Regulation of genetic flux between bacteria by restriction–modification systems

Pedro H. Oliveira, Marie Touchon, Eduardo P. C. Rocha

Regulation of genetic flux between bacteria by restriction–modification systems

Pedro H. Oliveira[a,b,1], Marie Touchon[a,b], and Eduardo P. C. Rocha[a,b]

[a]Microbial Evolutionary Genomics, Institut Pasteur, 75015 Paris, France; and [b]CNRS, UMR 3525, 75015 Paris, France

Edited by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved April 5, 2016 (received for review March 2, 2016)

Restriction–modification (R-M) systems are often regarded as bacteria's innate immune systems, protecting cells from infection by mobile genetic elements (MGEs). Their diversification has been recently associated with the emergence of particularly virulent lineages. However, we have previously found more R-M systems in genomes carrying more MGEs. Furthermore, it has been suggested that R-M systems might favor genetic transfer by producing recombinogenic double-stranded DNA ends. To test whether R-M systems favor or disfavor genetic exchanges, we analyzed their frequency with respect to the inferred events of homologous recombination and horizontal gene transfer within 79 bacterial species. Genetic exchanges were more frequent in bacteria with larger genomes and in those encoding more R-M systems. We created a recognition target motif predictor for Type II R-M systems that identifies genomes encoding systems with similar restriction sites. We found more genetic exchanges between these genomes, independently of their evolutionary distance. Our results reconcile previous studies by showing that R-M systems are more abundant in promiscuous species, wherein they establish preferential paths of genetic exchange within and between lineages with cognate R-M systems. Because the repertoire and/or specificity of R-M systems in bacterial lineages vary quickly, the preferential fluxes of genetic transfer within species are expected to constantly change, producing time-dependent networks of gene transfer.

homologous recombination | horizontal gene transfer | bacterial evolution

Prokaryotes evolve rapidly by acquiring genetic information from other individuals, often through the action of mobile genetic elements (MGEs) such as plasmids or phages (1). In bacterial population genetics, the events of gene transfer are usually termed horizontal gene transfer (HGT) when they result in the acquisition of new genes and homologous recombination (HR) when they result in allelic replacements. The distinction between the two evolutionary mechanisms (HGT and HR) is not always straightforward: incoming DNA may integrate the host genome by double crossovers at homologous regions, leading to allelic replacements in these regions and to the acquisition of novel genes in the intervening ones. HR takes place only between highly similar sequences, typically within species (2). As a result, it usually involves the exchange of few polymorphisms, eventually in multiple regions, between cells (3). It may also result in no change if the recombining sequences are identical, which leaves no traces and cannot be detected by sequence analysis. HGT may occur between distant species, resulting in the acquisition of many genes in a single event. The replication and maintenance of MGEs have fitness costs to the bacterial host and have led to the evolution of cellular defense systems. These systems can sometimes be counteracted by MGEs, leading to evolutionary arms races.

Restriction–modification (R-M) systems are some of the best known and the most widespread bacterial defense systems (4). They encode a methyltransferase (MTase) function that modifies particular DNA sequences in function of the presence of target recognition sites and a restriction endonuclease (REase) function that cleaves them when they are unmethylated (5). R-M systems are traditionally classified into three main types. Type II systems are by far the most abundant and the best studied (6). With the exception of the subType IIC, they comprise MTase and REase functions encoded on separate genes and are able to operate independently from each other. R-M systems severely diminish the infection rate by MGEs and have been traditionally seen as bacteria's innate immune systems (7). However, successful infection of a few cells generates methylated MGEs immune to restriction that can invade the bacterial population (8). Hence, R-M systems are effective as defense systems during short periods of time and especially when they are diverse across a population (9, 10). In particular, it has been suggested that they might facilitate colonization of new niches (11). Type II R-M systems are also addictive modules that can propagate selfishly in populations (12). Both roles of R-M systems, as defense or selfish systems, may explain why they are very diverse within species (13, 14). Accordingly, R-M systems endure selection for diversification and are rapidly replaced (15, 16).

Several recent large-scale studies of population genomics have observed more frequent HR within than between lineages (17, 18). This suggests that HR might favor the generation of cohesive population structures within bacterial species (19). Specific lineages of important pathogens that have recently changed their R-M repertoires show higher sexual isolation, such as *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Burkholderia pseudomallei*, and *Staphylococcus aureus* (20–22). For example, a Type I R-M system decreased transfer to and from a major methicillin-resistant *S. aureus* lineage (23). Diversification of R-M target recognition sites could thus reduce transfer between lineages with different systems while establishing preferential gene fluxes between those with R-M systems recognizing the same target motifs (cognate R-M). However, these results can be confounded by evolutionary distance: closely related genomes are more likely to encode similar R-M systems, inhabit the same environments (facilitating transfer between cells), and have

Significance

The role of restriction–modification (R-M) as bacteria's innate immune system, and a barrier to sexual exchange, has often been challenged. Recent works suggested that the diversification of these systems might have driven the evolution of highly virulent bacterial lineages. Here, we showed that R-M systems were more abundant in species enduring more DNA exchanges and that within-species flux of genetic material was higher when cognate systems were present. Presumably, bacteria enduring frequent infections by mobile elements select for the presence of more numerous R-M systems, but rapid diversification of R-M systems leads to varying patterns of sexual exchanges between bacterial lineages.

Author contributions: P.H.O. and E.P.C.R. designed research; P.H.O., M.T., and E.P.C.R. analyzed data; and P.H.O., M.T., and E.P.C.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: pcphco@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1603257113/-/DCSupplemental.

similar sequences (that recombine at higher rates). The advantages conferred by new genes might be higher when transfer takes place between more similar genetic backgrounds.

Here, we aimed at testing the effect of R-M systems on the genetic flux in bacterial populations. We concentrated on Type II R-M systems because they are the best studied, very frequent, and those for which we could predict sequence specificity. We inferred genome-wide counts of HR and HGT and tested their association with the frequency of R-M systems encoded in the genomes. We then made a more precise test of the key hypothesis that bacteria carrying similar R-M systems establish highways of gene transfer, independently of phylogenetic proximity and clade-specific traits.

Results

Quantification of Homologous Recombination, HGT, and Their Covariates. We analyzed a dataset of 79 core genomes and pangenomes (*SI Methods*) corresponding to a total of 884 complete genomes. These clades were based on taxonomy, i.e., the genomes of a named species were put together. They spanned many different bacterial phyla (Fig. 1*A* and *SI Methods*). The pangenomes varied between 466 and 18,302 gene families (Dataset S1), and correlated with genome size (Spearman's $\rho = 0.89$, $P < 10^{-4}$) and phylogenetic depth, defined as the average root-to-tip distances in the clade phylogenetic tree (*SI Methods* and Dataset S2) (Spearman's $\rho = 0.42$, $P < 10^{-4}$). Hence, our dataset represents a large diversity of bacteria in terms of taxonomy, genome size, and intraspecies diversity.

HR is notoriously difficult to quantify accurately (24). We used five different programs to detect HR in the core genome (*SI Methods*). These programs detect different types of signals, and together they should provide a thorough assessment of HR. Among the 79 core genomes, we found an average of 329 (NSS), 374 (MaxCHI), 264 (PHI), 504 (Geneconv), and 1,035 (ClonalFrameML, CFML) HR events per core genome (Datasets S1 and S3). Even if the different methods provided different numbers of events, their results were highly correlated (average Spearman's $\rho = 0.84$, all comparisons $P < 10^{-4}$). Accordingly, we focused our analysis on the results of Geneconv, which provides the positions of recombination tracts and directions of transfer necessary for the last part of this study.

We used Count (25) to infer the events of HGT from the patterns of presence and absence of gene families in the species' trees (*SI Methods* and Dataset S4). We identified 236,894 events of gene transfer in the 79 pangenomes (Dataset S1). These events were very

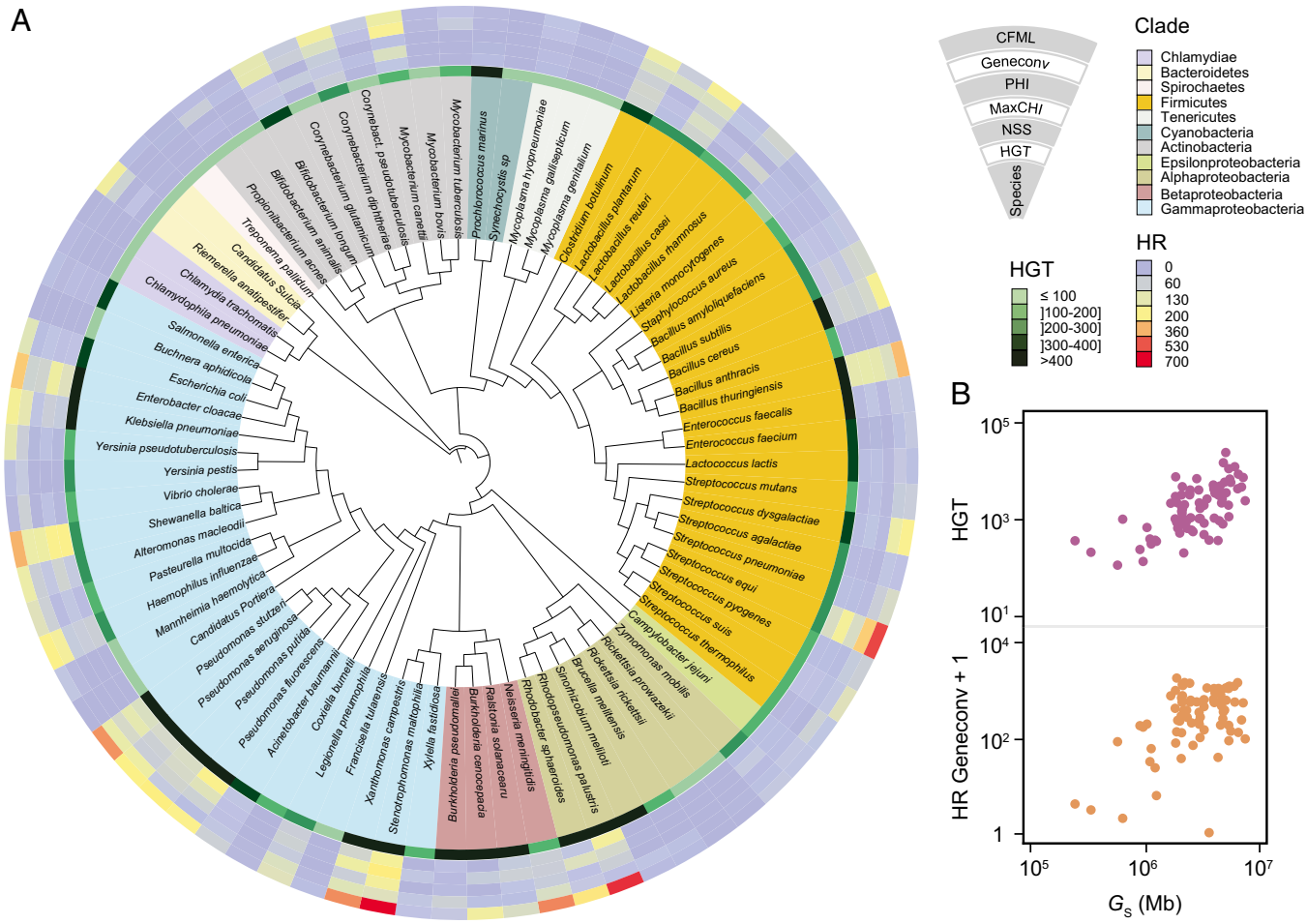

Fig. 1. Analysis of HR and HGT events. (*A*) 16S rRNA phylogenetic tree of the 79 bacterial species. The tree was drawn using the iTOL server (itol.embl.de/index.shtml) (40). The innermost circle layer indicates the species and associated clade. The six subsequent layers correspond (in an outwardly direction) to the average number of HGT events per genome computed using Count; the number of recombined genes per genome given by NSS, MaxChi, and PHI; and the number of recombination events per genome given by Geneconv and CFML (outermost layer), respectively. These values are given in Dataset S1. (*B*) Distribution of the average number of horizontal gene transfer (HGT) events and homologous recombination (HR) events (inferred by Geneconv) per clade according to genome size (G_S). Spearman's $\rho_{HGT} = 0.65$, $P_{HGT} < 10^{-4}$; Spearman's $\rho_{Geneconv} = 0.32$, $P_{Geneconv} < 10^{-2}$. Data obtained with the remaining recombination inference tools are shown in Fig. S1.

unevenly distributed among clades, from close to none in the genomes of obligatory endosymbionts to 1,538 events per genome in *Rhodopseudomonas palustris* (Fig. 1*A*).

The frequencies of HR and HGT were expected to depend on a number of variables, including the following: (*i*) genome size; (*ii*) phylogenetic depth (deeper lineages accumulate more events of exchange); and (*iii*) the number of genomes in the clade (larger samples capture more past events). We built stepwise linear models to assess the role of these variables in explaining the variance in HGT and HR (Table S1, part A). These showed that genome size had a strong direct effect on HR and HGT (Fig. 1*B* and Fig. S1). The remaining variables had significant, but less important, explanatory roles. HR also depended weakly on core genome size (Table S1, part B). Hence, studying the effect of R-M systems on HR and HGT requires control for phylogenetic depth, the number of genomes in the clade, and especially the genome size.

Association Between R-M Systems and Genetic Transfer. We identified 1,352 R-M systems among the 79 clades using a previously published methodology (4) (*SI Methods* and Dataset S1), including 233 Type II R-M systems (excluding Type IIC). The number of HGT events was higher in genomes with more R-M systems (Fig. 2*A*), and especially in those with Type II systems (Fig. 2*B*). The number of HR events increased with the number of R-M systems (Fig. 2*C*) and especially in the presence of Type II R-M systems (Fig. 2*D*). Similar results were obtained for the remaining HR inference tools (Fig. S2).

We then tested the effect of R-M systems on the number of HGT events and the rates of HR, while controlling for their covariates mentioned above. A stepwise regression showed that the numbers of Type II R-M systems were not significant predictors of HGT when the three previous variables were already introduced in the regression (the latter explaining ~76% of all variance; Table S1, part C). An analogous analysis for the frequency of HR showed that genome size and the number of Type II R-M systems were both significant predictors of HR ($R^2 = 0.42$, both variables $P < 10^{-4}$; Table S1, part C). These results show that genomes carrying more R-M systems acquire more genetic material by both HR and HGT, even if the latter association might be the result of clade-specific traits such as genome size.

Evolution of Target Motifs and Identification of Cognate R-M Systems. To test the hypothesis that R-M systems affect the genetic flux between genomes, one needs to identify the systems recognizing the same target recognition motif. Such systems are cognates, i.e., DNA methylation by one system will protect from the other. We could not identify a method to identify cognate R-M systems in the literature. Hence, we created one based on the sequence

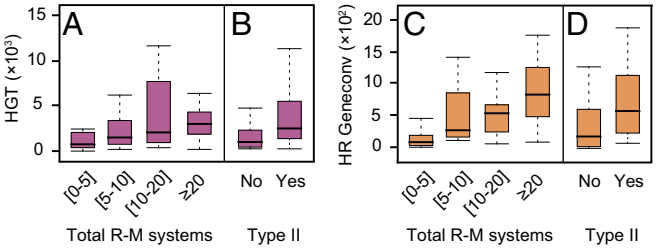

Fig. 2. Association between gene transfer and R-M systems. Distribution of the average HGT events (*A*) and homologous recombination (HR) events inferred by Geneconv (*C*) per clade according to the total number of R-M systems. Spearman's $\rho_{HGT} = 0.43$, Spearman's $\rho_{Geneconv} = 0.62$; both $P < 10^{-4}$. Distribution of the average HGT (*B*) and Geneconv HR events (*D*) per clade according to the presence (Yes)/absence (No) of Type II R-M systems (both $P < 10^{-4}$; Mann–Whitney–Wilcoxon test). We obtained similar qualitative results with the remaining recombination inference tools (Fig. S2).

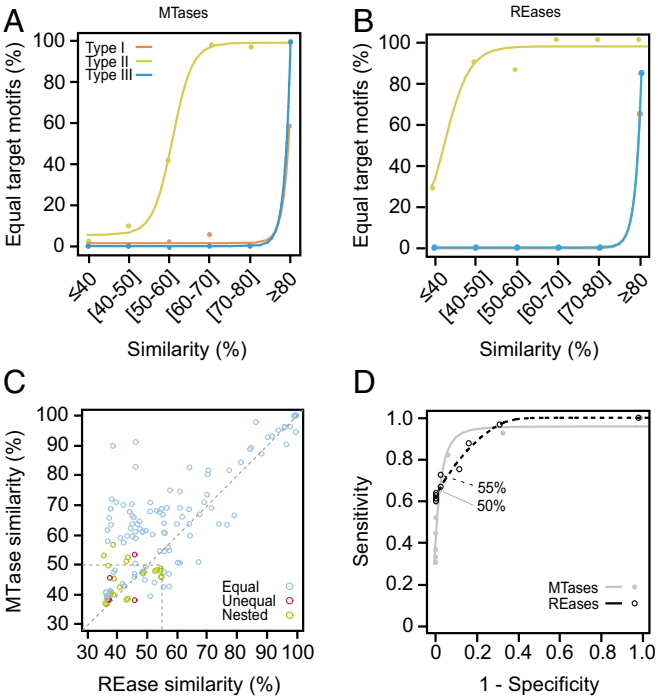

Fig. 3. Relation between target specificity and protein similarity in R-M components. Percentage of equal target motifs recognized by Types I, II, and III MTases (*A*) and REases (*B*) according to their pairwise protein sequence similarity. (*C*) Plot of all pairwise similarities of Type II MTases versus the cognate Type II REases of the REBASE gold standard. Blue dots correspond to equal target motifs, red dots to unequal target motifs, and green dots to nested motifs. The dashed horizontal and vertical lines indicate the threshold similarity limits for MTases and REases. (*D*) The same dataset was used to plot the corresponding receiver operating characteristic (ROC) curves. These curves depict the Sensitivity (true-positive rate) versus 1-Specificity (false-positive rate) for several values of percentage similarity of Type II MTases and REases. We selected the cutoff values of similarity that maximized the true-positive rate and minimized the false-positive rate. Details on the number of R-M proteins of each type can be found in Table S2. ROC data including curve-fitting equations can be found in Table S3.

conservation of MTases and REases. For this, we used the "gold-standard" component of REBASE (26) and plotted the frequency with which MTases or REases of a given type recognized the same motif (*SI Methods*) for a given bin of sequence similarity. Only nearly identical homologs of Types I and III MTases and REases recognized the same motifs (Fig. 3 *A* and *B*). The analysis of the Specificity (S) and target recognition domains (TRDs) led to similar conclusions (*SI Methods* and Fig. S3*A*). The small number of such systems in REBASE gold standard resulted in small statistical power for this analysis, but adding more recent data from REBASE PacBio database did not change these conclusions (*SI Methods* and Fig. S3 *B–E*). The rapid evolution of sequence target specificity precludes the identification of systems with similar restriction sites from the alignment of REases or MTases in both Type I and Type III R-M systems.

In contrast, homologs of Type II REases and MTases, which are much more numerous in the database, have different target motifs only when their sequence similarity is low (typically less than 50% for MTases and 55% for REases; Fig. 3). We used these thresholds to estimate the probability that two homologous systems recognize the same target recognition motif, and restricted our subsequent analyses to Type II systems.

R-M Systems Promote Preferential Genetic Transfer Fluxes. The observation of higher genetic fluxes in the presence of R-M systems might seem unexpected in the light of the role of the latter in

degrading exogenous DNA. To explain these results, we put forward three hypotheses.

Hypothesis 1: The relative abundance of R-M systems in a clade results from the selective pressure imposed by the abundance of MGEs in that clade. Selection for multiple R-M systems is expected to be stronger for clades enduring infections by many MGEs. R-M systems have limited efficiency and might not completely prevent MGE infection and transfer (8). This results in a weak positive association between transfer of genetic information and the abundance of R-M systems.

Hypothesis 2: R-M systems favor transfer of genetic material between cells by generating restriction breaks that stimulate recombination between homologous sequences.

Hypothesis 3: Type II R-M systems encoded in MGEs favor genetic transfer by selfishly stabilizing the element's presence in the new host (16). Genomes enduring more transfer would have more R-M systems if they were carried by MGEs. This last hypothesis is unlikely to explain our results, because we have shown that R-M systems are rare in MGEs (4). Furthermore, the association between genetic transfer and number of R-M systems remained significant when we excluded Type II R-M systems from the analysis (those more likely to act as selfish elements; Fig. S4). This fits recent findings that R-M systems occur and recombine in genomes in ways that are independent of the presence of MGEs (5).

To distinguish between the first two hypotheses, we analyzed the genetic flux between pairs of genomes with cognate Type II R-M systems. If R-M systems predominantly prevent genetic transfer (hypothesis 1), then the flux of genetic material between genomes encoding cognate R-M systems should be higher. If R-M systems predominantly stimulate genetic transfer (hypothesis 2), then pairs of genomes encoding cognate R-M systems should show lower than average genetic flux.

We tested the two hypotheses for HR and HGT separately. We selected the HR events that took place between terminal branches in the phylogenetic trees of the clades. Each terminal branch was then associated with the respective focal genome (the tip), which was labeled in terms of the target recognition motifs of the R-M systems encoded in the focal genome. We excluded HR or HGT occurring in the internal branches of the tree because of the high incertitude in the inference of ancestral R-M systems (Fig. S5). We then computed the number of HR events between terminal branches associated with genomes encoding cognate R-M systems and compared it with the other pairs of genomes encoding R-M systems. Similar analyses were performed for HGT events that simultaneously affected pairs of terminal branches, i.e., for genes transferred to two terminal branches in two independent events. In both cases, we observed that lineages represented by genomes encoding cognate R-M systems coexchanged more genetic information (Fig. S6 *A* and *B*).

Next, we restricted our analysis to clades having at least 10 comparisons between genomes encoding cognate R-M systems and 10 comparisons between genomes lacking cognate systems (but encoding R-M systems). This avoids the confounding effect of putting together in the same analysis clades with few R-M systems or with little diversity in these systems. This restricted our dataset to eight clades: *Bacillus amyloliquefaciens*, *Bifidobacterium longum*, *Escherichia coli*, *Haemophilus influenzae*, *Listeria monocytogenes*, *N. meningitidis*, *Salmonella enterica*, and *S. pneumoniae*. Within this restricted dataset, the results were qualitatively identical: lineages associated with genomes encoding cognate R-M systems coexchanged more genetic information (Fig. 4 *A*–*C*). We confirmed that these results were insensitive to uncertainties in phylogenetic reconstruction and to the effects of HR in phylogenetic inference (*SI Methods* and Fig. S7). The

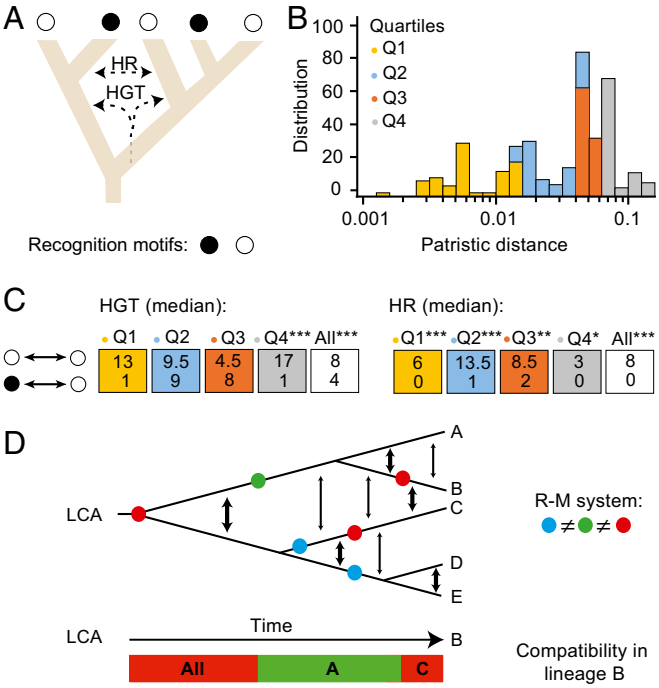

Fig. 4. Gene flux in bacteria encoding R-M systems. (*A*) We analyzed the patterns of HR and HGT in the tree of each clade, comparing the flux between tips ending in cognate (similar recognition motifs) or noncognate (different motifs) extant taxa. (*B*) Histogram of patristic distances (colored by quartiles) between bacteria with Type II R-M systems. (*C*) Median values of HGT and recombination events for each quartile (Q) and for the full dataset (All) between terminal branches of bacteria with Type II R-M systems recognizing (or not) the same target motif. We analyzed *Bacillus amyloliquefaciens*, *Bifidobacterium longum*, *Escherichia coli*, *Haemophilus influenza*, *Listeria monocytogenes*, *Neisseria meningitidis*, *Salmonella enterica*, and *Streptococcus pneumoniae*. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$ (see Fig. S6 *A* and *B* for the data including all clades). (*D*) Genetic flux in function of time and the presence of R-M systems. As lineages diverge and R-M systems change (circles indicate such changes), the lineages with cognate R-M systems (same color) share more genetic material than the other lineages. For example, the lineage B changes R-M systems twice since the last common ancestor (LCA). Initially transfer is favored with all lineages, then with the sister lineage A, and finally with the distantly related lineage C.

results on HR might be strongly affected by the ability of bacteria to engage in natural transformation. We restricted our analysis to the five naturally transformable species, following ref. 27, and found similar results ($P < 10^{-3}$).

We then tested whether the clade-associated traits covarying with HR and HGT—phylogenetic depth, average genome size, and number of genomes—were affecting our conclusions by making the comparisons on each clade separately. We observed more HGT and HR among pairs of genomes encoding cognate R-M systems in six of the eight clades, which was statistically significant (each $P = 0.035$, binomial test, $P = 0.01$ for the combined test). One species (*L. monocytogenes*) was an exception to the general trend both concerning HR and HGT. This species showed very low rates of HGT and HR, and the differences in HR and HGT between R-M cognate and R-M noncognate genomes were not significant.

We mentioned in the Introduction that closely related taxa are expected to exchange more genetic information independently of the R-M systems they encode. To verify that the presence of cognate R-M systems is associated with increased genetic exchange independently of evolutionary distance, we binned the comparisons between events occurring in terminal branches in terms of the phylogenetic distance between pairs of genomes. We

then ran the same analysis in each bin separately. These analyses showed more cotransfer between genomes encoding cognate R-M systems in nearly all bins, even if this analysis had lower statistical power (fewer comparisons per bin) (Fig. 4 *B* and *C* for the eight clades and Fig. S6 *A* and *B* for all of the data). Importantly, this difference was always significant for the most distant pairs of genomes. Hence, pairs of genomes encoding cognate R-M systems were associated with more frequent HR and HGT, independently of the evolutionary distances between them.

Discussion

Genome size is the result of the balance between accretion and deletion events moderated by natural selection. Larger bacterial genomes are expected to engage in more frequent HGT because this is the dominant mechanism of genetic accretion (28). However, there are remarkably few studies demonstrating an association between HGT and genome size (29). Here, we found that larger genomes exchange DNA at higher rates, both by HGT and by HR. This association is not just caused by sexually isolated endosymbiotic bacteria with very small genomes—e.g., *Chlamydiae*, *Buchnera*, or *Spirochaetes* (Fig. 1 and Dataset S1)—because it remains significant for genomes larger than 2 Mb, which include few obligatory endosymbionts. Many reasons might explain the association between HGT, HR, and genome size: bacteria with larger genomes might have more diverse lifestyles, select for more diverse types of genes, inhabit more environments, or accommodate more MGEs. Even if the test of these different hypotheses falls outside the scope of this work, this association is important and must be accounted for when assessing the impact of R-M systems in genetic fluxes. The higher frequency of HR and HGT among larger genomes suggests that the latter are more targeted by MGEs. Accordingly, larger bacterial genomes encode more transposable elements (30), more prophages (31), and more conjugative elements (32). If MGEs targeting bacteria with larger genomes are more abundant, they might lead to strong selection for R-M systems in their bacterial hosts. This might explain why we found more R-M systems in larger genomes (4). It might also explain the positive association between the frequencies of HR and HGT and the abundance of R-M systems (Fig. 2).

R-M systems have a well-known inhibitory effect on the transfer of genetic information (9). However, whether this trait is an important driver of their evolution has remained controversial (12, 33, 34). Our results contribute to the clarification of these two issues. R-M systems can function as a barrier to MGE infection when encoded in the chromosome or other MGE. They can also stabilize the presence of MGEs in cells by preventing infections by other competing MGEs. Our previous observation that MGEs encode few R-M systems and many solitary MTases (4), suggests that R-M systems are more frequently a chromosomal-encoded barrier to MGEs than an MGE-encoded tool for cell infection. The coassociation of MGEs, bacterial genome size, and R-M systems might thus result from increased selection for R-M systems in the face of abundant MGEs in large genomes.

Contrary to the popular view that R-M systems limit the flux of genetic material (9), it has been proposed that restriction actually favors evolvability by producing DNA double-stranded ends that are recombinogenic (33, 34). This hypothesis is compatible with the observation that genomes enduring more HGT and HR have more R-M systems. However, it is not in agreement with the observation that pairs of genomes encoding cognate R-M systems coexchange more DNA. It is also hardly reconcilable with the notorious deleterious effect of R-M systems on bacterial genetic transformation in the laboratory (35). Although R-M systems have been shown to favor intragenomic HR events (12), the overall effect of R-M systems on genetic exchange is to decrease both HR and HGT between bacteria encoding noncognate R-M systems.

Our statistical analyses could not explicitly account for the presence of the many other systems affecting genetic transfer between cells. Some of them facilitate transfer, e.g., MGEs or competence for natural transformation, and we checked that all of the clades in Fig. 4 have known phages and conjugative elements. Restricting the analysis to the five naturally transformable bacteria did not change our results. Importantly, all of these clades encoded the key enzymes involved in RecA-mediated homologous recombination, including the presynaptic pathways RecBCD/AddAB and RecOR (36). Hence, there is little ground to think that our results are strongly biased by lack of mechanisms for gene transfer. Some systems disfavor transfer between bacteria, including CRISPRs, abortive infection, or other R-M systems. It is not possible to account for all these factors in a statistical model, because of the lack of quantitative data. Nevertheless, we could verify that cognate genomes did not have fewer R-M systems than the other genomes. Even if other barriers to DNA exchange are certainly present in these species, our use of a diverse set of well-known species, numerous alternative analyses, and focus on intraspecies comparisons (in which lifestyles and other general traits are much less variable), suggests that our results are robust.

Our work shows that noncognate genomes have reduced DNA exchanges. This decreases the power of natural selection and increases the effect of drift, potentially leading to the accumulation of deleterious mutations. Importantly, R-M systems' diversification at the origin of a lineage may increase its genetic cohesion by disfavoring exchanges with the closest related ones, as previously suggested for some pathogens (20–22). Interestingly, diversification can also increase the genetic flux between distant bacteria encoding cognate R-M systems with which there were previously few genetic exchanges. Hence, R-M systems might shape population structure in complex ways depending on the repertoire of R-M systems in the other lineages.

The study of the flux of genetic information among bacteria using network-based approaches is rising in importance (37–39). Our work shows that R-M systems may carve preferential routes of DNA exchange between certain bacterial subpopulations. Their rapid diversification constantly changes these preferences, thereby producing complex patterns of genetic exchange with time.

Methods

Details on the data used, identification of core genomes and pangenomes, phylogenetic analyses, inference of HR, reconstruction of the evolution of gene families, identification of R-M systems, robustness of the target motif predictor, and robustness of the Count analysis to phylogenetic reconstruction can be found in *SI Methods*.

ACKNOWLEDGMENTS. We thank Vincent Daubin (Université de Lyon) for the suggestion to use Count. We also thank Florent Lassale (University College London) and the anonymous reviewers for critically reviewing the manuscript. This work was supported by European Research Council Grant EVOMOBILOME 281605.

1. Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: The agents of open source evolution. *Nat Rev Microbiol* 3(9):722–732.
2. Vulić M, Dionisio F, Taddei F, Radman M (1997) Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci USA* 94(18):9763–9767.
3. Didelot X, Wilson DJ (2015) ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 11(2):e1004041.
4. Oliveira PH, Touchon M, Rocha EP (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res* 42(16):10618–10631.
5. Mruk I, Kobayashi I (2014) To be or not to be: Regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res* 42(1):70–86.
6. Pingoud A, Wilson GG, Wende W (2014) Type II restriction endonucleases—a historical perspective and more. *Nucleic Acids Res* 42(12):7489–7527.
7. Vasu K, Nagamalleswari E, Nagaraja V (2012) Promiscuous restriction is a cellular defense strategy that confers fitness advantage to bacteria. *Proc Natl Acad Sci USA* 109(20):E1287–E1293.
8. Korona R, Korona B, Levin BR (1993) Sensitivity of naturally occurring coliphages to type I and type II restriction and modification. *J Gen Microbiol* 139(Pt 6):1283–1290.

9. Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 3(9):711–721.
10. Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. *Nat Rev Microbiol* 8(5):317–327.
11. Korona R, Levin BR (1993) Phage-mediated selection for restriction-modification. *Evolution* 47(2):565–575.
12. Kobayashi I (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res* 29(18):3742–3756.
13. Xu Q, Morgan RD, Roberts RJ, Blaser MJ (2000) Identification of type II restriction and modification systems in *Helicobacter pylori* reveals their substantial diversity among strains. *Proc Natl Acad Sci USA* 97(17):9671–9676.
14. Jeltsch A, Pingoud A (1996) Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J Mol Evol* 42(2): 91–96.
15. Seshasayee AS, Singh P, Krishna S (2012) Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Res* 40(15):7066–7073.
16. Kusano K, Naito T, Handa N, Kobayashi I (1995) Restriction-modification systems as genomic parasites in competition for specific sequences. *Proc Natl Acad Sci USA* 92(24):11095–11099.
17. Didelot X, et al. (2011) Recombination and population structure in *Salmonella enterica*. *PLoS Genet* 7(7):e1002191.
18. Doroghazi JR, Buckley DH (2010) Widespread homologous recombination within and between *Streptomyces* species. *ISME J* 4(9):1136–1143.
19. Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. *Science* 315(5811):476–480.
20. Budroni S, et al. (2011) *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc Natl Acad Sci USA* 108(11):4494–4499.
21. Croucher NJ, et al. (2014) Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun* 5:5471.
22. Nandi T, et al. (2015) *Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. *Genome Res* 25(1): 129–141, and erratum (2015) 25(4):608.
23. Roberts GA, et al. (2013) Impact of target site distribution for Type I restriction enzymes on the evolution of methicillin-resistant *Staphylococcus aureus* (MRSA) populations. *Nucleic Acids Res* 41(15):7472–7484.
24. Chan CX, Beiko RG, Ragan MA (2006) Detecting recombination in evolving nucleotide sequences. *BMC Bioinformatics* 7:412.
25. Csurös M (2010) Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26(15):1910–1912.
26. Roberts RJ, Vincze T, Posfai J, Macelis D (2010) REBASE—a database for DNA restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Res* 38(Database issue):D234–D236.
27. Johnston C, Martin B, Fichant G, Polard P, Claverys JP (2014) Bacterial transformation: Distribution, shared mechanisms and divergent control. *Nat Rev Microbiol* 12(3): 181–196.
28. Treangen TJ, Rocha EP (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7(1):e1001284.
29. Cordero OX, Hogeweg P (2009) The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc Natl Acad Sci USA* 106(51):21748–21753.
30. Touchon M, Rocha EP (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* 24(4):969–981.
31. Bobay LM, Rocha EP, Touchon M (2013) The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol* 30(4):737–751.
32. Guglielmini J, Quintais L, Garcillán-Barcia MP, de la Cruz F, Rocha EP (2011) The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet* 7(8):e1002222.
33. Arber W (2000) Genetic variation: Molecular mechanisms and impact on microbial evolution. *FEMS Microbiol Rev* 24(1):1–7.
34. Vasu K, Nagaraja V (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol Mol Biol Rev* 77(1):53–72.
35. Corvaglia AR, et al. (2010) A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical *Staphylococcus aureus* strains. *Proc Natl Acad Sci USA* 107(26):11954–11958.
36. Rocha EP, Cornet E, Michel B (2005) Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet* 1(2):e15.
37. Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E (2010) Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci USA* 107(1): 127–132.
38. Skippington E, Ragan MA (2011) Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol Rev* 35(5):707–735.
39. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* 21(4):599–609.
40. Letunic I, Bork P (2011) Interactive Tree of Life v2: Online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39(Web Server issue):W475–W478.
41. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue):D61–D65.
42. Touchon M, et al. (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 5(1):e1000344.
43. Miele V, Penel S, Duret L (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* 12:116.
44. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–321.
45. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
46. Schliep KP (2011) phangorn: Phylogenetic analysis in R. *Bioinformatics* 27(4):592–593.
47. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
48. 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.
49. Smith JM (1992) Analyzing the mosaic structure of genes. *J Mol Evol* 34(2):126–129.
50. Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172(4):2665–2681.
51. Sawyer S (1989) Statistical tests for detecting gene conversion. *Mol Biol Evol* 6(5): 526–538.
52. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20(2):289–290.
53. Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ (2014) PopGenome: An efficient Swiss army knife for population genomic analyses in R. *Mol Biol Evol* 31(7): 1929–1936.
54. Wolf YI, Makarova KS, Yutin N, Koonin EV (2012) Updated clusters of orthologous genes for Archaea: A complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol Direct* 7:46.
55. Cohen O, Pupko T (2010) Inference and characterization of horizontally transferred gene families using stochastic mapping. *Mol Biol Evol* 27(3):703–713.
56. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7):1575–1584.
57. Katoh K, Standley DM (2014) MAFFT: Iterative refinement and additional methods. *Methods Mol Biol* 1079:131–146.
58. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27(2):221–224.
59. Finn RD, Clements J, Eddy SR (2011) HMMER Web server: Interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server issue):W29–W37.
60. Furuta Y, Kobayashi I (2012) Mobility of DNA sequence recognition domains in DNA methyltransferases suggests epigenetics-driven adaptive evolution. *Mob Genet Elements* 2(6):292–296.
61. Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175(3):1251–1266.