The Adaptation of Temperate Bacteriophages to Their Host Genomes

M. Bobay, E. P. C. Rocha, M. Touchon

▶ **To cite this version:**

M. Bobay, E. P. C. Rocha, M. Touchon. The Adaptation of Temperate Bacteriophages to Their Host Genomes. Molecular Biology and Evolution, 2013, 30 (4), pp.737 - 751. 10.1093/molbev/mss279 . pasteur-01374945

HAL Id: pasteur-01374945

https://pasteur.hal.science/pasteur-01374945

Submitted on 2 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

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

© The Author(s) 2012. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

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

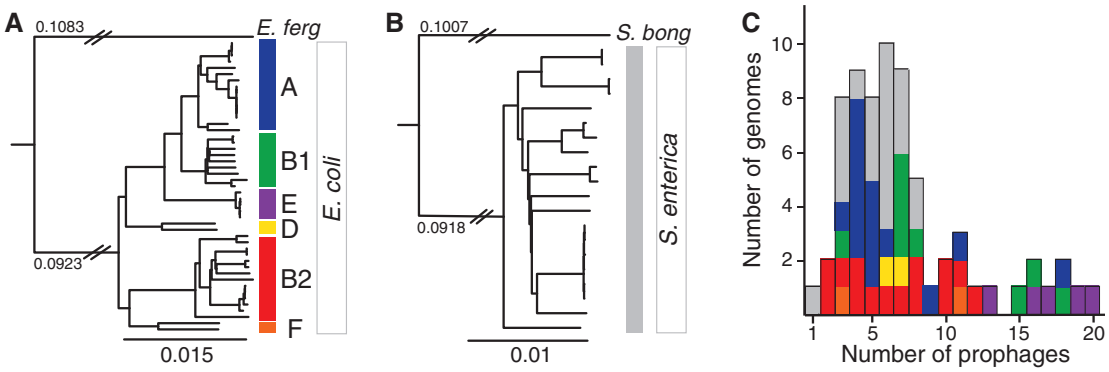

Fig. 1. Core genome phylogenies and prophage content of *Escherichia* and *Salmonella*. (*A*) Maximum likelihood phylogenetic tree of the 47 *Escherichia coli* strains. (*B*) Maximum likelihood phylogenetic tree of the 20 *Salmonella enterica* strains. *Escherichia fergusonii* and *S. bongori* were used 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

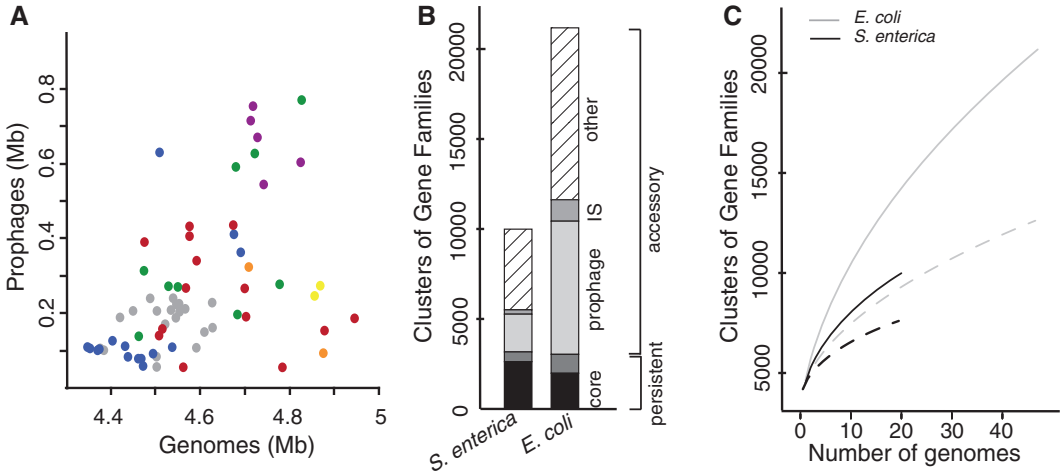

Fig. 2. Contribution of prophages to chromosome plasticity. (*A*) Scatter plot of cumulative size of resident prophages against the size of the host genome (Spearman's $\rho = 0.52$, $P < 0.0001$). Colors correspond to the phylogenetic groups as in figure 1. (*B*) Fraction of the core, persistent, and 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

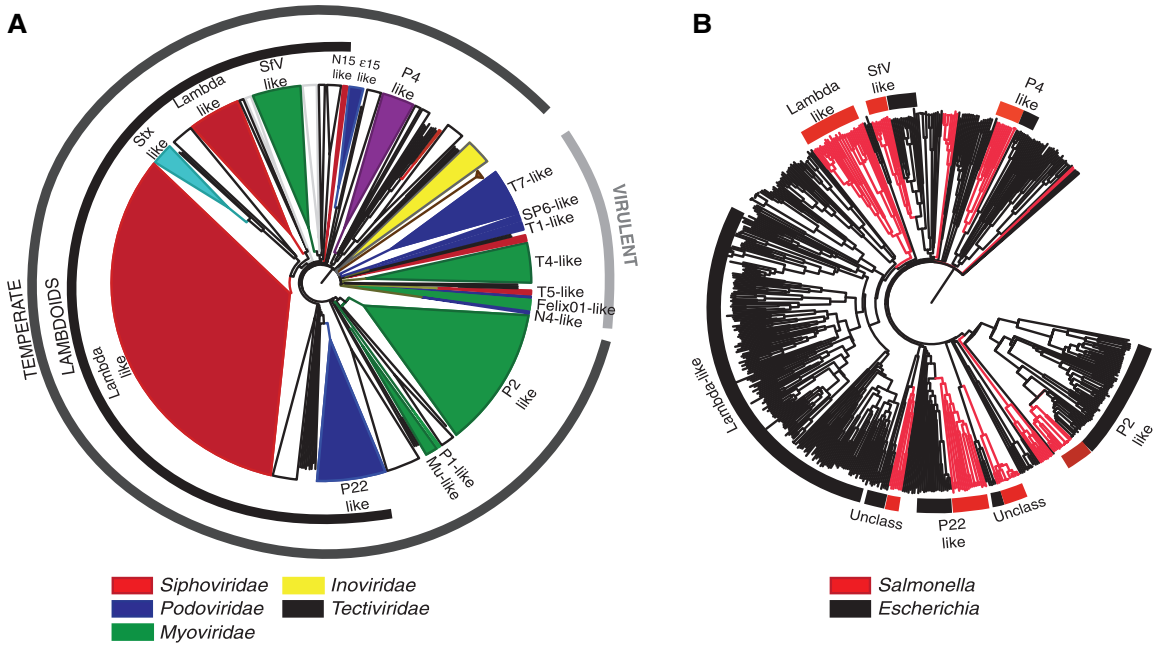

Fig. 3. Classification of prophages. (*A*) Phylogenetic tree of phages and prophages based on gene repertoire relatedness (see Materials and Methods). Phage/prophage families are colored according to the color key. The phage/prophage genus is indicated in the inner circle. The members of the "lambdoid" 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.

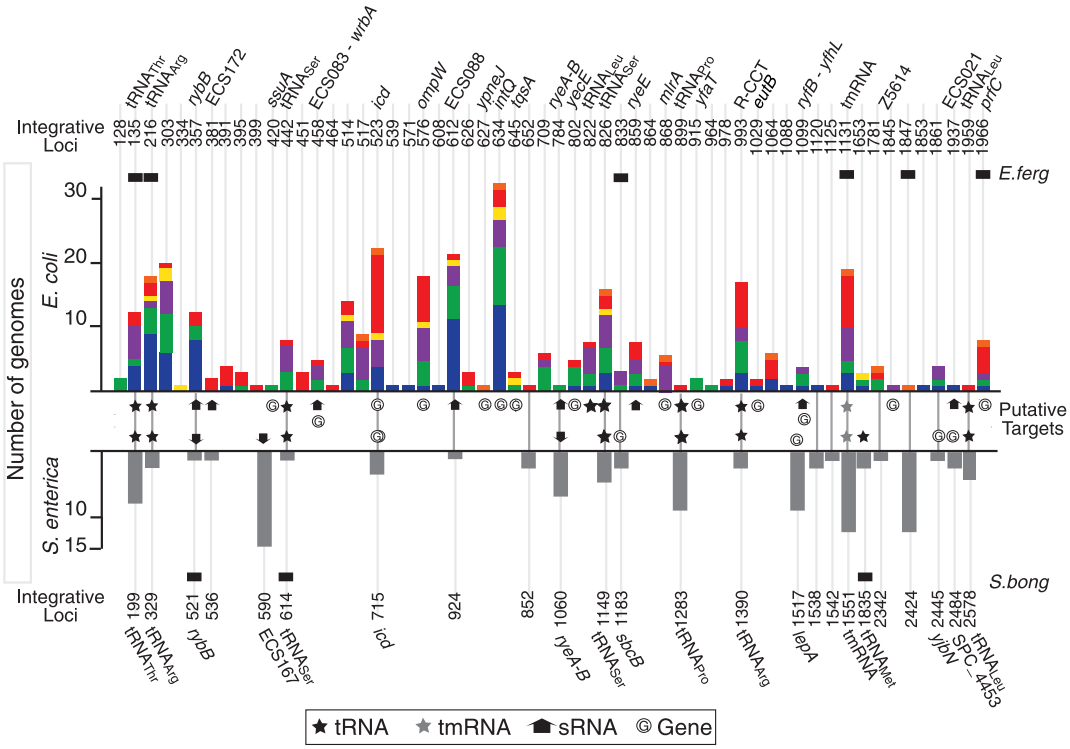

Fig. 4. Distribution of prophages at integration hotspots. The *x* axis indicates the position of the hotspots of phage integration in the genomes of *Escherichia coli* (top) and *Salmonella enterica* (bottom). The positions of the "integrative loci" (on top for *E. coli* and bottom for *S. enterica*) 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*, *yneJ*, *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*, *yneJ*, *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.

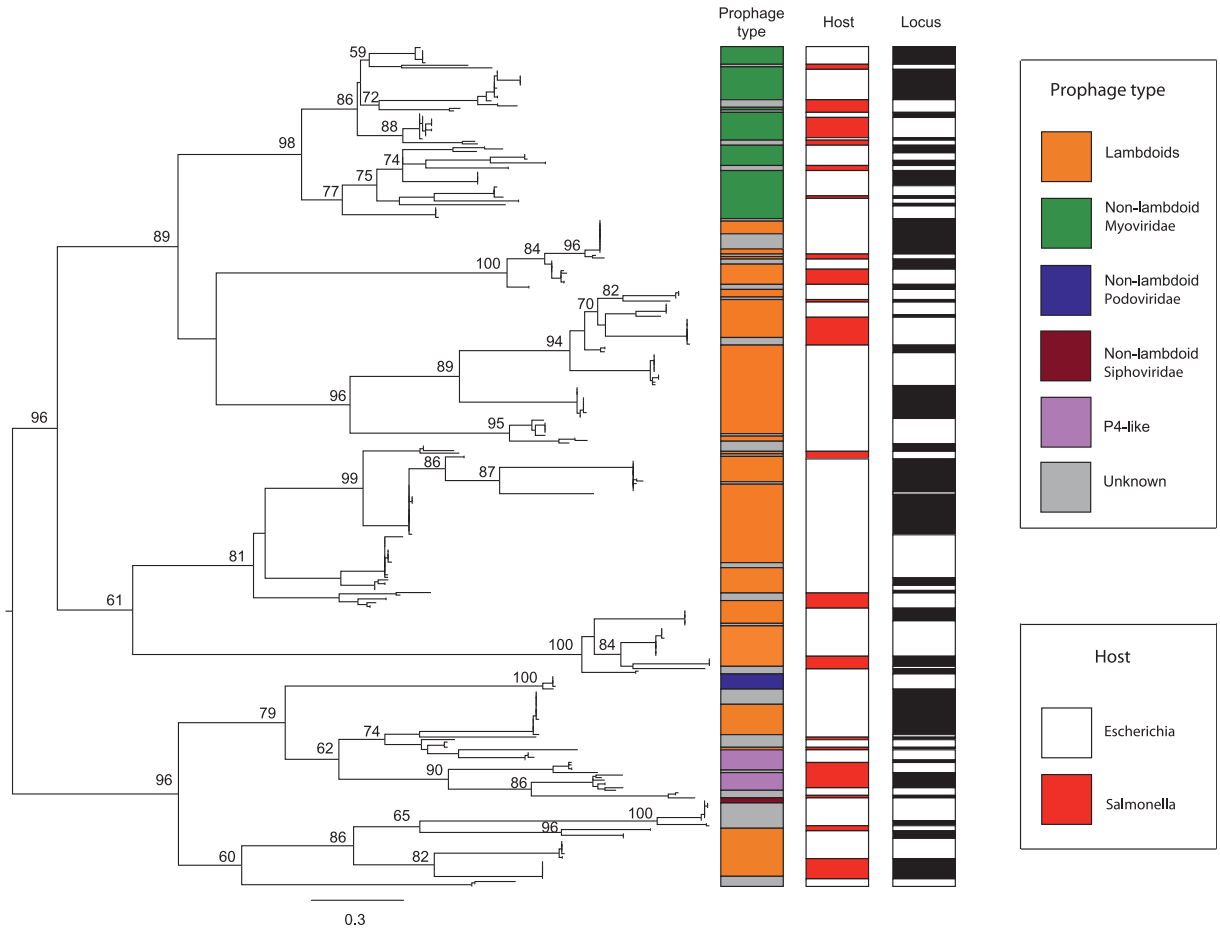

Fig. 5. Phylogeny of the integrases. The maximum likelihood tree was made from a trimmed alignment of 332 tyrosine recombinases and rooted using the midpoint root. Bootstrap values (out of 1,000 replicates) are given in percents in the tree and are shown when exceeding 50%. Prophage types are indicated 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.

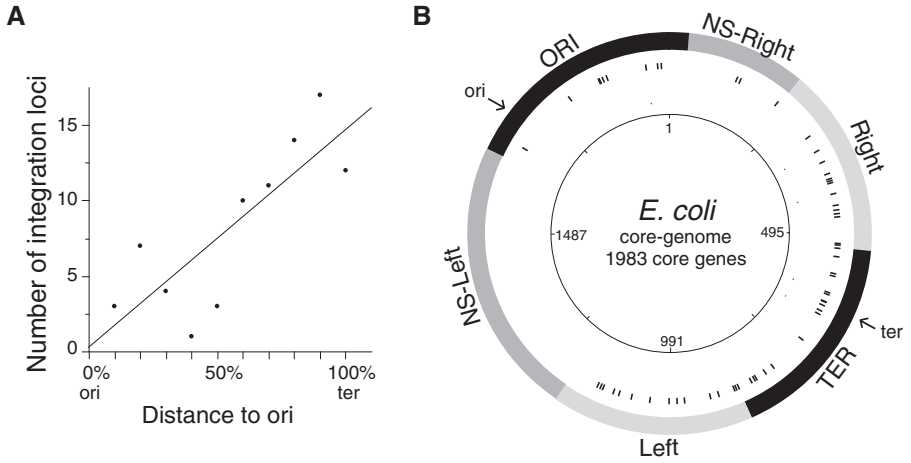

FIG. 6. Distribution of prophages in the chromosome. (*A*) Number of loci with prophages in function of the distance to the origin of replication. Distribution of integration loci in function of the distance to the origin of replication (ori: origin and ter: terminus). (*B*) Circular representation of 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

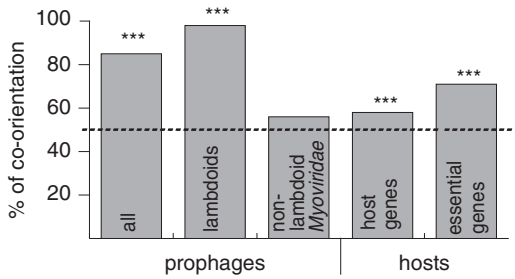

Fig. 7. Percentage of prophages and host genes co-oriented with the replication fork. The dotted line shows the polarization under random expectation (50%). $P < 0.05$ (*); $P < 0.01$ (**); $P < 0.001$ (***).

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.