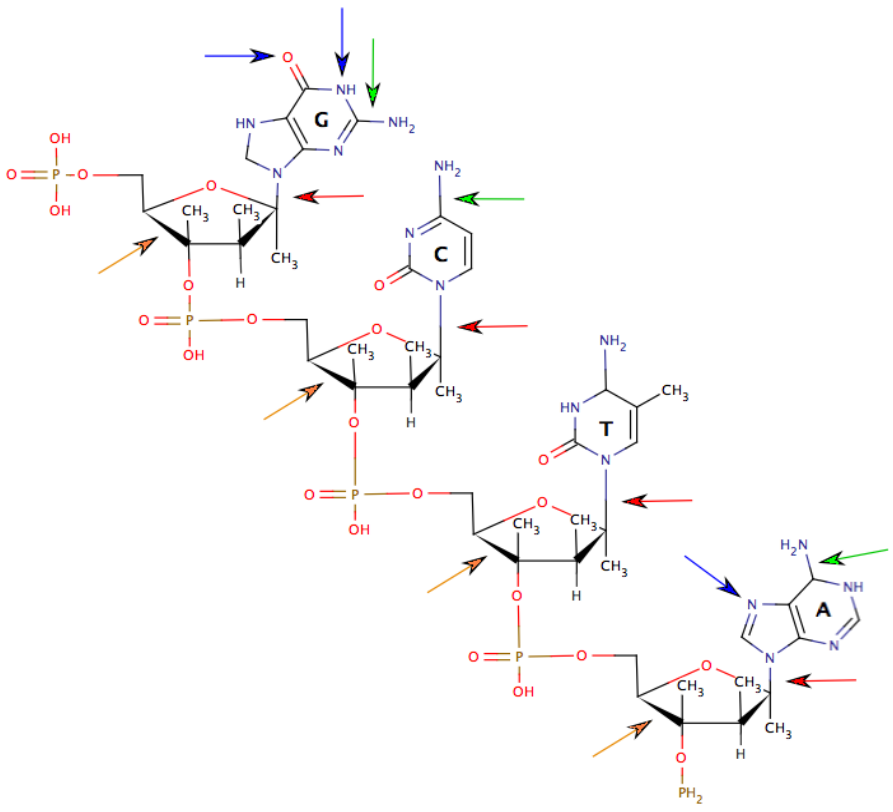

Fig. 3 : Localisation des principaux dommages pouvant affecter la molécule d'ADN. Un court segment d'un seul brin de la double hélice d'ADN est représenté avec les quatre bases azotées classiques. Les emplacements des dommages principaux sont indiqués par les flèches. Le type de mutation est symbolisé par la couleur. Des emplacements qui peuvent subir une attaque hydrolytique sont indiqués par les flèches rouges, les désaminations par des flèches vertes, les alkylations par des flèches bleues et ceux pouvant subir une oxydation sont indiquées par les flèches oranges. G : guanine ; C : cytosine ; T : thymine ; A : adénine (Cyril Langlois, com. pers.).

1.2.1. L'hydrolyse

Par ce processus, une molécule d'eau clive la liaison entre le désoxyribose et la base. Il s'agit d'un processus spontané (Hutchinson 1996). Ces sucres qui ont perdu leur base sont appelés les sites AP (apurinique/apyrimidique). Ils sont efficacement réparés par le système de réparation des bases excisées (BER). En fait, pour réparer certains nucléotides mutés, des ADN glycosylases transforment en sites AP la base problématique pour permettre son remplacement.

1.2.2. La désamination

La désamination (perte du groupement NH_2) spontanée de la cytosine est fréquente. Elle forme alors une uracile. L'ADN qui avait un appariement C-G normal présente maintenant un appariement U-G anormal, dit pré-mutagène. Il est dit pré-mutagène car s'il n'est pas réparé avant la réplication, il formera un appariement G-C d'un coté et A-U de l'autre. Le processus de réparation de ces erreurs est le BER, et implique lors de sa première phase l'Uracile N-glycosylase. Les autres désaminations (adénine en hypoxantine et guanine en xantine) sont moins fréquentes (Hutchinson 1996).

1.2.3. L'alkylation

Des agents mutagènes alkylants endogènes ou exogènes ont le pouvoir de transférer des groupements éthyle ou méthyle. Les atomes d'azote des purines (N_3 de l'adénine et N_7 de la guanine) et l'atome d'oxygène (O_6) de la guanine sont particulièrement susceptibles d'être méthylé par alkylation. Dans ce dernier cas, la guanine est transformée en O_6-méthylguanine, ce qui entraîne la rupture de liaisons hydrogène avec la cytosine. Par

ailleurs, la O_6-méthylguanine peut s'apparier par erreur avec la thymine, ce qui peut conduire à une transition : G : C -> A : T. La réparation de cette méthylation s'effectue sans coupure de l'ADN par la O_6-méthylguanine méthyltransférase qui transfère le groupement méthyle sur un de ces acides aminés et par là-même s'inactive irréversiblement (Verbeek, Southgate et al. 2008).

1.2.4. L'oxydation

L'oxydation des bases de l'ADN provoque des lésions, en fait des coupures franches simples et doubles de l'ADN. Les coupures simple brin sont réparées par le BER, les double brins, par la recombinaison homologue. L'oxydation peut aussi affecter une guanine, alors transformée en 8-oxo-guanine qui peut s'apparier avec une adénine. Cette erreur est habituellement réparée par le BER. Plus précisément, la protéine MutM excise les 8-oxo-guanines de la matrice, alors que MutY excise les adénines appariées avec elles (Miller 1996).

Le type de mutations le plus fréquent est la substitution de bases. La délétion ou l'insertion d'une ou de deux paires de bases est le second type de mutation le plus fréquent. Elle induit un décalage de phase (« frameshift ») lorsqu'elle est située sur une séquence codante (Hutchinson 1996).

1.3. Les erreurs réplicatives

Ces altérations chimiques peuvent induire la polymérase en erreur et la pousser à produire des mutations, mais la polymérase elle même, en l'absence de lésion, comme tout processus biologique, peut produire des erreurs. Ces erreurs peuvent résulter en des mutations ponctuelles, des insertions, des délétions ou des réarrangements. Pour certains de ces évènements un système de réparation post réplicatif (le SRM) peut intervenir et réduire le taux de mutation à 10^{-10} par base et par génération (Miller 1996).

1.3.1. Mutations ponctuelles

Les erreurs les plus fréquentes, conséquences ou non des dommages préexistants, sont les substitutions. Certaines des mutations ponctuelles résultent aussi de l'utilisation de polymérases de faible fidélité qui permettent d'outrepasser des lésions de l'ADN qui bloqueraient la polymérase réplicative. Ces polymérases induites par les stress qui

endommagent l'ADN et / ou bloquent la réplication, sont appelées des SOS polymérases ou polymérases translésionelles. Il s'agit de PolII, PolIV et PolV codées respectivement par les gènes *polB*, *dinB* et *umuCD* (Bjedov, Lecointre et al. 2003).

1.3.2. Dérapages lors de la réplication

Les décalages de phase peuvent être dus à un dérapage de l'ADN polymérase lors de la réplication lorsqu'il y a des bases répétées en tandem.

La deuxième façon d'obtenir un décalage de phase se produit lorsqu'une base a été abimée par un des mécanismes décrits précédemment. Dans ce cas, lors de la réplication, l'un des deux brins peut glisser par rapport à l'autre et engendrer un frameshift (ou décalage de phase). Le taux d'erreur de la machine réplicative par base et par réplication est d'environ 10^{-7}.

1.3.3. Excision d'une base endommagée

Pendant la réplication, il est possible qu'une nouvelle base soit placée en face d'une base endommagée. L'ajout de cette nouvelle base engendre une pause de la réplication. La base abimée pourra alors être excisée si la nouvelle base forme une liaison Watson-Crick avec la base juste derrière la base endommagée. Dans ce cas, un décalage de phase -1 sera créé.

1.4. Les réarrangements

L'altération à grande échelle la plus fréquente est la délétion. Elle peut concerner 10 bases comme plusieurs centaines de milliers de bases. Pour qu'il y ait délétion, il faut que l'ADN double brins soit rompu à deux endroits et qu'une ligation se produise entre les deux. Il arrive également que lors de la réplication, s'il y a des séquences répétées, elles s'apparient entre elles, en formant alors une boucle d'ADN qui sera excisée.

En conclusion, chez *E. coli*, les mutations spontanées les plus fréquentes sont les transitions C : G -> T : A. Pour ce qui est des décalages de phase, les délétions d'une seule base sont les plus fréquentes.

En le ramenant à un taux de mutation par génome et par réplication, Drake (Drake 1991) a montré que la valeur obtenue (0,003) est conservée chez les microorganismes à

ADN. Les causes évolutives à l'origine d'une telle stabilité du taux de mutation entre des espèces si différentes restent encore inconnues. Le taux de mutation génomique est donc contrôlé génétiquement par de nombreux mécanismes. L'inactivation de certains de ces mécanismes de réparation peut résulter en un taux de mutation plus élevé, on parle alors de souche mutatrice.

2. Les différentes échelles spatio-temporelles de la mutation

Comme nous venons de le voir, il existe de nombreuses sources mutagènes qui sont pour la plupart très efficacement corrigées par les systèmes de réparation de la bactérie. Si avoir un taux de mutation faible est sélectionné quand une bactérie est adaptée à son environnement, l'adaptation à de nouvelles conditions nécessite l'apparition de mutations dites bénéfiques qui permettent à la bactérie de mieux survivre ou exploiter ce nouvel environnement. Dans de telles conditions, la sélection peut favoriser une augmentation du taux de mutation. Différentes études (Jyssum 1960; Gross and Siegel 1981; LeClerc, Li et al. 1996; Matic, Radman et al. 1997) ont montré que les mutateurs sont présents en quantité non négligeables dans les populations naturelles d'*E. coli* (de 1 à 15 %). Cette augmentation peut être locale ou globale, constitutive ou transitoire (inductible).

2.1. Globale et permanente

Cette augmentation du taux de mutation peut être globale et permanente : on parle alors de souches mutatrices constitutives (Tenaillon, Toupance et al. 1999). L'allèle mutateur de ces souches serait fixé par auto-stop avec les rares allèles favorables qu'il a permis de générer. Le mécanisme impliqué est celui des gènes mutateurs. En fait il s'agit de gènes de réparation de l'ADN qui, lorsqu'ils sont mutés, augmentent le taux de mutations : *mut*, *dam*, *uvr*... Ce mode d'augmentation de la mutation a un fort coût car des gènes indispensables à la survie de la cellule peuvent subir des mutations délétères et cela, même en dehors des phases d'adaptation. Il a aussi été proposé que les bactéries passent par des phases où ces gènes sont mutés, puis réactivés grâce à la recombinaison avec un allèle sauvage (Denamur, Lecointre et al. 2000).

2.2. Globale et transitoire

Une augmentation globale mais transitoire du taux de mutation peut-être observée, on parle alors de souches mutatrices inductibles (Bjedov, Tenaillon et al. 2003). L'existence de ces systèmes génétiques inductibles (activés en condition de stress) peut être interprétée comme une stratégie évolutive pour s'adapter au stress et survivre ou bien comme le prix à payer pour la survie, une augmentation du taux de mutation n'étant que le sous-produit de mécanismes directement impliqués dans la survie (Tenaillon, Denamur et al. 2004). Mécanistiquement il semble que le système SOS soit impliqué dans ce type de contrôle du taux de mutation. Le système SOS est un ensemble de gènes co-régulés qui permettent de réinitialiser la fourche réplicative lorsqu'un stress tel que l'irradiation aux UVs la bloque. Une induction très forte ou persistante de ce système entraîne une plus forte mutagénèse. Il a été mis en évidence que le gène *rpoS* codant pour la sous-unité sigma S de l'ARN polymérase régulait positivement ce type de mutagénèse. Ce gène était préalablement connu pour réguler l'expression de gènes impliqués dans la réponse à divers stress dont la phase stationnaire, un déficit en nutriment, des chocs osmotiques, acides, thermiques et oxydatifs. La possibilité d'adapter le taux de mutation aux conditions environnementales semble intéressant d'un point de vue évolutif, car ce processus n'engendre pas de mutation délétère en dehors des phases d'adaptation. Mais l'ensemble des mécanismes impliqués dans ce type de mutagénèse étant aussi directement associé à la survie, il est difficile de savoir si le processus à l'œuvre est la sélection directe pour la survie ou pour une augmentation du taux de mutation.

2.3. Locale

Le taux de mutation peut aussi être variable le long du chromosome et certains locus peuvent exhiber un fort taux de mutation. On parle alors de locus contingent (Moxon, Rainey et al. 1994). Ces locus sont souvent impliqués dans la pathogénèse et codent des protéines membranaires qui entrainent une reconnaissance suivie d'une destruction de la bactérie par le système immunitaire. C'est pourquoi ces gènes sont soumis à de fortes pressions de sélection afin de pouvoir échapper au système immunitaire de l'hôte. L'augmentation du taux de mutation est alors associée à la présence de répétitions (une série de guanines par exemple) sur lesquels la polymérase a une forte propension à « déraper » et donc à introduire des décalages de cadre de lecture. Ces mutations ne sont pas distribuées aléatoirement sur le génome, mais par contre elles ne sont pas

programmées, c'est à dire que leur fréquence n'augmente pas quand elles pourraient être bénéfiques. Cette stratégie évolutive semble plus intéressante que la stratégie « mutateurs constitutifs » puisque d'autres gènes indispensables à la vie de la cellule ne risquent pas d'être endommagés et que des taux de mutation très élevés peuvent être atteints (Funchain, Yeung et al. 2000).

CONCLUSION :

Comme nous venons de le voir, malgré des systèmes de réparation efficace, l'ADN subit des mutations associées à des lésions physico-chimiques et aux erreurs de réplication. Les pressions de sélection agissant sur ce taux de mutation sont variables et contradictoires et peuvent sélectionner des modulations du taux de mutations globales, inductibles, ou spécifiques à chaque gène. De nombreuses études ont tenté de mesurer le taux de mutation intrinsèque d'une bactérie. Pour cela, les évolutionnistes examinent les substitutions affectant les sites qu'ils pensent être indemnes de pressions sélectives. Les bactéries étant généralement assez pauvres en pseudogènes, ils se sont servis des sites synonymes c'est à dire ne modifiant pas l'acide aminé codé lorsqu'ils sont mutés (Ochman 2003). Seulement, il a été montré que ces sites synonymes ne sont pas libres de toute sélection, au contraire, ils en subissent de nombreuses qui contraignent la séquence et masque la signature de la mutation. Par exemple, depuis bientôt 30 ans, les biologistes savent qu'il existe un biais de codons, résultat de la sélection naturelle agissant sur les codons synonymes pour améliorer la fidélité et l'efficacité de la traduction (Grantham, Gautier et al. 1981; Sharp and Li 1986). Kudla et collaborateurs ont récemment montré que la répartition des codons synonymes dans la séquence des gènes modifiait leur expression (Kudla, Murray et al. 2009). Grâce à 154 gènes synthétisés en laboratoire, en faisant varier aléatoirement les codons synonymes, les auteurs ont mis en évidence que la stabilité du repliement de l'ARNm au niveau du site de fixation du ribosome expliquait plus de la moitié de la variation de l'expression des protéines codées par ces gènes. On peut ainsi imaginer qu'il est fort probable que certains gènes réels codant des protéines nécessitant d'être traduites en grand nombre subissent une sélection affectant le choix des codons synonymes selon leur position.

Chapitre III : *E. coli* : Impact des processus de mutation et de recombinaison sur sa phylogénie.

La structure d'une population est la conséquence de la balance entre les mutations et les évènements de recombinaison. Lorsque les évènements de recombinaison sont rares, la population est dite clonale, lorsqu'ils sont beaucoup plus fréquents que les mutations, la population est dite panmictique. Comment peut-on qualifier la structure de la population d'*E. coli* ? Quels types de recombinaison affectent son génome ? Quel impact la recombinaison a-t-elle sur le génome? Peut-on reconstruire l'arbre phylogénétique de l'espèce ?

1. Clonalité versus panmixie

Maynard Smith et collaborateurs en 1993 ont classé plusieurs espèces bactériennes selon le type de structure de leur population. Ils observent que les bactéries sont réparties dans un continuum de structure populationnelle allant de la clonalité à la panmixie. Une population est dite clonale quand elle subit très peu de recombinaison par rapport à la mutation, elle est alors soumise à une évolution dite verticale. Au contraire, dans une population panmictique, tous les individus ont la même probabilité d'échanger du matériel génétique entre eux par recombinaison ou transfert horizontal (Fig 4). Contrairement aux eucaryotes pour lesquels la méiose est le siège du brassage génétique, chez les bactéries ce brassage est séparé de la reproduction (qui s'effectue par scissiparité).

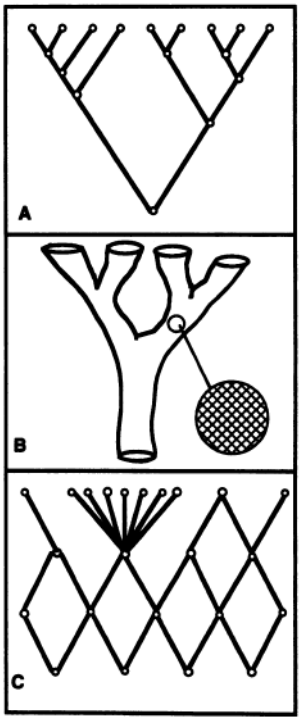

Fig. 4 : Les structures de population. A : La structure de la population est strictement clonale, il n'y a pas de recombinaison ni à l'intérieur d'une branche de l'arbre ni entre les deux branches de l'arbre. B : Il n'y a pas de recombinaison entre les deux principales branches de l'arbre, mais les évènements de recombinaison sont fréquents dans chacune de ces deux branches, c'est pourquoi elles sont figurées sous forme de réseau. C : La structure de la population est panmictique, elle est représentée sous la forme d'un réseau plutôt que d'un arbre. Occasionnellement, un individu très compétitif augmente rapidement dans la population pour produire un clone épidémique (Maynard-Smith, Smith et al. 1993).

2. Les transferts horizontaux affectant les procaryotes.

Un transfert horizontal consiste en une acquisition d'ADN qui ne s'est pas faite par la stricte voie verticale : de parent à descendant. Ainsi dans le cas des bactéries, l'acquisition d'un plasmide constitue un transfert horizontal. De même, l'entrée et la persistance de tout ADN parasite (transposons, bactériophages qui sont les virus des bactéries ...) dans la cellule est un transfert. Ces types de séquences sont adaptés à ce mode de diffusion. Certains transferts horizontaux impliquent des séquences qui ne sont pas dédiées à être mobiles. Ce sont ces derniers cas qui suscitent un débat, car ils peuvent affecter des gènes essentiels au fonctionnement cellulaire, qui d'ordinaire sont utilisés pour reconstruire l'histoire évolutive des organismes. Par exemple, Milkman et Bridges (Milkman and Bridges 1990; Milkman and Bridges 1993) ont montré qu'un certain nombre de régions de l'opéron tryptophane (*trp*) avait subit divers événements de recombinaison. Il existe trois mécanismes permettant aux bactéries d'intégrer de l'ADN étranger. La transformation qui consiste en l'internalisation d'une molécule d'ADN libre du milieu, la conjugaison qui nécessite un contact entre deux individus et la transduction qui met en œuvre une particule virale véhiculant l'ADN transformant. Quand l'ADN étranger a pénétré la cellule, il peut intégrer son génome par recombinaison homologue (se produisant entre deux séquences très similaires), par l'intermédiaire de crossing-over (recombinaison réciproque entre deux molécules d'ADN homologues) et/ou de conversion génique (transfert non réciproque de courtes séquences d'ADN entre séquences homologues). Il peut également être intégré dans le génome par recombinaison site spécifique (utilisant des recombinases reconnaissant certains sites de recombinaison particulier dans le génome) ou illégitime (aussi nommée non homologue, elle s'effectue en l'absence de toute séquence particulière).

2.1. Les mécanismes d'entrée de l'ADN étranger dans la bactérie

2.1.1. La transformation

La transformation est un transfert génétique au cours duquel de l'ADN bicaténaire, libre, nu et en solution est introduit dans une bactérie réceptrice, puis intégré au chromosome par recombinaison homologue.

Certaines souches bactériennes, sont naturellement compétentes, c'est à dire qu'elles ont la capacité de capturer de l'ADN présent dans l'environnement. Pour de nombreuses bactéries, dont *E. coli*, aucune transformation naturelle n'a été observée, mais il est possible d'obtenir une transformation artificielle. Dans ce cas, on utilise en laboratoire des techniques (l'électroporation par exemple) permettant de perméabiliser l'enveloppe bactérienne, afin de rendre la bactérie compétente.

Dans ce mode de transfert horizontal, l'étape de recombinaison homologue est critique. Généralement, seul un ADN présentant un fort degré de similarité avec le chromosome sur une portion de séquence de longueur variable selon les bactéries, pourra recombiner. Cependant, dans certaines conditions de stress important ou lorsque des gènes contrôlant la spécificité de l'appariement des deux brins d'ADN sont non-fonctionnels, des évènements de recombinaisons hétérologues peuvent avoir lieu, et provoquer alors l'intégration de l'ADN d'une espèce éloignée.

2.1.2. La conjugaison

Lors d'une conjugaison le transfert de gènes s'effectue dans un sens déterminé. Le caractère donneur est sous la dépendance d'un facteur de fertilité (facteur F). Le facteur F est en fait le premier plasmide conjugatif mis en évidence chez les bactéries. C'est un gros plasmide (94 500 pb) qui contrôle sa propre réplication, son nombre de copies, la répartition des copies dans les cellules filles et son transfert. Les nombreux gènes gouvernant sa transmission sont situés dans l'opéron *tra*. Le transfert s'effectue grâce à des pili qui permettent le contact entre les deux cellules. Ces pili sont qualifiés de sexuels car il existe des mécanismes pour empêcher que la conjugaison ne survienne entre deux bactéries F^+ : il s'agit de l'exclusion de surface et de l'exclusion d'entrée. Dans le cas de l'exclusion de surface, la bactérie F^+ présente à sa surface une lipoprotéine fixant le pilus avec une affinité

très forte. Une fois le pilus fixé, le processus de conjugaison est arrêté. Dans celui de l'exclusion d'entrée, une protéine de la membrane interne, bloque les signaux envoyés par le donneur après l'agrégation.

Entre une bactérie F^+ et une bactérie F^- reliées par un pili sexuel, le transfert de l'ADN est possible grâce à la mise en place d'un complexe protéique appelé relaxosome qui coupe le plasmide et le désenroule à partir de l'origine de transfert ou *oriT*. Un brin d'ADN pénètre dans la bactérie réceptrice puis, aussi bien chez la bactérie donatrice que chez la bactérie réceptrice, le brin complémentaire sera synthétisé. En conséquence, le facteur F persiste chez la bactérie donatrice (qui reste F^+) et une copie du facteur F est acquise par la bactérie réceptrice (qui devient F^+).

Des *E. coli* Hfr pour « high frequency of recombination » ont été décrites. Dans ces souches, le plasmide F est intégré au génome chromosomique. Cette intégration peut se faire à divers endroits. Lors d'une conjugaison entre une bactérie Hfr et une bactérie F^-, l'ADN simple brin est transmis à la bactérie réceptrice à partir de l'origine de transfert du plasmide intégré. Lors du transfert l'intégrité du génome de la bactérie donatrice est assurée par un processus de réplication asymétrique. L'ADN monocaténaire transféré est également répliqué dans la bactérie réceptrice et les gènes transférés peuvent être incorporés dans le chromosome de la bactérie réceptrice par crossing-over. Si la conjugaison n'est pas interrompue, le chromosome entier est transmis en environ 100 minutes. Mais généralement les deux bactéries se séparent avant et la bactérie receveuse reste F^- (Fig 5).

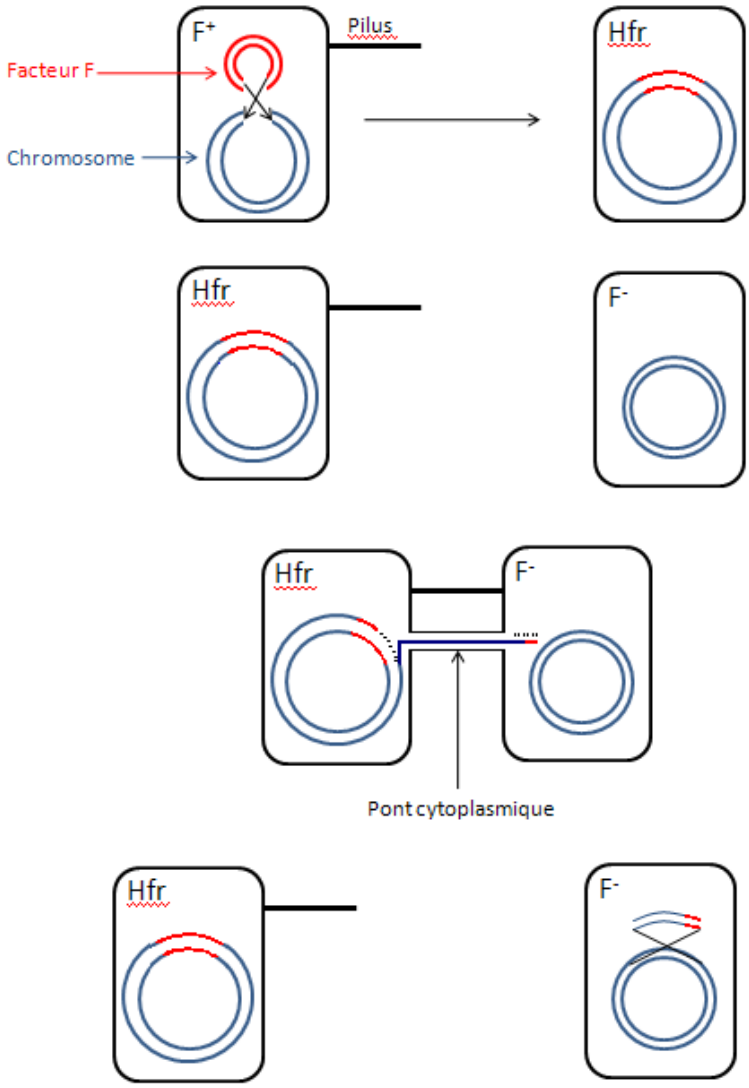

Fig. 5 : Formation d'une bactérie Hfr puis conjugaison entre une bactérie Hfr et une bactérie F- . La première étape présente une bactérie F+, c'est à dire ayant acquis le plasmide conjugatif appelé facteur F qui devient une bactérie Hfr, en intégrant le facteur F à son génome. Cette bactérie Hfr peut alors conjuguer avec une bactérie F-. Les pili (figurés par un trait noir sur le schéma) présentés par la bactérie Hfr permettent la reconnaissance de la bactérie F-. Un pont cytoplasmique permet le transfert d'un brin de l'ADN dont le complémentaire est aussitôt resynthétisé par chacune des bactéries protagonistes. Dans la dernière étape, les bactéries se sont séparées avant que le chromosome entier ne soit transféré, la bactérie receveuse reste F- et peut intégrer le nouveau fragment à son génome. L'ADN chromosomique est figuré en bleu, l'ADN du facteur F en rouge.

2.1.3. La transduction

Les bactériophages sont également un moyen efficace de transfert d'ADN. Il existe deux types de cycles : le cycle productif et le cycle lysogénique. Pendant le cycle productif, le génome du phage est exprimé et répliqué grâce à la machinerie bactérienne. Puis les nouveaux phages sont assemblés et enfin libérés à la suite d'une lyse de la cellule provoquée par le lysozyme d'origine phagique. Les cycles lysogéniques impliquent des phages dits tempérés, car ils peuvent établir des relations de longue durée, éventuellement réversibles avec la bactérie. Au cours de la lysogénie, l'ADN phagique est intégré au chromosome bactérien sous la forme d'un prophage incapable de se répliquer de façon autonome. Certains phages comme le phage λ, s'intègrent dans le génome d'*E. coli* toujours au même endroit alors que d'autres présentent plusieurs sites d'intégration (comme par exemple le phage P2). Ce mode d'intégration est appelé recombinaison site spécifique. Mu-1, quant à lui s'intègre totalement au hasard, il s'agit donc de recombinaison dite illégitime. Nous présenterons par la suite ces différents modes d'intégration dans le génome. Les fonctions virales sont alors réprimées par une protéine synthétisée par le bactériophage. L'induction (ou passage d'un cycle lysogénique à un cycle productif) est en général induite par un stress tel que les rayons UV, par exemple.

Il existe deux types de transduction, une transduction généralisée et une transduction localisée, dont les mécanismes reposent sur l'existence des deux types de réplication (cycle productif et cycle lysogénique).

2.1.3.1 La transduction généralisée

Lors du cycle productif, l'ADN bactérien peut être segmenté par une désoxyribonucléase d'origine phagique. Au moment de l'assemblage du nouveau virus, un fragment d'ADN bactérien peut être encapsidé par erreur à la place de l'ADN phagique pour peu qu'il soit d'une taille appropriée. Le phage ayant incorporé de l'ADN bactérien ne peut plus se répliquer (phage défectif), mais il peut transférer de l'ADN bactérien à une bactérie réceptrice (les étapes d'adsorption et de pénétration ne sont pas modifiées chez un phage défectif). Cette transduction est qualifiée de généralisée ou de non spécifique car elle concerne tous les fragments d'ADN (chromosomique ou extra-chromosomique) pourvu que leur taille soit compatible avec une encapsidation. Le fragment d'ADN transféré par le phage peut alors s'intégrer par recombinaison homologue et toute la descendance de la bactérie réceptrice portera l'ADN transféré. On dit que la transduction est complète. Le fragment

d'ADN peut également rester libre, il ne sera pas répliqué, mais les gènes transmis sont fonctionnels et peuvent être exprimés. Lors de la multiplication bactérienne, seule une des deux cellules filles acquiert le fragment d'ADN transféré et, au cours des divisions successives, ce fragment sera peu à peu dilué. On dit que la transduction est abortive.

2.1.3.2 Transduction localisée

La transduction localisée est réalisée par des phages tempérés. Elle correspond à une excision anormale du prophage. Lorsque le répresseur est inactivé, un cycle productif succède à un cycle lysogénique. Avec une fréquence de l'ordre de 10^{-6}, l'excision est anormale et on obtient la libération d'une molécule d'ADN hybride constituée d'un fragment d'ADN phagique et d'un fragment d'ADN bactérien. Ce fragment d'ADN bactérien est adjacent à la zone d'intégration du prophage d'où le nom de transduction localisée.

Le phage λ est un bon exemple de ce type de transduction. Il s'intègre toujours sous forme de prophage entre les régions *gal* et *bio* du chromosome de *E. coli*.

Lorsque le cycle lytique du phage λ se déclenche, si l'excision est anormale, l'ADN libéré est formé d'ADN viral et des régions *gal* ou *bio* (selon que l'excision se décale d'un côté ou de l'autre).

La transduction assuré par le phage λ est donc restreinte aux marqueurs *gal* et *bio*. Lorsque le phage possède plusieurs sites d'intégration (*Enterobacteria* phage P2, *Enterobacteria* phage P22, ...) la transduction localisée devient non spécifique. Avec le phage Mu-1, qui peut s'intégrer totalement au hasard sur une molécule d'ADN bactérien, de nombreux gènes bactériens peuvent être transférés et la transduction localisée a des allures de transduction généralisée.

2.2 Les mécanismes d'intégration de l'ADN étranger dans le génome bactérien

Que ce soit à la suite d'une transformation, d'une conjugaison ou d'une transduction, l'ADN étranger à la cellule, peut s'incorporer dans le génome, par recombinaison homologue, site spécifique ou illégitime. Lorsqu'on s'intéresse à la diversité génomique au niveau de l'espèce bactérienne, la recombinaison induit une hétérogénéisation dans le cas

où il y a incorporation d'ADN étranger à l'espèce. Il s'agit alors d'un échange de matériel génétique inter-espèce.

Il n'existe encore aucune définition universellement admise de l'espèce bactérienne, seules des recommandations ont été faites. Une espèce est constituée par sa souche type et par l'ensemble des souches considérées comme suffisamment proches de la souche type pour être inclus au sein de la même espèce. Le critère de similarité le plus couramment admis est un pourcentage d'hybridation ADN-ADN supérieur à 70% (Wayne, Brenner et al. 1987; Stackebrandt, Frederiksen et al. 2002). Avec la multiplication des génomes complets bactériens, il est maintenant possible de corréler le pourcentage d'hybridation ADN-ADN, l'identité nucléotidique moyenne des séquences communes et le pourcentage d'ADN conservés entre deux bactéries. Goris et collaborateurs ont effectués ces comparaisons sur 28 génomes bactériens, provenant de six groupes différents : *Bacillus cereus*, le genre *Burkhlderia*, les espèces *E. coli/Shigella*, le genre *Pseudomonas*, *Shewanella* et les espèces *Streptococcus agalactiae*. Ces trois estimateurs sont très fortement positivement corrélés entre eux. D'après cette étude, le seuil de 70% d'hybridation ADN-ADN correspond à une identité nucléotidique moyenne de 95%, et à une proportion d'ADN conservés de 69% (Goris, Konstantinidis et al. 2007).

Par contre, si la recombinaison implique un échange d'ADN intra-espèce, elle sera à l'origine d'une homogénéisation de la séquence du point de vue de l'espèce dans son ensemble. Mais si on s'intéresse à un unique individu par rapport à la population, ce même cas pourra être décrit comme une hétérogénéisation. Cela dépend donc du référentiel (l'individu ou l'espèce) ainsi que de la nature de la recombinaison (intra ou inter espèces).

Nous verrons dans ce chapitre que la recombinaison site spécifique et la recombinaison illégitime, ne nécessitant pas ou très peu d'homologie, engendrent, dans le cas général, une hétérogénéisation des séquences puisqu'elles permettent l'acquisition d'ADN étranger à l'espèce.

Le cas de la recombinaison homologue est nettement plus complexe. Par exemple, Brochet et collaborateurs étudient les traces de recombinaison dans huit génomes (répartis dans quatre complexes clonaux) de *Streptococcus agalactiae*, responsable d'infections néonatales. Ils ont mis en évidence un complexe clonal partageant un grand nombre de

régions conservés avec les autres complexes clonaux. En plus des opérons codant les ARN ribosomiques, ils décrivent 3 régions hautement conservées dans les huit souches. Plus précisément elles sont davantage conservées entre les isolats provenant d'intestins humains qu'avec les isolats provenant d'animaux. Les auteurs font l'hypothèse que ces régions comportent des loci présentant un avantage sélectif quant à l'adaptation à l'hôte humain. Elles auraient été acquises lors d'un transfert horizontal par recombinaison illégitime puis distribuées par recombinaison homologue entre les souches. Ainsi, dans les organismes sujets à un fort taux de recombinaison homologue, la sélection naturelle des loci présentant un avantage sélectif produirait un balayage sélectif et donc une réduction de la diversité génétique autour de ces loci (Brochet, Rusniok et al. 2008). La première étude démontrant cette possibilité a été établi chez *E. coli* sur le HPI (« High Pathogenicity Island »). Il s'agit du principal îlot de pathogénicité d'*E. coli*, il contient plusieurs gènes de virulence et peut à lui seul transformer une souche inoffensive en souche pathogène. Les auteurs, en comparant les phylogénies de plusieurs gènes du MLST, du HPI et de ses séquences flanquantes dans les souches de la collection ECOR sont parvenus à expliquer la répartition du HPI parmi les souches d'*E. coli* pathogènes extraintestinales (Schubert, Darlu et al. 2009). Le HPI chez *E. coli* s'est inséré au locus *asnT* comme chez les *Yersiniae.* Cette acquisition s'est fait par l'intermédiaire d'un plasmide conjugatif suivi par une recombinaison site spécifique. Cette étape constitue une diversification par acquisition de nouvelles fonctions provenant d'un autre genre bactérien. Nous verrons dans le prochain chapitre que la recombinaison site spécifique nécessite de courtes séquences répétées qui semblent avoir été perdues chez *E. coli*. La recombinaison site spécifique ne peut donc pas expliquer la dissémination du HPI parmi les différentes souches de cette espèce. Les auteurs ont montré que l'hypothèse la plus plausible d'après leurs observations est que la dispersion du HPI s'est effectuée grâce à un plasmide conjugatif suivi par une étape de recombinaison homologue dans les séquences flanquantes partagées par toutes les souches. Cette homogénéisation des souches d'*E. coli* entre elles semble s'être faite très rapidement car les séquences du HPI sont très proches les unes des autres. Actuellement plus de 80% des souches d'*E. coli* pathogènes extraintestinales portent le HPI.

Nous pouvons donc conclure que, dans le cas général, la recombinaison homologue induit une hétérogénéisation au niveau de l'individu, mais une homogénéisation de la population lorsque l'on s'intéresse à l'histoire évolutive d'une espèce bactérienne.

L'un des résultats de la recombinaison homologue est la conversion génique. Nous en discuterons les mécanismes et essaierons d'éclaircir les ambiguïtés de vocabulaire qui accompagnent cette notion.

2.2.1 La recombinaison homologue

Dans ce cas, une homologie de séquence est requise. La longueur de cette homologie de séquence est variable selon les mécanismes en jeu. La recombinaison est un processus très complexe qui implique plus d'une trentaine de gènes chez *E. coli* (Smith 1988; Lloyd and Low 1996). Parmi ceux-ci on peut citer *recA* qui contrôle l'appariement des homologues, les autres gènes *rec*, les gènes *ruv* impliqués dans l'échange des brins. Plusieurs modèles ont été proposés pour décrire la recombinaison homologue. Tous ont en commun l'invasion de la terminaison 3'-OH d'une séquence d'ADN simple brin par un deuxième ADN double brins (Smith 1988). L'appariement initial peut se produire à n'importe quelle position de la région homologue. La réaction d'échange entre brins commence quand les deux molécules sont alignées et que l'extrémité de l'ADN est libre. S'établit alors une des étapes clef de la recombinaison homologue : la jonction des deux molécules (jonction de Holliday) formant une région heteroduplex (Holliday 1964). Si les deux molécules ne sont pas identiques le système de réparation des mésappariements (SRM) que nous avons présenté dans le chapitre précédent peut corriger le produit de la recombinaison et donc en affecter le résultat. La résolution de la jonction se fait par deux nouvelles coupures simple brin ; suivant la position de ces coupures on aura ou non échange de marqueurs flanquants.

S'il y a coupure puis religation des deux brins qui ont subi l'échange il n'y aura pas de crossing-over visible, seul un très petit fragment sera échangé, sur un seul brin, ce phénomène est appelé conversion génique. La conversion génique a d'abord été mise en évidence chez la levure. Le locus *MAT* qui détermine le sexe de *Saccharomyces cerevisiae*, subit une modification grâce au remplacement d'un segment d'ADN (Y), par un autre. Ce fragment est apporté par deux sites donneurs non exprimés *HMLα* et *HMRa* respectivement homologues aux allèles *MATα et MATa.* Cette recombinaison s'effectue sans modification

des sites donneurs et le plus souvent sans échanges des marqueurs flanquants, donc dans ce cas sans crossing-over (Haber 1998). Le premier modèle proposé était le DSBR (cassure double brins et réparation), mais certaines incohérences avec les observations ont permis de proposer un meilleur modèle : le SDSA (« synthesis-dependent strand annealing » ou synthèse dépendante du brin apparié). Les loci *MAT* et *HML/HMR* se situent sur le même chromosome linéaire (le chromosome III).

En revanche si les coupures ont lieu sur les deux brins qui n'ont pas subi l'échange, le crossing-over se verra par échange des marqueurs flanquants. Ces deux possibilités coexistent dans les modèles de Holliday, de Meselson-Radding (Meselson and Radding 1975) et de DSBR (Szostak, Orr-Weaver et al. 1983; Kowalczykowski, Dixon et al. 1994). Certains modèles comme le SDSA (Allers and Lichten 2001) produisent exclusivement de la conversion génique (Fig 6). Tous ces modèles ont été développés pour les eucaryotes.

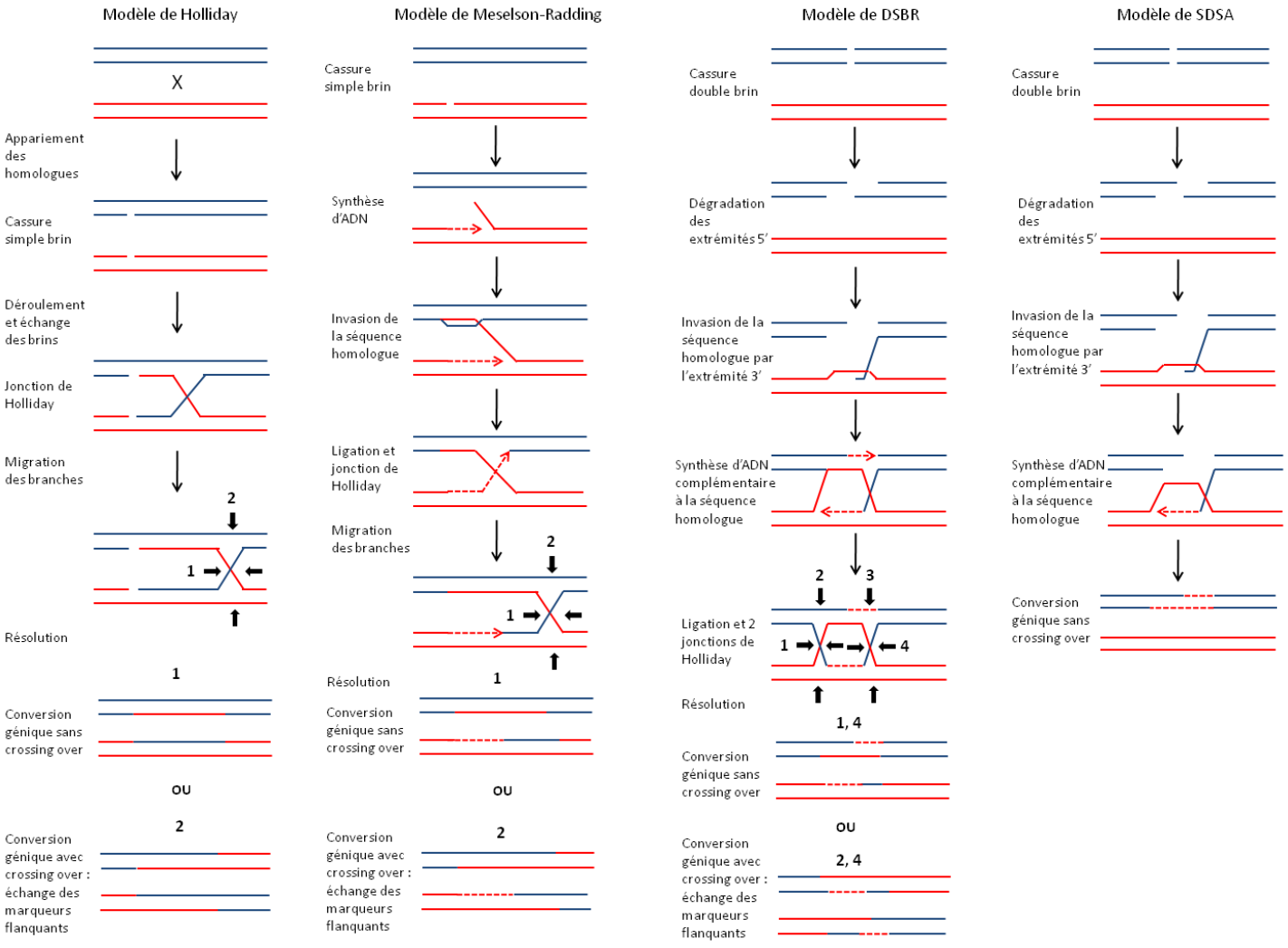

Fig. 6 : Modèles moléculaires de recombinaison permettant d'expliquer la conversion génique chez les eucaryotes. Présentation succincte des modèles de Holliday, Meselson-Radding, DSBR, et SDSA. Les séquences homologues sont présentées de deux couleurs différentes. La synthèse d'ADN est symbolisée par une flèche, et l'ADN nouvellement synthétisé par des pointillés.

Chez les procaryotes, le DSBR semble un bon modèle pour expliquer la recombinaison homologue, dans le cas où il y aurait cassure double brins de l'ADN (Cromie, Connelly et al. 2001). La mise en oeuvre du modèle SDSA produisant uniquement de la conversion génique, c'est à dire sans crossing-over, n'a jamais été démontrée chez *E. coli.* Par contre il a été proposé un modèle de recombinaison homologue permettant de rétablir une fourche de réplication interrompue par une cassure du brin direct (Fig. 7). Ce modèle implique également une jonction de Holliday qui est résolue par RuvC (Cromie, Connelly et al. 2001). Les auteurs distinguent les cassures double brins et les extrémités libres. Les premières sont deux extrémités proches, se faisant face, pouvant être observées lorsque la cellule est soumise à certains agents chimiques ou à certaines radiations. Les secondes ont une origine diverse : elles peuvent provenir d'un fragment linéaire introduit dans la cellule, une fourche réplicative interrompue... Ces extrémités libres sont dans tous les cas soit une seule extrémité double brins, soit deux extrémités indépendantes l'une de l'autre.

Comme nous avons pu le voir, la jonction de Holliday est un intermédiaire très souvent présent dans les modèles de recombinaison homologue. Elle peut être résolue de deux façons différentes : avec ou sans échanges des séquences flanquantes. Il a été montré que ce choix ne s'effectuait pas au hasard et qu'en fait la jonction de Holliday était polarisée. Chez *E. coli*, les propriétés du complexe RuvABC entraineront la formation d'un seul crossing-over lors de la réparation d'une cassure double brins (choix 1 et 3 pour résoudre le modèle DSBR de la Fig. 12). Par contre, une fourche réplicative interrompue à cause d'une cassure du brin direct sera résolue sans crossing-over (Fig. 7) (Cromie, Connelly et al. 2001).

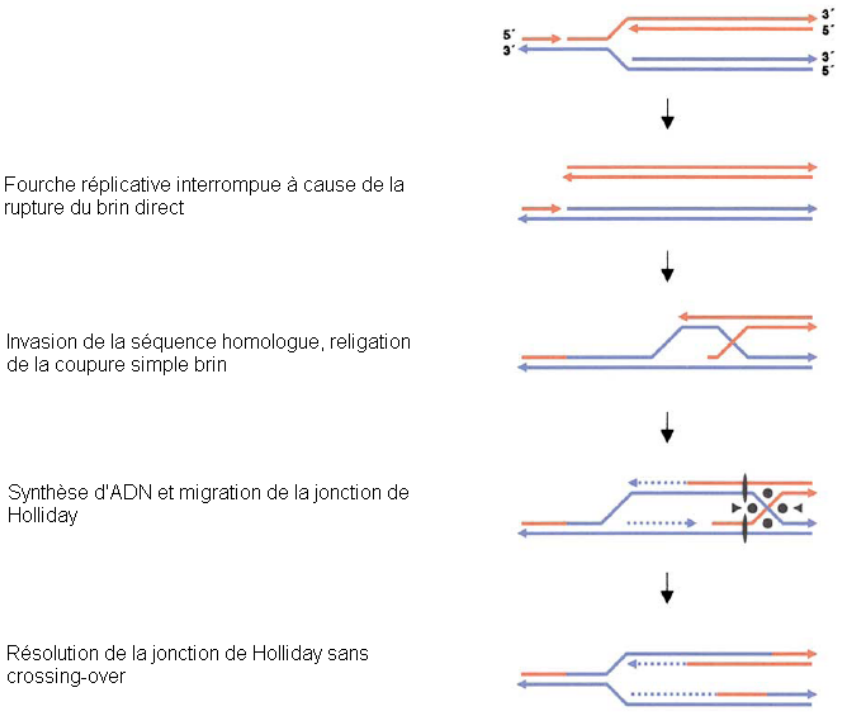

Fig. 7 : Modèle moléculaire de recombinaison homologue permettant d'expliquer la réparation des extrémités libres causées par une cassure du brin direct lors de la réplication chez *E. coli*. Les séquences homologues sont présentées de deux couleurs différentes. La synthèse d'ADN est symbolisée par une flèche, et l'ADN nouvellement synthétisé par des pointillés (d'après la figure 3C de (Cromie, Connelly et al. 2001)).

Chez les procaryotes, il a été prouvé qu'il existait une recombinaison dépendante de la réplication (Kogoma 1997). Un nouveau modèle émerge (BIR, « Break-induced replication ») : la réplication induite par une cassure (Fig. 8). La boucle crée par l'invasion de l'extrémité 3' du brin homologue sera suivi par l'établissement d'une fourche de réplication unidirectionnelle qui compléterait les molécules. Chez *E. coli*, un fragment d'ADN introduit par conjugaison, transformation ou transduction donnerait lieu à deux invasions indépendantes des brins homologues établissant alors chacun une fourche de réplication copiant les 5Mb restant du chromosome circulaire.

Il semble exister une forte dépendance entre la recombinaison homologue et la réplication (Haber 2007).

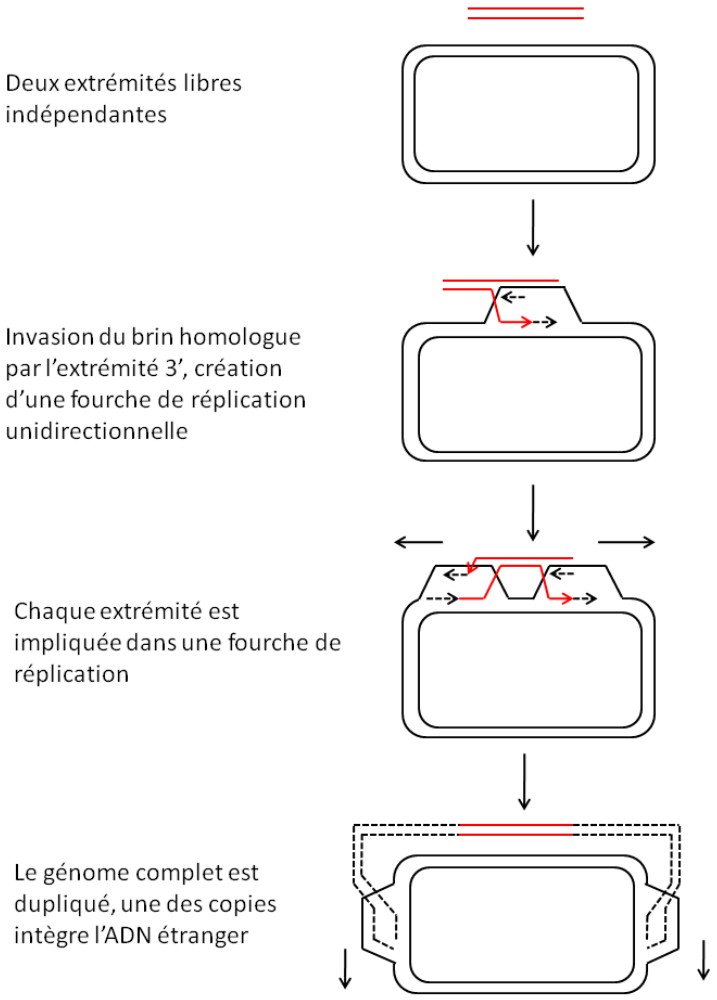

Fig. 8 : Double "break-induced replication". Modèle de recombinaison couplé à la réplication chez *E. coli* (d'après la Fig. 7 de (Motamedi, Szigety et al. 1999)) .

2.2.2 La recombinaison site spécifique

La recombinaison site spécifique décrit un processus de recombinaison très spécialisé qui implique un échange réciproque entre des sites bien définis. Les sites les plus simples ont une longueur comprise entre 20 et 30 nucléotides et sont composés de deux courtes répétitions inversées auxquelles se lie la recombinase entourant une zone centrale où se produit le crossing-over. Les principales étapes de ce mode de recombinaison sont : la reconnaissance des sites par la recombinase, la formation du complexe synaptique, le clivage de l'ADN, l'échange des brins, puis la religation de la molécule et enfin la rupture du complexe synaptique permet la libération des recombinants. L'énergie de la liaison phosphodiester est conservée après le clivage et est réutilisée pour la ligation. C'est

pourquoi, il n'y a pas nécessité de requérir à un cofacteur fournisseur d'énergie tel que l'ATP dans cette réaction (Grindley, Whiteson et al. 2006). Il a été décrit des sites plus complexes. Dans ce cas des séquences supplémentaires sont présentes, la plupart étant des sites de liaison protéique. Ils ont alors des fonctions de régulation ou structurelles (de stabilisation, de détermination de la direction de la recombinaison par exemple)

Selon la disposition initiale des sites de recombinaison, la recombinaison sites spécifiques peut engendrer trois résultats : l'intégration, l'excision ou l'inversion d'une molécule d'ADN. L'intégration résulte de la recombinaison entre deux molécules séparées (dont au moins une est circulaire). Par contre, lorsque les sites sont localisés sur le même chromosome, le résultat de la recombinaison sera une excision entre des sites ayant la même orientation, et une inversion lorsque l'échange se produit entre des sites dont l'orientation est inversée.

Il existe deux familles de recombinases, les tyrosines recombinases, appelées aussi intégrases et les sérines recombinases ou résolvases/invertases. Leur nom provient de l'acide aminé nucléophile qui sera lié à l'ADN pendant l'échange des brins d'ADN. Les sérines recombinases coupent les deux brins d'ADN aux deux sites d'échange, tous les brins sont clivés avant que l'échange soit initié. Les tyrosines recombinases, quant à elles, clivent un brin à la fois dans chaque duplex. Il doit être joint à son nouveau partenaire avant que le second brin soit coupé, ce qui produit une jonction de Holliday (Grindley, Whiteson et al. 2006). Ces dernières sont celles qui sont le plus souvent rencontrées chez les procaryotes, l'exemple le plus connu est l'intégration du phage λ dans le génome d'*E. coli* (Weisberg 1983). Mais on peut citer également certains transposons (comme le Tn*7*) dont l'intégration chromosomique se fait au site *attTn7* situé en aval (extrémité 3') du gène *glmS* (codant la glucosamine 6-phosphatase) (Peters and Craig 2001).

Le phage λ dans le cas où il suit un cycle lysogénique est en dormance sous la forme d'un prophage intégré dans le génome bactérien. La recombinaison site spécifique permettant cette intégration est une des plus connue. Le seul et unique site dans le chromosome d'*E. coli* où peut s'intégrer le phage λ est situé entre les opérons *gal* (galactose) et *bio* (biotine). Il est appellé *attB* pour site d'attachement bactérien. Ce site est constitué de seulement 30 nucléotides, la séquence centrale en comportant 15. Le site de recombinaison présent sur le phage (*attP*) est plus complexe. Il contient la même séquence de 15 nucléotides bornée par des sites de liaison protéiques appelés bras P et P'. L'intégrase (Int) phagique et le facteur d'intégration dans l'hôte (IHF) bactérien se lient aux sites P et P' pour former un complexe dans lequel les sites *attB* et *attP* seront alignés. Les brins sont alors coupés, échangés et joints à nouveau. Le résultat de la recombinaison est que le prophage intégré est flanqué par deux sites d'attachement modifiés (*attR* et *attL*). Xis (excisionase) et Fis sont deux protéines qui se lient à des sites spécifiques sur le bras P et qui, avec Int et IHF vont former un complexe autour de *attR* et *attL* alignés et catalyser la réaction inverse permettant ainsi l'excision du prophage (Fig 9).

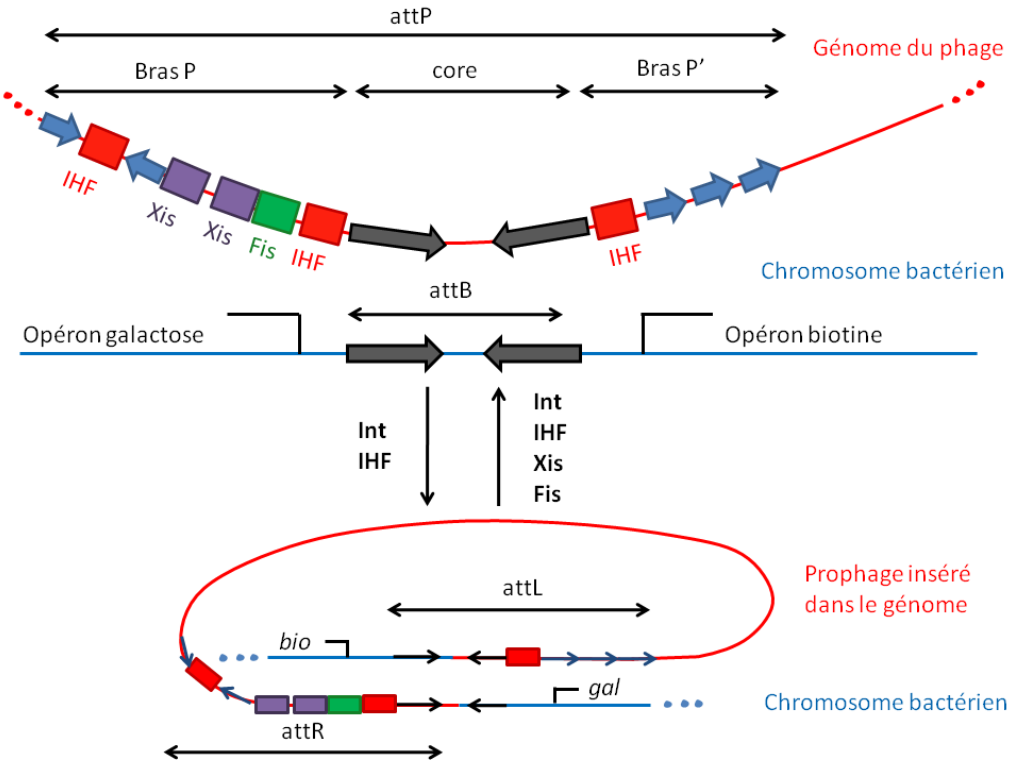

Fig. 9 : Réaction de recombinaison site spécifique entre le phage λ et le chromosome bactérien. Int et IHF sont nécessaires à l'intégration de l'ADN phagique dans le génome bactérien. Int, IHF, Xis et Fis sont requises lors de son excision.

2.2.3 La recombinaison illégitime ou RecA indépendante

Comme nous avons pu le voir précédemment, la recombinaison homologue est RecA dépendante, la recombinaison site spécifique implique quant à elle des recombinases particulières. La recombinaison illégitime diffère des deux premières car elle peut se faire à de nombreux sites entre des séquences ne présentant pas d'homologie ou présentant une homologie courte. Elle ne nécessite pas la protéine RecA. Les réarrangements qui mettent en œuvre ces mécanismes peuvent être des délétions, des inversions, des duplications et la translocation de certains phages (Michel 1999). La recombinaison illégitime regroupe en fait différents mécanismes moléculaires selon le contexte. Nous présenterons les différentes formes de recombinaison illégitime et ce que l'on en sait aujourd'hui. Elle semble en général assez peu fréquente chez *E. coli*, pourtant, sa fréquence augmente de manière importante lorsque la cellule est soumise à des agents préjudiciables à l'ADN comme les rayons ultraviolets.

2.2.3.1 La recombinaison illégitime entre de courtes séquences répétées

Il peut s'agir de recombinaison entre de courtes séquences homologues. Chez *E. coli*, elle serait mise en œuvre entre des séquences homologues comportant moins de 20 pb. Au dessus, ce serait la recombinaison homologue (RecA dépendante) qui interviendrait, cette limite étant dépendante des organismes. Les courtes séquences homologues sont dans ce cas toujours des répétitions intrachromosomiques relativement proches l'une de l'autre. Il a également été observé que la présence de palindromes (formant des structures secondaires en tige-boucle) augmente la fréquence de la recombinaison illégitime. Les protéines impliquées ne sont pas encore toutes connues, mais il semblerait que l'exonucléase RecJ intervienne. L'hélicase RecQ paraît quant à elle réprimer cette forme de recombinaison illégitime. DnaB, une autre hélicase semble elle aussi jouer un rôle (Shiraishi, Imai et al. 2005). Shiraishi et collaborateurs ont montré, grâce à l'induction de λ dans des bactéries incapables d'exciser le prophage, que lorsque l'hélicase Rep est inactivée, la recombinaison illégitime augmente d'un facteur 9 sans irradiation aux UV (Shiraishi, Imai et al. 2005). Or, lorsque cette enzyme est non fonctionnelle, il a été observé que la réplication était plus fréquemment interrompue. Plusieurs modèles ont été proposés pour expliquer ces observations. Tout d'abord, le modèle appelé « glissement de la réplication » : une pause lors de la réplication entrainerait une dissociation du brin néo-synthétisé et son appariement

avec une autre séquence répétée. Ceci entraine une délétion lorsque le site répété qui s'apparie se situe en aval du site de pause réplicative, et une duplication lorsque celui-ci se situe en amont. S'il y a un mésappariement entre les deux répétitions appariées, le système de réparation des mésappariements pourrait faire avorter la délétion en excisant le fragment comportant le mésappariement (Bzymek and Lovett 2001). Le deuxième modèle, modèle d'hybridation simple brin (SSA : « single strand annealing ») requiert une cassure double brins. Des exonucleases raccourcissent alors la molécule, lui permettant de se réapparier avec la séquence complémentaire appartenant à l'autre répétition. Ce modèle permet d'expliquer la délétion d'un fragment situé entre deux séquences répétées. Chez *E. coli*, ce mécanisme semble inefficace à cause de la dégradation très efficace de l'ADN double brins libre par RecBCD, sauf lorsque des palindromes seraient impliqués (Bzymek and Lovett 2001). On peut citer également un troisième modèle qui met en œuvre des répétitions en tandem présentes dans deux réplicons d'ADN circulaires. Deux fragments répétés recombineraient entre eux pour engendrer un dimère qui se rescinderait à un autre endroit et pourrait produire dans un cas une augmentation du nombre de répétitions et une diminution dans l'autre (Bzymek and Lovett 2001). Il a également été proposé que cet échange de chromosomes frères puisse s'effectuer lors de la réplication et être alors couplé à un glissement (correspondant au premier modèle).

2.2.3.2 La recombinaison illégitime associée à des éléments sites spécifiques

Il semblerait qu'en cas de stress important, l'activité des transposons puisse s'intensifier. Ce processus pourrait faire intervenir la recombinaison illégitime (Michel 1999). La transposition intramoléculaire induirait des réarrangements tels que des délétions, des insertions, des inversions. Il a été également observé que les recombinases impliquées dans la recombinaison site spécifique présentée ci-dessus peuvent se tromper de site d'action (on parle de pseudo-site ou de quasi-site). Par exemple, une erreur lors de la transposition de certain transposon (Mu, Tn*10*) induit une délétion d'un morceau de séquence adjacent au transposon. Dans ces cas, on parle également de recombinaison illégitime (Michel 1999).

2.2.3.3 La recombinaison illégitime entre des séquences non homologues

Certains réarrangements peuvent se produire sans qu'il n'y ait aucune homologie de séquence. Par exemple, les gyrases (topoisomérases II) coupent l'ADN double brins pour

modifier son enroulement puis religuent les fragments. Ces enzymes peuvent, lorsqu'elles sont liées à deux molécules d'ADN différentes, ou à deux endroits différents de la même molécule d'ADN, échanger leurs sous-unités entre elles ce qui entraine un réarrangement après liaison des deux fragments. Les topoisomérases de type I, permettent le relâchement du superenroulement de l'ADN par coupure transitoire et ligation d'un seul brin. Il a été proposé qu'une topoisomérase I puisse lier deux extrémités de deux fourches de réplication différentes bloquées aux sites de terminaison ce qui entrainerait la délétion d'un fragment d'ADN sous forme circulaire (Michel 1999).

Bien que la recombinaison illégitime semble peu efficace chez *E. coli*, l'analyse récente de Touzain et collaborateurs suggère que ce mécanisme est responsable d'une grande part de la diversité génomique observée chez *E. coli*, il serait à l'origine de 44% des petits segments variables (20 – 500 pb) dans un alignement de 20 souches et 54% de ceux observables dans un alignements de 5 souches du groupe B2 (Touzain, Denamur et al. 2010).

Il est intéressant de remarquer que quatre des mécanismes proposés : le glissement de la réplication, le SSA, l'échange entre chromosomes frères couplé au glissement réplicatif et les erreurs des topoisomérases sont tous plus fréquemment observés en cas de blocage de la réplication.

2.2.4 La conversion génique

La conversion génique est définie comme étant un transfert non réciproque d'information génétique entre plusieurs copies d'un gène. Elle semble avoir un rôle important dans l'évolution des familles multigéniques. Dans ce cas elle s'effectue entre des séquences homologues appartenant au même chromosome bactérien. Par exemple pour augmenter la diversité antigénique : dans ce cas, des fragments de séquences ADN (cassettes de gènes) sont transférés par conversion génique à partir de gènes non exprimés vers le gène homologue exprimé (Deitsch, Moxon et al. 1997). Nous pouvons citer deux exemples tirés de la littérature : PilE sous-unité indispensable au pilus de *Neisseiria gonorrhoea*, bactérie pathogène de l'homme et la VMP (« Variable Major Protein »), principal antigène reconnu lors de l'infection à *Borrelia hermsii*, qui est une bactérie spiralée du groupe des spirochètes pathogène des mammifères. Dans le premier cas, il a été montré que des fragments non exprimés homologues à *pilE* pouvaient recombiner et ainsi fournir un

nouveau variant. Cette recombinaison intragénomique est RecA dépendante et non réciproque : il s'agit bien de conversion génique. Dans le cas de la VMP, il y a plusieurs copies du gène. Elles se trouvent à l'extrémité de plusieurs plasmides linéaires. La copie exprimée est localisée sur un plasmide différent de ceux qui portent les copies silencieuses, elle est dans ce cas sous la régulation d'un promoteur actif. Le changement de sérotype s'effectue en remplaçant la copie exprimée par une des copies silencieuses. La copie donneuse n'est pas modifiée par le transfert. Le résultat semble donc équivalent à ce qu'on obtiendrait par conversion génique. Pourtant il semblerait qu'il s'agisse plutôt d'une translocation duplicative, suivi d'une délétion (Donelson 1995). Un second mécanisme permettant de modifier l'antigène peut alors intervenir. Dans ce cas, un fragment provenant d'un pseudogène homologue situé directement en amont du gène exprimé vient remplacer le fragment homologue dans le gène exprimé (Donelson 1995). Il s'agit donc d'une conversion génique partielle.

Nous avons pu voir précédemment que la conversion génique est un des dérivés de la recombinaison homologue. Nous avons, à cette occasion, discuté du fait que la recombinaison homologue homogénéisait les séquences entre elles. Or dans l'exemple ci-dessus, il s'agissait de diversification antigénique. En fait, il s'agit encore d'une différence de point de vue : la cellule, grâce à ces mécanismes, exprime effectivement un nouvel antigène, pourtant, si on compare les différentes copies du gène en question (exprimées et non exprimées) entre elles, il y a bien homogénéisation.

Un autre exemple est l'évolution « concertée » des ARN ribosomiques (Hashimoto, Stevenson et al. 2003). Les trois ARN ribosomiques procaryotes (16S, 23S et 5S) sont, en général, organisés en opérons (les opérons *rrn*). Ces derniers sont souvent présents en plusieurs exemplaires. Chez *E. coli*, par exemple, il y a 7 opérons *rrn*. Il a été proposé que ces copies étaient homogénéisées entre elles par conversion génique (Hashimoto, Stevenson et al. 2003). Par ce mécanisme, les paralogues deviennent plus similaires entre eux qu'avec les orthologues. Une mutation avantageuse peut alors être propagée aux différents membres de la famille par conversion génique (Santoyo and Romero 2005).

L'ARNr 16S, très conservé, est souvent utilisé pour établir des phylogénies entre espèce éloignée. Le fait que ces copies à l'intérieur du génome soit très proches est, dans ce

cas un avantage, car il n'y pas de risque de confondre orthologues et paralogues. Par contre, ce gène est inutilisable pour construire une phylogénie intra-espèce pour cette même raison.

Différents modèles ont été proposés pour expliquer le cas particulier de recombinaison homologue qu'est la conversion génique. Chez les eucaryotes, le modèle le plus communément accepté est le DSBR (Réparation de coupure double brins : « Double Strand Break Repair ») (Fig 6). Chez *E. coli*, il semblerait que ce mécanisme soit très peu efficace à cause de la dégradation de l'ADN double brins libre par RecBCD, le modèle émergent est donc dans ce cas le BIR (break-induced replication), qui correspond en fait à deux demi-DSBR couplés à deux fourches de réplication (Fig. 8). Les protéines RuvAB, RecG et RadA semblent être les principales responsables de la migration de la jonction de Holliday chez *Rhizobium etli*, une bactérie s'associant en symbiose avec les racines des légumineuses (Castellanos and Romero 2009). De plus, chez cet organisme, les évènements de conversion génique observés sont toujours couplés à un crossing-over. Il semblerait, que dans le cas où un fragment linéaire présentant deux extrémités libres s'insère, deux crossing-over (un de chaque coté) seraient mis en œuvre.

La communauté des généticiens des populations utilisent ce même terme de conversion génique pour désigner autre chose. Ils décrivent ainsi un ensemble de substitutions localisées qui aurait été transporté d'un génome à un autre (Fig. 10). Les généticiens des populations s'intéressent à l'observation de ces séquences insérées et non au mécanisme sous-jacent. Or nous avons pu conclure au fur et à mesure de ce chapitre, que dans les séquences bactériennes un fragment inséré dans un autre était, dans le cas général, issu d'un double crossing-over.

```
.CT..T...GTGGCGT.C.A........T.............................
.AG..C...AAATGCG.G.T........A.............................
.CT..T...GTGGCGT.C.A........T.............................
.CT..T...GTGGCGG.C.A........A.............................
```

Fig. 10 : Alignement présentant une succession de substitutions appelées conversion génique par les généticiens des populations. Les points figurent les nucléotides non polymorphes, le segment inséré est surligné en jaune.

3 Une phylogénie d'*E. coli* est-elle possible ?

Afin de comprendre l'histoire évolutive de l'espèce *E. coli* et comment elle s'est adaptée à ses différentes niches écologiques, il est nécessaire de statuer sur la structuration globale de la population. Est-elle plutôt panmictique (recombinaison forte par rapport à la mutation) ou clonale (recombinaison faible par rapport à la mutation) ?

3.1 Une population clonale ?

Dés 1947, Kaufman décrit la diversité antigénique de l'espèce (Kauffmann 1947). La technique utilisée, le sérotypage, consiste à caractériser les souches bactériennes selon leurs antigènes. Cent soixante treize antigènes O (somatiques, c'est à dire situé dans la paroi), 80 antigènes K (capsulaires) et 56 antigènes H (flagellaires) ont été décrits. Dans les années 1970, Orskov et collaborateurs montrent que ces antigènes ne sont pas associés de façon aléatoire. Par contre, ils n'observent pas de structuration géographique de la population (Orskov, Orskov et al. 1976). Les auteurs émettent alors l'hypothèse de lignées stables (clones) dont les gènes subissent peu de recombinaison. A peu près à la même époque, la technique d'électrophorèse enzymatique multilocus (MLEE : « multilocus enzyme electrophoresis ») a été développée. Cette technique permet de caractériser les isolats selon la mobilité électrophorétique relative d'un grand nombre d'enzymes de ménage hydrosolubles. Les différents allèles à chaque locus définissent un type électrophorétique. De plus il est possible de construire un dendrogramme représentant les relations entre les différents isolats à partir d'une matrice calculée à l'aide des différences deux à deux entre les types électrophorétiques. Milkman en 1973 analyse avec la technique du MLEE, 5 enzymes extraites de 839 clones d'*E. coli*. Il observe un allèle largement dominant par rapport aux autres. Il décrit également une variabilité importante entre les clones isolés d'un même hôte, il suggère donc que la recombinaison entre eux est importante. Le but de cette étude était de vérifier que la variabilité génétique d'une espèce haploïde peut être aussi forte que dans une espèce diploïde malgré l'absence de l' « overdominance » (sélection pour l'hétérozygote). Puisqu'un allèle est largement prédominant dans la population, et cela dans différents hôtes indépendants, cela favoriserait une sélection forte pour cet allèle (Milkman 1973). Plusieurs années plus tard, Selander et Levin étudient à nouveau la diversité génétique d'*E. coli* à l'aide de la technique du MLEE et montre ainsi que les conclusions de Milkman étaient fausses. Ils utilisent 20 enzymes et 109 clones provenant d'origines

diverses. Même s'ils observent une diversité génétique deux fois supérieure à celle décrite par Milkman, ils ne comptent que 98 types électrophorétiques différents parmi les 109 clones. Ils en concluent que les loci ne sont pas indépendants les uns des autres (il y a un déséquilibre de liaison) et que par conséquent, le modèle neutre utilisé par Milkman n'est pas approprié. Pour eux, les légères différences observées entre les clones sont davantage expliquées par les mutations que par les recombinaisons. Celles-ci seraient rares, la structure de la population clonale et la variabilité génétique observée proviendrait principalement des mutations et de processus de sélection périodique (Selander and Levin 1980). On parle de sélection périodique lorsque l'allèle prédominant envahit la population, ce qui diminue drastiquement sa diversité et lui confère une certaine stabilité (Atwood, Schneider et al. 1951). Ce fut la naissance d'un intense débat scientifique, qui perdure encore aujourd'hui : la structure de la population d'*E*. *coli* est-elle clonale ou panmictique ?

3.2 Une population panmictique ?

Dans les années 1980, les techniques de séquençage ont permis d'étudier la signature de la recombinaison dans les gènes. Milkman et Crawford, dés 1983, ont identifié des substitutions regroupées sous la forme de « cluster » dans l'opéron *trp*. Ils ont interprété cette observation comme de probables évènements de recombinaison (Milkman and Crawford 1983). Puis, plusieurs études parviennent à des conclusions similaires : Dubose et collaborateurs en 1988 séquencent le gène *phoA* de 8 isolats naturels d'*E*. *coli*. Quand ils tentent de reconstruire sa phylogénie, ils remarquent de nombreux sites en contradiction avec la phylogénie la plus parcimonieuse. Ils expliquent cette observation par la présence de recombinaison intragénique impliquant de courts fragments ne remettant pas en cause la clonalité des lignées au niveau chromosomique. Par contre ces auteurs commencent à douter de la fiabilité des arbres phylogénétiques intra-espèce, la recombinaison augmenterait la similarité génétique et nivellerait le signal (DuBose, Dykhuizen et al. 1988). Peu de temps après, d'autres auteurs démontrent que l'arbre le plus parcimonieux du locus *gnd* est différent de celui construit par les données du MLEE pour 35 enzymes (Bisercic, Feutrier et al. 1991). Les auteurs expliquent cette incongruence par la présence de recombinaisons. En fait, ce locus est situé à coté du site *rfb*, codant l'antigène O, sélectionné pour être très polymorphe afin de déjouer le système immunitaire de l'hôte. Cette proximité semble avoir une influence. Dykhuisen et Green comparent les arbres phylogénétiques de

ces trois gènes (*gnd*, *phoA* et *trp*). Leurs nombreuses différences sont expliquées par la présence de recombinaison (Dykhuizen and Green 1991). Pourtant cette recombinaison n'est pas, d'après eux, suffisante pour invalider l'utilisation d'algorithme de construction d'arbres. Les auteurs proposent alors un parallèle intéressant entre la notion d'espèce bactérienne et d'espèce biologique : en effet les individus semblent interféconds (puisqu'on observe de la recombinaison). En conséquence, les phylogénies de différents gènes intra-espèces sont différentes, alors qu'elles sont identiques inter-espèces. Ces différents exemples impliquaient le plus souvent la recombinaison de courts fragments en séries discontinues (Milkman and Bridges 1993). Ces courts fragments pouvant être rentrés dans la cellule sous cette forme ou avoir été coupés après entrée dans la cellule par des nucléases. La troisième hypothèse étant que la superposition de plusieurs évènements de recombinaison impliquant de longs fragments pourrait expliquer la mosaïque de petits fragments observée. Ces trois hypothèses ne sont absolument pas exclusives, il est tout à fait possible qu'elles coexistent.

Ces études ont donc démontré que la recombinaison affectait le génome d'*E. coli*. Le taux de recombinaison semble même être 50 fois supérieur à celui de la mutation (Guttman and Dykhuizen 1994).

Comme nous venons de le voir, la recombinaison est loin d'être négligeable lorsque l'on souhaite s'intéresser à l'histoire évolutive *d'E. coli*. Quels en sont les impacts sur l'organisation du génome et sur la phylogénie de l'espèce ?

3.2.4 Impact de la recombinaison sur l'organisation du génome

Tout d'abord, il est important de noter que la recombinaison n'affecte pas tous les gènes de la même manière. Certains gènes, comme *gnd*, sont fortement recombinés (Bisercic, Feutrier et al. 1991), alors que d'autres (*gapA*, *celC*, *crr*, *gutB*) n'en présentent aucune trace (Nelson, Whittam et al. 1991; Hall and Sharp 1992).

Sans présumer des mécanismes, on s'aperçoit au niveau génomique qu'il existe de nombreuses traces d'acquisition ou de perte de gènes, appelés aussi transfert horizontal de gènes (Bergthorsson and Ochman 1995). En effet, les comparaisons des génomes de différentes souches ont permis de montrer que le contenu en gènes variait de façon

importante entre elles. En 2001, la souche enterohémorragique EDL933 O157:H7 a été complètement séquencée : elle contient plus de 30% de gènes en plus que K-12, la première souche de *E. coli* séquencée (Hayashi, Makino et al. 2001). Avec l'évolution des techniques de séquençage et leur automatisation, d'autres génomes d'*E. coli* ont pu être séquencés et comparés. Ce chiffre illustre bien la très grande plasticité du génome en termes d'acquisition et de perte de gènes, ce qui, on peut l'imaginer, peut s'avérer être un avantage adaptatif certain (Ochman, Lawrence et al. 2000). Par conséquent, la sélection naturelle agit sur ces gènes transférés horizontalement en les maintenant dans le génome s'ils y apportent un avantage. Les séquences inutiles seront quand à elles éliminées. Il existe ainsi une balance entre acquisition et perte de gènes. Les caractères observés provenant d'un transfert horizontal chez *E. coli*, peuvent être, par exemple, une résistance à certains antibiotiques. Cette acquisition se fait le plus souvent par le truchement de plasmide. Si la séquence est entourée de deux séquences d'insertion, elle peut également se comporter comme un transposon. Un autre moyen de propager les gènes de résistance aux antibiotiques sont les intégrons qui sont des structures incorporant des gènes grâce à un site d'attachement et une intégrase. Ces structures contiennent également un promoteur contrôlant l'expression des séquences incorporées. Il a été observé également l'acquisition par transfert horizontal de facteurs de virulence sous la forme d'îlot de pathogénicité qui sont situés le plus souvent au niveau d'un tRNA. De plus, certaines propriétés métaboliques comme la fermentation du lactose ont été acquises par transfert horizontal (Ochman, Lawrence et al. 2000).

La répartition de ces transferts horizontaux dans le génome d'*E. coli* K-12 MG1655 a été étudiée par comparaison de génomes d'entérobactéries proches ou par l'observation de biais compositionnels (G+C%, usage des codons). Si on cumule les deux méthodes, il apparaît que 25% des gènes de cette souche semblent avoir été acquis horizontalement (Fig. 11).

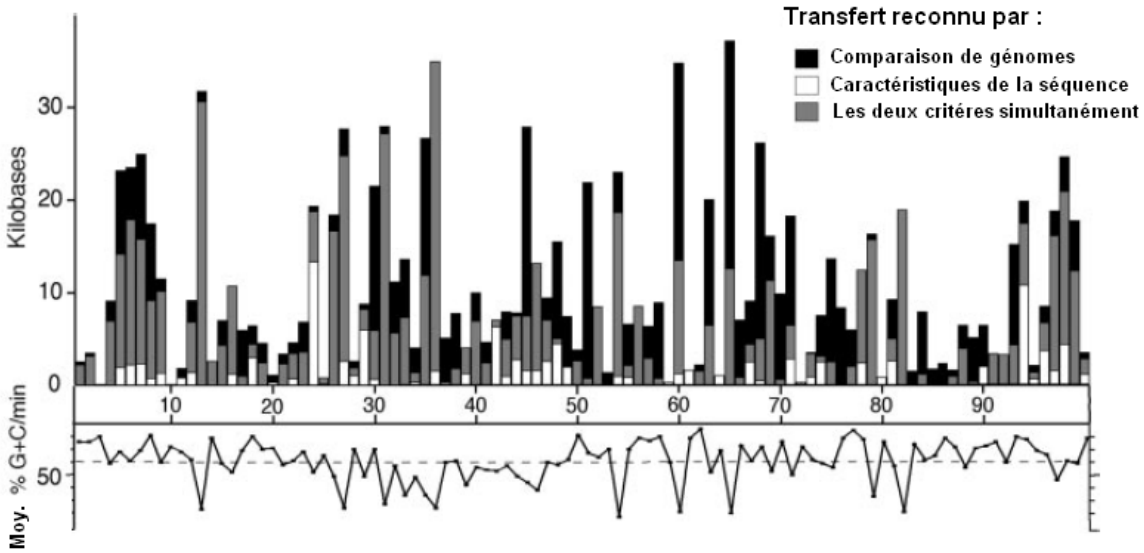

Fig. 11 : Représentation linéaire du chromosome d'*E. coli* K-12 MG1655 montrant la distribution d'ADN codant des protéines acquis horizontalement. Les barres verticales correspondent à la quantité d'ADN codant des protéines acquis horizontalement selon deux méthodes : en blanc : la composition en base, en noir : la comparaison de génomes entre *E. coli*, *Salmonella enterica* et *Klebsiella pneumoniae*. Si le gène est présent uniquement chez *E. coli*, il est figuré. En gris sont figurées les séquences trouvées par les deux méthodes. En bas, nous pouvons voir la composition en G+C de chaque minute du chromosome. La moyenne (51%) étant représentée par la ligne horizontale en pointillée (Ochman, Lerat et al. 2005).

De manière générale, ces acquisitions et pertes de gènes semblent n'affecter que très rarement les gènes du core génome de *E. coli* (Ochman, Lerat et al. 2005). Les gènes du core génome sont les gènes ayant un orthologue dans toutes les souches d'*E. coli.*

3.2.4 Impact de la recombinaison sur la phylogénie

Il est reconnu que la recombinaison présente un fort impact sur les méthodes de construction d'arbres.

La recombinaison a, plus particulièrement, de lourdes conséquences sur les longueurs de branches (Schierup and Hein 2000). Lorsqu'un fragment est transféré entre deux clones phylogénétiquement éloignés, la distance génétique qui les sépare diminue, et la distance entre la souche réceptrice et les souches proches augmente. L'arbre résultant présente donc de manière artéfactuelle des branches terminales longues ainsi que des branches internes courtes. Ce type d'arbre peut également être expliqué par une expansion

de la taille de la population. Pour faire la différence entre ces deux causes possibles, il suffit de calculer le D de Tajima (Tajima 1989). En effet, une expansion de la population est accompagnée par un excès d'allèles rares (D de Tajima négatif), ce qui n'est généralement pas observé chez *E. coli*.

Wirth et collaborateurs s'appuyant sur une approche de type MLST (« multi-locus sequence typing »), étudient 7 gènes de ménage dans 462 isolats d'*E. coli*. Ils concluent, peut-être à cause du biais affectant les longueurs de branches en présence de conversion génique, un peu rapidement à une expansion de la population (Wirth, Falush et al. 2006). Il est difficile de s'affranchir de cet artefact puisque, le polymorphisme expliquant les branches internes étant plus vieux, il a davantage de probabilité d'être impliqué dans un évènement de recombinaison. Enlever les sites recombinés de l'analyse biaiserait donc l'arbre de la même manière : les branches externes seront plus longues que les branches internes (Denamur, Picard et al. 2010). Pour Wirth et collaborateurs, l'expansion rapide de la population qu'ils observent et la recombinaison fréquente dans le génome d'*E. coli* interdisent l'utilisation des méthodes de phylogénie traditionnelles pour décrire les relations ancestrales entre les différents groupes (au nombre de 4 pour ces auteurs). C'est pourquoi, ils ont utilisé un modèle d'évolution en réseau pour décrire les groupes de souches (implémenté par le logiciel STRUCTURE). Un tiers des souches qu'ils ont utilisées ont été placées par cette méthode dans des groupes hybrides, c'est à dire qu'elles dériveraient de plusieurs ancêtres. Certaines études plus récentes (Gordon, Clermont et al. 2008; Jaureguy, Landraud et al. 2008) montrent qu'il faut augmenter le nombre des groupes jusqu'à 7 pour analyser l'espèce. De cette façon la plupart des souches classées comme étant recombinantes par Wirth se trouvent dans ces nouveaux groupes également retrouvés par les analyses phylogénétiques.

4 La phylogénie

Les premiers phénogrammes obtenus grâce aux données du MLEE identifiaient 4 groupes principaux (A, B1, B2 et D), puis 2 groupes accessoires (C et E) (Selander, Caugant et al. 1986; Goullet and Picard 1989; Herzer, Inouye et al. 1990). Puis, les auteurs se sont intéressés à la concaténation de différents gènes du MLST. Ils ont retrouvé les mêmes groupes, que ce soit par des approches phylogénétiques (en enlevant ou pas les séquences

contenant des traces de recombinaison) ou de génétique des populations. Ces dernières méthodes repèrent les événements de recombinaison qui interrompent l'héritage clonal (Milkman and Stoltzfus 1988; Milkman and Bridges 1990). Par contre, lorsqu'elles utilisent les profils alléliques au lieu des séquences brutes, elles pondèrent de la même façon une mutation et une recombinaison (Lecointre, Rachdi et al. 1998; Escobar-Paramo, Sabbagh et al. 2004; Johnson, Owens et al. 2006; Wirth, Falush et al. 2006). La phylogénie d'*E. coli* se précise alors, également à l'aide du MLST sur un nombre de souches plus important (Reid, Herbelin et al. 2000; Hershberg, Tang et al. 2007) et en utilisant *E. fergusonii* comme racine (Escobar-Paramo, Sabbagh et al. 2004). Utiliser *E. fergusonii* (plus proche d'*E. coli* (Lawrence, Ochman et al. 1991) au lieu de *Salmonella enterica* permet de limiter l'artefact d'attraction des longues branches vers la racine de l'arbre.

A la base de cet arbre, le groupe des B2 apparaît le plus diversifié avec au moins 9 sous-groupes phylogénétiques (Le Gall, Clermont et al. 2007). Puis, un sous-groupe du groupe D (appelé F) se distingue (Jaureguy, Landraud et al. 2008). Ensuite, le reste de l'espèce est figuré. Le reste du groupe D émerge d'abord, suivi par le groupe E. Finalement, les groupes frères A et B1 apparaissent (Fig. 12 et 13).

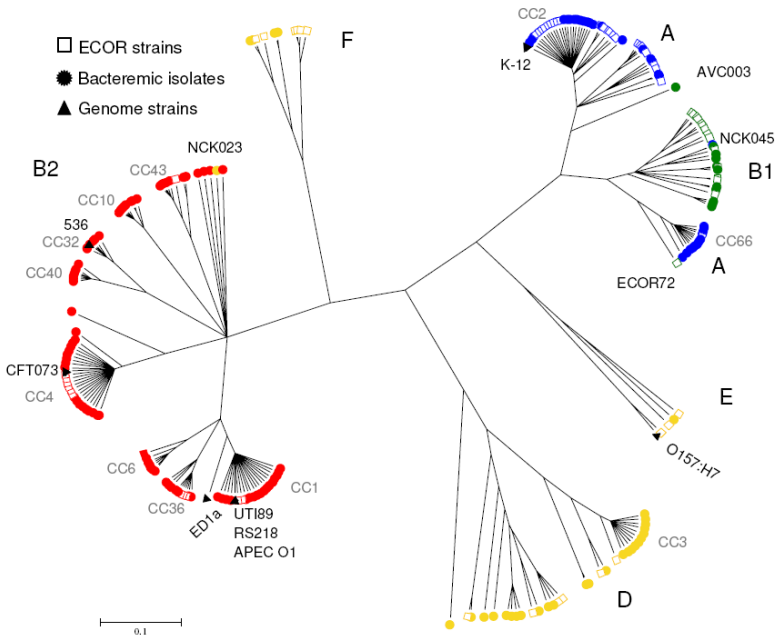

Fig. 12 : La phylogénie d'*E. coli* basée sur les données du MLST (Clonalframe). Analyse phylogénétique réalisée avec Clonalframe basée sur les séquences de 8 gènes de 161 isolats d'*E. coli* issus de bacteriémie (cercles) et 67 souches de la collection de référence

ECOR (carrés), ainsi que de 7 génomes de références (triangles) (d'après la fig. 1 de (Jaureguy, Landraud et al. 2008)).

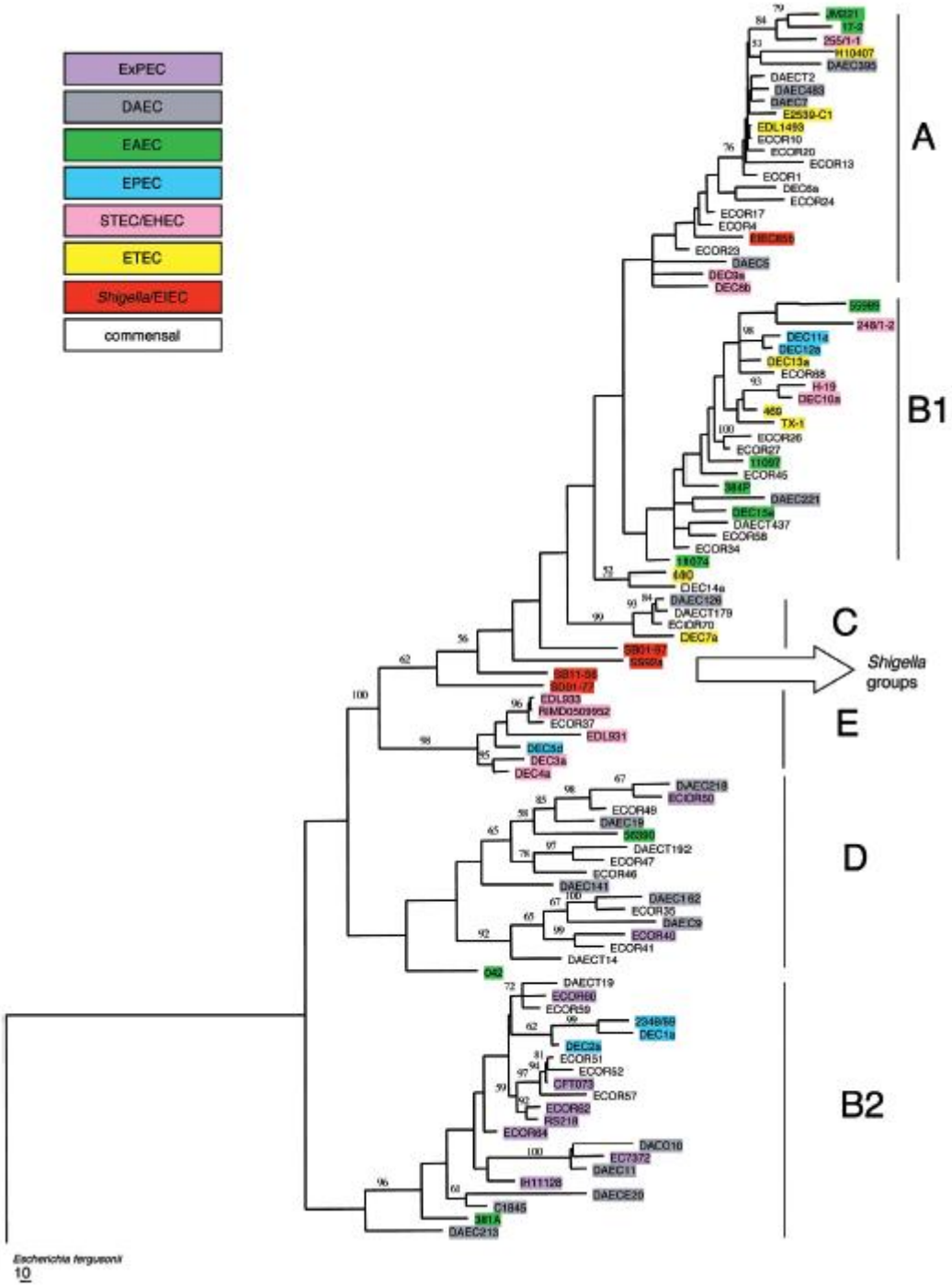

Fig. 13 : La phylogénie d'*E. coli* basée sur les données du MLST (consensus). Arbre consensus basé sur l'analyse de 6 gènes essentiels par maximum de parcimonie, raciné sur *E. fergusonii*. Seuls les bootstraps supérieurs à 50% sont figurés. (d'après la fig. 1 de (Escobar-Paramo, Sabbagh et al. 2004))

CONCLUSION :

Comme nous avons pu le voir, il semble que la structure de l'espèce *E. coli* soit plutôt clonale. Il faut pourtant prendre en considération la recombinaison lorsque l'on souhaite étudier son histoire évolutive. Le MLST ne permet pas de trancher de manière franche la question de la phylogénie. C'est le séquençage de nombreuses souches qui nous aidera à savoir si une phylogénie est possible, et si oui à la reconstruire de la manière la plus robuste possible.

PARTIE EXPERIMENTALE

Chapitre I : Une forme de mutation : la mutation transcriptionnelle et son influence sur le génome

1 Introduction

Comme nous l'avons rappelé dans la première partie de ce manuscrit, l'ADN peut subir des altérations chimiques qui génèrent alors des mutations. Ces mutations peuvent être coûteuses lorsqu'elles inactivent ou diminuent une fonction utile à l'organisme. Elles peuvent aussi être bénéfiques. Leur impact dépend des conditions environnementales. Plus généralement, lorsque les conditions de vie changent, une population a de meilleures chances de perdurer si elle abrite une importante diversité génétique. Car elle a ainsi une plus grande probabilité qu'un individu possède une mutation qui s'avérerait avantageuse dans ce nouvel environnement.

Il a été montré, pour certains gènes particuliers, que la transcription pouvait avoir une influence sur le taux de mutation (Wright, Reimers et al. 2002; Wright, Reschke et al. 2003). Le mécanisme proposé est le suivant : pendant la transcription, l'ADN est transitoirement simple brin. Pendant cette période il est davantage soumis aux altérations chimiques et donc aux mutations. De plus, il forme des structures secondaires dépendantes de la séquence nucléotidique. Dans ces structures, certaines bases vont se retrouver plus fréquemment appariées que d'autres, et donc davantage protégées des mutations. On entrevoit aisément que par ce mécanisme, il peut exister un contrôle temporel de cette forme de mutations en augmentant la transcription. On peut également imaginer qu'il puisse exister un contrôle local préventif de cette mutagénèse selon la nature de la séquence impliquée, contrairement aux systèmes de réparation de l'ADN qui agissent après que l'erreur ait été faite.

Afin d'étudier cette forme de mutabilité à l'échelle du génome, nous avons mis au point un indice de mutabilité transcriptionnelle basé sur la stabilité des structures secondaires dans lesquelles chacune des bases du gène est impliquée.

Les objectifs de cette étude étaient, tout d'abord, d'utiliser cet indice de mutabilité pour décrire l'ensemble des gènes d'*E. coli* en terme de mutagénèse transcriptionnelle, plus particulièrement de savoir s'il était possible de mettre en évidence des gènes présentant des

traces de sélection pour augmenter ou diminuer cette forme de mutabilité. Cette sélection, si elle existe, affecte les codons synonymes. D'autres formes de sélection agissent sur ces derniers : celle affectant le taux de bases G et C du gène, celle affectant le biais de codon ou encore celle affectant la stabilité de l'ARN messager. Un autre objectif de cette étude a été d'essayer de faire la part entre ces différentes formes de sélection dont les forces peuvent se contrecarrer ou au contraire s'additionner.

2 Article I

3 Principaux résultats et perspectives

Grâce à l'établissement d'un indice de mutabilité transcriptionnelle basé sur les principales propriétés thermodynamiques de l'ADN, nous avons pu étudier l'influence de cette mutagénèse sur l'évolution du génome et avons pu montrer que le contrôle de la mutabilité transcriptionnelle à travers les structures secondaires formées par l'ADN simple brin est sous sélection dans le génome d'*E. coli*.

En alignant deux à deux les gènes orthologues des génomes des souches K-12 MG1655, CFT073 et EDL933 O157:H7, nous avons mis en évidence que les sites synonymes variables entre chaque paire de génomes avaient un indice de mutabilité significativement plus élevé que les sites constants. En utilisant des régressions logistiques nous avons montré qu'en moyenne, un site ayant un indice de mutabilité transcriptionnelle maximum (+1) augmentait sa probabilité de varier entre K-12 MG1655 et CFT073 de plus de 20% par rapport à un site ayant un indice de mutabilité transcriptionnelle nul. Nous avons ensuite divisé les données en trois groupes selon une mesure du biais de codon (le « Major Codon Usage » : MCU) qui représente une bonne approximation du taux d'expression moyen du gène. Cette analyse nous a permis de montrer que, comme attendu, le niveau d'expression des gènes augmente l'impact de la mutabilité transcriptionnelle, sauf pour les gènes très fortement exprimés (MCU > 0,7) pour lesquels nous proposons que la sélection pour le biais de codon soit tellement forte qu'elle obscurcit le signal de la mutagénèse.

Par différentes approches de randomisation, nous avons mis en évidence que la sélection agissant pour moduler la mutabilité transcriptionnelle est assez forte pour laisser une empreinte significative sur le génome d'*E. coli* K-12 MG1655. Cette sélection s'effectue majoritairement pour augmenter la robustesse du génome. Plus un gène est riche en bases G et C, plus son indice de mutabilité transcriptionnelle moyen est faible. Les bases G et C sont particulièrement sensibles aux altérations chimiques, ce qui pourrait expliquer que la sélection pour diminuer la mutabilité transcriptionnelle soit plus importante dans ces gènes. De plus, l'appariement des bases G et C de l'ADN est plus fort que celui des bases A et T, il est donc possible que minimiser la mutabilité transcriptionnelle soit plus facile dans les gènes riches en GC, car les structures secondaires sont plus stables. En randomisant 1000 fois chaque gène de *E. coli* K-12 MG1655 en prenant soin de garder la même séquence en

acides aminés codée, le même taux de bases GC et le même biais de codons, nous avons pu estimer que 20% des gènes était significativement plus stable du point de vue de la mutabilité transcriptionnelle. Le facteur clef expliquant le fait que ces gènes soient significativement plus robustes à la mutation transcriptionnelle semble être le MCU. Cette sélection sur la mutabilité transcriptionnelle agit sur le positionnement des codons synonymes le long de la séquence du gène. Il est probable que la sélection agissant pour le biais de codons et celle agissant pour diminuer la mutabilité transcriptionelle entrent en compétition dans le cas des gènes les plus exprimés. De plus la sélection pour diminuer cette forme de mutabilité semble plus faible que celle agissant sur le biais de codons. Ces deux facteurs expliqueraient que les gènes les plus exprimés, ne présentent pas la plus faible mutabilité.

Dans ce chapitre, nous avons mis en évidence une forme de mutagénèse source de diversité dans les séquences. Une autre source de différences est la recombinaison. Celle-ci met en œuvre des fragments plus longs. La sélection naturelle et la dérive génétique agissent sur cette dernière comme sur les mutations et permettent l'évolution. Lorsque la recombinaison est fréquente, cette dernière empêcherait l'utilisation des méthodes phylogénétiques traditionnelles pour reconstruire les liens de parentalité entre les souches. Ce point sera abordé dans le chapitre suivant.

Chapitre II : Le génome d'*E. coli* : un désordre organisé

1 Introduction

Les processus mutationnels seuls ne suffisent pas à expliquer la totalité de la diversité observée entre les différentes souches de l'espèce *E. coli*. La recombinaison y est très importante (Guttman and Dykhuizen 1994). Or pour étudier la recombinaison à l'échelle du génome entier, la démarche la plus efficace est celle mettant en œuvre la génomique comparative. Pour cela, disposer d'un grand nombre de souches séquencées est nécessaire, si possible représentant au mieux la diversité de l'espèce.

La problématique sous-jacente à ce travail était d'essayer de comprendre comment *E. coli* s'est adapté à ses nombreux modes de vie. Pour répondre à cette question il a été nécessaire de reconstruire l'histoire évolutive de l'espèce, puis de caractériser les flux d'ADN dans le temps (le long de l'histoire évolutive) et dans l'espace (localisation sur le chromosome).

C'est dans ce but qu'a été initié le projet ColiScope en partenariat avec le Génoscope dans le cadre duquel nous avons procédé au séquençage de 6 nouvelles souches d'*E. coli* et de la souche type d'une espèce proche : *E. fergusonni* (Lawrence, Ochman et al. 1991), afin de l'utiliser comme racine. Les six souches d'*E. coli* ont été choisies dans le but de représenter le mieux possible les différentes situations épidémiologiques caractéristiques de l'espèce.

- 2 souches du groupe B1 : IAI1, une souche commensale et 55989, une souche entéroaggrégative responsable de diarrhée
- 2 souches du groupe D : IAI39, responsable de pyélonéphrite et UMN026, une souche multirésistante aux antibiotiques, isolée aux USA et appartenant au « clonal group A » (CGA) (Manges, Johnson et al. 2001)
- 2 souches du groupe B2 : S88, une souche hautement virulente responsable de méningites néonatales, de sérogroupe O45 et correspondant au clone Européen (Bonacorsi, Clermont et al. 2003) et

ED1a, une souche avirulente, exclusivement humaine et commensale (Clermont, Lescat et al. 2008).

A l'époque, dans la littérature 14 souches étaient disponibles. Parmi elles, la souche de laboratoire K-12 MG1655 (Blattner, Plunkett et al. 1997), des souches responsables d'infection urinaire : comme CFT073 (Welch, Burland et al. 2002), des souches déclenchant des infections diarrhéiques telles que les *Shigella sp* (Jin, Yuan et al. 2002; Wei, Goldberg et al. 2003; Nie, Yang et al. 2006) et deux clones O157:H7 responsables de diarrhée entérohémorragique (Hayashi, Makino et al. 2001; Perna, Plunkett et al. 2001). Le projet ColiScope consistait à faire l'étude comparative de ces 14 génomes avec les 7 génomes séquencés à son occasion. Ceci avec pour objectif de :

- Déterminer l'importance de la recombinaison par rapport à la mutation dans le génome
- Construire la phylogénie de l'espèce *E. coli* si cela est possible
- Caractériser l'impact de la mutation et de la recombinaison sur l'organisation du génome
- Trouver, s'ils existent, les gènes spécifiques de certains mode de vie comme le commensalisme, la pathogénicité ou encore la résistance aux antibiotiques

2 Article II

3 Principaux résultats et perspectives

Au sein du consortium réunissant 41 personnes, j'ai tout d'abord participé de manière importante à l'effort d'annotation manuelle fonctionnelle (10000 gènes annotés lors de l'ensemble de ce travail), ainsi qu'à la correction des codons d'initiation lorsque cela était possible. Par la suite, Olivier Tenaillon a estimé le taux de conversion génique sur le taux de mutation ainsi que la longueur moyenne des fragments impliqués par des méthodes d'approximation bayésienne. J'ai ensuite procédé aux différentes simulations s'appuyant sur le modèle de coalescence et ai mis en oeuvre les tests de comparaison de topologies montrant que le taux observé ne suffisait pas à brouiller le signal phylogénétique. La représentation de l'histoire évolutive de l'espèce sous forme d'arbre étant possible, j'ai effectué la totalité des analyses phylogénétiques qui ont permises de reconstruire l'arbre présenté. J'ai également participé, en collaboration avec Olivier Tenaillon, à l'analyse de la congruence phylogénétique de l'arbre global le long du chromosome. Enfin, j'ai mis en oeuvre les tests de détection de sélection, de mesure de recombinaison et de comparaison de topologies disponibles à l'époque sur l'ensemble des gènes du core génome d'*E. coli*, leur résultats étant peu intéressants nous avons choisi de ne pas les présenter en tant que tels dans l'article.

L'organisation du génome d'*E. coli* semble relativement stable au cours de l'évolution, en effet, les 21 génomes étudiés présentent assez peu de réarrangements. Ce sont les *Shigella* et *E. fergusonni* qui ont subit le plus grand nombre de remaniements.

L'ensemble des génomes a permis de constituer le « core » génome de l'espèce, représentant l'ensemble des gènes communs à tous les génomes séquencés de *E. coli* (n=1976), et le « pan » génome, comprenant l'ensemble des gènes de ces génomes (n=17838). Nous avons observé que le nombre de génomes pris en compte est maintenant suffisant pour donner une bonne estimation du « core » génome de l'espèce. Par contre le « pan » génome est encore loin d'être complet, ceci même si on enlève de l'analyse les paralogues potentiels, les éléments transposables et les gènes d'origine phagique (Fig 1 de l'article).

Comme nous l'avons mentionné dans le chapitre III de la partie bibliographique, pour certains auteurs, utiliser des méthodes de reconstruction d'arbres phylogénétiques intra-

espèce bactérienne est impossible à cause de l'impact important de la recombinaison homologue (Wirth, Falush et al. 2006). En effet, la recombinaison étant d'autant plus rare que les séquences sont divergentes, il semblerait, selon ces auteurs, qu'il soit possible de faire un parallèle entre espèce bactérienne et espèce biologique en termes d'échange de gènes, ce qui interdirait toute reconstruction d'arbre phylogénétique intra *E. coli*.

En utilisant l'impact de la conversion génique sur le déséquilibre de liaison, nous avons estimé que le ratio moyen du taux de recombinaison sur le taux de mutation était de 2,5. De plus, la longueur moyenne des fragments est très courte : seulement 50 pb. Nous avons ensuite effectué des simulations de coalescence, en utilisant le taux de mutation et la longueur des fragments précédemment estimés, mais avec des taux de conversion génique croissants. La comparaison des arbres directement dérivés des simulations et des arbres inférés par maximum de vraisemblance nous a permis de démontrer que l'important taux de conversion génique observé ne suffisait pas à obscurcir la phylogénie de l'espèce (Fig 3 de l'article).

L'arbre phylogénétique que nous avons alors pu construire à partir du « core » génome ou du « backbone » (régions homologues au sein de l'alignement de génomes entier) a confirmé l'existence des différents groupes A, B1, B2. Le groupe D, quant à lui est paraphylétique. B2 apparaît comme ancestral.

Lorsque nous avions fait des tests de comparaison de topologies (SU (Shimodaira–Hasegawa) et KH (Kishino–Hasegawa) et ELW (« Expected Likelihood Weight ») tests) (Kishino and Hasegawa 1989; Shimodaira and Hasegawa 1999; Goldman, Anderson et al. 2000; Strimmer and Rambaut 2002), nous avions trouvé que seulement 25% des gènes du core génome n'apparaissaient pas significativement différents de l'arbre global. Même s'il est tout à fait possible que certains gènes aient leur propre histoire évolutive, ce chiffre semblait faible. Or, il s'explique par un manque de signal phylogénétique (55% des gènes ont moins de 40 sites informatifs, et notre étude s'effectue sur 20 génomes). Les arbres sont donc en grande partie non résolus. Depuis la rédaction de l'article, une nouvelle approche a été implémentée, nommée arbres d'arbres (« TreeOfTrees ») qui permet de représenter sous forme d'arbre les différences topologiques entre plusieurs arbres (http://bioinformatics.lif.univ-mrs.fr/TreeOfTrees/index.html). Cette méthode consiste tout

d'abord à traduire les arbres qu'on désire comparer en matrice de distances. Plusieurs distances sont disponibles, j'ai choisi d'utiliser le nombre de branches non nulles qui séparent deux souches dans l'arbre. Puis la méthode compare les matrices d'arbres entre elles pour construire une matrice de distances d'arbres (en utilisant une simple distance euclidienne, par exemple). Cette matrice de distance peut ensuite être visualisée sous la forme d'un arbre en utilisant l'algorithme de neighbor-joining. On effectue cette même analyse un certain nombre de fois en pratiquant des « bootstraps » sur les données (rééchantillonnage aléatoire avec remise). On effectue alors un arbre consensus qui aura comme valeur de support aux nœuds le pourcentage des arbres effectués sur les données rééchantillonnées qui contiennent cette arête. Une forte valeur sépare deux ensembles d'arbres incongruents, une faible valeur suggère que les deux ensembles ne sont pas incongruents ou que le signal phylogénétique ne suffit pas pour rejeter l'hypothèse d'incongruence. Nous avons procédé à cette analyse sur chacun des 1000 gènes possédant le plus de sites informatifs rééchantillonnés par « bootstrap » 200 fois et seul un couple d'arbres de gènes était séparé par un nœud présentant une valeur de support supérieure à 50%. Contrairement aux SH, KH et ELW tests, cette analyse ne compare que les parties résolues des arbres (longueurs des branches non nulles). « TreeOfTrees » nous indique ici que nos arbres de gènes ne sont pas incongruents entre eux ou plutôt que le signal observé ne permet pas de conclure sur leur incongruence.

Nous nous sommes ensuite servis de l'arbre pour reconstruire les génomes ancestraux et inférer le scénario de gain et de perte de gènes au cours de l'histoire évolutive. Cette analyse a montré qu'il y a très peu de gènes spécifiques de clades, ce qui semble invoquer que les gènes transférés disparaissent rapidement de la population probablement à cause de leur conséquences délétères. Les acquisitions les plus récentes sont en général soit des gènes phagiques, soit des IS (séquences d'insertion). Par contre, l'acquisition de gènes ayant une fonction connue est peu fréquente, mais ils sont rarement perdus, ce qui semble indiquer qu'ils apportent un avantage adaptatif.

Nous n'avons pas pu mettre en évidence de gènes spécifiques de la virulence extraintestinale ce qui renforce l'hypothèse que ce type de virulence est en fait un sous-produit du commensalisme (Le Gall, Clermont et al. 2007).

L'organisation du génome bactérien est fortement liée à des processus cellulaires fondamentaux comme la réplication, la ségrégation et la transcription. Nous avons montré que le flux massif de gènes ne la perturbait pas car ces insertions se faisaient à des endroits précis du génome. C'est ainsi que 133 loci contiennent 71% des gènes du non « core ». Seulement, dans 83% des cas, il n'y a ni trace d'ARN de transfert, ni d'intégrase. Il ne semble donc pas s'agir de recombinaison site-spécifique. Nous proposons que ces points chauds sont formés grâce à l'acquisition d'un premier grand fragment dans une région permissive. Observer de nouvelles insertions à cet endroit est alors plus probable car elles sont neutres. Ces 133 loci contiennent 61% des ruptures de synténie, ce qui semble indiquer que ces points chauds d'insertion/délétion sont aussi des points chauds de réarrangements. Nous proposons l'hypothèse que l'ADN, une fois intégré dans un génome peut se propager dans la population par recombinaison homologue des régions flanquantes. Or, effectivement, nous observons 25 points chauds d'incongruence phylogénétiques correspondant à des gènes connus comme étant sous forte sélection diversifiante. Les deux plus importants étaient également des points chauds d'intégration : la région associée à l'opéron *rfb*, codant pour l'antigène O et celle associée à l'ARNt *leuX*, comprenant le locus *hsd*, précédemment identifiées par Milkman comme les 2 principaux « bastions de polymorphisme » de l'espèce (Milkman, Jaeger et al. 2003). Ces deux régions comportent, en effet, des points chauds d'intégration permettant l'arrivée de gènes par transfert horizontal inter-espèces et, à leurs bornes, des signes d'incongruence phylogénétique témoignant d'un important transfert horizontal intra-espèce comme c'est le cas pour le HPI (Schubert, Darlu et al. 2009).

L'analyse par le déséquilibre de liaison le long du « backbone » du ratio du taux de conversion génique sur le taux de mutation révèle une grande région, située autour du terminus de réplication et caractérisée par un taux de conversion génique sur un taux de mutation particulièrement faible. Cette région présente un moindre polymorphisme intra-espèce, un plus faible taux de GC mais par contre une divergence avec *Salmonella* légèrement supérieure que dans le reste du chromosome. Nous avons montré que le taux de divergence synonyme comme non synonyme avec *E. fergusonni* est deux fois plus important dans cette région que dans le reste du génome. L'enrichissement en bases A et T ainsi que l'augmentation de la divergence observée pourrait s'expliquer par une augmentation du taux de mutation, mais cette hypothèse est invalidée par le faible polymorphisme observé.

Il a été montré que lorsqu'il y a de nombreuses mutations moyennement délétères ainsi qu'un faible taux de recombinaison, il y a sélection d'arrière plan (« background selection »). Dans ce cas, une fraction de la population portant des allèles délétères est amenée à disparaitre à long terme, les sites liés génétiquement mais non délètères sont donc eux aussi peu à peu éliminés, ce qui se traduit par un excés d'allèles rares dans la population (Charlesworth, Morgan et al. 1993). En effet, le D de Tajima (Tajima 1989) calculé sur les sites non synonymes est négatif, celui calculé sur les sites synonymes nul. Ces observations sont en accord avec cette hypothèse et suggèrent que la plupart des mutations non synonymes sont délétères et donc éliminées de la population par sélection naturelle. Lorsqu'il y a sélection d'arrière plan et un faible taux de recombinaison, on s'attend à observer un plus faible polymorphisme, une augmentation de la proportion des allèles rares et une diminution de l'efficacité de la sélection naturelle. Or, dans la région du terminus nous avons observé moins de polymorphisme, un D de Tajima significativement plus faible (ce qui traduit une forte proportion d'allèles rares) ainsi qu'un ratio mutations non-synonymes sur synonymes supérieur au reste du génome signifiant une perte d'efficacité de la sélection naturelle.

Puisque les mutations s'effectuent préférentiellement des bases GC vers les bases AT, il est attendu que les fragments portant le plus de mutations soient plus riches en AT que les autres. Dans ce contexte, une autre hypothèse permettant d'expliquer nos observations dans la région du terminus pourrait être que l'enrichissement en bases A et T observée proviendrait du fait que le faible taux de recombinaison ne permettrait pas de remplacer les fragments portant des mutations délétères aussi efficacement que dans le reste du chromosome (Balbi, Rocha et al. 2009).

Il est également possible que la recombinaison est un effet mutagène direct. L'hypothèse de la conversion génique biaisée (utilisée pour expliquer l'hétérogénéité du contenu en GC des génomes de mammifères) stipule que les mésappariements présents dans les heteroduplex formés lors des recombinaisons seraient réparés en favorisant les bases G et C (Galtier, Piganeau et al. 2001). Par exemple, un mésappariement G-T sera plus fréquemment remplacé par une paire de bases G-C que A-T pas les systèmes de réparation. Cette hypothèse pourrait donc expliquer le plus faible taux de GC dans la région du

terminus, puisqu'à cet endroit du chromosome on observe également un taux de conversion génique moindre.

Le faible taux de conversion génique observé dans la région du terminus pourrait s'expliquer par le fait qu'elle comporte des régions compactées. Les liens entre la fréquence de la conversion génique, la composition de la séquence, le compactage de l'ADN et la sélection rappellent les relations fortes existant entre la dynamique du génome et l'organisation chromosomique.

Cette étude nous a permis de montrer que la recombinaison homologue s'effectuant sous la forme de conversion génique ne bouleversait pas la reconstruction de l'histoire évolutive par les techniques phylogénétiques classiques. De plus, il semble que le conflit entre dynamique du génome (en termes de flux de gènes) et organisation du chromosome ait été résolu par la création de points chauds concentrant les évènements. Le chapitre suivant traite d'un de ces points chauds.

Chapitre III : Caractérisation précise d'un des principaux points chauds d'intégration : la description de l'îlot de UMN026 et son application

1 Introduction

Parmi les deux plus importants points chauds d'incongruence phylogénétique et d'intégration présents dans tous les génomes séquencés d'*E. coli*, nous en avons détaillé un : la région associée à l'ARNt *leuX*. En effet, nous y avons découvert le support génétique de la résistance aux antibiotiques de la souche responsable d'infection urinaire UMN026. Cette souche représente le « clonal group A » ou CGA qui cause des infections extra-intestinales, principalement des infections urinaires. Ce groupe clonal est, de plus, caractérisé par le fait que les souches qui le constituent sont résistantes à de nombreux antibiotiques : ampicilline, chloramphénicol, streptomycine, sulfamides, tétracycline et triméthoprime (Manges, Johnson et al. 2001; Johnson, Manges et al. 2002). Bien que largement prévalent aux États-Unis, le CGA est pourtant distribué mondialement. Les souches du CGA appartiennent au groupe phylogénétique D et contiennent des facteurs de virulence typiques des souches responsables d'infection urinaire (allèle *papA* F16, allèle *papG* II, *iutA*, *kpsM* II, *traT* et *ompT*). De plus, elles peuvent être aussi pathogènes que des souches du groupe B2 dans un modèle murin d'infection urinaire (Johnson, Murray et al. 2005). Enfin, la comparaison de souches du clone par électrophorèse en champ pulsé montrait peu de profils et indiquait une dissémination récente du clone (Manges, Johnson et al. 2001). Ceci peut être expliqué par le fait qu'elles contiennent à la fois de nombreux facteurs de virulence extraintestinale et également des gènes leur permettant de résister à de nombreux antibiotiques. Cela représentait donc une très bonne occasion d'approfondir la structure et la composition de cet îlot génomique d'intégration et de recombinaison.

2 Article III

3 Principaux résultats et perspectives

Dans cette étude, je suis essentiellement intervenue en tant que formatrice et conseillère lors de l'annotation fine de cet îlot génomique.

L'annotation de la souche UMN026 a permis de mettre en évidence que presque tous ses gènes de résistance étaient localisés dans un unique îlot génomique de 105 kpbs porté par le chromosome. Cette région peut être considérée comme un îlot génomique car sa déviation en taux de GC par rapport à la moyenne le long de fenêtres de 1000 pbs est significativement plus grande que deux déviations standards. Il est, de plus, accolé à un gène d'ARNt (*leuX*). Et enfin, il inclut des gènes impliqués dans la mobilité tels que des intégrases, des transposases, des résolvases et des éléments à signature phagique. En fait, ces gènes de résistance sont groupés en amont de cette structure dans une région de 22,5 kpbs que nous avons appelé le GRM pour « genomic resistance module » (Fig 1 de l'article).

Le GRM contient 29 séquences codantes (25 complètes et 4 partielles) dont 8 sont associées à des résistances à certains antibiotiques et 2 à des résistances à des antiseptiques ou à des métaux lourds. Cinq des six marqueurs de résistances caractéristiques de CGA sont présents. Le GRM est caractérisé par un fort taux de GC (55% contre 50,7% en moyenne sur le génome entier). Il est borné par l'ARNt et par une transposase appartenant à la famille des IS*1*. Il contient deux gènes codant des intégrases, 10 transposases entières ou partielles, un transposon ressemblant au Tn*21* ainsi qu'un intégron. Il manque pourtant certaines des inversions répétées qui sont d'ordinaire observées autour des transposons et des intégrons. L'abondance des éléments impliqués dans la mobilité tels que les transposases pourrait expliquer les réarrangements observés. Cela pourrait également être à l'origine de la localisation chromosomique et non plasmidique de cet îlot de résistance.

La souche 042 isolée lors d'un épisode diarrhéique et appartenant au groupe D contient 11 des gènes du GRM incluant des résistances à 4 antibiotiques. Mais dans cette souche, les gènes sont localisés ailleurs dans le génome et tous dans le Tn*21* complet. Ceci suggère de multiples évènements récents d'acquisitions de ces résistances. Les éléments transposables observés dans ces îlots génomiques seraient à l'origine des remaniements et des transferts latéraux avec d'autres entérobactéries ou avec des bactéries de l'environnement.

L'approche comparative détaillée de l'îlot génomique de 105 kpbs dans 14 autres souches d'*E. coli* nous a permis de montrer sa structure composite. La présence et l'absence des sous-régions que nous avons décrites dans chacune des souches ne semblent ni corrélées à la phylogénie ni au pathotype. Il est intéressant de noter qu'à la même position, c'est à dire juste après l'ARNt *leuX* et *intB*, on trouve dans les souches UTI89 et 536, toutes deux appartenant au groupe B2 et engendrant des infections urinaires, l'îlot de pathogénicité II (PAI II). Cette étude illustre le fait que la plasticité des génomes d'*E. coli* s'effectue selon de multiples voies et permet, à un seul locus, la présence alternative de gènes de virulence et de résistance.

Les gènes de virulence peuvent être intégrés dans le génome par le biais de transfert latéraux et par la-même ne pas être informatifs lorsqu'on s'intéresse à la phylogénie de l'espèce. Nous nous sommes donc intéressés, dans le chapitre suivant, à un marqueur de virulence : l'estérase B afin de comprendre les sélections qui l'affectent et afin de comprendre son histoire évolutive en la comparant à celle de l'espèce.

Chapitre IV : Caractérisation d'un marqueur de virulence en tant que marqueur phylogénétique : *aes*, ou l'estérase B.

1 Introduction

Les quatre principaux groupes phylogénétiques (A, B1, B2 et D) d'*E. coli* ont été historiquement déterminés sur des critères phénotypiques puis génétiques. Parmi ces critères ont peut citer la séparation électrophorétique de protéines, le polymorphisme de fragments de restriction (RFLP « restriction fragment length polymorphism ») des ADN ribosomiques, le polymorphisme d'amplification aléatoire (RAPD « random amplified polymorphic DNA ») et le MLST. Sept types d'estérases nommées A, B, C, D, I, F et S ont été décrits dans l'espèce. Ils diffèrent par leur capacité d'hydrolyser différents substrats synthétiques et par leur sensibilité au di-isopropyl fluorophosphate. De plus ces différents types sont séparables sur gel de polyacrylamide-agarose (Goullet 1980). Le type le plus constamment observé est l'estérase B (E.C. 3.1.1.1). Or, il a été montré il y a une vingtaine d'année que cette estérase comporte deux niveaux de mobilité électrophorétique : le type B_2 migrant plus lentement que le type B_1 (Goullet and Picard 1989). Les souches du groupe phylogénétique B2 présentent les variants de type B_2, les autres ceux du type B_1. Le groupe B2 contient la majorité des souches pathogènes extra-intestinales, ce qui suggère un lien possible entre le polymorphisme de l'estérase B et la virulence extra-intestinale.

Cette enzyme est donc un marqueur de différenciation intra-spécifique entre les souches du groupe B2 et les autres, ainsi qu'un marqueur de virulence.

L'objectif de l'étude était d'identifier le gène codant l'estérase B puis de répondre aux questions suivantes : l'estérase B peut-elle être considérée comme un marqueur phylogénétique intra *E. coli* ? Et s'agit-il d'un facteur de virulence ?

2 Article IV

3 Principaux résultats et perspectives

Au cours de cette étude, j'ai effectué une partie des arbres phylogénétiques. J'ai, par contre, mis en oeuvre l'ensemble des tests de sélection utilisés.

Parmi plusieurs gènes candidats, nous avons pu montrer par inactivation et complémentation du gène que le gène *aes* (acétyl estérase) code pour l'estérase B. Ce gène a été décrit en 1997 (Peist, Koch et al. 1997). Son arbre phylogénétique établi sur les 72 souches de la collection ECOR, représentative de l'espèce (Ochman and Selander 1984) sépare parfaitement les deux variants B_1 et B_2. Différents tests de sélection ont montré qu'*aes* était sous sélection purifiante. Par une méthode de comparaison d'arbres décrites dans le chapitre II-3 p. 73 (« TreeOfTrees ») nous avons également mis en évidence qu'*aes* était, au même titre que les gènes traditionnellement utilisés dans les approches de MLST, un excellent marqueur de l'histoire évolutive de l'espèce.

La modélisation de la protéine a permis de localiser les acides aminés polymorphes dans les deux types de variant. Ils se situent en surface à des localisations différentes.

En utilisant un modèle murin de colonisation et de septicémie qui consiste à infecter des souris avec les souches d'*E. coli* CFT073 sauvage et mutée pour *aes* nous n'avons pas pu établir de lien entre la présence de l'estérase B et la virulence extra-intestinale. Son contexte génomique est, de plus, dépourvu de gènes impliqués dans la virulence. Nous avons également testé la croissance de mutants *aes* dans les souches K-12 MG1655 et CFT073 sur différentes sources de carbone. Nous n'avons pas observé de différences de croissance entre les souches mutantes et les souches sauvages.

En conclusion, cet article a établit que l'estérase B n'avait pas de lien direct avec la virulence, mais qu'il s'agissait d'un excellent marqueur phylogénétique. Certains gènes, comme *aes*, sont donc de bons marqueurs phylogénétiques, et permettent d'inférer la phylogénie sans disposer des génomes entiers. C'est pourquoi nous avons utilisé une sélection de gènes (MLST) pour étudier la répartition de la spécificité d'hôtes dans les groupes phylogénétiques à partir de 234 souches d'*E. coli*. Nous aborderons cette étude dans le chapitre suivant.

Chapitre V : Répartition de la spécificité d'hôte (humaine ou animale) dans les groupes phylogénétiques

1. Introduction

Comme nous l'avons déjà évoqué plus tôt dans ce manuscrit, *E. coli* est capable de nombreux modes de vie. Cette bactérie peut vivre de manière commensale dans l'intestin des vertébrés ou dans l'eau et les sédiments. Elle peut également être à l'origine de pathologies intra ou extra intestinales humaines ou animales. Nous avons montré que malgré l'important flux de gènes, et le taux important de conversion génique affectant le génome de cette bactérie, sa population était structurée et constituée d'au moins six groupes phylogénétiques bien distincts (Tenaillon, Skurnik et al. 2010). Dans ce contexte, il semble utile d'avoir une vision globale de la répartition dans ces groupes phylogénétiques des souches humaines et animales, commensales et pathogènes. Pour cela nous avons étudié 234 souches d'*E. coli*, les plus diversifiées possible. Parmi ces 234 souches, nous avions :

- Un panel de 35 souches pathogènes animales (8 oiseaux et 27 mammifères) engendrant des pathologies diverses : des ExPEC/APEC (Avian Pathogenic *E. coli*), et des InPEC (dont des ETEC, EPEC, EHEC et 3 non classées) en nombre approximativement équivalent.
- Un panel de 92 souches pathogènes humaines dont environ le même nombre d'ExPEC que d'InPEC (dont des ETEC, EPEC, EHEC, EAEC, EIEC, DAEC et 1 non classée).
- 45 souches commensales animales.
- 54 souches commensales humaines.
- 8 souches InPEC supplémentaires pour lesquelles leur génome complet était disponible.

Afin d'étudier la distribution des facteurs de virulence parmi ces souches, nous avons également recherché par PCR (« Polymerase Chain Reaction ») la présence de divers facteurs de virulence (impliqués dans les infections extra-

intestinales ou intra-intestinales) ainsi que d'adhésines classiquement associées aux souches pathogènes animales.

2. Article V (soumis à Applied and Environmental Microbiology)

3. Principaux résultats et perspectives

Une analyse factorielle des correspondances a été faite pour les 234 souches et les 35 variables suivantes : l'origine humaine ou animale, les caractères commensal, ExPEC et InPEC, l'appartenance à un des sept groupes phylogénétiques et la présence ou l'absence de 23 facteurs de virulence ou d'adhésines animales. Les ExPEC et les InPEC sont relativement bien séparés par le premier axe, par contre nous n'avons pas observé de séparation entre les souches d'origine animales et humaines. Les principales conclusions de cette analyse sont que les infections extra-intestinales sont principalement causées par les souches du groupe B2 qui possèdent un grand nombre de facteurs de virulence extra-intestinale alors que les infections intra-intestinales sont principalement causées par les souches des groupes A, B1 et E et contiennent certains facteurs de virulence intra-intestinale spécifiques (Fig 1 de l'article).

La phylogénie que j'ai construite à partir de 8 gènes du MLST par maximum de vraisemblance fait apparaître un nouveau groupe par rapport à ceux qui étaient préalablement décrits (Tenaillon, Skurnik et al. 2010) que nous avons appelé C (Moissenet, Salauze et al. 2010). Ce groupe correspond au groupe accessoire déjà nommé C en MLEE (Selander, Caugant et al. 1986) et MLST (Escobar-Paramo, Clermont et al. 2004) (Fig 13) et au CC66 (complexe clonal 66) (Jaureguy, Landraud et al. 2008) (Fig 12). Seule 5 des 234 souches n'appartiennent à aucun des 7 groupes principaux. Les groupes qui apparaissent les plus basaux sont les groupes F et B2. Le groupe F, proposé il y a peu par Jaureguy et collaborateurs, regroupe des souches préalablement assignées au groupe D, paraphylétique avant sa création (Jaureguy, Landraud et al. 2008). Les groupes ayant divergé le plus récemment semblent être les groupes A et B1/C. De plus, cette analyse à permis de définir 9 sous-groupes en plus du groupe EPEC1 dans le groupe phylogénétique B2 (Le Gall, Clermont et al. 2007). De la même manière, le groupe A a été divisé en 3 sous-groupes, le groupe B1 en 5 sous-groupes. Une autre conclusion importante de cette analyse est que nous avons clairement observé que les souches d'origines humaines et animales appartiennent aux mêmes sous-groupes phylogénétiques ou complexes clonaux. Certains des groupes sont en effet plutôt divers (par exemple les groupe D, A et B1), alors que d'autres apparaissent clonaux (par exemple les groupes C et E).

Les souches animales ExPEC non B2 semblent appartenir aux groupes D et C. Les souches animales InPEC, quant à elles, sont réparties dans les groupes A, B1, C et E. Les différents facteurs de virulence extra-intestinale présents ne permettent pas de séparer les souches animales ExPEC des autres, ni les souches animales et humaines de la même lignée. De la même façon, les facteurs de virulence intra-intestinale ne peuvent pas discriminer les souches animales et humaines. Par contre, les adhésines que nous avons recherchées sont essentiellement présentes dans les souches animales non B2.

Cette analyse montre donc clairement que les souches pathogènes humaines et animales partagent des fonds génétiques communs. Bien que les souches humaines et animales responsables d'une même pathologie dans les deux types d'hôtes partagent un pool commun de gènes de virulence, un ensemble d'adhésines spécifiques des souches animales non B2 a été identifié. Il semblerait donc qu'en plusieurs occasions, la spécification d'hôtes se soit effectuée par des changements génétiques subtils à partir d'un ancêtre commun proche. Il serait intéressant de tenter de les identifier par exemple à l'aide de séquençage haut-débit grâce aux méthodes de séquençage nouvelle génération.

SYNTHESE ET PERSPECTIVES

La diversité dans les génomes a pour origine deux mécanismes principaux : la mutation et la recombinaison. Nous avons étudié une forme de mutabilité qui n'avait pas encore été caractérisée à l'échelle d'un génome entier : la mutabilité transcriptionnelle. Nous avons montré que cette mutabilité avait une influence sur le génome et, de plus, qu'elle était soumise à sélection sur les sites synonymes. Cette sélection s'effectue globalement en privilégiant la robustesse du génome face à cette mutabilité. Dans le génome de la souche K-12 MG1655, 20% des gènes sont significativement plus stables qu'attendu vis à vis de la mutabilité transcriptionnelle. Nous avons mis en évidence que plus le MCU augmente (ceci jusqu'à une valeur de 0,7), plus la mutagénèse transcriptionnelle explique le fait qu'un site soit variable ou non entre K-12 MG1655 et CFT073 ainsi qu'entre K-12 MG1655 et EDL933 O157:H7. Après 0,7, on observe une diminution de l'impact de la mutagénèse transcriptionnelle. Parallèlement, plus le MCU augmente, plus la fraction de gènes significativement plus robuste qu'attendu en termes de mutagénèse transcriptionnelle augmente. Cette sélection sur la mutabilité transcriptionnelle agit sur le positionnement des codons synonymes le long de la séquence du gène. Il est donc probable que la sélection agissant pour le biais de codons et celle agissant pour diminuer la mutabilité transcriptionelle entrent en compétition dans le cas des gènes les plus exprimés. De plus, la sélection pour diminuer cette forme de mutabilité semble plus faible que celle agissant sur le biais de codons. Ces deux facteurs expliqueraient que les gènes les plus exprimés, ne présentent pas la plus faible mutabilité. Un autre facteur important est le taux de GC du gène. Les gènes à fort taux de GC présentent une mutabilité transcriptionnelle moindre. Ceci pourrait s'expliquer par le fait que ces bases sont davantage soumises aux altérations chimiques et que, par conséquent, la sélection pour diminuer la mutabilité transcriptionnelle y serait plus forte. Elle y serait également plus efficace car les bases G et C s'apparient avec 3 liaisons hydrogènes plutôt que 2 pour les bases A et T, ce qui permettrait la formation de structures secondaires thermodynamiquement plus stables.

La mutation est une des sources des différences entre les individus. Si elle est fixée dans la population, elle devient alors une des sources de différences entre les populations. Nous avons dans ce travail participé à la caractérisation d'une forme de sélection affectant les sites synonymes encore peu étudiée. Cette mutabilité présente l'avantage pour la bactérie de pouvoir être localement modifiée par le biais de la composition nucléotidique. La

sélection dont on peut observer la signature sur la séquence est la résultante des forces sélectives agissant sur celle-ci.

Ces forces sélectives varient selon différents paramètres. Par exemple, la taille de la population efficace est un paramètre important. En effet, les populations présentant une faible taille de population efficace sont caractérisées par une diminution de la sélection par rapport à la dérive. C'est pourquoi, par exemple, nous n'avons pas observé de sélection pour diminuer la mutabilité transcriptionnelle dans le génome de *Buchnera aphidicola* (pathogène intracellulaire). Il serait donc intéressant de faire ce même type d'analyse dans d'autres espèces présentant différentes tailles de population efficace ou différents modes de vie.

Très récemment, certains auteurs ont utilisé un autre indice pour étudier la mutagénèse transcriptionnelle en condition de stress uniquement (Kim, Lee et al. 2010). Par des approches de randomisations similaires à celles que nous avons utilisées, ces auteurs observent, comme nous, que la sélection affectant leur indice de mutabilité s'effectuait dans le sens de la robustesse pour les bases synonymes. Par contre lorsqu'ils effectuent la même analyse sur les sites non synonymes, ils observent une sélection dans le sens de l'augmentation de la mutabilité. En fait, cette analyse n'a d'intérêt que si on considère les situations de stress intense au cours desquelles, la nécessité de s'adapter est vitale. Dans ce cas, modifier une protéine pourrait effectivement être avantageux. Par contre, cette stratégie semble couteuse lorsque la bactérie n'est pas en condition de stress. Ils obtiennent cependant des conclusions intéressantes :

Par des corrélations de rang de Wilcoxon, ils comparent la sélection sur le codon d'initiation de la traduction et les bases qui suivent ; ils trouvent que la première méthionine est significativement plus robuste en termes de mutabilité transcriptionnelle liée au stress que les bases suivantes. Le phénomène est moins fort lorsqu'un codon d'initiation alternatif suit le premier de près (<10 codons). Ainsi, la pression de sélection pour augmenter la diversité protéique en condition de stress proposée par les auteurs, serait acompagnée d'une pression de sélection limitant le risque que ces changements affectent les codons d'initiation ce qui inactiverait le gène.

Ils regardent aussi l'impact d'une substitution sur l'indice de mutabilité (MI) de la base modifiée et des bases proches. Ils observent que si une base a un indice de mutabilité

fort, elle restera le plus souvent avec un indice de mutabilité fort lorsqu'elle mutera. Une mutation affecte la base et son entourage dans le même sens du point de vue de la valeur de l'indice de mutabilité, même si ce n'est pas avec la même intensité. Car, l'effet sur la base proprement dite est plus important que sur son entourage. Ils en concluent que les séquences codant les protéines chez *E. coli* ont évoluées pour contrôler la mutabilité transcriptionnelle liée au stress de manière à augmenter la diversité protéique tout en limitant l'inactivation des gènes par la mutation de leur codon d'initiation. De plus, une fois qu'une base a acquis une capacité à muter importante, celle-ci restera forte lorsqu'elle mutera.

Malgrès ces résultats intéressants, il subsiste, selon moi, une contradiction dans leur analyse. Le rééchantillonnage des séquences que les auteurs utilisent consiste, comme le notre, à échanger les codons synonymes entre eux. Donc les bases non synonymes sont restées identiques, seules les bases synonymes dans leur voisinage sont modifiées. Les auteurs montrent alors, comme nous, que les sites synonymes sont significativement plus robustes en termes de mutabilité transcriptionnelle que dans les séquences randomisées. Par contre, ils observent que les sites non synonymes sont quant à eux plus mutables qu'attendu. Pourtant, le signal qu'ils observent est lié uniquement aux changements des bases synonymes voisines. Or, ils ont eux-même montré que lorsqu'une base subissait une mutation, le changement de l'indice de mutabilité s'effectuait le plus souvent dans le même sens pour la base elle-même et pour les bases avoisinantes. Comment leur résultat est-il alors possible ? Je pense qu'il s'agit d'un biais dans leur calcul. En effet, nous avons montré que le signal majoritaire allait dans le sens de la diminution de la mutabilité transcriptionelle moyenne du gène, ceci avec notre indice de mutabilité, mais également avec l'indice de B. Wright que les auteurs utilisent (Wright, Reimers et al. 2002; Wright, Reschke et al. 2003). Ce qui se traduit par le fait que lorsque les séquences réelles sont significativement différentes en termes de mutabilité transcriptionnelle moyenne, elles apparaissent la plupart du temps plus stables (leur indice de mutabilité moyen est plus faible) que dans les séquences randomisées. Or les auteurs calculent des Z-score pour chacune des bases en effectuant le calcul suivant :

$(MI_{base} – moy(MI)_{gene})$ /écart-type

Puisqu'on s'attend à ce que la moyenne du gène réel soit inférieure à celle des séquences randomisées si elle est différente, le Z-score d'une base non silencieuse qui aurait un indice de mutabilité égale dans la séquence réelle et dans les séquences randomisées serait donc artéfactuellement supérieur dans la séquence réelle par rapports aux séquences contrôles. Ceci expliquerait leur résultat.

De plus, les auteurs précisent que leur indice ne permet pas de tenir compte du signal lié aux conditions favorables. En effet, l'indice de mutabilité qu'ils utilisent corrèle avec des données expérimentales de mutations obtenues en conditions de stress et pas avec les mutations spontanées observées chez *E. coli*. Ils critiquent rapidement notre indice de mutabilité transcriptionnelle qui englobe les mutations liées au stress et celles en conditions normales. En effet, ils nous reprochent de l'avoir validé grâce à la comparaison des bases variables et constantes entre deux *E. coli* et de ne pas avoir utilisé la séquence ancestrale pour cela. Je pense que le résultat aurait été très proche, mais qu'effectivement cette méthodologie aurait été conceptuellement plus appropriée.

Il serait intéressant de discriminer les effets des pressions de sélections liées aux conditions de stress de celles liées aux conditions normales, en élaborant un autre indice de mutabilité ne tenant pas compte des conditions de stress.

La recombinaison est également un des mécanismes à l'origine de la diversité génomique. Ce terme générique inclut différents mécanismes : les transferts latéraux interspécifiques mais également les événements de recombinaison homologues qui eux tendent à homogénéiser le génome de l'espèce. Ces deux types principaux désorganisent la phylogénie. Pourtant, nous avons montré que le flux important de gènes s'effectuait en des points précis : les points chauds d'intégration, ce qui limite leur impact en termes de désorganisation du génome. Au sein des gènes du « core » génome, la recombinaison homologue sous la forme de conversion génique est importante (2,5 fois le taux de mutation en moyenne). Nous avons pourtant mis en évidence que puisque les fragments mis en jeu sont très court (50 pbs) la phylogénie n'en était pas perturbée. Nous avons également mis en avant des points chauds d'incongruence dans lesquels ces évènements se concentrent. Il arrive que ces sites bornent les points chauds d'intégration. En effet, nous avons proposé que l'insertion par transferts horizontaux d'une séquence puisse ensuite être suivie de la

propagation de cette séquence par recombinaison homologue lorsque celle-ci procure un avantage sélectif (Schubert, Darlu et al. 2009).

Ces approches phylogénétiques et phylogénomiques posent le problème de la qualité des méthodes de reconstruction d'histoire évolutive. En effet, il est tout à fait probable que les modèles évolutifs actuels, même ceux ayant le plus grands nombre de paramètres, ne suffisent pas à prendre en compte toute les forces évolutives en jeu. Cette problématique est d'autant plus importante lorsqu'on souhaite reconstruire l'arbre de l'espèce avec son génome entier.

C'est pour cette raison que les approches basées sur le MLST sont encore largement utilisées. Pourtant ces approches posent d'autres biais liés au choix des gènes utilisés. Le fait que les deux approches (MLST et génomes entiers) soient cohérentes est très encourageant et traduit la robustesse de la structure clonale de l'espèce. Nous nous sommes d'ailleurs intéressés au gène codant l'estérase B dont le polymorphisme électrophorétique reflète la divergence entre B2 et non B2. Nous avons établi à partir des résultats du séquençage nucléotidique et du modèle murin de septicémie expérimentale, que l'estérase B est un excellent marqueur de phylogénie mais n'est pas un facteur de virulence extraintestinale. Le caractère basal du groupe B2, l'important flux de gène qui le caractérise (Touchon, Hoede et al. 2009) et une diversité nucléotidique importante sembleraient indiquer qu'il pourrait s'agir d'une sous-espèce. Le polymorphisme électrophorétique de l'estérase B étaye cette hypothèse.

Nous avons détaillé un des points chauds d'intégration car il comportait de nombreux gènes de résistance aux antibiotiques caractéristiques du CGA dans la souche UMN026. Cette analyse a mis en exergue la composition en mosaïque de ces îlots d'intégration. Tous les génomes comparés contiennent, à cette même position, différents gènes. Cette répartition ne correspond ni à la phylogénie ni au pathotype. Nous avons observé de nombreux éléments liés à la mobilité tels que des intégrases et des transposases qui pourraient expliquer ces nombreux réarrangements.

La population d'*E. coli* est structurée. Nous avons étudié 234 souches par une approche basée sur le MLST et nous avons finalement observé 7 groupes principaux pouvant pour certains d'entre eux contenir des sous-groupes. Le groupe B2 est celui qui en contient

le plus grand nombre : nous avons défini 10 sous-groupes. A l'intérieur de ces sous-groupes on trouve des souches animales et humaines, ce qui démontre que celles-ci partagent un fond génétique commun. L'abondance des sous-groupes du groupe B2 est un autre argument étayant l'idée que ce groupe pourrait constituer une sous-espèce. Les facteurs impliqués dans la spécificité d'hôte sont encore inconnus. Etant donné la proximité phylogénétique des souches animales et humaines, les facteurs génétiques, s'ils existent, doivent être en faible nombre ou constitués de modifications subtiles. Il serait donc également intéressant d'étudier les facteurs épigénétiques qui sont les informations héritables ne pouvant pas être expliqués par des modifications de la séquence.

Avec le développement des méthodes de séquençage à haut débit (dits de nouvelle génération), leur amélioration rapide ainsi que la diminution des coûts, il sera bientôt possible de séquencer des centaines de génomes par espèces. Pour analyser de tels types de données, des logiciels dédiés voient peu à peu le jour. Le MLST est donc voué à disparaître pour être remplacé par le SNP (« single nucleotide polymorphism ») à l'échelle du génome entier et l'étude de la structure des populations ne se fera plus par génétique des populations mais bien par génomique des populations. De la même façon, lorsqu'on s'intéressera aux relations phylogénétiques existant entre les populations, la phylogénomique des populations se développera.

BIBLIOGRAPHIE

Allers, T. and M. Lichten (2001). "Differential timing and control of noncrossover and crossover recombination during meiosis." Cell **106**(1): 47-57.

Anderson, G. G., S. M. Martin, et al. (2004). "Host subversion by formation of intracellular bacterial communities in the urinary tract." Microbes Infect **6**(12): 1094-101.

Atwood, K. C., L. K. Schneider, et al. (1951). "Periodic selection in Escherichia coli." Proc Natl Acad Sci U S A **37**(3): 146-55.

Balbi, K. J., E. P. Rocha, et al. (2009). "The temporal dynamics of slightly deleterious mutations in Escherichia coli and Shigella spp." Mol Biol Evol **26**(2): 345-55.

Bambou, J., A. Giraud, et al. (2006). "La flore intestinale commensale : la balance sans le glaive ?" Journal de la société de biologie **200**(2): 113-120.

Bentley, R. and R. Meganathan (1982). "Biosynthesis of vitamin K (menaquinone) in bacteria." Microbiol Rev **46**(3): 241-80.

Berg, R. D. (1996). "The indigenous gastrointestinal microflora." Trends Microbiol **4**(11): 430-5.

Bergthorsson, U. and H. Ochman (1995). "Heterogeneity of genome sizes among natural isolates of Escherichia coli." J Bacteriol **177**(20): 5784-9.

Bisercic, M., J. Y. Feutrier, et al. (1991). "Nucleotide sequences of the gnd genes from nine natural isolates of Escherichia coli: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic gnd locus." J Bacteriol **173**(12): 3894-900.

Bjedov, I., G. Lecointre, et al. (2003). "Polymorphism of genes encoding SOS polymerases in natural populations of Escherichia coli." DNA Repair (Amst) **2**(4): 417-26.

Bjedov, I., O. Tenaillon, et al. (2003). "Stress-induced mutagenesis in bacteria." Science **300**(5624): 1404-9.

Blattner, F. R., G. Plunkett, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12." Science **277**(5331): 1453-62.

Bonacorsi, S. and E. Bingen (2005). "Molecular epidemiology of Escherichia coli causing neonatal meningitis." Int J Med Microbiol **295**(6-7): 373-81.

Bonacorsi, S., O. Clermont, et al. (2003). "Molecular analysis and experimental virulence of French and North American Escherichia coli neonatal meningitis isolates: identification of a new virulent clone." J Infect Dis **187**(12): 1895-906.

Brochet, M., C. Rusniok, et al. (2008). "Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of Streptococcus agalactiae." Proc Natl Acad Sci U S A **105**(41): 15961-6.

Bzymek, M. and S. T. Lovett (2001). "Instability of repetitive DNA sequences: the role of replication in multiple mechanisms." Proc Natl Acad Sci U S A **98**(15): 8319-25.

Castellanos, M. and D. Romero (2009). "The extent of migration of the Holliday junction is a crucial factor for gene conversion in Rhizobium etli." J Bacteriol **191**(15): 4987-95.

Charlesworth, B., M. T. Morgan, et al. (1993). "The effect of deleterious mutations on neutral molecular variation." Genetics **134**(4): 1289-303.

Clermont, O., M. Lescat, et al. (2008). "Evidence for a human-specific Escherichia coli clone." Environ Microbiol **10**(4): 1000-6.

Conway, T., K. A. Krogfelt, et al. (2004). The life of commensal Escherichia coli in the mammalian intestine. Escherichia coli and Salmonella: cellular and molecular biology [3rd edition, Online.]. F. C. Neidhardt, Curtiss R. III, Ingraham L.et al. Washington DC, ASM Press.

Crockett, C. S., C. N. Haas, et al. (1996). "Prevalence of shigellosis in the U.S.: consistency with dose-response information." Int J Food Microbiol **30**(1-2): 87-99.

Cromie, G. A., J. C. Connelly, et al. (2001). "Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans." Mol Cell **8**(6): 1163-74.

Deitsch, K. W., E. R. Moxon, et al. (1997). "Shared themes of antigenic variation and virulence in bacterial, protozoal, and fungal infections." Microbiol Mol Biol Rev **61**(3): 281-93.

Denamur, E., G. Lecointre, et al. (2000). "Evolutionary implications of the frequent horizontal transfer of mismatch repair genes." Cell **103**(5): 711-21.

Denamur, E., B. Picard, et al. (2010). Population genetics of pathogenic Escherichia coli. In Bacterial population genetics in infectious disease. F. D. Robinson DA, Feil EJ, Wiley-Blackwell**:** 269-286.

Donelson, J. E. (1995). "Mechanisms of antigenic variation in Borrelia hermsii and African trypanosomes." J Biol Chem **270**(14): 7783-6.

Drake, J. W. (1991). "A constant rate of spontaneous mutation in DNA-based microbes." Proc Natl Acad Sci U S A **88**(16): 7160-4.

DuBose, R. F., D. E. Dykhuizen, et al. (1988). "Genetic exchange among natural isolates of bacteria: recombination within the phoA gene of Escherichia coli." Proc Natl Acad Sci U S A **85**(18): 7036-40.

Dykhuizen, D. E. and L. Green (1991). "Recombination in Escherichia coli and the definition of biological species." J Bacteriol **173**(22): 7257-68.

Escobar-Paramo, P., O. Clermont, et al. (2004). "A specific genetic background is required for acquisition and expression of virulence factors in Escherichia coli." Mol Biol Evol **21**(6): 1085-94.

Escobar-Paramo, P., C. Giudicelli, et al. (2003). "The evolutionary history of Shigella and enteroinvasive Escherichia coli revised." J Mol Evol **57**(2): 140-8.

Escobar-Paramo, P., A. Sabbagh, et al. (2004). "Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the Escherichia coli case study." Mol Phylogenet Evol **30**(1): 243-50.

Friedberg, E. C., G. C. Walker, et al., Eds. (1995). DNA Repair and Mutagenesis. ASM Press, Washington DC.

Funchain, P., A. Yeung, et al. (2000). "The consequences of growth of a mutator strain of Escherichia coli as measured by loss of function among multiple gene targets and loss of fitness." Genetics **154**(3): 959-70.

Galtier, N., G. Piganeau, et al. (2001). "GC-content evolution in mammalian genomes: the biased gene conversion hypothesis." Genetics **159**(2): 907-11.

Goldman, N., J. P. Anderson, et al. (2000). "Likelihood-based tests of topologies in phylogenetics." Syst Biol **49**(4): 652-70.

Gordon, D. M., O. Clermont, et al. (2008). "Assigning Escherichia coli strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method." Environ Microbiol **10**(10): 2484-96.

Goris, J., K. T. Konstantinidis, et al. (2007). "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities." Int J Syst Evol Microbiol **57**(Pt 1): 81-91.

Goullet, P. (1980). "Esterase electrophoretic pattern relatedness between Shigella species and Escherichia coli." J Gen Microbiol **117**(2): 493-500.

Goullet, P. and B. Picard (1989). "Comparative electrophoretic polymorphism of esterases and other enzymes in Escherichia coli." J Gen Microbiol **135**(1): 135-43.

Grantham, R., C. Gautier, et al. (1981). "Codon catalog usage is a genome strategy modulated for gene expressivity." Nucleic Acids Res **9**(1): r43-74.

Grindley, N. D., K. L. Whiteson, et al. (2006). "Mechanisms of site-specific recombination." Annu Rev Biochem **75**: 567-605.

Gross, M. D. and E. C. Siegel (1981). "Incidence of mutator strains in Escherichia coli and coliforms in nature." Mutat Res **91**(2): 107-10.

Guttman, D. S. and D. E. Dykhuizen (1994). "Clonal divergence in Escherichia coli as a result of recombination, not mutation." Science **266**(5189): 1380-3.

Haber, J. E. (1998). "Mating-type gene switching in Saccharomyces cerevisiae." Annu Rev Genet **32**: 561-99.

Haber, J. E. (2007). "Evolution of models of homologous recombination." Genome Dynamics & Stability.

Hall, B. G. and P. M. Sharp (1992). "Molecular population genetics of Escherichia coli: DNA sequence diversity at the celC, crr, and gutB loci of natural isolates." Mol Biol Evol **9**(4): 654-65.

Harrington, S. M., E. G. Dudley, et al. (2006). "Pathogenesis of enteroaggregative Escherichia coli infection." FEMS Microbiol Lett **254**(1): 12-8.

Hashimoto, J. G., B. S. Stevenson, et al. (2003). "Rates and consequences of recombination between rRNA operons." J Bacteriol **185**(3): 966-72.

Hayashi, T., K. Makino, et al. (2001). "Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12." DNA Res **8**(1): 11-22.

Hershberg, R., H. Tang, et al. (2007). "Reduced selection leads to accelerated gene loss in Shigella." Genome Biol **8**(8): R164.

Herzer, P. J., S. Inouye, et al. (1990). "Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of Escherichia coli." J Bacteriol **172**(11): 6175-81.

Hoeijmakers, J. H. (2001). "Genome maintenance mechanisms for preventing cancer." Nature **411**(6835): 366-74.

Holliday, R. (1964). "A mechanism for gene conversion." Genetic Research **5**: 282-304.

Hooper, L. V., M. H. Wong, et al. (2001). "Molecular analysis of commensal host-microbial relationships in the intestine." Science **291**(5505): 881-4.

Houdouin, V., S. Bonacorsi, et al. (2008). "[Clinical outcome and bacterial characteristics of 99 Escherichia coli meningitis in young infants]." Arch Pediatr **15 Suppl 3**: S138-47.

Hudault, S., J. Guignot, et al. (2001). "Escherichia coli strains colonising the gastrointestinal tract protect germfree mice against Salmonella typhimurium infection." Gut **49**(1): 47-55.

Hudault, S., O. B. Spiller, et al. (2004). "Human diffusely adhering Escherichia coli expressing Afa/Dr adhesins that use human CD55 (decay-accelerating factor) as a receptor does not bind the rodent and pig analogues of CD55." Infect Immun **72**(8): 4859-63.

Hutchinson, F. (1996). Mutagenesis. Escherichia coli and Salmonella: cellular and molecular biology. F. C. Neidhardt, Curtiss R. III, Ingraham L.et al. Washington DC, ASM Press. **118:** 2218–2235

Jaureguy, F., L. Landraud, et al. (2008). "Phylogenetic and genomic diversity of human bacteremic Escherichia coli strains." BMC Genomics **9**: 560.

Jin, Q., Z. Yuan, et al. (2002). "Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157." Nucleic Acids Res **30**(20): 4432-41.

Johnson, J. R. (1991). "Virulence factors in Escherichia coli urinary tract infection." Clin Microbiol Rev **4**(1): 80-128.

Johnson, J. R., A. R. Manges, et al. (2002). "A disseminated multidrug-resistant clonal group of uropathogenic Escherichia coli in pyelonephritis." Lancet **359**(9325): 2249-51.

Johnson, J. R., A. C. Murray, et al. (2005). "Distribution and characteristics of Escherichia coli clonal group A." Emerg Infect Dis **11**(1): 141-5.

Johnson, J. R., K. L. Owens, et al. (2006). "Phylogenetic relationships among clonal groups of extraintestinal pathogenic Escherichia coli as assessed by multi-locus sequence analysis." Microbes Infect **8**(7): 1702-13.

Jyssum, K. (1960). "Observations on two types of genetic instability in Escherichia coli." Acta Pathol Microbiol Scand **48**: 113-20.

Kauffmann, F. (1947). "The serology of the coli group." J Immunol **57**(1): 71-100.

Kim, H., B. S. Lee, et al. (2010). "Transcription-associated mutagenesis increases protein sequence diversity more effectively than does random mutagenesis in Escherichia coli." PLoS One **5**(5): e10567.

Kishino, H. and M. Hasegawa (1989). "Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea." J Mol Evol **29**(2): 170-9.

Kogoma, T. (1997). "Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription." Microbiol Mol Biol Rev **61**(2): 212-38.

Kowalczykowski, S. C., D. A. Dixon, et al. (1994). "Biochemistry of homologous recombination in Escherichia coli." Microbiol Rev **58**(3): 401-65.

Kudla, G., A. W. Murray, et al. (2009). "Coding-sequence determinants of gene expression in Escherichia coli." Science **324**(5924): 255-8.

Lawrence, J. G., H. Ochman, et al. (1991). "Molecular and evolutionary relationships among enteric bacteria." J Gen Microbiol **137**(8): 1911-21.

Le Gall, T., O. Clermont, et al. (2007). "Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group Escherichia coli strains." Mol Biol Evol **24**(11): 2373-84.

LeClerc, J. E., B. Li, et al. (1996). "High mutation frequencies among Escherichia coli and Salmonella pathogens." Science **274**(5290): 1208-11.

Lecointre, G., L. Rachdi, et al. (1998). "Escherichia coli molecular phylogeny using the incongruence length difference test." Mol Biol Evol **15**(12): 1685-95.

Lloyd, R. G. and K. B. Low (1996). Homologous Recombination. Escherichia coli and Salmonella: cellular and molecular biology, vol. 2. F. C. Neidhardt, Curtiss R. III, Ingraham L.et al. Washington DC, ASM Press**:** 2236-2255.

Maki, H. and M. Sekiguchi (1992). "MutT protein specifically hydrolyses a potent mutagenic substrate for DNA synthesis." Nature **355**(6357): 273-5.

Manges, A. R., J. R. Johnson, et al. (2001). "Widespread distribution of urinary tract infections caused by a multidrug-resistant Escherichia coli clonal group." N Engl J Med **345**(14): 1007-13.

Matic, I., M. Radman, et al. (1997). "Highly variable mutation rates in commensal and pathogenic Escherichia coli." Science **277**(5333): 1833-4.

Maynard-Smith, J., N. H. H. Smith, et al. (1993). "How clonal are bacteria?" Proc. Natl. Acad. Sci. USA **90**: 4384-4388.

Meselson, M. S. and C. M. Radding (1975). "A general model for genetic recombination." Proc Natl Acad Sci U S A **72**(1): 358-61.

Michel, B. (1999). Illegitimate recombination in bacteria. In Organization of the prokaryotic genome. R. I. Charlebois. ASM Press, Washington DC**:** 129-150.

Milkman, R. (1973). "Electrophoretic variation in Escherichia coli from natural sources." Science **182**(116): 1024-6.

Milkman, R. and M. M. Bridges (1990). "Molecular evolution of the Escherichia coli chromosome. III. Clonal frames." Genetics **126**(3): 505-17.

Milkman, R. and M. M. Bridges (1993). "Molecular evolution of the Escherichia coli chromosome. IV. Sequence comparisons." Genetics **133**(3): 455-68.

Milkman, R. and I. P. Crawford (1983). "Clustered third-base substitutions among wild strains of Escherichia coli." Science **221**(4608): 378-80.

Milkman, R., E. Jaeger, et al. (2003). "Molecular evolution of the Escherichia coli chromosome. VI. Two regions of high effective recombination." Genetics **163**(2): 475-83.

Milkman, R. and A. Stoltzfus (1988). "Molecular evolution of the Escherichia coli chromosome. II. Clonal segments." Genetics **120**(2): 359-66.

Miller, J. H. (1996). "Spontaneous mutators in bacteria: insights into pathways of mutagenesis and repair." Annu Rev Microbiol **50**: 625-43.

Mitsuoka, T. and K. Hayakawa (1973). "[The fecal flora in man. I. Composition of the fecal flora of various age groups]." Zentralbl Bakteriol Orig A **223**(2): 333-42.

Moissenet, D., B. Salauze, et al. (2010). "Meningitis caused by Escherichia coli producing TEM-52 extended-spectrum beta-lactamase within an extensive outbreak in a neonatal ward: epidemiological investigation and characterization of the strain." J Clin Microbiol **48**(7): 2459-63.

Motamedi, M. R., S. K. Szigety, et al. (1999). "Double-strand-break repair recombination in Escherichia coli: physical evidence for a DNA replication mechanism in vivo." Genes Dev **13**(21): 2889-903.

Moxon, E. R., P. B. Rainey, et al. (1994). "Adaptive evolution of highly mutable loci in pathogenic bacteria." Curr Biol **4**(1): 24-33.

Nataro, J. P. and J. B. Kaper (1998). "Diarrheagenic Escherichia coli." Clin Microbiol Rev **11**(1): 142-201.

Neidhart, F. C., R. I. Curtiss, et al. (1996). Escherichia coli and Salmonella typhimurium: cellular and molecular biology. Washington DC, ASM Press.

Nelson, K., T. S. Whittam, et al. (1991). "Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (gapA) in natural populations of Salmonella and Escherichia coli." Proc Natl Acad Sci U S A **88**(15): 6667-71.

Nie, H., F. Yang, et al. (2006). "Complete genome sequence of Shigella flexneri 5b and comparison with Shigella flexneri 2a." BMC Genomics **7**: 173.

Niyogi, S. K. (2005). "Shigellosis." J Microbiol **43**(2): 133-43.

Ochman, H. (2003). "Neutral mutations and neutral substitutions in bacterial genomes." Mol Biol Evol **20**(12): 2091-6.

Ochman, H., J. G. Lawrence, et al. (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**(6784): 299-304.

Ochman, H., E. Lerat, et al. (2005). "Examining bacterial species under the specter of gene transfer and exchange." Proc Natl Acad Sci U S A **102 Suppl 1**: 6595-9.

Ochman, H. and R. K. Selander (1984). "Standard reference strains of Escherichia coli from natural populations." J Bacteriol **157**(2): 690-3.

Ogura, Y., T. Ooka, et al. (2009). "Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli." Proc Natl Acad Sci U S A **106**(42): 17939-44.

Orskov, F., I. Orskov, et al. (1976). "Special Escherichia coli serotypes among enterotoxigenic strains from diarrhoea in adults and children." Med Microbiol Immunol **162**(2): 73-80.

Parsot, C. (2005). "Shigella spp. and enteroinvasive Escherichia coli pathogenicity factors." FEMS Microbiol Lett **252**(1): 11-8.

Peist, R., A. Koch, et al. (1997). "Characterization of the aes gene of Escherichia coli encoding an enzyme with esterase activity." J Bacteriol **179**(24): 7679-86.

Penders, J., C. Thijs, et al. (2006). "Factors influencing the composition of the intestinal microbiota in early infancy." Pediatrics **118**(2): 511-21.

Perna, N. T., G. Plunkett, 3rd, et al. (2001). "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7." Nature **409**(6819): 529-33.

Peters, J. E. and N. L. Craig (2001). "Tn7: smarter than we thought." Nat Rev Mol Cell Biol **2**(11): 806-14.

Power, M. L., J. Littlefield-Wyer, et al. (2005). "Phenotypic and genotypic characterization of encapsulated Escherichia coli isolated from blooms in two Australian lakes." Environ Microbiol **7**(5): 631-40.

Pupo, G. M., R. Lan, et al. (2000). "Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics." Proc Natl Acad Sci U S A **97**(19): 10567-72.

Rastegar Lari, A., F. Gold, et al. (1990). "Implantation and in vivo antagonistic effects of antibiotic-susceptible Escherichia coli strains administered to premature newborns." Biol Neonate **58**(2): 73-8.

Reid, S. D., C. J. Herbelin, et al. (2000). "Parallel evolution of virulence in pathogenic Escherichia coli." Nature **406**(6791): 64-7.

Rosen, D. A., T. M. Hooton, et al. (2007). "Detection of intracellular bacterial communities in human urinary tract infection." PLoS Med **4**(12): e329.

Russo, T. A. and J. R. Johnson (2000). "Proposal for a new inclusive designation for extraintestinal pathogenic isolates of Escherichia coli: ExPEC." J Infect Dis **181**(5): 1753-4.

Santoyo, G. and D. Romero (2005). "Gene conversion and concerted evolution in bacterial genomes." FEMS Microbiol Rev **29**(2): 169-83.

Savageau, M. A. (1983). "Escherichia coli habitats, cell types, and molecular mechanisms of gene control." The American Naturalist **122**: 732–744

Schierup, M. H. and J. Hein (2000). "Consequences of recombination on traditional phylogenetic analysis." Genetics **156**(2): 879-91.

Schubert, S., P. Darlu, et al. (2009). "Role of intraspecies recombination in the spread of pathogenicity islands within the Escherichia coli species." PLoS Pathog **5**(1): e1000257.

Selander, R. K., D. A. Caugant, et al. (1986). "Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics." Appl Environ Microbiol **51**(5): 873-84.

Selander, R. K. and B. R. Levin (1980). "Genetic diversity and structure in Escherichia coli populations." Science **210**(4469): 545-7.

Servin, A. L. (2005). "Pathogenesis of Afa/Dr diffusely adhering Escherichia coli." Clin Microbiol Rev **18**(2): 264-92.

Sharp, P. M. and W. H. Li (1986). "An evolutionary perspective on synonymous codon usage in unicellular organisms." J Mol Evol **24**(1-2): 28-38.

Shimodaira, H. and M. Hasegawa (1999). "Multiple comparisons of log-likelihoods with applications to phylogenetic inference." Mol Biol Evol **16**: 1114–1116.

Shiraishi, K., Y. Imai, et al. (2005). "Rep helicase suppresses short-homology-dependent illegitimate recombination in Escherichia coli." Genes Cells **10**(11): 1015-23.

Slanetz, L. W. and C. H. Bartley (1957). "Numbers of enterococci in water, sewage, and feces determined by the membrane filter technique with an improved medium." J Bacteriol **74**(5): 591-5.

Smith, G. R. (1988). "Homologous recombination in procaryotes." Microbiol Rev **52**(1): 1-28.

Solo-Gabriele, H. M., M. A. Wolfert, et al. (2000). "Sources of Escherichia coli in a coastal subtropical environment." Appl Environ Microbiol **66**(1): 230-7.

Stackebrandt, E., W. Frederiksen, et al. (2002). "Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology." Int J Syst Evol Microbiol **52**(Pt 3): 1043-7.

Strimmer, K. and A. Rambaut (2002). "Inferring confidence sets of possibly misspecified gene trees." Proc Biol Sci **269**(1487): 137-42.

Szostak, J. W., T. L. Orr-Weaver, et al. (1983). "The double-strand-break repair model for recombination." Cell **33**(1): 25-35.

Tajima, F. (1989). "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." Genetics **123**(3): 585-95.

Tenaillon, O., E. Denamur, et al. (2004). "Evolutionary significance of stress-induced mutagenesis in bacteria." Trends Microbiol **12**(6): 264-70.

Tenaillon, O., D. Skurnik, et al. (2010). "The population genetics of commensal Escherichia coli." Nat Rev Microbiol **8**(3): 207-17.

Tenaillon, O., B. Toupance, et al. (1999). "Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria." Genetics **152**(2): 485-93.

Touchon, M., C. Hoede, et al. (2009). "Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths." PLoS Genet **5**(1): e1000344.

Touzain, F., E. Denamur, et al. (2010). "Small variable segments constitute a major type of diversity of bacterial genomes at the species level." Genome Biol **11**(4): R45.

Verbeek, B., T. D. Southgate, et al. (2008). "O6-Methylguanine-DNA methyltransferase inactivation and chemotherapy." Br Med Bull **85**: 17-33.

Vollaard, E. J. and H. A. Clasener (1994). "Colonization resistance." Antimicrob Agents Chemother **38**(3): 409-14.

Wayne, L. G., D. J. Brenner, et al. (1987). "Report of the ad hoc committee on reconciliation of approaches to bacterial systematics." Int. J. Syst. Bacteriol. **37**: 463-464.

Wei, J., M. B. Goldberg, et al. (2003). "Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T." Infect Immun **71**(5): 2775-86.

Weisberg, R. L. A. (1983). Site specific recombination in phage lambda. Lambda II. C. S. H. Laboratory. Cold Spring Harbor, New York**:** 211-250.

Welch, R. A., V. Burland, et al. (2002). "Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli." Proc Natl Acad Sci U S A **99**(26): 17020-4.

Wirth, T., D. Falush, et al. (2006). "Sex and virulence in Escherichia coli: an evolutionary perspective." Mol Microbiol **60**(5): 1136-51.

Wright, B. E., J. M. Reimers, et al. (2002). "Hypermutable bases in the p53 cancer gene are at vulnerable positions in DNA secondary structures." Cancer Research **62**: 5641–5644.

Wright, B. E., D. K. Reschke, et al. (2003). "Predicting mutation frequencies in stem-loop structures of derepressed genes: Implications for evolution." Mol Microbiol **48**: 429-441.

Résumé

Escherichia coli constitue la majeure partie de la flore microbienne commensale aéro-anaérobie du tube digestif de l'hôte. Pourtant *E. coli* est aussi une des espèces les plus fréquemment rencontrées en pathologie humaine et animale. C'est l'une des espèces bactériennes les plus étudiées et les plus connues. L'évolution des génomes au sein de l'espèce repose sur deux mécanismes distincts : la mutation et la recombinaison, qui génèrent une diversité génétique sur laquelle la sélection naturelle peut opérer. Dans notre travail, nous nous sommes intéressés à ces processus et aux traces qu'ils laissent dans les génomes. Nous avons, en premier lieu, décrit une forme de mutabilité variable le long du génome car liée à l'existence de structure secondaire locale de l'ADN : la mutabilité transcriptionnelle. Nous avons pu d'une part quantifier cette mutagenèse et d'autre part révéler une réponse sélective au sein du génome pour en limiter les effets. La recombinaison, quant à elle, est connue pour brouiller le signal phylogénétique de manière importante. En second lieu, nous avons montré par une approche de génomique comparative que, malgré un taux relativement élevé de recombinaison, l'établissement d'une phylogénie robuste était possible. De plus, nous avons mis en évidence que les nombreuses acquisitions et pertes de gènes dans le génome des différentes souches d'*E. coli* se situaient préférentiellement à certains sites. Enfin, nous avons utilisé la structure phylogénétique de l'espèce à des applications taxonomiques et épidémiologiques.

Escherichia coli represents the major part of commensal aero-anaerobic microbiota of the host's digestive tract. Though, *E. coli* is also one of the most frequently encountered species in human and animal pathology. This is one of the most studied and the best known bacterial species. The evolution of genomes within the species is based on two distinct mechanisms: mutation and recombination that generate genetic diversity on which natural selection can operate. In our work, we were interested in those processes and the traces they leave in the genomes. We have at first described a form of variable mutability along the genome which is linked to the existence of local secondary DNA structure: the transcriptional mutability. We were able to quantify this mutagenesis and reveal a selective

response in the genome to limit its effects. Recombination is known to blur the phylogenetic signal significantly. Then, we have shown by a comparative genomics approach that, despite a relatively high recombination rate, the establishment of a robust phylogeny was possible. In addition, we have shown that the many acquisitions and loss of genes occurring in the genomes of different *E. coli* strains were located preferentially at certain sites. Lastly, we have used the phylogenetic structure of the species to study taxonomic and epidemiologic applications.

Mots clefs : mutagénèse, conversion génique, sélection naturelle, phylogénie, dynamique du génome.