L'évolution des phages tempérés d'entérobactéries

Louis-Marie Bobay

Université Pierre et Marie Curie

Ecole doctorale Complexité du Vivant

Génomique évolutive des microbes. Institut Pasteur. CNRS UMR 3525

L'évolution des phages tempérés d'entérobactéries

Par Louis-Marie Bobay

Thèse de doctorat de Biologie

Dirigée par Eduardo PC Rocha et Marie Touchon

Présentée et soutenue publiquement le 8 juillet 2014

Devant un jury composé de :

Dr. Mireille Ansaldi – Rapporteur
Dr. Vincent Daubin – Rapporteur
Prof. Patrick Forterre – Examinateur
Prof. Dominique Higuet – Examinateur
Prof. Sylvain Moineau – Examinateur

Résumé et mots clés

L'intégration et la dégradation des virus (ou phages) au sein des génomes bactériens (alors nommés prophages) constituent un flux de gènes promouvant la diversification génétique de leurs hôtes. Les mécanismes à l'origine de l'impact évolutif des prophages sur leurs hôtes restent cependant mal compris. Les génomes bactériens sont des entités fortement contraintes par différents niveaux d'organisation génétique et structurelle. La première partie de la thèse s'est attachée à comprendre comment les prophages sont adaptés aux génomes d'*Escherichia* et de *Salmonella*. Ces résultats ont mis en évidence une forte conservation des positions d'intégration des prophages, ainsi que différentes adaptations des phages tempérés à l'organisation chromosomique de leurs hôtes. L'origine de la diversité génétique des phages tempérés a été au centre de la deuxième partie de la thèse. L'étude des phages lambdoïdes d'entérobactéries a révélé l'existence de deux stratégies de recombinaison chez ces phages: utilisation du système de recombinaison RecBCD de l'hôte via la présence de sites Chi ou utilisation de leur propre système de recombinaison. Cette étude suggère que l'utilisation de l'une ou l'autre des stratégies de recombinaison a des impacts importants sur la diversification et le mosaïcisme génomique de ces phages. Enfin, la détection et l'analyse de prophages hérités verticalement au sein des génomes hôtes ont mis en évidence que de nombreux prophages partiellement dégradés sont conservés et évoluent sous sélection purificatrice. Ces résultats suggèrent que de nombreux prophages sont potentiellement des éléments fonctionnels domestiqués par la bactérie. L'ensemble de ces analyses permet de préciser un peu plus les mécanismes permettant aux prophages de contribuer à la diversification des répertoires de gènes de leurs hôtes.

Mots clés: procaryotes, bactéries, virus, phages tempérés, prophages, évolution, domestication, génomique comparative.

Table des matières

Introduction

1 Organisation et évolution du génome bactérien

1.1 Diversité des bactéries et de leurs génomes

Les bactéries sont très diverses sous différents aspects. Il a été estimé qu'il existerait entre 10^7 et 10^9 espèces bactériennes (Curtis, et al. 2002) et ces évaluations pourraient être largement sous-estimées (Schloss and Handelsman 2004). Les bactéries sont ubiquitaires et sont présentes dans des biotopes extrêmement variés: sols, océans, intestins animaux, etc. Les conditions physico-chimiques liées à ces environnements (salinité, température, etc) semblent influer sur la diversité bactérienne (Lozupone and Knight 2007). Les bactéries peuvent entretenir des associations diverses avec d'autres organismes, telles que le parasitisme, le mutualisme et le commensalisme. Il existe en outre différents types de mutualismes (symbiotiques, non-symbiotiques, facultatifs ou obligatoires) et de parasitismes (opportunistes ou obligatoires). Ces différents modes de vie se rencontrent souvent au sein d'un même genre bactérien, voire au sein de la même espèce (Dobrindt, et al. 2004; Vissa and Brennan 2001). Ceci souligne le fait que des bactéries évolutivement proches ont la capacité de s'adapter à des modes de vie très différents.

Le génome contient l'ensemble du matériel génétique d'un organisme et en particulier, ses gènes. L'origine de la diversité phénotypique bactérienne est largement liée à l'évolution et à la diversité de leurs génomes (Dobrindt, et al. 2004). La taille des génomes bactériens est extrêmement variable: allant d'environ 100kb jusqu'à 13Mb (Chang, et al. 2011; McCutcheon and Moran 2012). Ces génomes sont très compacts, et présentent une densité de gènes d'environ 85% (Mira, et al. 2001). La taille d'un génome est directement proportionnelle au nombre de gènes qu'il code. Cette variation de taille, et donc du nombre de gènes, semble en partie associée à des traits écologiques (Ochman and Davalos 2006). En effet, les bactéries environnementales, qui ne dépendent pas d'un hôte, présentent typiquement des génomes de grande taille. Les pathogènes facultatifs présentent en revanche des tailles de génome plus réduites (Ochman and Davalos 2006). Il a été établi que les symbiontes obligatoires tendent à évoluer vers une forte réduction génomique. Ceci s'explique par le fait que certaines fonctions cellulaires, notamment métaboliques, sont suppléées par celles de l'hôte. Suite à

l'établissement de telles relations, le génome du parasite ou du symbionte tend donc à perdre les gènes codants pour ces fonctions redondantes qui échappent ainsi aux pressions de sélection purificatrice (McCutcheon and Moran 2012). Citons en exemple le cas de la bactérie parasite obligatoire *Mycobacterium leprae*, dont la présence de très nombreux pseudogènes témoigne de la réduction récente de son génome (Gomez-Valero, et al. 2007). Il est intéressant de souligner qu'il existe de fortes variations de tailles de génome au sein d'une même espèce bactérienne (Touchon, et al. 2009). Ceci pourrait témoigner d'adaptations récentes et de diversifications écologiques. Je me suis concentré pendant ma thèse sur l'évolution des génomes de deux espèces d'entérobactéries: *Escherichia coli* et *Salmonella enterica*.

E. coli est probablement la bactérie modèle la plus étudiée et utilisée en génétique. Plus spécifiquement, l'étude de la souche K12, isolée il y a près d'un siècle (Bachmann 1972) est à l'origine de nombreuses découvertes majeures en génétique. De multiples travaux ont ainsi été effectués sur cette souche, bien que les conséquences de sa longue maintenance en conditions de laboratoire soulève quelques interrogations (Hobman, et al. 2007). *E. coli* appartient à la classe des gammaprotéobactéries (Williams, et al. 2010). Cette espèce est un objet d'étude intéressant car elle présente des biotypes, lysotypes, sérotypes variés et provoque diverses pathologies (Donnenberg 2002). Cette diversité phénotypique permet d'étudier les phénomènes adaptatifs sous-jacents. *S. enterica* est une autre entérobactérie très étudiée. Les genres *Escherichia* et *Salmonella* sont relativement proches et auraient divergé il y a environ 100 millions d'années (bien que ces estimations soient peu précises) (Battistuzzi, et al. 2004). Ces bactéries résident au sein du système digestif de leurs hôtes. *S. enterica* est également une préoccupation sanitaire importante car elle peut être à l'origine de diverses pathologies humaines. Enfin, de nombreux virus bactériens, les bactériophages (ou phages), infectant ces bactéries ont été décrits. L'abondance de données génomiques et de données expérimentales sur ces deux genres bactériens apparentés et de leurs virus, m'a conduit à choisir ces bactéries comme objet d'étude pour ma thèse.

La variabilité des génomes d'*E. coli* n'est pas seulement quantitative mais est aussi qualitative. En effet, des souches proches de cette espèce peuvent présenter des quantités de gènes fortement similaires tout en ayant des répertoires de gènes substantiellement différents. La comparaison du contenu génique de 20 souches d' *E. coli* a pu mettre en évidence de fortes disparités au sein de cette espèce (Touchon, et al. 2009). Si le génome moyen d'une souche d'*E. coli* contient environ 4.400 gènes, le nombre de gènes partagés par l'ensemble des 20 souches (génome core) est inférieur à 2.000 gènes (Fig 1). D'autre part, le nombre total de

gènes distincts (génome pan)présents au sein de ces génomes est d'environ 18.000 (Touchon, et al. 2009). Une autre étude a montré que le nombre de gènes essentiels *in vitro* chez la souche *E. coli* K12 MG1655 cultivée en milieu riche ne s'élève qu'à 300 gènes environ (Fig 1) (Baba, et al. 2006). Ces résultats soulignent l'importante diversité du répertoire génique d'une même espèce bactérienne. Ceci indique également que l'évolution des génomes bactériens se traduit en grande partie par une modification et par un renouvellement de leur contenu génique. L'une des principales causes de ce renouvellement réside dans la capacité des bactéries à transférer horizontalement des gènes, augmentant ainsi la plasticité de leurs génomes (Lerat, et al. 2005; Treangen and Rocha 2011). Enfin, il a été estimé qu'une part importante (~26%) de la diversité génétique d'*E. coli* est composée de gènes viraux (Touchon, et al. 2009). L'étude de cette fraction du génome est au centre de ma thèse. Les génomes bactériens présentent donc une diversité importante en répertoires de gènes qui est liée à leur dynamique évolutive. Pourtant, l'ADN bactérien, support physique du génome, est une structure organisée en réponse à différentes contraintes biologiques.

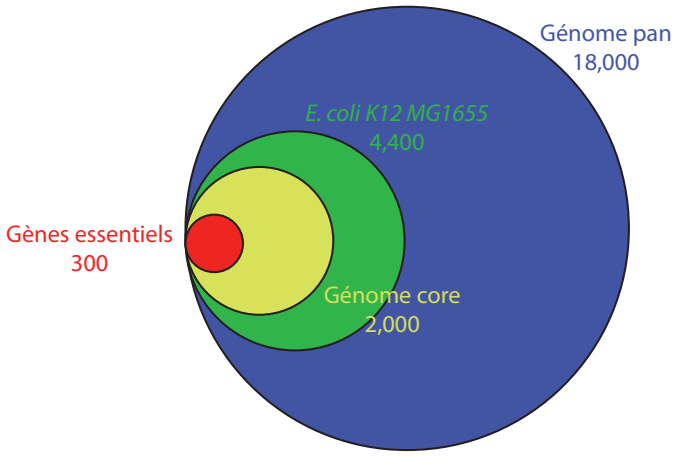

Figure 1: Diversité du répertoire génique d' *E. coli* (Baba, et al. 2006; Touchon, et al. 2009).

1.2 Le chromosome bactérien: une structure organisée

La majorité des bactéries présente la quasi-totalité de leurs gènes sur une seule molécule d'ADN généralement circulaire: le chromosome. Un sous-ensemble des gènes peut cependant être codé par des molécules d'ADN extra-chromosomique telles que les plasmides. Il existe des variantes à ce schéma avec certains genres bactériens, comme *Vibrio*, qui contiennent plusieurs chromosomes, ou encore comme *Borrelia*, qui ont la particularité d'avoir un chromosome linéaire (Chaconas and Kobryn 2010). Un gène d'*E. coli* est codé en moyenne

par 900pb. Le chromosome bactérien est très dense avec des séquences intergéniques de l'ordre de quelques dizaines de nucléotides (Mira, et al. 2001). De plus, des groupes de gènes impliqués dans des mêmes voies fonctionnelles sont fréquemment regroupés en opérons et sont alors transcrits et régulés simultanément. Enfin, le chromosome est une structure qui présente différents niveaux d'organisation imposés par divers processus cellulaires tels que la réplication, la ségrégation et la recombinaison. Je décrirai dans cette section les différents niveaux d'organisation du chromosome imposés par ces processus.

1.2.1 L'organisation associée à la réplication

La réplication du chromosome est un processus majeur du cycle de vie de la bactérie. Chez *E. coli*, la réplication est bidirectionnelle et débute à une origine unique appelée *ori*. Deux complexes protéiques vont ainsi parcourir le chromosome, séparer chaque brin et polymériser le brin complémentaire de chacun des brins. Les deux fourches progressent ainsi jusqu'au terminus de réplication (*ter*), où la séparation totale des deux chromosomes est catalysée par le complexe XerCD (Blakely, et al. 1993). Le chromosome possède ainsi un axe de symétrie par rapport à la réplication: l'axe *ori-ter*. Les deux moitiés de chromosome définies par cet axe sont nommées "réplichores". La position relative des sites *ori* et *ter* détermine donc la taille relative des deux réplichores. *E. coli* présente ainsi deux réplichores de longueurs semblables (Fig 2), ce qui permet une synchronisation du temps de réplication de chaque réplichore où les deux fourches se rencontrent de manière simultanée au site *ter* (Liu, et al. 2006). Des manipulations du chromosome d'*E. coli* ont mis en évidence, par déplacement relatif des sites *ori* et *ter*, que la bactérie peut supporter une certaine asymétrie des réplichores mais que cela allonge le temps de réplication (Liu, et al. 2006). Il a été suggéré que la symétrie des réplichores soit liée à des contraintes de ségrégation (Esnault, et al. 2007). La conservation d'une symétrie des réplichores chez de nombreuses espèces bactériennes (Matthews and Maloy 2010) souligne en tous cas l'importance de cette organisation du chromosome.

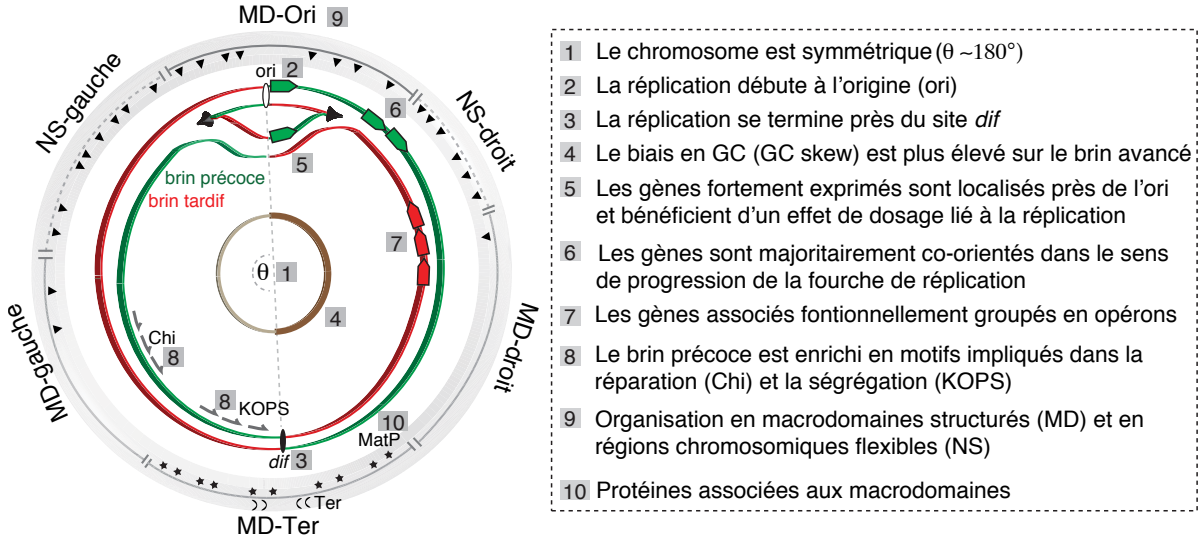

Figure 2: Organisation du chromosome d'*E.coli*.

La réplication complète du chromosome d'*E. coli* est d'environ 1h. Pourtant, *E. coli* peut, en conditions optimales de croissance, se diviser toutes les 20min (Cooper and Helmstetter 1968). Ce paradoxe s'explique par la capacité qu'a la bactérie d'initier de nouveaux cycles de réplication avant que le premier ne soit terminé. Ceci a une conséquence importante: les portions du chromosome situées près de l'*ori* sont présentes en copies plus nombreuses que celles proches du terminus, générant ainsi un effet de dosage de gènes (Fig 2). Il a été montré que cet effet de dosage affecte l'expression des gènes: les gènes sont d'autant plus exprimés qu'ils sont proches de l'*ori*, étant transitoirement présents en plusieurs copies au sein de la cellule (Schmid and Roth 1987; Sousa, et al. 1997). Le chromosome bactérien est organisé en conséquence, avec la présence des gènes fortement exprimés tels que les gènes impliqués dans la transcription et la traduction, près de l'*ori* (Couturier and Rocha 2006). Ceci semble représenter un avantage adaptatif pour les bactéries à croissance rapide (Couturier and Rocha 2006). Cette organisation du chromosome reflète donc une adaptation aux contraintes fonctionnelles liées à l'expression de certains gènes et à la réplication.

Au cours de la réplication, chacune des deux fourches est le lieu d'une double synthèse d'ADN. En effet, l'ADN est une double molécule polarisée. Suite à la séparation des deux brins par l'hélicase du complexe de réplication, la polymérase permet de reformer un ADN double-brin en utilisant chaque brin comme matrice. La synthèse de chaque brin complémentaire se fait uniquement dans l'orientation 5'-3'. Ceci a pour conséquence que l'un des deux brins est synthétisé de manière continue - puisqu'il est orienté avec le mouvement de la fourche de réplication - mais que l'autre brin est synthétisé de manière discontinue. La

synthèse du brin discontinu nécessite la réinitialisation de la réplication du brin à l'aide d'amorces ARN par l'action d'une primase. Ce brin est ainsi synthétisé en multiples séquences d'environ 1kb nommées fragments d'Okazaki (Kitani, et al. 1985). Ces fragments sont ensuite liés les uns aux autres au cours de la réplication par une ligase. Ainsi, pour chaque réplichore, le brin synthétisé de manière continue est qualifié de brin "avancé" ou "précoce" et le brin synthétisé de manière discontinue est nommé brin "retardé" ou "tardif". Cette asymétrie liée à la réplication des deux brins entraîne un biais mutationnel nommé "GC skew" puisque la matrice du brin tardif est sous forme simple brin plus longtemps. Le brin précoce est ainsi enrichi en G et le brin tardif en C (Lobry 1996). Une hypothèse suggère que ce biais pourrait être causé par la désamination des cytosines qui est plus fréquente lorsque l'ADN est sous forme simple brin (Lobry 1996). D'autres études ont proposé que d'autres processus mutationnels ou que des pressions sélectives pourraient entrer en jeu (Charneski, et al. 2011; Nikolaou and Almirantis 2005; Rocha, et al. 2006a).

Cette asymétrie réplicative engendre également un biais d'orientation des gènes. En effet, il a été montré que les gènes tendent à être co-orientés avec le sens de progression de la fourche de réplication et que cet effet est encore plus marqué pour les gènes essentiels (Fig 2) (Merrikh, et al. 2012; Rocha and Danchin 2003a; Rocha and Danchin 2003b). Cet enrichissement de gènes sur le brin précoce est lié à la transcription. La co-orientation de la fourche de réplication avec la transcription de l'ARN polymérase permettrait ainsi de diminuer la fréquence des collisions entre l'ADN polymérase et les ARN polymérases. Les gènes codés sur le brin tardif sont en effet transcrits dans le sens opposé à la réplication, ce qui augmente le nombre de collisions frontales entre les ARN polymérases et la fourche de réplication. Ces collisions peuvent provoquer l'arrêt de la fourche de réplication et générer des dommages et des réarrangements chromosomiques (Gan, et al. 2011; Mirkin and Mirkin 2005). Selon cette hypothèse, le biais de gènes sur le brin précoce serait une adaptation permettant de réduire le taux de mutations affectant les gènes. Cependant, la faible fréquence des arrêts de la fourche (<20% des tours de réplication) (Maisnier-Patin, et al. 2001) et l'absence de biais d'enrichissement des gènes fortement exprimés (qui sont plus contraints (Rocha and Danchin 2004)) sur le brin précoce ne supportent pas cette hypothèse d'un biais lié aux mutations (Rocha 2008). Alternativement, il a été proposé qu'il existe un avantage transcriptionnel à réduire ces collisions (Rocha 2008): i) Il en résulterait une légère augmentation du nombre de transcrits produits. ii) Cela diminuerait le nombre de transcrits tronqués et par conséquent le nombre protéines tronquées pouvant être toxiques pour la cellule. iii) La diminution du nombre de collisions avec les ARN polymérases diminuerait le

bruit transcriptionnel et permettrait une régulation plus fine de la transcription. Ce dernier effet peut se révéler important pour les gènes nécessitant une régulation précise tels que les régulateurs de transcription. Ce biais d'orientation des gènes peut être très important: 95% des gènes essentiels et 75% des autres gènes sont en effet codés sur le brin précoce chez *B. subtilis* (Rocha and Danchin 2003b). Il est en revanche plus faible (55% de l'ensemble des gènes et 76% des gènes essentiels) chez *E. coli* (Rocha and Danchin 2003b).

La réplication influence donc grandement l'organisation du chromosome. D'abord, ce processus impose la conservation d'une certaine symétrie liée à la division en deux réplichores. Les gènes situés près de l'*ori* bénéficient d'un effet de dosage, augmentant leur taux de transcription. Enfin, la synthèse asymétrique de l'ADN à chaque fourche de réplication engendre deux effets: un biais mutationnel enrichissant le brin précoce en G et le brin tardif en C et un enrichissement des gènes (principalement les gènes essentiels) sur le brin avancé.

1.2.2 La ségrégation

Faisant suite à la réplication, la ségrégation est le mécanisme qui dirige la séparation des deux chromosomes nouvellement formés au sein des deux futures cellules filles. Chez *E. coli*, ce processus implique la translocase FtsK qui interagit avec le chromosome par l'intermédiaire des sites KOPS (FtsK-orienting polar sequences) (Bigot, et al. 2005; Levy, et al. 2005). Le motif KOPS est un octamère polarisé: GGGNAGGG. Il est largement co-orienté avec la fourche de réplication (sur le brin précoce) et est présent tous les 12kb environ (Fig 2) (Touzain, et al. 2011). Il est en outre plus fréquent à proximité du terminus de réplication (Bigot, et al. 2005). La représentation fréquente des sites KOPS sur le chromosome et leur polarisation permettrait d'orienter la fixation de FtsK et par conséquent d'orienter la ségrégation du chromosome (Lowe, et al. 2008; Sivanathan, et al. 2006).

1.2.3 La recombinaison homologue

La recombinaison homologue nécessite la présence de séquences fortement similaires pour être initiée (Shen and Huang 1986). Chez *E. coli*, la principale voie utilisée est la voie

RecBCD, bien que d'autres voies, telles que la voie RecF, existent (Morimatsu and Kowalczykowski 2003). Je me contenterai ici de décrire la voie RecBCD car elle concerne directement mes travaux. La voie RecBCD est initiée lorsqu'une extrémité libre d'ADN double-brin est présente dans la cellule (Smith 2012). De telles molécules d'ADN peuvent être générées lorsqu'une cassure d'ADN double brin survient. L'extrémité d'ADN linéaire ainsi générée est reconnue par le complexe RecBCD (Dillingham and Kowalczykowski 2008). RecBCD présente une activité hélicase et exonucléase qui dégrade progressivement l'ADN à partir de son extrémité. La dégradation de l'ADN est inhibée lorsque RecBCD rencontre un motif Chi (Fig 3) (Henderson and Weil 1975; Myers and Stahl 1994). Plusieurs unités de la protéine RecA sont alors recrutées et se fixent sur ce brin. RecBCD se désassemble peu après et il en résulte une extrémité 3' d'ADN simple brin couverte de multiples copies de RecA (Fig 3). Cette extrémité nucléoprotéique représente l'intermédiaire réactionnel qui est capable d'initier la recombinaison proprement dite. En effet, RecA permet alors l'invasion de l'ADN simple brin au sein d'une séquence d'ADN double brin homologue, par déplacement du brin complémentaire originel (Fig 4) (Forget and Kowalczykowski 2012). Cet ensemble forme un complexe nommé boucle D ("D loop") (Kuzminov 1999). Deux scénarios sont alors possibles: i) En cas de coupure de la boucle D, l'ADN receveur peut, à son tour, envahir l'ADN donneur et former ainsi une jonction de Holliday. Cette structure sera résolue par des enzymes spécifiques: RuvABC ou RecG (Sharples, et al. 1999). ii) Le brin envahisseur initie une fourche de réplication sur le brin receveur qui resynthétise le brin dégradé. Il est intéressant de noter que ce mécanisme nécessite la présence de sites Chi, sans quoi RecBCD dégrade la totalité du double brin d'ADN. Ce complexe a donc une action défensive car il permet ainsi de dégrader de l'ADN exogène dépourvu de sites Chi tel que de l'ADN viral qui représente une menace pour la cellule (Dillingham and Kowalczykowski 2008). Le motif Chi (GCTGGTGG pour *E. coli*) est polarisé et présent sur le brin précoce de réplication (Fig 2). Il est présent environ tous les 5kb sur le chromosome (Touzain, et al. 2011).

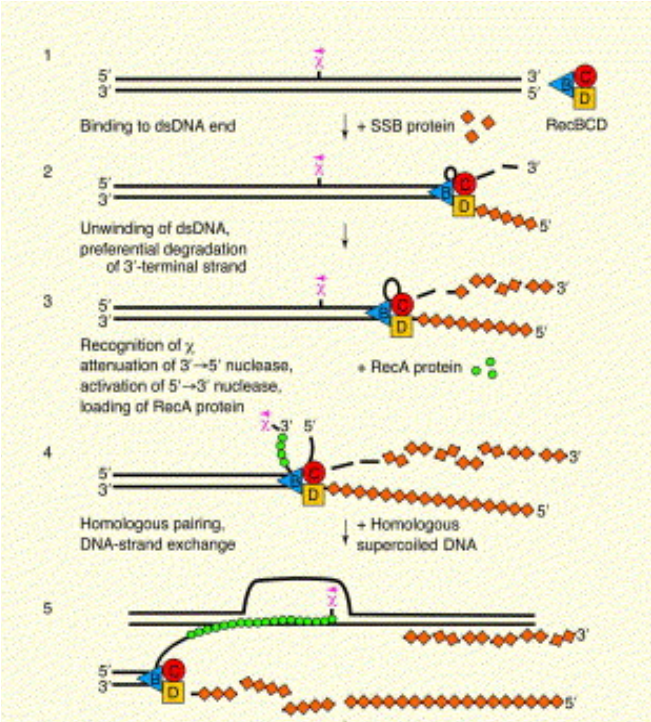

Figure 3: Initiation de la recombinaison homologue par RecBCD (Kowalczykowski 2000).

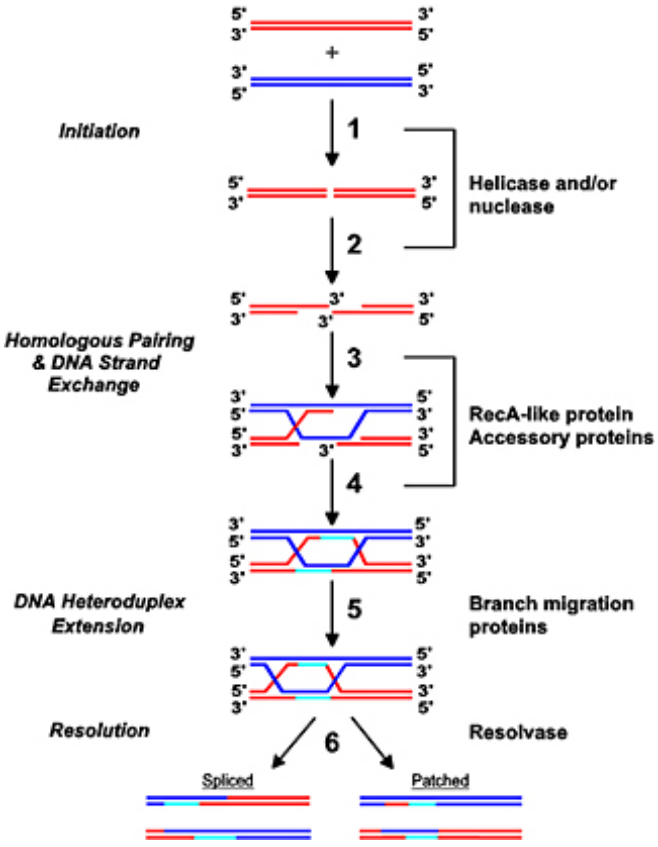

Figure 4: Recombinaison homologue chez *E. coli* (Bianco, et al. 1998).

1.2.4 La structure du chromosome

Le chromosome bactérien est associé, dans la cellule, à différentes protéines, telles que H-NS, qui forment un complexe protéonucléique nommé "nucléoïde". Cette structure présente différents niveaux de compaction. D'abord le nucléoïde est surenroulé négativement par l'action de topoisomérases (DiGate and Marians 1988; Higgins, et al. 1978; Roca 1995). Ensuite, l'interaction entre différentes protéines associées à l'ADN, permet d'augmenter la condensation du nucléoïde (Thanbichler, et al. 2005). Cette condensation n'est pas uniforme tout au long du chromosome. Il a été déterminé que le chromosome est composé de domaines moins condensés de 10 à 100kb, dont le but serait de diminuer les tensions physiques pouvant engendrer des cassures de l'ADN (Postow, et al. 2004; Worcel and Burgi 1972). L'estimation du nombre de ces domaines reste cependant imprécise (Postow, et al. 2004). A plus large échelle le chromosome d'*E. coli* est structuré en plus grands domaines nommés "macrodomaines" (Fig 2). Cette organisation en quatre macrodomaines et en deux régions non structurées a été déterminée grâce aux fréquences de recombinaison intra chromosomiques (Valens, et al. 2004). Le macrodomaine Ori contient l'origine de réplication et est entouré des deux régions non structurées. Le macrodomaine Ter contient le terminus de réplication et est adjacent aux macrodomaines "Droite" et "Gauche" (Fig 2). Les régions non structurées présentent une plus grande mobilité au sein de la cellule et peuvent interagir physiquement avec les macrodomaines adjacents (Valens, et al. 2004). Cela suggère que ces régions sont relativement peu condensées. Les quatre macrodomaines présentent des niveaux de condensation différents les uns des autres (Wiggins, et al. 2010). En outre, ils ne peuvent interagir physiquement qu'avec eux mêmes et très faiblement avec les régions chromosomiques adjacentes (Valens, et al. 2004). Il a également été montré que ces différentes régions chromosomiques occupent des positions définies au sein de la cellule (Meile, et al. 2011). Le macrodomaine Ter est situé le plus en périphérie du nucléoïde alors que les régions non structurées et Ori occupent une position plus centrale. Les macrodomaines Gauche et Droite sont quant à eux localisés aux pôles de la cellule. Les macrodomaines sont également associées avec différentes protéines (Dame, et al. 2011). La protéine MatP se fixe spécifiquement dans le macrodomaine Ter par l'intermédiaire du motif *matS* de 13bp: (GTGACA/GNT/CGTCAC) (Mercier, et al. 2008). Cette interaction entre MatP et *matS* est à l'origine de la structuration du macrodomaine Ter et *matS* est absent du reste du chromosome.

En effet, l'absence de MatP entraîne une déstructuration du macrodomaine Ter (Mercier, et al. 2008).

1.2.7 Implications évolutives

Les différents processus biologiques décrits au long de cette section mettent en avant l'importance de l'organisation du chromosome pour le bon fonctionnement de ces différents mécanismes cellulaires majeurs. Le chromosome est une structure qui s'est adaptée à des contraintes de symétrie, de polarité, de compaction et d'organisation des gènes. Ces différentes contraintes ont donc un impact direct sur la plasticité de la structure chromosomique (Esnault, et al. 2007). Par exemple, la nécessité de conserver des motifs polarisés tels que les KOPS peut fortement contraindre les inversions chromosomiques (Hendrickson and Lawrence 2006). Les grands réarrangements chromosomiques sont relativement rares et généralement symétriques à l'axe *ori-ter*, respectant ainsi la symétrie du chromosome et la polarité de ses motifs (Eisen, et al. 2000; Rocha 2008; Tillier and Collins 2000). De plus, ces contraintes pourraient largement limiter l'acquisition d'ADN exogène non adapté à ce contexte chromosomique. Pourtant, nous avons vu précédemment que le contenu des génomes bactériens est fortement variable. Le génome bactérien est donc une structure qui apparaît étonnamment plastique compte tenu de l'existence de nombreuses contraintes d'organisation et de structure.

1.3 Plasticité et évolution du génome bactérien

1.3.1 La sélection

Les modifications des génomes par mutation peuvent avoir différentes conséquences. Elles peuvent engendrer des modifications qui affectent la séquence des protéines, modifient leur taux d'expression ou encore engendrent des réarrangements chromosomiques. Elles peuvent également engendrer des modifications qui n'ont pas d'effets sur le fonctionnement cellulaire ou des effets négligeables. A cause de la redondance du code génétique, les substitutions

ponctuelles affectant les gènes codants n'entraînent pas toujours des modifications de la séquence protéique correspondante. Les substitutions ne modifiant pas la séquence protéique sont dîtes "synonymes", tandis que celles affectant la séquence protéique sont dîtes "non synonymes". La comparaison du taux de substitutions synonymes (dS) au taux de substitutions non synonymes (dN) permet de fournir des informations sur le type de sélection affectant un gène codant pour une protéine. En effet, la majorité des modifications de la séquence protéique engendrent des pertes ou des diminutions de fonctionnalité, ce qui conduit à l'élimination de nombreuses mutations non synonymes par sélection purificatrice (Yang and Bielawski 2000). A l'inverse, les mutations synonymes n'ont pas d'effet sur la fonction protéique et peuvent donc s'accumuler beaucoup plus librement. Il en résulte qu'un gène sous sélection purificatrice augmente plus rapidement en dS qu'en dN (dN/dS < 1). Si le gène est sous évolution neutre, les substitutions synonymes et non synonymes s'accumulent théoriquement à la même vitesse (dN/dS = 1). Enfin, certaines modifications protéiques peuvent apporter un avantage fonctionnel à l'organisme qui évolue alors sous sélection positive ou diversifiante. Cela peut alors conduire à une accumulation plus rapide des substitutions non synonymes par rapport aux substitutions synonymes (dN/dS > 1). Il est important de souligner que les substitutions synonymes pourraient avoir des coûts métaboliques associés à la composition en GC du génome (Rocha and Danchin 2002). De plus, l'usage du code génétique n'est pas aléatoire et certains codons synonymes seraient privilégiés afin de moduler l'efficacité et la régulation de la traduction (Grantham, et al. 1981; Novoa and Ribas de Pouplana 2012; Rocha 2004; Sharp 1991). Considérer les substitutions synonymes comme des événements complètement neutres d'un point de vue évolutif est donc une approximation mais la comparaison des taux de substitutions synonymes et non synonymes reste informative et largement utilisée.

1.3.2 La dérive génétique

La dérive génétique désigne l'évolution d'une population dirigée par des phénomènes aléatoires. Typiquement, la dérive est d'autant plus importante que la taille effective de la population est faible (Kimura 1968; Masel 2011). Contrairement à la sélection, la dérive peut conduire à la fixation de mutations modérément délétères. Différentes études ont montré l'effet important de la dérive sur la structuration des génomes, notamment sur l'accumulation de séquences supposées non fonctionnelles (Charlesworth and Barton 2004; Daubin and

Moran 2004; Lynch and Conery 2003). Les effets relatifs de la dérive et de la sélection sont généralement estimés à l'aide du taux de polymorphisme d'une population ou d'une espèce. Le polymorphisme est cependant difficile à estimer chez les bactéries où les définitions de populations et d'espèces sont problématiques (Kuo, et al. 2009). Néanmoins, il a été suggéré que les fortes tailles effectives des populations bactériennes conduiraient à une faible dérive génétique, et donc, à une faible accumulation de séquences non fonctionnelles (Lynch 2006). Alternativement, puisque les bactéries présentent un biais mutationnel qui favorise les délétions (Andersson and Andersson 2001; Mira, et al. 2001), une dérive génétique importante pourrait conduire à l'accumulation de délétions, et donc, à la diminution de la taille des génomes (Kuo, et al. 2009).

1.3.3 La recombinaison

Outre son rôle dans le fonctionnement cellulaire, la recombinaison a des conséquences évolutives majeures car elle permet l'échange de matériel génétique entre organismes, ou plus généralement entre réplicons. Elle permet ainsi de générer des nouvelles associations de gènes et d'allèles et d'augmenter l'efficacité de la sélection naturelle (Barton and Charlesworth 1998; Feil, et al. 2001). La recombinaison peut également modifier l'organisation des génomes via des réarrangements chromosomiques plus ou moins importants (Sun, et al. 2012). Enfin, ce processus permet d'incorporer des séquences d'ADN exogènes au sein d'un réplicon (Ochman, et al. 2005). La recombinaison joue donc un rôle important dans la diversification et dans l'évolution de l'architecture des génomes.

1.3.4 Les pertes de gènes

Les gènes peuvent être inactivés et éliminés des génomes par des processus de mutations ponctuelles et de délétions. Lors de variations des pressions sélectives liées à des changements du milieu, ou par l'acquisition de gènes aux fonctions analogues, un gène peut ne plus être requis (ou indispensable) au fonctionnement de la cellule. L'absence de sélection purificatrice agissant sur ce gène va alors permettre l'accumulation de mutations non synonymes, de modifications du cadre de lecture et l’apparition de codons stop prématurés.

Les gènes non fonctionnels ainsi altérés sont alors qualifiés de pseudogènes. Il a été observé que les pseudogènes sont rapidement éliminés des génomes bactériens (Lerat and Ochman 2004, 2005). Il a été suggéré que les pseudogènes peuvent souvent présenter des effets toxiques, notamment par la production de protéines tronquées qui interagiraient avec d'autres composants cellulaires de manière néfaste (Kuo and Ochman 2010). Cette hypothèse expliquerait leur élimination rapide qui pourrait ainsi être promue par sélection positive (Kuo and Ochman 2010). Les pertes de gènes non fonctionnels pourraient également être facilitées par le biais mutationnel favorisant les délétions chez les bactéries (Andersson and Andersson 2001; Mira, et al. 2001).

1.3.5 Les gains de gènes

Les bactéries peuvent acquérir de nouveaux gènes par différents mécanismes. D'abord, la duplication de gènes pré-existants suivie de leur diversification par mutations peut permettre l'apparition de gènes aux fonctions nouvelles (Lynch and Conery 2000; Serres, et al. 2009). Ces processus sont cependant lents et le biais de délétion pourrait rapidement éliminer les copies additionnelles des gènes aux fonctions redondantes (Isambert and Stein 2009). Le transfert horizontal semble être un moyen beaucoup plus rapide et prépondérant que la duplication pour l'acquisition de nouveaux gènes (Treangen and Rocha 2011). En effet, ce processus permet l'acquisition de nouvelles fonctions génétiques pré-existantes chez d'autres organismes. Cependant, différentes barrières au transfert de gènes existent. La présence de promoteurs de gènes différents pourrait par exemple empêcher l'utilisation d'un gène provenant d'une espèce éloignée (Sorek, et al. 2007). Il a également été suggéré que les gènes codant pour des protéines impliquées dans des complexes multi-protéiques seraient moins susceptibles d'être transférés (Jain, et al. 1999; Sorek, et al. 2007). Ceci serait du à la co-évolution des gènes d'un même complexe qui donnerait lieu à des interactions protéiques incompatibles entre espèces (Jain, et al. 1999). Les transferts horizontaux de gènes sont médiés par différents mécanismes tels que la transformation, la conjugaison et la transduction. D'autres mécanismes de transferts de gènes ont été proposés: les vésicules membranaires (Kolling and Matthews 1999; Rumbo, et al. 2011) et les nanotubes (Dubey and Ben-Yehuda 2011). Enfin, les génomes bactériens contiennent de nombreux gènes sans homologues identifiés: les ORFans (Daubin and Ochman 2004a; Fischer and Eisenberg 1999). Différents

travaux suggèrent qu'une proportion importante de ces gènes orphelins serait acquise par des transferts d'origine virale (Cortez, et al. 2009; Daubin and Ochman 2004b).

1.3.6 Les éléments mobiles

Les génomes bactériens présentent divers éléments mobiles. Ces éléments ont la capacité de se transférer de cellule à cellule de manière autonome ou semi-autonome. Certains ont en outre la capacité de s'intégrer au sein du chromosome. D'autres se mobilisent de manière autonome au sein du chromosome mais ne disposent du matériel génétique permettant leur transfert. Différents travaux de génomique comparative ont mis en évidence l'existence de blocs ou "îlots" de gènes synténiques transférés horizontalement entre chromosomes bactériens (Juhas, et al. 2009). De manière générale, ces éléments sont nommés îlots génomiques et leur taille varie typiquement de 10kb à 200kb (Juhas, et al. 2009). Leurs mécanismes d'intégration et d'excision au sein du chromosome, ainsi que leurs mécanismes de transfert permettent de classer ces éléments en différentes catégories (Fig 5) (Langille, et al. 2010). Certains de ces éléments tels que les ICEs (Integrated Conjugative Elements) et les phages tempérés peuvent s'intégrer de manière stable au sein du chromosome bactérien et ont également la capacité de se mobiliser de manière autonome vers d'autres cellules (Burrus, et al. 2002; Campbell 2007; Wozniak and Waldor 2010). Les ICEs et les phages se mobilisent respectivement par conjugaison et par transduction. D'autres éléments sont typiquement dépourvus de gènes permettant leur propre transfert et pourraient utiliser les systèmes de transfert codés par d'autres éléments mobiles (Chen, et al. 2005; Jain, et al. 2002; Novick, et al. 2010; Ruzin, et al. 2001). Ils pourraient en outre être transférés par transformation (Juhas, et al. 2009). Les îlots génomiques peuvent apporter différents avantages sélectifs à la bactérie et peuvent alors être nommés îlots de pathogénicité, de symbiose, de métabolisme, de fitness ou de résistance (Dobrindt, et al. 2004; Schmidt and Hensel 2004). Un grand nombre de ces éléments ne peut pas être classé dans une catégorie particulière (plasmide intégratif, ICE, etc) et ils semblent être issus d'origines évolutives variées (Juhas, et al. 2007; Vernikos and Parkhill 2008). Ils pourraient ainsi être le résultat de la dégradation d'éléments autonomes tels que les ICEs ou de phages (Burrus and Waldor 2004; Hsiao, et al. 2005; Juhas, et al. 2009) (Fig 5). Différents critères permettent de détecter ces différents éléments. Ils présentent typiquement une distribution sporadique au sein des souches d'une même espèce (Middendorf, et al. 2004). Ces éléments présentent souvent des compositions en

oligonucléotides différentes de celles de leurs bactéries hôtes (Karlin 2001; Langille, et al. 2010). Différents gènes associés avec leur mobilité peuvent également permettre d'identifier certains types d'îlots (Langille, et al. 2010). Enfin, ils semblent intégrés préférentiellement au sein de certaines séquences telles que les ARNt et sont souvent bordés par des séquences répétées (Hacker, et al. 1997). Aucun de ces différents critères d'identification n'est cependant universel et il reste difficile d'identifier et de distinguer les différents types d'îlots génomiques (Langille, et al. 2010).

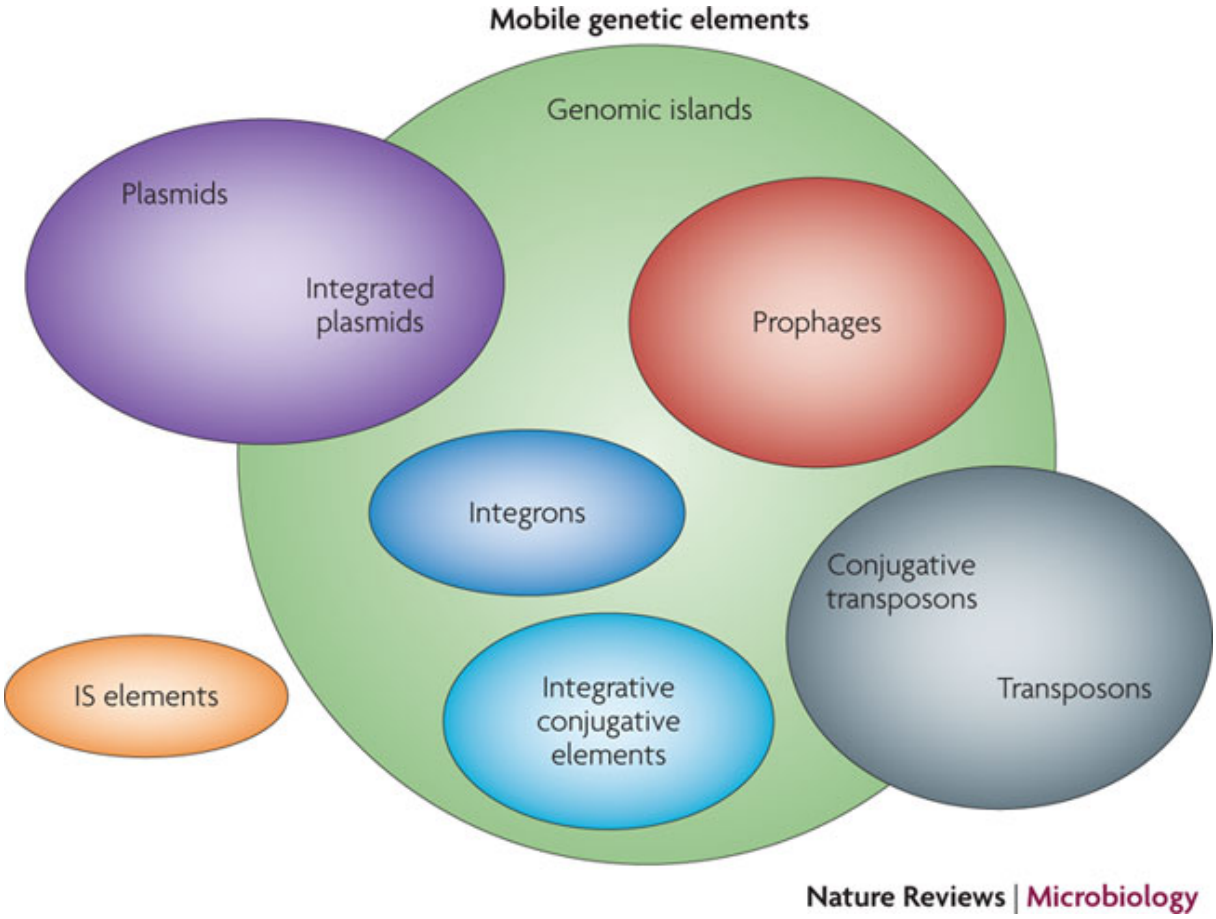

Figure 5: Différents types d'éléments génétiques mobiles (Langille, et al. 2010).

1.3.7 Points chauds de transfert

Il a été observé que les éléments mobiles tendent à être intégrés dans un faible nombre de loci du génome bactérien (Touchon, et al. 2009). Ces positions, nommées "points chauds" d'intégration, semblent en outre très conservées entre souches et espèces différentes et il est fréquent d'observer plusieurs éléments à un même locus (Touchon, et al. 2009). La plupart des

éléments mobiles s'intègrent au niveau d'un site spécifique via une recombinase site-spécifique (Napolitano, et al. 2011). Certaines séquences semblent être des cibles privilégiées de ces enzymes comme les ARNs de transfert (ARNt) (Williams 2002). Cependant, de telles séquences ne sont pas toujours présentes au niveau des points chauds et la dynamique de formation de ces loci est assez mal comprise (Fouts 2006). Il est possible que l'intégration initiale d'un élément mobile au sein du génome facilite l'incorporation d'éléments additionnels dans la mesure où ces séquences ne sont pas essentielles pour la bactérie (Hassan, et al. 2010). Il a également été suggéré que ces points chauds d'intégration soient transférés horizontalement entre cellules et intégrés par recombinaison homologue (Schubert, et al. 2009). Il est probable que des pressions sélectives conduisent au regroupement des éléments mobiles ainsi qu'à une position biaisée de ces points chauds au sein du chromosome. Ces îlots génomiques tendent en effet à être localisés à proximité du terminus de réplication et dans les chromosomes secondaires (Andersson, et al. 2010; Flynn, et al. 2010; Okada, et al. 2005). Chez certaines espèces, la majorité des éléments mobiles est localisée à l'extérieur du chromosome en de nombreux plasmides ou encore à l'extrémité des chromosomes linéaires (Casjens, et al. 2000; Choulet, et al. 2006). Il semble donc que l'organisation des éléments mobiles au sein du chromosome hôte soit un processus en partie forgé par la sélection naturelle en réponse à l'organisation chromosomique bactérienne.

2 Diversité et évolution des bactériophages tempérés

2.1 Diversité des phages

2.1.1 De la découverte de Twort et d'Hérelle à la biologie synthétique

Les bactériophages (ou phages) sont des virus infectants spécifiquement les bactéries. La découverte de ces entités est attribuée à Frederick Twort pour avoir observé l'action toxique des phages sur des cultures bactériennes (1915) et à Félix d'Hérelle pour les avoir isolés (1917) (D'Herelle 2007). La nature de ces entités était alors inconnue. Peu de temps après cette découverte, d'Hérelle utilisa les propriétés de ces éléments à des applications thérapeutiques anti-bactériennes: la phagothérapie. Cette thérapie fut largement abandonnée en occident suite au démarrage de la production industrielle des antibiotiques dans les années 1940s (Summers 2001). Cependant, les phages s'avérèrent par la suite être un outil génétique très puissant et l'étude de ces virus a permis de comprendre de nombreux mécanismes fondamentaux tels que, entre autres, l'apparition des mutations en l'absence de sélection (Luria and Delbruck 1943), la transduction (Zinder and Lederberg 1952), la régulation génétique (Ptashne 1992), ou encore la nature nucléique du support de l'information génétique (Hershey and Chase 1952). Le premier organisme séquencé fut également un phage: phiX174 (Sanger, et al. 1977). Enfin, les phages sont une source d'outils génétiques privilégiée. En effet, l'importante diversité des phages permet la découverte fréquente de systèmes génétiques inédits dont l'utilisation est exploitée pour le développement de biotechnologies. Citons par exemple leur utilisation à des fins de thérapie génique (Nafissi and Slavcev 2014) ou de bioingénierie telles que la technique de Recombineering (recombination-mediated genetic engineering) (Ellis, et al. 2001).

2.1.2 Classification de Baltimore et classification de l'ICTV

Il existe une très vaste diversité de phages. Deux systèmes de classification complémentaires des virus ont été développés: la classification de Baltimore et la classification de l'ICTV (International Commitee on Taxonomy of Viruses). La classification de Baltimore divise les virus en sept groupes distincts d'après la nature du support de leur information génétique (par exemple: groupe I: ADN à double brin ou groupe V: ARN à simple brin négatif) (Baltimore 1971). La classification de l'ICTV repose sur une méthode similaire à la systématique classique et définit les virus en ordres, familles et genres (http://www.ictvonline.org/). Le principal critère de l'ICTV repose sur la morphologie du virion qui subdivise les virus en ordres et en familles. Il existe actuellement 94 familles virales référencées sur le site de l'ICTV. A un niveau plus fin, cette classification prend également en compte les homologies des génomes viraux pour définir des sous-familles et des genres. Il est à noter que la diversité des phages est largement dominée par un groupe de phages particulier: les *Caudovirales*, qui représentent 96% des virus procaryotes décrits (Ackermann and Prangishvili 2012). C'est pour cette raison que je me concentrerai plus particulièrement sur ces éléments. Les *Caudovirales*, aussi nommés phages caudés, composent le seul ordre phagique décrit et sont définis par un virion à tête icosaédrique assorti d'une queue. Cet ordre est subdivisé en trois familles selon la structure de la queue du virion (Fig 6). Les *Siphoviridae* possèdent une queue longue, flexible et non contractile. Les *Podoviridae* présentent une queue courte, rigide et non contractile. Enfin, les *Myoviridae* sont définis par une queue longue, rigide et contractile. D'autres critères structurels, basés sur des observations en microscopie électronique, sont parfois utilisés pour subdiviser ces familles en sous-familles et en genres. Les ornements sur le virion (comme le plateau à l'extrémité de la queue) ou le ratio entre la longueur et la largeur de la tête du virion peuvent permettre de différencier les phages caudés d'une même famille. Des méthodes de génomique comparative sont de plus en plus utilisées pour classer ces entités en sous-familles et en genres: i) Par phylogénie moléculaire (Serwer, et al. 2004), mais cette dernière méthode est cependant limitée par l'absence fréquente de marqueurs phylogénétiques conservés. ii) Par comparaison du contenu en gènes (Lavigne, et al. 2009; Lima-Mendez, et al. 2008b; Rohwer and Edwards 2002). Cette dernière approche sera au centre de certaines de mes analyses et je la décrirai plus en détail par la suite (Résultats, section 1).

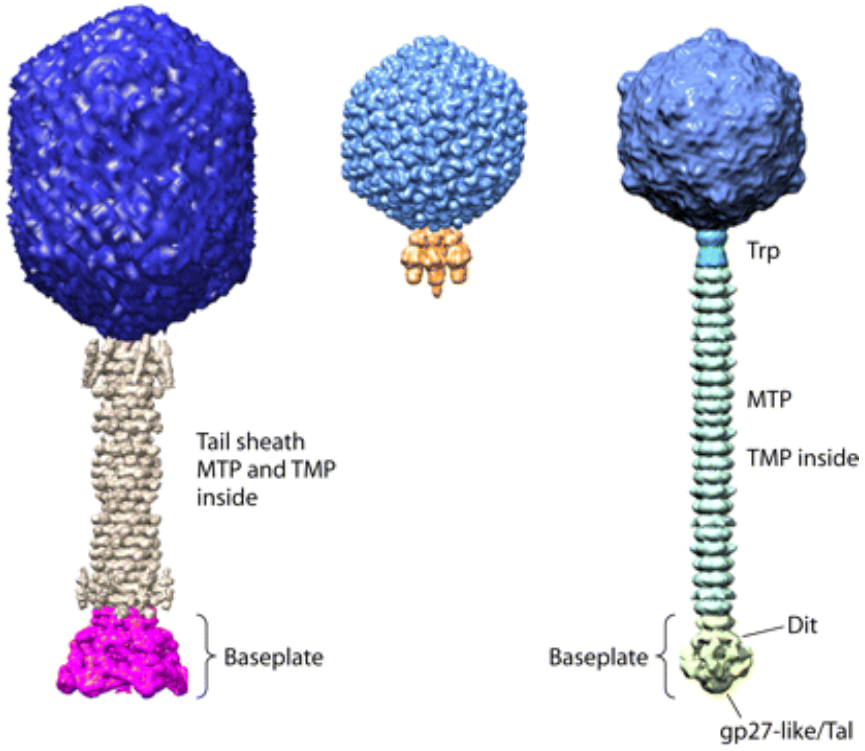

Figure 6: Particules virales des trois familles de l'ordre *Caudovirales*. Gauche: *Myoviridae*, mileu: *Podoviridae*, droite: *Siphoviridae* (Veesler and Cambillau 2011).

2.1.3 Phages virulents et phages tempérés

En tant que virus, les phages sont des parasites obligatoires. Ces organismes dépendent de leurs hôtes – les bactéries – pour accomplir leur cycle infectieux et ainsi se multiplier. Concernant cet aspect, on peut différencier deux types de phages: les phages virulents et les phages tempérés. i) Les phages virulents se multiplient uniquement par cycle lytique. Le cycle lytique consiste, dans un premier temps, en la réplication du génome phagique et en la production de capsides virales (Fig 7). Les génomes phagiques sont ensuite assemblés avec les différents éléments de la capside pour former des virions. Enfin, les virions sont libérés de la cellule, généralement en détruisant cette dernière (la lyse à proprement dite). Certains phages, tels que les phages filamenteux de la famille des *Inoviridae*, ont la capacité d'être sécrétés par les pores de la cellule sans lyser celle-ci (Rakonjac, et al. 2011). ii) Les phages tempérés ont la capacité de se multiplier par des cycles lytiques mais également par lysogénie (Fig 7). La lysogénie consiste à maintenir le génome du phage en état de dormance au sein de la cellule hôte. Le phage est alors nommé "prophage" et celui-ci est intégré dans le chromosome bactérien ou, plus rarement, reste à l'état d'épisome dans la cellule (Lobocka, et al. 2004; Ravin 2011). La bactérie contenant le prophage est alors qualifiée de "lysogène".

Lors de la lysogénie, le prophage n'exprime quasiment aucun gène et peut être répliqué passivement lorsque le chromosome hôte se réplique. Ainsi, la multiplication de la bactérie lysogène conduit directement à la multiplication du prophage. De cette manière, le prophage peut profiter de la multiplication de son hôte durant plusieurs générations. Suite à une induction ou de manière stochastique, un prophage peut sortir du cycle lysogénique et entamer un cycle lytique (Ball and Johnson 1991; Gottesman and Yarmolinsky 1968).

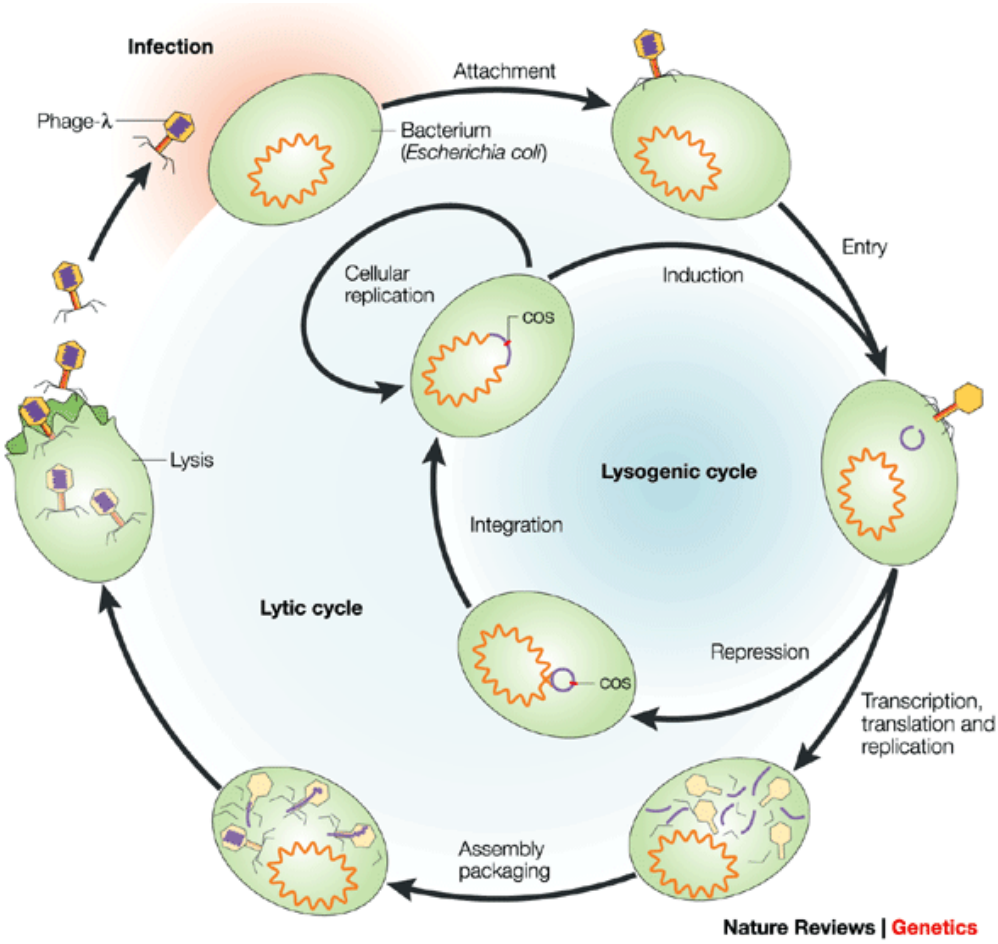

Figure 7: Cyclé lytique et cycle lysogénique (Campbell 2003).

2.2 Description d'un phage tempéré modèle: Lambda

2.2.1 Généralités

Découvert en 1950 par Esther Lederberg (Lederberg and Lederberg 1953), le phage Lambda a joué un rôle primordial dans l'essor de la génétique moléculaire. Ce phage tempéré d'*E. coli*

possède un génome de 48.502 nucléotides et code pour 73 protéines (Hendrix and Casjens 2006). Son génome est linéaire aux extrémités cohésives au sein de la capside. Il est circularisé et surenroulé négativement par l'action des ligases et gyrases hôtes suite à l'infection (Furth and WIckner 1983). Il appartient à la famille des *Siphoviridae* et possède ainsi une longue queue flexible non contractile prolongée de quatre fibrilles (protéines de fibre). J'ai décidé de me focaliser sur ce phage dans cette section car Lambda et d'autres phages apparentés ont occupé une place prépondérante dans mes travaux.

2.2.2 Un génome très organisé

Le génome du phage Lambda est très organisé. Deux niveaux d'organisation distincts co-existent à cause de contraintes différentes. Je parlerai ici du premier niveau d'organisation qui reflète l'organisation fonctionnelle liée à l'expression des gènes de Lambda. Je décrirai dans une autre section l'organisation forgée par l'évolution de Lambda.

Le génome de Lambda est structuré en trois opérons principaux (Fig 8): les opérons "early left", "early right" et "late". Les noms de ces opérons reflètent l'ordre d'expression de ces gènes lors du cycle lytique. Les deux opérons "early" codent pour les gènes impliqués dans la réplication et la recombinaison. Ces fonctions sont nécessaires à la multiplication du génome de Lambda au début de l'infection ou lors de l'induction du prophage. L'opéron "late" correspond à environ 50% du génome et code pour les protéines structurelles qui composent la particule virale, pour les protéines nécessaires à l'encapsidation du génome au sein de la particule et pour les protéines de lyse qui permettront de libérer les virions produits dans le milieu extracellulaire. Lors d'une infection, l'opéron "late" n'est pas systématiquement activé après les opérons "early". En effet, suivant la concentration relative des protéines régulatrices, l'opéron "late" peut rester inactif (Ptashne 1992). Dans ce cas, l'intégrase codée par l'opéron "early left" permettra au génome de Lambda de s'intégrer au sein du génome hôte si le site correspondant à l'intégrase est disponible, conduisant à l'établissement de la lysogénie.

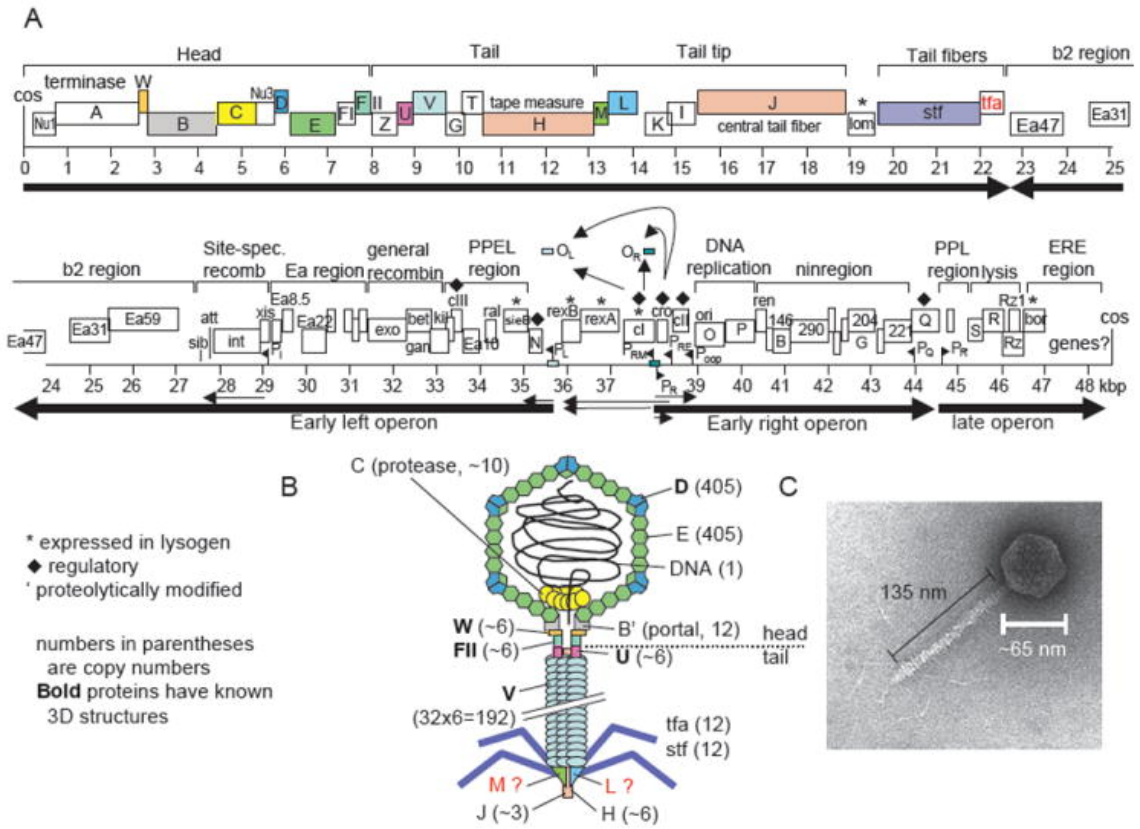

Figure 8: **A** Représentation schématique du génome du phage Lambda. **B** Architecture de la particule du phage Lambda. **C** Micrographie électronique du phage Lambda (Hauser, et al. 2012).

2.2.3 Régulation et cycle

Le cycle lytique de Lambda peut s'activer directement après infection ou suite à l'induction du prophage. Dans les deux cas, l'excision ou la circularisation de l'ADN de Lambda aboutissent à la formation d'un épisome circulaire au sein de la cellule (Guarneros and Echols 1970). La réplication est alors initiée grâce aux protéines O et P codées par l'opéron "early right" de Lambda et nécessite la machinerie de réplication d'*E. coli* (Furth and WIckner 1983). Le site d'initiation *ori* de la réplication se situe au sein même de la séquence du gène *O*. La réplication de Lambda débute dans un premier temps par plusieurs cycles de réplication bidirectionnelle "théta", similaire à la réplication du chromosome d'*E. coli* (Fig 9) (Furth and WIckner 1983). Il en résulte la production de plusieurs monomères circulaires du génome de Lambda. Dans un deuxième temps, la réplication se produit par des cycles "sigma" ou "cercles roulants" (Enquist and Skalka 1973). Ce mécanisme est initié par une coupure simple brin du génome circulaire de Lambda. Le brin coupé va permettre d'amorcer une réplication

unidirectionnelle dans le sens 5'-3' en utilisant le brin complémentaire comme matrice. Le brin non matriciel est alors déplacé au fur et à mesure de la réplication. La réplication en cercles roulants est donc unidirectionnelle et le même brin circulaire sert de matrice, ce qui permet d'effectuer plusieurs tours de réplication sans interruption (Fig 9). Il en résulte la synthèse de concatémères de génomes phagiques (Skalka, et al. 1972). Le mécanisme de transition entre les réplications théta et sigma reste assez mal compris (Furth and WIckner 1983). Il semblerait qu'il ne s'agisse pas d'une transition précisément contrôlée mais que les deux modes de réplication coexistent (Better and Freifelder 1983). La réplication en cercles roulants semble être privilégiée lorsque de nombreuses copies du génome de Lambda sont produites du fait de la moindre disponibilité des protéines O et P, nécessaires pour l'initiation (Furth and WIckner 1983).

De manière concomitante à la réplication, des particules de tête et de queue se forment dans la cellule grâce à l'expression de l'opéron "late" (Casjens 2011). Le mécanisme d'encapsidation est en revanche directement lié à la réplication de l'ADN de Lambda. Ce processus consiste à intégrer un monomère d'ADN phagique au sein d'une particule de tête vide grâce à une enzyme ATP-dépendante: la terminase (Feiss and Becker 1983). Lambda, comme de nombreux autres phages, présente la particularité de ne pouvoir encapsider son ADN qu'à partir d'ADN concatémérique (Szpirer and Brachet 1970). Les concatémères produits par la réplication en cercles roulants sont injectés dans la tête phagique et coupés par la terminase au niveau de sites spécifiques nommés *cos* (Fig 9) (Casjens 2011). D'autres phages apparentés à Lambda, tels que le phage P22, utilisent un autre système d'encapsidation: le système *pac* ou à "tête plein" (Jardine and Anderson 2006). Dans ce cas, l'initiation de l'encapsidation se fait au niveau du site *pac* (Backhaus 1985), puis l'ADN phagique est clivé lorsque la capside est remplie et non pas au niveau d'un site précis (Casjens and Hayden 1988). Ces mécanismes permettent d'intégrer un monomère linéaire d'ADN phagique au sein de chaque capside. Les queues se fixent ensuite sur les têtes phagiques remplies. La cellule est enfin lysée par l'action de la holine et de la lysine qui dégradent respectivement la membrane plasmique et la couche de peptidoglycanes (Campbell 1996).

Le processus d'encapsidation nécessite l'assistance des protéines du module *red* (Enquist and Skalka 1973). Ce module code pour trois protéines: Redα (ou Exo), Redβ (ou Bet) et Gam et est présent au sein de l'opéron "early left". Nous avons vu précédemment que la cellule hôte possède une exonucléase très puissante: RecBCD. Les concatémères d'ADN de Lambda produits par réplication sigma sont donc sujets à la dégradation par RecBCD (Enquist and Skalka 1973). Lambda code ainsi pour un inhibiteur de RecBCD: Gam (Fig 9). Gam est une

petite protéine dont la structure mime la structure de l'ADN et vient se fixer sur RecBCD (Court, et al. 2007). Les mutants Lambda *gam*$^{-}$ ne peuvent donc pas répliquer leur ADN par cercles roulants (Smith 1983). Chez ces mutants, la production de concatémères par réplication sigma peut être restaurée chez les hôtes *recBCD*$^{-}$ (Smith 1983). Il a été montré que Lambda peut également former des particules viables en l'absence de réplication sigma (Enquist and Skalka 1973). Cette deuxième voie est donc entièrement dépendante de la réplication théta mais nécessite l'action des protéines Redα et Redβ du module *red* (Fig 9). Redα est une exonucléase et Redβ une recombinase et leur action conjointe permet la recombinaison homologue des monomères d'ADN de Lambda. La recombinaison de plusieurs monomères circulaires d'ADN conduit ainsi à l'obtention d'ADN concatémérique (Smith 1983). Ce substrat peut alors être encapsidé au sein de particules phagiques par la terminase.

:

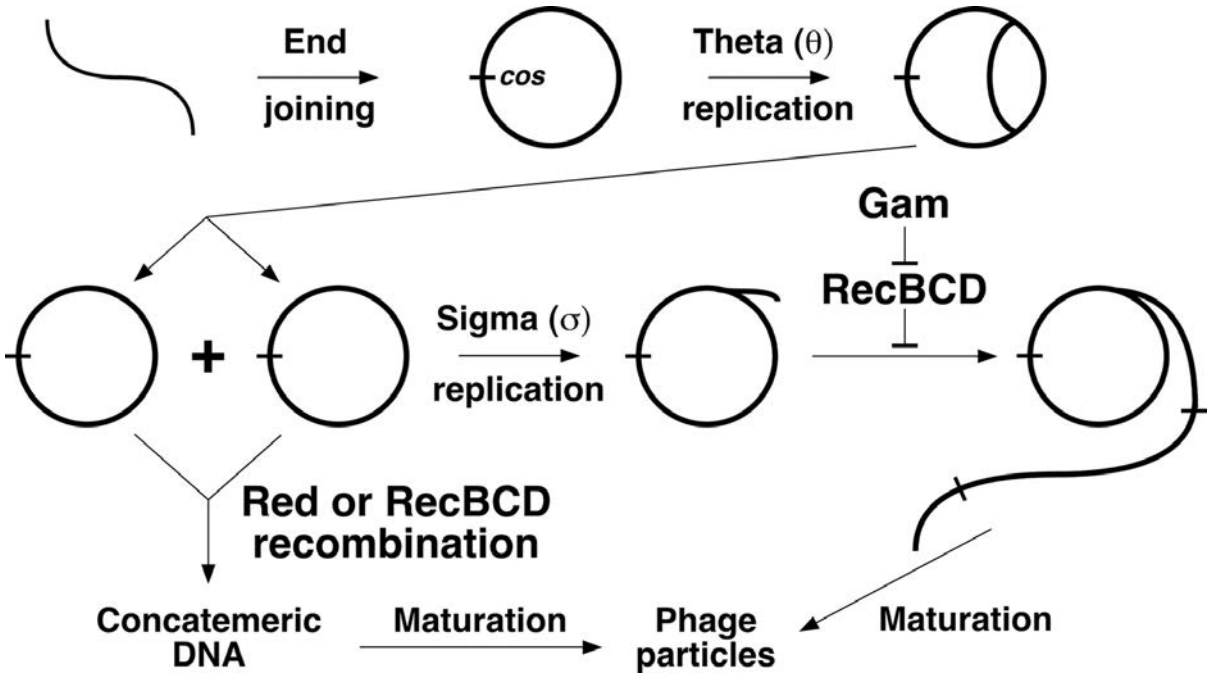

Figure 9: Réplication et recombinaison du phage Lambda (Smith 1983; Smith 2012).

2.2.4 Intégration et excision

Lambda peut s'intégrer au sein du chromosome hôte par recombinaison site-spécifique médiée par l'intégrase (Int). Cette enzyme promeut l'échange d'ADN au niveau de séquences particulières: les sites d'attachement (*att*) (Fig 10) (Campbell 1969). Les intégrases ont la

capacité de se fixer sur ces sites et de catalyser une réaction de recombinaison, ce qui conduit à l'intégration du phage. La reconnaissance du site se fait par l'intermédiaire d'une courte séquence similaire entre le génome du phage (*att*P) et celui de la bactérie (*att*B) (Weisberg and Landy 1983). Des séquences annexes peuvent être nécessaires à la fixation de l'intégrase (Radman-Livaja, et al. 2006). L'intégration donne lieu au prophage bordé des sites *att*L et *att*R (Weisberg and Landy 1983). Il existe deux familles distinctes d'intégrases site-spécifiques: les tyrosine recombinases et les serine recombinases qui utilisent respectivement un résidu tyrosine et sérine comme site réactionnel (Hallet and Sherratt 1997; Smith and Thorpe 2002). Lambda code pour une intégrase de la famille des tyrosine recombinases (Landy 1989). L'excision s'effectue par la même réaction de recombinaison entre les sites *att*L et *att*R qui vont ainsi circulariser le génome du phage (Weisberg and Landy 1983). Cette réaction est catalysée par l'intégrase mais nécessite l'action d'une enzyme phagique supplémentaire: l'excisionase (Xis) (Landy 1989). Les réactions d'intégration et d'excision nécessitent également la présence de protéines hôtes telles que IHF (Integration Host Factor) et FIS (Landy 1989).

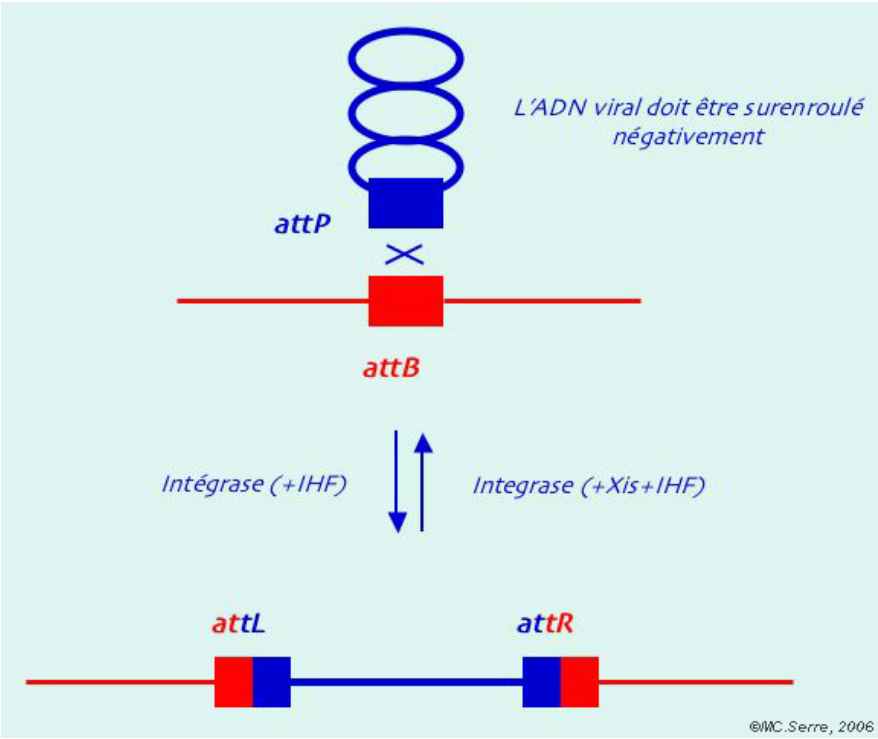

Figure 10: Intégration et excision de Lambda au sein du chromosome d'*E. coli* (www.igmors.u-psud.fr).

2.2.5 Stabilité de la lysogénie

Une fois intégré dans le chromosome sous forme de prophage, Lambda n'exprime qu'une seule protéine essentielle de manière constitutive: le répresseur CI. Des dimères de cette protéine vont ainsi maintenir le cycle lysogénique en se fixant sur les sites opérateurs O_R et O_L (Fig 11) (Gussin, et al. 1983; Ptashne 1992). Par leur fixation sur les opérateurs, les dimères de CI empêchent à la protéine Cro d'accéder à ces sites. Or, c'est la fixation de Cro sur les sites opérateurs O_R qui va induire la transcription de l'opéron "early right" et ainsi initier le cycle lytique et réprimer l'expression de *cI* (Ptashne 1992).

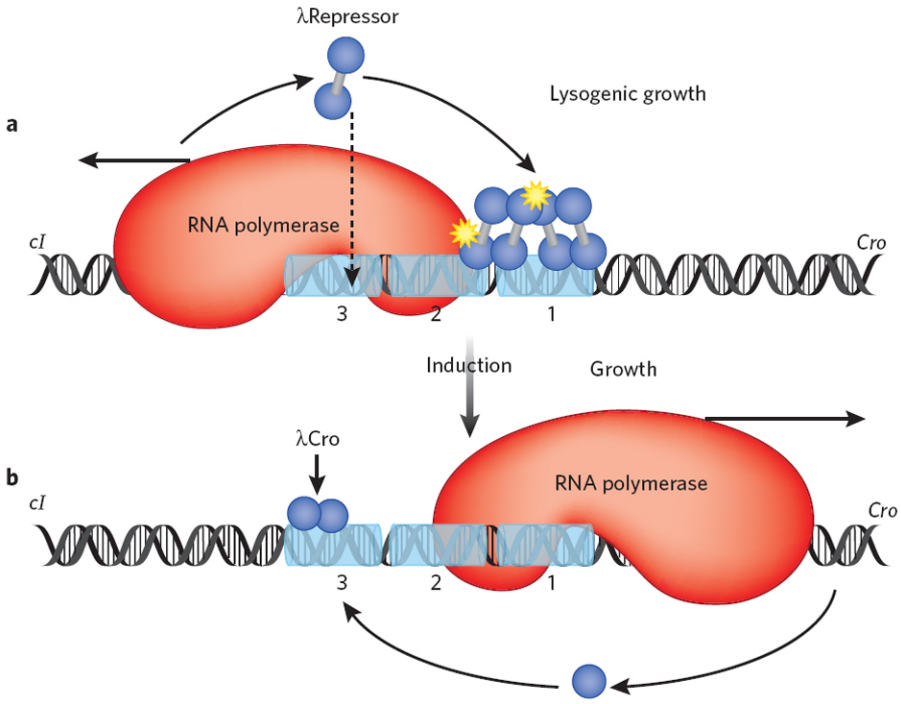

Figure 11: Maintenance de la lysogeny (**a**) et activation du cycle lytique (**b**) (Ptashne 2011). Les rectangles bleus 1, 2 et 3 représentent les sites opérateurs O_R. La fixation de dimères du répresseur CI sur O_R1 et O_R2 empêche la transcription de l'opéron "early right" et active la transcription du gène *cI*. Lorsque la concentration en répresseurs CI diminue, la transcription de l'opéron "early right" (codant pour Cro) est activée. La fixation de Cro sur O_R3 inhibe la transcription de *cI* et maintient la transcription de l'opéron "early right".

D'autres protéines secondaires de Lambda sont impliquées dans la régulation telles que les anti-terminateurs N et Q et les protéines CII et CIII. Au début de l'infection, CII est notamment impliquée dans la décision lyse/lysogénie via l'activation de la transcription de *cI* et de *int* (Ptashne 1992). Le fonctionnement du régulateur CIII reste relativement mal compris mais il semble stabiliser le taux de régulateurs CII (Ptashne 1992). Certains travaux suggèrent que les protéines hôtes H-NS auraient pour effet de réprimer la transcription des éléments

exogènes (Navarre, et al. 2006). Enfin, il a été montré récemment que le facteur de terminaison hôte Rho peut également jouer un rôle important dans le maintien de la lysogénie (Menouni, et al. 2013).

Il est possible d'induire le cycle lytique des prophages par différentes méthodes. Ces différents procédés tels que l'irradiation à la lumière UV ou l'action de la mitomycine C reposent sur l'endommagement de l'ADN. Ces agents physiques ou chimiques vont causer des lésions à l'ADN bactérien, ayant pour effet d'activer la réponse SOS. Lors de la réponse SOS, la protéine RecA est activée et conduit au clivage du répresseur CI (Ptashne 1992). En conséquence, la concentration de CI va drastiquement diminuer dans la cellule, ce qui va permettre à Cro d'être exprimé et de se fixer sur le site opérateur O_R et donc d'activer le cycle lytique de Lambda. Il a été observé que le taux d'excision spontané de Lambda est de l'ordre de 10^{-6} par division cellulaire en conditions non stressantes (Ball and Johnson 1991; Gottesman and Yarmolinsky 1968). Ce phénomène peut être imputé à des variations stochastiques du taux d'expression de CI, ainsi qu'à la stabilité des ARNm et des protéines produits (Arkin, et al. 1998). Enfin, le niveau de condensation de l'ADN chromosomique peut influencer le taux d'induction de Lambda (Norregaard, et al. 2014).

2.3 Evolution des lambdoïdes et le mosaïcisme phagique

2.3.1 Définition des lambdoïdes et conflits taxonomiques

La définition du terme "lambdoïde" porte souvent à confusion. Ceci est dû en grande partie au caractère empirique de la définition originelle de ce terme. D'abord, les lambdoïdes ont été définis comme les phages ayant la capacité de former des recombinants viables avec le phage Lambda, témoignant ainsi d'une proximité évolutive et fonctionnelle entre ces phages (Casjens 2008). Cependant, il a par la suite été observé que certains phages pouvaient recombiner avec des phages lambdoïdes mais pas avec Lambda lui-même. Cette observation a permis d'étendre la définition des lambdoïdes à d'autres phages, de telle sorte que différents phages possédant peu de gènes communs pouvaient se retrouver classés comme lambdoïdes. Finalement, une définition moins empirique des lambdoïdes peut être donnée sur la base de

leur organisation génomique commune. Les lambdoïdes partagent en effet certaines caractéristiques génomiques: taille du génome, organisation des opérons et organisation des gènes (Hendrix and Casjens 2006). Ces caractéristiques permettent, en théorie, aux lambdoïdes d'échanger des gènes par recombinaison (directement ou indirectement).

Cette particularité des lambdoïdes permet ainsi de regrouper des phages apparemment très différents au sein d'un même groupe. En effet, alors que Lambda est un *Siphoviridae*, P22 un *Podoviridae* et SfV un *Myoviridae*, ces trois phages sont des lambdoïdes et partagent un certain nombre de gènes (Juhala, et al. 2000; Mmolawa, et al. 2003). A l'inverse, les méthodes usuelles de détection d'homologie de séquence suggèrent qu'un phage tel que P22 ne partage pratiquement aucun gène commun avec d'autres *Podoviridae* comme Epsilon15 ou T7 (Rohwer and Edwards 2002). L'absence de similarité de séquences ne témoigne pas forcément d'une absence d'homologie puisque cela pourrait également traduire une divergence très ancienne de ces virus ou une évolution rapide de leurs séquences. Il y a donc un conflit entre la classification de l'ICTV et les méthodes de classification basées sur la similarité génomique. La comparaison de deux phages lambdoïdes tels que Lambda et P22 permet de mettre en évidence que ces deux phages partagent de nombreux gènes très similaires (gènes de réplication, recombinaison et régulation) mais présentent aussi des gènes très différents (gènes de structure) (Casjens 2008). Le terme "lambdoïde" n'est donc pas un rang taxonomique à proprement dit mais traduit plutôt une certaine promiscuité évolutive de ces phages liée à la recombinaison.

2.3.2 Organisation génomique modulaire conservée et morons

Nous avons vu que les lambdoïdes peuvent être définis comme des entités qui partagent une même organisation génomique. Cette caractéristique permet ainsi aux lambdoïdes d'échanger des gènes par recombinaison. Il en résulte une caractéristique typique des lambdoïdes: le mosaïcisme génomique (Botstein 1980). La comparaison des génomes de différents phages lambdoïdes a mis en évidence que ces phages partagent typiquement des régions de très grande identité de séquences et des régions très différentes (Fig 12) (Juhala, et al. 2000). Ces régions très similaires ou très dissimilaires correspondent généralement à des modules de gènes impliqués dans des fonctions semblables (Casjens and Hendrix 1974). En effet, on peut ainsi décrire le génome des lambdoïdes par différents modules fonctionnels majeurs: intégration/excision, recombinaison généralisée, régulation, réplication, lyse, encapsidation,

tête et queue. Outre le regroupement des gènes impliqués dans des fonctions identiques au sein d'un même module, les motifs d'ADN du génome phagique ciblés par ces gènes sont généralement situés au sein du module correspondant (Casjens and Hendrix 1974). Ainsi, le site *attP* ciblé par l'intrégrase se situe juste en amont du gène codant pour l'integrase. L'origine de réplication *ori* du phage est située au sein de la séquence du gène de réplication *O*. Le site de clivage *cos* de la terminase nécessaire à l'encapsidation est situé à proximité des gènes codants pour cette enzyme. Les différents modules codent typiquement pour des protéines qui interagissent entre elles et peu avec les autres protéines ou séquences phagiques (Casjens and Hendrix 1974). Cela favorise l'autonomie fonctionnelle de ces modules et donc leur transmission d'un génome à l'autre. Il est probable que cette organisation modulaire des génomes de lambdoïdes soit le résultat du taux important d'échange de séquences promu par une recombinaison peu spécifique. Le regroupement en modules des séquences dont les produits interagissent ensemble leur procurerait ainsi un plus grand succès évolutif.

Il est à noter qu'à part les modules fonctionnels essentiels liés à la régulation, la réplication et la morphogénèse, il existe d'autres gènes ou modules accessoires. La fonction de ces gènes est souvent inconnue et généralement non essentielle (voir (Court and Oppenheim 1983)). Le génome de Lambda possède par exemple la région *nin* (dans l'opéron "early right") codant potentiellement pour 10 protéines (Fig 8). Il a été montré que la délétion de cette grande région n'a pas d'impact important sur le taux de production de particules de Lambda (Hendrix and Casjens 2006). La taille et le contenu de ces régions varient beaucoup d'un lambdoïde à un autre (Juhala, et al. 2000).

J'ai mentionné précédemment que la majorité des gènes des lambdoïdes sont regroupés en trois grands opérons et que les différents gènes sont regroupés en modules fonctionnels au sein de ces opérons. Il existe cependant des gènes qui ont la particularité d'avoir leur propre promoteur de transcription et qui se situent fréquemment au sein de modules de gènes aux fonctions très différentes des leurs. Ces gènes ont été nommés "morons" parce qu'ils constituent une addition de matériel génétique qui interrompent des opérons et des modules fonctionnels ("more on" en anglais) (Hendrix and Casjens 2006). Ces gènes ont souvent des fonctions qui confèrent un avantage à la bactérie hôte et sont typiquement exprimés durant la lysogénie. Le fait de posséder leur propre promoteur transcriptionnel permet à ces gènes de ne pas être réprimés par le répresseur CI durant la lysogénie. Il est possible d'observer un même moron situé dans différents modules de phages lambdoïdes (Hendrix and Casjens 2006). Ce dernier point souligne la capacité des lambdoïdes à échanger ces gènes dans des contextes génomiques différents. Le phénomène de mosaïcisme n'est pas limité aux lambdoïdes et

d'autres phages semblent également présenter un mosaïcisme génomique important (Petrova, et al. 2013; Pope, et al. 2013). Cependant, certains phages tempérés, tels que les phages P2-like, présentent des contenus en gènes beaucoup moins divers (Nilsson and Haggard-Ljungquist 2007). Malgré la présence de gènes accessoires variables, la comparaison de leurs génomes ne suggère pas une évolution mosaïque telle qu'elle est observée chez les lambdoïdes (Nilsson and Haggard Ljungquist 2006; Nilsson and Haggard-Ljungquist 2007).

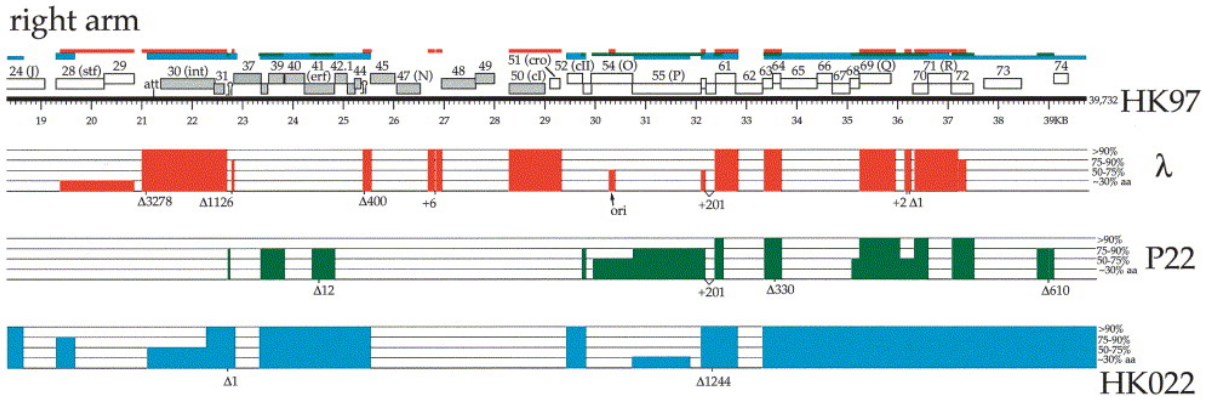

Figure 12: Exemple de mosaïcisme entre le phage HK97 et d'autres phages lambdoïdes: Lambda (rouge), P22 (vert) et HK022 (bleu) (Juhala, et al. 2000). Les régions homologues sont indiquées par les couleurs. La hauteur des barres de couleur indique le taux d'identité (entre 30 et 100%) entre les protéines des régions homologues chez Lambda, P22 et HK022. L'absence de couleurs indique une très faible identité de séquences protéiques (<30%).

2.3.3 Mécanismes à l'origine du mosaïcisme

J'ai décrit précédemment comment l'organisation modulaire des lambdoïdes facilite leur mosaïcisme génomique. Les mécanismes moléculaires à l'origine de ce mosaïcisme restent assez débattus. Néanmoins, les lambdoïdes codent généralement pour des enzymes de recombinaison généralisée (Lopes, et al. 2010). Nous avons vu que ces enzymes sont impliquées dans la réplication et l'encapsidation des génomes dans les particules de tête (Smith 1983). Différentes familles de recombinases ont été décrites et plusieurs d'entre elles ont été observées au sein des génomes de lambdoïdes (Lopes, et al. 2010; Murphy 2012). De même, il existe un deuxième type d'inhibiteur de RecBCD parmi les lambdoïdes (décrit chez le phage P22): Abc2 (Murphy 2012). Ces recombinases détectées chez les lambdoïdes médient des réactions de recombinaison homologue. Une analyse expérimentale de la recombinaison des phages lambdoïdes a mis en évidence que la recombinase Redβ de Lambda peut agir sur des séquences assez divergentes (Martinsohn, et al. 2008). L'évolution

des phages lambdoïdes pourrait ainsi être promue par un mécanisme de recombinaison homologue plus permissif, qualifié de recombinaison "homéologue" (Martinsohn, et al. 2008). De plus, un nouveau mécanisme de recombinaison de l'enzyme Redβ a récemment été proposé et permettrait des échanges importants d'ADN (Maresca, et al. 2010). Ce modèle a d'importantes implications pour la compréhension du mosaïcisme phagique puisqu'il suggère qu'il est possible d'échanger de longues séquences non homologues flanquées par de courtes régions homologues (ou homéologues) (Maresca, et al. 2010). Cependant, le mosaïcisme génomique des phages est généralement très contrasté: des régions de forte similarité sont typiquement entourées de régions très divergentes. D'autres travaux ont souligné que ce schéma est assez peu compatible avec un mécanisme de recombinaison homologue qui nécessiterait la conservation de séquences similaires pour permettre l'échange de modules différents. Si la conservation de telles séquences en bordures des modules a effectivement été observée dans certains génomes (Clark, et al. 2001), ceci ne semble pas pouvoir expliquer la majorité des cas (Hendrix and Casjens 2006). Pour cette raison, il a été suggéré que la recombinaison illégitime joue un rôle prépondérant dans le mosaïcisme phagique (Hendrix, et al. 1999). La recombinaison illégitime permet également d'expliquer la présence de morons homologues au sein de modules fonctionnels différents. L'origine du mosaïcisme génomique n'est donc pas une question résolue. Finalement, il est important de souligner que les deux hypothèses ne sont pas mutuellement exclusives. De plus, la forte diversité des phages et de leurs gènes permet de supposer que notre connaissance actuelle des stratégies et des mécanismes de recombinaison phagiques est encore très lacunaire.

3 Impact des bactériophages sur l'évolution bactérienne

3.1 Ecologie et course aux armements

3.1.1 L'organisme le plus abondant et son impact

Différents travaux ont permis d'estimer que les phages sont les organismes les plus abondants sur terre (Biers, et al. 2008; Rohwer and Thurber 2009; Suttle 2007). Il est généralement considéré qu'il existe environ 10^{31} particules phagiques sur terre contre 10^{30} cellules bactériennes. Cependant, ces estimations sont probablement surestimées à cause de la confusion de ces particules virales avec des vésicules membranaires ou des Agents de Transfert de Gènes (ATGs) (Forterre, et al. 2013). Quoiqu'il en soit, les phages représentent des entités que les bactéries ont une très forte probabilité de rencontrer. Il a ainsi été suggéré que les phages représentent la principale pression de sélection agissant sur les bactéries. Le taux de cellules continuellement lysées par les phages pourrait même avoir une influence importante sur les cycles géochimiques océaniques (Fuhrman 1999). Les phages permettraient également de maintenir une certaine diversité bactérienne selon l'hypothèse "Kill the winner" (Thingstad and Lignell 1997). Selon ce modèle, une augmentation du nombre de bactéries hôtes (le vainqueur) conduirait à une augmentation de ses phages, ce qui aurait pour conséquence d'accroître le taux de mortalité de l'hôte. En définitive, les phages empêcheraient les populations bactériennes dominantes de s'imposer et pourraient ainsi maintenir une diversité bactérienne plus importante.

3.1.2 La co-évolution et la course aux armements

De part leur abondance, les phages représentent une menace importante pour les bactéries. Non seulement la lyse mène à la destruction d'une bactérie mais ce processus conduit à la multiplication du nombre de phages dans l'environnement, ce qui peut détruire une grande partie des populations bactériennes. Face à cette menace, les bactéries ont développé de nombreux systèmes de défense contre les infections phagiques.

i) Elles peuvent modifier l'expression, la diversité ou l'accessibilité des protéines membranaires ciblées par les phages par un processus nommé "variation de phase". Ces modifications peuvent être programmés par différents mécanismes moléculaires: inversions site-spécifiques, glissement programmé de l'ADN polymérase, sécrétion de polysaccharides formant une capsule ou rétro-transcription (Bikard and Marraffini 2012; Labrie, et al. 2010).
ii) Les systèmes de restriction-modification sont utilisés pour détruire l'ADN phagique par dégradation des séquences qui ne possèdent pas les mêmes motifs de méthylation grâce à une action méthylase et endonucléase (Kobayashi 2001).
iii) Un autre système peut également dégrader l'ADN phagique: les CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats) associés aux gènes *cas* (CRISPR associated proteins) (Deveau, et al. 2010). Ce système permet de cibler spécifiquement des séquences d'ADN phagique par production de petites séquences d'ARN complémentaires couplée à l'action d'une endonucléase Cas et d'incorporer de nouvelles séquences phagiques au sein du CRISPR (Garneau, et al. 2010).
iv) Les systèmes d'avortement d'infection (systèmes Abi) permettent à la bactérie de stopper l'infection phagique par divers mécanismes. Typiquement, ces systèmes ciblent certaines étapes clés du cycle phagique telles que la recombinaison généralisée, empêchant ainsi l'encapsidation d'ADN phagique (Bouchard and Moineau 2004). Cette stratégie défensive implique parfois le suicide de la bactérie grâce aux systèmes toxine-antitoxine (Koga, et al. 2011). En se suicidant, la bactérie éviterait la destruction de l'ensemble de la population (Fukuyo, et al. 2012).
v) Enfin, des travaux récents sur les vésicules membranaires bactériennes ont montré que ces structures sont ciblées par les phages (Biller, et al. 2014). Les auteurs de cette étude ont ainsi suggéré que la production de vésicules membranaires par les bactéries pourrait diminuer significativement le taux d'infection. Ces vésicules possèdent en effet les récepteurs cellulaires ciblés par les phages. Selon cette hypothèse, les phages ne pourraient pas engager leur cycle lytique après infection de vésicules et ces éléments constitueraient donc de véritables leurres cellulaires.
Les phages s'adaptent continuellement à ces différents systèmes de défense. La plupart de ces systèmes peuvent être contournés par l'apparition spontanée de mutations ponctuelles. Par exemple, les mutations de gènes codants des protéines structurelles peuvent rétablir l'affinité du phage pour un récepteur ou lui permettre de s'attacher à de nouveaux récepteurs (Drexler, et al. 1989; Hofnung, et al. 1976). L'échange de modules génétiques par recombinaison donnant lieu au mosaïcisme phagique semble faciliter l'adaptation des phages à leurs hôtes.

De plus, ce phénomène permet également aux phages de modifier leurs répertoires géniques. Ce dernier point est d'autant plus important que les phages possèdent fréquemment des gènes leur permettant de contourner les défenses bactériennes (Samson, et al. 2013). Par exemple, les phages peuvent parfois coder pour l'antitoxine et empêcher ainsi l'avortement prématuré du cycle lytique (Otsuka and Yonesaki 2012). Il est intéressant de souligner que les phages peuvent parfois utiliser les mêmes systèmes que les bactéries pour contourner leurs défenses. Il a ainsi été montré récemment que les phages peuvent posséder leur propre système CRISPR-Cas pour échapper au système défensif de l'hôte (Seed, et al. 2013).

L'adaptation des phages conduit à l'adaptation permanente des bactéries pour leur échapper dans un processus de course aux armements comme développé dans l'hypothèse de la reine rouge (Van Valen 1973). Les bactéries peuvent ainsi s'adapter par mutations ponctuelles mais également via le renouvellement de leurs répertoires de gènes médié par transferts horizontaux (Avrani, et al. 2011). Les phages et leurs hôtes sont donc en co-évolution constante puisqu'ils entretiennent une relation hôte-parasite. Cette relation a certainement un impact important sur la diversification de ces deux entités.

3.1.3 L'ambigüité des phages tempérés: faux ennemis ou faux amis?

Bien que les parasites aient un impact délétère sur leurs hôtes, leur survie dépend de ces derniers. Il est par conséquent dans l'intérêt des parasites de minimiser leur impact sur leurs hôtes ou de compenser cet impact en contribuant temporairement à la fitness de l'hôte. Par exemple, le cyanophage virulent S-PM2 augmente momentanément la fitness de son hôte en codant pour des gènes de photosynthèse (Mann, et al. 2003). La présence de ces gènes permet ainsi à la cellule d'assurer la production d'énergie suffisante à la reproduction du virus durant l'infection.

Les phages tempérés entretiennent une relation encore plus complexe avec leurs hôtes. En effet, ces phages ont la capacité de stabiliser leur ADN au sein des bactéries et généralement au sein de leur chromosome. Il en résulte que ces phages partagent un intérêt commun avec leur hôte sur une plus longue période de temps. En d'autres termes, l'ADN phagique peut être perçu comme une séquence bactérienne à temps partiel. Il est donc de l'intérêt du phage de contribuer à la fitness de son hôte puisque la multiplication bactérienne est alors directement corrélée à la multiplication du génome phagique. Il est ainsi très fréquent que les phages tempérés codent pour diverses fonctions bénéfiques pour leurs hôtes (Brussow, et al. 2004;

Wang, et al. 2010). Les gènes phagiques codant pour ces fonctions ne sont généralement pas impliqués dans des fonctions phagiques essentielles (protéines de structures, de lyse, etc) mais contribuent indirectement à la fitness du phage en contribuant à celle de l'hôte. Les bactéries lysogènes peuvent ainsi être privilégiées par rapport à leurs semblables dépourvues de prophages. Cependant, ces bénéfices portés par les prophages peuvent s'avérer délétères à plus long terme. En effet, si aucune mutation n'a affecté de gènes essentiels durant la résidence du prophage au sein de son hôte, il est probable que le prophage soit induit et conduise à la lyse de la cellule. Les prophages agissent tels des cadeaux empoisonnés qui peuvent bénéficier à la bactérie dans un premier temps mais peuvent aussi la détruire à plus long terme.

3.2 Impact des prophages à court terme: un génotype bactérien étendu

3.2.1 La transduction

Le transfert d'ADN par transduction est promu par les phages (Zinder 1955). Ce mécanisme consiste en l'encapsidation d'ADN bactérien dans la capside virale pouvant être ensuite transféré à une autre cellule par infection. Les bactéries peuvent donc bénéficier de l'acquisition de nouveaux gènes bactériens par ce processus tels que des gènes de résistance aux antibiotiques (Muniesa, et al. 2013). Il existe deux mécanismes de transduction: la transduction spécialisée et la transduction généralisée. i) La transduction spécialisée est due aux prophages intégrés au sein du chromosome bactérien. Lors de l'excision, il arrive parfois que la recombinaison se ne fasse pas entre les deux sites de recombinaison originels (*att*) mais entre des sites secondaires (Campbell 1969). L'excision imparfaite du prophage peut alors mener à l'encapsidation d'une molécule d'ADN hybride constituée d'ADN phagique et d'un fragment d'ADN bactérien (Campbell 1969). Etant donné la limitation du volume d'ADN encapsidable par le virus, ce phénomène ne permet le transfert que de petites quantités d'ADN bactérien (typiquement <5kb) (Campbell 2006). Le fragment ainsi transféré via la particule virale, peut intégrer le chromosome de la cellule receveuse par intégration du virus via son intégrase. ii) La transduction généralisée résulte d'un empaquetage accidentel de fragments d'ADN exclusivement bactériens dans la capside du phage (Susskind and Botstein 1978). Elle permet donc le transfert de plus grandes quantités d'ADN hôte (environ 93kb pour le phage P1 par exemple) (Lennox 1955; Lobocka, et al. 2004). Ce phénomène ne se produit que pour

les phages utilisant un système d'encapsidation à tête pleine comme chez le lambdoïde P22 (Susskind and Botstein 1978) (voir section 2.2.3). Ne dépendant pas de motifs spécifiques d'encapsidation il est possible, lorsque de l'ADN bactérien non circulaire est disponible dans la cellule, d'encapsider aléatoirement certains fragments du chromosome. Jusqu'à 2% des phages P22 peuvent ainsi remplir leur capside avec de l'ADN bactérien (Ebel-Tsipis, et al. 1972a). L'ADN bactérien pourra alors être transféré à d'autres cellules par infection. L'ADN ainsi injecté peut intégrer le chromosome de la cellule receveuse par recombinaison homologue. Il est important de noter que le taux de particules virales contenant de l'ADN bactérien est relativement faible. La majorité des particules encapside uniquement de l'ADN viral. Néanmoins, ces taux d'encapsidation d'ADN bactérien peuvent varier de manière importante (jusqu'à 200 fois plus pour P22) chez certains virus mutants (Casjens, et al. 1992; Schmieger 1972; Wall and Harriman 1974).

3.2.2 Protection contre la sur-infection

La présence d'un prophage confère un avantage direct à la cellule hôte en lui conférant une immunité contre d'autres infections phagiques: "l'exclusion". Chez Lambda, le répresseur CI exprimé par le prophage va ainsi bloquer l'induction du cycle lytique par toute infection supplémentaire de Lambda (Court and Oppenheim 1983; Gussin, et al. 1983). Ce mécanisme est essentiel à l'établissement de la lysogénie, car les particules virales étant libérées simultanément par la cellule dans l'environnement, les sur-infections par un même phage sont très fréquentes. La répression agissant par la fixation du répresseur sur les sites opérateurs, ce mécanisme permet de réprimer l'infection par des phages apparentés. Ce mécanisme d'exclusion est aussi qualifié d'homoimmunité.

Des mécanismes d'hétéroimmunité, c'est à dire d'exclusion de phages taxonomiquement différents, sont fréquemment observés. Ces systèmes d'exclusion impliquent alors la présence de gènes accessoires non essentiels (Court and Oppenheim 1983). Certains systèmes semblent cibler certains mécanismes généraux du cycle phagique, ce qui permet d'exclure de nombreux phages différents, aussi bien tempérés que virulents. Le phage Lambda possède les gènes *rexA* et *rexB* qui protègent la cellule contre les infections de phi80, P22 et certains mutants des phages T4, T5 et T7 (Benzer 1955; Jacquemin-Sablon and Lanni 1973; Pao and Speyer 1975; Susskind and Botstein 1980; Toothman and Herskowitz 1980). L'action des gènes *rex* n'est pas connue en détails mais ils semblent agir durant la réplication (Court and Oppenheim

1983). Lambda et d'autres lambdoïdes codent également pour le gène *sieB* qui permet l'exclusion de différents phages tempérés (Ranade and Poteete 1993). Le gène accessoire *old* du phage P2 code pour une exonucléase qui dégrade l'ADN de Lambda (Myung and Calendar 1995).

La présence de systèmes d'exclusion codés par les prophages représente un bénéfice important pour les bactéries hôtes. L'acquisition d'un prophage permet ainsi d'éviter la lyse de la cellule et parfois l'infection par d'autres phages hétérologues. Ceci bénéficie en retour au phage qui porte ces gènes mais nuit aux phages exclus. Il y a ainsi une véritable compétition entre phages basée sur l'acquisition de systèmes d'exclusions et sur l'apparition de résistance à ces systèmes (Refardt 2011).

3.2.3 Gènes augmentant la croissance de l'hôte

Il a été suggéré que les prophages codent parfois pour des gènes augmentant la fitness de leur hôte en stimulant directement leur taux de croissance (Edlin, et al. 1977; Lin, et al. 1977). Des tests de croissance ont été effectués *in vitro* en présence de divers prophages: Lambda, Mu, P1 et P2. Les résultats de ces travaux suggèrent une augmentation de la fitness des lysogènes sur les non lysogènes lorsque le milieu de culture est appauvri en glucose. Les mécanismes à l'origine de cette augmentation de croissance restent cependant mal compris.

3.2.4 Gènes liés à la pathogenèse

Les prophages peuvent également avoir un impact positif sur leurs hôtes en augmentant leur pathogénicité ou leur résistance. L'introduction de tels gènes au sein de l'hôte peut entraîner des modifications phénotypiques importantes telles que la capacité d'infecter différents organismes ou la résistance à certains antibiotiques. Par exemple, la toxine shiga est codée par les gènes *stx* de divers phages lambdoïdes et cause des diarrhées sanguinolentes et des colites hémorragiques humaines (Allison 2007). Un cas d'épidémie récent a eu lieu en Allemagne et le prophage ayant introduit la toxine a pu être identifié (Laing, et al. 2012). Il semble cependant que la toxine Stx soit originellement utilisée comme système de défense contre les protistes (Los, et al. 2012). La présence de tels gènes au sein des phages tempérés

témoignerait d'un alignement entre les intérêts du prophage et du lysogène car la multiplication bactérienne promeut directement la multiplication du prophage. De nombreuses toxines codées par des phages ont été référencées (voir (Boyd, et al. 2012; Brussow, et al. 2004)). Les prophages peuvent également coder pour des fonctions accessoires augmentant la pathogénicité bactérienne. Le gène *lom* de Lambda permet ainsi d'améliorer l'adhérence d'*E. coli* aux cellules de mammifères, ce qui peut alors faciliter les infections (Hendrix and Casjens 2006). Une étude récente suggère que les phages seraient également des vecteurs de gènes de résistance aux antibiotiques (Modi, et al. 2013).

3.2.5 Utilisation des prophages comme armes

Il a été suggéré que l'induction d'un prophage peut permettre de tuer les bactéries ne contenant pas cet élément (Brown, et al. 2006; Gama, et al. 2013; Joo, et al. 2006). A l'inverse les bactéries lysogènes sont immunisées contre ce phage. Au sein d'une population bactérienne, l'induction d'un prophage va donc permettre d'éliminer les compétiteurs sensibles, sans affecter les individus apparentés issus de la multiplication clonale. L'induction de l'élément viral va engendrer la lyse de la bactérie émettrice mais peut bénéficier aux autres clones de la population par sélection de parentèle (Hamilton 1964a, b). Ce mécanisme a cependant un effet limité dans la mesure où une partie des bactéries non lysogènes pourront devenir lysogènes par intégration chromosomique du phage et ainsi devenir résistant à ce même phage (Gama, et al. 2013). Néanmoins, bien que temporaire, cet avantage compétitif pourrait permettre aux bactéries lysogènes de coloniser de nouvelles niches (Brown, et al. 2006).

3.2.6 Effets délétères liés à l'intégration

Bien que les prophages puissent apporter des avantages sélectifs, plusieurs effets délétères peuvent être attribués à leur présence. Leur impact sur la fitness bactérienne pourrait donc résulter d'un équilibre entre leurs effets avantageux et délétères. Premièrement, l'induction de ces éléments conduit à la lyse de la cellule. Nous avons vu que les prophages sont généralement induits par la réponse SOS. La réponse SOS est déclenchée lorsque la bactérie a subi des lésions génomiques importantes (Bjedov, et al. 2003). Elle ne permet cependant pas de réparer ces dommages systématiquement. Il est donc probable que cette induction du cycle

lytique par la réponse SOS diminue les effets délétères de la lyse sur la bactérie dans la mesure où celle-ci aurait peut-être succombé à ces dommages. L'induction stochastique des prophages (10^{-6} par division cellulaire) représente en revanche un fardeau direct diminuant la fitness du lysogène (Gottesman and Yarmolinsky 1968). Il est à noter également que les phages tempérés s'intègrent souvent au sein de gènes d'ARNt sans en altérer leur séquence (Campbell 2002). Néanmoins, de nombreux prophages (28%) semblent être intégrés au sein de gènes codants, ayant généralement pour effet d'interrompre la phase codante de ces gènes (Fouts 2006). Ceci suggère donc que l'intégration des phages peut avoir des effets délétères pour son hôte en inactivant certains gènes. Enfin, nous avons vu précédemment que les génomes bactériens sont des entités compactes et organisées. Or, certains génomes bactériens peuvent contenir de nombreux prophages (18 chez *E. coli* O157:H7 Sakaï) (Asadulghani, et al. 2009). Ceci laisse envisager que des pressions sélectives agissent sur l'intégration des phages afin de minimiser les coûts de fitness pour l'hôte (Lawrence and Hendrickson 2003). Enfin, il a été montré récemment que le prophage Mu est structuré en un domaine surenroulé à l'instar du reste du chromosome d'*E. coli* (Saha, et al. 2013). Cela suggère que les phages peuvent s'adapter passivement ou activement aux contraintes structurelles du chromosome hôte de manière à réduire les effets délétères liés à leur intégration.

3.3 Impact des prophages à long terme: source de nouveaux gènes et domestication

3.3.1 Les prophages cryptiques: reliques ou entités fonctionnelles?

Les infections phagiques sont très fréquentes. Il a été estimé qu'il se produit environ 10^{23} infections virales par seconde dans les océans (Suttle 2007). Des analyses conduites sur des échantillons marins ont permis d'estimer que 43% des bactéries contiennent des prophages inductibles (Jiang and Paul 1994). D'autre part, bien que les prophages puissent coder pour des fonctions bénéfiques pour leurs hôtes, la majorité des gènes phagiques codent pour des fonctions essentielles au phage telles que des protéines constituant la particule phagique et les protéines de réplication. Ces gènes n'étant pas fonctionnels pour la bactérie, il est attendu qu'ils disparaissent du génome hôte au cours du temps. J'ai en effet mentionné précédemment

que les séquences d'ADN non fonctionnelles disparaissent rapidement des génomes bactériens par délétions (Kuo and Ochman 2010) (section 1.3.2).

Différents travaux ont décrit le contenu en prophages de divers génomes bactériens (Canchaya, et al. 2004; Casjens 2003). Si certains éléments correspondent à des prophages *a priori* complets et fonctionnels, de nombreux prophages semblent être en cours de dégradation (Asadulghani, et al. 2009) et sont généralement désignés prophages "cryptiques" (Wang, et al. 2010). Il a été montré que ces prophages cryptiques peuvent apporter un avantage sélectif pour la bactérie grâce à la présence de gènes accessoires, tels ceux décrits précédemment (Wang, et al. 2010). Les prophages cryptiques semblent ainsi représenter un intermédiaire de dégradation des prophages, où les séquences phagiques non fonctionnelles pour l'hôte sont en cours de dégradation. La dégradation de séquences non fonctionnelles étant rapide chez les bactéries (Kuo and Ochman 2010), la présence fréquente de ces éléments pourrait s'expliquer par le renouvellement fréquent des prophages au sein des génomes bactériens.

Plusieurs études ont également étudié de manière exhaustive la fonctionnalité des prophages présents au sein de certains génomes bactériens (Asadulghani, et al. 2009; Matos, et al. 2013). Ces études ont mis en évidence que de nombreux prophages de grande taille sont fréquemment non fonctionnels. Par exemple, parmi les 18 prophages détectés chez *E. coli* O157:H7 Sakaï, un seul (sp5) s'avère entièrement fonctionnel (Asadulghani, et al. 2009). Les 17 prophages restants présentent différents niveaux de défauts génétiques: incapacité à s'exciser, défaut de réplication, incapacité d'encapsidation ou formation de particules virales non infectieuses. Certains de ces prophages ont la capacité de former des particules infectieuses à faible taux, probablement par complémentation fonctionnelle entre différents prophages. Ces études suggèrent qu'il existe une forte pression de sélection menant à la désactivation rapide de la majorité des prophages, y compris de prophages de grande taille. Il semble donc que le nombre de prophages cryptiques (généralement identifiés sur la base de leur taille) soit largement sous-estimé. De nombreux prophages d'apparence complète ne seraient ainsi pas fonctionnels non plus.

Enfin, il est possible que des prophages partiellement dégradés soient mobilisables par complémentation fonctionnelle, à l'instar des phages satellites tels que P4 (Six and Klug 1973). Le phage satellite P4 est hautement spécialisé dans le parasitisme du phage P2. P4 et d'autres phages satellites codent pour des protéines leur permettant d'interférer spécifiquement avec leur phage "helper" et sont typiquement dépourvus de gènes codant pour la particule virale (Christie and Dokland 2012). Ces caractéristiques suggèrent que ces éléments sont

hautement spécialisés dans un mode de vie parasitique. Il est cependant envisageable que des populations mixtes de phages non spécialisés dans ce mode de vie entrent en compétition pour l'encapsidation du génome phagique au sein de la particule. Il a en effet été montré que des génomes hautement dégradés de phages peuvent utiliser la particule de phages apparentés pour se mobiliser (Enea, et al. 1977). Dans quelle mesure ce phénomène est-il fréquent et stable dans le temps reste indéterminé.

3.3.2 Le renouvellement des prophages: une source d'ADN pour innover

Le taux élevé de renouvellement de séquences phagiques pourrait être à l'origine de différentes innovations génétiques bactériennes par exaptation. L'exaptation consiste en l'utilisation d'un gène ou d'un caractère dans un but différent de sa fonction originelle (Gould and Vrba 1982). Par exemple, les plumes des oiseaux présentaient initialement des fonctions de thermorégulation, puis ont été exaptées en structures impliquées dans le vol (Gould and Vrba 1982). Ce phénomène pourrait jouer un rôle important dans l'acquisition de nouvelles fonctions génétiques (Barve and Wagner 2013). Plusieurs cas d'exaptation de séquences phagiques ont conduit à leur domestication par leurs hôtes: les bactériocines de types R et F, le système de sécrétion de type VI (T6SS) et les agents de transfert de gènes (ATGs). Le terme "domestication" désigne ici la stabilisation et l'utilisation de séquences mobiles exogènes par l'hôte à ses propres fins. Dans le reste de cette section, je vais décrire ces différents systèmes bactériens qui semblent résulter de la domestication de prophages.

Les bactériocines de type R et F. Ces éléments ont été originellement décrits chez *Pseudomonas aeruginosa* sous le nom de pyocines de type R et pyocines de type F (Michel-Briand and Baysse 2002). Ils constituent des structures protéiques de poids moléculaire important qui sont structurellement similaires aux queues de phages *Myoviridae* (bactériocines de type R) (Fig 13) et de phages *Siphoviridae* (bactériocines de type F) (Nakayama, et al. 2000). Ces structures de forme cylindrique possèdent la capacité de perforer la membrane de cellules bactériennes, ce qui aboutit à leur mort. Les bactériocines sont ainsi utilisées dans la compétition bactérienne. Ces éléments correspondent typiquement à la domestication de gènes de queue phagiques formant la structure protéique en forme d'aiguille, de gènes de lyse permettant la libération de ce complexe et de gènes de régulation. De telles entités ont été décrites chez diverses bactéries Gram positives et Gram négatives (Coetzee, et

al. 1968; Gebhart, et al. 2012; Strauch, et al. 2001; Thompson and Pattee 1981; Zink, et al. 1995).

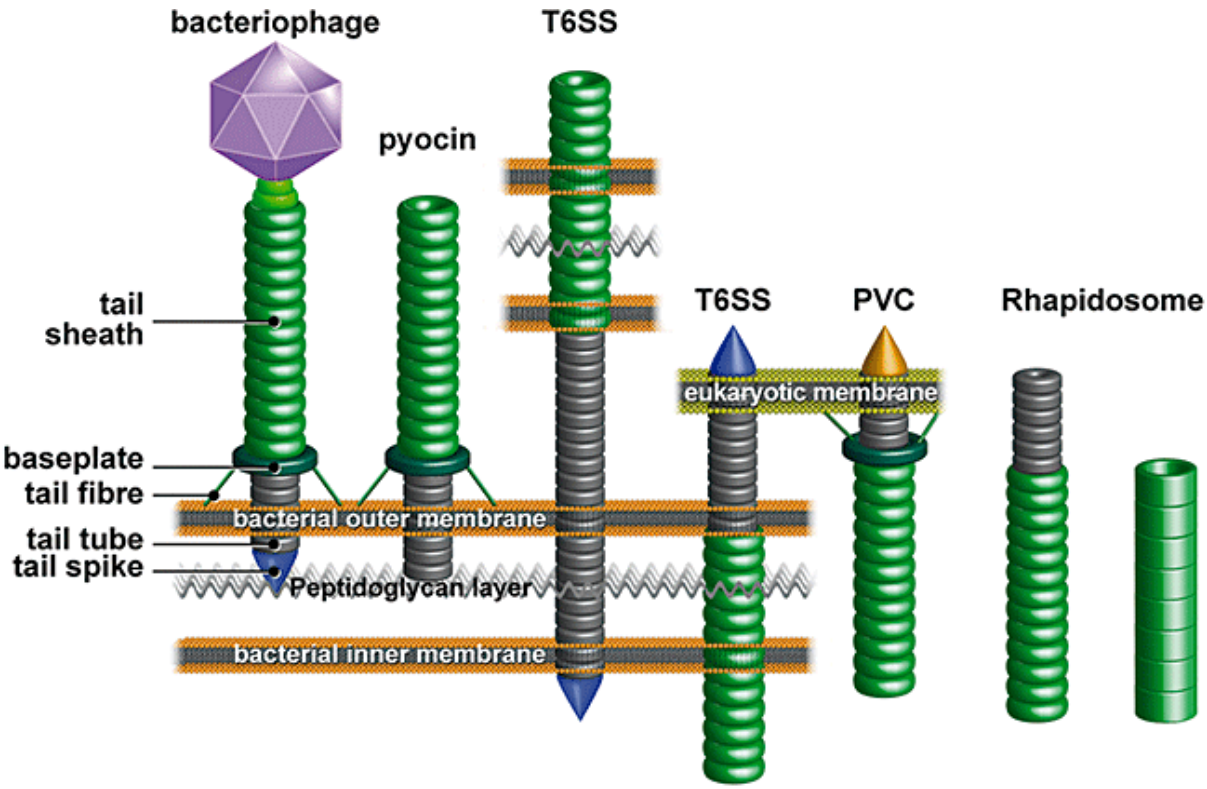

Figure 13: Similarité entre phages *Myoviridae* et différents systèmes bactériens (Bonemann, et al. 2010). Le système PVC (*Photorhabdus* Virulence Cassette) est introduit section 3.3.3.

Le système de sécrétion de type VI. Cet élément est un système de sécrétion observé chez diverses bactéries (Boyer, et al. 2009). Plusieurs protéines majeures de cet élément sont homologues à des gènes phagiques (Bonemann, et al. 2010). L'aspect général de sa structure, son mécanisme de sécrétion ainsi que la séquence de certains de ses composants suggèrent que cet élément est issu de la domestication de gènes de queue de phages *Myoviridae* (Fig 13) (Bonemann, et al. 2010).

Le rhapidosome. Cette structure ressemble aussi très fortement à une queue contractile de *Myoviridae*. Elle a été découverte il y a plus de 40 ans (Yamamoto 1967). Depuis, elle a été observée chez diverses bactéries mais sa fonction reste inconnue (Bonemann, et al. 2010). Contrairement aux bactériocines de type R, auxquelles ils ressemblent, les rhapidosomes sont des structures de plus grande taille, ne présentent pas d'activité bactéricide et semblent être exprimés spontanément (Delk and Dekker 1972).

Les agents de transfert de gènes. Ces éléments permettent le transfert d'ADN entre cellules bactériennes via la formation d'une particule très similaire aux particules phagiques. Ils semblent donc également être d'origine phagique (Lang and Beatty 2007). Leur aspect général est en effet très similaire à celui de particules de phages caudés (Fig 14). Peu d'ATGs ont été décrits mais ils ont été identifiés chez différents clades bactériens et archéens, ce qui laisse suggérer que ces systèmes sont potentiellement présents chez de très nombreux procaryotes. Ces structures sont codées par des groupes de gènes majoritairement synténiques de 14kb à 30kb (Lang and Beatty 2007). Il est intéressant de noter que ces éléments ont généralement la capacité d'encapsider des séquences d'ADN de longueur inférieure à celle de leur propre cluster de gènes. Ceci permet donc de ne pas les considérer comme des éléments mobiles. L'analyse des taux relatifs de substitutions synonymes et non synonymes semble indiquer que ces systèmes évoluent sous sélection purificatrice (Lang and Beatty 2007). L'aspect général des ATGs laisse envisager que leur mécanisme d'encapsidation et de transfert d'ADN est très similaire à celui des phages mais le fonctionnement précis de ces systèmes n'a été que peu étudié.

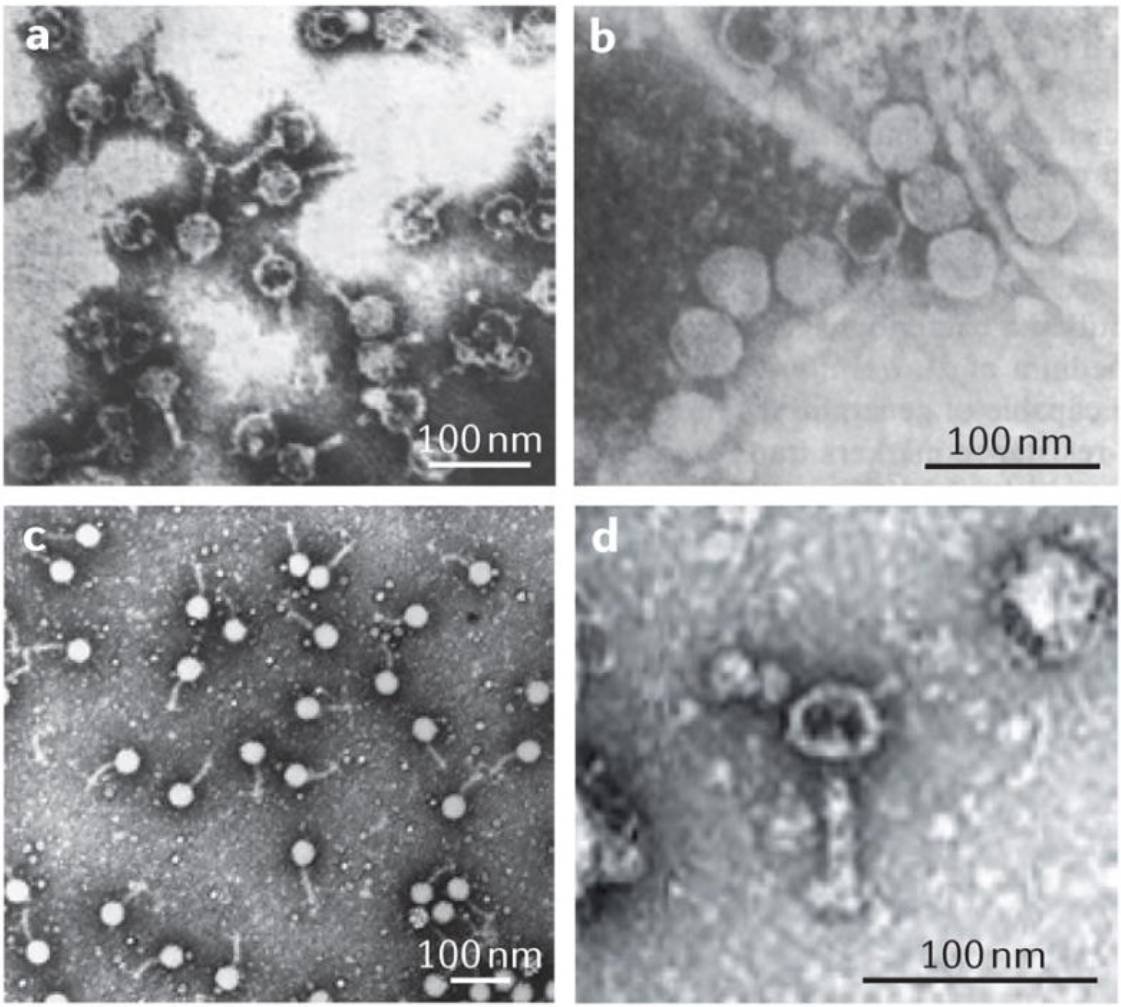

Figure 14: Micrographies électroniques d'agents de transfert de gènes (ATGs) observés chez différentes espèces (Lang, et al. 2012). **a** *Rhodobacter capsulatus* (Yen, et al. 1979). **b** *Desulfovibrio desulfuricans* (Rapp and Wall 1987). **c** *Brachyspira hyodysenteriae* (Humphrey, et al. 1997). **d** *Methanococcus voltae* (Eiserling, et al. 1999).

3.3.3 Ambigüités entre prophages défectifs et systèmes domestiqués

J'ai décrit dans la section précédente différents cas de domestication de séquences phagiques. Cependant, il existe différents éléments génétiques bactériens d'origine phagique dont la nature et la fonction sont plus ambiguës.

Bien que les bactériocines de type R et F constituent un exemple classique de la domestication des gènes phagiques de queue par la bactérie, de nombreux autres éléments agissant comme bactériocines ont été décrits. En effet, plusieurs études ont décrit la production de "particules phagiques tueuses" par diverses bactéries (Bradley 1967; Garro and Marmur 1970). Par stimulation à la mitomycine C ou à la lumière UV, de nombreuses

bactéries présentent la capacité de produire des particules phagiques tueuses mais non infectieuses. Ces particules ont la capacité de tuer des cellules sensibles sans les infecter et agissent ainsi comme des bactériocines (Campbell 1977). Contrairement aux bactériocines de type R et F, les particules phagiques tueuses ont une structure plus proche de phages complets (Fig 15). En effet, ces éléments ressemblent largement à des particules phagiques défectueuses. Différents niveaux d'altération peuvent être observés (Bradley 1967): i) la particule est complète et contient de l'ADN phagique mais est incapable d'injecter l'ADN au sein de la cellule cible. ii) La particule est complète mais ne contient pas d'ADN. iii) L'assemblage de la tête et de la queue est défectueux. Il est à noter que la production d'éléments présentant simultanément différents niveaux de dysfonctionnement peut être observée. Par exemple, *Listeria monocytogenes* peut produire simultanément des particules complètes tête-queue sans ADN et des particules ne correspondant qu'à des queues phagiques (Fig 15) (Bradley 1967). Malgré ces différents niveaux de dysfonctionnement, ces particules sont définies par leur capacité à tuer les bactéries sensibles (ne codant pas l'élément) et leur incapacité à les infecter (Campbell 1977). Il est important de noter que ces éléments semblent être conservés au sein de différentes souches isolées indépendamment à travers le monde (Campbell 1977). Ceci suggère que ces éléments sont des systèmes conservés par leur hôte à l'instar des bactériocines de type R et F. La particule tueuse PBSX de *Bacillus subtilis* est probablement l'élément le mieux décrit. Cet élément a été caractérisé génétiquement et ressemble fortement à un prophage défectueux (Wood, et al. 1990). La présence de tels éléments au sein des génomes bactériens souligne l'ambigüité entre prophages défectueux et systèmes domestiqués. Il semble ainsi qu'il existe un continuum allant des prophages fonctionnels jusqu'aux bactériocines de types R ou F. Ceci suggère également que le recrutement de séquences de prophages en tant que bactériocines est assez fréquent. Une étude a montré que l'altération chimique de particules du phage virulent T4 permet de former spontanément des particules tueuses non infectieuses (Duckworth 1970). Cela permet d'envisager que la modification de prophages fonctionnels en bactériocines ne pourrait impliquer que quelques mutations altérant la structure de la particule phagique. L'apparition de telles entités à partir de prophages fonctionnels pourrait donc être assez fréquente.

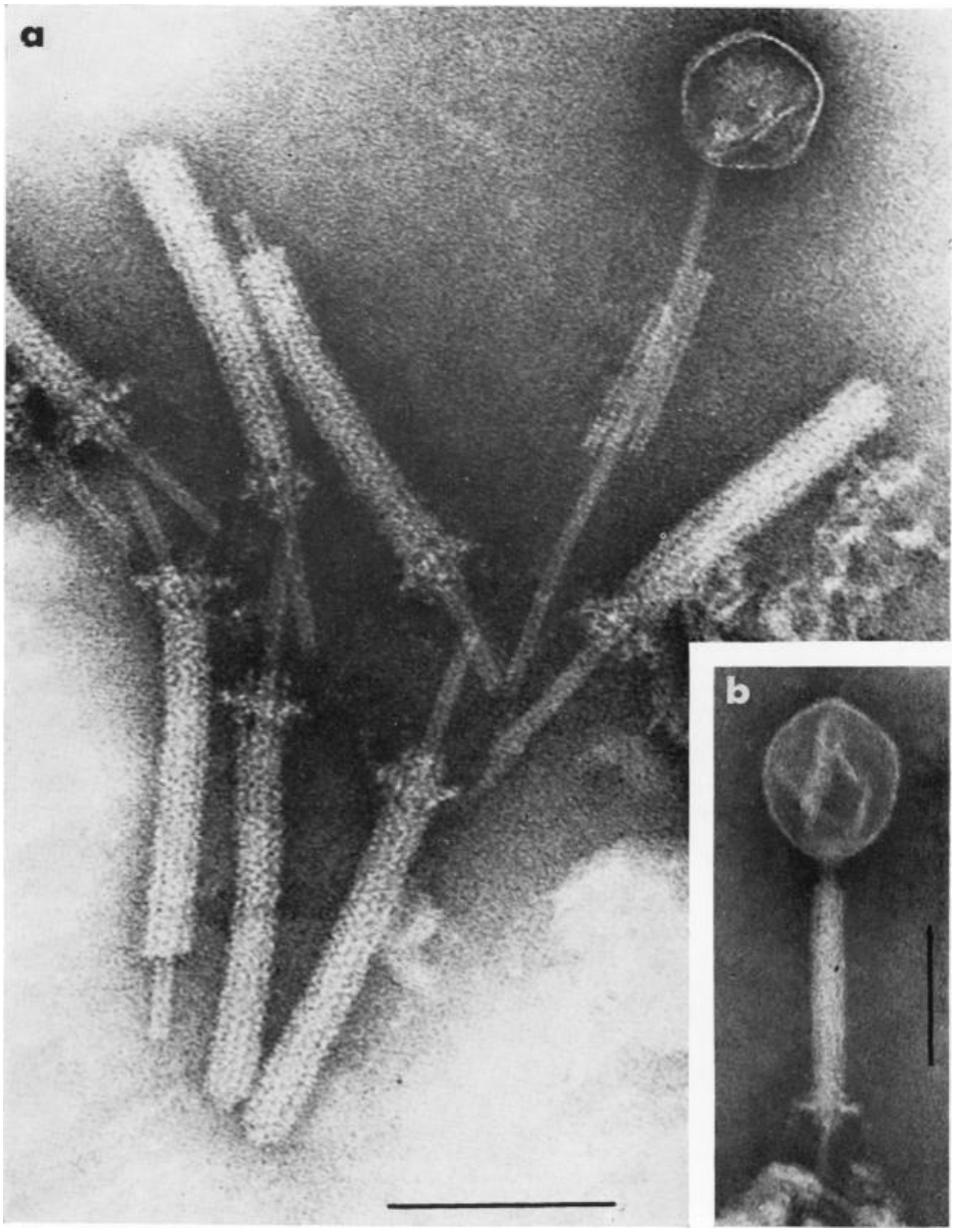

Figure 15: Micrographies électroniques de particules phagiques tueuses de *L. monocytogenes* (**a** et **b**) (Bradley 1967).

Il est à noter que les particules phagiques défectueuses semblent être domestiquées pour d'autres fonctions. En effet, il a été montré que certaines bactéries ont la capacité de produire des structures nommées PVC (*Photorhabdus* Virulence Cassette) correspondant à des queues phagiques et qui agissent contre certains insectes (Fig 13) (Hurst, et al. 2004; Yang, et al. 2006). De manière intrigante, il a été montré récemment que la bactérie *Pseudoalteromonas luteoviolacea* induit la métamorphose du ver tubulaire *Hydroides elegans* via la production de structures ressemblant fortement à des faisceaux de queues contractiles de phages *Myoviridae*

(Fig 16) (Shikuma, et al. 2014). Ces études suggèrent que les bactéries sont capables de recruter des structures phagiques pour un large éventail de fonctions.

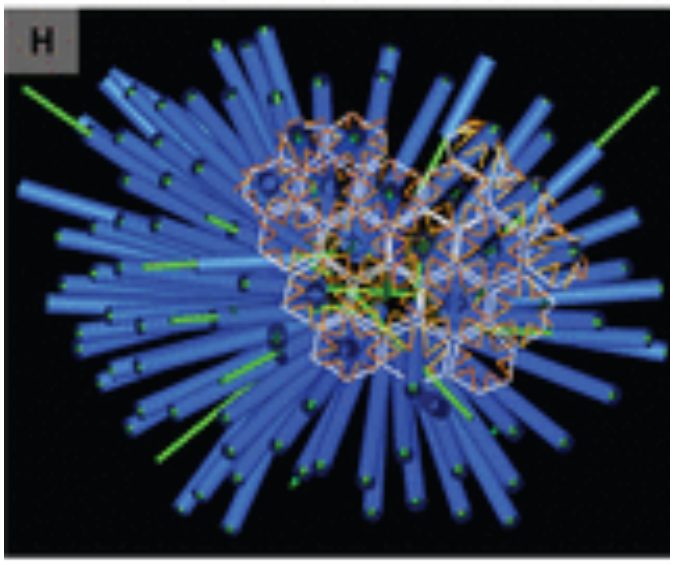

Figure 16: Modélisation des structures MAC (Metamorphosis-Associated Contractile structure) durant leur assemblage intracellulaire chez *P. luteoviolacea* (Shikuma, et al. 2014). Cette structure a été modélisée à partir de micrographies électroniques. Ces éléments (vert) ressemblent à des queues phagiques entourées de gaines contractiles (bleu). Chaque structure est fortement similaire à une queue contractile de *Myoviridae*. Ces éléments sont organisés en un faisceau régulier au sein de la cellule grâce à l'interaction de protéines similaires aux protéines de fibres phagiques (orange). Ces dernières interagissent entre elles et forment des hexagones relativement réguliers.

Nous avons vu que les ATGs permettent le transfert horizontal de gènes bactériens (Lang, et al. 2012). J'ai également mentionné que certains phages fonctionnels sont naturellement capables d'accomplir cette fonction à faible taux (jusqu'à 2%) via la transduction généralisée (Ebel-Tsipis, et al. 1972a; Ebel-Tsipis, et al. 1972b; Lang and Beatty 2007). Néanmoins, certaines mutations affectant la morphogénèse de la capside phagique permettent d'obtenir des phages présentant des capacités de transduction généralisée fortement augmentées (jusqu'à 200 fois) (Casjens, et al. 1992; Iida, et al. 1998). Ceci laisse envisager, qu'à l'instar des particules phagiques tueuses, il existe un continuum allant des prophages fonctionnels, transducteurs occasionnels d'ADN bactériens, aux ATGs, transducteurs spécialisés d'ADN hôte.

Les précédents cas de domestication envisagés impliquent l'exaptation de différents gènes phagiques formant des structures complexes. Il est cependant envisageable que la domestication d'un seul ou de quelques gènes phagiques soit plus fréquente. Il a été observé

que deux gènes de queue ont été domestiqués par *Streptomyces mitis* et *Enterococcus faecalis*. Ces gènes permettent de faciliter l'adhésion des bactéries hôtes aux plaquettes sanguines, favorisant ainsi les endocardites infectieuses (Bensing, et al. 2001; Matos, et al. 2013). Il est intéressant de souligner que l'avantage sélectif fourni par ces deux gènes phagiques implique la conservation de nombreux gènes phagiques additionnels. Les protéines de structure étant codées par les gènes de l'opéron "late", l'expression de ces protéines nécessite la transcription de l'opéron "late" et donc l'induction du cycle lytique (Matos, et al. 2013). Ainsi, l'expression de ces gènes requiert la conservation de gènes de régulation et de réplication. La conservation de ces différentes fonctions annexes, nécessaires à l'expression de protéines phagiques domestiquées, est probablement une source d'ambigüité entre prophages cryptiques et éléments domestiqués. Le recrutement de protéines de structure phagiques pour l'adhésion cellulaire pourrait en outre être un phénomène assez fréquent. Il a en effet été observé que les phages caudés contiennent fréquemment (~25%) des protéines de structure à domaines immunoglobulines (Ig-like) (Fraser, et al. 2006). Ces domaines protéiques permettent aux particules virales d'adhérer à la surface des muqueuses de divers métazoaires, favorisant ainsi leur probabilité de rencontrer un hôte bactérien (Barr, et al. 2013). Il est probable que cette capacité d'adhésion de nombreuses protéines phagiques puisse avoir été domestiquée à plusieurs reprises par les bactéries pour coloniser de nouvelles niches ou pour augmenter leur infectivité.

Objectifs

De nombreuses études ont mis en avant l'impact des phages sur la diversité bactérienne, en soulignant notamment que les phages sont eux-mêmes très divers. Les étapes allant de la formation de cette diversité des gènes phagiques jusqu'à la prise de contrôle de ces gènes par la bactérie hôte restent cependant largement incomprises. Les phages sont des prédateurs cellulaires et leur intégration dans le génome hôte est contrainte par l'organisation du génome bactérien. Les processus à l'origine de la diversification des phages restent mal compris. Enfin, le recrutement des gènes phagiques par la bactérie est une étape clé qui reste largement inconnue.

L'objectif général de la thèse a été d'étudier la dynamique des prophages au sein des génomes d'entérobactéries. Comment les phages contribuent-ils à l'enrichissement de la diversité des répertoires de gènes bactériens?

Pour cela trois objectifs ont été fixés:

i) Le premier objectif a eu pour but de détecter et classer les prophages d'*Escherichia* et de *Salmonella* afin d'étudier leur organisation par rapport à l'architecture des génomes hôtes.

ii) Dans un deuxième temps, je me suis focalisé sur les différentes stratégies de recombinaison utilisées par les phages tempérés afin de comprendre comment cela affecte leur diversité génétique.

iii) Enfin, le troisième objectif de la thèse a consisté à étudier la dynamique de dégradation des prophages au sein des génomes d'entérobactéries et des pressions sélectives affectant ces éléments.

Résultats

I L'adaptation des phages tempérés à leurs génomes hôtes.

Contexte

Les phages tempérés s'intègrent fréquemment au sein du chromosome hôte et certaines bactéries peuvent contenir des quantités importantes d'ADN phagique (Asadulghani, et al. 2009). Pourtant, le génome bactérien est une structure organisée qui répond à d'importantes contraintes génétiques et structurelles (section 1.2). L'objectif de cette analyse a été d'étudier comment l'intégration des prophages est organisée par rapport à l'architecture du génome hôte. Quels sont les processus adaptatifs qui permettent aux phages de s'accommoder aux contraintes du génome hôte?

Approche

Détection des prophages

La détection des prophages a constitué une étape majeure de la thèse. Les méthodes de détection de ces éléments sont en constante évolution et aucune méthode ne semble entièrement satisfaisante (Srividhya, et al. 2007). Deux types d'approches sont typiquement utilisées: i) les détections basées sur la recherche d'homologies de séquences. ii) les détections basées sur la composition en oligonucléotides. Dans les deux cas, ces méthodes visent à détecter des blocs de gènes contigus (îlots génomiques) spécifiquement d'origine phagique. La difficulté réside dans le fait que les phages sont extrêmement divers (Mokili, et al. 2012) et qu'ils partagent des caractéristiques communes avec d'autres éléments génomiques (section 1.3.6). Les prophages présentent également des gènes homologues à des composants de divers éléments mobiles (intégrase, systèmes de restriction/modification, éléments IS, etc). Enfin, les phages peuvent aussi présenter des gènes homologues avec la bactérie hôte.

Le premier type de méthodes, telles celles développées avec Phage Finder, Prophinder ou PHAST (Fouts 2006; Lima-Mendez, et al. 2008a; Zhou, et al. 2011), reposent sur la détection de gènes phagiques connus. Ces méthodes ont pour avantage de ne rarement confondre les

prophages avec d'autres éléments lorsqu'elles s'appuient sur la construction de banques de données et sur l'identification de gènes phagiques spécifiques tels que les gènes impliqués dans la lyse, la réplication ou la morphogénèse de la particule (Phage Finder et Prophinder). Ces approches dépendent donc fortement de notre état des connaissances actuelles. Nous pouvons donc en déduire que ces méthodes sont peu performantes lorsqu'elles sont confrontées à la présence de prophages taxonomiquement éloignés des éléments décrits jusqu'à présent. Les phages infectant des genres bactériens différents partagent généralement une faible similarité de séquence entre eux. Il est donc attendu que ces méthodes soient peu adaptées aux genres bactériens pour lesquels les phages n'ont été que peu étudiés. Les méthodes de détection basées sur la recherche d'homologies sont également peu sensibles aux prophages dégradés. Ces éléments ont perdu une part importante de leurs gènes et cela réduit ainsi la probabilité que ces approches détectent des gènes reconnus comme spécifiquement phagiques.

La deuxième catégorie d'outils de détection est basée sur les biais de composition en oligonucléotides des génomes (Nicolas, et al. 2002; Srividhya, et al. 2007). Ces approches permettent de détecter des prophages fortement dégradés et des prophages appartenant à des taxons jusqu'alors non décrits. Les biais de composition étant communs à différents types d'îlots génomiques (section 1.3.6), la capacité de ces méthodes à différencier les prophages des autres éléments génomiques reste peu convaincante. De plus, il est possible que la composition en oligonucléotides de ces éléments converge vers celle du génome hôte au cours du temps (Lawrence and Ochman 1997). Améliorer ces approches par une définition empirique précise des biais de composition des prophages reviendrait en outre à réintroduire le biais taxonomique existant chez les approches par recherche d'homologies de séquences. Finalement, une méthode récente a combiné l'approche par biais de composition avec la recherche d'homologies: PhiSpy (Akhter, et al. 2012). Malgré l'intérêt apparent que représente l'élargissement méthodologique de cet outil, il est à noter que la combinaison des deux approches n'efface pas les écueils propres à chacune d'entre elles. La synthèse des deux méthodes conduirait soit à ajouter une contrainte supplémentaire (mais la détection par homologies de séquences est déjà très restrictive), soit à considérer l'ensemble des résultats produits par les deux méthodes (mais la recherche par biais de composition ne garantit pas la détection de régions nécessairement d'origine phagique).

De part la plus grande restriction des approches de recherche d'homologies et parce que les phages d'entérobactéries sont relativement bien décrits, j'ai choisi de principalement appuyer ma détection sur Phage Finder, Prophinder et PHAST (Fouts 2006; Lima-Mendez, et al.

2008a; Zhou, et al. 2011). Dans cette première étape, j'ai utilisé une approche conservative afin d'éviter de détecter des régions non phagiques: je n'ai sélectionné que les prophages candidats supérieurs à 10kb et je n'ai conservé que ceux qui présentaient au moins un gène "core" (spécifique des phages) comme défini par Phage Finder (Fig 17). Par des méthodes de génomique comparative, j'ai ensuite pu valider ou invalider ces prophages candidats et délimiter leurs bordures en m'appuyant principalement sur la fréquence des gènes orthologues partagés par les différentes souches (Fig 18). La détection des intégrases (situées à l'extrémités des prophages) et des éléments IS m'a aussi permis de préciser leurs limites et d'éliminer les îlots d'éléments IS (Fig 17). Les prophages intégrés en tandem (consécutifs) ont également pu être délimités grâce à différents critères: résultats incongruents entre les différents programmes de détection, redondance de certaines fonctions phagiques essentielles et taille importante de l'élément. Enfin, deux procédures de recherche de gènes orthologues et de gènes homologues m'ont permis de rechercher d'éventuels prophages non détectés (délétion de gènes phagiques core) et d'harmoniser la définition de leurs bordures.

Les *Inoviridae* présentant une faible taille, j'ai spécifiquement cherché la présence de ces éléments par blast. J'ai inspecté toutes les régions génomiques contenant deux hits significatifs situés à moins de 10kb l'un de l'autre. Ceci m'a permis de détecter trois Inovirus présentant chacun quatre gènes fortement similaires aux gènes d'*Inoviridae* de GenBank.

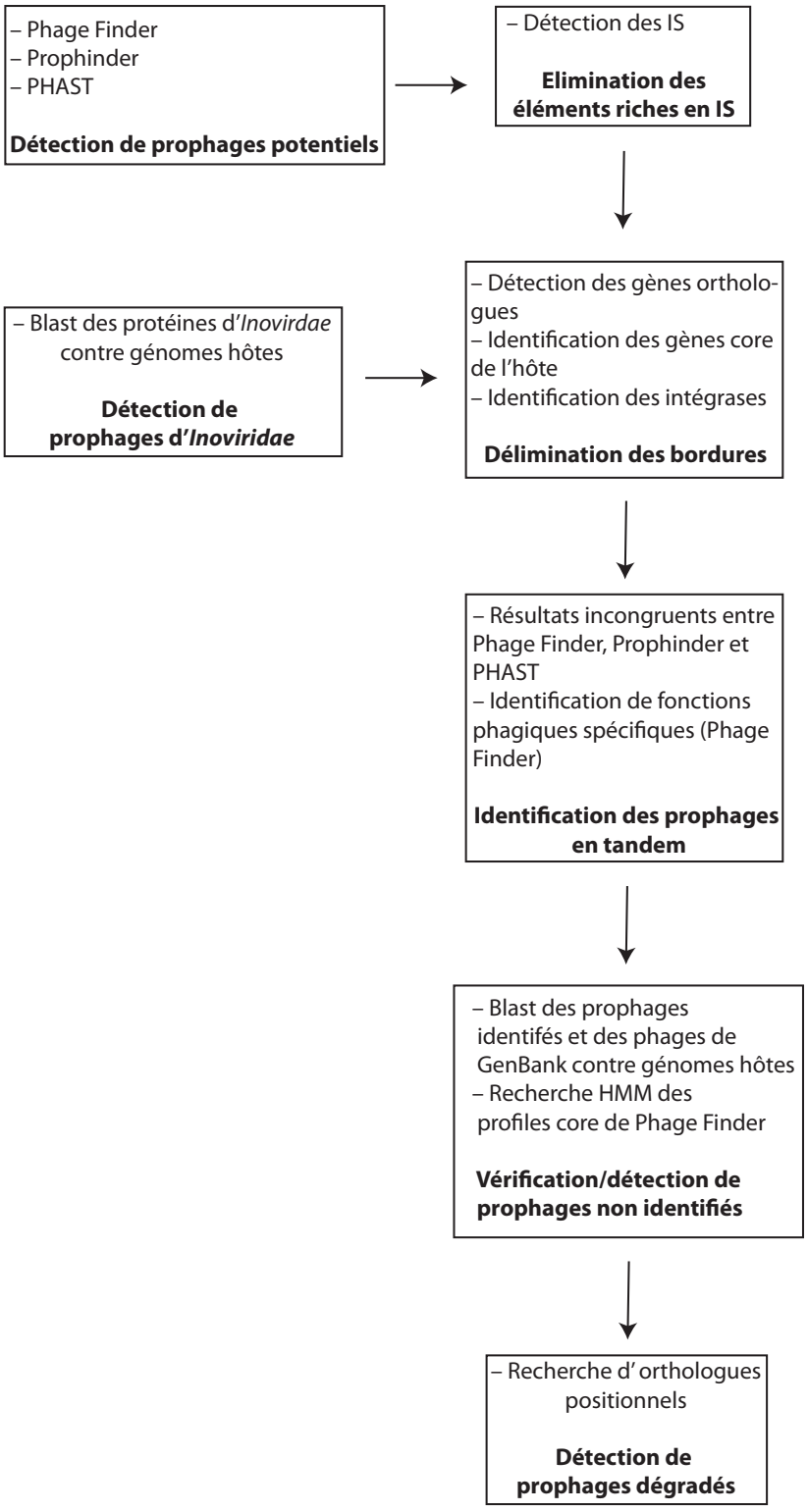

Figure 17: Procédure de détection des prophages.

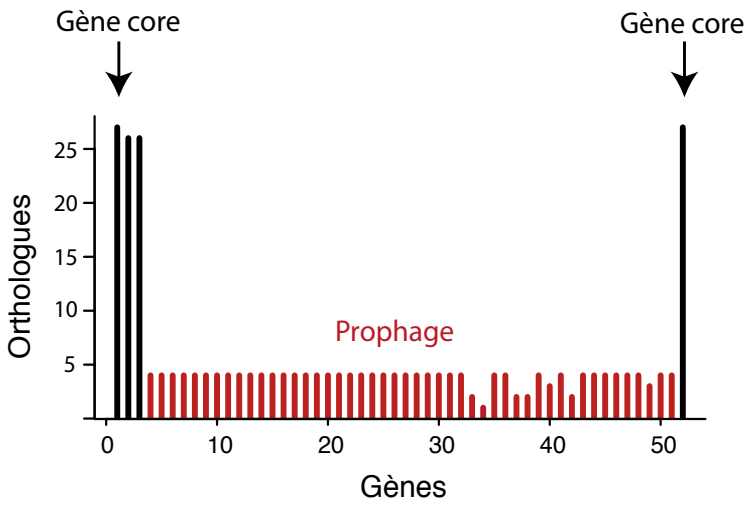

Figure 18: Exemple de délimitation des bordures d'un prophage de *S. enterica*. Les limites du locus du prophage sont définies par les gènes core de l'hôte. La fréquence des gènes orthologues au sein du locus des 27 souches de *S. enterica* permet de différencier les gènes du prophage (rouge) des gènes hôtes (noir).

Classification des prophages

J'ai décrit dans l'introduction les limites méthodologiques et conceptuelles liées à la classification des phages. Le mosaïcisme génomique des phages semble être un obstacle important au principe même de ces méthodes (Lawrence, et al. 2002). Des méthodes de phylogénie peuvent être utilisées pour comparer et classer des génomes sur la base des répertoires de gènes (Rohwer and Edwards 2002; Snel, et al. 2005). Les phages échangeant des modules fonctionnels de gènes, de nouvelles approches tentant de décrire les phages comme des combinaisons de modules ont été développées (Lima-Mendez, et al. 2008b). La comparaison et la représentation des génomes phagiques en réseaux permettent ainsi de définir des entités évolutives échangeant des gènes (Lima-Mendez, et al. 2008b). Cette approche rend possible le regroupement de phages ne partageant pas ou très peu de gènes au sein d'une même entité évolutive s'ils partagent des gènes avec d'autres phages communs. Il reste cependant difficile d'estimer la robustesse des regroupements générés par réseaux et différents niveaux de regroupements peuvent être obtenus par ces méthodes suivant les paramètres utilisés (Enright, et al. 2002). Ces méthodes sont donc particulièrement sujettes à des choix liés à l'utilisateur, ce qui affecte la reproductibilité et l'utilisation systématisée de ces méthodes. J'ai donc privilégié l'utilisation d'une représentation en dendrogramme (Rohwer and Edwards 2002) avec une méthode de phylogénie par Neighbour Joining avec BIONJ (Gascuel 1997).

L'établissement d'une classification basée sur l'utilisation de réseaux ou de méthodes de phylogénie sur les contenus génétiques repose avant tout sur la création d'une matrice de distances (ou scores) permettant de comparer les génomes phagiques entre eux. La définition du score est ainsi essentielle et son choix peut fortement affecter la classification. Le score peut être défini comme le nombre moyen de gènes partagés entre deux génomes de phages (Rohwer and Edwards 2002). J'ai utilisé une méthode similaire afin de comparer les prophages détectés aux phages classés et annotés d'entérobactéries présents dans GenBank. L'obtention d'un dendrogramme mêlant les prophages et les phages classés de GenBank m'a ainsi permis d'attribuer des familles et des genres à la majorité des prophages détectés. Le score défini par Rohwer et Edwards présente l'inconvénient d'être sensible aux délétions qui peuvent survenir durant la dégradation des prophages. Pour l'appliquer à la classification des prophages, il est donc essentiel de développer un score qui soit robuste aux délétions. Dans un premier temps, j'ai donc modifié le score: pour chaque paire de (pro)phages le score a été normalisé par le nombre de gènes du plus petit des prophages (et non plus pas le nombre de gènes moyen des deux (pro)phages). Un autre problème du score originel est lié au mosaïcisme phagique. Cette méthode présente l'inconvénient de ne pas différencier les gènes homologues hérités récemment de ceux issus d'une parenté ancienne. Durant mes travaux, j'ai ainsi pondéré ce score par la similarité propre à chaque paire de protéines homologues. Il est ainsi défini par $\sum_{i=1}^{M} \frac{S_{(Ai,Bi)}}{\min(n_A, n_B)}$ où $S_{(Ai,Bi)}$ est le score de similarité de la paire de gènes homologues *i* entre le (pro)phage A et le (pro)phage B, *M* est le nombre total de gènes homologues partagés par les (pro)phagse A et B et n_A et n_B représentent le nombre total de gènes des (pro)phages A et B respectivement. Suivant l'utilisation ou non de cette pondération, j'ai pu observer un changement majeur de la classification obtenue avec le positionnement des phages lambdoïdes SfV-like auprès d'autres phages *Myoviridae* non lambdoïdes. Si ces groupes de phages partagent effectivement une majorité de gènes en commun (les gènes de structure), ces gènes présentent de faibles similarités de séquences ce qui témoigne plus probablement d'un échange ancien. En utilisant un score du contenu en gènes pondéré par les similarités de séquence j'ai ainsi obtenu un seul groupe composé exclusivement de phages lambdoïdes.

Positions d'intégration des prophages

Afin de comparer les positions d'intégration des différents prophages j'ai utilisé l'ensemble des gènes orthologues partagés par l'ensemble des génomes d'*Escherichia* et de *Salmonella* respectivement: les génomes core. Ces gènes représentent plus de 25% de ces génomes bactériens, ce qui permet de définir les positions d'intégration des prophages de manière assez précise. Deux gènes core consécutifs définissent un locus. Les prophages situés entre deux gènes core ayant subi un réarrangement chromosomique n'ont pas été considérés comme situés au sein du même locus car il n'est alors pas possible d'attribuer l'appartenance au locus avec certitude. Afin de comparer les positions d'intégration des prophages entre *Salmonella* et *Escherichia*, j'ai construit le génome core des deux genres. Les prophages situés entre deux gènes core identiques ont été considérés comme appartenant à un même locus.

Article 1

The Adaptation of Temperate Bacteriophages to Their Host Genomes

Louis-Marie Bobay,*[,‡,1,2,3] Eduardo P.C. Rocha,[1,2] and Marie Touchon[1,2]

[1]Microbial Evolutionary Genomics Group, Institut Pasteur, Paris, France

[2]CNRS, UMR3525, Paris, France

[3]Université Pierre et Marie Curie, Cellule Pasteur UPMC, rue du Docteur Roux, Paris, France

[‡]Present address: Institut Pasteur, 25 rue du Docteur Roux, Paris, France

***Corresponding author:** E-mail: lbobay@pasteur.fr.

Associate editor: Csaba Pal

Abstract

Rapid turnover of mobile elements drives the plasticity of bacterial genomes. Integrated bacteriophages (prophages) encode host-adaptive traits and represent a sizable fraction of bacterial chromosomes. We hypothesized that natural selection shapes prophage integration patterns relative to the host genome organization. We tested this idea by detecting and studying 500 prophages of 69 strains of *Escherichia* and *Salmonella*. Phage integrases often target not only conserved genes but also intergenic positions, suggesting purifying selection for integration sites. Furthermore, most integration hotspots are conserved between the two host genera. Integration sites seem also selected at the large chromosomal scale, as they are nonrandomly organized in terms of the origin–terminus axis and the macrodomain structure. The genes of lambdoid prophages are systematically co-oriented with the bacterial replication fork and display the host high frequency of polarized FtsK-orienting polar sequences motifs required for chromosome segregation. *matS* motifs are strongly avoided by prophages suggesting counter selection of motifs disrupting macrodomains. These results show how natural selection for seamless integration of prophages in the chromosome shapes the evolution of the bacterium and the phage. First, integration sites are highly conserved for many millions of years favoring lysogeny over the lytic cycle for temperate phages. Second, the global distribution of prophages is intimately associated with the chromosome structure and the patterns of gene expression. Third, the phage endures selection for DNA motifs that pertain exclusively to the biology of the prophage in the bacterial chromosome. Understanding prophage genetic adaptation sheds new lights on the coexistence of horizontal transfer and organized bacterial genomes.

Introduction

Bacterial viruses, commonly known as bacteriophages or phages, are numerous and have an important impact in the regulation of bacterial populations in the environment and in the human microbiome (Weinbauer 2004; Suttle 2005; Breitbart et al. 2008; Reyes et al. 2010). Bacteriophages are very abundant and very diverse. Their genomes can be single stranded or double stranded, made of DNA or RNA, in one or several linear or circular molecules (Abedon and Calendar 2005). The International Committee on Taxonomy of Viruses (ICTV) bases phage taxonomy on the shape of virion particle (King et al. 2011). However, distinct families can exchange large DNA fragments blurring classical taxonomical definitions (Hendrix et al. 1999). Exchange of functional modules between phages leads to reticulate evolution and may favor their evolvability (Botstein 1980). Modularity and genetic compaction lead to highly organized genomes of phages, where genes involved in related functions or expressed at the same moment in the phage infectious cycle are generally clustered together and expressed within the same operon (Ptashne 1992). A large group of otherwise unrelated phages (called "lambdoid" phages) share phage Lambda's genomic organization (Campbell and Botstein 1983). This is thought to facilitate viable genome assortment by recombination (Juhala et al. 2000). The rapid evolution of phages by mutation and recombination and their lack of universal genes (contrary to prokaryotes) render classical phylogenetic approaches of little use. Alternative methods based on gene repertoire relatedness have thus been proposed (Rohwer and Edwards 2002; Lima-Mendez et al. 2008b). Our understanding of phages is largely derived from the study of a few clades, most notably phages of enterobacteria. Accordingly, metagenomic studies find few sequences homologous to known phages (Edwards and Rohwer 2005; Angly et al. 2006; Reyes et al. 2010).

Phages are bacterial parasites whose transmission involves, with rare exceptions, the death of the host by completion of a lytic cycle. However, some phages, so-called temperate phages, have the ability to enter a lysogenic state and replicate vertically with the host (Kourilsky 1973; St-Pierre and Endy 2008). Most temperate phages integrate into the chromosome. Under specific physiological conditions, the prophage excises from the chromosome and enters the lytic cycle. Integration and excision are usually mediated by a site-specific tyrosine or serine recombinase (Nunes-Duby et al. 1998;

Article

Fast Track

Smith and Thorpe 2002). Some temperate phages remain in the cell under the extrachromosomal form, for example, phage N15 of *Escherichia coli* (Ravin 2011). Other prophages integrate and transpose randomly in genomes using DDE transposases, for example, Mu (Mizuuchi 1992). Satellite phages code for the information necessary to subvert virions from other phages but not for their own virion particle, for example, the P4 phage subverts virions from the P2 phage (Six and Klug 1973). Finally, *Inoviridae* are small single-stranded DNA (ssDNA) phages that integrate as prophages in the chromosome using the host recombinases (Huber and Waldor 2002). Thus, although the temperate Lambda phage model was instrumental in our understanding of phages (Ptashne 1992), the genetics of temperate phages is very diverse.

Prophages express very few genes. Among genes essential to their biology, they typically express a repressor of the lytic cycle (Ptashne 1992). Prophages and their bacterial hosts have aligned interests in avoiding further infection by mobile genetic elements. Hence, elements that are important in phage warfare are also useful to the host (Shinedling et al. 1987; Nechaev and Severinov 2008; Van Melderen and Saavedra De Bast 2009; Labrie et al. 2010). Some prophages carry cargo genes encoding traits adaptive to the host, among which are virulence factors in many bacterial pathogens (Ohnishi et al. 2001; Banks et al. 2002; Boyd and Brussow 2002; Brussow et al. 2004; Thomson et al. 2004; Abedon and Lejeune 2005; Winstanley et al. 2008). Not only do prophages encode traits that can increase the host fitness, they can also be used as biological weapons against other bacteria (Bossi et al. 2003; Brown et al. 2006). Several prophages have been shown to increase the growth rates of their hosts under particular conditions, even in the absence of competing mobile genetic elements (Edlin et al. 1977). These examples suggest a symbiotic association between phages and bacteria (Roossinck 2011). However, most intact prophages kill the bacterial cell upon induction of the lytic cycle. There is thus a delicate balance between lysogeny and induction of the lytic cycle, and this has important consequences in the interaction between phages and hosts. Understanding the way prophages integrate and remain in genomes is important to understand this balance and to quantify the contribution of prophages to bacterial fitness.

The integration of phages may affect a number of the organizational traits of the bacterial chromosome (Reyes-Lamothe et al. 2008; Rocha 2008). 1) Genes encoding functional neighbors or interacting proteins cluster in operons and superoperons (Lathe et al. 2000; Zaslaver et al. 2006). 2) The transcription of most genes, and especially essential genes, is co-oriented with the replication fork (Rocha and Danchin 2003). 3) Highly expressed genes concentrate near the origin of replication in fast growing bacteria to enjoy replication-associated gene dosage effects (Couturier and Rocha 2006). 4) *Escherichia coli*'s chromosome is structured in four macrodomains and two nonstructured regions (Valens et al. 2004). Physical interactions are frequent within and rare between macrodomains. This chromosome structure has not yet been extensively investigated in other bacterial species. 5) The genome is packed with regulatory signals involved in cell processes such as translation, transcription, replication, chromosome structure, and segregation (Touzain et al. 2011). All these five organizational features are expected to constrain changes in bacterial genomes (Rocha 2004). Thus, large changes in chromosome structure are tolerated only when its organization is respected (Itaya et al. 2005; Cui et al. 2007; Esnault et al. 2007; Val et al. 2012). As a result, one would expect strong natural selection for phage integration in sites where it least affects the host fitness (Lawrence and Hendrickson 2003). Prophages are part of the chromosome. Thus, one would also expect selection for gene orientation and DNA motifs in the prophage matching the local and global chromosomal organization. Selection for such traits in phages is possible because most phages integrate at specific well-defined sites in the chromosome leaving reproducible prophage structures. Also, prophages and chromosomes have aligned interests whenever prophage organization within the genome improves, or at least does not negatively affect, the host fitness.

There have been indications that prophages are not randomly distributed in genomes. Notably, prophages encoding integrases of the tyrosine recombinase family tend to integrate at or close to the 3′ of transfer RNA (tRNA) or transfer-messenger RNA (tmRNA) genes possibly due to a preference for palindromic structures (Campbell 1992, 2003; Williams 2002, 2003). The current availability of very large data sets of complete genomes for *Escherichia*, *Salmonella*, and their phages opens up the possibility to study with a strong statistical basis the adaptation of prophages to the chromosome background. In this work, we focus on the patterns of phage integration and how these relate with local and global organizational features of the bacterial chromosome.

Results and Discussion

Identification of Prophages

We analyzed 47 completely sequenced genomes of *E. coli*, one from *E. fergusonii*, 20 from *Salmonella enterica*, and 1 from *S. bongori* (for details see supplementary table S1, Supplementary Material online). We identified prophages using Phage Finder (Fouts 2006), Prophinder (Lima-Mendez et al. 2008a), and PHAST (Zhou et al. 2011). We compared these independent predictions in the light of published information (Ohnishi et al. 2001; Casjens 2003; Canchaya et al. 2004; Thomson et al. 2004; Asadulghani et al. 2009). We precised prophage boundaries using sequence similarity to phages and the patterns of presence and absence of genes in the bacterial strains of the same species (see Materials and Methods). The few tandem prophages were curated manually. Smaller prophage remnants (putative defectives) are often very difficult to distinguish from other integrative elements. Therefore, we removed prophages smaller than 10 kb, as in Canchaya et al. (2003) and Casjens (2003). We removed 49 prophages with more than 25% of transposases in their gene repertoires. These elements are degraded and thus

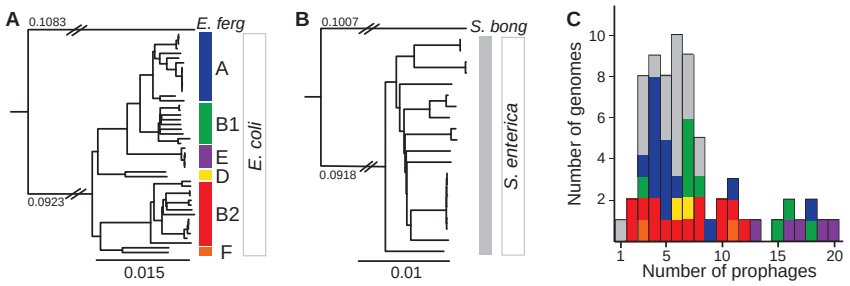

Fig. 1. Core genome phylogenies and prophage content of *Escherichia* and *Salmonella*. (*A*) Maximum likelihood phylogenetic tree of the 47 *Escherichia coli* strains. (*B*) Maximum likelihood phylogenetic tree of the 20 *Salmonella enterica* strains. *Escherichia fergusonii* and *S. bongori* were used to root the trees of each species. The branch length separating *E. fergusonii* from the *E. coli* strains is not to scale (same for *S. bongori*); the numbers above the branch indicate the respective substitution rates per site. All nodes of the trees were supported with high bootstrap values (>97%), the few exceptions correspond to some terminal branches connecting very closely related strains. Phylogenetic groups of the strains are indicated with colors on the right part of each panel. (*C*) Distribution of the number of prophages per genome. Colors correspond to the phylogenetic groups of panels *A* and *B*.

difficult to distinguish from other mobile elements. This resulted in the main data set of 500 prophages.

Prophages tend to be recently integrated in bacterial genomes and thus strain specific (Canchaya et al. 2003). Nevertheless, our data set includes some very closely related bacterial strains (fig. 1*A* and *B*), and some of the prophages may have arisen from the same integration event in an ancestral genome (henceforth named orthologous prophages). To control for pseudoreplication in the statistical analyses, we identified these prophages using similarity and positional scores (see Materials and Methods). This nonredundant data set (NRall) includes 418 prophages that have similarity scores lower than 90%. We also created an even smaller data set including 301 prophages in NRall that are larger than 30 kb (NRlong). These prophages are nonredundant and less affected by accumulation of mutations and pseudogenization events. By default, we present the statistics obtained using the main data set. Other data sets are mentioned only when relevant, for example, when leading to different conclusions. Comparison of the size of the main and the NRall data set suggests that most prophages are not orthologous.

The number of prophages in genomes is highly variable regardless of their phylogenetic group (fig. 1*C*). It ranges from 2 to 20 in *Escherichia* (up to 13.5% of the genome of O157:H7 str. EC4115) and from 1 to 8 in *Salmonella* (up to 4.9% of the genome of Newport str. SL254) (see supplementary table S1, Supplementary Material online). On average, *Escherichia* genomes have more prophage genes than *Salmonella*'s (5.6% vs. 3.5%; Student's test, $P < 0.0005$). Independent of this effect, larger genomes have more prophages (fig. 2*A*; Spearman's $\rho = 0.52$, $P < 0.0001$). To investigate how prophages contribute to the diversity of the repertoire of gene families in both *E. coli* and *S. enterica*, we computed the pan genomes of these species (see Materials and Methods). In both species, we found approximately 3,000 genes present in more than 90% of the strains (persistent genes), although the fraction of core genes (present in 100% of the strains) is smaller in *E. coli* (1,983 genes vs. 2,628 in *S. enterica*) (fig. 2*B*). The accessory genome, consisting of the genes present in less than 90% of the strains, is much larger in *E. coli* (~18,100 genes) than in *S. enterica* (~6,800 genes). Importantly, *E. coli* pan genomes remain larger when analyzing the same number of genomes of the two species (fig. 2*C*). The larger *E. coli* accessory genome is consistent with the high abundance of prophages in this species. Indeed, prophages account for 41% and 31% of the accessory genes in *E. coli* and *S. enterica*, respectively. A total of 75% of prophage genes are present in less than two strains in *E. coli* (80% in *S. enterica*), suggesting that upon acquisition, they tend to be rapidly lost, contributing to the open pan genome of these two species (fig. 2*C*). Prophages are important contributors to genome plasticity (Ohnishi et al. 2001; Banks et al. 2002; Casjens 2003; Canchaya et al. 2004). In these clades, they account for a large fraction of the accessory genome determining variations in genome size.

The Diversity of Prophages

We made sequence similarity analyses between the proteomes of all phages of enterobacteria and all detected prophages of *Escherichia* and *Salmonella*. With these results, we built phage classification schemes based on trees and on graphs (see Materials and Methods). In the following, we use the tree representation because it is easier to compare with classical protein phylogenies and does not involve the choice of clustering parameters. Prophages were classified by comparing their position in the cladogram with those of a set of 147 phages and 50 prophages classified in GenBank or in the literature (Casjens 2003) (see Materials and Methods and supplementary fig. S1, Supplementary Material online). Six different features were thus attributed to each prophage, when possible: 1) the nucleic acid type (double stranded DNA [dsDNA] or ssDNA), 2) the life style (temperate or virulent), 3) the type lambdoid or nonlambdoid, 4) the order, 5) the viral family (based on the particle structure), and 6) the genus (see supplementary table S2, Supplementary Material online). The nucleic acid type and the life style were confidently

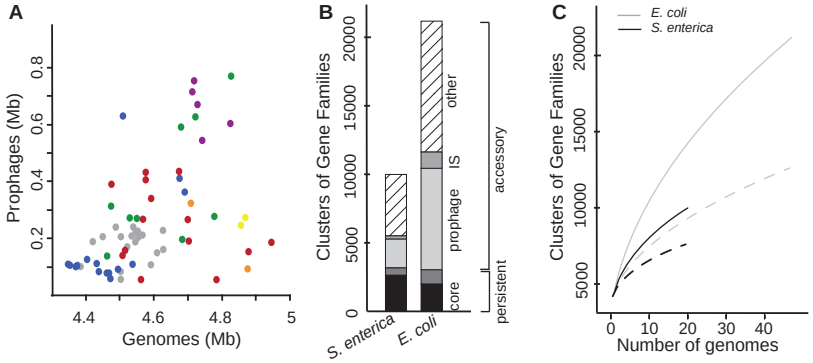

Fig. 2. Contribution of prophages to chromosome plasticity. (*A*) Scatter plot of cumulative size of resident prophages against the size of the host genome (Spearman's $\rho = 0.52$, $P < 0.0001$). Colors correspond to the phylogenetic groups as in figure 1. (*B*) Fraction of the core, persistent, and accessory genes in the pan genome of *Salmonella enterica* (left) and *Escherichia coli* (right). The core genome corresponds to the genes present in all strains, the persistent genome to the genes present in more than 90% of the strains. The accessory genome is split in three categories: the prophages, the insertion, sequences (IS), and the other genes. (*C*) *Escherichia coli* (in gray) and *S. enterica* (in black) pan genomes according to the number of sequenced genomes. The dotted lines correspond to pan genomes after removing prophage elements.

determined for all the prophages. The taxonomic order, a family, and a genus were attributed to 75% of the prophages (supplementary table S2, Supplementary Material online). The remaining 25% prophages are on average much smaller (median size of 19 kb vs. 40 kb for classed prophages, $P < 0.0001$, Wilcoxon test). Almost one third of unclassified prophages lack an integrase (vs. 12% in the NRlong data set, see later). These traits suggest that many unclassified elements are prophage relics, which might justify their unreliable classification. Some of the few large unclassified prophages may be previously nondescribed classes or chimeras. Indeed, the Stx-like group of prophages is related to both Lambda-like (*Siphoviridae*) and P22-like (*Podoviridae*) phages (Garcia-Aljaro et al. 2009) and was classed apart from both. A second group of prophages was classed independently of the genera defined by the ICTV: the "SfV-like" phages. Such elements display unique features as they are lambdoid and have a *Myoviridae* tail structure (Allison et al. 2002; Mmolawa et al. 2003). Importantly, our method of classification can be sensitive to the inclusion of small genomes in the data set (Wolf et al. 2002; Snel et al. 2005). To test the robustness of the classification tree, we applied the same procedure to the 301 NRlong prophages. We found identical classifications for 90% of the prophages. Hence, small phage genomes may affect the topology of the cladogram but do not introduce major changes in the classification. In the following analyses, we use the classification based on the entire data set as this allows classing all prophages.

Temperate and virulent phages form clearly distinct clades in our classification. Accordingly, no single prophage was positioned among virulent phages in the tree (fig. 3*A*). The majority of prophages are from the *Myoviridae*, *Siphoviridae*, and *Podoviridae* families (126, 223 and 30 prophages, respectively), with only three occurrences of *Inoviridae*. Two thirds of the prophages are lambdoid. *Escherichia coli* and *S. enterica* have significantly different distributions of phage genera ($P < 0.0001$, χ^2 test), with the latter lacking Inoviruses, Epsilon15-like, Mu-like, and phiC31-like prophages. However, a wide diversity of viruses, including filamentous phages, were previously observed in *Salmonella* (Ackermann 2007), suggesting that a larger sampling will partially correct for this effect. The most noticeable difference between the species is the very high fraction of Lambda-like prophages in *E. coli* (50%) relative to *S. enterica* (23%) ($P < 10^{-6}$, χ^2 test). Interestingly, within a few groups (Lambda, SfV, P22, and P2), the phages of *E. coli* and *S. enterica* are well separated in the classification (fig. 3*B*). This suggests that host switching happens rarely and/or that it is accompanied with rapid evolution of specific gene repertoires.

Integration Hotspots

Comparative analyses of prophage locations are complicated by the high plasticity of the genomes of *Escherichia* and *Salmonella* (Vernikos et al. 2007; Touchon et al. 2009). To facilitate this analysis, we localized prophages relative to the closest flanking core genes. *Escherichia coli* and *S. enterica* genomes are mostly collinear (see supplementary table S1, Supplementary Material online), and only 4% of prophages are within a rearrangement breakpoint region. These few elements were removed from the analysis of integration loci. The remaining 369 *E. coli* prophages were found in 58 distinct integrative loci and the 102 *S. enterica* prophages in 24 distinct integrative loci (fig. 4). Loci are shared by an average of 6.4 and 4.2 prophages within *E. coli* and within *S. enterica* genomes, respectively. Importantly, similar trends are found with the NRlong data set (5.4 and 3 in *E. coli* and *S. enterica*, respectively). We simulated 1,000 times the expected number of integration locations if they took place at random. In this case, one would expect to find 336.2

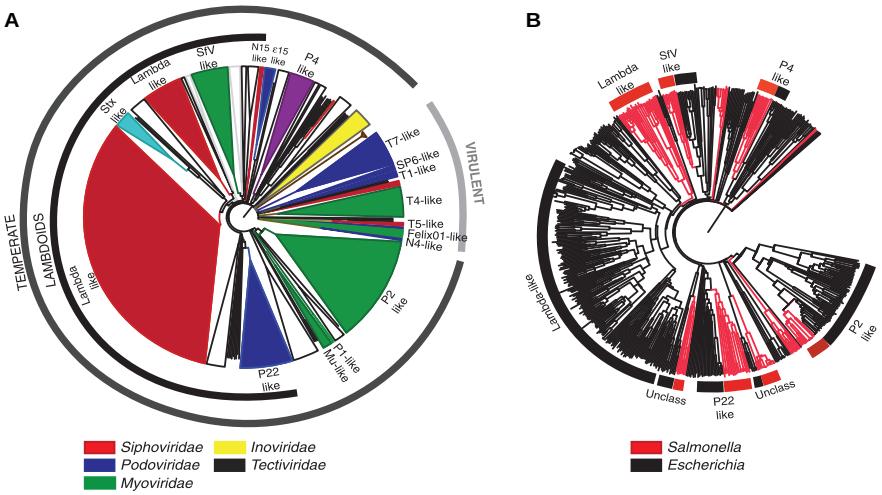

FIG. 3. Classification of prophages. (*A*) Phylogenetic tree of phages and prophages based on gene repertoire relatedness (see Materials and Methods). Phage/prophage families are colored according to the color key. The phage/prophage genus is indicated in the inner circle. The members of the "lambdoid" group are indicated in the second circle. The classification of phages/prophages into temperate and virulent is indicated in the third circle. White clusters correspond to unclassified clades. (*B*) Phylogenetic tree as in (*A*) but restricted to temperate phages/prophages. Red branches correspond to *Salmonella* phages/prophages and black branches to *Escherichia* phages/prophages. Labels indicate some types of phages/prophages of interest and mentioned in the text.

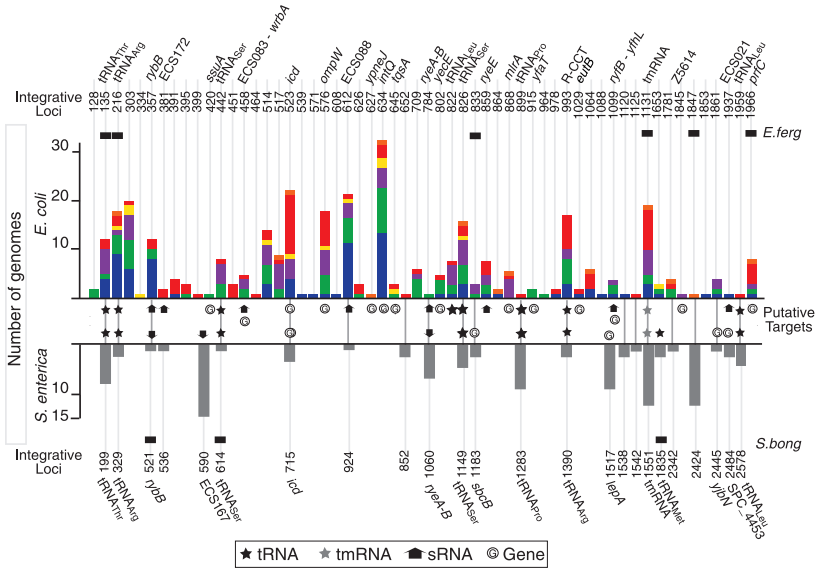

FIG. 4. Distribution of prophages at integration hotspots. The *x* axis indicates the position of the hotspots of phage integration in the genomes of *Escherichia coli* (top) and *Salmonella enterica* (bottom). The positions of the "integrative loci" (on top for *E. coli* and bottom for *S. enterica*) are indicated as positions in the core genome. For example, position 634 in *E. coli* refers to prophages integrated 3′ of the 634th core gene in the reference genome of *E. coli* (MG1655 see Materials and Methods). The bars indicate the number of genomes with at least one prophage integrated among *E. coli* (top) and *S. enterica* (bottom). Colors in the bars correspond to the phylogenetic group of the genomes as in figure 1. The presence of prophages in *E. fergusonii* and in *S. bongori* is represented by a black rectangle above (respectively below) the bars of *E. coli* (respectively *S. enterica*). The 19 integrative loci conserved between *E. coli* and *S. enterica* genomes are connected in the middle of the figure. "Putative targets" of integration are also indicated in the middle part of the figure (details in the keys). The identification of tRNA (amino acid), sRNA, and protein coding genes are reported at the top and the bottom of the graphs, next to the indication of the flanking core gene (details in supplementary table S3, Supplementary Material online).

(95% interval of confidence [CI]: ±0.3) distinct loci in *E. coli* (1.1 prophage per locus) and 99.8 (95% CI: ±0.1) in *S. enterica* (1 prophage per locus). Hence, prophages have significant integration hotspots in the genomes. A total of 19 of the 24 integrative loci of *S. enterica* (80%) are also integration loci in *E. coli* (fig. 4). Hence, the turnover of prophages is very high but restricted to a few sites in the bacterial chromosome that are often conserved for many millions of generations.

Hotspots flanking tRNA or tmRNA genes have often been described and could result from integrases targeting conserved palindromic sequences (Williams 2002). However, these genes flank only 15% of *E. coli* and 37% of *S. enterica* integration sites (fig. 4 and supplementary table S3, Supplementary Material online) and only 8 of the 19 conserved hotspots between the two species. The tRNA gene pool is highly variable in these two species (Withers et al. 2006), but the tRNA genes flanking these integration loci are present in a single copy in all strains of *E. coli* and *S. enterica*. These tRNAs are not a random sample of the tRNAs of *E. coli* and *S. enterica*: They are present in all genomes in one single copy and they decode the least used anticodon of 4- or 6-codon amino acids (supplementary table S4, Supplementary Material online). This might represent selection for elements that are lowly expressed (the case of rarely used tRNAs [Dong et al. 1996]), highly conserved in genomes (core genes), and present in unique positions (allowing coevolution between the temperate phage and the host).

Many recently identified small RNA (sRNA) genes also include palindromes forming hairpins (Waters and Storz 2009). Hence, we analyzed the colocalization of prophages with 441 sRNAs identified in recent large-scale studies of *Escherichia* and *Salmonella* (Huang et al. 2009; Raghavan et al. 2011; Shinhara et al. 2011; Kroger et al. 2012) (see Materials and Methods). A total of 11 (19%) and 4 (17%) additional integration sites (after removing the overlap with tRNA genes) are located close (<1 kb) to conserved sRNA genes in *E. coli* and *S. enterica*, respectively (fig. 4 and supplementary table S3, Supplementary Material online). No further sRNAs were identified when the detection window was extended to 5 kb. We found that eight sRNA genes form stable secondary structures (i.e., more stable than 90% of random sequences with same size and composition, see Materials and Methods). Two of these genes (*ryeB* in *Salmonella* and *ryeE* in *E. coli*) were previously known to be targeted by phages (Wassarman et al. 2001; Balbontin et al. 2008). Therefore, sRNAs might also be important integration sites.

We investigated the specific features of the 64% (*E. coli*) and 46% (*S. enterica*) of integration loci that are not associated with tRNAs, tmRNAs, ribosomal RNAs (rRNAs), or sRNAs (henceforth named noncoding RNA [ncRNAs]). Integration into protein coding sequences has been described within *icd* (Wang et al. 1997) and *ompW* in *E. coli* (Creuzburg et al. 2011) and *lepA* in *S. enterica* (Hermans et al. 2006). Indeed, we find these three loci among the most occupied hotspots (fig. 4). Integration leads to duplication of the 3′-end without affecting the length of the ORF in the first case, whereas the gene is disrupted in the second case (supplementary table S3, Supplementary Material online). We identified 15 additional protein encoding genes disrupted due to phage integration (*ssuA*, *ynel*, *wrbA*, *intQ*, *tqsA*, *intR*, *mlrA yecE*, *yfaT eutB*, *yfhL*, *prfC*, *yjbN*, SPC_4453, and *Z5614*) (fig. 4 and supplementary table S3, Supplementary Material online). The *intQ* and *intR* genes encode integrases and might correspond to pseudogenes of previous prophages. Surprisingly, the other genes are well conserved within *E. coli*, and eight of them (*ssuA*, *wrbA*, *prfC*, *yecE*, *ynel*, *tqsA*, *yfaT*, and *eutB*) would be part of the *E. coli* core genome if they had not been disrupted by phage integration. These cases correspond to sites less frequently occupied by prophages (3.5 prophages per site on average). Two of them were disrupted by Mu-like prophages that integrate randomly in the host genome (Bukhari and Metlay 1973). Thus, some protein encoding genes are hotspots even though this leads to their disruption. However, most of these integration loci are poorly populated suggesting that these are secondary integration sites.

Strikingly, 50% of *E. coli* and 25% of *S. enterica* integrative loci are neither next to ncRNA genes nor within protein coding genes. Many of these loci have few or even one single prophage and may represent secondary integration sites. However, five of these loci are occupied at higher frequencies than the average loci (11.8 prophages, $P < 0.02$, Wilcoxon test). This is the case of the integration site of phage Lambda (Otsuka et al. 1988). Contrary to ncRNA genes, intergenic regions are under few constraints, and integration sites in these regions are expected to evolve fast. Nevertheless, we observe four such hotspots shared by *E. coli* and *S. enterica* (i.e., 21% of all conserved loci). Conservation of intergenic sequences at such large evolutionary distances requires strong purifying selection. This may result from selection for lysogeny, which is adaptive for the host, and for constancy of integration sites, which favors coevolution of phage and bacterial genome structures.

Tropism of Phage Integration

We also studied the tropism of phage integration from the point of view of the phage. In *E. coli*, Inovirus, Epsilon15-like, and phiC31-like phages integrate each at one single site (supplementary table S5, Supplementary Material online). Stx-like, P4-like, P22-like, and SfV-like phages integrate at a small number of different sites (2, 3, 5, and 5 sites, respectively). On the other hand, P2-like and lambda-like phages integrate into many sites (13 and 21 sites, respectively). Expectedly, we found Mu-like phages integrated randomly in the chromosome. Integration loci tend to be genus specific because few sites (8/4 in *E. coli*/*S. enterica*) include more than one phage genus. Of these, two sites show an extreme prophage diversity including almost all genera of prophages and even other mobile genetic elements such as integrative conjugative elements and pathogenicity islands (i.e., sites flanking $tRNA_{Thr}$ and tmRNA, supplementary fig. S2, Supplementary Material online). We found no obvious association between phage genus and target type (i.e., tRNA, tmRNA, sRNA, or protein

coding gene) (supplementary table S5, Supplementary Material online). We found 15 integrative sites containing only unclassified prophages in *E. coli* (4 in *S. enterica*) (supplementary table S5, Supplementary Material online), which typically correspond to small elements ongoing genetic degradation. This suggests that some integration sites provide a more favorable genetic background than others.

We then tested whether integration tropisms were associated with the phylogeny of the phage integrases. We found that 413 of the 500 prophages (83%) contained an integrase, all tyrosine recombinases. This percentage rose to 89% among NRlong prophages. Phages lacking integrases may have lost them after integration or use other means to integrate. Accordingly, Mu-like prophages and Inoviruses lacked such integrases (1% of the NRlong prophages). We constructed a phylogenetic tree of the integrases to associate integrase similarity with integration tropism. The deeper nodes of the tree are poorly supported limiting the conclusions that can be taken from ancient evolutionary events (fig. 5). The more recent nodes show clusters of phages of the same genus. This includes P2-like, P4-like, and Epsilon15-like prophages. Lambdoid prophages are intermingled in the tree as expected because they showed no commonalities in terms of integration sites. Importantly, integrases from elements integrated at the same locus form terminal clades in the tree, that is, closely related integrases tend to integrate at the same sites. The few apparent exceptions were all examined in detail and concern loci with multiple close integrations where one element is correctly grouped in the tree and the other is inserted in a nearby sequence and clusters elsewhere in the tree (supplementary fig. S3, Supplementary Material online).

Distribution of Prophages in the Chromosome

The propensity for integration by site-specific recombination varies with genomic regions in *S. enterica* (Garcia-Russell et al. 2004). Unfortunately, there are no data available on the large-scale structure of the chromosome of *Salmonella*. In *E. coli*, the chromosome is structured in domains and macrodomains that are associated with specific local properties, such as DNA compactedness (Wiggins et al. 2010). This might affect patterns of prophage integration or excision.

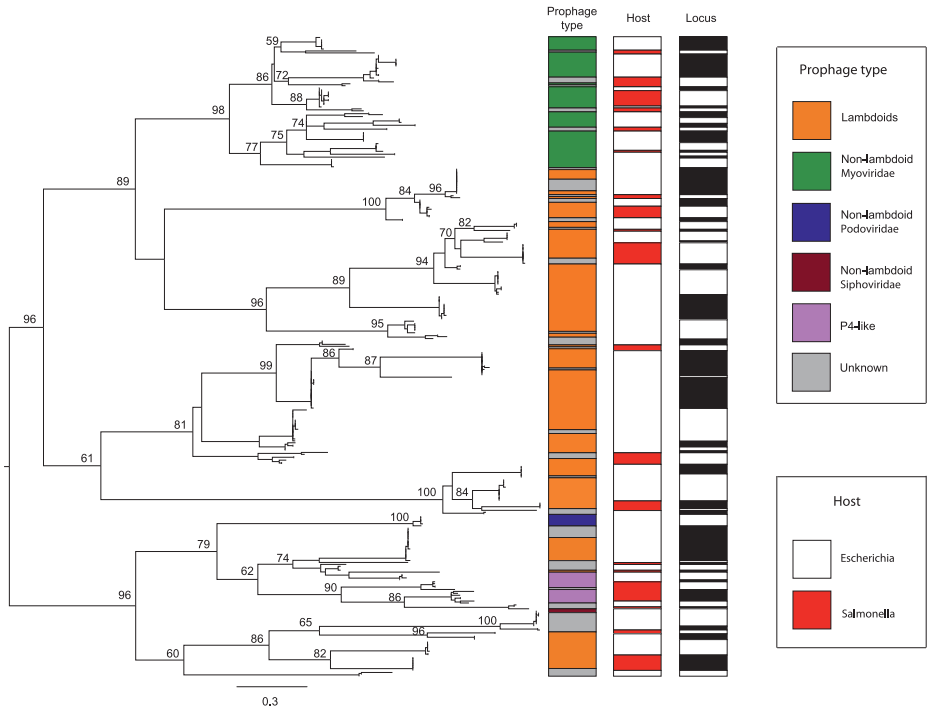

Fig. 5. Phylogeny of the integrases. The maximum likelihood tree was made from a trimmed alignment of 332 tyrosine recombinases and rooted using the midpoint root. Bootstrap values (out of 1,000 replicates) are given in percents in the tree and are shown when exceeding 50%. Prophage types are indicated in the first column. The species hosting the prophage is shown in the second column. The third column shows that blocks of closely related integrases correspond to phages integrated at the same loci. One given block puts together a given number of integrases that are together in the phylogenetic tree and are associated with a single locus.

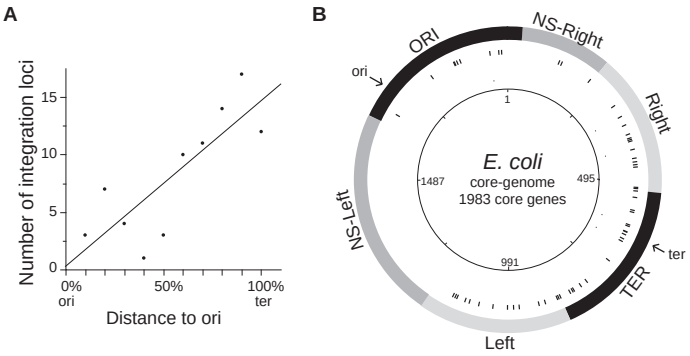

Fig. 6. Distribution of prophages in the chromosome. (*A*) Number of loci with prophages in function of the distance to the origin of replication. Distribution of integration loci in function of the distance to the origin of replication (ori: origin and ter: terminus). (*B*) Circular representation of the distribution of the prophages in function of the macrodomains of *Escherichia coli*. Circles represent the following (from the inside out): 1, position in the core genome; 2, location of the integration locus; and 3, location of the four macrodomains and the two nonstructured (NS-right and NS-left) domains of the *E. coli* chromosome.

Accordingly, prophages and their integration loci are not randomly distributed among the four macrodomains and the two nonstructured (NS-left and NS-right) domains of the *E. coli* chromosome (both $P < 0.0005$, χ^2 test). The latter have the lowest number of prophage loci (3 in NS-right and 0 in NS-left), followed by the origin of replication (Ori) macrodomain (nine loci) (fig. 6*B*). Contrary to the four macrodomains, NS regions show high intracellular mobility and interact with their surrounding domains (Valens et al. 2004). This should not disfavor integration events and indeed we find that the frequency of transposases in this region is not significantly different from the rest of the genome ($P = 0.77$, χ^2 test). Furthermore, NS regions integrate some well-known pathogenicity islands encoding tyrosine recombinases (Napolitano et al. 2011), for example, PAI-LEE, PAI-I_{CFT073}, and PAI-III_{EDL933} (Blum et al. 1994; McDaniel et al. 1995; Dobrindt et al. 2002). Core genes in these regions have sequence compositions similar to the rest of the genome (51% in GC content, $P = 0.2$, Wilcoxon test) suggesting this is not the cause of a putative integration bias. The frequency of tRNA or sRNA genes in these regions is also not different from expected ($P > 0.05$, χ^2 test). Essential genes are 50% more abundant than expected in NS regions ($P < 10^{-7}$, χ^2 test), but their density (10% vs. 6% in the entire genome) seems too low to lead to a general avoidance of prophages in these regions because of the over-representation of genes for which inactivation is lethal. Interestingly, the average Codon Adaptation Index of genes in the NS regions and the Ori macrodomain is higher than in the rest of the genome (0.414 vs. 0.396, $P < 10^{-6}$, Wilcoxon test). High expression of neighboring genes might render prophages less stable. On the other hand, macrodomains are located in different regions of the cell. Notably, the NS-right region is closer to the cell center, followed by the Ori, the Right, and the terminus of replication (Ter) macrodomain that is the closest to the cell poles (NS-left and left were not tested) (Meile et al. 2011). This might render the Ter and the nearby macrodomains more susceptible to integration by phages, especially because phage infection might preferentially take place at cell poles (Edgar et al. 2008; Guerrero-Ferreira et al. 2011).

The frequency of prophages (and integration loci) increases with the distance to the origin of replication both in *Escherichia* and *Salmonella* (fig. 6*A*, respectively, Spearman's $\rho = 0.79$, $P < 0.006$ and $\rho = 0.82$, $P < 0.005$). The frequency of ncRNA genes is not higher in this region ($P > 0.6$, χ^2 test) and cannot justify the observed pattern. We then tested whether macrodomain structure was sufficient to explain these patterns. For this, we analyzed the abundance of prophages within each macrodomain. We divided each macrodomain in equally sized terminus-proximal and terminus-distal regions. The intra-macrodomain regions nearer the terminus have 24% more prophages and 24% more integration loci than the intra-macrodomain regions nearer the origin of replication (respectively, $P < 10^{-6}$ and $P = 0.055$, χ^2 tests). Hence, prophages are more abundant in certain macrodomains, and within the macrodomains, they are more abundant in regions closer to the terminus of replication.

Prophage Polarization

The genes of lambdoid prophages show a preference for co-orientation with the bacterial replication fork, and this is not explained by their tropism toward some tRNAs (Campbell 2002). Indeed, we found no loci specificity toward lambdoid phages after accounting for phage genus. Bacterial genes, and especially essential genes, are also predominantly co-orientated with the replication fork, presumably to minimize effects of the collisions between the replication fork and the RNA polymerase (Rocha and Danchin 2003). To study these patterns, we defined prophage transcription polarity as the fraction of the prophage coding sequences in the most gene-rich strand of the prophage. We analyzed two subsets: the lambdoids (330 prophages) and the

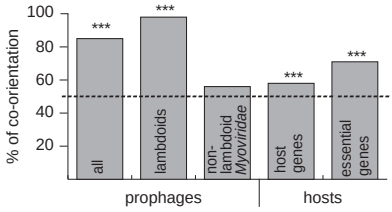

FIG. 7. Percentage of prophages and host genes co-oriented with the replication fork. The dotted line shows the polarization under random expectation (50%). $P < 0.05$ (*); $P < 0.01$ (**); $P < 0.001$ (***).

nonlambdoid *Myoviridae* (104 prophages), which are the largest clade of the remaining prophages. Together these groups make 87% of our data set. Most prophages were highly polarized with an average of 77% of the coding nucleotides in the most gene-rich strand (76% in lambdoids and 79% in nonlambdoid *Myoviridae*).

The co-orientation of a large fraction of the prophage genome does not necessarily entail co-orientation of prophage genes with the bacterial replication fork. We defined prophage replication polarization as the predominant orientation of genes relative to the direction of the bacterial replication fork. We found that 85% of prophages are predominantly co-oriented with the bacterial replication fork ($P < 10^{-15}$ in the three data sets: all, NRall, and NRlong, χ^2 test) (fig. 7). The effect is much stronger in lambdoid prophages (98% of prophages, $P < 10^{-15}$, χ^2 test) than for the average host gene (~57% both in *E. coli* and *S. enterica*) and for the *E. coli* essential genes (71%). Replication polarization of nonlambdoid *Myoviridae* is not significant (56%, $P > 0.05$, χ^2 test). Hence, replication polarity, contrary to transcription polarity, is specific to lambdoids. Interestingly, among lambdoid phages, the smaller and presumably more degraded prophages are less often co-oriented with the replication fork than the NRlong prophages (88% vs. 100%, $P < 10^{-6}$, χ^2 test). Lambdoid prophages might thus degrade faster when antioriented with the replication fork.

If the replication polarity of lambdoids is caused by collisions between the bacterial RNA polymerases and replication forks, as proposed for bacteria, then the transcription of genes expressed in the prophage should be preferentially co-oriented with the replication fork. Most genes are silent in the prophage state, with the notable exception of the repressor of the lytic cycle. We thus identified a total of 115 *cI* repressors of the lytic cycle among the 330 lambdoid prophages (see Materials and Methods). A majority of these (90%) were found antioriented with the replication fork. This result is in stark contradiction with the hypothesis that collisions between RNA polymerase and the replication fork cause co-orientation of prophage genes with the bacterial replication fork. Inversion of Lambda prophages in *E. coli* lacks strong phenotypes in terms of bacterial growth or genetic instability (Campbell 2002). This suggests that prophage polarization does not have a strong impact on the cell's physiology. Co-orientation of lambdoids with the replication fork might thus be associated with their particular genetic organization and how it accommodates in the bacterial chromosome, for example, in terms of DNA motifs (see later). Alternatively, this might be due to some association between the mechanism of phage integration and the bacterial replication fork. This association was found in several DDE recombinases (Peters and Craig 2001) but to the best of our knowledge not in integrases using tyrosine recombinase activity.

Distribution of DNA Motifs in Prophages

The genomes of *Escherichia* and *Salmonella* are packed with signals that regulate cellular processes affecting the chromosome at large scales such as macrodomain formation (*matS*) and chromosome segregation (FtsK-orienting polar sequences [KOPS]) (Touzain et al. 2011). The MatP protein interacts with the 13 bp *matS* sites to organize the terminus of replication of the chromosome into the Ter macrodomain (Mercier et al. 2008). The motif *matS* is thus concentrated in the Ter macrodomain and absent from the rest of the chromosome. We found no single *matS* motif in any of the prophages. This is statistically unexpected given the motif size and composition (see Materials and Methods, $P < 0.004$, χ^2 test). The absence of *matS* in the prophages of the Ter macrodomain is not statistically significant but might simply result from the lack of statistical power ($P = 0.1$, χ^2 test). Indeed, prophages of the Ter macrodomain of *E. coli* display a strong underrepresentation of *matS* motifs when compared with the host Ter macrodomain ($P < 10^{-15}$, χ^2 test). The density of *matS* in the Ter macrodomain of *E. coli* K12 MG1655 is low (1 every 49 kb). The average size of the NRlong prophages is 44 kb. Therefore, integration of a prophage lacking *matS* probably has no disruptive effect in the formation of the macrodomain. However, this does not explain the significant avoidance of *matS* in prophages. The *matS* motif defines the Ter macrodomain and is absent from the rest of the chromosome (Mercier et al. 2008). Avoidance of *matS* in prophages outside the Ter macrodomain might be caused by its potential disruptive effect. Phages recombine frequently to produce mosaic structures. Hence, lack of *matS* in phages integrating at the Ter macrodomain could increase the probability of producing viable recombinant genomes with phages integrating at other chromosomal sites. These results suggest that motifs can be strongly counter selected in prophages when they disrupt chromosomal structure.

KOPS motifs are octamers that orient the transport of DNA by FtsK at the last stages of chromosome segregation (Bigot et al. 2005; Levy et al. 2005). KOPS are more frequent in the ter-proximal regions and in co-orientation with the replication fork (Bigot et al. 2005). KOPS are more abundant than expected in the chromosome (9.6×10^{-5} KOPS/nt) and in lambdoid prophages (9.5×10^{-5} KOPS/nt, both $P < 0.01$, χ^2 test). They are also strongly co-oriented with the replication fork (respectively, 90% and 86%). We observed lower density of KOPS in nonlambdoid *Myoviridae* prophages (5.1×10^{-5} KOPS/nt, $P > 0.1$, χ^2 test) and even lower in virulent phages (3.1×10^{-5} KOPS/nt, $P > 0.7$, χ^2 test).

Interestingly, the density of KOPS in lambdoids mirrors the trends of the rest of the genome: KOPS density is lower in prophages in the Ori-proximal half of the chromosome than in Ter-proximal half (7.2×10^{-5} vs. 1.0×10^{-4} KOPS/nt, $P < 10^{-5}$, χ^2 test). Furthermore, the density of KOPS in the Ter-proximal half and in its lambdoid prophages is very similar (9.6×10^{-5} vs. 1.0×10^{-4} KOPS/nt, $P > 0.4$, χ^2 test). This suggests selection for the over-representation of polarized KOPS in lambdoids to match the chromosomal organization.

Conclusion

Our study shows that phages integrate in ways that minimize their negative effects on the chromosome organization. This coevolution of phages and bacteria involves selection for integration sites, gene order, and DNA motifs that affect the biology of the bacterial chromosome. Phage integration is restricted to a few sites that are conserved over very long evolutionary periods. Targeting slow evolving sequences (especially RNA genes) is adaptive for phages. However, many prophages integrate at sites in intergenic regions that are conserved between *E. coli* and *S. enterica*. This suggests selection for the conservation of integration sites as a means of promoting lysogeny over lysis and facilitating long-term coevolution of temperate phages and bacteria. Prophage organization is also important at the chromosome scale because prophage density increases along the replichores and differs markedly among macrodomains. This might result from integration biases caused by different accessibility of chromosomal regions to prophages. It might also result from selection for regions of low gene expression. Accordingly, phage abundance increases along the ori–ter axis. The expression of the tmRNA gene, an important integration hotspot, is important for the function of the neighboring P22-like phages and pathogenicity islands (Julio et al. 2000). This suggests that integration sites might provide other functions besides a site-specific recombination point, for example, regulation of gene expression. Accordingly, we find that prophages avoid integration in the most expressed tRNA genes and in the chromosomal regions with the highest fraction of highly expressed genes. This suggests that they avoid proximity to regions highly transcribed. Transcriptional spillover from nearby genes could lead to expression of phage genes and destabilization of the lysogen. Importantly, temperate phages show avoidance and over-representation of DNA motifs that are relevant only at the prophage state in the context of the biology of the host. This adds a constraint to the evolution of temperate phages that is absent from virulent phages. Learning the way prophages minimize their impact on genome organization might provide key information on how to modify genomes with minimal impact on bacterial fitness.

Materials and Methods

Data

A data set of 69 complete genomes of *Escherichia* and *Salmonella* was downloaded from the NCBI RefSeq (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/, last accessed January 2012). It consists of 20 *S. enterica*, 1 *S. bongori*, 47 *E. coli*, and 1 *E. fergusonii* genomes. A total of 299 complete genomes of phages infecting enterobacterial hosts were also downloaded from the NCBI RefSeq (ftp://ftp.ncbi.nih.gov/genomes/Viruses/, last accessed December 2011).

Identification of Prophages

Prophages were detected using Phage Finder (Fouts 2006), PHAST (Zhou et al. 2011), and Prophinder (Lima-Mendez et al. 2008a). These three phage-finding programs combine sequence comparisons to known phage or prophage genes, comparisons to known bacterial genes, tRNA genes, dinucleotide analysis, and identification of integration sites. Phages infecting the enteric bacteria *E. coli* and *S. enterica* are the most intensively studied, many sequences are available, and it is therefore less probable to miss prophages due to a gap of knowledge on phages for these genera. We removed small prophages (<10 kb) and elements with a large number of insertion sequences (IS; >25% of the predicted ORFs). IS elements were detected as in Touchon and Rocha (2007). Prophage borders and the few prophages coded in tandem were manually validated using different types of information: gene annotation, PFAM protein functions, and core/pan genome definition in bacterial genomes (see later). Prophage genes integrate together and are thus expected to share similar patterns of presence/absence in bacterial genomes. The frequency of gene families in pan genomes (see later) follows a U-distribution, where most families are present in either very few or many genomes (Touchon et al. 2009). Families of genes in prophages, because they tend to be strain specific, are among the low-frequency genes. On the other hand, genes involved in the core functions of the bacterial cell tend to be among high-frequency genes. Hence, when a bordering gene corresponds to a persistent gene (present in at least 90% of strains), it was removed of the predicted prophage. A blastp (with an *e* value < 0.001) of the detected prophages was performed against the rest of the bacterial hosts to check for the presence of further undetected elements. Any cluster of 10 or more genes (with a maximal distance of 3 kb between two consecutive genes) was further inspected. Because of their small sizes (typically <10 kb), Inoviruses were detected using a dedicated procedure. They were searched by similarity to known phages by blastp (with an *e* value < 0.001). When at least four proteins of the reference genomes (GenBank IDs NC_001332, NC_001954, NC_002014 and NC_003287) were detected in a 10 kb window, the putative prophage was checked with GenBank annotations and its borders were manually confirmed as described earlier.

Classification of Phages

Prophages were classed by comparison to previously classed phages by building a common gene content matrix. First, homologous proteins were identified as unique reciprocal best hits with >40% similarity in amino acid sequence and <20% of difference in protein length as in Touchon et al.

(2009). The similarity score was determined with the BLOSUM60 matrix and the Needleman–Wunsch end gap free alignment algorithm. We measured gene repertoire relatedness between pairs of (pro)phages as: $\sum_{i=1}^{M} \frac{S_{(Ai,Bi)}}{\min(n_A,n_B)}$ with $S_{(Ai,Bi)}$ the similarity score of the pair i of homologous proteins shared by (pro)phage A and (pro)phage B (varying from 0.4 to 1), M the total number of homologs between (pro)phages A and B and n_A and n_B the total number of proteins of (pro)phage A and B, respectively.

The gene repertoire relatedness matrix between all pairs of phage/prophages was used to calculate a tree using BioNJ (Gascuel 1997). We then classed phages/prophages using the gene repertoire similarity tree. Prophages were classed according to the phages/prophages with known classification with which they branched together (forming a monophyletic subtree with the classified (pro)phages branching basally, see supplementary fig. S1, Supplementary Material online). For many prophages, we consistently inferred different features: 1) the nucleic acid type: dsDNA/ssDNA/ssRNA; 2) the ICTV taxonomic order: *Caudovirales*/non-*Caudovirales*; 3) the life style: temperate/virulent; 4) the type: lambdoid/non-lambdoid; 5) the ICTV family: *Siphoviridae*/*Podoviridae*/*Myoviridae*; and 6) the ICTV genus: Lambda-like/P22-like/P2-like/Epsilon15-like/PhiC31-like/Mu-like/P4-like/Inovirus. Temperate/virulent life styles and the lambdoid membership could be determined from literature data for most phages of the databank. In addition to the genera defined by the ICTV, two supplementary groups were considered as a genus due to their unique features: "SfV-like" phages that can be defined as lambdoid *Myoviridae* (Allison et al. 2002; Mmolawa et al. 2003) and considered as an independent *Myoviridae* group, albeit not officially elevated to the rank of genus (Lavigne et al. 2009) and "Stx-like" phages as they constitute a group of very closely related lambdoid phages carrying the Stx toxin and displaying *Siphoviridae*/*Podoviridae* hybrid structures (Garcia-Aljaro et al. 2009). The identification of P4 prophages is more complicated because these satellite phages lack structural genes, and there is only one reference sequence in GenBank. P4 encodes one characteristic protein, Sid, which is responsible for its parasitic behavior. Sid functions as a head size determination of phage P2, preventing P2 to integrate its genome within its own capsids (Dearborn et al. 2012). Prophages were classed as P4-like when branching next to P4 (GenBank ID NC_001609) in the tree and if they contained the Sid protein (blastp e value < 0.001). Sid is a good marker of P4-like phages because it was not found in prophages distant from P4 in the tree.

Identification of Core and Pan Genomes

A preliminary set of orthologs was defined by identifying unique pairwise reciprocal best hits, with at least 60% similarity in amino acid sequence and less than 20% of difference in protein length. The list was then refined using information on the distribution of similarity of these putative orthologs and data on gene order conservation (as in Touchon et al. [2009]). The analysis of orthology was made for every pair of genomes of each clade (*E. coli* and *S. enterica*). The core genome consists of genes found in all strains of a clade and was defined as the intersection of pairwise lists of positional orthologs.

Definition of Integration Loci

The *E. coli* and *S. enterica* core genomes were used to define the integration loci of the detected prophages. Each prophage was localized relative to the two closest flanking core genes of the species. By convention, an integration locus was defined by the relative position of the left core gene among the core genome of the species. For example, the locus 135 in *E. coli* corresponds to a prophage located between the 135th and the 136th core genes of the *E. coli* core genome. The relative positions of the loci were defined by the order of the core genes in *E. coli* K12 MG1655 strain and *S. enterica* LT2 strain. These strains were used as references for *E. coli* and *S. enterica* gene orders, respectively, because they represent the most likely configuration of the chromosome in the ancestor of each species. Few rearrangements were observed (respectively, 2.3 and 2 in average for *E. coli* and *S. enterica* genomes) compared with the two reference genomes. Integration loci located between two nonsuccessive core genes, that is, with rearrangements in between them were removed.

Clades Phylogenetic Trees

We extended the species core genomes by adding genomes of the two earliest diverging available species, *E. fergusonii* and *S. bongori*. We made multiple alignments of each family of core proteins using muscle v3.6 (Edgar 2004) with default parameters and back-translated these alignments to DNA. The concatenated alignments of core genes were given to Tree-puzzle 5.2 (Schmidt et al. 2002) to compute the distance matrix between genomes using maximum likelihood under the Hasegawa–Kishino–Yano + G(8) + I model. The tree of the core genome was built from the distance matrix using BioNJ (Gascuel 1997). We made 1,000 bootstrap experiments on the concatenated sequences to assess the robustness of the topology. The topology of these trees is congruent with previous whole-genome phylogenetic analyses of *E. coli* and *S. enterica* (Touchon et al. 2009; Touchon and Rocha 2010). Groups' terminology is based on the latest update of *E. coli* strains classification (Tenaillon et al. 2010).

Identification of Integrases, cI Repressors, and Phylogenetic Analysis

Integrase and cI repressor proteins were searched using PFAM protein profiles for tyrosine recombinase (PF00589), serine recombinase (PF07508 and PF00239), and cI repressor (PF07022) obtained from the PFAM database, version 26.0 (http://pfam.sanger.ac.uk/, last accessed January 2012). Prophages were searched with these profiles using hmmpfam (e value < 0.001, coverage of $>50\%$ of the profile) (Eddy 2011). The multiple alignment of the 413 tyrosine recombinase proteins was made with muscle v3.6 (Edgar 2004). Informative regions were selected using BMGE with the BLOSUM30 matrix (Criscuolo and Gribaldo 2010). Poorly aligned sequences were manually removed from the

alignment. The final alignment of 332 sequences was used to reconstruct the phylogenetic tree using the maximum likelihood method implemented in TREEFINDER (Jobb et al. 2004) under a mixed + G(5) model, which was estimated as the best-fit model with the Akaike information criterion. The tree topology was assessed with 1,000 bootstrap replicates using the same model.

Identification of ncRNA

The tRNA genes were identified using tRNAscan-SE 1.23 (Lowe and Eddy 1997). The tmRNA genes were detected by sequence similarity search using blastn, having at least 90% of identity sequence and less than 20% of difference in sequence length with the original sequence identified in *E. coli* (Lee et al. 1978). A single tmRNA gene was thus identified in each genome of *Escherichia* and *Salmonella*. Other sRNA genes were identified using two recent published data sets from *E. coli* (Raghavan et al. 2011; Shinhara et al. 2011) and one from *Salmonella* (Kroger et al. 2012). The 328 sRNA sequences reported in *E. coli* K12 MG1655 strain and the 113 sRNA sequences identified in *S. enterica* SL1344 strain were then blasted against all genome sequences analyzed in this study. For each sRNA, only the best match within each host genome with at least 80% of identity sequence and length coverage of 50% was considered. We found 326 and 195 sRNAs in *Escherichia* and *Salmonella* genomes, respectively, with 153 nonredundant sRNA genes shared by all *Escherichia* strains, 123 shared by all *Salmonella* genomes, and 73 shared between all *Escherichia* and all *Salmonella* genomes. RNA genes were considered as putative integration targets of a prophage when found at less than 1 kb of prophages borders. A sRNA was not considered as a putative integration target if a core gene of the host was found between the sRNA and the prophage. RNA genes located within a prophage (or a neighboring prophage) were not considered as potential integration targets. sRNA secondary structures were predicted with RNAfold (Gruber et al. 2008). Each sequence was shuffled 1,000 times keeping nucleotide composition constant, and the distribution of minimum free energies was computed with the 1,000 randomized sequences. For each sRNA, the predicted structure was considered as reliable when its minimum free energy was found among the 10% most stable structures of the distribution of minimum energy for the random sequences.

Identification of Targeted CDS

The identification of putative integration targets within protein coding genes was made by searching for homologies between the sequences flanking the prophage and proteins in the pan genome using tblastn (Altschul et al. 1997) (e value < 0.001). We took 1 kb sequences around each prophage limit. When both prophage flanking regions matched the same protein, we aligned them independently to the corresponding gene with needle (Rice et al. 2000) using the end gap free option. Two cases were then considered: 1) phage integration led to the duplication of one end of the CDS and 2) the CDS was disrupted due to phage integration. The first situation was identified when one hit corresponded to the entire CDS and the other hit to a smaller fragment. The second case was recognized when none of the hits corresponded to the entire query CDS and when they were found aligned to complementary parts of the query CDS (i.e., nonoverlapping but converging at the same position).

Identification of Macrodomains, Essential Genes, Origin and Terminus of Replication, KOPS, and *matS* Motifs

Macrodomain borders were delineated as in Scolari et al. (2011). Essential genes were defined as in Baba et al. (2006). We used the sequences patterns reviewed by Touzain et al. (2011) to identify KOPS (GGG[ATGC]AGGG) (Bigot et al. 2005) and *matS* (GTGAC[AG][AGTC][TC]GTCAC) (Mercier et al. 2008) sequences in the 69 *Escherichia* and *Salmonella* complete genomes using Fuzznuc (http://emboss.bioinformatics.nl/cgi-bin/emboss/fuzznuc, last accessed January 3, 2013). To identify the origin of replication, we searched using blastn, the best hit with the known *oriC* sequence of *E. coli* K12 MG1655 of 378 bp in the other genomes. This sequence is well conserved in *Salmonella* (>86% of identity sequence in all length) and *Escherichia* (98.7% of identity sequence) replicons. To identify the terminus of replication, we searched using Fuzznuc the known *dif* site sequence (GGTGCGCATAATGTATATTATGTTAAAT) (Hendrickson and Lawrence 2007) and also the *terC* sequence of *E. coli* K12 MG1655 (GGATGTTGTAACTA) in all the genomes analyzed (Duggin and Bell 2009). Both sequences are well conserved between the two species and are close to each other along the chromosome (<20 kb). Cumulative GC and AT skews analysis in 10 kb sliding windows (Greub et al. 2003), the switch of KOPS orientation (Bigot et al. 2005), and the identification of the *dnaA* gene (Mackiewicz et al. 2004) close to the origin were used to confirm/support the predictions. We then classed all prophage genes and KOPS motifs according to their orientation relative to the replication fork movement.

Statistics on Oligonucleotide Usage

Over-representation of KOPS and *matS* motifs was determined by comparison to the expected frequencies of these motifs in the different genomes. The expected frequencies of KOPS were calculated using a Markov maximal order model as in Schbath (1997). As KOPS motifs display a degenerate nucleotide at position 4, random expectation was calculated for each one of the four possible KOPS motifs independently. The degenerate *matS* motifs are longer (13 nucleotides), and their random frequencies cannot be estimated confidently with the Markov maximum order model because such long motifs are expected at very low frequencies. Random expectation of these motifs was then estimated using the hosts' or (pro)phages' nucleotide content:

$F(matS) = f(G)^3 \cdot f(C)^3 \cdot f(A)^2 \cdot f(T)^2 \cdot f(A/G) \cdot f(T/C)$, with $f(X)$ the frequency of nucleotide X in the genome.

Supplementary Material

Supplementary tables S1–S5 and figures S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a European Research Council starting grant (EVOMOBILOME no. 281605) and a grant from the Ministère de l'enseignement supérieur et de la recherche to L.-M.B.

References

Abedon ST, Calendar RL. 2005. The bacteriophages. New York: Oxford University Press.
Abedon ST, Lejeune JT. 2005. Why bacteriophage encode exotoxins and other virulence factors. *Evol Bioinform Online.* 1:97–110.
Ackermann HW. 2007. *Salmonella* phages examined in the electron microscope. *Methods Mol Biol.* 394:213–234.
Allison GE, Angeles D, Tran-Dinh N, Verma NK. 2002. Complete genomic sequence of SfV, a serotype-converting temperate bacteriophage of *Shigella flexneri*. *J Bacteriol.* 184:1974–1987.
Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.
Angly FE, Felts B, Breitbart M, et al. (18 co-authors). 2006. The marine viromes of four oceanic regions. *PLoS Biol.* 4:e368.
Asadulghani M, Ogura Y, Ooka T, Itoh T, Sawaguchi A, Iguchi A, Nakayama K, Hayashi T. 2009. The defective prophage pool of *Escherichia coli* O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog.* 5: e1000408.
Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol.* 2:2006.0008.
Balbontin R, Figueroa-Bossi N, Casadesus J, Bossi L. 2008. Insertion hot spot for horizontally acquired DNA within a bidirectional small-RNA locus in *Salmonella enterica*. *J Bacteriol.* 190:4075–4078.
Banks DJ, Beres SB, Musser JM. 2002. The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. *Trends Microbiol.* 10:515–521.
Bigot S, Saleh OA, Lesterlin C, Pages C, El Karoui M, Dennis C, Grigoriev M, Allemand JF, Barre FX, Cornet F. 2005. KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase. *EMBO J.* 24:3770–3780.
Blum G, Ott M, Lischewski A, Ritter A, Imrich H, Tschape H, Hacker J. 1994. Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. *Infect Immun.* 62:606–614.
Bossi L, Fuentes JA, Mora G, Figueroa-Bossi N. 2003. Prophage contribution to bacterial population dynamics. *J Bacteriol.* 185:6467–6471.
Botstein D. 1980. A theory of modular evolution for bacteriophages. *Ann N Y Acad Sci.* 354:484–490.
Boyd EF, Brussow H. 2002. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol.* 10:521–529.
Breitbart M, Haynes M, Kelley S, et al. (13 co-authors). 2008. Viral diversity and dynamics in an infant gut. *Res Microbiol.* 159:367–373.
Brown SP, Le Chat L, De Paepe M, Taddei F. 2006. Ecology of microbial invasions: amplification allows virus carriers to invade more rapidly when rare. *Curr Biol.* 16:2048–2052.
Brussow H, Canchaya C, Hardt WD. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev.* 68:560–602.
Bukhari AI, Metlay M. 1973. Genetic mapping of prophage Mu. *Virology* 54:109–116.
Campbell A. 2003. Prophage insertion sites. *Res Microbiol.* 154:277–282.
Campbell A, Botstein D. 1983. Evolution of the lambdoid phages. In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA, editors. Lambda II. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory. p. 365–380.
Campbell AM. 1992. Chromosomal insertion sites for phages and plasmids. *J Bacteriol.* 174:7495–7499.
Campbell AM. 2002. Preferential orientation of natural lambdoid prophages and bacterial chromosome organization. *Theor Popul Biol.* 61:503–507.
Canchaya C, Fournous G, Brussow H. 2004. The impact of prophages on bacterial chromosomes. *Mol Microbiol.* 53:9–18.
Canchaya C, Proux C, Fournoux G, Bruttin A, Brussow H. 2003. Prophage genomics. *Microbiol Mol Biol Rev.* 67:238–276.
Casjens S. 2003. Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol.* 49:277–300.
Couturier E, Rocha EPC. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol Microbiol.* 59: 1506–1518.
Creuzburg K, Heeren S, Lis CM, Kranz M, Hensel M, Schmidt H. 2011. Genetic background and mobility of variants of the gene nleA in attaching and effacing *Escherichia coli*. *Appl Environ Microbiol.* 77: 8705–8713.
Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol.* 10: 210.
Cui T, Moro-oka N, Ohsumi K, Kodama K, Ohshima T, Ogasawara N, Mori H, Wanner B, Niki H, Horiuchi T. 2007. *Escherichia coli* with a linear genome. *EMBO Rep.* 8:181–187.
Dearborn AD, Laurinmaki P, Chandramouli P, Rodenburg CM, Wang S, Butcher SJ, Dokland T. 2012. Structure and size determination of bacteriophage P2 and P4 procapsids: function of size responsiveness mutations. *J Struct Biol.* 178:215–224.
Dobrindt U, Blum-Oehler G, Nagy G, Schneider G, Johann A, Gottschalk G, Hacker J. 2002. Genetic structure and distribution of four pathogenicity islands (PAI I(536) to PAI IV(536)) of uropathogenic *Escherichia coli* strain 536. *Infect Immun.* 70:6365–6372.
Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J Mol Biol.* 260:649–663.
Duggin IG, Bell SD. 2009. Termination structures in the *Escherichia coli* chromosome replication fork trap. *J Mol Biol.* 387:532–539.
Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol.* 7: e1002195.
Edgar R, Rokney A, Feeney M, Semsey S, Kessel M, Goldberg MB, Adhya S, Oppenheim AB. 2008. Bacteriophage infection is targeted to cellular poles. *Mol Microbiol.* 68:1107–1116.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
Edlin G, Lin L, Bitner R. 1977. Reproductive fitness of P1, P2, and Mu lysogens of *Escherichia coli*. *J Virol.* 21:560–564.
Edwards RA, Rohwer F. 2005. Viral metagenomics. *Nat Rev Microbiol.* 3: 504–510.
Esnault E, Valens M, Espeli O, Boccard F. 2007. Chromosome structuring limits genome plasticity in *Escherichia coli*. *PLoS Genet.* 3:e226.
Fouts DE. 2006. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* 34:5839–5851.
Garcia-Aljaro C, Muniesa M, Jofre J, Blanch AR. 2009. Genotypic and phenotypic diversity among induced, stx2-carrying bacteriophages from environmental *Escherichia coli* strains. *Appl Environ Microbiol.* 75:329–336.
Garcia-Russell N, Harmon TG, Le TQ, Amaladas NH, Mathewson RD, Segall AM. 2004. Unequal access of chromosomal regions to each other in *Salmonella*: probing chromosome structure with phage

lambda integrase-mediated long-range rearrangements. *Mol Microbiol.* 52:329–344.
Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* 14:685–695.
Greub G, Mege JL, Raoult D. 2003. *Parachlamydia acanthamoebae* enters and multiplies within human macrophages and induces their apoptosis. *Infect Immun.* 71:5979–5985.
Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. 2008. The Vienna RNA websuite. *Nucleic Acids Res.* 36:W70–W74.
Guerrero-Ferreira RC, Viollier PH, Ely B, Poindexter JS, Georgieva M, Jensen GJ, Wright ER. 2011. Alternative mechanism for bacteriophage adsorption to the motile bacterium *Caulobacter crescentus*. *Proc Natl Acad Sci U S A.* 108:9963–9968.
Hendrickson H, Lawrence JG. 2007. Mutational bias suggests that replication termination occurs near the dif site, not at Ter sites. *Mol Microbiol.* 64:42–56.
Hendrix RW, Smith MCM, Burns RN, Ford ME, Hatfull GF. 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci U S A.* 96: 2192–2197.
Hermans AP, Beuling AM, van Hoek AH, Aarts HJ, Abee T, Zwietering MH. 2006. Distribution of prophages and SGI-1 antibiotic-resistance genes among different *Salmonella enterica* serovar Typhimurium isolates. *Microbiology* 152:2137–2147.
Huang HY, Chang HY, Chou CH, Tseng CP, Ho SY, Yang CD, Ju YW, Huang HD. 2009. sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic Acids Res.* 37:D150–D154.
Huber KE, Waldor MK. 2002. Filamentous phage integration requires the host recombinases XerC and XerD. *Nature* 417:656–659.
Itaya M, Tsuge K, Koizumi M, Fujita K. 2005. Combining two genomes in one cell: stable cloning of the *Synechocystis* PCC6803 genome in the *Bacillus subtilis* 168 genome. *Proc Natl Acad Sci U S A.* 102: 15971–15976.
Jobb G, von Haeseler A, Strimmer K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol.* 4:18.
Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW. 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J Mol Biol.* 299: 27–51.
Julio SM, Heithoff DM, Mahan MJ. 2000. ssrA (tmRNA) plays a role in *Salmonella enterica* serovar Typhimurium pathogenesis. *J Bacteriol.* 182:1558–1563.
King AMQ, Lefkowitz E, Adams MJ, Carstens EB. 2011. Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses. Waltham (MA): Elsevier.
Kourilsky P. 1973. Lysogenization by bacteriophage lambda. I. Multiple infection and the lysogenic response. *Mol Gen Genet.* 122:183–195.
Kroger C, Dillon SC, Cameron AD, et al. (21 co-authors). 2012. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. *Proc Natl Acad Sci U S A.* 109: E1277–E1286.
Labrie SJ, Samson JE, Moineau S. 2010. Bacteriophage resistance mechanisms. *Nat Rev Microbiol.* 8:317–327.
Lathe WC, Snel B, Bork P. 2000. Gene context conservation of a higher order than operons. *Trends Biochem Sci.* 25:474–479.
Lavigne R, Darius P, Summer EJ, Seto D, Mahadevan P, Nilsson AS, Ackermann HW, Kropinski AM. 2009. Classification of *Myoviridae* bacteriophages using protein sequence similarity. *BMC Microbiol.* 9: 224.
Lawrence JG, Hendrickson H. 2003. Lateral gene transfer: when will adolescence end? *Mol Microbiol.* 50:739–749.
Lee SY, Bailey SC, Apirion D. 1978. Small stable RNAs from *Escherichia coli*: evidence for the existence of new molecules and for a new ribonucleoprotein particle containing 6S RNA. *J Bacteriol.* 133: 1015–1023.
Levy O, Ptacin JL, Pease PJ, Gore J, Eisen MB, Bustamante C, Cozzarelli NR. 2005. Identification of oligonucleotide sequences that direct the movement of the *Escherichia coli* FtsK translocase. *Proc Natl Acad Sci U S A.* 102:17618–17623.
Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. 2008a. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* 24:863–865.
Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. 2008b. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol.* 25:762–777.
Lowe T, Eddy S. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25: 955–964.
Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek MR, Cebrat S. 2004. Where does bacterial replication start? Rules for predicting the oriC region. *Nucleic Acids Res.* 32:3781–3791.
McDaniel TK, Jarvis KG, Donnenberg MS, Kaper JB. 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proc Natl Acad Sci U S A.* 92:1664–1668.
Meile JC, Mercier R, Stouf M, Pages C, Bouet JY, Cornet F. 2011. The terminal region of the *E. coli* chromosome localises at the periphery of the nucleoid. *BMC Microbiol.* 11:28.
Mercier R, Petit MA, Schbath S, Robin S, El Karoui M, Boccard F, Espeli O. 2008. The MatP/matS site-specific system organizes the terminus region of the *E. coli* chromosome into a macrodomain. *Cell* 135: 475–485.
Mizuuchi K. 1992. Transpositional recombination: mechanistic insights from studies of Mu and other elements. *Annu Rev Biochem.* 61: 1011–1051.
Mmolawa PT, Schmieger H, Heuzenroeder MW. 2003. Bacteriophage ST64B, a genetic mosaic of genes from diverse sources isolated from *Salmonella enterica* serovar typhimurium DT 64. *J Bacteriol.* 185: 6481–6485.
Napolitano MG, Almagro-Moreno S, Boyd EF. 2011. Dichotomy in the evolution of pathogenicity island and bacteriophage encoded integrases from pathogenic *Escherichia coli* strains. *Infect Genet Evol.* 11:423–436.
Nechaev S, Severinov K. 2008. The elusive object of desire—interactions of bacteriophages and their hosts. *Curr Opin Microbiol.* 11: 186–193.
Nunes-Duby SE, Kwon HJ, Tirumalai RS, Ellenberger T, Landy A. 1998. Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res.* 26:391–406.
Ohnishi M, Kurokawa K, Hayashi T. 2001. Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? *Trends Microbiol.* 9:481–485.
Otsuka AJ, Buoncristiani MR, Howard PK, Flamm J, Johnson C, Yamamoto R, Uchida K, Cook C, Ruppert J, Matsuzaki J. 1988. The *Escherichia coli* biotin biosynthetic enzyme sequences predicted from the nucleotide sequence of the bio operon. *J Biol Chem.* 263: 19577–19585.
Peters JE, Craig NL. 2001. Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein TnsE. *Genes Dev.* 15:737–747.
Ptashne M. 1992. Genetic switch: phage lambda and higher organisms. Cambridge (MA): Blackwell.
Raghavan R, Groisman EA, Ochman H. 2011. Genome-wide detection of novel regulatory RNAs in *E. coli*. *Genome Res.* 21: 1487–1497.
Ravin NV. 2011. N15: the linear phage-plasmid. *Plasmid* 65:102–109.
Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466:334–338.
Reyes-Lamothe R, Wang X, Sherratt D. 2008. *Escherichia coli* and its chromosome. *Trends Microbiol.* 16:238–245.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16:276–277.
Rocha EPC. 2004. Order and disorder in bacterial genomes. *Curr Opin Microbiol.* 7:519–527.
Rocha EPC. 2008. The organisation of the bacterial genome. *Annu Rev Genet.* 42:211–233.

Rocha EPC, Danchin A. 2003. Essentiality, not expressiveness, drives gene strand bias in bacteria. *Nat Genet.* 34:377–378.
Rohwer F, Edwards R. 2002. The phage proteomic tree: a genome-based taxonomy for phage. *J Bacteriol.* 184:4529–4535.
Roossinck MJ. 2011. The good viruses: viral mutualistic symbioses. *Nat Rev Microbiol.* 9:99–108.
Schbath S. 1997. An efficient statistic to detect over- and under-represented words in DNA sequences. *J Comput Biol.* 4:189–192.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
Scolari VF, Bassetti B, Sclavi B, Lagomarsino MC. 2011. Gene clusters reflecting macrodomain structure respond to nucleoid perturbations. *Mol Biosyst.* 7:878–888.
Shinedling S, Parma D, Gold L. 1987. Wild-type bacteriophage T4 is restricted by the lambda rex genes. *J Virol.* 61:3790–3794.
Shinhara A, Matsui M, Hiraoka K, Nomura W, Hirano R, Nakahigashi K, Tomita M, Mori H, Kanai A. 2011. Deep sequencing reveals as-yet-undiscovered small RNAs in *Escherichia coli*. *BMC Genomics* 12:428.
Six EW, Klug CA. 1973. Bacteriophage P4: a satellite virus depending on a helper such as prophage P2. *Virology* 51:327–344.
Smith MC, Thorpe HM. 2002. Diversity in the serine recombinases. *Mol Microbiol.* 44:299–307.
Snel B, Huynen MA, Dutilh BE. 2005. Genome trees and the nature of genome evolution. *Ann Rev Microbiol.* 59:191–209.
St-Pierre F, Endy D. 2008. Determination of cell fate selection during phage lambda infection. *Proc Natl Acad Sci U S A.* 105:20705–20710.
Suttle CA. 2005. Viruses in the sea. *Nature* 437:356–361.
Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol.* 8:207–17.
Thomson N, Baker S, Pickard D, et al. (15 co-authors). 2004. The role of prophage-like elements in the diversity of *Salmonella enterica* serovars. *J Mol Biol.* 339:279–300.
Touchon M, Hoede C, Tenaillon O, et al. (41 co-authors). 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* 5:e1000344.
Touchon M, Rocha EP. 2007. Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol.* 24:969–981.
Touchon M, Rocha EP. 2010. The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and Salmonella. *PLoS One* 5:e11126.
Touzain F, Petit MA, Schbath S, El Karoui M. 2011. DNA motifs that sculpt the bacterial chromosome. *Nat Rev Microbiol.* 9:15–26.
Val ME, Skovgaard O, Ducos-Galand M, Bland MJ, Mazel D. 2012. Genome engineering in *Vibrio cholerae*: a feasible approach to address biological issues. *PLoS Genet.* 8:e1002472.
Valens M, Penaud S, Rossignol M, Cornet F, Boccard F. 2004. Macrodomain organization of the *Escherichia coli* chromosome. *EMBO J.* 23:4330–4341.
Van Melderen L, Saavedra De Bast M. 2009. Bacterial toxin-antitoxin systems: more than selfish entities? *PLoS Genet.* 5:e1000437.
Vernikos GS, Thomson NR, Parkhill J. 2007. Genetic flux over time in the *Salmonella* lineage. *Genome Biol.* 8:R100.
Wang H, Yang CH, Lee G, Chang F, Wilson H, del Campillo-Campbell A, Campbell A. 1997. Integration specificities of two lambdoid phages (21 and e14) that insert at the same attB site. *J Bacteriol.* 179: 5705–5711.
Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S. 2001. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* 15:1637–1651.
Waters LS, Storz G. 2009. Regulatory RNAs in bacteria. *Cell* 136:615–628.
Weinbauer MG. 2004. Ecology of prokaryotic viruses. *FEMS Microbiol Rev.* 28:127–181.
Wiggins PA, Cheveralls KC, Martin JS, Lintner R, Kondev J. 2010. Strong intranucleoid interactions organize the *Escherichia coli* chromosome into a nucleoid filament. *Proc Natl Acad Sci U S A.* 107:4991–4995.
Williams KP. 2002. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.* 30:866–875.
Williams KP. 2003. Traffic at the tmRNA gene. *J Bacteriol.* 185: 1059–1070.
Winstanley C, Langille MGI, Fothergill JL, et al. (19 co-authors). 2008. Newly introduced genomic prophage islands are critical determinants of in vivo competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res.* 19:12–23.
Withers M, Wernisch L, dos Reis M. 2006. Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome. *RNA* 12: 933–942.
Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. *Trends Genet.* 18:472–479.
Zaslaver A, Mayo A, Ronen M, Alon U. 2006. Optimal gene partition into operons correlates with gene functional order. *Phys Biol.* 3: 183–189.
Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool. *Nucleic Acids Res.* 39:W347–W352.

Conclusions et perspectives

Cette étude a mis en évidence que les prophages sont intégrés de manière organisée au sein de leur génome hôte. Ils sont structurés en loci d'intégration stables au cours du temps. Ils sont répartis de manière non-aléatoire le long de l'axe de réplication et parmi les différents macrodomaines chromosomiques. Ils sont fortement polarisés par rapport à l'axe de réplication. Enfin, les prophages présentent un contenu en motifs chromosomiques similaire à la densité locale en motifs du chromosome hôte. Ces résultats suggèrent que les phages tempérés présentent des signes d'adaptation qui leurs permettent de minimiser leur impact délétère sur l'organisation chromosomique de l'hôte. Les conclusions de ce travail permettent de soulever certaines questions.

i) Les prophages sont intégrés au sein d'un faible nombre de loci. Ces loci d'intégration sont stables dans le temps et un grand nombre d'entre eux sont même partagés entre les genres *Escherichia* et *Salmonella*. Quels processus sont à l'origine de la conservation de ces sites d'intégration? Deux hypothèses sont envisageables: a) Les sites *att*B sont conservés par sélection purificatrice. b) Les phages s'intègrent dans des éléments fonctionnels bactériens contraints par la sélection purificatrice. Dans le cadre de la première hypothèse, une pression sélective maintiendrait ces sites d'intégrations directement. Lors d'une infection par un phage tempéré, le phage a la possibilité de s'intégrer au sein du génome hôte ou d'entamer le cycle lytique. Si le site d'intégration du phage ne peut pas être reconnu par l'intégrase, la lysogénie ne pourra pas s'établir, ce qui augmentera la probabilité du phage d'entrer en cycle lytique (Shimada, et al. 1972). Ce processus pourrait alors conduire à l'élimination rapide des bactéries ayant muté au niveau de leurs sites d'intégration par sélection purificatrice. Selon la deuxième hypothèse, les phages pourraient s'intégrer préférentiellement au sein des ARNt et des ARNtm car ces séquences sont fortement contraintes sur l'ensemble de leur séquence, ce qui empêcherait les bactéries d'accumuler des mutations au niveau de ces sites. Les séquences codantes peuvent en revanche accumuler des substituions synonymes sans affecter la fonction protéique. L'analyse conduite ici montre que des sites d'intégration situés au sein de gènes codants sont également conservés dans le temps. D'autres sites d'intégration situés au sein de séquences intergéniques semblent également conservés au cours du temps. Ces observations sont donc plutôt en faveur d'une sélection purificatrice maintenant les sites d'intégration (première hypothèse). Néanmoins, les gènes ARN restent des cibles privilégiées. Il est donc

probable que les deux explications soient valides. On peut notamment supposer que, dans le cadre d'une évolution sur le long terme où les phages sélectionnent les bactéries ayant conservés les sites *att*B, cette stratégie sera plus stable dans le temps lorsque les phages ciblent des séquences à évolution lente. Cette question pourrait éventuellement être résolue par des travaux d'évolution expérimentale. Il est envisageable qu'une étude de la stabilité des prophages par déplacement de sites *att*B de phages dans des régions non fonctionnelles puisse répondre à cette question.

ii) Cette étude a mis en évidence que les prophages sont structurés par rapport à la réplication de leurs hôtes (distance *ori-ter*) et par rapport à la structure du chromosome en macrodomaines. Quelle est l'origine de ce biais? Ces observations semblent refléter l'accessibilité du chromosome au sein de la cellule car les macrodomaines les plus riches en prophages sont situés plus en périphérie et aux pôles de la cellule. L'augmentation du nombre de prophages le long de l'axe *ori-ter* et le biais d'intégration dans les macrodomaines peut également refléter un évitement lié à l'expression de certaines régions. En effet, l'expression des gènes est inégale selon les macrodomaines et l'axe *ori-ter*. L'intégration à proximité de gènes fortement exprimés pourrait déstabiliser la lysogénie et favoriser l'entrée en cycle lytique. En accord avec cette hypothèse, nous avons montré dans cette étude que les phages tendent à s'intégrer au sein des ARNt les moins fortement exprimés. Il n'est pas non plus exclu que ces biais d'intégrations reflètent à la fois une accessibilité inégale selon les régions chromosomiques et une stabilité inégale de la lysogénie entre certaines régions. L'origine de cette intégration inégale pourrait être résolue par l'étude des positions d'intégration du phage Mu. Ce phage a en effet la spécificité de s'intégrer de manière apparemment aléatoire au sein du génome hôte car il s'intègre vie une transposase DDE (Mizuuchi 1992). Des travaux expérimentaux pourraient ainsi permettre de déterminer si cet élément présente des biais d'intégration qui refléteraient alors un biais d'accessibilité de certaines régions chromosomiques. Le faible nombre de prophages Mu (5) détectés dans notre analyse ne permet pas de tester statistiquement cette hypothèse.

iii) La quasi-totalité des prophages lambdoïdes sont polarisés par rapport à la fourche de réplication. Cette polarisation n'est pas observée chez la majorité des prophages non lambdoïdes. Quelle est l'origine de ce biais? Deux hypothèses peuvent expliquer cette forte polarisation: a) la polarité est liée au mécanisme de recombinaison des intégrases de lambdoïdes qui serait dépendant de la fourche de réplication. b) les lambdoïdes sont moins stables lorsqu'ils sont anti-orientés par rapport à la fourche de réplication. Outre certains gènes accessoires, le répresseur *cI* est le seul gène des prophages lambdoïdes à être exprimé

constitutivement durant la lysogénie (Ptashne 1992). Son expression en quantités suffisantes est requise pour le maintient de la lysogénie. Or, il a été montré que le biais d'orientation des gènes par rapport à l'axe de réplication résulterait d'une adaptation favorisant la régulation transcriptionnelle des gènes (section 1.2.1). Pourtant, le gène régulateur *cI* est presque systématiquement anti-orienté à l'axe de réplication, ce qui va à l'encontre de l'hypothèse d'un polarisation favorisant la stabilité de la lysogénie. L'hypothèse alternative d'un biais mécanistique d'intégration des lambdoïdes par rapport à la fourche de réplication n'est cependant pas supportée par nos connaissances actuelles du mécanisme de recombinaison site-spécifique des intégrases. Il est envisageable que l'accessibilité de la fourche de réplication ou l'hémiméthylation du brin néosynthétisé aient un impact sur l'intégration des lambdoïdes. D'autres causes peuvent potentiellement mener à contre-sélectionner les prophages anti-orientés tels que la recombinaison intra-chromosomique entre prophages (De Paepe, et al. 2014). En effet, la recombinaison homologue entre prophages anti-orientés pourrait provoquer des réarrangements génétiques délétères pour ces éléments. La forte polarisation des prophages lambdoïdes reste cependant inexpliquée. L'inversion de la polarité des prophages associée à des tests de fitness en laboratoire pourrait éventuellement permettre de répondre à cette question.

II Manipuler ou remplacer les fonctions de recombinaison de l'hôte: un dilemme qui façonne l'évolvabilité des phages.

Contexte

Le flux important de gènes phagiques au sein des génomes bactériens contribue fortement au renouvellement de la diversité des répertoires génétiques bactériens. Les phages sont des entités extrêmement diverses génétiquement. Les phages lambdoïdes, qui sont largement majoritaires parmi les phages tempérés d'entérobactéries, présentent un mosaïcisme génomique important, qui est associé à un taux élevé de recombinaison et à une forte diversité génétique. Il a été montré que cette recombinaison importante des lambdoïdes est liée à l'utilisation de recombinases moins spécifiques que le système de recombinaison de l'hôte RecA/RecBCD-Chi (Martinsohn, et al. 2008). Les systèmes de recombinaison des phages lambdoïdes ont principalement été étudiés chez le phage Lambda qui code pour le système Redα-Redβ-Gam (Smith 1983). D'autres lambdoïdes tels que le phage P22 utilisent des gènes différents codant pour les protéines Erf, Abc1 et Abc2 (Murphy 2012). Bien que ces systèmes de recombinaison relativement permissifs contribuent à augmenter la diversité des génomes phagiques, ils constituent avant tout des systèmes essentiels permettant d'obtenir des concatémères d'ADN phagiques (Enquist and Skalka 1973). La production de concatémères d'ADN est une étape nécessaire à l'encapsidation des génomes des lambdoïdes et d'autres phages (Enquist and Skalka 1973) (section 2.2.3).

Il est intéressant de noter que la découverte des sites Chi, promouvant la recombinaison hôte via la voie RecBCD, a été faite grâce à l'obtention de mutants du phage Lambda (Stahl 2005). Les phages Lambda défectueux pour leurs gènes de recombinaison sont incapables de produire des concatémères d'ADN et ne peuvent donc pas encapsider leur génome (Enquist and Skalka 1973). Cependant, il a été montré que l'introduction d'octamères Chi au sein de tels mutants permettait de restaurer leur faculté à produire des concatémères. La découverte des sites Chi et de la voie de recombinaison RecBCD a donc été permise par l'absence de Chi au sein du génome de Lambda. Lors de la précédente étude, je me suis intéressé à la présence de certains motifs chromosomiques tels que KOPS qui pourraient résulter d'une adaptation des phages tempérés à la structure du chromosome hôte. Lors de ce travail, j'ai également recherché les motifs Chi parmi les prophages. J'ai pu observé une distribution très inégale de

ces motifs parmi les différents prophages. Certains éléments présentaient un grand nombre de motifs Chi (~10) alors que d'autres n'en contenaient pratiquement aucun. Cette observation préliminaire suggérait donc que les motifs Chi n'étaient pas liés à l'accommodation des phages au chromosome. La présence de motifs impliqués dans la recombinaison homologue soulevait ainsi la possibilité qu'ils aient un rôle fonctionnel chez ces phages.

La détection et la classification d'un très grand nombre (500) de prophages chez *Escherichia* et *Salmonella* m'a permis d'augmenter substantiellement le nombre de génomes de phages tempérés disponibles (70 sont disponibles sur GenBank) pour étudier la diversité et le mosaïcisme chez ces phages.

L'objectif de cette étude a été d'identifier les différents mécanismes de recombinaison utilisés par les phages lambdoïdes afin d'étudier l'impact de ces différents systèmes sur la diversité génétique de ces phages et sur leur mosaïcisme.

Approche

Détection des systèmes de recombinaison et annotation fonctionnelle

Il existe une forte diversité de systèmes de recombinaison chez les phages. Ces systèmes peuvent cependant être classés en un faible nombre de familles partageant une origine commune (Lopes, et al. 2010). Dans cette étude, je me suis concentré sur la détection des recombinases chez les lambdoïdes. Des profiles HMM ont été construits dans une étude antérieure pour un très grand nombre de recombinases phagiques (Lopes, et al. 2010). J'ai également construit des familles de protéines homologues (et les profiles HMM correspondants) au sein de l'ensemble du protéome des phages et prophages lambdoïdes. L'utilisation de l'outil de comparaison de profiles contre profiles HHsearch (Remmert, et al. 2012) permet d'effectuer des comparaisons de séquences avec une forte sensibilité. La forte sensibilité de cette approche a néanmoins l'inconvénient de détecter un fort nombre de faux positifs. Cet écueil est d'autant plus important ici que certaines recombinases partagent des domaines hélicases ou ATPases que l'on retrouve chez d'autres enzymes telles que DnaB (Lopes, et al. 2010). Pour éviter la surprédiction de recombinases, j'ai alors comparé l'ensemble des familles protéiques des (pro)phages à l'ensemble des familles protéiques de PFAM-A. Cette étape m'a ainsi permis d'éliminer un certain nombre de protéines similaires à des recombinases phagiques mais correspondant à d'autres enzymes annotées dans PFAM-A. Les inhibiteurs de RecBCD (Gam et Abc2) ont pu être détectés grâce aux profiles PFAM

correspondants. Puisque certains prophages peuvent présenter des délétions ou des événements de pseudogénisation (Asadulghani, et al. 2009; Casjens 2003), je n'ai considéré que les prophages d'une taille supérieure à 30kb (cette taille correspond à la taille minimale des phages tempérés non satellites d'entérobactéries présents dans GenBank). Enfin, je n'ai pas considéré les prophages présentant une forte similarité de génome car il est probable qu'ils soient issus du même événement d'intégration. La comparaison des profiles de familles protéiques des (pro)phages avec la banque de profiles annotés de PFAM-A m'a également permis d'annoter et de catégoriser les fonctions des familles protéiques des (pro)phages comme schématisé (Fig 19).

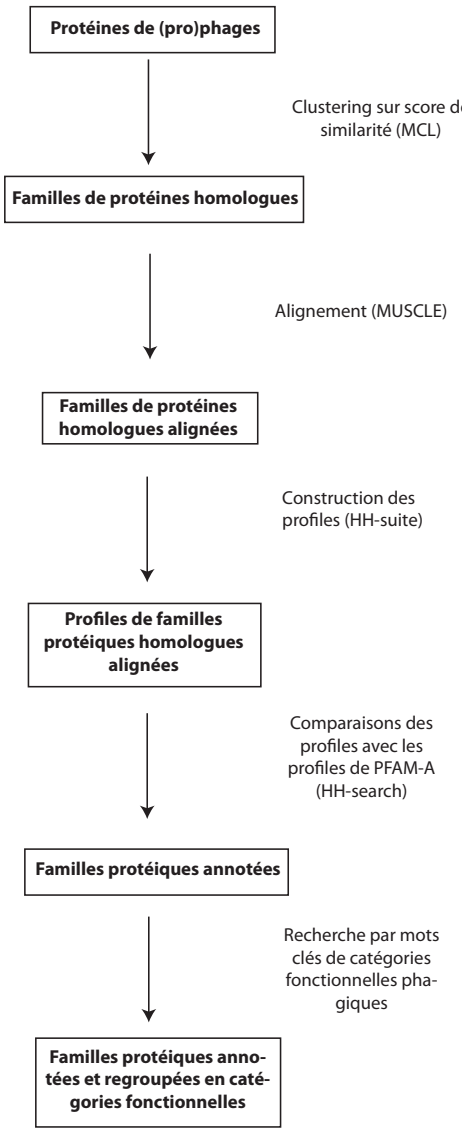

Figure 19: Annotation fonctionnelle des familles protéiques des (pro)phages.

Détection des motifs Chi

Les statistiques sur l'octamère Chi ont pu être effectuées à l'aide du programme R'MES qui permet de calculer la fréquence attendue de motifs par rapport à la composition en oligonucléotides des génomes (Schbath and Hoebeke 2011). J'ai ainsi estimé la surreprésentation du motif Chi par rapport à des compositions en oligonucléotides de différentes tailles: la composition en nucléotides, en tri-nucléotides, en penta-nucléotides et en hetpa-nucléotides (modèle maximal). Les fréquences attendues de motifs Chi ont été calculées pour chaque génome phagique individuellement car des phages différents peuvent potentiellement présenter des compositions en oligonucléotides différentes.

Estimation du mosaïcisme phagique

Le mosaïcisme se caractérise, lors de comparaisons de génomes phagiques, par l'alternance de séquences fortement similaires et de séquences fortement dissimilaires (Juhala, et al. 2000) (section 2.3.1). Différentes méthodes de quantification du mosaïcisme ont été envisagées. J'ai d'abord envisagé d'utiliser des méthodes de reconstruction phylogénétiques basées sur l'incongruence entre arbres de gènes et arbre des phages. Ces méthodes ne se sont cependant pas avérées applicables à cause de la difficulté d'obtenir un arbre des phages robuste. En effet, j'ai été confronté au faible nombre de marqueurs phylogénétiques et au mosaïcisme important de ces éléments. Ces difficultés sont d'autant plus importantes que les phages présentent des génomes de taille relativement réduite (typiquement <50kb). J'ai donc utilisé une méthode plus restrictive mais qui a l'avantage de ne pas nécessiter la construction d'un arbre des phages. Cette méthode consiste à se focaliser sur les gènes mosaïques récents. Les gènes mosaïques ont ainsi été définis comme des gènes fortement similaires présents dans des génomes phagiques fortement dissimilaires. J'ai donc calculé le taux de divergence *d* des gènes phagiques orthologues par maximum de vraisemblance (Schmidt, et al. 2002). J'ai également estimé le taux de divergence *D* de chaque paire de (pro)phages défini comme le taux de divergence moyen de leurs gènes orthologues. Différents seuils de similarités de gènes *d* et de dissimilarités de phages *D* ont alors été testés pour estimer la proportion de gènes mosaïques.

Article 2

OPEN ACCESS Freely available online

PLOS | GENETICS

Manipulating or Superseding Host Recombination Functions: A Dilemma That Shapes Phage Evolvability

Louis-Marie Bobay[1,2,3]*, Marie Touchon[1,2], Eduardo P. C. Rocha[1,2]

1 Microbial Evolutionary Genomics, Institut Pasteur, Paris, France, **2** CNRS, UMR3525, Paris, France, **3** Université Pierre et Marie Curie, Cellule Pasteur UPMC, Paris, France

Abstract

Phages, like many parasites, tend to have small genomes and may encode autonomous functions or manipulate those of their hosts'. Recombination functions are essential for phage replication and diversification. They are also nearly ubiquitous in bacteria. The *E. coli* genome encodes many copies of an octamer (Chi) motif that upon recognition by RecBCD favors repair of double strand breaks by homologous recombination. This might allow self from non-self discrimination because RecBCD degrades DNA lacking Chi. Bacteriophage Lambda, an *E. coli* parasite, lacks Chi motifs, but escapes degradation by inhibiting RecBCD and encoding its own autonomous recombination machinery. We found that only half of 275 lambdoid genomes encode recombinases, the remaining relying on the host's machinery. Unexpectedly, we found that some lambdoid phages contain extremely high numbers of Chi motifs concentrated between the phage origin of replication and the packaging site. This suggests a tight association between replication, packaging and RecBCD-mediated recombination in these phages. Indeed, phages lacking recombinases strongly over-represent Chi motifs. Conversely, phages encoding recombinases and inhibiting host recombination machinery select for the absence of Chi motifs. Host and phage recombinases use different mechanisms and the latter are more tolerant to sequence divergence. Accordingly, we show that phages encoding their own recombination machinery have more mosaic genomes resulting from recent recombination events and have more diverse gene repertoires, i.e. larger pan genomes. We discuss the costs and benefits of superseding or manipulating host recombination functions and how this decision shapes phage genome structure and evolvability.

Citation: Bobay L-M, Touchon M, Rocha EPC (2013) Manipulating or Superseding Host Recombination Functions: A Dilemma That Shapes Phage Evolvability. PLoS Genet 9(9): e1003825. doi:10.1371/journal.pgen.1003825

Editor: Josep Casadesús, Universidad de Sevilla, Spain

Received May 21, 2013; **Accepted** August 8, 2013; **Published** September 26, 2013

Funding: This work was supported by an European Research Council starting grant [EVOMOBILOME n°281605]; and a grant from the Ministère de l'enseignement supérieur et de la recherche to LMB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: louis-marie.bobay@pasteur.fr

Introduction

Genetic recombination plays key roles in biology. Recombinases are required for essential cellular functions such as repair of stalled or collapsed replication forks, DNA repair and chromosome segregation [1,2]. Recombination also drives genetic diversification and increases the efficiency of natural selection [3,4]. Inter-genomic recombination allows horizontal gene transfer between organisms and exchange of sequences between viruses infecting the same cell [5]. Illegitimate and homologous recombination events between bacterial viruses (phages) are frequent and result in strongly mosaic genomes, i.e. genomes with strong internal phylogenetic incongruences [6], but the relative importance of each recombination mechanism remains unclear [7–9]. The group of lambdoid phages provides a striking example of this phenomenon. These temperate phages account for more than two thirds of *E. coli* prophages [10], and are extremely diverse from the genetic, structural and physiological point of view. Nevertheless, they all have similar genetic organization and this allows the production of viable hybrids by inter-genomic recombination [11,12]. Lambdoid genomes are organized in relatively autonomous gene clusters with genes being encoded next to their interactants, i.e. genes encoding an interacting protein or the targeted DNA site [13]. Moreover, the organization of morphogenesis genes strikingly reflects the order of the proteins forming the virion structure, suggesting a direct link between gene order and function or structure within each module [13]. The extent and phylogenetic range of genetic exchange can be very large: lambdoids include phages with different virion structures such as Siphovirus Lambda, Podovirus P22 or Myovirus SfV, showing that recombination blurs the traditional taxonomy (based on virion morphology). Nevertheless, two thirds of the lambdoid phages in *E. coli* are closely related to phage Lambda and display a *Siphoviridae*'s virion structure (Lambda-like elements) [10]. Phages and bacteria are in constant evolutionary arms races [14]. Accordingly, bacterial outer membrane structures that are phage attachment sites evolve very fast because of the selective pressure imposed by phages [15]. Reciprocally, phage proteins involved in attachment to the host cell, such as tail-fiber proteins, evolve fast in response to these changes [16]. Recombination both in the bacteria and in the phage facilitates these diversifying selection processes, accelerating the rate of evolution [17].

Efficient encapsidation of phage Lambda requires concatemeric DNA (reviewed in [18]). These concatemers can be produced by homologous recombination or rolling-circle (sigma) replication (Figure 1). However, rolling-circle replication is inhibited by the exonucleolytic activity of the host RecBCD enzyme from the major homologous recombination pathway [19]. Hence, the

Author Summary

Bacterial viruses, called bacteriophages, are extremely abundant in the biosphere. They have key roles in the regulation of bacterial populations and in the diversification of bacterial genomes. Among these viruses, lambdoid phages are very abundant in enterobacteria and exchange genetic material very frequently. This latter process is thought to increase phage diversity and therefore facilitate adaptation to hosts. Recombination is also essential for the replication of many lambdoid phages. Lambdoids have been described to encode their own recombination genes and inhibit their hosts'. In this study, we show that lambdoids are split regarding their capacity to encode autonomous recombination functions and that this affects the abundance of recombination-related sequence motifs. Half of the phages encode an autonomous system and inhibit their hosts'. The trade-off between superseding and manipulating the hosts' recombination functions has important consequences. The phages encoding autonomous recombination functions have more diverse gene repertoires and recombine more frequently. Viruses, as many other parasites, have small genomes and depend on their hosts for several housekeeping functions. Hence, they often face trade-offs between supersession and manipulation of molecular machineries. Our results suggest these trade-offs may shape viral gene repertoires, their sequence composition and even influence their evolvability.

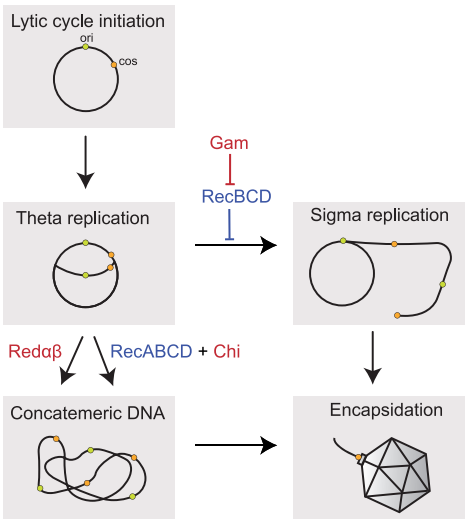

Figure 1. Implication of recombination in the replication of Lambda phage. Packaging of Lambda chromosomes requires concatemeric DNA. The induction of the lytic cycle leads to a number of rounds of theta replication (circle-to-circle). Concatemeric DNA can be formed directly from these newly replicated chromosomes by homologous recombination using the Red pathway, which requires the recombinase Redβ and the exonuclease Redα, or the host RecBCD pathway of recombination specifically enhanced by Chi sites. Concatemers can also be produced by rolling-circle (sigma) replication if the host RecBCD exonuclease is inhibited (e.g. by Gam encoded in Lambda). Concatemers are cleaved by the phage-encoded terminase at their *cos* sites (represented in orange) as they are packaged into the capsid. Lambda encoded sequences are indicated in red, the host encoded genes in blue. *ori* indicates the origin of replication.
doi:10.1371/journal.pgen.1003825.g001

phage needs to either block this exonucleolytic activity or produce concatemers by homologous recombination. Phage Lambda encodes its own homologous recombination toolkit under the form of a 3-genes operon [20]: *exo*, *bet* and *gam*, that encode Redα, Redβ and Gam respectively. Redα is a double strand specific 5′ to 3′ exonuclease and Redβ is a recombinase of the Rad52 superfamily that mediates strand annealing and exchange reactions starting from DNA extremities. Redβ and RecA (the host recombinase) have different recombination mechanisms, substrates and rates [21]. The protein Gam inhibits the host RecBCD exonuclease activity thus allowing efficient rolling-circle replication [22]. Thus, Lambda blocks the host recombination, superseding it with its own encoded recombination machinery. Other phages use evolutionarily related (e.g. Erf in P22) or unrelated recombinases (Sak4 in HK620, related to RecA) as well as other inhibitors of the exonucleolytic activity of RecBCD (e.g. Abc2 in P22 or gp5.9 in T7) [23,24].

Lambda and most of its mutants cannot produce concatemers from monomers using the host RecABCD pathway of homologous recombination because Gam inhibits RecBCD. When *gam* is experimentally inactivated, RecBCD prevents phage replication by degrading its genome. However, Lambda mutants that include a chromosomal sequence with the octamer Chi motif (GCTGGTGG) are viable [25]. This is because the destructive nuclease-helicase activity of RecBCD shifts to repair mode when it meets a Chi site by recruiting the RecA recombinase onto nascent Chi-containing ssDNA [26]. The single strand annealing protein RecA then promotes strand invasion and recombination. Chi sites are very abundant in *E. coli*, found in average every 5 kb, and much more frequently in the core genome than in recently acquired genes [27,28]. Chi sites are absent from the wild-type genome of Lambda and this prevents the use of RecBCD to produce phage concatemers. The high frequency of Chi in the *E. coli* genome and its rarity in Lambda and phage T4 led to the hypothesis that Chi is implicated in the discrimination between self and non-self and that the RecBCD-Chi system also functions to protect the genome from mobile genetic elements [29–31].

Phage fitness depends on its ability to control its host and on what it pays for that in terms of genome space and production costs [32]. Phages encoding their own recombination mechanisms gain an advantage by using proteins that co-evolved with the phage for a long period of time and are thus adapted to it in terms of processivity and tolerance to sequence divergence. However, the expression of recombination functions takes up resources. Encoding these functions also takes up genome space. Lambdoids rarely exceed 60 kb in size and most are between 40 kb and 50 kb [10]. This suggests the existence of an optimal size beyond which further accretion of genetic material lowers the phage fitness. Loss of the recombination module might facilitate acquisition of other functions with higher adaptive value in certain ecological contexts as long as recombination functions can be found in the host and manipulated by the phage. Increase in phage genome size might also be costly because of the replication cost and because such genomes require larger virions [33]. Phages that manipulate host recombination functions do not pay these additional costs, but they must use machineries adapted to their hosts. These proteins might not fit optimally the phage requirements and may have a cost in terms of host range. On the other hand, these mechanisms are well adapted to the host genetic background. Here, we study

phage recombination functions to understand how the dilemma between encoding and manipulating them shapes phage evolution.

Results

Chi sequences are abundant in lambdoids

We analyzed recombination functions encoded by lambdoid phages. These phages account for the majority *E. coli* prophages [10], and their recombination mechanisms have been thoroughly studied [18]. The classification of phages in the group of lambdoids is itself motivated by their ability to produce viable hybrids by recombination at high frequency. We identified Chi motifs in a set of 275 lambdoid phages of *Escherichia* and *Salmonella* (see Materials and Methods). We computed the expected number of the 8-nucleotide Chi motifs using four different statistical models: accounting for the frequency of nucleotides, tri-nucleotides, penta-nucleotides and hepta-nucleotides (see Materials and Methods). The different models gave concordant results (Table S1 and S2). We present the results for the tri-nucleotides model, which is the most adequate for the slightly degenerated Chi motif and the small genomes of phages (see Materials and Methods). We computed the number of Chi motifs observed/expected (O/E) ratio separately for each phage genome. Surprisingly, we found that, as a whole, lambdoids have more Chi motifs than expected (median O/E = 2.30, $p<0.0001$, Mann-Whitney test). In fact, most Lambda-like phages encode Chi motifs (85%), which are significantly more frequent in these phages than expected given sequence composition (median O/E = 2.43, $p<0.0001$, Mann-Whitney test). These results show that Chi sites are far from rare in phage genomes. In fact, they are much more abundant than expected given genome size and composition.

Phage recombinases and RecBCD inhibitors shape the abundance of Chi sites

Phage genomes lacking recombinases require the host machinery to engage in homologous recombination. To test the hypothesis that this leads to selection for the presence of Chi sites to recruit RecBCD, we detected phage recombinases using protein clustering and profile-profile alignments (see Materials and Methods). We identified a recombinase in 141 genomes of lambdoids, i.e. approximately half of our dataset (Rec^+ phages, 51%) (Figure 2A). Most of the identified recombinases (68%) are from the Redβ family, the one encoded by Lambda (Figure S1). Phage genomes lacking recombinases (Rec^- phages) display a significant over-representation of Chi sites (median O/E = 3.12, $p<0.0001$, Mann-Whitney test). These results are well in agreement with our hypothesis that phages lacking recombination functions select for the presence of Chi sites to recruit the host recombination machinery.

Phages encoding recombination functions but no RecBCD inhibitory functions could select for the presence of Chi motifs in their genomes to protect themselves from RecBCD exonucleolytic activity. To test this hypothesis, we searched for RecBCD inhibitors from the Gam and Abc2 families and identified 95 of these (see Materials and Methods). We found no single phage lacking a recombinase and encoding a RecBCD inhibitor. Red^-Gam^+ Lambda mutants are viable [19], showing that recombinases are not strictly required for phage replication when RecBCD is inhibited. On the other hand, RecBCD inactivation in the absence of phage recombinases has a very strong fitness cost in *E. coli* [34]. Cells where phages inhibit RecBCD without superseding it with their own recombinases lack tools to efficiently repair DNA double strand breaks. The fitness cost associated with this impairment might explain the lack of Rec^-Inh^+ phages in our dataset.

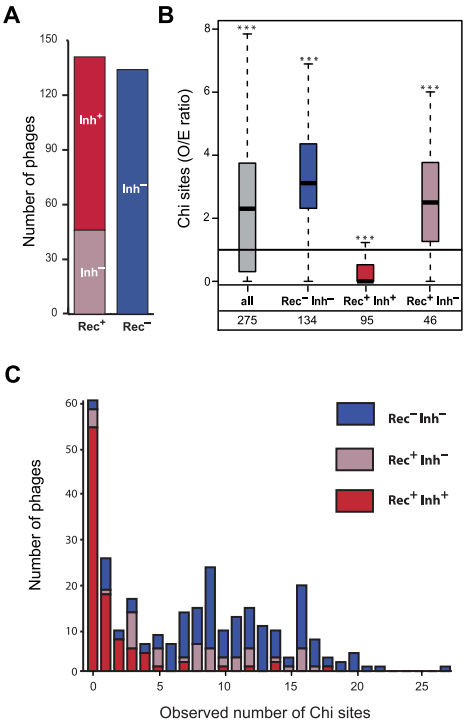

Figure 2. Association between the presence of phage recombination functions and the abundance of Chi sites. (*A*) Number of lambdoid phages encoding RecBCD inhibitors (Inh^+/Inh^-) and recombinases (Rec^+/Rec^-). (*B*) Distribution of the number of Chi sites observed/expected (O/E) ratios among lambdoid phages. Inh^+ and Inh^- indicate the presence or the absence of a RecBCD inhibitor protein respectively. For each box, the lower and upper horizontal edges represent respectively the first and the third quartile. The middle bar of each box indicates the median value. The central vertical lines indicate the data range, with a maximal distance of 1.5 interquartile ranges (i.e. the distance between the first and third quartile values). The number of phages is indicated for every class. For each class, we tested if the median value of the O/E ratio among phages was significantly different from 1 with the Mann-Whitney test (* $p<0.05$, ** $p<0.01$, *** $p<0.001$). (*C*) Number of Chi sites among lambdoid phage genomes lacking a recombinase (Rec^-) or encoding a recombinase with (Rec^+ Inh^+) or without (Rec^+ Inh^-) an inhibitor of RecBCD (Gam or Abc2). We found no phages Rec^- Inh^+.
doi:10.1371/journal.pgen.1003825.g002

We found 95 phage genomes encoding a recombinase and a recombination inhibitor (Rec^+Inh^+). Among Rec^+ phages, Inh^+ phages display a significant under-representation of Chi sites (median O/E = 0, $p<0.0001$, Mann-Whitney test), whereas Inh^- over-represent Chi motifs (median O/E = 2.50, $p<0.0001$, same test) (Figure 2B and 2C). Importantly, while both Rec^+Inh^- and Rec^- phages over-represent Chi, the latter show stronger over-representation ($p<0.03$, Wilcoxon test). Gam-like proteins inhibit

RecBCD activity, whereas Abc2-like RecBCD inhibitors subvert RecBCD functions rendering them Chi-insensitive [35]. We tested if phages encoding Gam-like RecBCD inhibitors showed different degrees of avoidance of Chi motifs relative to those encoding Abc2-like RecBCD inhibitors. While there is a slightly stronger avoidance of Chi sites in Abc2 encoding phages ($p = 0.030$, Wilcoxon test), both Gam-like and Abc2-like RecBCD inhibitors are strongly associated with Chi motifs under-representation (median O/E of 0.30 and 0 respectively, both $p<0.0001$, Mann-Whitney tests). Hence, phages encoding recombinases but not RecBCD inhibitors have more Chi sites than expected, whereas phages with RecBCD inhibitors strongly avoid Chi sites. This suggests that Rec^+Inh^- phages select for the presence of Chi sites, whereas Rec^+Inh^+ phages select for the absence of Chi sites. Phage Lambda is thus a typical representative of the Rec^+Inh^+ class of phages. These results show a strong link between the ability of a phage to inhibit the exonucleolytic activity of RecBCD and the presence or absence of Chi motifs.

Chi motifs in phages and their hosts

We compared the frequency of Chi motifs in phages and their hosts. As observed previously [27,28], Chi motifs are over-represented in the genomes of *E. coli* K12 and *S. enterica* Typhimurium (O/E = 2.29, $p<0.0001$ and O/E = 2.40, $p<0.0001$, Z score), and slightly more in the core genome of each species (resp. O/E = 2.36 and 2.38, both $p<0.0001$, same test, see Table S3 for the different models). The density of Chi sites in Rec^- phages is not significantly different from the host bacterial genome (0.2 Chi motifs/kb, $p = 0.103$, Mann-Whitney test). However, given their composition, Chi motifs are more over-represented in these phages than in the core genome of *E. coli* ($p<0.0001$, Mann-Whitney test). The over-representation of Chi sites in Rec^+Inh^- phages is not significantly different from that of the core genome of *E. coli* ($p = 0.30$, same test, see Table S4 for the other models). These results suggest that phages lacking RecBCD inhibitors endure similar or even stronger selection for Chi motifs than their hosts.

Some of the phages in our dataset were sequenced from virions whereas others were identified from bacterial chromosomes. We tested if inaccurate delimitation of the latter might have affected the number of Chi motifs found in our dataset. The median O/E number of Chi sites was not significantly different between Rec^- phages and Rec^- prophages (resp. 4.76 and 3.08, $p = 0.45$, Wilcoxon test). This ratio was almost indistinguishable among Rec^+Inh^- phages and prophages (resp. 2.42 and 2.52, $p = 0.58$, same test) and among Rec^+Inh^+ phages and prophages (both medians equal to 0, $p = 0.84$, same test). Thus, the trends we observe in the frequency of Chi motifs do not reflect biases associated with prophage detection. We also verified that Chi motifs in phages were not concentrated at the cargo region, typically at the edge of the element opposing the integrase [36]. Interestingly, we found that Chi motifs were concentrated far from this region and between the genes encoding the replication functions and the terminase, before the structural genes. In Lambda this corresponds to the region between the origin of replication (in gene O) and the *cos* site (before the terminase gene Nu1) where DNA is cut during packaging (Figure 3). The distribution of Chi sites along the chromosomes of Rec^+Inh^- phages and Rec^- phages is different ($p<0.0001$, Kolmogorov-Smirnov test). Chi motifs are more concentrated near the origin of replication of Rec^- phages, and towards the *cos* site in Rec^+Inh^- phages. These results show that Chi over-representation in lambdoids cannot result from inaccuracies in the delimitation of prophages and suggests a tight association between recombination, replication and packaging in phages.

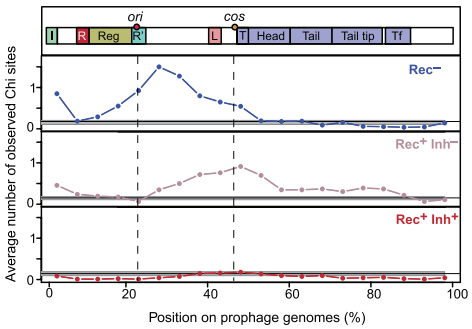

Figure 3. Distribution of Chi sites in lambdoid genomes. Lambdoid genomes (average length of 45 kb) were divided in 5% non-overlapping contiguous intervals (i.e. 2.25 kb). We plotted for each interval the average number of observed Chi motifs per phage. Phages were divided in three classes according to the encoded recombination functions: Rec^-, Rec^+Inh^- and Rec^+Inh^+. Integrases of the Tyrosine recombinase family were detected as in [10]. Each genome was polarized with the integrase on the left end (the few genomes lacking tyrosine recombinase integrases were discarded). A schematic representation of Lambda-like phages is given on top. Label "*ori*" indicates the median position of the homologs of Lambda gene O, which includes the origin of replication. The label "*cos*" indicates the median position of the start of the first terminase gene, which is where the *cos* site is located in Lambda.
doi:10.1371/journal.pgen.1003825.g003

Phage recombinases promote gene repertoire diversification and mosaicism

Recombination between different phages leads to genetic mosaicism and increases the diversity of gene repertoires. Redβ catalyzes recombination at higher rates and is more tolerant to sequence divergence than RecA [8]. We thus hypothesized that phages encoding recombination functions have more diverse gene repertoires. We built the pan genomes (i.e. the set of all different gene families) of Rec^+ and Rec^- lambdoids (see Materials and Methods). The pan genome of Rec^+ phages is systematically ~22% larger than the pan genome of Rec^- phages for the same number of genomes (Figure 4). This effect could not be explained by genome size, which is indistinguishable between the two types of phages (average of 45 kb, $p = 0.85$, Wilcoxon test). Hence, the permissivity of phage recombinases might allow faster diversification of gene repertoires in phages encoding their own recombination functions.

We then tested the hypothesis that these phages are also more mosaic, i.e. exchange homologous genes at higher rates. For this, we identified highly similar homologous genes present in highly dissimilar phage genomes (see Materials and Methods). This is a conservative subset of the genes that have recently undergone recombination between distinct phages. We restricted the analysis to the 163 Lambda-like phages of *E. coli* since broader taxonomic groups share too few homologous proteins for reliable inference of distances between phages. We computed the distance matrices between homologous proteins (d) and between phages (D) and identified proteins for which d is small and D is large using a range of thresholds T_d and T_D (see Materials and Methods). The results consistently show that genes with low d encoded in phages of high

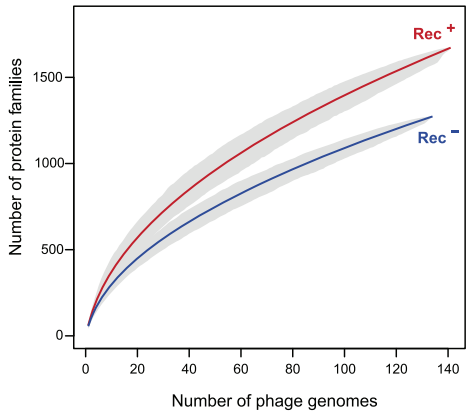

Figure 4. Pan genomes of the lambdoid phages encoding recombination functions (Rec$^+$) are larger than those lacking them (Rec$^-$). The pan genome size (*y*-axis) of each type of phage genome was computed for increasing numbers of genomes (x-axis). For each value of x we draw x genomes randomly and compute the pan genome. This is repeated 1000 times for each value of x to draw the 95% interval of confidence of the pan genome size (grey zone).
doi:10.1371/journal.pgen.1003825.g004

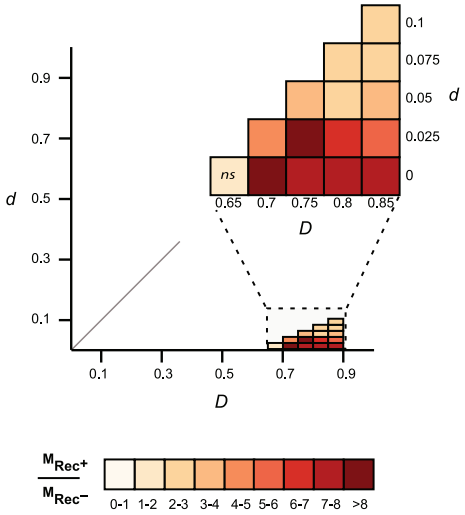

Figure 5. Comparison of gene mosaicism in 163 Lambda-like phages encoding (Rec$^+$) or lacking (Rec$^-$) recombinases. Mosaic genes are pairs of homologous genes with low evolutionary distances (low *d*) in phages with high evolutionary distances (high *D*). For each threshold T_d and T_D, we compared the frequency of mosaic genes of Rec$^+$ phages (M_{Rec}^+) and Rec$^-$ phages (M_{Rec}^-). The color scale gives the ratio of the frequency of mosaic genes between phages encoding and lacking recombination functions (M_{Rec}^+/M_{Rec}^-). Non-significant ($p>0.05$, Wilcoxon test) differences on the frequency of mosaic genes are indicated on the graph (*ns*).
doi:10.1371/journal.pgen.1003825.g005

D are very significantly over-represented in Rec$^+$ phages (Figure 5). Rec$^+$ phages have up to 8 times more such genes than Rec$^-$ phages and this difference is particularly high for the most recent transfers (corresponding to the lowest values of *d*). We tested if these results could be explained by the nature of the genes undergoing recombination. We analyzed the functional categories of the transferred genes (Text S2), and found no significant differences between them and the remaining genes ($p>0.1$, χ^2 test). We conclude that the higher mosaicism of phages encoding recombinases is independent of its phage gene repertoire size or content.

Discussion

In this work we studied the presence in phage genomes of genes and DNA motifs involved in homologous recombination. We showed that some phages encode a large number of Chi motifs and are thus able to manipulate RecBCD. This provides certain advantages. First, for similar genome size, and thus capsid volume, this allows the genome to encode other potentially adaptive functions. Second, Chi sites protect from the exonucleolytic activity of RecBCD and thus also from restriction-modification systems [37]. Third, RecABCD recombination is less frequent between very divergent sequences and could lead to fewer non-viable hybrid genomes. Finally, Chi motifs being important for genome maintenance, the presence of Chi in prophages might stabilize the element and lower its fitness cost for the host. Prophages make up to 35% of the pan genome of *E. coli* and we have shown that they encode motifs associated with their local context in the bacterial chromosome [10]. Hence, prophages with Chi motifs might integrate more seamlessly in the host chromosome.

Some phages encode their own recombination machinery, inhibit the host's and avoid Chi motifs. Recombination autonomous to the host machinery also has some advantages. First, recombination machineries co-evolving with the phage should be better adapted to its specificities, e.g. in terms of recombination frequency, sequence composition or homology requirements. For example, RecT, a Redβ homolog from prophage Rac, shows preference for AT rich regions [38], which are typical of phages. Second, reduced dependence on the host's machinery might broaden the range of possible hosts. Even if the composition of the machinery of homologous recombination is similar in most non-intracellular γ-Proteobacteria [39], the Chi motifs of *E. coli* and *Haemophilus influenzae* show a number of differences [40]. Hence, phages relying on host recombination functions may be at a disadvantage in a new host encoding different Chi motifs. Third, Red recombination is more permissive to sequence divergence and this may enlarge the mutational landscape of the phage, accelerating its diversification.

The dilemma of encoding or manipulating host recombination functions may also impact ecological interactions between mobile genetic elements. For example, the protein Old encoded by phage P2 targets Redβ [41] and the AbiK system of *Lactococcus lactis* plasmids targets different phage recombinase families [42]. On the other hand, encoding autonomous recombination functions may render the phage less susceptible to mobile elements that compete to manipulate host recombination. During co-infection, phages encoding RecBCD inhibitors might therefore have an important advantage over Chi-dependent phages by reducing the number of concatemeric chromosomes they can produce.

The chromosomes of *E. coli* strains are packed with prophages, some of which contribute to important adaptive functions. Different temperate phages may recombine in the bacterial cell. These cells may thus work as 'phage factories', releasing a wide variety of recombinant phages in the environment [43]. We have shown that phages carrying their own recombination functions have more mosaic genomes and larger pan genomes. The gene repertoires of bacteria are in constant genetic flux partly due to the action of phage transduction. For example, the recent epidemic of *E. coli* in Germany was the direct consequence of toxins encoded by prophages [44]. Adaptive associations between phage and bacteria can be very complex, e.g. a bacterial endosymbiont prophage protects aphids from parasitoid wasps [45]. As mentioned above, recombination is also important in the context of the ongoing arms races between phages and their hosts. Hence, the way phages recombine may impact their rates of diversification, but also those of their bacterial hosts.

The absence of Chi in phage Lambda was instrumental to the discovery of the function of this motif [46]. It was also interpreted as lack of selection for the presence of Chi sites in phages carrying their own recombination systems [29]. Here, we showed that contrary to common belief Chi sites are very abundant in most phages. Yet, these results also put forward a puzzling observation. RecBCD inhibitors render Chi sites useless either by blocking the activity of the protein or by rendering it insensitive to Chi. Hence, phages encoding RecBCD inhibitors should have a number of Chi sites close to the random expectation given sequence length and composition. Surprisingly, we show that these phages strongly avoid Chi sites, i.e. they have fewer sites than expected. Chi is thus selected *against* in phages encoding RecBCD inhibitors and *for* in the other phages. This suggests that carrying simultaneously Chi sites and RecBCD inhibitors is deleterious for the phage. We have no good explanation for these intriguing results at the moment. One might speculate that Chi sites affect the efficiency of RecBCD inhibitors, but this is at odds with the observation that the *E. coli* chromosome is packed with Chi motifs. Chi avoidance might be related to the chromosomal context of the prophage and how it affects chromosome maintenance processes, e.g. selection for recombination outside the prophage element to avoid chromosomal rearrangements [47]. But this would suggest that Chi are deleterious to integrative elements, which seems at odds with the large number of Chi sites found in the majority of prophages. Understanding selection against Chi sites will require further experimental work.

We showed that Chi sites in phages are concentrated between the origin of replication (especially in Rec^- phages) and the packaging sites (especially in Rec^+Inh^- phages). Naturally, the origin and *cos* (or *pac*) sites are unknown for the majority of phages and this result must be interpreted with care since it assumes that among lambdoids these positions are relatively unchanged. Nevertheless, the high density of Chi in the origin and packaging site regions, and the differences between the two regions in terms of phage recombination repertoires suggest some sort of selection for Chi sites in these locations. In fact, the very high frequency of Chi motifs in such a small region, up to three times the density in the *E. coli* core genome, might explain why this region is unusually variable among lambdoid genomes (the *nin* region [7,48]). The association between replication and recombination is pervasive in cellular organisms [1] and phages lacking recombinases might thus select for Chi sites near the origin of replication to process stalled replication forks. In phages encoding a recombinase able to process stalled replication forks, Chi sites might be more important for protection of free DNA ends from degradation by RecBCD than for its recruitment for recombination, explaining the fewer Chi sites and their location close to the packaging site in these phages. Hence, the study of the roles of Chi sites in phages might enlighten further functional associations between recombination, phage replication and packaging.

To check on the generality of our observations, we made some preliminary analyses of non-lambdoid *E. coli* phages in GenBank (Table S5 and Text S3). These analyses are hampered by the small dataset for each phage family and the lack of available information on the mechanisms of recombination in most genera. Yet, we could verify that phages requiring concatemers for packaging over-represent Chi motifs relative to phages able to encapsidate monomers ($p<0.0001$, Wilcoxon test). The two phage genera requiring concatemers for packaging and lacking recombinases (T5-like and P1-like) exhibit the strongest over-representation of Chi motifs (Table S5). The Chi abundance in P1-like phages shows that Chi sites can also be abundant in non-integrative temperate phages. T5 is a virulent phage showing that Chi over-representation is not limited to temperate phages. The reliable identification of presence or absence of specific RecBCD inhibitors is difficult in non-lambdoids because of the phage diversity and the tendency of RecBCD inhibitors to be small family-specific and fast-evolving proteins. Yet, these results suggest that Chi-dependent recombination might be widespread among phages packaging concatemeric DNA, for which recombination is important, even among virulent phages and non-integrative temperate phages.

Dilemmas between manipulation and supersession of host functions are probably common in viruses. For example, some phages encode tRNAs to complement the host's repertoire [49] and some filamentous phages encode their own secretion apparatus whereas others manipulate their host's secretion systems [50]. In fact, pathogenic bacteria or protozoa manipulating host functions might also face similar trade-offs [51]. Understanding why different parasites evolved to manipulate host functions or to encode their own, can provide important clues on their mechanisms of virulence and, as we showed, of their evolvability.

Materials and Methods

Genome data

The complete genomes of *Escherichia* (47 *E. coli*, 1 *E. fergusonii*) and *Salmonella* (20 *S. enterica* and 1 *S. bongori*) were downloaded from NCBI RefSeq (ftp://ftp.ncbi.nih.gov/genomes/). We analyzed a total of 275 phages including 38 lambdoid phages infecting enterobacteria (downloaded from RefSeq) and 237 long (>30 kb) non-redundant lambdoid prophages from the genomes of the abovementioned species identified in [10] with different mobile element detections [52–55] (see also Text S1 and Table S6). Among the 131 non-lambdoid genomes, 80 phage genomes of the *Caudovirales* order (69 virulent and 11 temperate) were downloaded from RefSeq (when classified in a genus defined by the ICTV). And 51 non-lambdoid prophages were identified in [10] with the same criteria (>30 kb and non-redundant).

Core and pan genomes

The core genomes of *E. coli* and *S. enterica* were computed as described previously [10]. The pan genomes were computed from the 141 Rec^+ lambdoid phages (9108 proteins), the 134 Rec^- lambdoid phages (7554 proteins), and also the 163 Lambda-like phages of *E. coli* (9856 proteins). Homologous proteins were defined as pairs of proteins with more than 40% sequence similarity, computed using a Needleman-Wunsch end gap free alignment algorithm with the BLOSUM62 matrix, and with less than 50% of difference in length. Protein families were built from

the pairwise analyses by transitivity, i.e. a protein is included in the family if it shares a relation of homology to a protein already in the family. The pan genome is the set of all different protein families. We excluded Genbank entries NC_004913, NC_004914 and NC_003525 from this analysis because their annotations over-predict the number of genes (nearly three times more genes per kilobase than the average lambdoid phage).

Identification of recombinases

We compared all lambdoid phage proteins to each other using blastp (*e*-value$<$0.001). The resulting blast bit score was used to cluster the proteins with MCL [56]. After testing the MCL inflation parameters in the range [1.2 to 5.0], we used I = 3.0 because it was the smallest that produced protein clusters where all proteins of each cluster could be analyzed in a single multiple alignment. A total of 1812 protein clusters were obtained for the 16662 proteins analyzed. We aligned the proteins of each cluster with MUSCLE v3.6 [57] and built protein profiles with the HH-suite v2.0.9 [58]. The protein profiles of recombinases were initially found by comparison with published profiles [24] using HHsearch (profile-profile comparison, p$>$95% in local and global alignments and $>$50% of profile coverage). We identified initially a subset of 14 protein clusters. To exclude helicases with ATPase domains from recombinases [24] we also made profile-profile comparisons with PFAM-A profiles (downloaded the 11/25/2011) using HHsearch (same parameters). We excluded the clusters matching PFAM-A profiles annotated as helicases (e.g. DnaB, helicase-ATPase domain, DEAD/DEAH box helicase, PIF1-like helicase), producing a final set of 8 protein clusters of recombinases. This corresponds to 141 proteins found in 141 lambdoids. Our procedure was able to find all of the recombinases previously identified in lambdoid phages of enterobacteria [24].

Identification of RecBCD inhibitors

We searched lambdoid phage genomes for hits of PFAM profiles of Gam (PF06064) and Abc2 (PF11043) proteins using HMMER v3.0 (*c*-value$<10^{-5}$) [59]. A total of 95 RecBCD inhibitors were detected: 56 Gam and 39 Abc2 proteins. The families of RecBCD inhibitors from T7 (gp5.9, NP_041987), *Enterococcus* phage BC-611 (ORF41, BAM44931), *Clostridium* phage phi8074-B1 (phi8074-B1_00044, AFC61976) and the DNA end protector from T4 (gp2, NP_049754), have not been described among enterobacterial temperate phages. Indeed, we found no significant BLASTP hits (at a threshold of *e*-value$<$0.001) to these proteins in our dataset of lambdoid phages.

Detection of Chi motifs

We used R'MES v3.1.0 to search the non-degenerated Chi motif 5′GCTGGTGG3′ and to compute significance of Z scores [60]. We computed the number of expected and observed Chi motifs accounting for the oligonucleotide composition separately for each genome. This was done to avoid putting together different phage genomes, which differ extensively in terms of nucleotide composition [61]. Four statistical models were analyzed for each genome. 1) The simplest model (M0) accounts only for nucleotide composition. 2) The M2 model accounts for the composition in tri-nucleotides. 3) The M4 model accounts for the composition in penta-nucleotides. 4) The maximal model (M6) accounts for the frequency of the maximal sub-strings of Chi motifs, i.e. hepta-nucleotides. The four models produced concordant statistics (Table S1). The M0 model is a poor predictor of random usage of large oligonucleotides because these are also affected by selection on other smaller oligonucleotides such as codons [62]. Phage genomes are small ($<$50 kb on average) and the Chi motif is slightly degenerated [63]. These two traits hinder the statistical power of the M6 and M4 models. Therefore we show in the text the results of the M2 model. The statistical significance of Chi sites over or under-representation in a given set of phages was computed using the Mann-Whitney test. Chi sites over-representation per genome was assessed by the Z score computed with R'MES. We computed all models under the compound Poisson approximation that is more adequate for low counts [60].

Analysis of gene mosaicism

We initially aimed at using classical phylogenetic approaches to identify recombination events. Unfortunately, no proteins are ubiquitous to the whole set of 163 Lambda-like phages of *E. coli*. We therefore designed a method to find highly similar pairs of homologous proteins in two otherwise distantly related phages, which are likely the result of recent recombination events (mosaic genes). This approach resembles closely that of [64]. First, we constructed the multiple alignment of each protein family of the pan genome of Lambda-like phages of *E. coli* with MUSCLE v3.6 [57]. Second, we extracted the informative positions in the alignments using BMGE with the BLOSUM30 matrix [65]. The 19 (4%) protein families with trimmed alignments shorter than 50 sites were excluded due to the lack of phylogenetic signal. Third, we computed the protein distances ($d_{i,j}^{P}$) of each pair of homologous proteins between two phages i and j in every protein family using TREE-PUZZLE v5.2 [66]. The distance matrix was computed using maximum likelihood under automatic estimation of the best substitution model and a $\Gamma(8)$ correction for rate heterogeneity. Fourth, the distance matrix between phages $D_{i,j}$ was defined as the mean value of $d_{i,j}$ for the orthologs shared by each pair of phages i and j. For each pair of phages, orthologous proteins were defined as unique reciprocal best hits with more than 40% similarity in amino acid sequence and less than 50% of difference in protein length. Finally, mosaic genes were identified as the ones encoding highly similar homologous proteins in highly dissimilar genomes for different thresholds T_d and T_D. More precisely, a pair of homologous genes between two phages i and j was regarded as mosaic if the encoded proteins were closey related ($d_{ij}<T_d$) and the two phages were distantly related ($D_{i,j}>T_D$). The different thresholds tested T_d and T_D showed qualitatively similar results. We did not analyze recombination events in genes encoding recombination functions, because they are absent from Rec$^-$ phages. We also ignored transposable elements, because they are self-mobilizable.

Supporting Information

Figure S1 Recombinase families identified in lambdoid phages. Recombinases were identified by profile-profile comparisons with HHsearch (see Materials and Methods). Most of the identified recombinases belong to the Rad52 superfamily (Redβ, Erf and Sak). Sak4 recombinases are part of the Rad51 superfamily and are remote homologs of RecA [24]. Gp2.5 represents the last superfamily of phage recombinases and is found much more frequently in virulent phages [24].
(EPS)

Table S1 Chi sites Observed/Expected ratio for lambdoid phages and their bacterial hosts computed with models M0, M4 and M6. The expected number of Chi sites has been determined with three additional models: M6, M4 and M0. For each category, we tested if the ratio O/E of Chi composition in the set of phages was significantly different from random expectation (O/E = 1) with the Mann-Whitney test.
(XLS)

Table S2 Chi sites Z score statistics for lambdoid phages and their bacterial hosts. The expected number of Chi sites has been determined with the M2 model. For each category, we tested if the Z score of Chi composition in the set of phages was significantly different from random expectation (Z = 0) with the Mann-Whitney test. The "Skew" column indicates if the phage category over-represents (+) or under-represents (−) Chi sites.
(XLS)

Table S3 Chi sites Observed/Expected ratio for *E. coli* and *S. enterica* with models M0, M4 and M6. The expected number of Chi sites has been determined with three additional models: M6, M4 and M0. For each core or complete genome, we tested if the Chi composition was significantly different from random expectation with the Z score. The analysis was run on *E. coli* K12 MG1655 and *S. enterica* Typhimurim LT2 genomes respectively.
(XLS)

Table S4 Comparison of the Chi sites Observed/Expected ratio of *E. coli* lambdoid phages and all lambdoid phages to the Chi sites Observed/Expected ratio of *E. coli* core genome with models M0, M4 and M6. The median value "M" of the Chi sites Oberved/Expected ratio is given for lambdoid coliphages and for all lambdoid phages for each category. For each category and model, we tested if the Chi composition was significantly different from the Chi composition of *E. coli*'s core genome with the Mann-Whitney test. The analysis has been done on the core genes of *E. coli* K12 MG1655.
(XLS)

Table S5 Chi sites Observed/Expected ratio and Z scores for different genera of phages and prophages infecting enterobacteria. We used the non-lambdoid phage genera of the *Caudovirales* order defined by the ICTV. Prophages were identified and classified as in [10]. Phage's life style, i.e. virulent (v) and temperate (t) is indicated in the "Type" column. The type of DNA substrate used for packaging, i.e. concatemeric (C) of monomeric (M) is indicated in the "Packaging" column.
(XLS)

Table S6 Description of lambdoid phages and prophages. For each phage (ph) or prophage (pro) used for the analysis, the RefSeq ID of the host genome is given for prophages and the RefSed ID is given for phage genomes directly. The recombinase family (Redβ, ERF, Sak, Sak4 and Gp2.5) is indicated in the "Rec" column. The type of RecBCD inhibitor (Gam or Abc2) is indicated in the "Inh" column. The expected number of Chi sites is given for the tri-nucleotides model (M2). For each genome, Chi composition significantly different from random expectation is given by the "*pvalue* (Z score)" column. The GC content of each phage or prophage genome is given in the "GC" column.
(XLS)

Text S1 Identification and classification of prophages.
(DOC)

Text S2 Function assignment.
(DOC)

Text S3 Detection of Chi sites in non-lambdoid phages.
(DOC)

Acknowledgments

We thank Miriem El Karoui, David Leach and Ivan Matic for discussions, suggestions and criticisms.

Author Contributions

Conceived and designed the experiments: LMB MT EPCR. Performed the experiments: LMB. Analyzed the data: LMB MT EPCR. Wrote the paper: LMB MT EPCR.

References

1. Michel B, Grompone G, Flores MJ, Bidnenko V (2004) Multiple pathways process stalled replication forks. Proc Natl Acad Sci USA 101: 12783–12788.
2. Perals K, Capiaux H, Vincourt JB, Louarn JM, Sherratt DJ, et al. (2001) Interplay between recombination, cell division and chromosome structure during chromosome dimer resolution in Escherichia coli. Mol Microbiol 39: 904–913.
3. Barton NH, Charlesworth B (1998) Why sex and recombination? Science 281: 1986–1990.
4. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, et al. (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A 98: 182–187.
5. Ochman H, Lerat E, Daubin V (2005) Examining bacterial species under the specter of gene transfer and exchange. Proc Natl Acad Sci U S A 102 Suppl 1: 6595–6599.
6. Hendrix RW, Smith MCM, Burns RN, Ford ME, Hatfull GF (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci USA 96: 2192–2197.
7. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, et al. (2000) Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol 299: 27–51.
8. Martinsohn JT, Radman M, Petit MA (2008) The lambda Red proteins promote efficient recombination between diverged sequences: Implications for bacteriophage genome mosaicism. PLoS Genet 4: e1000065.
9. Botstein D (1980) A theory of modular evolution for bacteriophages. Ann N Y Acad Sci 354: 484–490.
10. Bobay LM, Rocha EP, Touchon M (2013) The Adaptation of Temperate Bacteriophages to Their Host Genomes. Mol Biol Evol 30: 737–751.
11. Campbell A, Botstein D (1983) Evolution of the lambdoid phages. In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA, editors. Lambda II. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory pp. 365–380.
12. Casjens SR (2008) Diversity among the tailed-bacteriophages that infect the Enterobacteriaceae. Res Microbiol 159: 340–348.
13. Casjens S, Hendrix R (1974) Comments on the arrangement of the morphogenetic genes of bacteriophage lambda. J Mol Biol 90: 20–25.
14. Kashiwagi A, Yomo T (2011) Ongoing phenotypic and genomic changes in experimental coevolution of RNA bacteriophage Qbeta and Escherichia coli. PLoS Genet 7: e1002188.
15. Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R (2007) Genes under positive selection in Escherichia coli. Genome Res 17: 1336–1343.
16. Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, et al. (2010) Antagonistic coevolution accelerates molecular evolution. Nature 464: 275–278.
17. Weinbauer MG (2004) Ecology of prokaryotic viruses. FEMS Microbiol Rev 28: 127–181.
18. Smith GR (1983) General Recombination. In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA, editors. Lambda II. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory. pp. 175–210.
19. Enquist LW, Skalka A (1973) Replication of Bacteriophage-Lambda DNA-Dependent on Function of Host and Viral Genes .1. Interaction of Red, Gam and Rec. J Mol Biol 75: 185–212.
20. Kuzminov A (1999) Recombinational repair of DNA damage in Escherichia coli and bacteriophage lambda. Microbiol Mol Biol Rev 63: 751–813.
21. Maresca M, Erler A, Fu J, Friedrich A, Zhang YM, et al. (2010) Single-stranded heteroduplex intermediates in lambda Red homologous recombination. BMC Mol Biol 11: 54.
22. Unger RC, Clark AJ (1972) Interaction of the recombination pathways of bacteriophage lambda and its host Escherichia coli K12: effects on exonuclease V activity. J Mol Biol 70: 539–548.
23. Iyer LM, Koonin EV, Aravind L (2002) Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52. BMC Genomics 3: 8.
24. Lopes A, Amarir-Bouhram J, Faure G, Petit MA, Guerois R (2010) Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. Nucleic Acids Res 38: 3952–3962.
25. Myers RS, Stahl FW (1994) Chi and the RecBC D enzyme of Escherichia coli. Annu Rev Genet 28: 49–70.
26. Dillingham MS, Kowalczykowski SC (2008) RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiology and molecular biology reviews : MMBR 72: 642–671.
27. El Karoui M, Biaudet V, Schbath S, Gruss A (1999) Characteristics of Chi distribution on different bacterial genomes. Res Microbiol 150: 579–587.

28. Halpern D, Chiapello H, Schbath S, Robin S, Hennequet-Antier C, et al. (2007) Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling. PLoS Genet 3: 1614–1621.
29. Kuzminov A, Schabtach E, Stahl FW (1994) Chi sites in combination with RecA protein increase the survival of linear DNA in Escherichia coli by inactivating exoV activity of RecBCD nuclease. Embo J 13: 2764–2776.
30. Anderson DG, Kowalczykowski SC (1998) Reconstitution of an SOS response pathway: derepression of transcription in response to DNA breaks. Cell 95: 975–979.
31. Kobayashi I (1998) Selfishness and death: raison d'être of restriction, recombination and mitochondria. Trends Genet 14: 368–374.
32. Bull JJ, Badgett MR, Springman R, Molineux IJ (2004) Genome properties and the limits of adaptation in bacteriophages. Evolution 58: 692–701.
33. De Paepe M, Taddei F (2006) Viruses' Life History: Towards a Mechanistic Basis of a Trade-Off between Survival and Reproduction among Phages. PLoS Biol 4: e193.
34. Capaldo FN, Ramsey G, Barbour SD (1974) Analysis of Growth of Recombination-Deficient Strains of Escherichia-Coli-K-12. J Bacteriol 118: 242–249.
35. Murphy KC (2012) Phage Recombinases and Their Applications. Advances in Virus Research, Vol 83: Bacteriophages, Pt B 83: 367–414.
36. Thomson N, Baker S, Pickard D, Fookes M, Anjum M, et al. (2004) The role of prophage-like elements in the diversity of Salmonella enterica serovars. J Mol Biol 339: 279–300.
37. Handa N, Ichige A, Kusano K, Kobayashi I (2000) Cellular responses to postsegregational killing by restriction-modification genes. J Bacteriol 182: 2218–2229.
38. Noirot P, Gupta RC, Radding CM, Kolodner RD (2003) Hallmarks of homology recognition by RecA-like recombinases are exhibited by the unrelated Escherichia coli RecT protein. Embo J 22: 324–334.
39. Rocha EPC, Cornet E, Michel B (2005) Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems. PLoS Genet 1: e15.
40. Source S, Biaudet V, Karoui ME, Ehrlich SD, Gruss A (1998) Identification of the Chi site of Haemophilus influenzae as several sequences related to the Escherichia coli Chi site. Mol Microbiol 27: 1021–1029.
41. Myung H, Calendar R (1995) The Old Exonuclease of Bacteriophage-P2. J Bacteriol 177: 497–501.
42. Bouchard JD, Moineau S (2004) Lactococcal phage genes involved in sensitivity to AbiK and their relation to single-strand annealing proteins. J Bacteriol 186: 3649–3652.
43. Ohnishi M, Kurokawa K, Hayashi T (2001) Diversification of Escherichia coli genomes: are bacteriophages the major contributors? Trends Microbiol 9: 481–485.
44. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, et al. (2011) Origins of the E. coli Strain Causing an Outbreak of Hemolytic-Uremic Syndrome in Germany. N Engl J Med 365: 709–717.
45. Oliver KM, Degnan PH, Hunter MS, Moran NA (2009) Bacteriophages Encode Factors Required for Protection in a Symbiotic Mutualism. Science 325: 992–994.
46. Stahl FW (2005) Chi: a little sequence controls a big enzyme. Genetics 170: 487–493.
47. Canchaya C, Fournous G, Brussow H (2004) The impact of prophages on bacterial chromosomes. Mol Microbiol 53: 9–18.
48. Hendrix RW, Casjens S (2006) Bacteriophage Lambda and its Genetic Neighborhood. In: Abedon ST, Calendar RL, editors. The Bacteriophages. 2nd ed. New York: Oxford University Press. pp. 409–447.
49. Bailly-Bechet M, Vergassola M, Rocha E (2007) Causes for the intriguing presence of tRNAs in phages. Genome Res 17: 1486–1495.
50. Davis BM, Lawson EH, Sandkvist M, Ali A, Sozhamannan S, et al. (2000) Convergence of the secretory pathways for cholera toxin and the filamentous phage, CTX phi. Science 288: 333–335.
51. Brown SP (2005) Do all parasites manipulate their hosts? Behav Process 68: 237–240.
52. Fouts DE (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res 34: 5839–5851.
53. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS (2011) PHAST: a fast phage search tool. Nucleic Acids Res 39: W347–352.
54. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics 24: 863–865.
55. Touchon M, Rocha EP (2007) Causes of insertion sequences abundance in prokaryotic genomes. Mol Biol Evol 24: 969–981.
56. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.
57. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
58. Remmert M, Biegert A, Hauser A, Soding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9: 173–175.
59. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7: e1002195.
60. Schbath S, Hoebeke M (2011) R'MES: a tool to find motifs with a significantly unexpected frequency in biological sequences. Elnitski L, Piontkivska O, Welch L, editors. Singapore: World Scientific.
61. Rocha EPC, Danchin A (2002) Competition for scarce resources might bias bacterial genome composition. Trends Genet 18: 291–294.
62. Schbath S, Prum B, Turckheim Ed (1995) Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. J Comput Biol 2: 417–437.
63. Cheng KC, Smith GR (1987) Cutting of chi-like sequences by the RecBCD enzyme of Escherichia coli. J Mol Biol 194: 747–750.
64. Novichkov PS, Omelchenko MV, Gelfand MS, Mironov AA, Wolf YI, et al. (2004) Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J Bacteriol 186: 6575–6585.
65. Criscuolo A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10: 210.
66. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.

Conclusions et perspectives

Cette étude a mis en évidence l'utilisation de deux stratégies de recombinaison au sein d'un même groupe de phages: les phages lambdoïdes. Certains phages codent pour leur propre système de recombinaison. D'autres phages semblent manipuler le système de recombinaison de l'hôte via la présence de motifs Chi au sein de leurs génomes. Enfin, certains phages semblent utiliser une stratégie intermédiaire où la présence de sites Chi semble nécessaire pour inhiber l'activité exonucléase de RecBCD mais codent pour leur propre recombinase. Ironiquement, la majorité des phages Lambda-like présentent des sites Chi alors que l'absence de sites Chi chez Lambda a justement permis la découverte de ces motifs et de leur fonction (Stahl 2005). En 2005, Franklin Stahl écrivait: "Musing: If wild-type Lambda (48.5kb) contained one GCTGGTGG (which occurs every 5kb in *E. coli*), Chi might still be undiscovered" (Stahl 2005). Ces observations permettent d'étendre un peu plus notre connaissance du fonctionnement des phages lambdoïdes et ouvrent de nouvelles perspectives.

i) Ces résultats ont montré que les phages codant à la fois pour leur propre recombinase et leur inhibiteur de RecBCD sous-représentent significativement les sites Chi. Les raisons de cet évitement des sites Chi par ces phages ne sont pas claires. Ceci suggère qu'il y a un antagonisme fonctionnel entre l'utilisation de sites Chi et l'utilisation d'un inhibiteur de RecBCD.

ii) Il a été montré expérimentalement que l'inhibition de RecBCD par Gam permettait à un mutant de Lambda déficient pour Redβ de répliquer et d'encapsider son génome (Enquist and Skalka 1973; Smith 1983). Pourtant les résultats de notre analyse n'ont révélé aucun génome de phages lambdoïdes codant pour un inhibiteur de RecBCD en l'absence de recombinases. Bien que cette stratégie semble fonctionnelle *in vitro*, son absence parmi l'ensemble des phages et prophages étudiée suggère que l'inhibition de RecBCD par Gam n'est pas un mécanisme viable à long terme *in vivo* sans recombinase phagique. Cela permet de proposer l'hypothèse que la recombinase serait donc systématiquement requise *in vivo* pour la formation de concatémères chez les phages qui n'utilisent pas RecBCD. L'inhibition de RecBCD ne serait pas suffisante et surviendrait alors uniquement lorsque le phage utilise son propre système de recombinaison.

iii) Les lambdoïdes dépourvus de recombinases et utilisant des sites Chi pour recombiner présentent une forte concentration de sites Chi au sein de la région accessoire *nin* de leur

génome. Cette région est située entre l'origine de réplication et le site de packaging. Elle apparaît donc comme une zone privilégiée de recombinaison et d'inhibition de RecBCD pour satisfaire la fonction de la recombinaison chez ces phages: promouvoir la formation de concatémères nécessaires à l'encapsidation. La forte variabilité génétique et la faible fréquence de gènes essentiels dans cette région peuvent être mises en relation avec la recombinaison. En effet, la recombinaison pourrait favoriser l'acquisition et la formation de nouveaux gènes dans cette région. De plus, il est probable que la nécessité fonctionnelle de recombiner dans cette région façonne l'architecture des génomes lambdoïdes. Un taux élevé de recombinaison dans la région *nin* pourrait conduire à la contre-sélection des gènes essentiels dans cette région. En effet, la recombinaison homologue chez ces phages étant peu spécifique (Martinsohn, et al. 2008), elle pourrait conduire à accroître le taux de mutation dans la région *nin*. Il pourrait être intéressant de comparer expérimentalement les taux de mutation et de recombinaison entre cette région et les autres régions des génomes de lambdoïdes.

III Domestication des prophages défectueux par les bactéries.

Contexte

L'intégration et la dégradation des phages tempérés au sein de leurs hôtes sont à l'origine du renouvellement d'une partie importante de la diversité génétique bactérienne. De nombreux prophages défectifs peuvent être identifiés dans les génomes bactériens et témoignent de ce phénomène (Asadulghani, et al. 2009; Wang, et al. 2010). La dynamique de dégradation de ces séquences reste très peu étudiée. Ce renouvellement important de séquences phagiques au sein des génomes bactériens permet également la domestication occasionnelle de ces séquences. Plusieurs exemples de systèmes phagiques domestiqués ont été décrits (section 3.3.2). Cependant, la fréquence de ce phénomène et la fonction des prophages domestiqués sont largement inconnues.

L'objectif de ce dernier travail a été de comprendre la dynamique de dégradation des prophages. Comment sont-ils dégradés au sein des génomes hôtes? Les événements de délétion et de pseudogénisation affectent-ils différemment les diverses fonctions phagiques? Enfin, certains gènes ou groupes de gènes sont-ils conservés par l'hôte? Peut-on détecter des séquences de prophages domestiquées?

Approche

Détection de prophages hérités verticalement.

La difficulté majeure de cette étude repose sur la détection de prophages issus d'un même événement d'intégration. Ces prophages issus d'un même événement d'intégration ont été nommés "prophages orthologues". A l'inverse, les prophages issus d'intégrations indépendantes ont été qualifiés de "prophages non orthologues". Parce qu'ils ont évolué au sein du génome hôte, les prophages orthologues doivent présenter certaines caractéristiques permettant de les identifier. J'ai utilisé ici une méthode de détection des prophages orthologues basée sur quatre critères (Fig 20).

i) Les prophages orthologues proviennent du même événement d'intégration. Ces éléments doivent donc être intégrés au même locus. J'ai donc utilisé, comme précédemment, les gènes core d'*E. coli* et de *S. enterica* pour définir ces loci. L'ordre de ces gènes core a été basé

arbitrairement par rapport à celui d'une souche d'*E. coli* et d'une souche de *S. enterica*. Je n'ai pas retenu les éléments situés au niveau de réarrangements chromosomiques (entre des gènes core de l'hôte qui sont consécutifs dans la souche de référence mais non consécutifs dans la souche où le prophage a été identifié).

ii) Les prophages orthologues doivent présenter une forte similarité de répertoires de gènes. Ils peuvent cependant présenter des événements de délétion et de pseudogénisation. J'ai ainsi utilisé le score (R), comme défini pour la classification, afin d'estimer la similarité générale des prophages. Ce score est robuste aux délétions et m'a permis de distinguer les prophages divergents intégrés à un même locus

iii) Les prophages orthologues ont évolué au sein de leur chromosome hôte. Ceci implique donc qu'ils doivent présenter un taux de substitutions synonymes proche de celui de leurs hôtes. J'ai donc comparé les valeurs de dS de chaque paire de prophages aux valeurs de dS des gènes core de la paire de génomes hôtes correspondants. Les prophages présentant des taux de dS trop faibles ou trop importants par rapport à leurs hôtes ont été exclus.

iv) Chaque famille de prophages orthologues est issue d'un même génome ancestral. La diversité totale des répertoires de gènes de chaque famille de prophages orthologues ne doit donc pas excéder la diversité de gènes du prophage ancestral. J'ai donc construit le génome pan de chaque famille de prophages orthologues. J'ai alors comparé la diversité génétique totale de chaque famille de prophages orthologues au nombre de gènes du plus grand prophage de la famille. Les familles présentant une diversité génétique très élevée par rapport au plus grand élément de la famille ont alors été éliminées ou subdivisées à l'aide d'une phylogénie moléculaire réalisée sur un concaténât des gènes des prophages.

L'ensemble de ces critères m'a permis de définir des familles de prophages potentiellement hérités verticalement. Je me suis ensuite intéressé à l'évolution des séquences de ces prophages orthologues.

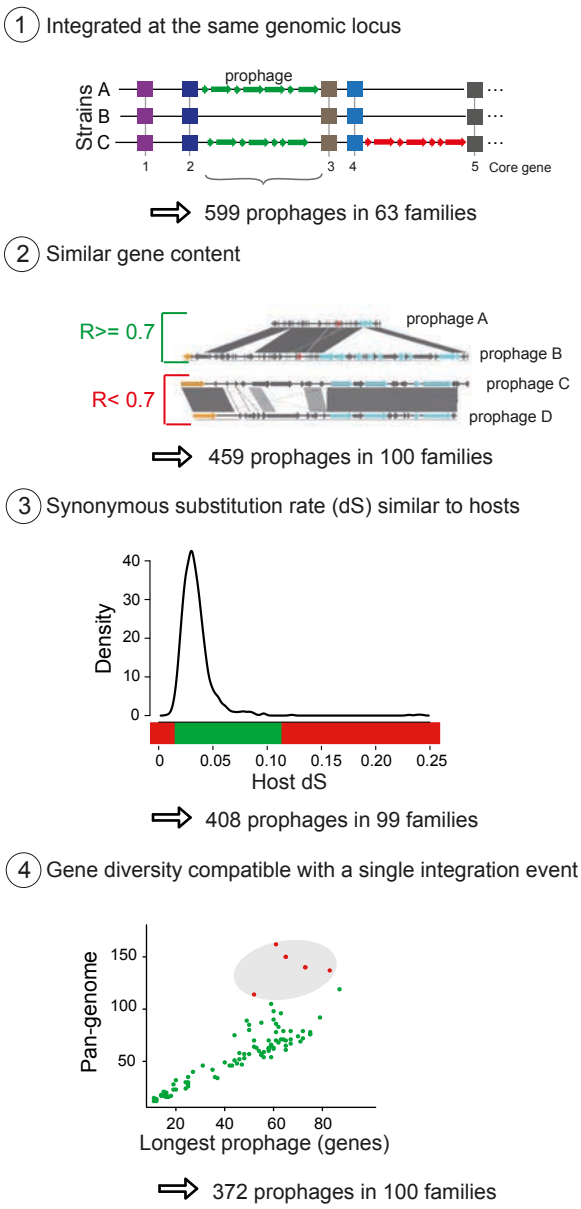

Figure 20: Procédure d'identification des prophages hérités verticalement (Article 3, Fig S3).

Détection des éléments sous pression sélective

Parmi les familles de prophages orthologues, j'ai calculé les ratios dN/dS de chaque paire de gènes orthologues avec la méthode yn00 implémentée dans PAML (Yang and Nielsen 2000). Certains prophages ayant de faibles taux de divergence, il est probable que la présence de nombreux gènes avec un très faible ratio dN/dS soit liée à des valeurs nulles de dN. J'ai donc également calculé les ratios dN/dS sur le concaténât des gènes orthologues partagés par

chaque paire de prophages orthologues afin d'obtenir un signal plus robuste (Rocha, et al. 2006b). Enfin, de tels ratios pouvant être très bruités en cas de faibles valeurs, j'ai vérifié que les faibles ratios dN/dS ne sont pas dus aux éléments présentant un faible taux de divergence. L'obtention d'un ratio dN/dS assez constant pour les différentes valeurs de dS, suggère que le signal obtenu n'est pas un artefact lié à une faiblesse du signal.

Article 3

Pervasive domestication of defective prophages by bacteria

Louis-Marie Bobay[1,2,3,*], Marie Touchon[1,2], Eduardo PC Rocha[1,2]

[1]Microbial Evolutionary Genomics, Institut Pasteur, 75724 Paris, France

[2]CNRS, UMR3525, 75724 Paris, France

[3]Sorbonne universités, UPMC Univ Paris06, IFD, 4 place Jussieu, 75252 Paris cedex05, France.

* Corresponding author. Tel: 33 1 45 68 87 68; Fax: 33 1 44 27 97 79; Email: lbobay@pasteur.fr; Present Address: Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris

Keywords: Evolution, prokaryotes, viruses, comparative genomics.

Abstract

Integrated phages (prophages) are major contributors to the diversity of bacterial gene repertoires. Domestication of their components is thought to have endowed bacteria with molecular systems involved in secretion, defense, warfare and gene transfer. However, the rates and mechanisms of domestication remain unknown. We used comparative genomics to study the evolution of prophages within the bacterial genome. We identified over 300 vertically inherited prophages within enterobacterial genomes. Some of these elements are very old and might pre-date the split between *Escherichia coli* and *Salmonella enterica*. The size distribution of prophage elements is bimodal; suggestive of rapid prophage inactivation followed by much slower gene degradation. Accordingly, we observed a pervasive pattern of systematic counter-selection of non-synonymous mutations in prophage genes. Importantly, such patterns of purifying selection are observed not only on accessory regions, but also in core phage genes, such as those encoding structural and lysis components. This suggests that bacterial hosts select for phage-associated functions. Several of these conserved prophages have gene repertoires compatible with described functions of adaptive prophage-derived elements such as bacteriocins, killer particles, gene transfer agents or satellite prophages. We suggest that bacteria frequently domesticate their prophages. Most such domesticated elements end-up deleted from the bacterial genome because they are replaced by analogous functions carried by new prophages. This puts the bacterial genome in a state of continuous flux of acquisition and loss of phage-derived adaptive genes.

Significance

Several molecular systems with important adaptive roles have originated from the domestication of integrated phages (prophages). However, the evolutionary mechanisms and extent of prophage domestication remains poorly understood. In this work, we detected several hundred prophages originating from common integration events and described their dynamics of degradation within their hosts. Surprisingly, we observed strong conservation of the sequence of most vertically inherited prophages, including selection for genes encoding phage-specific functions. These results suggest pervasive domestication of parasites by the bacterial hosts. Since prophages account for a large fraction of bacterial genomes, phage domestication may drive bacterial adaptation.

Introduction

The ubiquity and abundance of bacteriophages (or phages) makes them key actors in bacterial population dynamics (1). While all phages are able to propagate horizontally between cells, temperate phages also propagate vertically in bacterial (lysogenic) lineages, typically by integrating into the bacterial chromosome as prophages. Very few genes are expressed in the prophage, which replicates with the bacterial chromosome (2). The evolutionary interests of integrated phages (prophages) are partly aligned with those of the host chromosome since rapid proliferation of the later effectively increases prophage population. Accordingly, prophages protect the host against further phage infection (3), from phagocytosis (4) and provide bacterial pathogens with virulence factors (5). Temperate phages also encode accessory genes that increase host fitness under certain conditions, such as increased growth under nutrient limitation (6), biofilm formation (7), and antibiotic tolerance (8). Some prophages provide bacteria with regulatory switches (9). Functional prophages might also be used as biological weapons by lysogens, since their induction can counteract or delay colonization by non-lysogens (10-12). High diversity and high turnover of temperate phages result in a constant input of new genes in the host genome (13, 14). For example, *E. coli* prophages contribute to more than 35% of the gene diversity (pan-genome) of the species (15). Most temperate phages integrate into a few very specific and conserved integration hotspots in chromosomes and their sequences are adapted to the local frequency of DNA motifs, suggesting adaptation of the phage sequence to the requirements of the prophage state (15).

Independently of occasional contributions to bacterial fitness, intact prophages are molecular time bombs that kill their hosts upon activation of the lytic cycle (2). It has been shown that bacterial pseudogenes are under selection for rapid deletion from bacterial genomes (16, 17). Prophage inactivation should be under even stronger selection because these elements can kill the cell. One might thus expect rapid genetic degradation of prophages: either they activate the lytic cycle and kill the cell before accumulating inactivating mutations or they are irreversibly degraded and deleted from the host genome. Bacterial chromosomes have numerous cryptic (defective) prophages and other prophage-derived elements that might result from this evolutionary dynamics (13, 14). Accordingly, functional studies of the full

repertoire of prophages of bacterial genomes suggest that the majority of prophages are defective at some level: excision, virion formation, lysis or infective ability (18, 19).

Bacterial genomes encode many molecular systems presumably derived from defective prophages. These include gene transfer agents (GTA) that transfer random pieces of chromosomal DNA to other cells (20), and bacteriocins and type 6 secretion systems (T6SS) that are involved in bacterial antagonistic associations (21, 22). Model phage-derived elements, like GTAs or T6SSs, are streamlined and genetically very stable. Yet, genomes contain a number of elements derived from prophages that fit less neatly in the above categories and perform a number of functions with diverse degrees of efficiency: they parasite other phages, kill other bacteria or transfer host DNA (23). Finally, prophage-derived structures are also involved in complex animal-bacteria associations (24, 25). These different elements blur the distinction between stable phage-derived elements and prophages ongoing genetic degradation, suggesting that some defective prophages provide adaptive functions to bacteria.

Temperate phages and their hosts develop complex antagonistic and mutualistic interactions: depending on the circumstances, prophages can either kill bacteria or increase their fitness. There are no systematic studies of which trend dominates the evolutionary dynamics of prophage-bacteria interactions. Here, we bring to the fore a key related question: how do prophages evolve **within** the bacterial genome? To answer it, we identified the repertoire of vertically inherited prophages of *Escherichia coli* and *Salmonella enterica*. The analysis of these prophages revealed unexpected evolutionary patterns suggesting widespread contribution of prophages to bacterial fitness.

Results

Prophages display signs of degradation. In this study we analyzed a dataset of phages and prophages of *E. coli* and *S. enterica* that we have previously identified (15), to which we added prophages from recently published genomes. Prophages were identified using several tools and their precise limits were determined by comparative genomics and expert curation (see Materials and Methods). We identified 624 prophages among 58 and 27 fully assembled genomes of *E. coli* (474 prophages) and *S. enterica* (150 prophages) respectively. Since intact prophages are likely to kill the cell upon induction of the lytic cycle, there should be strong selection for mutations leading to prophage inactivation. We therefore expected to find few large prophages, corresponding to recent integrations, and then a gradient of smaller and smaller prophages having endured diverse levels of genome degradation. Surprisingly, the distribution of the genome size of prophages is clearly bimodal with a class of small and another of large elements (Fig. 1). A near complete separation between the two classes occurs for prophage genome size of ~30kb, the size of the smallest autonomous dsDNA phage infecting enterobacteria in GenBank (Fig. 1). Many prophages (37%) are smaller than 30kb, even though we excluded from the analysis the very small prophages difficult to distinguish from other mobile genetic elements (see Materials and Methods). This suggests either the presence of two different populations of prophages or rapid degradation of large prophages and then stabilization of the resulting elements in the genome.

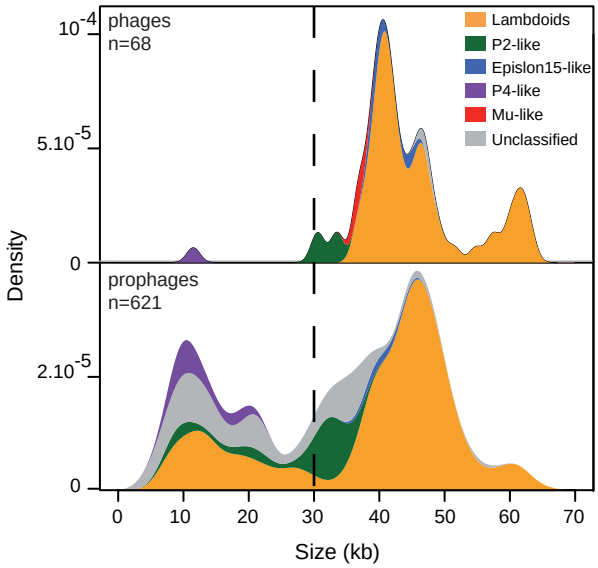

Figure 21. Probability distributions of the genome size of the 68 dsDNA temperate caudophages infecting enterobacteria (top) and of the *Caudovirales* prophages (bottom). The taxonomic groups are indicated on the right of the figure.

The bimodal distribution of prophages could be due to differences in taxonomic groups, e.g., small prophages and large prophages could derive from different types of phages. To test this hypothesis, we classified phages and prophages in taxa using a genome similarity score that includes information on the patterns of gene presence/absence and sequence similarity (following (15, 26)). Most of the small prophages (74%) could be assigned to known taxa. These prophages are systematically smaller than the phages of the same taxa in GenBank, a clear indication that they have endured some genetic degradation ($P<0.0001$, Wilcoxon test). Moreover, 48% of the small prophages could be classified as lambdoid or P2-like phages (Fig. 1), for which all known representatives are larger than 30kb. Few small prophages are clearly distinct from autonomous dsDNA phages (13%): we identified 3 ssDNA Inoviruses and 28 P4-like satellite prophages. The remaining small prophages lack homologs of the characteristic proteins of P4-like or SaPI satellite phages (27, 28): Sid, Pif, CpmA and CpmB (blastp, e-value>0.001) (29, 30). To study the gene repertoires of small prophages, and their differences, we built protein families for the whole dataset of prophages and temperate phages of enterobacteria (see Materials and Methods). Small prophages are significantly enriched in tail genes when compared to temperate caudophages of enterobacteria ($P<0.0001$, χ^2 test) (Fig. S1A). Hence, small prophages rarely encode characteristic satellite phage proteins, they are often phylogenetically close to large prophages and they often encode structural proteins. This shows that most small prophages are not satellite phages, but prophages resulting from the genetic degradation of larger elements.

The observed bimodality of prophage size could result from systematic large neutral deletions of genetic material within the prophage. The deletion spectrum in the chromosome of *Salmonella* is not consistent with this pattern, showing a clear predominance of small deletions (31). Nevertheless, we tested this hypothesis by simulating genetic deletions of different sizes within prophages while requiring conservation of the flanking bacterial core genes. Our results show that a pattern dominated by large deletions leads to unbalanced deletion of genes in the prophage: genes encoded in the central part of the element (like lysis or packaging genes) are much more frequently deleted than genes at the edges (integrases and cargo genes) (Fig. S2). On the contrary, small deletions (and the experimental deletion spectrum) lead to a much weaker dependency of the probability of deletion with the position in the prophage. The comparison of the results of the simulations with the functions encoded in small lambdoid prophages relative to known lambdoid phages does not show significant

differences in most functional classes (Fig. S1B). It is thus not compatible with the predominance of large neutral deletions in prophage evolution.

Many prophages are vertically inherited. Our dataset includes an average of 8.2 and 5.6 prophages per genome for *E. coli* and *S. enterica*, respectively. Prophages at similar loci in different genomes can derive from a single ancestral prophage (orthologous prophages) or from multiple independent integrations at the same loci (15). Hence, prophages in a given chromosomal locus are a mixture of orthologous and non-orthologous prophages. We used a conservative set of four criteria to distinguish them (SI Methods and Fig. S3). First, two orthologous prophages must be integrated into the same chromosomal locus (flanked by the same bacterial core genes). Second, orthologous prophages must have a high genomic similarity score ($R \geq 0.7$). Third, orthologous prophages are replicated like any part of the bacterial chromosome and should thus exhibit similar neutral substitution rates. Synonymous positions are under weak selection and we require prophages to have average synonymous substitution rates similar to the genes of the core genome. In a few cases the inferred ancestral prophage genome was much larger than expected given the distribution of genome lengths of temperate phages infecting enterobacteria. This suggested that we needed a fourth criterion: as orthologous families derive from a single ancestral prophage, the gene diversity (pan-genome) of a given family of orthologous prophages must not be much larger than the number of genes present in the prophage encoding the highest number of genes (see Materials and Methods). With these strict filters our step-by-step method eliminated non-orthologous prophages and removed or split families into multiple smaller families. In the end, we identified 100 families of orthologous prophages (71 in *E. coli* and 29 in *S. enterica*). The majority of the integration events (72%) was observed in one single strain, presumably because they occurred very recent. Around 28% of the inferred integration events involved a prophage that has left remnants in more than one strain, i.e., produced a family of orthologous prophages. These families include 372 prophages (60% of the total) and contain from 2 to 15 orthologous elements (Fig. S4). This suggests that many prophages in a species are derived from a single ancestral integration event in spite of frequent prophage loss. The elements of a given family of orthologous prophages have remained in the bacterial chromosome the same number of years and should exhibit comparable levels of genetic degradation. Indeed, nearly all families of orthologous prophages (90%) include elements of either the large or the small classes of prophage size, but not both.

The groups of prophages with and without orthologous elements do not have significantly different taxonomic distributions (P=0.2, χ^2 test, Fig S5), suggesting no particular taxonomic bias in the prophages that reside in bacterial chromosomes for longer periods of time (Fig. S5). As prophages in general, the orthologous prophages include mostly groups of lambdoids (59%) and a smaller number of P2-like (13%) elements. As expected, orthologous prophages are found in more closely related bacterial strains when compared to non-orthologous prophages integrated into the same loci (P<0.0001, Wilcoxon test). Prophages with orthologous elements are also shorter than prophages lacking orthologs (30.9kb vs 36.7kb on average. P<0.0001 Wilcoxon test), which likely results from their longer residence time. The number of genes lost in all elements of a family cannot be precisely quantified since we ignore the genome of the ancestral phage. However, the comparison of pairs of orthologous prophages allows the quantification of the patterns of differential gene loss in the prophage family, i.e., losses that did not take place in all elements. Prophages have endured an average of 7.8 such gene losses per pair (5.6kb on average, see Materials and Methods). The median deletion is 500nt long and only 40% of pairs of orthologous prophages do not display any indels.

Spontaneous excision rates of complete prophages are of the order of 10^{-6}/cell division for Lambda under non-stressful conditions and are otherwise orders of magnitude higher (32). Therefore, fully functional prophages are unlikely to remain in the chromosome for a long time. To assess how many of the prophage families might constitute functional prophages we analyzed the only strain in our dataset for which the function of all prophages has been experimentally studied (O157:H7 Sakaï) (18). We found orthologous elements for 15 of the 16 defective prophages in this genome. We found no single orthologous prophage for the only fully functional phage. This restricted analysis supports the claim that prophages endure rapid genetic degradation.

Vertically inherited prophages are under purifying selection. To investigate the action of natural selection on the genes of prophages, we computed the ratio of non-synonymous over synonymous substitution rates (dN/dS) for the pools of orthologous genes within bacterial core genes and within orthologous prophages. As expected, bacterial core genomes display very low dN/dS values (median dN/dS=0.06, P<0.0001, Wilcoxon test). Prophage genes display higher dN/dS genes. However, and very surprisingly, most orthologous prophages display a dN/dS ratio much lower than one (median dN/dS=0.22, P<0.0001, Mann-Whitney test). The preferential purge of non-synonymous mutations by natural selection suggests

selection for maintaining the function of the genes encoded in prophages (Fig. 2). A similar dN/dS distribution was observed for a subset of prophages for which orthology was defined using even more stringent criteria (Fig. S6, see Materials and Methods). Nevertheless, prophage genes are under weaker purifying selection than the genes of the hosts' core genomes (median dN/dS=0.06, $P<0.0001$, Wilcoxon test). Very few pairs of orthologous prophages (6%) show dN/dS values consistent with neutral, positive or diversifying selection ($dN/dS \geq 1$). The low dN/dS values are not an artifact associated with the small number of genes or the low density of SNPs in the dataset, since dN/dS values are similar for the most divergent genomes where the signal is the strongest (Fig. S7). In fact, dN/dS is constant along the range of dS values, consistent with the rapid imprint of purifying selection in dN/dS in *E. coli* (33). Small prophages are as constrained by purifying selection as the large prophages (median dN/dS=0.23 and 0.22 respectively, $P=0.09$, Wilcoxon test).

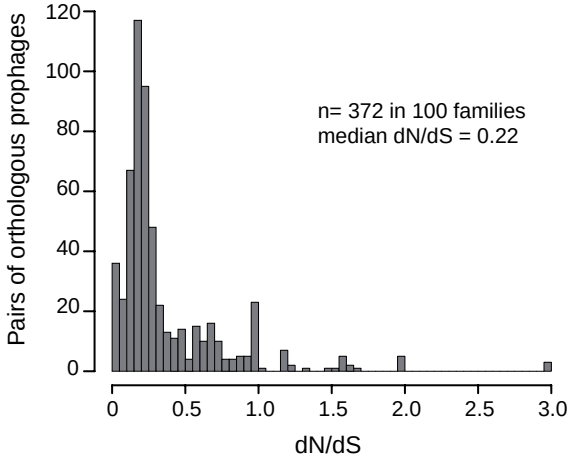

Figure 22. Histogram of the ratio of non-synonymous to synonymous substitutions (dN/dS) between orthologous prophages.

Recombination with incoming phages can imprint a signal of purifying selection on prophages by introducing an over-abundance of synonymous polymorphisms resulting from purifying selection on phages. To test if this effect was producing our unexpected findings, we detected recombinant genes among orthologous prophages using seven different methods and a combination of them. In each analysis we removed the gene families showing significant evidence of recombination or phylogenetic incongruence. This led to the rejection of between 8% and 32% of the genes (Table S1). A joint analysis using PHI/NSS/MaxChi, Prunier and MaxChi on concatenates rejected 16% of the genes (34-36). In all the eight variants of the analysis we observed dN/dS values for the non-recombining genes very

significantly smaller than 1 (Table S1). Hence, while our results confirm the existence of recombination between prophages and phages (or other prophages), as previously shown (37), this effect is not the major cause of the observed low dN/dS values.

To investigate the patterns of substitution rates and gene loss in prophages we analyzed each gene in function of its position in the prophage genome. Gene positions are highly conserved in lambdoids and in P2-like phages (38, 39). We restricted our attention to lambdoid and P2-like large prophages (55% of the dataset) because they can be mapped accurately in this genetic organization. Overall, there is a nearly constant high degree of gene conservation in orthologous prophages (Fig. 3). Comparison with Fig S2 confirms our previous observation that prophage degradation in our dataset does not result from random neutral large deletions. The high degree of gene conservation might partly result from the strict rules used to define prophage orthology. In lambdoids, the region encoding the tail proteins is slightly more conserved and the cargo region (extreme end after the tail genes) is less conserved. P2-like prophages are also well conserved along the genome except for two regions ("morons" 1 and 2), which typically contain fast evolving accessory genes (39). The regions of the prophages where gene loss is less frequent are also those where genes have lower dN/dS values (except P2-like moron region 1, that has low dN/dS values) (Fig. 3). This is consistent with purifying selection on these genes. Regions encoding infection-related functions, such as replication or capsid proteins, are very conserved in prophages. These results were confirmed independently by the analysis of prophage gene families in terms of functional categories, which are not limited to large P2-like and lambdoid prophages (Fig. S8). As a whole, the data suggest pervasive positive contribution of prophage genes to bacterial fitness.

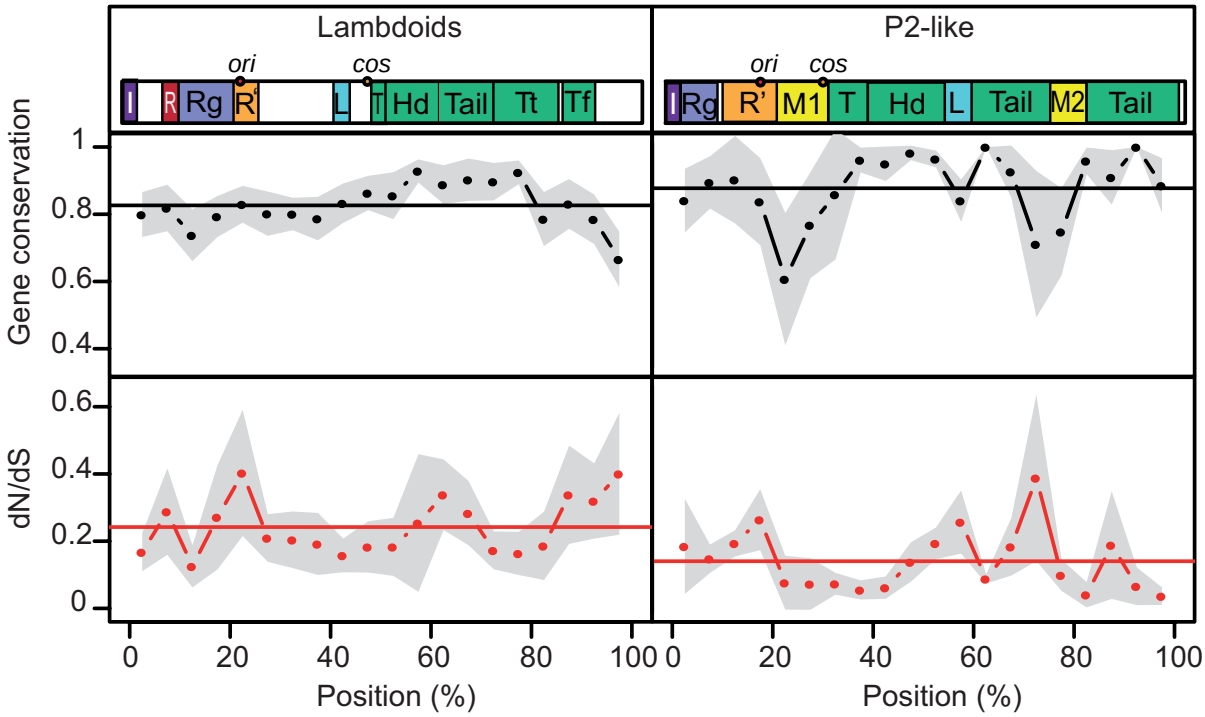

Figure 23. Average gene conservation and non-synonymous to synonymous substitution ratios (dN/dS) along prophage genomes. Left: Lambdoids. Right: P2-like prophages. Only large (≥30kb) prophages were considered in this analysis since small prophages can't be confidently represented on the normalized genome map. Schematic representations of lambdoid and P2-like genomes are given on top and are based on Lambda (left) and P2 (right) genome architectures. I: integration, R: recombination, Rg: regulation, R': replication, T: terminase(s), Hd: head, L: lysis, Tf: tail fiber, Tt: tail tip, M1: moron region 1, M2: moron region 2. The grey contours represent the 95% confidence intervals of gene conservation and dN/dS ratios. Horizontal lines indicate median values of gene conservation and dN/dS ratios.

Discussion

We found that orthologous prophages are numerous, representative of the diversity of enterobacterial phages and have endured genetic degradation since the ancestral integration into the genome. This has presumably rendered them defective. Surprisingly, most orthologous prophages show strong signs of purifying selection. Previous studies have shown that many recently acquired genes in bacteria are under purifying selection (40). We have mentioned above that many prophages carry accessory genes that are adaptive to bacteria (6-8). Our results differ from previous works in one essential aspect: in our study many of the prophage genes under stronger purifying selection encode core phage-related functions, like tail and lysis proteins. This suggests that prophage functions are under selection in the bacterial chromosome.

We observed a strongly bimodal distribution of prophage genome size, which is neither caused by phage taxonomic biases nor by large neutral deletions of genetic material. Bi-modality could result from rapid inactivating gene losses followed by much slower genetic degradation of the remaining genes (14). The slow-down of genetic inactivation could result from purifying selection on certain genes as observed in the dN/dS analysis. This raises the question of why bacteria are not accumulating even larger numbers of prophage genes. We suggest that analog/homolog gene replacements may lead to frequent gene loss. Genomes of enterobacteria are constantly acquiring prophages of a relatively small number of taxonomic groups. Extant prophage genes may thus suddenly become under relaxed selection when homologous or analogous genes arrive in the host. This may lead to the replacement of prophage genes by others performing similar functions.

Our results show that many phage-derived functions are under purifying selection. We suggest this is because they are adaptive for their host. Previous studies have shown that prophages unable to produce viable virions upon infection can protect from superinfection, excise, package DNA and even infect other cells (8, 18, 19). Therefore, partly degraded prophage elements can have a number of adaptive functions, as described below (Fig. 4).

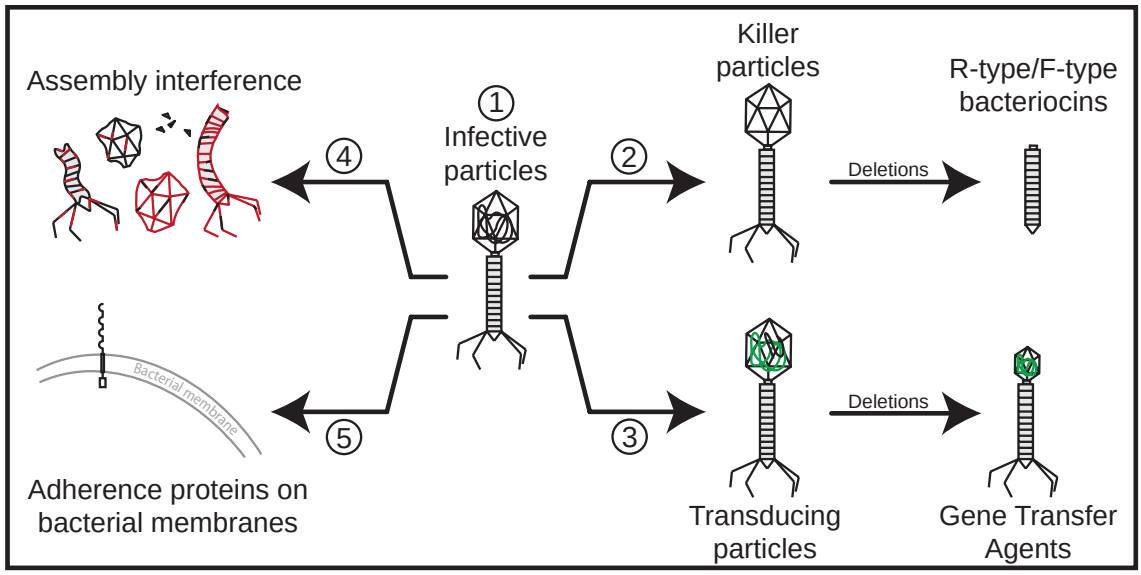

Figure 24. Putative functions of orthologous prophages conserved in their hosts. i) Functional prophages can be used as molecular weapons to kill non-lysogens through the production of infective particles. ii) Defective prophages can produce non-infective particles (phage killer particles and R/F-type bacteriocins) that kill sensitive cells. iii) Prophages can form transducing particles and Gene Transfer Agents (GTAs) that promote host DNA exchange (displayed in green). iv) Degraded prophages might interfere with the assembly of other phages (represented in red) leading to the formation of defective particles. v) Prophage structural proteins often display Ig-like domains that might be used by their hosts for adherence in niche colonization.

Functional prophages allow populations of lysogens to kill non-lysogen competitors (41). However, this advantage rapidly disappears by the creation of lysogens in the susceptible populations (12), and comes at the cost of cell lysis in a fraction of the population. Our results indicating that intact prophages are rapidly lost suggest that this strategy might be very short-lived or rarely used.

Phage-derived bacteriocins kill cells whose genomes lack their cognate immunity genes (21, 42). Contrary to fully functional phages they do not re-infect other cells neither do they produce lysogens, preventing the creation of immunity in other populations. R-type and F-type bacteriocins are typically composed of domesticated tail and lysis genes from myophages and siphophages respectively. This fits the observed gene repertoires of some families of small prophages in our dataset. Notably the largest orthologous prophage family has a phylogenetic tree that mirrors that of the bacterial host, lacks integrases and seems to have been stabilized in a large number of strains of *Salmonella* (Fig. S9). This element is related to P2-like phages (*Myoviridae*) and could therefore correspond to a R-type bacteriocin. The putative domestication of this prophage might even pre-date the split between *Escherichia* and *Salmonella*, since we identified a very similar small prophage also

missing an integrase at the same position in two *E. coli* strains (Fig. S9). Most prophage families do not fit so well the description of R-type or F-type bacteriocins but could correspond to phage killer particles. These elements behave like bacteriocins but are very diverse genetically, presenting characteristics ranging from streamlined elements like R-type and F-type bacteriocins to nearly complete phages (23, 43-45). Defective phages can easily give rise to phage killer particles. For example, PBSX and other non-infectious defective phage particles were termed phage killer particles or protophages and act *de facto* as bacteriocins (43, 45). Some of these systems are conserved among different isolates and might be widespread among bacterial species (23, 44). Altered particles of T-even phages also display a bacteriocin activity (47). It is possible that many of the orthologous prophages correspond to R/F-type bacteriocins or phage killer particles.

GTAs are found in diverse prokaryotic clades and are thought to have originated from domestication of general transducing phages (48). They typically encode structural genes, lack integrases and evolve under purifying selection. General transducing phages P22 and P1 also produce and transfer phage particles that contain exclusively bacterial DNA (49, 50). An increased rate of general transduction by these phages can be obtained by a small number of different mutations altering the head morphogenesis process (51, 52). Therefore, a defective phage may lead to a transducing agent in very few mutational steps. Accordingly, several prophage families in our dataset encode a nearly complete set of structural and lysis genes while lacking many functions associated with regulation, replication and integration/excision. Also, the size of GTAs, between 14 kb and 30kb (20), is in close agreement with the lower peak of prophage size distribution observed in our dataset (Fig. 1). GTAs package unspecific DNA and this can be easily achieved by *pac*-based phage terminases but not by *cos*-based terminases (53). Lambdoids use either *pac* or *cos* systems (54), and the former might act as GTAs. However, P2-like phages use *cos* packaging systems (55) and are unlikely to become GTAs. Nevertheless, the mechanism by which GTAs package random fragments of DNA is not well understood and it remains possible that trans-encoded packaging systems fill phage particles with host DNA.

Our results suggest that bacteria select for the conservation of components of virion particles. Virions of lambdoids are composed of over 500 protein units and the architecture of the phage particle depends on their precise interaction (56). If the assembly of different phage particles in the cell results in protein interactions between components of the different virions this might lead to the production of defective phage particles. If a phage infects a cell containing prophages, their expression may interfere with the incoming phage and diminish its viable

progeny. This would result in higher fitness for bacterial populations carrying prophages. This idea is supported by the observation that double co-infections by lambdoids lead to fewer virions when compared to single infections (57). A similar type of molecular interference has been proposed to explain why proteins forming large complexes are less prone to horizontal gene transfer and duplication (58, 59). Given the high complexity of phage particles (providing ample potential for interference) and the abundance of phage infections in natural environments (providing strong fitness gains for bacteria avoiding phage infections), we speculate that selection for the maintenance of interfering prophages might be advantageous for the host. This advantage might, however, disappear in the long-term due to rapid phage diversification.

There may be many other uses of phage proteins that for the moment are purely speculative. For example, about 25% of caudophages encode proteins with Ig-like domains (60). These phages are abundant in mucosal surfaces of Metazoans, protecting them from bacterial infections and allowing phages to have a constant supply of hosts (61). Conversely, expressing prophage structural proteins at the surface of bacterial cells could aid bacterial infection or niche colonization. For example, prophage tail proteins expressed after treatment with mitomycin C or UV light mediate *Streptococcus mitis* and *Enterococcus faecalis* platelet binding, favoring infective endocarditis (19, 62). Prophage structural proteins might thus favor physical interactions of bacteria with their environment.

The functions encoded by some prophages are compatible with selection for the use of prophages in antagonistic relations with other bacteria (as biological weapons, phage-derived bacteriocins or killer particles), for horizontal transfer (GTAs), for protection against other phages or for bacterial colonization. Several cases have been described in the literature of degraded phages performing such functions. Yet, no single type of phage-derived element fits the gene repertoires of all orthologous prophages that we have identified. This suggests that prophages provide several different functions. These prophage-derived functions may be very generic. Since closely related phages are constantly arriving at the bacterial genome, prophage-derived genes are likely to be frequently superseded by other incoming prophage-derived genes. As mentioned above, this should lead to frequent analogous/homologous gene replacements. Occasionally prophage-derived elements may evolve towards a new highly specialized function. This will lead to their enduring domestication.

Materials and Methods

Data. We downloaded 58 and 27 complete genomes of *E. coli* and *S. enterica* respectively from the Bacterial section of NCBI RefSeq (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/, last accessed February 2013), and 68 temperate caudophage genomes infecting enterobacteria from the Virus section of NCBI RefSeq (ftp://ftp.ncbi.nih.gov/genomes/Viruses/, last accessed February 2013).

Prophage detection and classification. Prophages were detected as in (15), except that prophages with no match to core phage genes were discarded. This resembles the implementation of the stringent option of Phage Finder (63) but is doesn't exclude *a priori* the smallest prophages. Prophages detected at rearrangement breakpoints were removed from the analysis since their positions couldn't be confidently defined in all genomes. To avoid gene loss over-prediction due to imprecise delimitation of prophages, we built the families of homologous proteins found in all genomes between the same two flanking core genes. Each protein family was then considered as phage-related or host-related but not both. Information about the 624 prophages is detailed in Dataset S1.

The resulting 624 prophages were classified by comparing their gene repertoires with those of phages infecting enterobacteria (15). Gene repertoire relatedness (R) between pairs of (pro)phages was defined as: $\sum_{i=1}^{M} \frac{S_{(Ai,Bi)}}{\min(n_A, n_B)}$ where $S_{(\mathrm{A}i,\mathrm{B}i)}$ is the similarity score of the pair *i* of homologous proteins between (pro)phage A and (pro)phage B (varying from 0.4 to 1), *M* the total number of homologs shared by (pro)phages A and B and n_{A} and n_{B} the total number of proteins of (pro)phages A and B, respectively.

Other Materials and Methods. Details on the computation of core genomes and bacterial distances, the definition of orthologous prophages, the functional assignment of prophage proteins, deletion simulations, estimation of recombination and the computation of synonymous and non-synonymous substitution rates are provided in SI Materials and Methods.

Acknowledgments

We thank Jean-François Gout for helpful comments on an earlier version of this manuscript. This work was supported by an European Research Council starting grant [EVOMOBILOME n°281605] to EPCR; and a grant from the Ministère de l'enseignement supérieur et de la recherche to LMB.

References

1. Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* 399(6736):541-548.
2. Ptashne M (1992) *Genetic Switch: Phage Lambda and Higher Organisms*, (Blackwell, Cambridge, MA), 2nd Ed p 192.
3. Campbell AM (1996) Bacteriophages. *Escherichia coli and Salmonella: cellular and molecular biology*, (ASM Press, Washington DC), pp 2325-2338.
4. Steinberg KM & Levin BR (2007) Grazing protozoa and the evolution of the Escherichia coli O157:H7 Shiga toxin-encoding prophage. *Proc Biol Sci* 274(1621):1921-1929.
5. Waldor MK & Friedman DI (2005) Phage regulatory circuits and virulence gene expression. *Curr Opin Microbiol* 8(4):459-465.
6. Edlin G, Lin L, & Bitner R (1977) Reproductive fitness of P1, P2, and Mu lysogens of Escherichia coli. *J Virol* 21(2):560-564.
7. Godeke J, Paul K, Lassak J, & Thormann KM (2011) Phage-induced lysis enhances biofilm formation in Shewanella oneidensis MR-1. *ISME J* 5(4):613-626.
8. Wang X, *et al.* (2010) Cryptic prophages help bacteria cope with adverse environments. *Nat Commun* 1:147.
9. Rabinovich L, Sigal N, Borovok I, Nir-Paz R, & Herskovits AA (2012) Prophage Excision Activates Listeria Competence Genes that Promote Phagosomal Escape and Virulence. *Cell* 150(4):792-802.
10. Bossi L, Fuentes JA, Mora G, & Figueroa-Bossi N (2003) Prophage contribution to bacterial population dynamics. *J Bacteriol* 185:6467-6471.
11. Brown SP, Le Chat L, De Paepe M, & Taddei F (2006) Ecology of microbial invasions: Amplification allows virus carriers to invade more rapidly when rare. *Curr Biol* 16(20):2048-2052.
12. Gama JA, *et al.* (2013) Temperate Bacterial Viruses as Double-Edged Swords in Bacterial Warfare. *PLoS One* 8(3):e59043.
13. Canchaya C, Fournous G, & Brussow H (2004) The impact of prophages on bacterial chromosomes. *Mol Microbiol* 53:9-18.
14. Casjens S (2003) Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol* 49:277-300.
15. Bobay LM, Rocha EP, & Touchon M (2013) The Adaptation of Temperate Bacteriophages to Their Host Genomes. *Mol Biol Evol* 30:737-751.
16. Lawrence JG, Hendrix RW, & Casjens S (2001) Where are the pseudogenes in bacterial genomes? *Trends Microbiol* 9(11):535-540.
17. Kuo CH & Ochman H (2010) The Extinction Dynamics of Bacterial Pseudogenes. *PLoS Genet* 6(8):e1001050.
18. Asadulghani M, *et al.* (2009) The defective prophage pool of Escherichia coli O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog* 5(5):e1000408.
19. Matos RC, *et al.* (2013) Enterococcus faecalis prophage dynamics and contributions to pathogenic traits. *PLoS Genet* 9(6):e1003539.
20. Lang AS, Zhaxybayeva O, & Beatty JT (2012) Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Microbiol* 10(7):472-482.
21. Michel-Briand Y & Baysse C (2002) The pyocins of Pseudomonas aeruginosa. *Biochimie* 84(5-6):499-510.

22. Leiman PG, *et al.* (2009) Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin. *Proc Natl Acad Sci U S A* 106(11):4154-4159.
23. Campbell A (1977) Defective Bacteriophages and Imcomplete Prophages. *Regulation and Genetics*, eds Fraenkel-Conrat H & Wagner RR (Springer US, New-York), Vol 8, pp 259-328.
24. Hurst MR, Glare TR, & Jackson TA (2004) Cloning Serratia entomophila antifeeding genes--a putative defective prophage active against the grass grub Costelytra zealandica. *J Bacteriol* 186(15):5116-5128.
25. Shikuma NJ, *et al.* (2014) Marine tubeworm metamorphosis induced by arrays of bacterial phage tail-like structures. *Science* 343(6170):529-533.
26. Rohwer F & Edwards R (2002) The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage. *J Bacteriol* 184(16):4529-4535.
27. Novick RP & Subedi A (2007) The SaPIs: mobile pathogenicity islands of Staphylococcus. *Chem Immunol Allergy* 93:42-57.
28. Christie GE & Dokland T (2012) Pirates of the Caudovirales. *Virology* 434(2):210-221.
29. Ubeda C, *et al.* (2009) Specificity of staphylococcal phage and SaPI DNA packaging as revealed by integrase and terminase mutations. *Mol Microbiol* 72(1):98-108.
30. Damle PK, *et al.* (2012) The roles of SaPI1 proteins gp7 (CpmA) and gp6 (CpmB) in capsid size determination and helper phage interference. *Virology* 432(2):277-282.
31. Sun S, Ke R, Hughes D, Nilsson M, & Andersson DI (2012) Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. *PLoS One* 7(8):e42639.
32. Gottesman ME & Yarmolinsky MB (1968) Integration-negative mutants of bacteriophage lambda. *J Mol Biol* 31(3):487-505.
33. Rocha EPC, *et al.* (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* 239(2):226-235.
34. Abby SS, Tannier E, Gouy M, & Daubin V (2010) Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics* 11:324.
35. Bruen TC, Philippe H, & Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172(4):2665-2681.
36. Martin DP, *et al.* (2010) RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26(19):2462-2463.
37. De Paepe M, *et al.* (2014) Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-like recombinases. *PLoS Genet* 10(3):e1004181.
38. Campbell A & Botstein D (1983) Evolution of the lambdoid phages. *Lambda II*, eds Hendrix RW, Roberts JW, Stahl FW, & Weisberg RA (Cold Spring Harbor Laboratory Cold Spring Harbor, NY), pp 365-380.
39. Nilsson A & Haggard Ljungquist E (2006) The P2-Like Bacteriophages. *The Bacteriophages*, eds Calendar RL & Abedon ST (Oxford University Press, New York), 2nd Ed Vol 1, pp 365-390.
40. Daubin V & Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome Res* 14(6):1036-1042.
41. Brown SP, Le Chat L, De Paepe M, & Taddei F (2006) Ecology of microbial invasions: amplification allows virus carriers to invade more rapidly when rare. *Curr Biol* 16(20):2048-2052.
42. Hardy KG (1975) Colicinogeny and related phenomena. *Bacteriol Rev* 39(4):464-515.
43. Bradley DE (1967) Ultrastructure of bacteriophage and bacteriocins. *Bacteriol Rev* 31(4):230-314.

44. Garro AJ & Marmur J (1970) Defective bacteriophages. *J Cell Physiol* 76(3):253-263.
45. Wood HE, Dawson MT, Devine KM, & McConnell DJ (1990) Characterization of PBSX, a defective prophage of Bacillus subtilis. *J Bacteriol* 172(5):2667-2674.
46. Gerdes JC & Romig WR (1975) Complete and Defective Bacteriophages of Classical Vibrio cholerae: Relationship to the Kappa Type Bacteriophage. *J Virol* 15(5):1231-1238.
47. Duckworth DH (1970) The metabolism of T4 phage ghost-infected cells. I. Macromolecular synthesis and ransport of nucleic acid and protein precursors. *Virology* 40(3):673-684.
48. Lang AS & Beatty JT (2007) Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol* 15(2):54-62.
49. Schmieger H (1972) Phage P22-mutants with increased or decreased transduction abilities. *Mol Gen Genet* 119(1):75-88.
50. Wall JD & Harriman PD (1974) Phage P1 mutants with altered transducing abilities for Escherichia coli. *Virology* 59(2):532-544.
51. Casjens S, *et al.* (1992) Molecular genetic analysis of bacteriophage P22 gene 3 product, a protein involved in the initiation of headful DNA packaging. *J Mol Biol* 227(4):1086-1099.
52. Iida S, Hiestand-Nauer R, Sandmeier H, Lehnherr H, & Arber W (1998) Accessory genes in the darA operon of bacteriophage P1 affect antirestriction function, generalized transduction, head morphogenesis, and host cell lysis. *Virology* 251(1):49-58.
53. Ebel-Tsipis J, Botstein D, & Fox MS (1972) Generalized transduction by phage P22 in Salmonella typhimurium. I. Molecular origin of transducing DNA. *J Mol Biol* 71(2):433-448.
54. Casjens SR (2011) The DNA-packaging nanomotor of tailed bacteriophages. *Nat Rev Microbiol* 9(9):647-657.
55. Nilsson AS & Haggard-Ljungquist E (2007) Evolution of P2-like phages and their impact on bacterial evolution. *Res Microbiol* 158(4):311-317.
56. Hauser R, *et al.* (2012) Bacteriophage protein-protein interactions. *Adv Virus Res* 83:219-298.
57. Refardt D (2011) Within-host competition determines reproductive success of temperate bacteriophages. *ISME J* 5(9):1451-1460.
58. Jain R, Rivera MC, & Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* 96(7):3801-3806.
59. Baker CR, Hanson-Smith V, & Johnson AD (2013) Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science* 342(6154):104-108.
60. Fraser JS, Yu Z, Maxwell KL, & Davidson AR (2006) Ig-like domains on bacteriophages: a tale of promiscuity and deceit. *J Mol Biol* 359(2):496-507.
61. Barr JJ, *et al.* (2013) Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc Natl Acad Sci U S A* 110(26):10771-10776.
62. Bensing BA, Siboo IR, & Sullam PM (2001) Proteins PblA and PblB of Streptococcus mitis, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. *Infect Immun* 69(10):6186-6192.
63. Fouts DE (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* 34(20):5839-5851.

Conclusions et perspectives

Ce travail suggère premièrement que de nombreux prophages sont des éléments en cours de dégradation. De manière intéressante, l'ensemble des éléments phagiques présente une distribution de taille bimodale. De nombreux éléments ont une taille proche de celle des phages complets présents dans GenBank (>30kb) alors que beaucoup d'autres éléments ont une taille bien inférieure (~15kb). Ces petits éléments sont majoritairement apparentés à des phages lambdoïdes et P2-like qui sont typiquement supérieurs à 30kb lorsqu'ils sont fonctionnels. La présence de nombreux gènes de structure parmi ces petits prophages suggère qu'ils ne correspondent pas à des phages satellites qui sont typiquement dépourvus de tels gènes. Les petits prophages semblent donc correspondre à des prophages non mobiles en cours de dégradation. De nombreux prophages de grande taille semblent également correspondre à des éléments non fonctionnels comme suggéré par des données expérimentales (Asadulghani, et al. 2009). Les prophages hérités verticalement semblent en outre plus dégradés que les prophages non orthologues. Ceci suggère qu'une contre-sélection des prophages fonctionnels par mutations inactivatrices a lieu rapidement après leur intégration. Une deuxième phase, plus lente, d'accumulation de délétions dégraderait ensuite ces éléments. L'étude des substitutions synonymes et non synonymes (dN/dS) affectant les prophages orthologues suggère que la vaste majorité de ces éléments évolue sous sélection purificatrice. De manière très étonnante, ces signaux ne sont pas uniquement détectés au sein des gènes accessoires mais également au sein des gènes phagiques essentiels pour la réplication et la morphogénèse virale. Ces résultats suggèrent donc l'utilisation de nombreux prophages défectifs par la bactérie. Ces prophages orthologues semblent cependant être éliminés relativement rapidement des génomes hôtes. Ceci supporte l'idée que ces systèmes sont fréquemment perdus et remplacés, sous doute par l'acquisition de nouveaux prophages homologues ou analogues. Différentes hypothèses concernant le rôle fonctionnel de ces prophages domestiqués ont été avancées dans ce travail. Je me contenterai ici d'approfondir certains aspects.

i) Le passage de prophage fonctionnel à un système domestiqué pourrait se faire par étapes graduelles. En effet, il a été suggéré que les prophages fonctionnels peuvent apporter un avantage aux bactéries lysogènes en leur permettant d'éliminer leurs compétitrices. Bien que cet avantage sélectif tende à disparaître rapidement, cela a pour avantage d'empêcher la

contre-sélection immédiate des prophages. Les particules phagiques tueuses présentent différents degrés de dysfonctionnement par rapport aux phages infectieux. De plus, l'altération chimique de particules phagiques infectieuses permet d'obtenir des particules non infectieuses aux propriétés similaires (Duckworth 1970). Ceci suggère que des particules présentant des propriétés bactéricides peuvent être obtenues par dégénération de prophages fonctionnels. La stabilisation de tels éléments au cours du temps conduirait ensuite à l'obtention de bactériocines plus compactes telles que les pyocines de types R et F, correspondant à des structures de queues phagiques. La domestication de prophages fonctionnels en bactériocines pourrait donc suivre un continuum initié par des mutations altératrices. Il est probable qu'un tel scénario de mutations altératrices et de dégradation puisse également expliquer l'émergence d'Agents de Tranfert de Gènes à partir de prophages permettant la transduction généralisée; la fréquence de particules transductrices étant modulable par mutations (Casjens, et al. 1992). Selon ce modèle, la domestication de prophages par mutations altératrices serait un phénomène assez reproductible et moins rare qu'originellement attendu. Il pourrait être intéressant de tester la reproductibilité de ces phénomènes *in vitro*. Peut-on obtenir des mutants produisant des particules phagiques non infectieuses à partir de lysogènes contenant un prophage fonctionnel? De telles particules ont-elles des propriétés bactéricides?

ii) L'expression des génomes des phages tempérés est contrôlée par un faible nombre de grands opérons (typiquement trois opérons). Chez les phages lambdoïdes, environ la moitié des gènes sont codés sur l'opéron "late". La domestication d'un seul gène de cet opéron nécessiterait donc la conservation de nombreux autres gènes phagiques permettant l'expression de cet opéron. Cette contrainte est illustrée par l'utilisation de deux gènes structurels de prophages par leurs hôtes: *E. faecalis* et *S. mitis* (Bensing, et al. 2001; Matos, et al. 2013). Il est probable que l'hôte s'affranchisse ultérieurement du reste du prophage grâce à l'apparition d'un promoteur transcriptionnel indépendant contrôlant le gène d'intérêt. Ce modèle suggère que la domestication d'un seul gène nécessite d'abord la conservation d'un grand fragment du prophage. Il est donc envisageable que certains prophages de petite ou grande taille sous sélection purificatrice ne soient conservés que pour la présence d'un seul gène bénéfique pour l'hôte.

Conclusion

L'ensemble de ces travaux s'est attaché à comprendre la dynamique spatiale et temporelle des phages tempérés au sein des génomes d'entérobactéries, ainsi que leurs processus de diversification. Les phages tempérés représentent une part importante de la diversité génétique des entérobactéries. Malgré l'impact néfaste des infections virales sur leurs hôtes, la relation entre ces entités se révèle bien plus complexe. En effet, les phages tempérés montrent des signes d'adaptation aux contraintes de leur chromosome hôte. De manière intéressante, les sites d'intégration des phages semblent fortement conservés au cours du temps. Ceci semble refléter l'avantage sélectif que représente la lysogénie dans certaines conditions. Les prophages peuvent ainsi apporter des avantages sélectifs à leurs hôtes à trois niveaux différents: i) par protection contre la surinfection. ii) par l'expression de gènes accessoires d'intérêts pour l'hôte. iii) par leur domestication occasionnelle en systèmes bactériens. Ce dernier processus est probablement plus fréquent, plus dynamique et plus reproductible qu'attendu.

La présence de prophages hérités verticalement et évoluant sous sélection purificatrice suggère que ces éléments ont un rôle fonctionnel pour la bactérie. J'ai mentionné différentes fonctions que les prophages domestiqués peuvent remplir dans la cellule (section 3.3.2). Il serait donc intéressant de tester expérimentalement l'expression et l'activité de ces systèmes. Peut-on induire la formation de ces structures de type phagique chez ces bactéries? Peut-on définir leur fonction? Une ambigüité semble exister entre ces systèmes et les phages fonctionnels (section 3.3.3). Peut-on distinguer expérimentalement ces éléments des particules phagiques infectieuses? Outre leur caractérisation fonctionnelle, de tels travaux pourraient permettre de mieux définir les frontières entre prophages fonctionnels, prophages dégradés et prophages domestiqués.

J'ai décrit plus haut les limites méthodologiques liées à la détection des prophages. Le développement et l'amélioration des approches de détection de prophages sont cependant envisageables. Le récent essor de la métagénomique permet le séquençage d'une grande quantité de génomes viraux (viromes) (Edwards and Rohwer 2005). Ces données peuvent potentiellement être utilisées à très large échelle afin d'améliorer la sensibilité et l'universalité des approches de détection basées sur la recherche d'homologies de séquences. La

contamination expérimentale ou la transduction peuvent cependant introduire des séquences bactériennes au sein de ces viromes et restent un obstacle important à l'utilisation de ces données à des fins de détection de prophages. Des méthodes récentes permettent néanmoins l'identification de ces contaminants (Roux, et al. 2013). Le séquençage et la caractérisation de génomes de phages additionnels permettraient également d'améliorer ces approches.

La vaste diversité des répertoires de gènes des phages tempérés, additionnée à la forte conservation des sites d'intégrations phagiques, permet de considérer les phages tempérés comme un génotype étendu bactérien, par analogie au phénotype étendu défini par Richard Dawkins (Dawkins 1982). Selon l'hypothèse du phénotype étendu, les traits phénotypiques codés par le génome d'un organisme vivant peuvent s'exprimer en dehors de cet organisme (y compris chez d'autres organismes). Par symétrie, les phages tempérés représentent des génomes étendus ou "satellites" extracellulaires gravitant autour de leur hôte. Ils peuvent être perçus comme des modules génomiques permettant des modifications phénotypiques bactériennes de par leurs gènes accessoires et leur domestication potentielle. Ils semblent ainsi être fréquemment recrutés par exaptation. Cette utilisation des prophages semble facilitée par le fait que les sites d'intégrations phagiques sont conservés au sein des génomes bactériens et que l'ADN phagique est pré-adapté aux contraintes du chromosome hôte. L'utilisation fréquente des phages tempérés par les bactéries pourrait être liée à la forte capacité qu'ont les phages à produire de nombreuses innovations génétiques. Il est probable que cette capacité à générer du matériel génétique innovant peut être mise en relation avec différentes propriétés spécifiques aux phages.

Je vais me concentrer sur les spécificités qui font des phages tempérés des acteurs privilégiés de l'évolvabilité bactérienne et de potentielles "usines de gènes".

i) *Des entités extracellulaires*. Les phages présentent une différence importante avec les autres éléments mobiles: ils peuvent être maintenus à l'état de dormance de manière extracellulaire (il n'est cependant pas exclu que d'autres éléments utilisent fréquemment la transduction comme moyen de transfert). Les phages, et la diversité génétique qu'ils contiennent, sont donc probablement moins affectés par la dynamique des populations hôtes que d'autres éléments tels que les plasmides. Les particules virales contiennent ainsi des collections de gènes bactériens qui sont maintenus extracellulairement à l'état de dormance. La particule virale permet également une plus forte dispersion dans le milieu car leur transmission ne nécessite pas de contacts entre cellules.

ii) *Un double cycle*. L'ambivalence des phages tempérés en tant que prédateurs et modules chromosomiques bactériens pourrait également être à l'origine de leur diversification rapide.

En effet, l'alternance des croissances lytiques et lysogéniques permet de modifier cycliquement les contraintes sélectives agissant sur les différents gènes de ces éléments. Il est attendu que les gènes accessoires soient affranchis de sélection purificatrice lors du cycle lytique. A l'inverse, les gènes de morphogénèse nécessaires au cycle lytique évolueraient de manière neutre lors de la lysogénie. Ces épisodes périodiques de diminution des contraintes sélectives pourraient ainsi permettre aux gènes phagiques d'explorer plus largement des paysages adaptatifs par dérive génétique. Additionné au fort mosaïcisme phagique promu par la recombinaison, ce phénomène pourrait être à l'origine de la diversification importante des répertoires de gènes phagiques.

iii) *La capside*. Des régions d'ADN non essentiel sont fréquemment observées chez les phages (section 2.3.2). Ces séquences codent souvent pour des protéines accessoires bénéfiques pour l'hôte et sont donc indirectement avantageuses pour le phage. La présence de ces régions pourrait cependant résulter de contraintes propres aux phages. En effet, malgré la forte diversité des virus, certains travaux ont souligné que les protéines de capside des virus des trois domaines du vivant présentent un nombre de conformations structurelles relativement réduit (Bamford 2003; Bamford, et al. 2005; Krupovic and Bamford 2011). Ceci souligne les contraintes importantes liées à la formation de particules virales viables. Si les phages de l'ordre des *Caudovirales* présentent une très vaste diversité génétique, l'architecture générale de leurs capsides est beaucoup plus limitée. L'ADN de ces phages est en outre fortement concentré dans la capside à un taux assez similaire chez de nombreux phages (~500mg par ml) (Molineux and Panja 2013). Cette importante condensation pourrait refléter une adaptation permettant une meilleure résistance de l'ADN aux altérations chimiques ou pourrait résulter de contraintes liées à l'encapsidation ou à l'éjection (Molineux and Panja 2013). Il est probable que cette double contrainte liée à l'architecture de la particule phagique et à la condensation de l'ADN contraigne fortement la taille des génomes phagiques. Les phages pourraient donc être contraints de conserver de l'ADN non essentiel sur de longues périodes afin de maintenir une forte condensation au sein de la capside. La conservation de telles séquences, en l'absence de contraintes sélectives fortes, pourrait alors constituer une source importante d'innovation génétique.

Références

Ackermann HW, Prangishvili D 2012. Prokaryote viruses studied by electron microscopy. Arch Virol 157: 1843-1849.

Akhter S, Aziz RK, Edwards RA 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res.

Allison HE 2007. Stx-phages: drivers and mediators of the evolution of STEC and STEC-like pathogens. Future Microbiol 2: 165-174.

Andersson AF, Pelve EA, Lindeberg S, Lundgren M, Nilsson P, Bernander R 2010. Replication-biased genome organisation in the crenarchaeon Sulfolobus. BMC Genomics 11: 454.

Andersson JO, Andersson SG 2001. Pseudogenes, junk DNA, and the dynamics of Rickettsia genomes. Mol Biol Evol 18: 829-839.

Arkin A, Ross J, McAdams HH 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. Genetics 149: 1633-1648.

Asadulghani M, Ogura Y, Ooka T, Itoh T, Sawaguchi A, Iguchi A, Nakayama K, Hayashi T 2009. The defective prophage pool of Escherichia coli O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. PLoS Pathog 5: e1000408.

Avrani S, Wurtzel O, Sharon I, Sorek R, Lindell D 2011. Genomic island variability facilitates Prochlorococcus-virus coexistence. Nature 474: 604-608.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2: 2006 0008.

Bachmann BJ 1972. Pedigrees of some mutant strains of Escherichia coli K-12. Bacteriol Rev 36: 525-557.

Backhaus H 1985. DNA packaging initiation of Salmonella bacteriophage P22: determination of cut sites within the DNA sequence coding for gene 3. J Virol 55: 458-465.

Ball CA, Johnson RC 1991. Efficient excision of phage lambda from the Escherichia coli chromosome requires the Fis protein. J Bacteriol 173: 4027-4031.

Baltimore D 1971. Expression of animal virus genomes. Bacteriol Rev 35: 235-241.
Bamford DH 2003. Do viruses form lineages across different domains of life? Res Microbiol 154: 231-236.

Bamford DH, Grimes JM, Stuart DI 2005. What does structure tell us about virus evolution? Curr Opin Struct Biol 15: 655-663.

Barr JJ, Auro R, Furlan M, Whiteson KL, Erb ML, Pogliano J, Stotland A, Wolkowicz R, Cutting AS, Doran KS, Salamon P, Youle M, Rohwer F 2013. Bacteriophage adhering to mucus provide a non-host-derived immunity. Proc Natl Acad Sci U S A 110: 10771-10776.

Barton NH, Charlesworth B 1998. Why sex and recombination? Science 281: 1986-1990.

Barve A, Wagner A 2013. A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature 500: 203-+.

Battistuzzi FU, Feijao A, Hedges SB 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evol Biol 4: 44.

Bensing BA, Siboo IR, Sullam PM 2001. Proteins PblA and PblB of Streptococcus mitis, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. Infect Immun 69: 6186-6192.

Benzer S 1955. Fine Structure of a Genetic Region in Bacteriophage. Proc Natl Acad Sci U S A 41: 344-354.

Better M, Freifelder D 1983. Studies on the replication of Escherichia coli phage lambda DNA. I. The kinetics of DNA replication and requirements for the generation of rolling circles. Virology 126: 168-182.

Bianco PR, Tracy RB, Kowalczykowski SC 1998. DNA strand exchange proteins: a biochemical and physical comparison. Front Biosci 3: D570-603.

Biers EJ, Wang K, Pennington C, Belas R, Chen F, Moran MA 2008. Occurrence and expression of gene transfer agent genes in marine bacterioplankton. Appl Environ Microbiol 74: 2933-2939.

Bigot S, Saleh OA, Lesterlin C, Pages C, El Karoui M, Dennis C, Grigoriev M, Allemand JF, Barre FX, Cornet F 2005. KOPS: DNA motifs that control E. coli chromosome segregation by orienting the FtsK translocase. Embo J 24: 3770-3780.

Bikard D, Marraffini LA 2012. Innate and adaptive immunity in bacteria: mechanisms of programmed genetic variation to fight bacteriophages. Curr Opin Immunol 24: 15-20.

Biller SJ, Schubotz F, Roggensack SE, Thompson AW, Summons RE, Chisholm SW 2014. Bacterial vesicles in marine ecosystems. Science 343: 183-186.

Bjedov I, Tenaillon O, Gerard B, Souza V, Denamur E, Radman M, Taddei F, Matic I 2003. Stress-induced mutagenesis in bacteria. Science 300: 1404-1409.

Blakely G, May G, McCulloch R, Arciszewska LK, Burke M, Lovett ST, Sherratt DJ 1993. Two related recombinases are required for site-specific recombination at dif and cer in E. coli K12. Cell 75: 351-361.

Bonemann G, Pietrosiuk A, Mogk A 2010. Tubules and donuts: a type VI secretion story. Mol Microbiol 76: 815-821.

Botstein D 1980. A theory of modular evolution for bacteriophages. Ann N Y Acad Sci 354: 484-490.

Bouchard JD, Moineau S 2004. Lactococcal phage genes involved in sensitivity to AbiK and their relation to single-strand annealing proteins. J Bacteriol 186: 3649-3652.

Boyd EF, Carpenter MR, Chowdhury N 2012. Mobile effector proteins on phage genomes. Bacteriophage 2: 139-148.

Boyer F, Fichant G, Berthod J, Vandenbrouck Y, Attree I 2009. Dissecting the bacterial type VI secretion system by a genome wide in silico analysis: what can be learned from available microbial genomic resources? BMC Genomics 10: 104.

Bradley DE 1967. Ultrastructure of bacteriophage and bacteriocins. Bacteriol Rev 31: 230-314.

Brown SP, Le Chat L, De Paepe M, Taddei F 2006. Ecology of microbial invasions: amplification allows virus carriers to invade more rapidly when rare. Current Biology 16: 2048-2052.

Brussow H, Canchaya C, Hardt WD 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68: 560-602.

Burrus V, Pavlovic G, Decaris B, Guedon G 2002. Conjugative transposons: the tip of the iceberg. Mol Microbiol 46: 601-610.

Burrus V, Waldor MK 2004. Shaping bacterial genomes with integrative and conjugative elements. Res Microbiol 155: 376-386.

Campbell A. 1977. Defective Bacteriophages and Imcomplete Prophages. In: Fraenkel-Conrat H, Wagner RR, editors. Regulation and Genetics. New-York: Springer US. p. 259-328.

Campbell A. 1969. Episomes. New York: Harper and Row.

Campbell A. 2002. Eubacterial genomes. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington DC: American Society of Microbiology. p. 1024-1039.

Campbell A 2003. The future of bacteriophage biology. Nat Rev Genet 4: 471-477.

Campbell A. 2006. General Aspects of Lysogeny. In: Calendar RL, Abedon ST, editors. The Bacteriophages. New York: Oxford University Press. p. 66-73.

Campbell A 2007. Phage integration and chromosome structure. A personal history. Annu Rev Genet 41: 1-11.

Campbell AM. 1996. Bacteriophages. In. Escherichia coli and Salmonella: cellular and molecular biology. Washington DC: ASM Press. p. 2325-2338.

Canchaya C, Fournous G, Brussow H 2004. The impact of prophages on bacterial chromosomes. Mol Microbiol 53: 9-18.

Casjens S 2003. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol 49: 277-300.

Casjens S, Hayden M 1988. Analysis in vivo of the bacteriophage P22 headful nuclease. J Mol Biol 199: 467-474.

Casjens S, Hendrix R 1974. Comments on the arrangement of the morphogenetic genes of bacteriophage lambda. J Mol Biol 90: 20-25.

Casjens S, Palmer N, van Vugt R, Huang WM, Stevenson B, Rosa P, Lathigra R, Sutton G, Peterson J, Dodson RJ, Haft D, Hickey E, Gwinn M, White O, Fraser CM 2000. A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete Borrelia burgdorferi. Mol Microbiol 35: 490-516.

Casjens S, Sampson L, Randall S, Eppler K, Wu H, Petri JB, Schmieger H 1992. Molecular genetic analysis of bacteriophage P22 gene 3 product, a protein involved in the initiation of headful DNA packaging. J Mol Biol 227: 1086-1099.

Casjens SR 2008. Diversity among the tailed-bacteriophages that infect the Enterobacteriaceae. Res Microbiol 159: 340-348.

Casjens SR 2011. The DNA-packaging nanomotor of tailed bacteriophages. Nat Rev Microbiol 9: 647-657.

Chaconas G, Kobryn K 2010. Structure, function, and evolution of linear replicons in Borrelia. Annu Rev Microbiol 64: 185-202.

Chang YJ, Land M, Hauser L, Chertkov O, Del Rio TG, Nolan M, Copeland A, Tice H, Cheng JF, Lucas S, Han C, Goodwin L, Pitluck S, Ivanova N, Ovchinikova G, Pati A, Chen A, Palaniappan K, Mavromatis K, Liolios K, Brettin T, Fiebig A, Rohde M, Abt B, Goker M, Detter JC, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk HP, Lapidus A 2011. Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium Ktedonobacter racemifer type strain (SOSP1-21). Stand Genomic Sci 5: 97-111.

Charlesworth B, Barton N 2004. Genome size: does bigger mean worse? Current Biology 14: R233-235.

Charneski CA, Honti F, Bryant JM, Hurst LD, Feil EJ 2011. Atypical at skew in Firmicute genomes results from selection and not from mutation. PLoS Genet 7: e1002283.

Chen I, Christie PJ, Dubnau D 2005. The ins and outs of DNA transfer in bacteria. Science 310: 1456-1460.

Choulet F, Aigle B, Gallois A, Mangenot S, Gerbaud C, Truong C, Francou FX, Fourrier C, Guerineau M, Decaris B, Barbe V, Pernodet JL, Leblond P 2006. Evolution of the terminal regions of the Streptomyces linear chromosome. Mol Biol Evol 23: 2361-2369.

Christie GE, Dokland T 2012. Pirates of the Caudovirales. Virology 434: 210-221.

Clark AJ, Inwood W, Cloutier T, Dhillon TS 2001. Nucleotide sequence of coliphage HK620 and the evolution of lambdoid phages. J Mol Biol 311: 657-679.

Coetzee HL, De Klerk HC, Coetzee JN, Smit JA 1968. Bacteriophage-tail-like particles associated with intra-species killing of Proteus vulgaris. J Gen Virol 2: 29-36.

Cooper S, Helmstetter CE 1968. Chromosome replication and the division cycle of Escherichia coli B/r. J Mol Biol 31: 519-540.

Cortez D, Forterre P, Gribaldo S 2009. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol 10: R65.

Court D, Oppenheim A. 1983. Phage Lambda's accessory Genes. In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA, editors. Lambda II. Cold Spring Harbour: CSHL. p. 251-278.
Court R, Cook N, Saikrishnan K, Wigley D 2007. The crystal structure of lambda-Gam protein suggests a model for RecBCD inhibition. J Mol Biol 371: 25-33.

Couturier E, Rocha EP 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol 59: 1506-1518.

Curtis TP, Sloan WT, Scannell JW 2002. Estimating prokaryotic diversity and its limits. Proc Natl Acad Sci U S A 99: 10494-10499.

D'Herelle F 2007. On an invisible microbe antagonistic toward dysenteric bacilli: brief note by Mr. F. D'Herelle, presented by Mr. Roux. 1917. Res Microbiol 158: 553-554.

Dame RT, Kalmykowa OJ, Grainger DC 2011. Chromosomal macrodomains and associated proteins: implications for DNA organization and replication in gram negative bacteria. PLoS Genet 7: e1002123.

Daubin V, Moran NA 2004. Comment on "The origins of genome complexity". Science 306: 978; author reply 978.

Daubin V, Ochman H 2004a. Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res 14: 1036-1042.

Daubin V, Ochman H 2004b. Start-up entities in the origin of new genes. Curr Opin Genet Dev 14: 616-619.

Dawkins R. 1982. The extended phenotype. Oxford, New York: Oxford University Press.

De Paepe M, Hutinet G, Son O, Amarir-Bouhram J, Schbath S, Petit MA 2014. Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-like recombinases. PLoS Genet 10: e1004181.

Delk AS, Dekker CA 1972. Characterization of rhapidosomes of Saprospira grandis. J Mol Biol 64: 287-295.

Deveau H, Garneau JE, Moineau S 2010. CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol 64: 475-493.

DiGate RJ, Marians KJ 1988. Identification of a potent decatenating enzyme from Escherichia coli. J Biol Chem 263: 13366-13373.

Dillingham MS, Kowalczykowski SC 2008. RecBCD enzyme and the repair of double-stranded DNA breaks. Microbiol Mol Biol Rev 72: 642-671.

Dobrindt U, Hochhut B, Hentschel U, Hacker J 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2: 414-424.

Donnenberg MS. 2002. Escherichia coli. Virulence mechanisms of a versatile pathogen.: Baltimore: Academic press, Elsevier Science.

Drexler K, Riede I, Montag D, Eschbach ML, Henning U 1989. Receptor specificity of the Escherichia coli T-even type phage Ox2. Mutational alterations in host range mutants. J Mol Biol 207: 797-803.

Dubey GP, Ben-Yehuda S 2011. Intercellular nanotubes mediate bacterial communication. Cell 144: 590-600.

Duckworth DH 1970. The metabolism of T4 phage ghost-infected cells. I. Macromolecular synthesis and ransport of nucleic acid and protein precursors. Virology 40: 673-684.

Ebel-Tsipis J, Botstein D, Fox MS 1972a. Generalized transduction by phage P22 in Salmonella typhimurium. I. Molecular origin of transducing DNA. J Mol Biol 71: 433-448.

Ebel-Tsipis J, Fox MS, Botstein D 1972b. Generalized transduction by bacteriophage P22 in Salmonella typhimurium. II. Mechanism of integration of transducing DNA. J Mol Biol 71: 449-469.

Edlin G, Lin L, Bitner R 1977. Reproductive fitness of P1, P2, and Mu lysogens of Escherichia coli. J Virol 21: 560-564.

Edwards RA, Rohwer F 2005. Viral metagenomics. Nature Rev Microbiol 3: 504-510.

Eisen JA, Heidelberg JF, White O, Salzberg SL 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 1: RESEARCH0011.

Eiserling F, Pushkin A, Gingery M, Bertani G 1999. Bacteriophage-like particles associated with the gene transfer agent of methanococcus voltae PS. J Gen Virol 80: 3305-3308.

Ellis HM, Yu D, DiTizio T, Court DL 2001. High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. Proc Natl Acad Sci U S A 98: 6742-6746.

Enea V, Horiuchi K, Turgeon BG, Zinder ND 1977. Physical map of defective interfering particles of bacteriophage f1. J Mol Biol 111: 395-414.

Enquist LW, Skalka A 1973. Replication of Bacteriophage-Lambda DNA-Dependent on Function of Host and Viral Genes .1. Interaction of Red, Gam and Rec. J Mol Biol 75: 185-212.

Enright AJ, Van Dongen S, Ouzounis CA 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575-1584.

Esnault E, Valens M, Espeli O, Boccard F 2007. Chromosome Structuring Limits Genome Plasticity in Escherichia coli. PLoS Genet 3: e226.

Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, Zhou J, Spratt BG 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A 98: 182-187.

Feiss M, Becker A. 1983. DNA Packaging and Cutting. In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA, editors. Lambda II. Cold Spring Harbour: CSHL. p. 305-330.

Fischer D, Eisenberg D 1999. Finding families for genomic ORFans. Bioinformatics 15: 759-762.

Flynn KM, Vohr SH, Hatcher PJ, Cooper VS 2010. Evolutionary rates and gene dispensability associate with replication timing in the archaeon Sulfolobus islandicus. Genome Biol Evol 2: 859-869.

Forget AL, Kowalczykowski SC 2012. Single-molecule imaging of DNA pairing by RecA reveals a three-dimensional homology search. Nature 482: 423-427.

Forterre P, Soler N, Krupovic M, Marguet E, Ackermann HW 2013. Fake virus particles generated by fluorescence microscopy. Trends Microbiol 21: 1-5.

Fouts DE 2006. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res 34: 5839-5851.

Fraser JS, Yu Z, Maxwell KL, Davidson AR 2006. Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J Mol Biol 359: 496-507.

Fuhrman JA 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399: 541-548.

Fukuyo M, Sasaki A, Kobayashi I 2012. Success of a suicidal defense strategy against infection in a structured habitat. Sci Rep 2: 238.

Furth M, WIckner S. 1983. Lambda DNA Replication. In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA, editors. Lambda II. Cold Spring Harbour: CSHL. p. 145-174.

Gama JA, Reis AM, Domingues I, Mendes-Soares H, Matos AM, Dionisio F 2013. Temperate Bacterial Viruses as Double-Edged Swords in Bacterial Warfare. PLoS One 8: e59043.

Gan W, Guan Z, Liu J, Gui T, Shen K, Manley JL, Li X 2011. R-loop-mediated genomic instability is caused by impairment of replication fork progression. Genes Dev 25: 2041-2056.

Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH, Moineau S 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 468: 67-71.

Garro AJ, Marmur J 1970. Defective bacteriophages. J Cell Physiol 76: 253-263.

Gascuel O 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14: 685-695.

Gebhart D, Williams SR, Bishop-Lilly KA, Govoni GR, Willner KM, Butani A, Sozhamannan S, Martin D, Fortier LC, Scholl D 2012. Novel High-Molecular-Weight, R-Type Bacteriocins of Clostridium difficile. J Bacteriol 194: 6240-6247.

Gomez-Valero L, Rocha EP, Latorre A, Silva FJ 2007. Reconstructing the ancestor of Mycobacterium leprae: the dynamics of gene loss and genome reduction. Genome Res 17: 1178-1185.

Gottesman ME, Yarmolinsky MB 1968. Integration-negative mutants of bacteriophage lambda. J Mol Biol 31: 487-505.

Gould SJ, Vrba E 1982. Exaptation - a missing term in the science of form. Paleobiology 8: 4-15.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9: r43-74.

Guarneros G, Echols H 1970. New mutants of bacteriophage lambda with a specific defect in excision from the host chromosome. J Mol Biol 47: 565-574.

Gussin G, Johnson A, Pabo C, Sauer R. 1983. Repressor and Cro Protein: Structure, Function, and Role in Lysogenization. In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA, editors. Lambda II. Cold Spring Harbour: CSHL. p. 93-122.

Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 23: 1089-1097.

Hallet B, Sherratt DJ 1997. Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. Fems Microbiol Rev 21: 157-178.

Hamilton WD 1964a. The genetical evolution of social behaviour. I. Journal of Theoretical Biology 7: 1-16.

Hamilton WD 1964b. The genetical evolution of social behaviour. II. Journal of Theoretical Biology 7: 17-52.

Hassan F, Kamruzzaman M, Mekalanos JJ, Faruque SM 2010. Satellite phage TLCphi enables toxigenic conversion by CTX phage through dif site alteration. Nature 467: 982-985.

Hauser R, Blasche S, Dokland T, Haggard-Ljungquist E, von Brunn A, Salas M, Casjens S, Molineux I, Uetz P 2012. Bacteriophage protein-protein interactions. Advances in Virus Research, Vol 83: Bacteriophages, Pt B 83: 219-298.

Henderson D, Weil J 1975. Recombination-deficient deletions in bacteriophage lambda and their interaction with chi mutations. Genetics 79: 143-174.

Hendrickson H, Lawrence JG 2006. Selection for chromosome architecture in bacteria. J Mol Evol 62: 615-629.

Hendrix RW, Casjens S. 2006. Bacteriophage Lambda and its Genetic Neighborhood. In: Abedon ST, Calendar RL, editors. The Bacteriophages. New York: Oxford University Press. p. 409-447.

Hendrix RW, Smith MCM, Burns RN, Ford ME, Hatfull GF 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci USA 96: 2192-2197.

Hershey AD, Chase M 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol 36: 39-56.

Higgins NP, Peebles CL, Sugino A, Cozzarelli NR 1978. Purification of subunits of Escherichia coli DNA gyrase and reconstitution of enzymatic activity. Proc Natl Acad Sci U S A 75: 1773-1777.

Hobman JL, Penn CW, Pallen MJ 2007. Laboratory strains of Escherichia coli: model citizens or deceitful delinquents growing old disgracefully? Mol Microbiol 64: 881-885.

Hofnung M, Jezierska A, Braun-Breton C 1976. lamB mutations in E. coli K12: growth of lambda host range mutants and effect of nonsense suppressors. Mol Gen Genet 145: 207-213.

Hsiao WW, Ung K, Aeschliman D, Bryan J, Finlay BB, Brinkman FS 2005. Evidence of a Large Novel Gene Pool Associated with Prokaryotic Genomic Islands. PLoS Genet 1: e62.

Humphrey SB, Stanton TB, Jensen NS, Zuerner RL 1997. Purification and characterization of VSH-1, a generalized transducing bacteriophage of Serpulina hyodysenteriae. J Bacteriol 179: 323-329.

Hurst MR, Glare TR, Jackson TA 2004. Cloning Serratia entomophila antifeeding genes--a putative defective prophage active against the grass grub Costelytra zealandica. J Bacteriol 186: 5116-5128.

Iida S, Hiestand-Nauer R, Sandmeier H, Lehnherr H, Arber W 1998. Accessory genes in the darA operon of bacteriophage P1 affect antirestriction function, generalized transduction, head morphogenesis, and host cell lysis. Virology 251: 49-58.

Isambert H, Stein RR 2009. On the need for widespread horizontal gene transfers under genome size constraint. Biol Direct 4: 28.

Jacquemin-Sablon A, Lanni YT 1973. Lambda-repressed mutants of bacteriophage T5. I. Isolation and genetical characterization. Virology 56: 230-237.

Jain R, Rivera MC, Lake JA 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96: 3801-3806.

Jain R, Rivera MC, Moore JE, Lake JA 2002. Horizontal gene transfer in microbial genome evolution. Theor Popul Biol 61: 489-495.

Jardine PJ, Anderson DL. 2006. DNA Packaging in Double-Stranded DNA Phages. In: Abedon ST, Calendar RL, editors. The Bacteriophages. New York: Oxford University Press. p. 49-65.

Jiang SC, Paul JH 1994. Seasonal and Diel Abundance of Viruses and Occurrence of Lysogeny/Bacteriocinogeny in the Marine-Environment. Marine Ecology Progress Series 104: 163-172.

Joo J, Gunny M, Cases M, Hudson P, Albert R, Harvill E 2006. Bacteriophage-mediated competition in Bordetella bacteria. Proc Biol Sci 273: 1843-1848.

Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol 299: 27-51.

Juhas M, Power PM, Harding RM, Ferguson DJ, Dimopoulou ID, Elamin AR, Mohd-Zain Z, Hood DW, Adegbola R, Erwin A, Smith A, Munson RS, Harrison A, Mansfield L, Bentley S, Crook DW 2007. Sequence and functional analyses of Haemophilus spp. genomic islands. Genome Biol 8: R237.

Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. Fems Microbiol Rev 33: 376-393.

Karlin S 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol 9: 335-343.

Kimura M 1968. Evolutionary rate at the molecular level. Nature 217: 624-626.

Kitani T, Yoda K, Ogawa T, Okazaki T 1985. Evidence that discontinuous DNA replication in Escherichia coli is primed by approximately 10 to 12 residues of RNA starting with a purine. J Mol Biol 184: 45-52.

Kobayashi I 2001. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res 29: 3742-3756.

Koga M, Otsuka Y, Lemire S, Yonesaki T 2011. Escherichia coli rnlA and rnlB compose a novel toxin-antitoxin system. Genetics 187: 123-130.

Kolling GL, Matthews KR 1999. Export of virulence genes and Shiga toxin by membrane vesicles of Escherichia coli O157:H7. Appl Environ Microbiol 65: 1843-1848.

Kowalczykowski SC 2000. Initiation of genetic recombination and recombination-dependent replication. Trends Biochem Sci 25: 156-165.

Krupovic M, Bamford DH 2011. Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. Curr Opin Virol 1: 118-124.

Kuo CH, Moran NA, Ochman H 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res 19: 1450-1454.

Kuo CH, Ochman H 2010. The Extinction Dynamics of Bacterial Pseudogenes. PLoS Genet 6: e1001050.

Kuzminov A 1999. Recombinational repair of DNA damage in Escherichia coli and bacteriophage lambda. Microbiol Mol Biol Rev 63: 751-813.

Labrie SJ, Samson JE, Moineau S 2010. Bacteriophage resistance mechanisms. Nature Rev Microbiol 8: 317-327.

Laing CR, Zhang Y, Gilmour MW, Allen V, Johnson R, Thomas JE, Gannon VP 2012. A comparison of Shiga-toxin 2 bacteriophage from classical enterohemorrhagic Escherichia coli serotypes and the German E. coli O104:H4 outbreak strain. PLoS One 7: e37362.

Landy A 1989. Dynamic, structural, and regulatory aspects of lambda site-specific recombination. Annu Rev Biochem 58: 913-949.

Lang AS, Beatty JT 2007. Importance of widespread gene transfer agent genes in alpha-proteobacteria. Trends Microbiol 15: 54-62.

Lang AS, Zhaxybayeva O, Beatty JT 2012. Gene transfer agents: phage-like elements of genetic exchange. Nat Rev. Microbiol 10: 472-482.

Langille MG, Hsiao WW, Brinkman FS 2010. Detecting genomic islands using bioinformatics approaches. Nat Rev Microbiol 8: 373-382.

Lavigne R, Darius P, Summer EJ, Seto D, Mahadevan P, Nilsson AS, Ackermann HW, Kropinski AM 2009. Classification of Myoviridae bacteriophages using protein sequence similarity. BMC Microbiol 9: 224.

Lawrence JG, Hatfull GF, Hendrix RW 2002. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol 184: 4891-4905.

Lawrence JG, Hendrickson H 2003. Lateral gene transfer: when will adolescence end? Mol Microbiol 50: 739-749.

Lawrence JG, Ochman H 1997. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44: 383-397.

Lederberg EM, Lederberg J 1953. Genetic Studies of Lysogenicity in Escherichia Coli. Genetics 38: 51-64.

Lennox ES 1955. Transduction of linked genetic characters of the host by bacteriophage P1. Virology 1: 190-206.

Lerat E, Daubin V, Ochman H, Moran NA 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol 3: e130.

Lerat E, Ochman H 2004. Psi-Phi: exploring the outer limits of bacterial pseudogenes. Genome Res 14: 2273-2278.

Lerat E, Ochman H 2005. Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Res 33: 3125-3132.

Levy O, Ptacin JL, Pease PJ, Gore J, Eisen MB, Bustamante C, Cozzarelli NR 2005. Identification of oligonucleotide sequences that direct the movement of the Escherichia coli FtsK translocase. Proc Natl Acad Sci U S A 102: 17618-17623.

Lima-Mendez G, Van Helden J, Toussaint A, Leplae R 2008a. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics 24: 863-865.

Lima-Mendez G, Van Helden J, Toussaint A, Leplae R 2008b. Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol 25: 762-777.

Lin L, Bitner R, Edlin G 1977. Increased reproductive fitness of Escherichia coli lambda lysogens. J Virol 21: 554-559.

Liu GR, Liu WQ, Johnston RN, Sanderson KE, Li SX, Liu SL 2006. Genome plasticity and ori-ter rebalancing in Salmonella typhi. Mol Biol Evol 23: 365-371.

Lobocka MB, Rose DJ, Plunkett G, 3rd, Rusin M, Samojedny A, Lehnherr H, Yarmolinsky MB, Blattner FR 2004. Genome of bacteriophage P1. J Bacteriol 186: 7032-7068.

Lobry JR 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13: 660-665.

Lopes A, Amarir-Bouhram J, Faure G, Petit MA, Guerois R 2010. Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. Nucleic Acids Res 38: 3952-3962.

Los JM, Los M, Wegrzyn A, Wegrzyn G 2012. Altruism of Shiga toxin-producing Escherichia coli: recent hypothesis versus experimental results. Front Cell Infect Microbiol 2: 166.

Lowe J, Ellonen A, Allen MD, Atkinson C, Sherratt DJ, Grainge I 2008. Molecular mechanism of sequence-directed DNA loading and translocation by FtsK. Mol Cell 31: 498-509.

Lozupone CA, Knight R 2007. Global patterns in bacterial diversity. Proc Natl Acad Sci U S A 104: 11436-11440.

Luria SE, Delbruck M 1943. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics 28: 491-511.

Lynch M 2006. Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol 60: 327-349.

Lynch M, Conery JS 2000. The evolutionary fate and consequences of duplicate genes. Science 290: 1151-1155.

Lynch M, Conery JS 2003. The origins of genome complexity. Science 302: 1401-1404.

Maisnier-Patin S, Nordstrom K, Dasgupta S 2001. Replication arrests during a single round of replication of the Escherichia coli chromosome in the absence of DnaC activity. Mol Microbiol 42: 1371-1382.

Mann NH, Cook A, Millard A, Bailey S, Clokie M 2003. Marine ecosystems: bacterial photosynthesis genes in a virus. Nature 424: 741.

Maresca M, Erler A, Fu J, Friedrich A, Zhang YM, Stewart AF 2010. Single-stranded heteroduplex intermediates in lambda Red homologous recombination. BMC Molecular Biology 11: 54.

Martinsohn JT, Radman M, Petit MA 2008. The lambda Red proteins promote efficient recombination between diverged sequences: Implications for bacteriophage genome mosaicism. PLoS Genet 4: e1000065.

Masel J 2011. Genetic drift. Current Biology 21: R837-838.

Matos RC, Lapaque N, Rigottier-Gois L, Debarbieux L, Meylheuc T, Gonzalez-Zorn B, Repoila F, Lopes Mde F, Serror P 2013. Enterococcus faecalis prophage dynamics and contributions to pathogenic traits. PLoS Genet 9: e1003539.

Matthews TD, Maloy S 2010. Fitness effects of replichore imbalance in Salmonella enterica. J Bacteriol 192: 6086-6088.

McCutcheon JP, Moran NA 2012. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10: 13-26.

Meile JC, Mercier R, Stouf M, Pages C, Bouet JY, Cornet F 2011. The terminal region of the E. coli chromosome localises at the periphery of the nucleoid. BMC Microbiol 11: 28.

Menouni R, Champ S, Espinosa L, Boudvillain M, Ansaldi M 2013. Transcription termination controls prophage maintenance in Escherichia coli genomes. Proc Natl Acad Sci U S A 110: 14414-14419.

Mercier R, Petit MA, Schbath S, Robin S, El Karoui M, Boccard F, Espeli O 2008. The MatP/matS site-specific system organizes the terminus region of the E. coli chromosome into a macrodomain. Cell 135: 475-485.

Merrikh H, Zhang Y, Grossman AD, Wang JD 2012. Replication-transcription conflicts in bacteria. Nat Rev Microbiol 10: 449-458.

Michel-Briand Y, Baysse C 2002. The pyocins of Pseudomonas aeruginosa. Biochimie 84: 499-510.

Middendorf B, Hochhut B, Leipold K, Dobrindt U, Blum-Oehler G, Hacker J 2004. Instability of pathogenicity islands in uropathogenic Escherichia coli 536. J Bacteriol 186: 3086-3096.

Mira A, Ochman H, Moran NA 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet 17: 589-596.

Mirkin EV, Mirkin SM 2005. Mechanisms of transcription-replication collisions in bacteria. Mol Cell Biol 25: 888-895.

Mizuuchi K 1992. Transpositional recombination: mechanistic insights from studies of mu and other elements. Annu Rev Biochem 61: 1011-1051.

Mmolawa PT, Schmieger H, Heuzenroeder MW 2003. Bacteriophage ST64B, a genetic mosaic of genes from diverse sources isolated from Salmonella enterica serovar typhimurium DT 64. J Bacteriol 185: 6481-6485.

Modi SR, Lee HH, Spina CS, Collins JJ 2013. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature 499: 219-222.

Mokili JL, Rohwer F, Dutilh BE 2012. Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2: 63-77.

Molineux IJ, Panja D 2013. Popping the cork: mechanisms of phage genome ejection. Nat Rev Microbiol 11: 194-204.

Morimatsu K, Kowalczykowski SC 2003. RecFOR proteins load RecA protein onto gapped DNA to accelerate DNA strand exchange: a universal step of recombinational repair. Mol Cell 11: 1337-1347.

Muniesa M, Colomer-Lluch M, Jofre J 2013. Potential impact of environmental bacteriophages in spreading antibiotic resistance genes. Future Microbiol 8: 739-751.

Murphy KC 2012. Phage Recombinases and Their Applications. Advances in Virus Research, Vol 83: Bacteriophages 83: 367-414.

Myers RS, Stahl FW 1994. Chi and the RecBC D enzyme of Escherichia coli. Annu Rev Genet 28: 49-70.

Myung H, Calendar R 1995. The Old Exonuclease of Bacteriophage-P2. J Bacteriol 177: 497-501.

Nafissi N, Slavcev R 2014. Bacteriophage recombination systems and biotechnical applications. Appl Microbiol Biotechnol.

Nakayama K, Takashima K, Ishihara H, Shinomiya T, Kageyama M, Kanaya S, Ohnishi M, Murata T, Mori H, Hayashi T 2000. The R-type pyocin of Pseudomonas aeruginosa is related to P2 phage, and the F-type is related to lambda phage. Mol Microbiol 38: 213-231.

Napolitano MG, Almagro-Moreno S, Boyd EF 2011. Dichotomy in the evolution of pathogenicity island and bacteriophage encoded integrases from pathogenic Escherichia coli strains. Infect Genet Evol 11: 423-436.

Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, Fang FC 2006. Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Science 313: 236-238.

Nicolas P, Bize L, Muri F, Hoebeke M, Rodolphe F, Ehrlich SD, Prum B, Bessieres P 2002. Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models. Nucleic Acids Res 30: 1418-1426.

Nikolaou C, Almirantis Y 2005. A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species. Nucleic Acids Res 33: 6816-6822.

Nilsson A, Haggard Ljungquist E. 2006. The P2-Like Bacteriophages. In: Calendar RL, Abedon ST, editors. The Bacteriophages. New York: Oxford University Press. p. 365-390.

Nilsson AS, Haggard-Ljungquist E 2007. Evolution of P2-like phages and their impact on bacterial evolution. Res Microbiol 158: 311-317.

Norregaard K, Andersson M, Sneppen K, Nielsen PE, Brown S, Oddershede LB 2014. Effect of supercoiling on the lambda switch. Bacteriophage 4: e27517.

Novick RP, Christie GE, Penades JR 2010. The phage-related chromosomal islands of Gram-positive bacteria. Nat Rev Microbiol 8: 541-551.

Novoa EM, Ribas de Pouplana L 2012. Speeding with control: codon usage, tRNAs, and ribosomes. Trends Genet 28: 574-581.

Ochman H, Davalos LM 2006. The nature and dynamics of bacterial genomes. Science 311: 1730-1733.

Ochman H, Lerat E, Daubin V 2005. Examining bacterial species under the specter of gene transfer and exchange. Proc Natl Acad Sci U S A 102: 6595-6599.

Okada K, Iida T, Kita-Tsukamoto K, Honda T 2005. Vibrios commonly possess two chromosomes. J Bacteriol 187: 752-757.

Otsuka Y, Yonesaki T 2012. Dmd of bacteriophage T4 functions as an antitoxin against Escherichia coli LsoA and RnlA toxins. Mol Microbiol 83: 669-681.

Pao CC, Speyer JF 1975. Mutants of T7 bacteriophage inhibited by lambda prophage. Proc Natl Acad Sci U S A 72: 3642-3646.

Petrova M, Shcherbatova N, Kurakov A, Mindlin S 2013. Genomic characterization and integrative properties of phiSMA6 and phiSMA7, two novel filamentous bacteriophages of Stenotrophomonas maltophilia. Arch Virol.

Pope WH, Jacobs-Sera D, Best AA, Broussard GW, Connerly PL, Dedrick RM, Kremer TA, Offner S, Ogiefo AH, Pizzorno MC, Rockenbach K, Russell DA, Stowe EL, Stukey J, Thibault SA, Conway JF, Hendrix RW, Hatfull GF 2013. Cluster J mycobacteriophages: intron splicing in capsid and tail genes. PLoS One 8: e69273.

Postow L, Hardy CD, Arsuaga J, Cozzarelli NR 2004. Topological domain structure of the Escherichia coli chromosome. Genes Dev 18: 1766-1779.

Ptashne M. 1992. Genetic Switch: Phage Lambda and Higher Organisms. Cambridge, MA: Blackwell.

Ptashne M 2011. Principles of a switch. Nat Chem Biol 7: 484-487.

Radman-Livaja M, Biswas T, Ellenberger T, Landy A, Aihara H 2006. DNA arms do the legwork to ensure the directionality of lambda site-specific recombination. Curr Opin Struct Biol 16: 42-50.

Rakonjac J, Bennett NJ, Spagnuolo J, Gagic D, Russel M 2011. Filamentous bacteriophage: biology, phage display and nanotechnology applications. Curr Issues Mol Biol 13: 51-76.

Ranade K, Poteete AR 1993. Superinfection exclusion (sieB) genes of bacteriophages P22 and lambda. J Bacteriol 175: 4712-4718.

Rapp BJ, Wall JD 1987. Genetic Transfer in Desulfovibrio-Desulfuricans. Proceedings of the National Academy of Sciences of the United States of America 84: 9128-9130.

Ravin NV 2011. N15: the linear phage-plasmid. Plasmid 65: 102-109.

Refardt D 2011. Within-host competition determines reproductive success of temperate bacteriophages. ISME J 5: 1451-1460.

Remmert M, Biegert A, Hauser A, Soding J 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9: 173-175.

Roca J 1995. The mechanisms of DNA topoisomerases. Trends Biochem Sci 20: 156-160.

Rocha EP 2004. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res 14: 2279-2286.

Rocha EP 2008. The organization of the bacterial genome. Annu Rev Genet 42: 211-233.

Rocha EP, Danchin A 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 21: 108-116.

Rocha EP, Danchin A 2002. Base composition bias might result from competition for metabolic resources. Trends Genet 18: 291-294.

Rocha EP, Danchin A 2003a. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res 31: 6570-6577.

Rocha EP, Touchon M, Feil EJ 2006a. Similar compositional biases are caused by very different mutational effects. Genome Res 16: 1537-1547.

Rocha EPC, Danchin A 2003b. Essentiality, not expressiveness, drives gene strand bias in bacteria. Nature Genet 34: 377-378.

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ 2006b. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239: 226-235.

Rohwer F, Edwards R 2002. The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage. J Bacteriol 184: 4529-4535.

Rohwer F, Thurber RV 2009. Viruses manipulate the marine environment. Nature 459: 207-212.

Roux S, Krupovic M, Debroas D, Forterre P, Enault F 2013. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biol 3: 130160.

Rumbo C, Fernandez-Moreira E, Merino M, Poza M, Mendez JA, Soares NC, Mosquera A, Chaves F, Bou G 2011. Horizontal transfer of the OXA-24 carbapenemase gene via outer membrane vesicles: a new mechanism of dissemination of carbapenem resistance genes in Acinetobacter baumannii. Antimicrob Agents Chemother 55: 3084-3090.

Ruzin A, Lindsay J, Novick RP 2001. Molecular genetics of SaPI1--a mobile pathogenicity island in Staphylococcus aureus. Mol Microbiol 41: 365-377.

Saha RP, Lou Z, Meng L, Harshey RM 2013. Transposable prophage Mu is organized as a stable chromosomal domain of E. coli. PLoS Genet 9: e1003902.

Samson JE, Magadan AH, Sabri M, Moineau S 2013. Revenge of the phages: defeating bacterial defences. Nat Rev Microbiol 11: 675-687.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M 1977. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265: 687-695.

Schbath S, Hoebeke M. 2011. R'MES: a tool to find motifs with a significantly unexpected frequency in biological sequences. Singapore: World Scientific.

Schloss PD, Handelsman J 2004. Status of the microbial census. Microbiol Mol Biol Rev 68: 686-691.

Schmid MB, Roth JR 1987. Gene location affects expression level in Salmonella typhimurium. J Bacteriol 169: 2872-2875.

Schmidt H, Hensel M 2004. Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev 17: 14-56.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502-504.

Schmieger H 1972. Phage P22-mutants with increased or decreased transduction abilities. Mol Gen Genet 119: 75-88.

Schubert S, Darlu P, Clermont O, Wieser A, Magistro G, Hoffmann C, Weinert K, Tenaillon O, Matic I, Denamur E 2009. Role of intraspecies recombination in the spread of pathogenicity islands within the Escherichia coli species. PLoS Pathog 5: e1000257.

Seed KD, Lazinski DW, Calderwood SB, Camilli A 2013. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature 494: 489-491.

Serres MH, Kerr AR, McCormack TJ, Riley M 2009. Evolution by leaps: gene duplication in bacteria. Biol Direct 4: 46.

Serwer P, Hayes SJ, Zaman S, Lieman K, Rolando M, Hardies SC 2004. Improved isolation of undersampled bacteriophages: finding of distant terminase genes. Virology 329: 412-424.

Sharp PM 1991. Determinants of DNA sequence divergence between Escherichia coli and Salmonella typhimurium: codon usage, map position, and concerted evolution. J Mol Evol 33: 23-33.

Sharples GJ, Ingleston SM, Lloyd RG 1999. Holliday junction processing in bacteria: insights from the evolutionary conservation of RuvABC, RecG, and RusA. J Bacteriol 181: 5543-5550.

Shen P, Huang HV 1986. Homologous recombination in Escherichia coli: dependence on substrate length and homology. Genetics 112: 441-457.

Shikuma NJ, Pilhofer M, Weiss GL, Hadfield MG, Jensen GJ, Newman DK 2014. Marine tubeworm metamorphosis induced by arrays of bacterial phage tail-like structures. Science 343: 529-533.

Shimada K, Weisberg RA, Gottesman ME 1972. Prophage lambda at unusual chromosomal locations. I. Location of the secondary attachment sites and the properties of the lysogens. J Mol Biol 63: 483-503.

Sivanathan V, Allen MD, de Bekker C, Baker R, Arciszewska LK, Freund SM, Bycroft M, Lowe J, Sherratt DJ 2006. The FtsK gamma domain directs oriented DNA translocation by interacting with KOPS. Nat Struct Mol Biol 13: 965-972.

Six EW, Klug CA 1973. Bacteriophage P4: a satellite virus depending on a helper such as prophage P2. Virology 51: 327-344.

Skalka A, Poonian M, Bartl P 1972. Concatemers in DNA replication: electron microscopic studies of partially denatured intracellular lambda DNA. J Mol Biol 64: 541-550.

Smith GR. 1983. General Recombination. In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA, editors. Lambda II. Cold Spring Harbour: CSHL. p. 175-210.

Smith GR 2012. How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. Microbiol Mol Biol Rev 76: 217-228.

Smith MC, Thorpe HM 2002. Diversity in the serine recombinases. Mol Microbiol 44: 299-307.

Snel B, Huynen MA, Dutilh BE 2005. Genome trees and the nature of genome evolution. Annu Rev Microbiol 59: 191-209.

Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science 318: 1449-1452.

Sousa C, de Lorenzo V, Cebolla A 1997. Modulation of gene expression through chromosomal positioning in Escherichia coli. Microbiology 143: 2071-2078.

Srividhya KV, Alaguraj V, Poornima G, Kumar D, Singh GP, Raghavenderan L, Katta AV, Mehta P, Krishnaswamy S 2007. Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. PLoS One 2: e1193.

Stahl FW 2005. Chi: a little sequence controls a big enzyme. Genetics 170: 487-493.

Strauch E, Kaspar H, Schaudinn C, Dersch P, Madela K, Gewinner C, Hertwig S, Wecke J, Appel B 2001. Characterization of enterocoliticin, a phage tail-like bacteriocin, and its effect on pathogenic Yersinia enterocolitica strains. Appl Environ Microbiol 67: 5634-5642.

Summers WC 2001. Bacteriophage therapy. Annu Rev Microbiol 55: 437-451.

Sun S, Ke R, Hughes D, Nilsson M, Andersson DI 2012. Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. PLoS One 7: e42639.

Susskind MM, Botstein D 1978. Molecular Genetics of Bacteriophage-P22. Microbiol Rev 42: 385-413.

Susskind MM, Botstein D 1980. Superinfection exclusion by lambda prophage in lysogens of Salmonella typhimurium. Virology 100: 212-216.

Suttle CA 2007. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol 5: 801-812.

Szpirer J, Brachet P 1970. [Physiological relationship between the temperate phages lambda and phi80]. Mol Gen Genet 108: 78-92.

Thanbichler M, Wang SC, Shapiro L 2005. The bacterial nucleoid: a highly organized and dynamic structure. J Cell Biochem 96: 506-521.

Thingstad T, Lignell R 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquat Microb Ecol 13: 19-27.

Thompson NE, Pattee PA 1981. Genetic-Transformation in Staphylococcus-Aureus - Demonstration of a Competence-Conferring Factor of Bacteriophage Origin in Bacteriophage-80-Alpha Lysates. J Bacteriol 148: 294-300.

Tillier ER, Collins RA 2000. Genome rearrangement by replication-directed translocation. Nat Genet 26: 195-197.

Toothman P, Herskowitz I 1980. Rex-dependent exclusion of lambdoid phages. I. Prophage requirements for exclusion. Virology 102: 133-146.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, El Karoui M, Frapy E, Garry L, Ghigo J, Gilles A, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit M, Pichon C, Rouy Z, Saint Ruf C, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha E, Denamur E 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 5: e1000344.

Touzain F, Petit MA, Schbath S, El Karoui M 2011. DNA motifs that sculpt the bacterial chromosome. Nat Rev Microbiol 9: 15-26.

Treangen TJ, Rocha EP 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet 7: e1001284.

Valens M, Penaud S, Rossignol M, Cornet F, Boccard F 2004. Macrodomain organization of the Escherichia coli chromosome. Embo J 23: 4330-4341.

Van Valen L 1973. A new evolutionary law. Evolutionary Theory 1: 1-30.

Veesler D, Cambillau C 2011. A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries. Microbiol Mol Biol Rev 75: 423-433.

Vernikos GS, Parkhill J 2008. Resolving the structural features of genomic islands: a machine learning approach. Genome Res 18: 331-342.

Vissa VD, Brennan PJ 2001. The genome of Mycobacterium leprae: a minimal mycobacterial gene set. Genome Biol 2: REVIEWS1023.

Wall JD, Harriman PD 1974. Phage P1 mutants with altered transducing abilities for Escherichia coli. Virology 59: 532-544.

Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, Sturino JM, Wood TK 2010. Cryptic prophages help bacteria cope with adverse environments. Nat Commun 1: 147.

Weisberg R, Landy A. 1983. Site-specific Recombination in Phage Lambda. In: Hendrix RW, Roberts JW, Stahl FW, Weisberg RA, editors. Lambda II. Cold Spring Harbour: CSHL. p. 211-250.

Wiggins PA, Cheveralls KC, Martin JS, Lintner R, Kondev J 2010. Strong intranucleoid interactions organize the Escherichia coli chromosome into a nucleoid filament. Proc Natl Acad Sci U S A 107: 4991-4995.

Williams KP 2002. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res 30: 866-875.

Williams KP, Gillespie JJ, Sobral BW, Nordberg EK, Snyder EE, Shallom JM, Dickerman AW 2010. Phylogeny of gammaproteobacteria. J Bacteriol 192: 2305-2314.

Wood HE, Dawson MT, Devine KM, McConnell DJ 1990. Characterization of PBSX, a defective prophage of Bacillus subtilis. J Bacteriol 172: 2667-2674.

Worcel A, Burgi E 1972. On the structure of the folded chromosome of Escherichia coli. J mol Biol 71: 127-147.

Wozniak RA, Waldor MK 2010. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol 8: 552-563.

Yamamoto T 1967. Presence of rhapidosomes in various species of bacteria and their morphological characteristics. J Bacteriol 94: 1746-1756.

Yang G, Dowling AJ, Gerike U, ffrench-Constant RH, Waterfield NR 2006. Photorhabdus virulence cassettes confer injectable insecticidal activity against the wax moth. J Bacteriol 188: 2254-2261.

Yang Z, Bielawski JP 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496-503.

Yang ZH, Nielsen R 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32-43.

Yen HC, Hu NT, Marrs BL 1979. Characterization of the gene transfer agent made by an overproducer mutant of Rhodopseudomonas capsulata. J Mol Biol 131: 157-168.

Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS 2011. PHAST: a fast phage search tool. Nucleic Acids Res 39: W347-352.

Zinder ND 1955. Bacterial transduction. J Cell Physiol Suppl 45: 23-49.

Zinder ND, Lederberg J 1952. Genetic exchange in Salmonella. J Bacteriol 64: 679-699.

Zink R, Loessner MJ, Scherer S 1995. Characterization of Cryptic Prophages (Monocins) in Listeria and Sequence-Analysis of a Holin/Endolysin Gene. Microbiology 141: 2577-2584.

Annexes

Matériel supplémentaire – Article 1

Supplementary material for the manuscript

The Adaptation of Temperate Bacteriophages to their Host Genomes

by

Louis-Marie Bobay, Eduardo PC Rocha, Marie Touchon

Table of contents:

Table S1. General features of host genomes and their prophage content.

Accession Number	Strain	Phylum	Host Size (bp)	Prophages Size (bp)	Proportion of Prophages (%)	Number of Prophages	Number of breakpoints
NC_000913.2	MG1655	A	4562619	77056	1.66	4	0
AC_000091.1	W3110	A	4569276	77056	1.66	4	4
NC_010473.1	DH10B	A	4595601	90536	1.93	4	0
CP001637.1	DH1	A	4573241	57466	1.24	3	2
AP012030.1	DH1 (ME8569)	A	4539447	81983	1.77	5	0
NC_010468.1	ATCC8739	A	4638437	107781	2.27	4	6
NC_012759.1	BW2952	A	4475503	102656	2.24	4	0
NC_012967.1	REL606	A	4504846	124966	2.70	6	0
NC_012947.1	BL21-Gold-DE3	A	4471463	99475	2.18	5	2
NC_012892.1	BL21	A	4448859	108182	2.37	5	0
CP001509.3	BL21-DE3	A	4454438	104515	2.29	5	0
NC_009800.1	HS	A	4533075	110463	2.38	4	10
FN649414.1	H10407	A	4791769	361666	7.02	9	0
CP002729.1	UMNK88	A	4776537	409879	7.90	11	0
CP002890.1	UMNF18	A	4610149	629058	12.00	18	0
NC_011741.1	IAI1	B1	4563618	136942	2.91	3	0
NC_013361.1	11368	B1	4928070	769170	13.50	18	14
NC_009801.1	E24377A	B1	4784392	195227	3.92	8	0
NC_011748.1	55989	B1	4878769	276093	5.35	7	4
NC_011415.1	SE11	B1	4575256	312259	6.39	7	0
NC_013364.1	11128	B1	4780864	590213	10.99	16	16
NC_013353.1	12009	B1	4823225	626089	11.49	15	8
CP002185.1	W	B1	4630097	270871	5.53	7	0
CP002516.1	KO11	B1	4651116	269052	5.47	7	2
NC_002695.1	Sakai	E	4829253	669197	12.17	17	0
NC_002655.2	EDL933	E	4925825	602620	10.90	16	4
NC_011353.1	EC4115	E	4819256	752819	13.51	20	0
NC_013008.1	TW14359	E	4814113	714023	12.92	19	0
NC_013941.1	CB9615	E	4843122	543230	10.08	13	0
NC_011751.1	UMN026	D	4957071	245019	4.71	6	0
FN554766.1	O42	D	4969664	272313	5.19	7	0
NC_004431.1	CFT073	B2	5046473	184955	3.53	5	4
NC_007946.1	UTI89	B2	4800674	265067	5.23	7	10
NC_008253.1	536	B2	4885060	53860	1.09	2	8
NC_008563.1	APEC01	B2	4677179	404846	7.96	10	8
NC_011745.1	ED1a	B2	4775056	434492	8.34	11	0
NC_011742.1	S88	B2	4693074	339194	6.74	8	4
AP009378.1	SE15	B2	4663196	54142	1.15	2	0
CP001671.1	ABU83972	B2	4979233	152164	2.96	4	4
CP001855.1	NRG857C	B2	4608993	138826	2.92	3	4
CP001969.1	IHE3034	B2	4677196	431187	8.44	12	4
CP002167.1	UM146	B2	4803663	189350	3.79	6	14
CU651637.1	LF82	B2	4616442	156666	3.28	4	4
NC_011601.1	E2348-69	B2	4576391	389162	7.83	10	4
CP002797.1	NA114	B2	4669051	266190	5.39	8	46
NC_011750.1	IAI39	F	4809772	322296	6.28	11	22
NC_010498.1	SMS-35	F	4976575	91814	1.81	3	4
NC_011740.1	esfe	*E. fergusonii*	4365498	223213	4.86	6	101
NC_015761.1	NCTC-12419	*S. bongori*	4379885	80220	1.80	3	0

NC_003197.1	LT2	*S. enterica*	4654816	202616	4.17	6	0
NC_003198.1	CT18	*S. enterica*	4604245	204792	4.26	6	6
NC_004631.1	Ty2	*S. enterica*	4622693	169268	3.53	6	8
NC_006905.1	SC-B67	*S. enterica*	4550832	204868	4.31	6	0
NC_010102.1	SPB7	*S. enterica*	4710495	148392	3.05	5	2
NC_011080.1	SL254	*S. enterica*	4827641	238485	4.93	8	0
NC_011083.1	SL476	*S. enterica*	4728763	160005	3.27	6	0
NC_011094.1	CVM19633	*S. enterica*	4521656	187419	3.98	5	0
NC_011149.1	SL483	*S. enterica*	4692433	106227	2.21	4	2
NC_011147.1	AKU_12601	*S. enterica*	4477255	104542	2.28	3	2
NC_011205.1	CT_02021853	*S. enterica*	4635678	207230	4.28	5	2
NC_011274.1	287/91	*S. enterica*	4603765	54932	1.18	1	6
NC_011294.1	NC_011294	*S. enterica*	4603350	82498	1.76	3	2
NC_006511.1	ATCC9150	*S. enterica*	4485549	99680	2.17	3	2
NC_012125.1	RKS4594	*S. enterica*	4646604	186476	3.86	6	5
FN424405.1	D23580	*S. enterica*	4640480	238920	4.90	7	3
FQ312003.1	SL1344	*S. enterica*	4653893	224119	4.59	8	0
AP011957.1	T000240	*S. enterica*	4727844	226970	4.58	7	0
CP001363.1	14028S	*S. enterica*	4647761	222504	4.57	6	0
CP002487.1	4/74	*S. enterica*	4667560	210453	4.31	7	0

Table S2. Summary table of the classification of the detected prophages.

Order	Type	Family	Genus (well-known classified*)	Number of Prophages
Caudovirales	Lambdoids (330)	*Siphoviridae*	Lambda-like (30)	221
		mixed	Stx-like (7)	5
		Podoviridae	P22-like (11)	24
		Myoviridae	SfV-like (3)	22
		ND	Unknown	58
	Non-Lambdoids (167)	*Myoviridae*	P2-like (10)	70
			Mu-like (3)	5
			Unknown	29
		Podoviridae	Epsilon15-like (2)	6
		Siphoviridae	phiC31-like (2)	2
		ND	Unknown	38
		None	P4-like (4)	17
..	None	*Inoviridae*	Inovirus	3

Note: (*) well-known classified phages and prophages were obtained from Genbank and from literature data (Casjens 2003).

Table S3. Putative integration targets in *E.coli* and *S. enterica.*

E. coli				*S. enterica*			
Integrative Locus	Putative target gene			Integrative Locus	Putative target gene		
	tRNA-tmRNA	sRNA	*Protein coding gene*		tRNA-tmRNA	sRNA	*Protein coding gene*
135	Thr CGT	ECS209 – 074 (IGR)		199	Thr CGT		
216	Arg TCT	ECS074 (IGR)		329	Arg TCT		
357		*rybB* (IGR)		521		*rybB* (IGR)	
381		ECS172 (IGR)		590		ECS167 (IGR)	
420			*ssuA*	614	Ser TGA		
442	Ser TGA			715			*lcd**
458		ECS083 (IGR)	*wrbA*	924			*intR*
523		C0293 (IGR)	*lcd**	1060		*ryeA ryeB* (IGR)	
576			*ompW**	1149	Ser CGA		
612		ECS088 (IGR)		1183			*sbcB**
627			*yneJ*	1283	Pro GGG		
634		ECS151 (CA *intQ*,P)	*intQ*	1390	Arg CCT		
645			*tqsA*	1517		ECS060 (CA *lepA*, 5'-UTR)	*lepA*
784		*ryeA ryeB* (IGR)		1551	tmRNA		
802			*yecE*	1835	Met CAT		
822	Leu TAA			2445			*yjbN*
826	Ser CGA			2484			SPC_4453
859		*cyaR/ryeE* (IGR)		2578	Leu CAA		
868			*mlrA*	18	9	4	6
899	Pro GGG						
915			*yfaT*				
993	Arg CCT						
1029		ECS170-125 (CA *eutB*, G)	*eutB*				
1099		*ryfB* (IGR)	*yfhL*				
1131	tmRNA						
1845			Z5614				
1937		ECS021 (CA *ytfP*, G)					
1959	Leu CAA						
1966			*prfC*				
29	9	13	14				

Note: For tRNA genes, the amino acid and the anticodon are indicated. (IGR) indicates that the candidate sRNA is located in the intergenic region; *'CA = cis-antisense'* indicates that the candidate sRNA is located at the same genomic locus but on the opposite strand to known genes (G) or known pseudogenes (P). For more details see (Shinhara, et al. 2011). The star corresponds to the presence of an alternative 3' or 5' end in the target gene.

Table S4. Summary table of tRNA genes present in *E. coli.*

N AA	Anticodon	codon	RSCU-K12	Integration Locus	min trna detected	max trna detected	Conservation	Copy number/genome
5 Arg	CCT	AGG	0	Yes	1	1	core	Not-variable
5 Arg	TCG	CGA	0	No	0	5	Not-core	variable
5 Leu	TAG	CTA	0.0129032	No	0	1	Not-core	variable
3 Gly	TCC	GGA	0.0284192	No	1	5	core	variable
5 Arg	CCG	CGG	0.0332103	No	1	1	core	Not-variable
5 Arg	TCT	AGA	0.0332103	Yes	1	9	core	variable
3 Pro	GGG	CCC	0.0487805	Yes	1	1	core	Not-variable
3 Gly	CCC	GGG	0.0568384	No	1	1	core	Not-varibale
5 Ser	CGA	TCG	0.0627178	Yes	1	1	core	Not-varibale
5 Leu	TAA	TTA	0.141935	Yes	1	1	core	Not-varibale
1 SeC(p)	TCA	TGA	0.15	No	1	1	core	Not-varibale
3 Thr	CGT	ACG	0.156863	Yes	1	1	core	Not-variable
3 Thr	TGT	ACA	0.179272	No	0	5	Not-core	variable
5 Ser	TGA	TCA	0.188153	Yes	1	1	core	Not-variable
5 Leu	GAG	CTC	0.232258	No	1	1	core	Not-variable
5 Leu	CAA	TTG	0.245161	Yes	1	1	core	Not-variable
2 Val	GAC	GTC	0.290323	No	2	2	core	Not-variable
2 Ala	GGC	GCC	0.316779	No	1	2	core	variable
2 Gln	TTG	CAA	0.373333	No	0	2	Not-core	variable
3 Pro	TGG	CCA	0.520325	No	1	2	core	variable
1 Met	CAT	ATG	1	Yes	5	18	core	variable
1 Trp	CCA	TGG	1	No	1	1	core	Not-varaible
2 Ala	TGC	GCA	1.08456	No	1	6	core	variable
2 Val	TAC	GTA	1.13548	No	3	6	core	variable
1 Cys	GCA	TGC	1.22581	No	1	1	core	Not-variable
1 Asp	GTC	GAC	1.3141	No	3	3	core	Not-variable
5 Ser	GCT	AGC	1.37979	No	1	1	core	Not-variable
3 Gly	GCC	GGC	1.39254	No	3	4	core	variable
1 His	GTG	CAC	1.40909	No	1	1	core	Not-variable
1 Lys	TTT	AAA	1.44983	No	2	7	core	variable
1 Tyr	GTA	TAC	1.50365	No	1	4	core	variable
1 Glu	TTC	GAA	1.51293	No	1	6	core	variable
5 Ser	GGA	TCC	1.52613	No	1	2	core	variable
1 Phe	GAA	TTC	1.5468	No	1	2	core	variable
2 Gln	CTG	CAG	1.62667	No	1	2	core	variable
1 Asn	GTT	AAC	1.74892	No	3	5	core	variable
3 Thr	GGT	ACC	1.7591	No	1	2	core	variable
1 Ile	GAT	ATC	2.27456	No	1	6	core	variable
5 Ser	AGA	TCT	2.55052	No	1	1	core	Not-variable
3 Pro	CGG	CCG	2.84553	No	1	1	core	Not-variable
5 Arg	ACG	CGT	4.14022	No	3	5	core	variable
5 Leu	CAG	CTG	5.07097	No	2	4	core	variable

Note : (core) present in at least one copy in each strain of *E. coli.*

Table S5. Summary table of the integrative loci, prophage genera and their putative targets of integration.

E. coli						*S. enterica*				
IL	Genus	N	NR	Target	C	IL	Genus	N	NR	Target
128	Mu-like	2	1	x						
135	Lambda	1	1	$tRNA_{Thr}$	1	199	P4-like	1	1	$tRNA_{Thr}$
	P22-like	1	1	$tRNA_{Thr}$			P22-like	6	5	$tRNA_{Thr}$
	SfV-like	1	1	$tRNA_{Thr}$						
	Unknown	5	5	$tRNA_{Thr}$						
	Unknown -- P4-like	5	2	ECS074						
	SfV-like -- Unknown	1	1	$tRNA_{Thr}$						
216	Lambda-like	9	9	$tRNA_{Arg}$	2	329	P22-like	1	1	$tRNA_{Arg}$
	Unknown	9	4	$tRNA_{Arg}$			Lambda-like -- Unknown	1	1	$tRNA_{Arg}$
303	P22-like	1	1	x						
	SfV-like	2	2	x						
	Unknown	2	2	x						
	Lambda-like	15	13	x						
334	Unknown	1	1	x						
357	P2-like	12	9	*rybB*	3	521	P2-like	1	1	*rybB*
							Unknown	1	1	*rybB*
381	Unknown	2	1	ECS172		536	Lambda-like	1	1	x
391	P2-like	4	4	x		590	Unknown	3	3	ECS167
395	P2-like	3	3	x			Lambda-like	9	5	ECS167
399	Unknown	1	1	x						
420	Mu-like	1	1	*ssuA*						
442	Unknown	1	1	$tRNA_{Ser}$	4	614	Unknown	2	2	$tRNA_{Ser}$
	Lambda-like	7	7	$tRNA_{Ser}$						
451	Lambda-like	1	1	x						
	P22-like	2	2	x						
458	Unknown	1	1	ECS083 *wrbA*						
	Stx-like	3	3	ECS083 *wrbA*						
464	Lambda-like	1	1	x						
514	Unknown	1	1	x						
	Lambda-like	12	12	x						
517	phiC31-like	2	2	x						
	Unknown	7	6	x						
523	Unknown	5	3	C0293 *icd*	5	715	Lambda-like	2	2	x
	Lambda-like	16	15	x			SfV-like-- Unknown	1	1	x
	Lambda-like * 2	1	1	x						
539	Unknown	1	1	x						
571	Unknown	1	1	x						
576	Lambda-like	15	14	*ompW*						
	Lambda-like * 2	2	2	*ompW*						
608	Unknown	1	1	x						
612	Lambda-like	21	12	ECS088	6	924	Lambda-like	1	1	x
626	Inovirus	3	3	x						
627	P2-like	1	1	*ypneJ*						
634	Lambda-like	28	17	x						
	Lambda-like * 2	4	4	ECS151						
645	Unknown	3	3	*tqsA*						
652	Unknown	1	1	x	7	852	Unknown	2	1	x
709	P2-like	6	6	x						
784	Lambda-like	1	1	ryeA-B	8	1060	Lambda-like	3	3	*ryeA-B*
							Unknown	3	3	*ryeA-B*
802	SfV-like	1	1	*yecE*						
	Lambda-like	4	4	*yecE*						

822	Unknown	1	1	$tRNA_{Leu}$						
	P2-like	7	6	$tRNA_{Leu}$						
826	Unknown	3	2	$tRNA_{Ser}$	9	1149	SfV-like	4	2	$tRNA_{Ser}$
	Lambda-like * 3	1	1	$tRNA_{Ser}$						
	Lambda-like	12	12	$tRNA_{Ser}$						
833	Lambda-like	2	2	x	10	1183	SfV-like	1	1	*sbcB*
	P22-like	2	2	x			Unknown	1	1	*sbcB*
859	P2-like	8	6	*cyaR/ryeE*						
864	Unknown	2	2	x						
868	Lambda-like	5	5	*mlrA*						
	Lambda-like * 2	1	1	x						
899	Unknown	2	2	$tRNA_{Pro}$	11	1283	Unknown	8	5	$tRNA_{Pro}$
915	Lambda-like	2	2	*yfaT*						
964	Mu-like	1	1	x						
978	P2-like	2	2	x						
993	Stx-like	2	2	$tRNA_{Arg}$	12	1390	P22-like	1	1	x
	Unknown	5	5	$tRNA_{Arg}$			Unknown	1	1	$tRNA_{Arg}$
	P22-like	10	7	$tRNA_{Arg}$						
1029	Unknown	2	2	ECS170-125 *eutB*		1517	Unknown	1	1	ECS060
1064	Epsilon15-like	6	6	x			Lambda-like	7	5	ECS060
1088	Unknown	1	1	x						
1099	Unknown	4	4	*ryfB yfhL*						
1120	P2-like	1	1	x	13	1538	Unknown	2	2	x
1125	P2-like	1	1	x	14	1542	P2-like	1	1	x
1131	P2-like	2	2	tmRNA	15	1551	P4-like	2	2	tmRNA
	P2-like -- SfV-like	1	1	tmRNA			Unknown – P4-like	1	1	tmRNA
	P4-like	2	2	tmRNA			P2-like	4	3	tmRNA
	Lambda-like	4	4	tmRNA			P2-like -- Unknown	2	1	tmRNA
	SfV-like	5	3	tmRNA			P2-like -- P4 like	2	2	tmRNA
	Unknown	6	6	tmRNA						
1653	Unknown	3	3	x		1835	Unknown	1	1	$tRNA_{Met}$
							P2-like	2	2	$tRNA_{Met}$
1781	P2-like	4	2	x	16	2342	P2-like	1	1	x
1845	Mu-like	1	1	Z5614						
1847	Unknown	2	1	x	17	2424	Unknown	11	1	x
1853	Lambda-like	1	1	x						
1861	Lambda-like	1	1	x	18	2445	SfV-like	1	1	*yjbN*
	SfV-like	1	1	x						
	Unknown	2	2	x						
1937	P2-like	1	1	ECS021		2484	P2-like	2	1	SPC_4453
1959	P4-like	1	1	$tRNA_{Leu}$	19	2578	P4-like	4	3	$tRNA_{Leu}$
1966	Lambda-like	9	8	*prfC*						

Note : (IL) = integrative locus; (N) = number of prophages; (NR) = number of non-redundant prophages; (C) = IL conserved in both species; (Target) = putative target detected in 1kb out of both sides of prophages. Grey boxes indicate integration loci conserved between *Salmonella* and *Escherichia*.

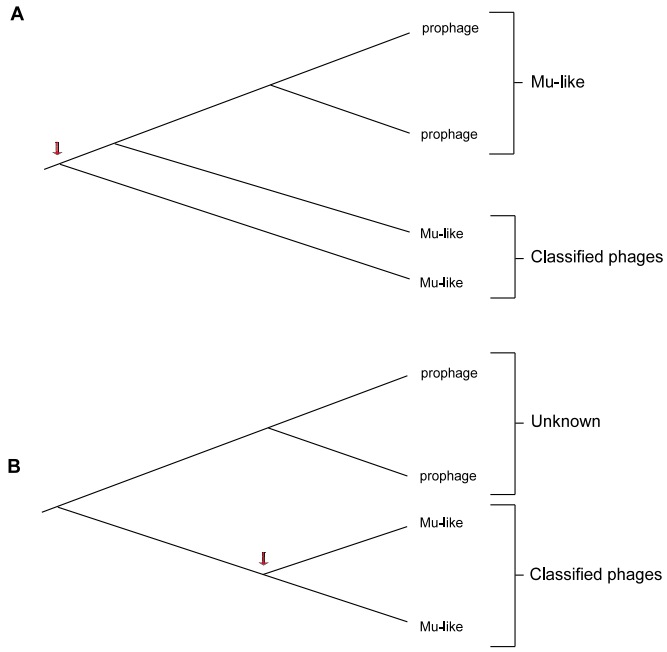

Fig. S1. Illustration of the classification method of the detected prophages. Within the classification cladogram of (pro)phages based on gene repertoire relatedness (see Methods), prophages were attached to an order, a type, a family and a genus when local topologies relative to classified (pro)phages were informative. **(A)** Schematic representation of a topology where prophages are considered as Mu-like prophages. Two Mu-like (pro)phages are found branching basally to the detected prophages. The red arrow indicates the predicted Mu-like ancestor. In this case, the ancestor of the prophages is considered as a Mu-like phage. The two detected prophages are then classified as Mu-like prophages **(B)** Illustration of a topology where the two prophages are not classified as Mu-like viruses. Mu-like (pro)phages expected ancestor is shown by the red arrow. In this topology, it is not known if the Mu-like viruses share a Mu-like ancestor with the prophages. The prophages are then classified as "unknown".

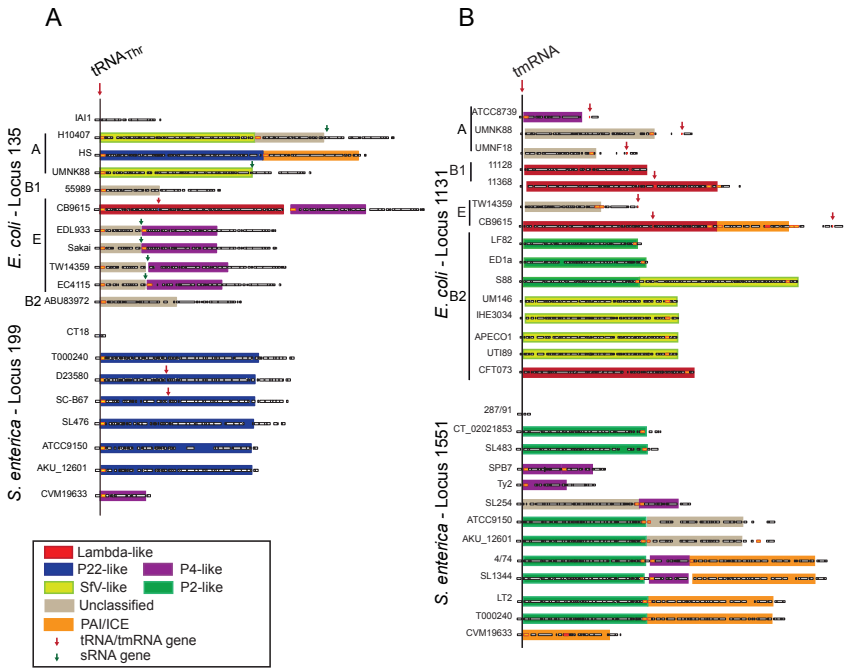

Fig. S2. Two integration loci with exceptionally diverse integrative elements. **(A)** Integrative loci 135 and 199 are conserved in both species. **(B)** Integrative loci 1131 and 1551 are conserved also between the 2 species. Each line represents gene organization for each strain containing at least one detected prophage in these loci. Other strains are devoid of prophages at these loci (e.g. IAI1 and CT18 strains show the basic gene organization of these loci without prophage). Red and green arrows correspond to tRNA/tmRNA genes and sRNA genes respectively. Yellow rectangles with red borders represent integrases.

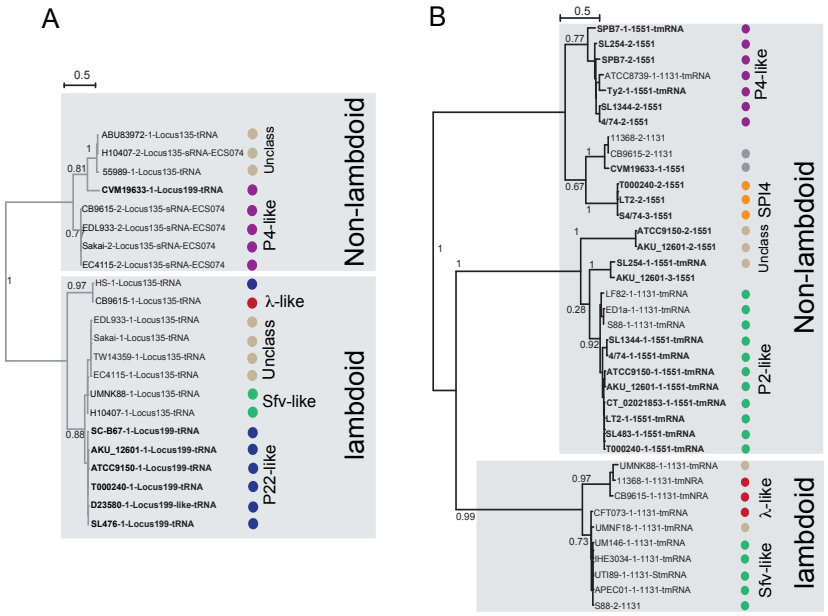

Fig. S3. Molecular phylogeny of the integrase proteins located in the two conserved loci 135-199 and 1131-1551 in *E. coli-S. enterica*. Phylogenetic trees for the integrase proteins were performed using PhyML with the WAG+G model. Values correspond to aLRT values. In bold = integrase of *S. enterica* strains.

References:

Casjens S 2003. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol 49: 277-300.
Shinhara A, Matsui M, Hiraoka K, Nomura W, Hirano R, Nakahigashi K, Tomita M, Mori H, Kanai A 2011. Deep sequencing reveals as-yet-undiscovered small RNAs in Escherichia coli. BMC Genomics 12: 428. doi: 1471-2164-12-428 [pii]
10.1186/1471-2164-12-428

Matériel supplémentaire – Article 2

Manipulating or Superseding Host Recombination Functions: a Dilemma that Shapes Phage Evolvability

Louis-Marie Bobay, Marie Touchon, Eduardo P.C. Rocha

Supporting information

Table of contents

Text S1. Identification and classification of prophages.

Identification of prophages. Prophages were detected as in [1]. i) The initial detection process used three prophage-detection programs: Phage Finder [2], PHAST [3] and Prophinder [4]. These programs combine sequence comparisons to known phage or prophage genes, comparisons to known bacterial genes, identification of tRNA genes, dinucleotide frequency analysis and identification of attachment sites. ii) We then removed putative prophages with a large number of Insertion Sequences (>25% of the predicted genes). IS elements were detected as in [5]. iii) In the third step, prophage borders and the few tandem were manually curated using gene annotation, hits to PFAM and the definition of core/pan bacterial genomes. iv) From the resulting 500 prophages we removed the shorter than 30 kb to avoid partially degraded prophages. The threshold of 30kb has been chosen as it represents the minimum size range of a functional temperate - and non-satellite - phage genome infecting enterobacteria within our *Caudovirales* phage dataset (*Salmonella* phage PsP3, [6]). v) Finally, pairs of prophages with a repertoire relatedness score over 0.9 (see [1]), were considered as redundant and only the longer prophage was kept for further analysis. A total of 301 prophages were thus obtained and correspond to the non-redundant long dataset (NRlong) of [1].

Phage classification. Using taxonomic information from the ICTV and the literature we were able to classify most of the detected prophages (i.e. to attribute them an order, a family, a genus and the membership to the lambdoid group) [1]. The majority of temperate phages were attributed to the lambdoid group which contains the following genera/groups: Lambda-like, P22-like, SfV-like, Stx-like, N15-like and unclassified lambdoids. The non-lambdoid temperate phages were classified in P2-like, Mu-like, P1-like, Epsilon15-like and unclassified non-lambdoids genera/groups. The rest of the analysis focuses on these 237 lambdoid prophages thus classified and the 38 lambdoid phages downloaded from RefSeq.

Text S2. Function assignment.

We used HHsearch to search for genes in phages matching PFAM-A protein profiles (parameters: p>95% in local and global alignments and >50% of profile coverage). Phage functions were assigned to the protein clusters using this information. Functions that are specific of phage were automatically attributed to the protein clusters from PFAM-A profile descriptions. The following functional classes were defined: "integrase", "excisionase", "resolvase", "transcription regulator", "replication protein", "lysis", "terminase", "portal protein", "packaging protein", "head protein", "head tail connector", "tail protein", "tail fiber protein", "methyl tranferase", "virulence protein" and "transposase". Around 63% of protein clusters didn't display a significant match with any of the functions listed before and were classified as "unknown".

Text S3. Detection of Chi sites in non-lambdoid phages

We computed the Chi sites O/E ratio for the genomes of 131 non-lambdoid phages and prophages, including temperate and virulent phage genera infecting enterobacteria. The details of the results are given table S5. We estimated the over- or under-representation of Chi sites of these genomes individually by using the Z score statistics with the tri-nucleotides (M2) model. Two genera significantly and consistently over-represent Chi sites: the non-integrative temperate P1-like phages and the virulent T5-like phages. To the best of our knowledge and according to our detection of recombinases and RecBCD inhibitors, these phages don't encode recombinases. Moreover, these elements only package their genomes from concatemeric DNA. It is therefore likely that they manipulate the host recombination functions through the presence of Chi sites in their genome.

The other phages show less clear patterns of over or under-represent Chi motifs. The small sample sizes preclude taking very solid conclusions from these trends. For example some groups have no single Chi motif, but even this extreme under-representation is not statistically significant because of the small sample size. This is the case of P2-like, P4-like and Mu-like phages. Interestingly, all these phages are able to package monomeric DNA and therefore might be under weaker selection for using the host recombination functions. T7-like and Epsilon15-like phages seem to slightly over-represent Chi sites and both genera were found to encode a recombinase (Gp2.5 and Redβ respectively). We could not find a RecBCD inhibitor in Epsilon15-like phages. A minority (32%) of T7-like phages encodes an homolog of the RecBCD inhibitor protein Gp5.9, but the sample size precludes a comparison between Inh^+ and Inh^- phages in this case. It is therefore likely that T7-like and Epsilon15-like phages use Chi sites as a protection from the host RecBCD exonuclease like Rec^+Inh^- lambdoids. But for these and other clades, the identification of clear patterns will require a much larger sample size and a better knowledge of their RecBCD inhibitors.

.

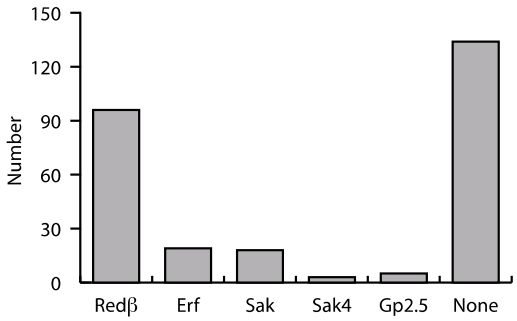

Figure S1. Recombinase families identified in lambdoid phages. Recombinases were identified by profile-profile comparisons with HHsearch (see Materials and Methods). Most of the identified recombinases belong to the Rad52 superfamily (Redβ, Erf and Sak). Sak4 recombinases are part of the Rad51 superfamily and are remote homologs of RecA [7]. Gp2.5 represents the last superfamily of phage recombinases and is found much more frequently in virulent phages [7].

Table S1. Chi sites Observed/Expected ratio for lambdoid phages and their bacterial hosts computed with models M0, M4 and M6. The expected number of Chi sites has been determined with three additional models: M6, M4 and M0. For each category, we tested if the ratio O/E of Chi composition in the set of phages was significantly different from random expectation (O/E=1) with the Mann-Whitney test.

	Maximal (M6)		Penta-nucleotides (M4)		No memory (M0)	
Category	**Median Obs/Exp**	***pvalue* Mann-Whitney**	**Median Obs/Exp**	***pvalue* Mann-Whitney**	**Median Obs/Exp**	***pvalue* Mann-Whitney**
All lambdoids	1.09	0.869	1.09	0.043	4.66	<0.00001
Rec$^-$	1.29	<0.00001	1.54	<0.00001	6.69	<0.00001
Rec$^+$ Inh$^+$	0.0	<0.00001	0.0	<0.00001	0.0	<0.00001
Rec$^+$ Inh$^-$	1.12	0.146	1.49	0.001	5.0	<0.00001

Table S2. Chi sites Z score statistics for lambdoid phages and their bacterial hosts. The expected number of Chi sites has been determined with three M2 model. For each category, we tested if the Z score of Chi composition in the set of phages was significantly different from random expectation (Z=0) with the Mann-Whitney test. The "Skew" column indicates if the phage category over-represents (+) or under-represents (–) Chi sites.

	Tri-nucleotides (M2)		
Category	**Skew**	**Median Z scores**	***pvalue* Mann-Whitney**
All lambdoids	+	1.83	<0.00001
Rec$^-$	+	3.00	<0.00001
Rec$^+$ Inh$^+$	–	–1.03	<0.00001
Rec$^+$ Inh$^-$	+	2.20	<0.00001

Table S3. Chi sites Observed/Expected ratio for *E. coli* and *S. enterica* with models M0, M4 and M6. The expected number of Chi sites has been determined with three additional models: M6, M4 and M0. For each core or complete genome, we tested if the Chi composition was significantly different from random expectation with the Z score. The analysis was run on *E. coli* K12 MG1655 and *S. enterica* Typhimurim LT2 genomes respectively.

	Maximal (M6)		**Penta-nucleotides (M4)**		**No memory (M0)**	
Category	**Median Obs /Exp**	***pvalue* Z score**	**Median Obs /Exp**	***pvalue* Z score**	**Median Obs /Exp**	***pvalue* Z score**
Core genome *E. coli*	1.142	<0.01	1.72	<0.00001	7.01	<0.00001
Core genome *S. enterica*	1.09	<0.05	1.48	<0.00001	5.21	<0.00001
Complete genome *E. coli*	1.138	<0.00001	1.69	<0.00001	6.84	<0.00001
Complete genome S. enterica	1.07	<0.00001	1.43	<0.00001	5.01	<0.00001

Table S4. **Comparison of the Chi sites Observed/Expected ratio of *E. coli* lambdoid phages and all lambdoid phages to the Chi sites Observed/Expected ratio of *E. coli* core genome with models M0, M4 and M6**. The median value "M" of the Chi sites Oberved/Expected ratio is given for lambdoid coliphages and for all lambdoid phages for each category. For each category and model, we tested if the Chi composition was significantly different from the Chi composition of *E. coli*'s core genome with the Mann-Whitney test. The analysis has been done on the core genes of *E. coli* K12 MG1655.

	Maximal (M6)				**Pentanucleotides (M4)**				**No memory (M0)**			
	Coliphages		**All phages**		**Coliphages**		**All phages**		**Coliphages**		**All phages**	
Category	**M**	**pvalue**	**M**	**pvalue**	**M**	**pvalue**	**M**	**pvalue**	**M**	**pvalue**	**M**	**pvalue**
Rec$^-$	1.28	<0.00001	1.29	<0.00001	1.44	<0.00001	1.54	0.076	6.09	0.114	6.69	0.964
Rec$^+$ Inh$^+$	0	<0.00001	0.0	<0.00001	0	<0.00001	0.0	<0.00001	0	<0.00001	0.0	<0.00001
Rec$^+$ Inh$^-$	1.40	0.33	1.12	0.624	0.80	0.055	1.49	0.258	2.78	0.021	5.0	0.018

Table S5: Chi sites Observed/Expected ratio and Z scores for different genera of phages and prophages infecting enterobacteria. We used the non-lambdoid phage genera of the *Caudovirales* order defined by the ICTV. Prophages were identified and classified as in [1]. Phage's life style, i.e. virulent (v) and temperate (t) is indicated in the "Type" column. The type of DNA substrate used for packaging, i.e. concatemeric (C) of monomeric (M) is indicated in the "Packaging" column.

Family	Subfamily	Genus	Type	Genomes	Rec	Packaging	Median O/E	Median Z scores
Myoviridae	*Tevenvirinae*	T4-like	V	22	UvsX	C	1.26	0.32
	–	Felix01-like	V	4	–	NA	0.47	-0.72
	–	Vi1-like	V	1	UvsX	NA	1.16	0.36
	–	Mu-like	T	5	–	M	0.00	-1.70
	–	P1-like	T	2	–	C	10.81	8.20
	Peduovirinae	P2-like	T	46	–	M	0.00	-1.39
Siphoviridae	–	T1-like	V	5	ERF	C	0.55	-0.11
	–	T5-like	V	3	–	C	2.60	2.21
Podoviridae	*Autographivirinae*	SP6-like	V	5	1/5 (UvsX)	C	1.18	0.22
		T7-like	V	25	Gp2.5	C	2.49	1.96
	–	N4-like	V	2	Sak4	M	1.80	1.23
	–	Phieco32-like	V	2	–	NA	0.72	-0.70
	–	Epsilon15-like	T	8	Redβ	NA	2.69	2.19
–	–	P4-like	T	1	–	M	0.00	0.19

References

1. Bobay LM, Rocha EP, Touchon M (2013) The Adaptation of Temperate Bacteriophages to Their Host Genomes. Mol Biol Evol 30: 737-751.
2. Fouts DE (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res 34: 5839-5851.
3. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS (2011) PHAST: a fast phage search tool. Nucleic Acids Res 39: W347-352.
4. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics 24: 863-865.
5. Touchon M, Rocha EP (2007) Causes of insertion sequences abundance in prokaryotic genomes. Mol Biol Evol 24: 969-981.
6. Bullas LR, Mostaghimi AR, Arensdorf JJ, Rajadas PT, Zuccarelli AJ (1991) Salmonella phage PSP3, another member of the P2-like phage group. Virology 185: 918-921.
7. Lopes A, Amarir-Bouhram J, Faure G, Petit MA, Guerois R (2010) Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. Nucleic Acids Res 38: 3952-3962.

Matériel supplémentaire – Article 3

Supporting information

SI materials and methods

Computation of core genomes and distance matrices. The core genome is the set of all ubiquitous orthologous protein families identified within a species. We built the core genomes of *E. coli* and of *S. enterica*. For each pair of genomes, orthologous proteins were identified as unique reciprocal best hits with more than 60% similarity in amino acid sequence and less than 20% of difference in protein length. *E. coli* K12 MG1655 and *S. enterica* LT2 were used as references for the order of the core genome in each species. Core proteins were aligned with MUSCLE v3.6 (1) (default parameters), and multiple alignments merged. We computed the protein distance (D) for pairs of prophages using concatenated alignments of their core protein sets with TREE-PUZZLE v5.2 (2). We computed matrix distances using maximum likelihood under automatic estimation of the best substitution model and a $\Gamma(8)$ correction for rate heterogeneity.

Definition of orthologous prophages. We followed a four-step approach to determine orthologous prophages.

1) Orthologous prophages descend from a single integration event and must therefore be integrated at the same locus (they are flanked by the same homologous core genes of the host). Prophages integrated at a core genome breakpoint (i.e. a genome rearrangement) were removed from the analysis since the location condition cannot be ascertained (4% of all prophages).

2) Orthologous prophages are expected to display a high gene repertoire relatedness R (defined above). Different thresholds of R (T_R) were tested (0.6, 0.65, 0.7, 0.75, 0.8, 0.85 and 0.9). For each threshold, families of orthologous prophages were built by transitivity, i.e. a pair of prophages belongs to the same family if $R \geq T_R$. The size and the connectivity of the resulting prophage families were visualized using BioLayout (3). Most thresholds gave similar numbers of groups. The intermediate value T_R=0.7 was chosen because lower thresholds gave larger families but with a lower connectivity of prophages and higher thresholds were too stringent.

3) Vertically inherited prophages should display evolutionary rates similar to their host-specific genes at synonymous sites because they endure similar mutation rates and because selective pressures at such sites are weak even in the majority of core non-highly expressed

genes. For each family of orthologous prophages, homologous proteins were defined as unique reciprocal best hits with >40% similarity in amino acid sequence and <50% difference in protein length and aligned with MUSCLE v3.6 (1) (default parameters). IS elements were detected as in (4) and were removed from all the analyses in the study. Aligned proteins were back-translated into the corresponding codon-aligned nucleotide sequences and the synonymous substitution rates (dS) were computed with the yn00 method implemented in PAML (5). Synonymous substitution rates of *E. coli* and *S. enterica* core genes were computed using the same method. For each pair of prophages sharing *n* homologous proteins, we made 1000 experiments where we randomly picked *n* pairs of homologous proteins from the hosts' core genome. A pair of prophages was considered as orthologous when the average dS value of their homologous genes was found within the 99% confidence interval of the distribution of the 1000 random experiments. The pairs of orthologous prophages passing the test were classed by transitivity into 99 families.

4) To remove potential false positives from the set of vertically inherited prophages we eliminated or subdivided families of orthologous prophages based on their gene diversity. Orthologous prophages descend from a single integration event and should therefore display a gene diversity that doesn't greatly exceed the gene content of the ancestral prophage. For example, the largest lambdoid temperate phage in GenBank encodes 83 genes. Hence, a family of lambdoid prophages with inferred ancestral gene content much larger than 83 is unlikely to correspond to a group of orthologous prophages. Also, orthologous prophages have endured genetic degradation for a similar period of time. Therefore they are expected to show roughly similar levels of genetic degradation. Thus, we built the protein families of each family of orthologous prophages (the prophage family pan-genome). We then compared the number of protein families of the prophage family to the number of protein coding genes found in the largest prophage of the prophage family. We then plotted the pan-genome size versus the size of the largest prophage (Fig. S3) and did an analysis of outliers. When compared with the average trend, only five families displayed an atypically large pan-genome ($P<0.05$, Dixon's Q test). For each of these five families, homologous proteins present in more than 75% of the prophages were aligned with MUSCLE v3.6 (default parameters) and concatenated. A maximum likelihood tree was built using PhyML3 with default parameters (6). The families were manually subdivided or removed from the analysis based on their corresponding phylogenetic tree topology. At the end of the procedure, we identified 100 families of orthologous prophages, including 372 distinct prophages.

In summary, prophages are orthologous if they are integrated at the same locus, have high gene repertoire relatedness, display synonymous substitution rates compatible with their hosts, and display gene diversity compatible with a single integration event. Albeit stringent, this procedure can't completely rule out the possibility that some of these prophages have resulted from independent integrations. A second subset of orthologous prophages was therefore defined with even more stringent parameters. Under this definition, we restricted the dataset to the orthologs evolving very closely to the rate of their corresponding host core genes at synonymous positions. Thus, we selected pairs of orthologous prophages displaying average synonymous substitution rates comprised between the average dS of the core genome of the host (dS_H) and lower that two times dS_H ($dS_H \leq dS_P \leq 2.dS_H$). This procedure identified 187 pairs of orthologous prophages (41 families) that showed similar low values of dN/dS.

The number of deletions was computed for each pair of orthologous prophages. For each pair of orthologous prophages A and B, each gene of prophage A was compared to the complete nucleotide sequence of prophage B (blastn, *e* value<0.001). Conversely, all the genes of prophage B were compared to the complete nucleotide sequence of prophage A. All genes lacking significant hits were considered as indels. Consecutive genes with no hits were merged into larger indels. This is an underestimate of the deletion sizes, since elements lost in both prophages cannot be accounted for.

Functional assignment. We assigned functions to prophage proteins by sequence comparison to PFAM-A profiles (downloaded the 07/29/2013) with HMMER v3.0 (7) (*e* value<0.001). Phage specific functions were classed in broad functional categories: "integration", "regulation", "recombination", "replication", "lysis", "packaging", "head", "tail" and "accessory". Accessory proteins represent various phage functions such as: toxin/virulence, restriction/modification, surinfection avoidance, *nin* genes and antibiotic resistance. PFAM profiles with no clear relationship to phage functions were not considered and the corresponding proteins were classified as "unknown or other". Proteins matching two or more different functional categories were also categorized as "unknown or other". Homologous proteins were defined by unique reciprocal best hits with more than 40% similarity in amino acid sequence and less than 50% difference in protein length. They were grouped into protein families by transitivity. We attributed a function to the protein family when more than 50% of the proteins were assigned to the same functional category.

Computation of synonymous and non-synonymous substitution rates. Protein families of orthologous prophages were aligned as described above and back-translated into the

corresponding nucleotide sequences. When bacterial strains are very closely related the corresponding orthologous prophages endured very few nucleotide substitutions. This introduces imprecision on the estimation of the ratios of non-synonymous (dN) versus synonymous (dS) substitution rates. Therefore, for each pair of orthologous prophages, all the alignments of orthologous genes were concatenated to compute the synonymous and non-synonymous substitution rates of the set of families. This was done using yn00 as implemented in PAML (5). The same method was also used to compute synonymous and non-synonymous substitution rates of individual homologous prophage genes and of the host core genes.

Simulation of deletions. In order to assess how neutral deletions are expected to affect the patterns of gene loss along the genome, we generated one deletion among 1000 prophages following a Gaussian distribution for different average deletion sizes *s* (SD±10%) ranging from 1kb to 30kb. In a separate analysis, we extracted from (8) the distribution of deletion sizes for *Salmonella* and used it to simulate deletions in the same way. The position of each deletion was then randomly placed on a 45kb-long prophage flanked by two core genes of the host. We stipulated that any deletion of the host core genes is forbidden because it should be strongly counter-selected. We then measured the probability of deletion of each region in the phage genome.

Estimation of recombination. We used three different approaches to identify the presence of recombination in orthologous prophages:

i) We estimated recombination for each family of orthologous genes independently. We used four different methods implemented in the PhiPack program (9): the Neighbor Similarity Score (NSS) (10), the Maximum χ^2 (MaxChi) (11) and the Pairwise Homoplasy Index (PHI) (9) (with 1000 permutations and without permutations). A gene family was considered recombining when $P<0.05$ (permutation test). Only gene families composed of a minimum of three genes were considered in this analysis.

ii) To account for larger recombination tracts we estimated recombination among concatenates of orthologous prophage genes using the MaxChi method implemented in RDP3 (12) with default parameters. The analysis was performed on the genes conserved in all prophages of each family of orthologous prophages containing at least three prophages. We identified the positions of recombining sequences within the multiple alignment of each concatenate. The corresponding genes of the concatenate were considered as recombining when overlapping the predicted recombination tracts.

iii) In a third analysis we estimated recombination by a phylogenetic approach. For each family of orthologous prophages (≥ 4 prophages) we built a maximum likelihood reference tree on the ubiquitous genes of the family with RaxML (13) under a GTR + Γ(4) model. Individual gene trees were built with the same parameters with 100 bootstrap replicates. Gene trees were then individually compared to the reference tree with Prunier (14) with default parameters. Since the prediction of incongruent genes can differ depending on the root of the reference tree, we estimated the number of recombining genes by rooting reference trees in a way that either minimize (Min) or maximize (Max) the predicted number of recombining genes.

Finally, we combined the recombination prediction of the three approaches. Since the first approach doesn't identify which sequences are the recombinants within each alignment, we only considered the results of Prunier (Min) and/or MaxChi on concatenates if the alignment was positive for one of these two last methods. Gene families displaying significant signal of recombination with the fist approach (NSS, Maxchi or PHI) but not with the two last approaches (Prunier or MaxChi on concatenates) were considered recombining sequences as a whole. Only the orthologous gene families that could be tested by all three approaches have been considered for this last analysis. Overall, we detected recombining sequences by seven parallel methods and a combination of these methods.

SI references

1. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797.
2. Schmidt HA, Strimmer K, Vingron M, & von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502-504.
3. Enright AJ & Ouzounis CA (2001) BioLayout - an automatic graph layout algorithm for similarity visualization. *Bioinformatics* 17(9):853-854.
4. Touchon M & Rocha EP (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* 24(4):969-981.
5. Yang ZH & Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17(1):32-43.
6. Guindon S & Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696-704.
7. Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* 7(10):e1002195.
8. Sun S, Ke R, Hughes D, Nilsson M, & Andersson DI (2012) Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. *PLoS One* 7(8):e42639.
9. Bruen TC, Philippe H, & Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172(4):2665-2681.
10. Jakobsen IB & Easteal S (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput Appl Biosci* 12(4):291-295.
11. Smith JM (1992) Analyzing the Mosaic Structure of Genes. *J Mol Evol* 34(2):126-129.
12. Martin DP, *et al.* (2010) RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26(19):2462-2463.
13. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688-2690.
14. Abby SS, Tannier E, Gouy M, & Daubin V (2010) Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics* 11:324.
15. Sullivan MJ, Petty NK, & Beatson SA (2011) Easyfig: a genome comparison visualizer. *Bioinformatics* 27(7):1009-1010.

SI figures

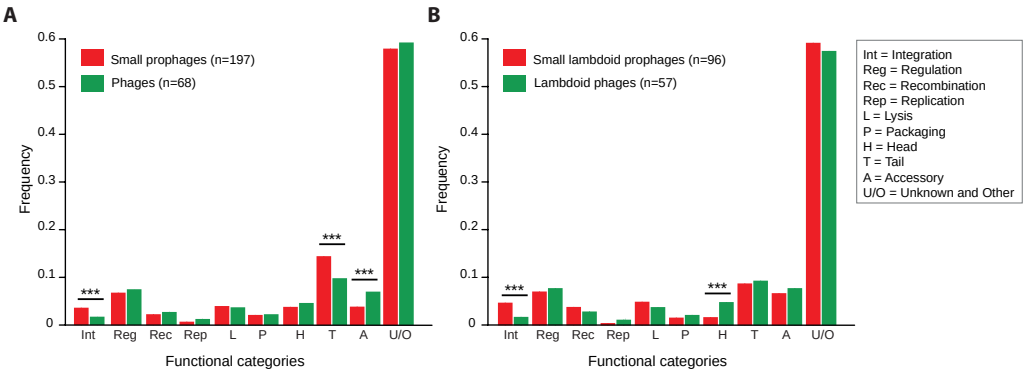

Figure S1. Distribution of protein functions encoded by small prophages (<30kb, red) and temperate caudophages of enterobacteria from GenBank (green). (**A**) all phages and small prophages. (**B**) lambdoid phages and small lambdoid prophages. Functional categories are indicated in the right insert. n: number of prophages. ***: $P<0.001$, χ^2 test.

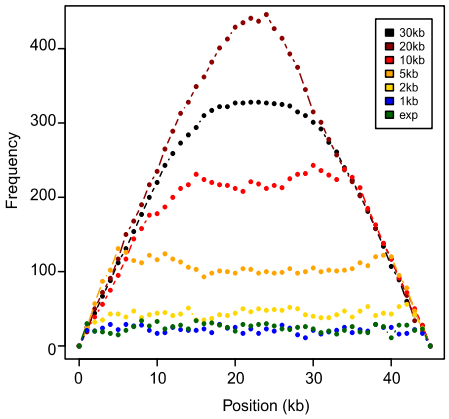

Figure S2. Distribution of the probability of deletion of genes along the prophage genome if deletions were neutral. The frequency of deleted positions is represented on a hypothetical 45kb-long prophage flanked by two core genes of the host (see SI Method). We considered that any deletion affecting a surrounding host core gene would be counter-selected. We also ran similar simulations using experimental data on the size of deletions (8) (labeled "exp").

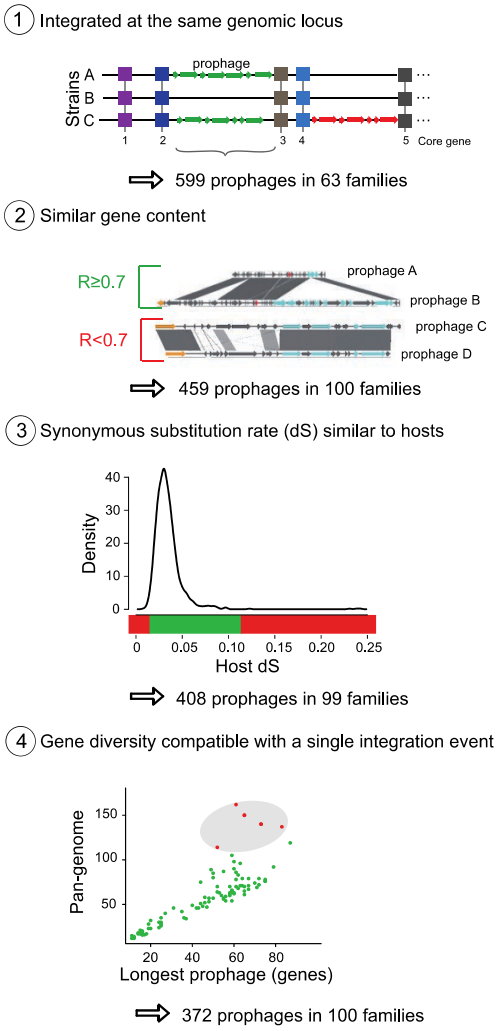

Figure S3: Four-step procedure used to detect orthologous prophages. 1) Orthologous prophages must be integrated between the same core genes of their hosts. 2) Orthologous prophages must be highly similar at the genomic level. The gene repertoire relatedness score R was developed to assess the similarity of prophage genomes and is robust to deletions (see Materials and Methods). Prophage pairs were not considered as orthologous when R<0.7. 3) Orthologous prophages must display a substitution rate in the range of the corresponding host core genes at synonymous sites. Two prophages are considered as orthologous when they display an average dS ratio comprised in the 99% interval of the dS distribution of the core genes of the corresponding hosts. 4) The resulting orthologous prophage families must display gene diversity (pan-genome size) compatible with a single integration event. For each family of orthologs, the size of pan-genome was compared to the largest prophage of the family. Outliers were split into smaller families or removed from the analysis (see Materials and Methods).

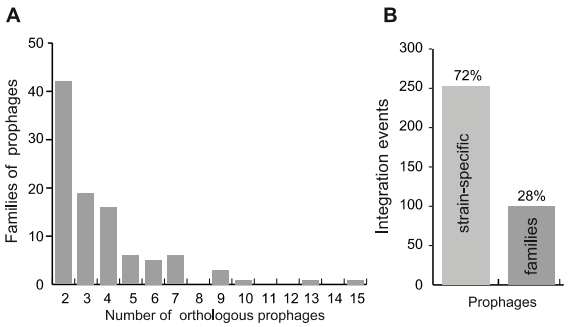

Figure S4. Frequencies of orthologous and non-orthologous prophages. (A) Histogram of the number of orthologous prophages in the 100 prophage families. **(B)** Number of strain-specific prophage families (i.e. prophages with no orthologs detected) and number of prophage families with more than one member.

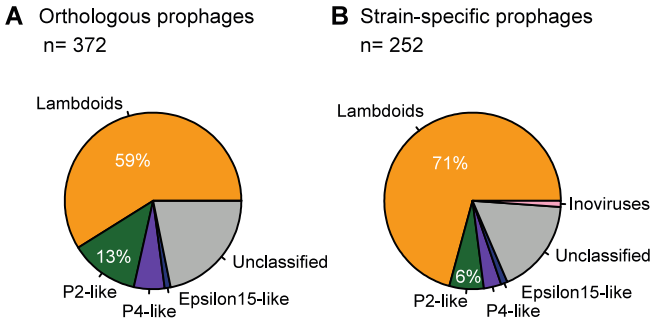

Figure S5. Taxonomic distribution of orthologous prophages (A) and prophages with no orthologs detected (B). Taxonomic groups were attributed by comparison of the gene content between prophages and classified phages of GenBank (see Materials and Methods). The two distributions are not significantly different (P=0.2, $\chi 2$ test). n corresponds to the number of prophages.

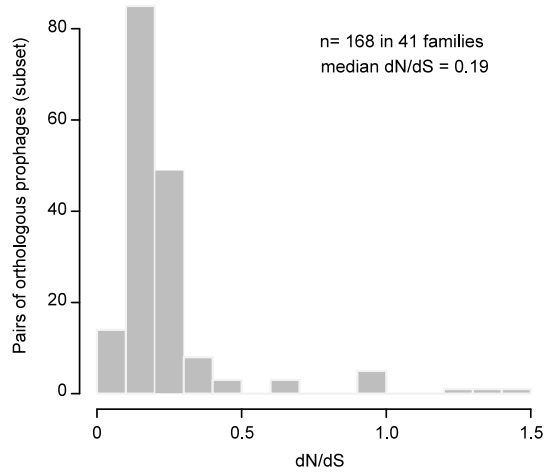

Figure S6. Histogram of the ratios of non-synonymous to synonymous substitutions (dN/dS) in orthologous prophages defined with more stringent parameters. We defined a subset of orthologous prophages that evolve very closely at the rate of their hosts. Additionally to the previous parameters, prophages have been defined as orthologs when $dS_H \leq dS_P \leq 2.dS_H$ with dS_P the synonymous substitution rate of a pair of prophages and dS_H the average synonymous rate of the core genes of the corresponding hosts. Orthologous prophages defined with these more stringent parameters display a slightly lower distribution of dN/dS ratios than our main data set of orthologous prophages (median values are 0.19 and 0.22 respectively, $P<0.01$, Wilcoxon Test).

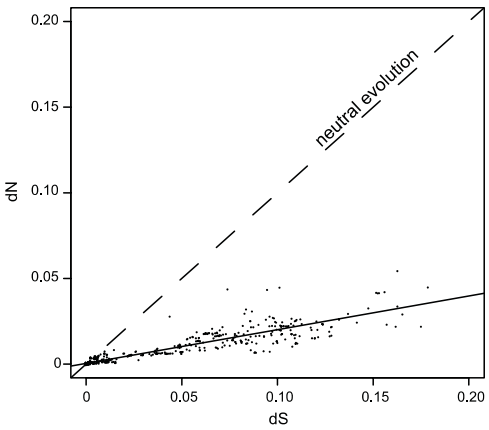

Figure S7. Correlation between dN and dS of orthologous prophages. The dashed line dN = dS represents the correlation between dN and dS under neutral evolution. The solid line corresponds to the linear regression (least squares method) of dN~dS (slope=0.2, $P<0.0001$). This plot shows that orthologous prophages with robust signal of substitutions (high dS) also display low dN/dS ratios.

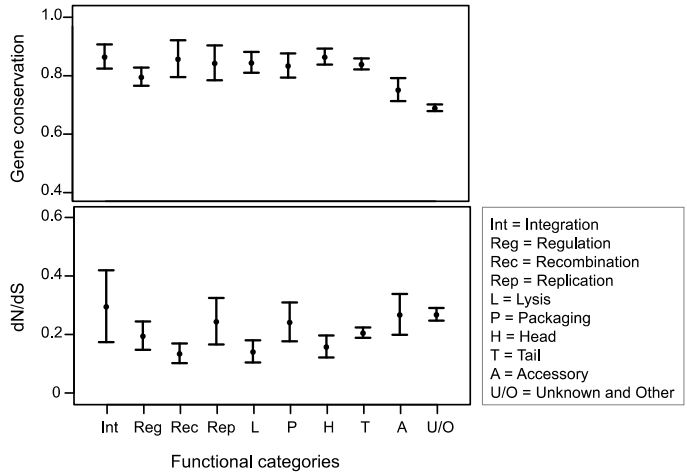

Figure S8. Average gene conservation (top) and dN/dS values (bottom) of the different functional categories of orthologous prophages. All the orthologous prophages were considered. Functional categories are indicated in the right insert.

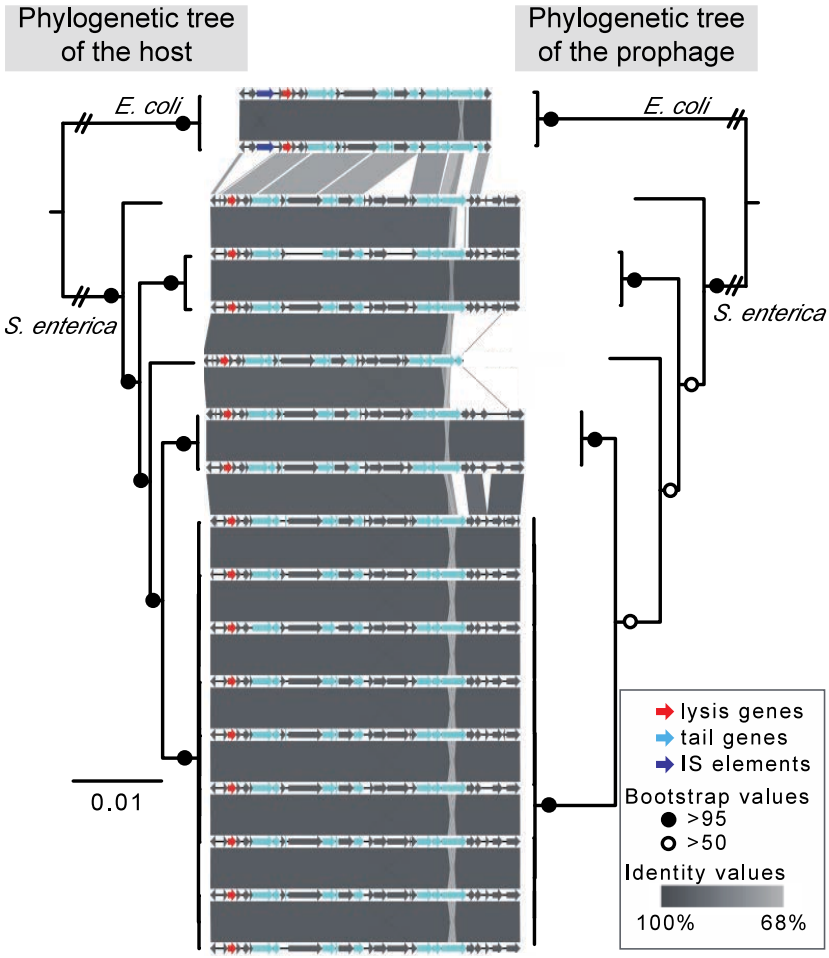

Figure S9. Representation of a putative case of prophage domestication into a R-type bacteriocin. The orthologous prophage family is present among 15 strains of *S. enterica*. Two additional orthologous prophages are integrated within two strains of *E. coli* at the same locus. Both prophage families are integrated between flanking core genes with homologs in both *S. enterica* and *E. coli*. Prophage genomes were represented with easyfig (15). Nucleotide sequence identity has been computed with blastn implemented in easyfig with default parameters. The maximum likelihood tree of the 17 host core genomes (2466 concatenated core genes) is represented on the left. The maximum likelihood tree of the 17 prophages is represented on the right and was performed on 12 concatenated genes shared by all 17 prophages. Both trees were built using RaxML (13) with the GTR + GAMMA model. Bootstrap replicates were computed with the rapid bootstrap analysis mode and are displayed at the nodes. The long branches between *E. coli* and *S. enterica* are not represented and are indicated by broken lines. Both trees were rooted at the *E. coli*/*S. enterica* junction.

SI table

Table S1. Results of the analysis of recombination in prophages using different methods. The table shows the number of orthologous prophage families, the number of gene families and the number of genes tested. It shows the number of gene families where recombination was identified. The remaining gene families were used to compute average dN/dS values and these were compared with the expected neutral value (dN/dS=1) with a Wilcoxon signed rank test for non-recombinant genes. Neutrality is rejected in all cases with high confidence. Description of the methods is in SI Materials and Methods. The last column (Σ) describes the results of the combination of the three approaches (PHI/NSS/MaxChi, Prunier and MaxChi on concatenates).

	Approach							
	Individual gene families				Concatenate	Prunier		Σ
Method	NSS (p*)	MaxChi (p*)	PHI (p*)	PHI	MaxChi	Min	Max	Σ
Orthologous prophage families tested	58	58	58	58	57	39	39	24
Orthologous gene families tested	313	313	313	280	404	219	219	131
Genes tested	2037	2037	2037	1856	6246	2624	2624	1017
% recombinants	18%	32%	13%	20%	8%	10%	13%	16%
Median dN/dS of non-recombinant genes	0.16	0.15	0.16	0.16	0.14	0.12	0.12	0.08
Different dN/dS in non-recombining vs recombining genes (P^+)	0.86	0.027	0.893	0.850	0.023	0.953	0.953	0.0565
dN/dS<1 in non-recombining genes (P^+)	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

*: 1000 permutations
+: P-value computed using the Wilcoxon test

Article 4