TORRANCE, ELLIS L. Ph.D. Estimating Homologous Recombination Rates Across Bacterial Lineages and Genomes. (2024)
Directed By Dr. Louis-Marie Bobay. 147 pp.

The study of bacteria has become increasingly important to agriculture, healthcare, and industry. However, the evolutionary forces that enable their unparalleled ability to adapt and persist in new environments have yet to be thoroughly determined. Due to their reproduction by binary fission, most evolutionary models consider bacteria to be clonally evolving. However, this ignores the contribution of genetic material from lateral genetic transfer (*i.e.* homologous recombination) which may be more impactful to their species and genomic evolution than mutation alone. Furthermore, tools developed to compare the impact of homologous recombination to mutation in bacteria are often based on strong assumptions and have been used to analyze only few species represented by few genomes. This, combined with a lack of standardization across methodologies and highly inconsistent measurements between studies makes determining the true impact of homologous recombination on bacterial evolution difficult. In this dissertation, I estimate the evolutionary impact of lateral genetic exchange via homologous recombination in 162 bacterial and one archaeal species under a unified framework based on Approximate Bayesian Computation (ABC). Using this data, I was able to map the evolution of recombination rate – as a trait – across many bacterial species represented by thousands of genomes, as well as estimate recombination rate variation on a gene-by-gene basis across bacterial chromosomes. Overall, this study provides insight into the diversity of recombination rates across bacterial species – a key step in understanding how homologous recombination plays a role in bacterial speciation, adaptation, evolution, and population diversity.

ESTIMATING HOMOLOGOUS RECOMBINATION RATES ACROSS BACTERIAL LINEAGES AND GENOMES

by

Ellis L. Torrance

A Dissertation
Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro

2024

Approved by

Dr. Louis-Marie Bobay
Committee Chair

DEDICATION

I dedicate my PhD Dissertation to my mom and my dog. Neither of them fully understands *what* I've written or *why* I've written it, but both have been unconditionally loving and supportive, nevertheless.

APPROVAL PAGE

This dissertation written by Ellis L. Torrance has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. Louis-Marie Bobay

Committee Members

Dr. Dan Schrider

Dr. Kasie Raymann

Dr. Malcolm Schug

March 11, 2024

Date of Acceptance by Committee

November 3, 2023

Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to formally thank my family, friends, mentors, and colleagues for contributions to my research, professional development, and personal life. Specifically, I would like to thank my research advisor Dr. Louis-Marie Bobay for his incredible kindness, patience, and guidance throughout my Ph.D. and the development of this dissertation. Also, thank you to Tracey Schwartz, Tyler Lacy, David Schwartz, and Zach Ostrum for their help in proof-reading this document. I would also like to thank Dr. Louis-Marie Bobay, Corey Burton, Awa Diop, and Matthew Miller for research contributions towards the contents of Chapter 2&3. Lastly, I would like to thank my committee members Dr. Kasie Raymann, Dr. Dan Schrider, and Dr. Malcolm Schug for their valuable advice and support in the development of my research aims and goals.

Financially, this material is based upon work supported by the National Institutes of Health grant R01GM132137 awarded to Dr. Louis-Marie Bobay as well as the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship (DOE CSGF) awarded to me under Award Number DE-SC0021110. This report was prepared as an account of work sponsored by an agency of the United States Government and requires the following disclaimer: Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the

United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I: INTRODUCTION

I.1 An Introduction to DNA Exchange in Bacteria

Bacteria and archaea reproduce clonally via binary fission. However, they are also capable of transferring DNA between individuals through a variety of mechanisms (*e.g.*, conjugation, transformation, and transduction). Genomic material acquired from outside of parent-to-offspring (*i.e.*, *vertical*) inheritance is said to be *laterally* (or, *horizontally*) transferred. For prokaryotes, lateral genetic transfer is thought to contribute considerably to bacterial and archaeal evolution and their unparalleled ability to quickly adapt to new environments. Yet, the pervasiveness of lateral DNA exchange has been evaluated in few species, across few genomes, and often with inconsistent methodology (1–3). As such, it is still unclear whether DNA exchange contributes significantly, in contrast to the contributions of mutations alone, to the evolution and adaptation of bacteria. Understanding the extent of lateral transfer will enable discernment of the fundamental drivers of bacterial evolution and the roles that it plays in their speciation, population structure, adaptation, genome architecture, and gene function, among others. Nevertheless, as I will explore in this review, quantification of the impact of DNA transfer events relative to random mutation on bacterial evolution has been somewhat difficult to ascertain and thus, its relative contribution to genomic evolution across bacterial species and genomes remains largely unexplored (4).

The overarching goal of my dissertation is to: quantify the impact of DNA transferred via homologous recombination relative to mutation across many bacterial species, compare its variation between species, and to map and compare its variation on a gene-by-gene basis across their genomes. By quantifying the extent of genomic exchange across numerous species and

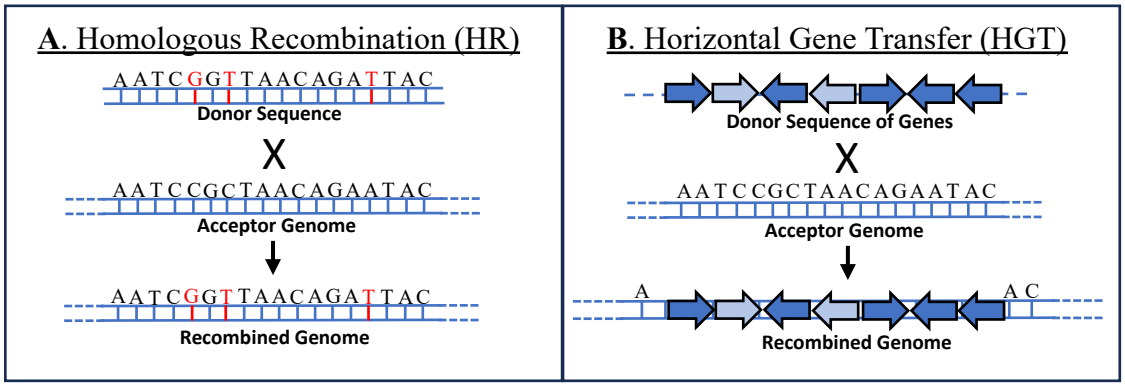
genomes, this study provides some of the data and computational tools necessary to fundamentally unravel some of the ways in which bacteria have evolved and persisted for millions of years on Earth. This introduction chapter aims to provide the reader with a summary of relevant background information and begins with defining the primary differences between homologous recombination and horizontal gene transfer and continues with the mechanisms and modes by which genetic information can be transferred between Prokaryotes. I next provide an overview of the different methods that have been used to quantify the impact of HR and the potential problems inherent therein. I then summarize the results of prior studies that have quantified recombination rate, its variation across some bacteria, and outline evolutionary hypotheses which have been developed to explain why HR occurs in Prokaryotes. Next, I review the genomic organization of bacteria and summarize findings of the few previous studies that have explored intragenomic HR variation in bacteria. Finally, I conclude this introduction by describing the organization of the following chapters and questions addressed by my dissertation research.

I.2 Homologous Recombination (HR) *vs.* Horizontal Gene Transfer (HGT)

The primary focus of this dissertation is to quantify the rates of homologous recombination and their variations across bacterial species and genomes. Thus, horizontal gene transfer will be largely ignored. However, it is worth noting the differences and similarities between the processes and their genomic contributions as some analyses within this study explore the overlap between the two. Both HR and HGT contribute to the genomic plasticity of prokaryotic organisms and describe the processes by which genes and alleles are exchanged between individuals of a species, between species, and across wider phylogenetic relationships. Specifically, homologous recombination refers to the exchange of small (~50-2,000bp (5–7))

highly similar DNA fragments which are thought to occur with higher frequency in relation to sequence identity (4, 8). As such, it is expected that homologous recombination occurs most readily between highly similar genes from highly related individuals and imparts comparatively subtle changes (relative to HGT) in the genome of its recipient (9, 10). Alternatively, the definition of HGT is less nuanced and can be used to describe imports that originate from a variety of organisms which do not necessarily require homology to be incorporated into a genome (7, 11, 12). These events are most often characterized by their contribution to a prokaryotes *accessory* genome which is the collection of genes which are *not* restricted to the species and *not* shared by all its members (13). Although HGT is an important process, this study instead focuses primarily on quantifying genomic change imparted by HR which, alongside mutation, defines the primary mechanism by which change may accrue and be selected upon in the *core genome* (*i.e.,* the set of genes which are present in all members of a species) (14). Figure 1 provides a cartoon illustrating the difference between HR (Figure 1A) and HGT (Figure 1B) where HR is the exchange of a highly similar DNA sequence that generates only allelic differences as evidence of its occurrence. Whereas in HGT (Figure 1B), the exchange results in the insertion of genes in the acceptor genome.

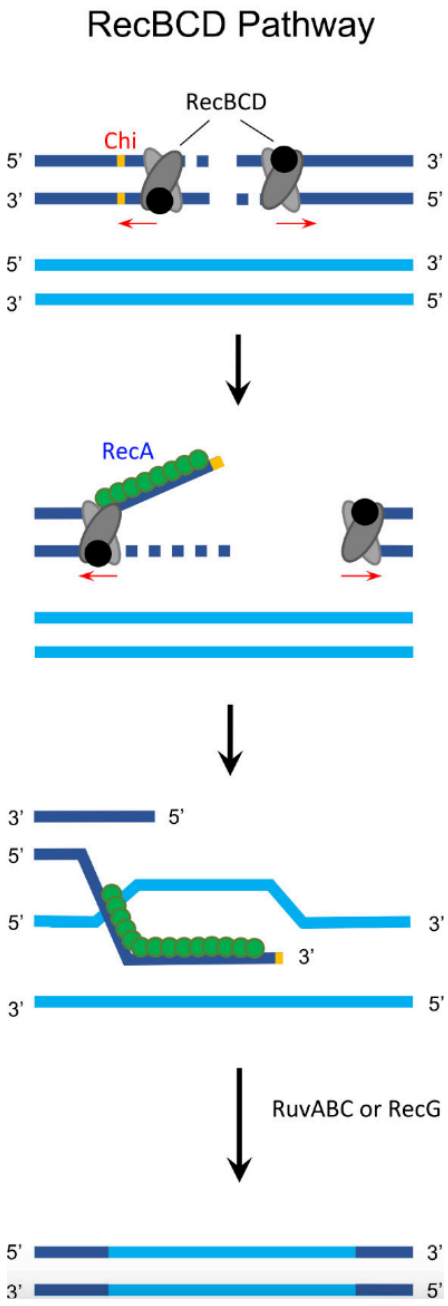
Figure I—1. Cartoon Representing the Difference Between Homologous Recombination (HR) and Horizontal Gene Transfer (HGT).



I.3 Mechanisms of Homologous Recombination

Homologous recombination in Prokaryotes is the unidirectional incorporation of a DNA sequence from a donor molecule of DNA into a highly similar, or homologous region, of an acceptor genome. This process is akin to gene conversion in Eukaryotes and is thought to occur primarily as a housekeeping mechanism by which DNA damage – such as double stranded breaks (DSBs) – occurring in the host's DNA may be repaired (15). Though the proteins that facilitate homologous recombination vary between organisms, the general pathway is essentially the same: enzymes or enzymatic complexes (such as RecBCD or RecFOR) make a portion of the genome accessible to strand-invasion of a donor sequence and subsequent strand-exchange or integration of the donor sequence into the host genome (16). Specifically, HR strand-invasion and strand-exchange is catalyzed by RecA (or a RecA homolog) (Figure 2 (17)). Notably, RecA also has a role in the regulation of the SOS response in bacteria in addition to its role in the repair of DNA lesions (16).

Figure I—2. A cartoon representing the RecBCD homologous recombination pathway (17). This figure is reprinted from "The Impact of Lateral Gene Transfer in Chlamydia" by H. Marti et al., 2022, Frontiers in Cellular and Infection Microbiology, 12, © Frontiers in Cellular and Infection Microbiology (2022).



The most well-characterized recombination pathway is the RecBCD pathway in *E. coli* which repairs DSBs (Figure 2) (18). Here, the RecBCD enzyme complex is catalyzed by a free DNA end characteristic of a DSB and is comprised of DNA helicases (RecB (3' helicase) and RecD (5' helicase)) and a DNA nuclease (RecC) (19). When a blunted end or nonpaired DNA fragment is recognized, the RecBCD complex binds to the DNA end and begins unwinding and digesting the DNA strand. A crossover hotspot instigator, better known as a *Chi* site, is an octameric DNA motif which is recognized by the RecBCD complex; its presence inhibits the DNA degradation activity of RecBCD and then promotes recombination instead (20, 21). During the resolution of a DSB, the interaction between RecC and the 3' Chi site halts nuclease activity in the 5' to 3' direction (22). RecB then assists the binding of RecA protein units to the non-digested 3' DNA tails and RecBCD disassembles. RecA then facilitates the formation of heteroduplex DNA with a homologous sequence via strand invasion (23). DNA polymerase then copies the template single stranded DNA to synthetize the complementary strand. The Holliday junction that results from strand-invasion is resolved via the hexameric protein complex RuvAB and RuvC which unwind and cleave the DNA resulting in either a hybrid or recombinant DNA configuration (19, 24) (Figure 2). Though the RecBCD is the most well-characterized homologous recombination pathway in bacteria, it is only one example of how DNA can be incorporated into the host's genome through homologous recombination. In fact, many other homologous recombination pathways exist or are predicted to exist across prokaryote species, in bacteriophages, and in plasmids (16).

I.4 Types of DNA Transfer in Prokaryotes

DNA has been hypothesized to enter the prokaryotic cell to engage in homologous recombination via various processes and mechanisms (*e.g.*, cell vesicles, and nanotubes) (25,

26). However, this introduction will only briefly describe the best characterized modes of DNA transfer which are: *i*) conjugation, *ii*) transduction, and *iii*) transformation.

I.4.1 Conjugation

Bacterial conjugation, first described in 1946 (27), is the exchange of DNA from a donor to a recipient cell that are in physical contact with each other. This exchange is both site and strand specific and requires a conjugative system consisting of a coupling system (*e.g.* the pilus or pilin) containing an ATPase which facilitates the active process of moving DNA into the acceptor cell via the conjugative secretion system (*e.g.* type-IV secretion system) (28). Conjugative elements are most commonly plasmids or integrative plasmids (and, less frequently, transposons, or chromosomal regions) capable of being exchanged via conjugation (29). Briefly, transfer of these elements is initiated at the *nic* (*i.e.,* origin of transfer (*oriT*)) by the *relaxosome* - a protein complex which binds and cleaves the *nic* site and remains covalently bound to its 5' end (30). Components of the relaxosome then bind with transport machinery to form a complex known as the *transferome*. The transferome then mediates transfer of the conjugative element to the conjugative system which transfers the DNA to the donor cell through the type IV secretion system (28, 30).

I.4.2 Transduction

Transduction is the transfer of bacterial DNA through the action of bacteriophages (*i.e.,* phages) or phage-like elements (31). Phages are viruses of bacteria which can insert their DNA into a host cell to facilitate its replication by the host's machinery. Once inside the host cell the phage DNA may enter a replicative phase known as the lytic cycle where the phage genome is replicated, and its proteins are expressed, to allow for the formation of phage particles and cell lysis (32). Alternatively, the phage may enter a temperate or lysogenic phase in which the viral

DNA integrates into the bacterial chromosome and remains dormant within the host genome until induction. The induction event triggers the excision of the phage DNA from the host genome and its subsequent replication, particle assembly, and release from its host's cell (32). Phages whose genome is integrated into the genome of the host are referred to as *prophages*. Transduction is generally categorized as either *i*) generalized transduction or *ii*) specialized transduction (33, 34).

In current models of generalized transduction, the packaging proteins of phages mistakenly recognize a signaling motif in the bacterial genome as viral which leads to the packaging of bacterial DNA rather than phage DNA (35). By way of this phage particle, the bacterial DNA may then be transferred to another bacterial cell where it may be integrated into the genome of this recipient cell (32). In contrast, specialized transduction is the transfer of a hybridized bacteria-phage DNA fragment resulting from aberrant excision of the phage from the bacterial genome. Here, bacterial genes flanking the phage encoding region of the bacterial genome are excised with the viral genome as a concatemer and can be packaged into a phage particle which may then infect and transfer the bacterial DNA to another bacterial cell(32).

Though transduction generally involves the activity of a complete phage, it can also involve phage-like elements. Prokaryotes may occasionally "domesticate" phages. Prophage domestication entails the cooption of genes that convey a fitness advantage to the bacteria and the deletion or pseudogenization of the remainder of the prophage (36). In some cases, these can be phage genes such as those involved in DNA packaging and transfer (36, 37). An example of prophage domestication are gene transfer agents (GTA) which may arise when bacteria evolutionarily coopt phage genes for DNA packaging and transfer to assist in the transmission of its own genome (36, 37).

I.4.3 Transformation

Some bacteria can gain DNA through the mechanism of transformation. Transformation is the uptake and incorporation of exogenous DNA directly from the environment (38). As opposed to transduction (which requires viral infection) or conjugation (which requires direct contact with another cell), the process of transformation is entirely mediated by the acceptor cell. To participate in the uptake and incorporation of extracellular DNA, a bacteria must first be competent (39). *Competence* is conferred by the expression of a transient set of proteins (*i.e., Com* proteins) that coordinate the uptake of extracellular dsDNA into intracellular ssDNA. This DNA can then be bound by RecA and incorporated into the genome through HR like other DNA imports (40). However, very few species encode *com* genes and thus most species are not believed to be capable of transformation (7). For species which *do* encode *com* genes, competence is generally not constitutive (41). Instead, competence genes are transiently expressed in response to environmental stimuli (*e.g.,* antibiotics or mitomycin-C (42)) or signaling factors produced by other cells, such as the peptides produced by *Streptococcus* in response to population quorum (38, 41, 43).

I.5 Quantifying Homologous Recombination

Most bacteria readily exchange genomic material and gene flow has been found to be a pervasive force shaping the evolution of bacteria (4). However, the relative impact of homologous recombination vs. mutation to sequence diversity remains largely unexplored for most species. Detection of the number of alleles introduced by recombination (r) relative to mutation (m) provides a simple metric by which we may determine the evolutionary impact of homologous recombination on genome evolution relative to mutations (typically expressed as r/m) (44). As an example: if a fragment of DNA is exchanged by homologous recombination

resulting in the substitution of two bases and, separately, two point-mutations occur, then the contribution of recombination relative to mutation in genome diversity is essentially equal (*i.e., r/m*=1). It should be emphasized that this is an *effective* measurement. Meaning that, in cases where a recombination event transfers no polymorphisms, the ratio *r/m* will be zero. As such, a low value of *r/m* (*r/m*<1) is not necessarily an indicator that little recombination has occurred but instead that the *impact* of recombination on sequence diversity was effectively low (45).

Differentiating polymorphisms imparted by mutation vs. homologous recombination is a difficult task because homologous recombination events may leave few, or no polymorphisms, and thus may appear identical to mutational events. Several methods have been proposed to measure HR by leveraging the fact that, over time, the occurrence of HR is expected to leave genomic signatures in a population such as: uneven clusters of polymorphisms, decay in linkage disequilibrium between sites across the genome, and instances of homoplasy (*i.e.,* alleles whose distribution across strains is incompatible with vertical inheritance from a direct common ancestor (phylogenetic incongruencies)). Several programs have been developed to exploit these patterns in effort to determine the relative impact of homologous recombination to bacterial evolution. A select few which have been widely adopted or recently developed are outlined below:

i) *ClonalFrameML*: Perhaps the most popular modern tool for assessing recombination rate in bacterial populations is ClonalFrameML (CFML) (45) which is an improvement on the tool ClonalFrame (46). CFML takes the input of a maximum likelihood phylogeny and the multiple sequence alignment (MSA) used to build the phylogeny. From this input, CFML attempts to identify the clonal portion of the alignment and then constructs a maximum likelihood

phylogeny of the clonal genome (*i.e*., the inferred phylogeny in the absence of the recombined regions). The software then attempts to identify recombination breakpoints in the MSA by detecting regions of homoplasy and regions which contain a high number of polymorphisms which are not consistent with the clonal genealogy.

ii) *Gubbins*: Gubbins (47) requires only a MSA as input from the user. As a first step, the software defines all SNPs (single nucleotide polymorphisms) which are consistent with a constant per-site mutation rate and all clusters of SNPs which are not. It infers the clonal portions of the MSA, then removes all SNP clusters and constructs a clonal phylogeny from the remaining aligned sites. Gubbins then quantifies recombination rates from all the SNPs assumed to derive from imports to the phylogeny as homologous recombination events.

iii) *fastGEAR*: The recombination detection program fastGEAR (48) takes an MSA as input and then groups all taxa within into lineages which it defines as being identical over 50% of aligned sites. Clusters of polymorphisms are assumed to be imports. The software then attempts to locate the origin of the import. If the origin is found to be in another lineage than the one being analyzed, it is considered a recombination.

iv) *mCorr*: The tool mCorr (49) was the first that claimed to be capable of detecting recombination events in raw non-assembled reads and metagenomic datasets. To do this, the software employs an approach based on linkage-disequilibrium where it scans each pair of homologous sequences in the dataset and computes a correlation profile based on substitution probability relative to the distances

between sites. The data is then compared to simulated populations evolved under a coalescent model with different recombination rates and fragment lengths.

v) *Rhometa*: Rhometa (50) is another tool which aims to detect homologous recombination in metagenomic shotgun sequencing reads. This tool partially implements models from tools developed for recombination analysis in diploid organisms such as LDhat (51) and pyrho (52) and methodologically is very similar to mCorr with the exception that it is capable of handling non-coding as well as coding sequences and that the recombination tract length in Rhometa simulations is fixed.

Each of the homologous recombination detection programs outlined above usually utilizes a *single* genomic signal to determine whether homologous recombination has occurred: either the decay of linkage disequilibrium, the density and distribution of polymorphism, or the patterns of homoplasies. The assumption that HR results in clusters of polymorphisms as in (47, 48) was initially motivated by the scarcity of sequenced genomes. This approach attempts to detect imports from more distant strains of species however, it lacks in accuracy when detecting transfers between more related genomes (*i.e.*, transfers which are more prone to result in ungrouped polymorphisms). Additionally, homoplasies may arise because of parallel or convergent evolution which are ignored in (45) and may result in erroneous inference of recombination in some cases (8). Tools that analyze the decay of linkage disequilibrium (LD) as evidence of recombination such as (49, 50) ignore that, especially when using a small or biased genomic population such as in a metagenomic sample, one may expect to see dramatic difference in linkage decay between sampled populations across different genomic regions, different

populations, and different growth phases (53). Thus, the complexity of parameters at play compromises the inference of accurate rates of recombination from LD patterns alone.

Aside from the use of only one metric to infer recombination, these tools generally have other problematic assumptions. For instance, it is commonly assumed that substitution introduced by mutation occurs at a uniform rate across the genome. This assumption ignores that genomic sites may evolve at different rates and are under different selective pressures and, as such, not all clusters of polymorphic sites result from homologous recombination alone (54–56). Further, many tools rely on the inference of a "clonal" or non-recombining portion of the genome which assumes there is a region of the genome which does not recombine (or recombines very infrequently) (45, 47). In fact, the inference of the wrong clonal frame can have drastic consequences for these methods by causing recombined regions to be entirely ignored. Lastly, one inherent issue with using the detection of signatures of HR directly (*e.g.,* homoplasy, polymorphism clusters, LD) to infer homologous recombination events is that these signals may be saturated in rapidly recombining loci. Therefore, inference of recombination based directly on the detection of HR signatures may cause the recombination rate to be underestimated in regions that frequently recombine.

I.6 Homologous Recombination Rates Inferred in Previous Studies

Though the methods available for recombination identification and rate estimation carry several biases and assumptions it has been used to assess recombination in a variety of prokaryotic organisms. However, most studies have focused on a single or very few species. Nevertheless, there are two studies where recombination rate was evaluated in many prokaryotic species under a unified methodology (1, 2). The first is by Vos & Didelot (2009) (1) who used ClonalFrame (46) (the predecessor of ClonalFrameML (45)) to determine recombination rates in

multi locus sequence typing (MLST) datasets (typically seven gene markers per species) to predict recombination rates for 46 bacterial species and 2 archaea (see Table 1 in (1) for data regarding the species and number of loci and sequence types analyzed for each). This study found *r/m* across species to range from very low (*Leptospira interrogans r/m*=0.02) to very high (*Flavobacterium psychrophilum r/m*=63) with 56% of species having *r/m*>1. The more recent study by González-Torres et al. (2019) (2) analyzed whole bacterial genomes to infer recombination rates in 54 species by using a composite of methods including ClonalFrameML. In this study, the authors considered a recombined fragment to be a region denoted as being recombined by three-out-of-five detection programs. This method was used to analyze the whole genome of 54 bacterial and archaeal species but only processed a total of 338 genomes (less than seven genomes per species, on average). Unfortunately — rather than denoting *r/m* as the number of polymorphic sites imported by recombination relative to those imported by mutation as in other studies— the authors stated that their metric *r/m* instead represents a ratio of probabilities that a given site was altered by recombination relative to mutation — making these values difficult to compare to other studies. The authors found the highest recombining species to be *Burkholderia pseudomallei* (*r/m*=973.8) and the lowest to be a group of genomes from the genus *Frankia* (*r/m* = 0.0). Interestingly, the authors found that recombination was low in endosymbionts. The authors also evaluated the number of recombination events relative to gene function and found genes associated with cellular defense to have high incidences of recombination and that genes associated with conserved house-keeping functionalities tended to recombine less.

Homologous recombination rate has been estimated in a multitude of previous studies for more than ninety prokaryotic species (57). Though, with exception to the two outlined in the

paragraph above, each study has separately estimated recombination rates for only one or a few species. Moreover, these studies were conducted using disparate methodologies, genomic sites, species definitions, alternate definitions for what constitutes a recombination event, as well as different metrics for defining recombination rate itself (3, 57). An example of this is shown in Table 1 of (3) where the authors describe the wide variety of methodologies and metrics from prior studies of recombination rate variation in *Escherichia coli.* Here, *E. coli* has alternately been described as panmictic (or, having such a high recombination rate that the true value is not readily discernable) to more recent valuations of $r/m \leq 1$. These vast differences in recombination rate estimates across a single species are indicative of: *i*) the need for a singular robust method to define recombination rate, *ii*) the need for a singular definition for recombination rate itself, and *iii*) a study of multiple species performed under a unified framework to determine what patterns of recombination rate variation exist amongst bacterial species.

I.7 Homologous Recombination as an Evolutionary Process

DNA strands incorporated by HR into an acceptor sequence may be *effectively silent* or, impart no genetic change to the genetic makeup of the acceptor strand. Alternatively, polymorphisms may be exchanged and contribute to the divergence of that sequence and be the target of selective pressures. The duality of homologous recombination is that it may be considered a force which favors the cohesiveness of a genomic population *or* alternatively, HR may increase diversity by conveying novel alleles.

The question of why homologous recombination occurs in bacteria is akin to the question of why sex occurs in eukaryotes as they appear to fulfill the same general function: to increase genetic divergence and unlink alleles on which selection may then act more efficiently. As such, some of the hypotheses for why sex exists in Eukaryotes may be applied to bacteria as outlined

in Vos (2009) (15). Here, the author outlines five theories of the benefit of sexual reproduction as they pertain to bacterial recombination that I've paraphrased as follows:

1) In the "Tangled Bank Model" bacterial populations which have reached peak fitness in each biome will begin to deplete that environment of resources. To enhance population growth and escape competition of related individuals it benefits the bacteria to exchange alleles which allow the population to escape the resource-depleted environment and colonize new niches.

2) The "Sign epistasis model" posits that the landscape of fitness is not a single idyllic peak, but instead many hills and valleys and that genomic mutations come with both costs and benefits that may force bacteria into suboptimal niches. Homologous recombination may then allow alleles to be transferred from a population occupying one peak to a population occupying another — allowing peak or niche traversal.

3) The "Lottery Model" proposes that, instead of environmental landscapes being static — as in the first two hypotheses — that the fitness landscape is constantly changing and that it benefits the population to maintain as much diversity as possible to better their chances at exploitation of any peaks that should arise in the fitness landscape.

4) The "Red Queen Hypothesis" model considers not just the changing of the environment as in the previous two models, but also the changes in all other organisms present. Here, bacteria also must co-evolve and adapt to other species in their environment: their predators, their prey, and their competitors. As such, the primary role of homologous recombination is to speed up the acquisition of novel alleles which could help them adapt to the constantly evolving organisms present in their environment.

5) Finally, the "Muller's Ratchet" hypothesis posits that deleterious mutations, which are much more frequent than beneficial ones, accumulate and reach fixation in non-recombining populations. In this scenario, homologous recombination allows the purge of deleterious alleles that would otherwise accumulate in the genome.

There is likely no one model of evolution listed here that accurately encompasses all the reasons why homologous recombination occurs and some of these models are not incompatible with one another. However, one might expect to see recombination rate amongst species vary in accordance with some of these postulates. For instance, species with reduced interaction with the greater population (*e.g.*, endosymbionts) may be more predisposed to accumulating deleterious mutations through Muller's Ratchet. Previous studies in *Wolbachia* (a maternally-inherited bacterial symbiont in insects) have found that they are deficient in recombination machinery which may be a determining factor in their genome reduction (58). This is because, without recombination to counter Muller's Ratchet, genes are more likely to acquire deleterious mutations over time, lose their functionality, and eventually be altogether lost from the genome (58). *Chlamydia*, which are obligate intracellular bacteria that may occasionally be transmitted horizontally between hosts (which can be entirely different species such as *Chlamydia psittaci* transmission between birds and humans (59)), have been estimated by a previous study to have a relatively high incidence of recombination but a low overall contribution of polymorphisms by recombination relative to mutation (*Chlamydia trachomatis*, 12 genomes) (60). This indicates that, even though they are also endosymbionts, cell transfer between hosts or coinfection may allow homologous recombination between nonidentical individuals to occur and counteract Muller's Ratchet. However, in the same study, genes associated with interaction with the immune system of their host and pathogenicity were found to have higher incidences of

recombination relative to other genes indicating there is also a role of HR in adapting to and exploiting a changing host environment (60).

Some species are capable of thriving in a variety of environments, such as *E.coli* which can exist commensally or pathogenically in the gastrointestinal tract of humans, but also in water and soil. Due to their persistence in a variety of environments, *E. coli* may have much larger population sizes and encounter individuals with which they can share genomic content with much more often and therefore are capable of recombining more (61, 62). Here, homologous recombination may be comparatively elevated and play a role in their exemplary ability to adapt quickly to new environments (63). Additionally, variation in recombination has been found in previous studies (using ClonalFrameML) to exist across phylogroups indicating variation in recombination exists not only at the species level, but also at the population level (6, 61). However, the extent at which variation can be found across subpopulations in other species or even whether variation observed is simply due to lack of precision in methodology is unknown.

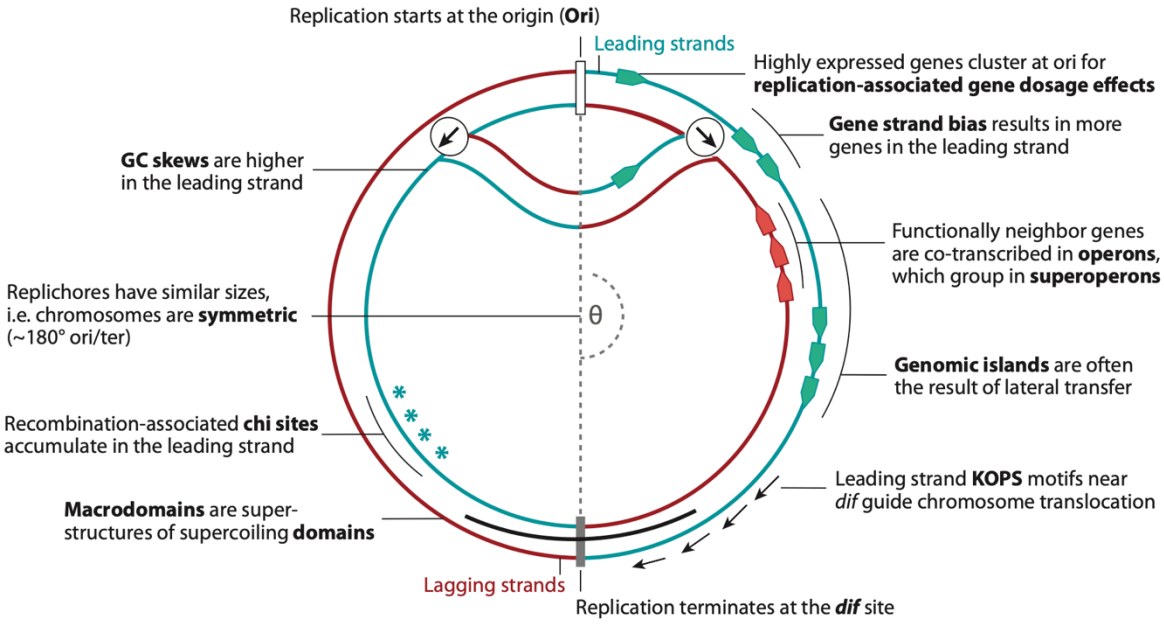
I.8 Homologous Recombination across the Genome

The genomes of prokaryotes are highly organized entities, which reflects the constraints of major genetic and cellular mechanisms such as replication, transcription, and cell division. Bacteria and archaea may have one or multiple chromosomes which can be circular or—more rarely—linear (as in *Streptomyces* or *Borrelia* species), or a combination of both (64). However, the overall process of initiating genomic replication is thought to be essentially the same (65–67). Genomic replication in bacterial chromosomes begins at the origin(s) of replication (Ori or *oriC*) where double-stranded DNA is separated by initiator proteins and replication machinery is recruited (68). Helicase enzymes then unwind the DNA and single-stranded binding proteins bind to the ssDNA to promote stability and prevent degradation. Strand synthesis proceeds

bidirectionally (i.e., in two replichores) outwards towards opposing replication forks (Figure 3, (69)). Within each replichore, the leading strand is continuously synthesized in the 5'-3' direction with the same orientation as the replication fork, whereas the synthesis of the lagging strand is not continuous and resultant Okazaki fragments must be joined together with a DNA ligase. In circular chromosomes, the two replication forks meet at the terminus of replication (Ter) which is approximately equidistant from the Ori and occurs in temporal conjunction with septum formation of the cell during cellular replication (70). Though replication fork resolution is less studied in linear chromosomes, literature suggests the presence of hairpin loops and/or telomeres at the terminus, and that replication resolution is likely somewhat like that of Eukaryotes and may, in some cases, confer a fitness advantage (64, 65, 71).

Transcription co-occurs with genome replication. As such, genes with higher transcription rates—such as those required for transcription and translation—are often located near Ori, providing a "replication-associated gene dosage effect" (Figure 3, (69)) (67, 72). As replication and transcription co-occur, collision between the replication machinery and transcriptional machinery may take place and can result in replication-fork collapse. To minimize collisions, a gene strand bias has evolved: most genes (55-70%, depending on the species) are co-oriented with the leading strand so that transcription is co-oriented with replication fork movement (Figure 3, (69)) (67, 70, 73). Replication-fork collapses are resolved by the formation of dsDNA breaks which can be repaired by the RecBCD complex using homologous recombination as described above (70).

Figure I—3. A cartoon representing chromosome replication in prokaryotes and some of the structural and functional organizations of the Prokaryotic genome (69)**. This figure is reprinted from "The Organization of the Bacterial Genome" by E. Rocha, 2008, Annual Review of Genetics, 42, p. 211-233. © Annual Reviews of Genetics (2008).**



Additional compositional and organizational properties have been observed in Prokaryotic chromosomes. For instance, guanine (G) is overrepresented in the leading strand and cytosine (C) is proportionally overrepresented in the lagging strand with a sharp transition in bias occurring at-the Ori and Ter which allows the approximate location of the Ori and Ter to be determined by plotting cumulative GC skew for many species (70, 74). This pattern is due to the discontinuous replication of the lagging strand, which remains single-stranded for longer periods of time and is therefore more likely to accumulate certain types of mutations (*e.g.,* cytosine deamination occurs more readily on single-stranded DNA). Prokaryotic genomes tend to range widely in size between 50kb to 13Mb and are much more compact relative to Eukaryotes with

gene densities across the genome of near 87% (67, 75). Additionally, the number of total coding sequences typically varies in proportion to the length of the genome (67, 76). Structurally, genes tend to be clustered into operons (69). Operons are groups of genes that are co-transcribed under the action of a shared promoter(s) and are often functionally related units of a metabolic pathway (69, 77). The organization of genes into operons ensures that all units of a given pathway are expressed nearly simultaneously in a dosage-dependent manner (78). This is thought to be particularly advantageous for the survival of genes during HGT because the transfer of an individual gene that is part of a pathway is unlikely to be retained by selection in the recipient genome, whereas the transfer of a cluster of genes encoding a fully functional metabolic pathway is much more likely to provide a fitness advantage to the recipient genome (69, 77).

The number of genes associated with housekeeping functions such as replication and translation tend to vary little across species, except for some obligate intracellular prokaryotes which may be missing some subsets of these genes and rely on the host to fulfill their roles (76). In larger genomes—which are expected to engage more frequently in HGT—most of the variability in gene content and gene diversity is observed in genes related to mobility, cellular transport, secretion, and those attributed to accessory metabolic functions and pathways (79). Thus, species with larger genomes and wider gene repertoires are thought to be generalists who are more readily capable of niche traversal (67). Further, compared to their more sexually isolated and reduced-genome counterparts (*e.g.*, obligate intracellular bacteria), population size is expected to be larger and thus selection is expected to be more efficient (75, 76).

Prior studies have explored variation of homologous recombination rate and its signatures in the context of prokaryotic genome organization. However, these studies are few and still contain the methodological problems addressed previously. As mentioned above, genes with

essential roles in replication and transcription are often found near the origin of replication (72). Studies have found genes in this region to display elevated signatures of HR in *Staphylococcus aureus* (80, 81) but similar patterns have not been found in other species (80). One might expect signatures of recombination to be lower in highly conserved genes as purifying selection is expected to dominate, and so polymorphisms should be less likely to be tolerated. Thus, the trend observed in *S. aureus* is somewhat unexpected and might be the consequence of the low number of strains analyzed or inaccuracies in methods for assessing *r/m*. Alternatively, less-conserved genes associated with cellular defense, pathogenicity, and virulence may be under positive selection as that would allow quicker adaptation of the organisms to a changing and competitive environment (82). For this reason, it might be expected that signatures of recombination would be higher in genes associated with these functions. In fact, this was observed in González-Torres et al (2019) (2). These genes have also been associated with *hotspots* of recombination (*i.e.,* a region of the genome with significantly elevated recombination rate relative to the genomic average) in Oliveira et al. (2017) (13). Hotspots of recombination have also been associated with the acquisition of horizontally transferred accessory genes and the regions flanking accessory genes have been found to have higher rates of recombination in some species (13, 83, 84). Overall, these genomic patterns have been evaluated in few species and using varying methodologies and it is therefore difficult to determine whether observed patterns are robust and if they are generalizable across bacterial species (57).

I.9 Overview of Dissertation Questions and Chapter Organization

As addressed in this introduction chapter, current methods for determining HR rate make many assumptions which may result in inaccurate estimation of rate variation across species. Further, few studies have examined many bacterial species using a large amount of genomic data

under a single methodological framework and definition for recombination rate. Thus, the relative contribution of HR to mutation in the evolution and vast diversity observed across bacterial species remains largely unexplored. In the following chapters, I; *i*) leverage a new approach for inferring homologous recombination rates and explore its potential biases, inconsistencies, and compare our estimates to those of previous studies, *ii*) explore the variation in HR rate observed across the species in the context of the evolution of recombination rate itself and whether variation is observed in correlation with bacterial pathogenicity and ecotype, and *iii*) compare HR rates across the genome and within the functional context of genomic organization to determine whether genomic trends in recombination exist across the species genome and whether those trends persist across species.

The next chapter is formatted for publication and thus the Aims 1 and 2 of my proposed research aims are both included in this single but *dense* chapter. The objective of Aim 1 was to *Accurately Estimate Homologous Recombination Rates in Bacteria* and included the curation of the original dataset – which was aided in part by Matthew Miller and Dr. Awa Diop – as well as the assessment of bias in our methodology. Contributions to analyzing variations in HR rate across intra-species populations in *E. coli* were made by Corey Burton during the completion of their master's degree. The manuscript also contains a detailed description of the computational pipeline co-developed by my mentor and committee chair Dr. Louis-Marie Bobay, which was leveraged by me to estimate recombination rate across a large variety of bacterial species and genomes and assess potential biases and inaccuracies in the methodology. The objective of Aim 2 of my proposal was to *Determine how Homologous Recombination Rates Have Evolved in Bacteria* which specifically referred to: *i*) mapping the rate of recombination across a multi-species phylogeny to determine whether recombination rates varied with any detectable patterns

across species and *ii*) identifying potential correlations between recombination rate and several genomic, metabolic, and environmental traits.

The last chapter before the conclusion presents the work conducted on Aim 3 whose objective was to *Determine Bacterial Recombination Rate Variation along the Chromosome* using a modified version of the computational pipeline from Aim 1. This enabled me to quantify the rates of homologous recombination across the bacterial chromosome of many species. Mapping gene-by-gene variation in HR allowed me to begin to determine how HR varied in the context of genome organization, whether any trends in HR were found in genes proximal to accessory genes, if hotspots of recombination were present in the genomes, and how HR varied by gene function. Lastly, by mapping the chromosomal variation of homologous recombination rate across the genomes of different species, I was able to determine whether any patterns of recombination persisted across species.

Overall, the aim of my dissertation is to characterize the impact of homologous recombination, relative to mutation alone, in the evolution of bacteria. By quantifying the rate of recombination across multiple species and genomes, I provide the first steps in discerning the impact of recombination on bacterial evolution and the roles that it plays in their population structure, genomic structure, and adaptation.

CHAPTER II: EVOLUTION OF HOMOLOGOUS RECOMBINATION RATES ACROSS BACTERIA

Ellis L. Torrance[1], Corey Burton[1], Awa Diop[2], Louis-Marie Bobay[1,2]

[1] Dept. of Biology, University of North Carolina at Greensboro, Greensboro, NC 27412, USA.

[2] Department of Biological Sciences, North Carolina State University, Raleigh, NC 27695, USA.

DISCLAIMER: This chapter has been submitted and accepted to the journal "Proceedings of the National Academy of Science" (PNAS) on 2/19/2024 and is currently under editor revisions. Please seek revised work.

II.1 Abstract

Bacteria are non-sexual organisms but are capable of exchanging DNA at diverse degrees through homologous recombination. Intriguingly, the rates of recombination vary immensely across lineages where some species have been described as purely clonal and others as "quasi-sexual". However, estimating recombination rates has proven a difficult endeavor and estimates often vary substantially across studies. It is unclear whether these variations reflect natural variations across populations or are due to differences in methodologies. Consequently, the impact of recombination on bacterial evolution has not been extensively evaluated and the evolution of recombination rate—as a trait—remains to be accurately described. Here, we developed a new approach based on Approximate Bayesian Computation (ABC) that integrates multiple signals of recombination to estimate recombination rates. We inferred the rate of recombination of 162 bacterial species and one archaeon and tested the robustness of our approach. Our results confirm that recombination rates vary drastically across bacteria, however, we found that recombination rate—as a trait—is rather conserved amongst several lineages.

Although some traits are thought to be associated with recombination rate (*e.g.*, GC-content), we found no clear association between genomic or phenotypic traits and recombination rate. Overall, our results provide an overview of recombination rate, its evolution, and its impact on bacterial evolution.

II.2 Statement of Significance

Homologous recombination is a fundamental mechanism driving the evolution of bacteria. Recombination rates have been found to vary tremendously across species but quantifying the rates of recombination in bacteria is a difficult task. Moreover, comparing estimates across studies is difficult due to the diversity of methodologies and datasets. Using a new methodological framework based on Approximate Bayesian Computation, we estimated the impact of recombination relative to mutation on bacterial genomic evolution for 162 species. We found that rates of homologous recombination do vary greatly across lineages but tend to be underestimated by prior studies. We further reconstructed the evolution of recombination rate across bacterial lineages and found that this trait is rather conserved.

II.3 Introduction

Bacteria adapt rapidly to changing environments and their capacity to survive in new environments is owed in part to their propensity for genetic exchange (3, 85). Specifically, genetic exchange in bacteria is defined as either resulting from horizontal gene transfer or homologous recombination. Horizontal gene transfer leads to the gain of a new DNA sequence that was not present in the recipient genome and these processes do not necessarily require close relatedness between the donor and recipient DNA sequences. Akin to allelic gene conversion events in eukaryotes, homologous recombination is characterized by the unidirectional genetic exchange of short homologous DNA sequences. As such, homologous recombination events lead

to the replacement of alleles within a homologous sequence, thereby modifying an existing sequence. Due to these characteristics, homologous recombination events may leave few—or no—genomic signatures of having occurred. This lack of clear genetic evidence makes distinguishing homologous recombination events from mutation events a challenging endeavor (1, 3, 4, 85).

Despite the central role of homologous recombination as a basic biological mechanism, its impact on bacterial evolution remains poorly understood. In addition, whether recombination rate is a fast-evolving or slow-evolving trait remains to be determined. Several analyses have quantified the rate of homologous recombination (recombination rate) across bacteria and studies have revealed staggering variations of recombination rate estimates across species (1, 2). Whether a species is evolving primarily clonally or sexually has key implications for understanding their biology and designing tools and models applicable to bacteria (86). However, estimating recombination rate has proven a difficult task and studies have often reported inconsistent results for the same species (3). For instance, *Escherichia coli* has been found to vary from a clonal species (no recombination) to panmictic (highly recombining) across studies (3). The lack of consistent estimates can be attributed to *i*) the diversity of methods developed to assess recombination rates, *ii*) the different assumptions made to model bacterial evolution, and *iii*) the inconsistency of datasets used to generate these estimates (1–3, 87).

The various methodologies that have been developed to estimate recombination rates in bacteria rely on the analyses of different signatures of recombination (45, 47, 48, 88–93). Some approaches rely on the detection of direct evidence of homologous recombination through the detection of homoplasies (*i.e.,* alleles that are incongruent with the vertical evolution of the species), while others infer recombination rates through the signals of linkage disequilibrium

(*i.e.,* the decay of linkage between alleles across genomic distances) or on the distribution of polymorphisms (87). Quantifying the direct signal of recombination is challenging due to fact that recombination can occur between identical—or nearly identical—sequences. In addition, the same sequence can be affected by multiple recombination events throughout its evolution, which can lead to a decayed signal of recombination. For these reasons, many recombination events can remain undetected and recombination rate can be difficult to quantify. Moreover, virtually all approaches aim at quantifying an *effective* rate of recombination *r/m*, which is defined as the number of times alleles have been exchanged by recombination (*r*) divided by the number of alleles that have been introduced by mutation (*m*). These approaches are quantifying an effective rate of recombination because many sites may have recombined without resulting in any transfer of alleles (i.e., the donor sequence was identical to the recipient sequence), and these recombination events are not effectively impacting genome evolution. However, other types of genetic events may lead to underestimation of recombination rates: *i*) alleles present in a single genome will not impact patterns of homoplasies or linkage disequilibrium; even though these alleles might have been gained and lost multiple times through recombination. *ii*) alleles can be exchanged by recombination but need not generate patterns of homoplasy or impact linkage disequilibrium.

Here, we developed a novel approach to quantify the rate of homologous recombination across a large set of bacterial species. Using Approximate Bayesian Computation (ABC) and by integrating multiple signatures of recombination, we inferred recombination rates in the core genomes of 162 species. By using population simulation to model homologous recombination we are not limited by direct extrapolation of recombination signal, thus allowing us to predict recombination rates more accurately than other methods when recombination rates are

particularly high and when the signal of recombination is more complex. In addition, our approach allows us to not only estimate the effective rate of recombination relative to mutation, but also to estimate the total number of nucleotides that have been exchanged by recombination relative to point mutations. We found that recombination rates vary widely across species, and on average, recombination promotes the exchange of alleles six times more frequently relative to mutation and is more impactful to species evolution than mutation alone for >80% of the species included in this analysis. Importantly, we reconstructed the evolution of recombination rate across bacteria using a phylogenomic framework, and we found evidence that recombination rates are phylogenetically conserved within several genera.

II.4 Results

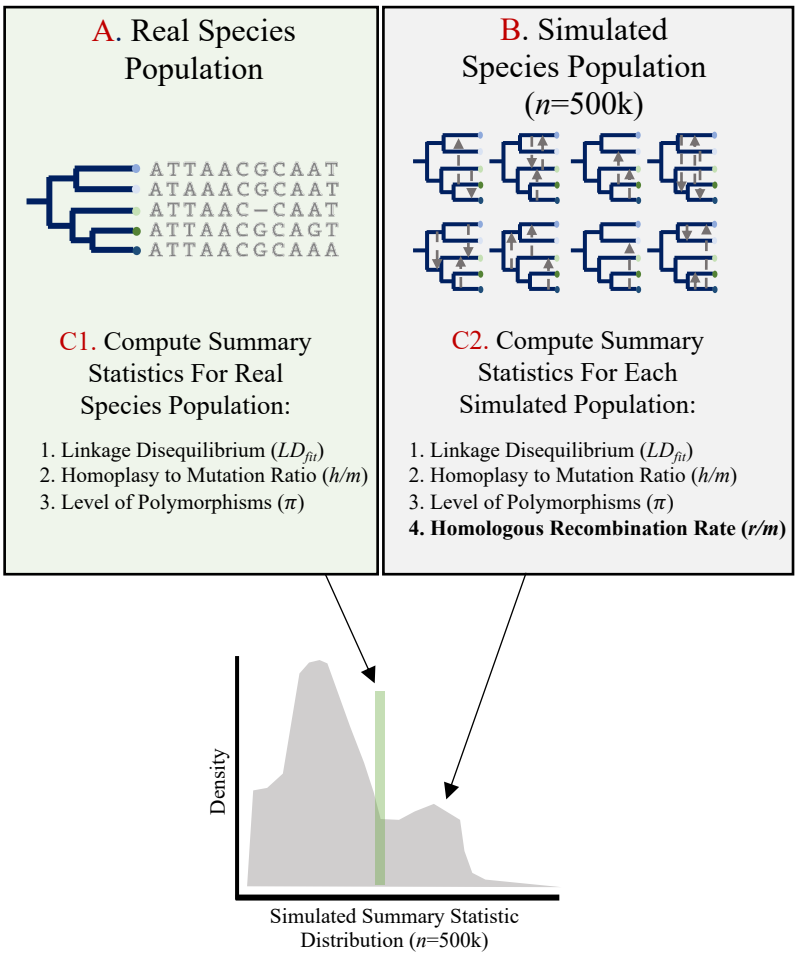
II.4.1 Description of the ABC approach

Most methods designed to infer homologous recombination rates rely either on incongruencies in the phylogenetic signal (i.e., homoplasies) *or* on the patterns of linkage disequilibrium (LD) (45, 47, 48, 90, 94). This is because these metrics are expected to be strongly impacted by recombination and represent a signal that can be directly quantified. However, very few methods have integrated multiple metrics and these approaches typically rely on the direct inference of effective recombination rates from these signatures of recombination. To estimate recombination rates across bacteria we developed an approach based on Approximate Bayesian Computation (ABC) that integrates patterns of homoplasies, polymorphisms, and linkage disequilibrium. Our conceptual framework allows us to improve estimates of recombination rate by accounting for recombination events that would be typically overlooked by most methods. In contrast to direct approaches, our methodology consists of simulating a wide array of genomes with various rates of recombination and a large set of

parameters. Rates of recombination are then inferred by identifying the simulated genomes that present signatures of recombination that most resemble those observed in the real dataset. In brief, our approach consists in simulating the evolution of a core genome ~500,000 times with a wide array of recombination rates and various levels of polymorphisms (see Methods) (Fig.1, B). For each simulated genome population, we quantify: *i)* the ratio of homoplasic alleles relative to non-homoplasic alleles (*h/m*), *ii)* the decline of allelic linkage across short genomic distances (LD_{fit}), and *iii)* the levels of polymorphisms (π) (Fig. 1, C2). We then use ABC to infer which of the 500,000 simulation sets of these three summary statistics most resemble the statistics of the real population. For each species, the posterior distribution generated with ABC represents the 0.01% of simulations which present the most similar signatures of recombination based on *h/m,* LD_{fit} and π (Fig. 1, D). The recombination rate *r/m* is then estimated as the average of the posterior population which best recapitulated the summary statistics of the species genomic sample.

Species were included in this analysis if: *i)* the species had 15 or more non-identical assembled genomes available from GenBank, and *ii)* the species retained at least 15 genomes after redefinition of species borders by ANI and inferred signal of gene flow between members computed by *ConSpeciFix* as in (4, 95). After applying these criteria, a final dataset composed of 162 bacterial species and one archaeon was used for estimating recombination rates with our ABC approach. Before further analyzing these data, we first conducted several tests to assess the accuracy of our estimates.

Figure II—1. Description of the method used to infer rates of homologous recombination (*recABC*) for 162 bacteria and one archaeal species in this study. A) For each species, the core genome is aligned and concatenated. A phylogeny is built from the alignment of the core genome. B) Using the nucleotide length and GC content of the core genome alignment from the real species, a single ancestral genome randomly generated to initiate the ancestral genome of each simulation. This ancestral genome is then evolved in a forward-in-time simulation following the phylogenetic topology of the real species under varied recombination rates (*Rho*) and recombination tract lengths (*delta*) using *CoreSimul* (96). The corresponding effective recombination rate *r/m* is also computed during the simulations. A total of 500,000 simulations are generated for each species. C1) Three summary statistics are computed: i) the decay of genetic linkage relative to genomic distance (LD_{fit}), ii) the ratio of homoplasy to mutation (*h/m*), and iii) the average nucleotide diversity (*Pi*) computed for the core genome alignment of the real species. C2) The same summary statistics (LD_{fit}, *h/m, pi*) are calculated for each of the 500k simulations. D) Approximate Bayesian Computation (ABC) is used to compare the summary statistics from the real species data to the distribution of the same statistics generated by simulation under known recombination rates (the prior distribution in grey). The simulations with statistics which most closely match the summary statistics from the real species are selected using ABC (the posterior distribution in green) with a tolerance threshold of 0.01% (*n*=50). The average rate of recombination of the posterior distribution is then used as an estimate of recombination rate for each species.



D. **Approximate Bayesian Computation** (ABC) used to compare real and simulated summary statistics (1-3) to infer the most probable rate of recombination (r/m) for the real species population.

II.4.2 **Assessment of the ABC approach**

To evaluate the performance of our method, we first tested the ability of our ABC approach to predict the recombination rate of simulated datasets. To do this, we used the 500,000 simulations of core genome alignments generated under known recombination rates for each species and set aside 10% of the simulated data from each species to use as a test population. From each test population, the summary statistics from 1% of the simulations (n=5,000) were used to infer the rate of recombination using ABC against a population of 90% of the remaining simulations (n=450,000) with a tolerance threshold of 0.01%. Results show that our approach accurately predicts the rate of recombination r/m for each simulated dataset as the average of the combined posterior for each of the 5,000 subpopulations converged to the known average r/m for the 5,000 simulated populations (r^2=1, $P<10^{-15}$) (Supp. Table 1I, Supp. Fig. 1). However, the inference of r/m was visibly less precise for several simulated datasets (n=19, Supp. Table 1I). Interestingly, the species with less accurate predictions shared several characteristics: their core genomes presented very few polymorphisms ($\pi<0.005$), few homoplasies were inferred ($h/m<0.38$), and linkage was high ($LD_{fit}>0.15$) (Supp. Table 1I). These results indicate that—as expected—the precision of our method is lower for species presenting few polymorphisms and low signal of recombination. Nevertheless, our approach did estimate an accurate range of recombination rates for these simulations, which was usually low (Supp. Table 1J).

To further assess the accuracy of our approach, we systematically verified that each of the three summary statistics—h/m, LD_{fit} and π—was precisely inferred for each of the 162 bacterial species, and one species of archaea. For each summary statistics, the predicted value was plotted against the value measured in the real datasets. The estimated values of the metrics were typically very close to their real values, except for several outliers. We observed a strong

correlation for each metric close to $y=x$ (Spearman's *Rho*: h/m=0.99, LD_{fit}=0.78, and π=0.87) (Supp. Fig. 2, A-C), which increased after removal of outliers (Spearman's rho without outliers: h/m=0.99, LD_{fit}=0.89, and π=0.97) (Supp. Fig. 2, D-F). Species were considered outliers when the Cook's Distance difference in real vs. simulated statistics influenced the expected regression slope of one with a cutoff of $4/n$. For the statistic h/m, LD_{fit} and π, we identified eight, seven, and nine outliers, respectively (a total of ~18% of the dataset). We were unable to estimate LD_{fit} for an additional 13 species due to low levels of polymorphisms in the simulated genomic samples. Three species were found to have outliers in two summary statistics (*Streptococcus salivarus, Haemophilus influenzae,* and the archaeon *Methanosarcina mazei*) and those species were marked with an asterisk on most graphs to reflect the potential inaccuracy of r/m estimates (Supp. Table 1A, Fig. 2). Notably, no species were found to be outliers in all three summary statistics.

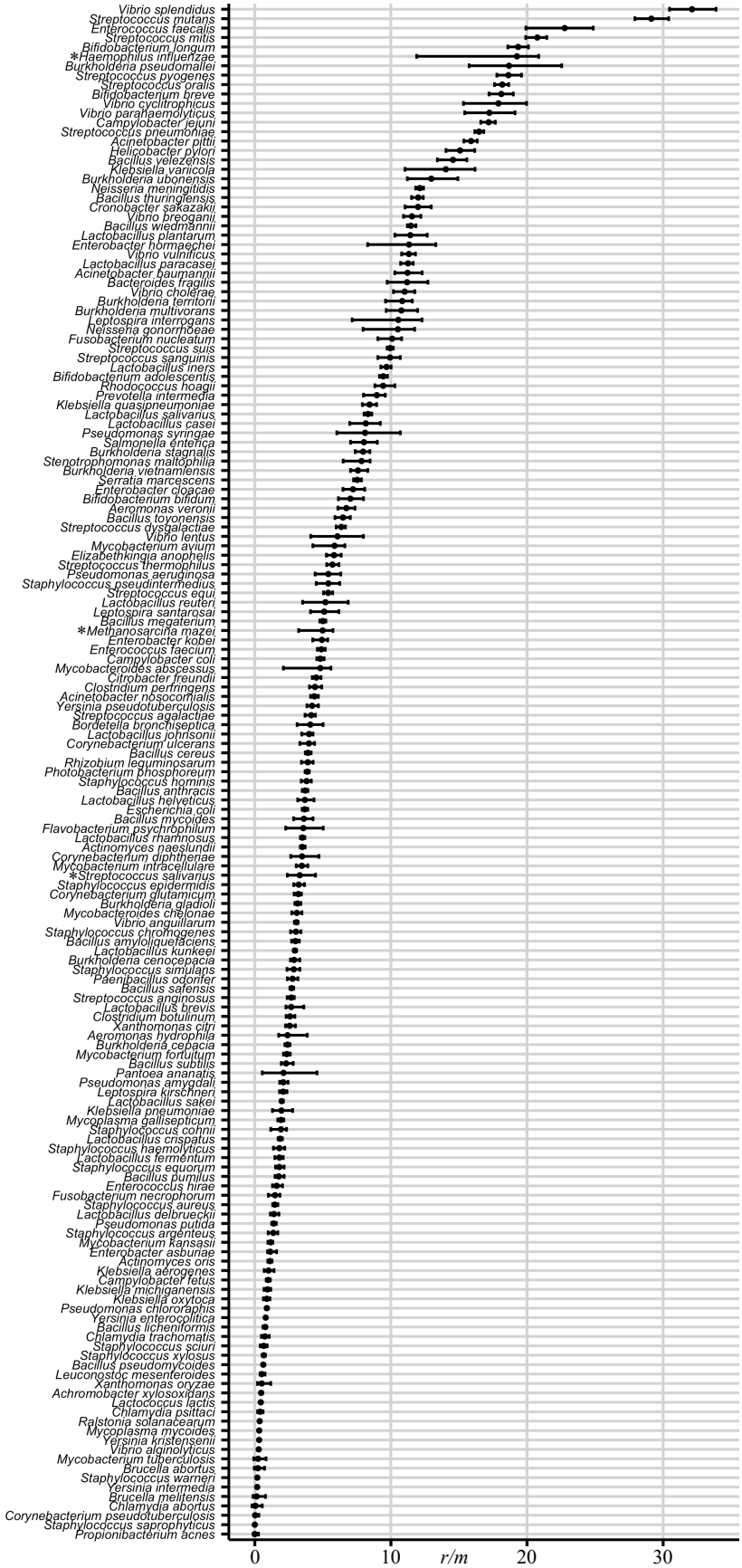
Finally, we tested whether one of the summary statistics was predominantly driving the inference of recombination rates. We found that r/m estimates most strongly correlate to h/m (Spearman's *Rho*=0.72, $P<10^{-14}$), especially for lower recombination rates (r/m<5) (Supp. Fig. 3, A). We did not find significant correlations of r/m with LD_{fit} and with π (Spearman's *Rho*=-0.12, *P*=0.11 and Spearman's *Rho*=-0.17, *P*=0.031, respectively) (Supp. Fig. 3, B&C). This result is somewhat expected since the proportion of nucleotide polymorphisms inferred as homoplasic is expected to increase with more frequent recombination, and the signal of homoplasy is expected to saturate for higher recombination rates.

II.4.3 **Rates of homologous recombination across bacteria**

The rate of effective homologous recombination (r/m) was found to vary widely across the 162 bacterial species analyzed with a median rate of 3.84. For most species (82%), the

homologous recombination rate was found to be more impactful on genomic evolution than mutations alone (r/m>1). The remaining 18% of species were found to acquire polymorphisms more frequently via mutation than homologous recombination (r/m<1). The lowest recombination rate was observed for *Staphylococcus saprophyticus* (r/m = 0.003) and the highest for *Vibrio splendidus* (r/m = 32.18). The average recombination rate across bacteria in this study was r/m = 5.98±5.89 (Fig. 2, Supp. Table 1A).

Figure II—2. Estimates of homologous recombination rate (*r/m*) across species. Each dot represents the estimate of recombination rates (r/m) for each of the 162 bacterial species and one archaeon in our dataset. These estimates represent the median estimated from the posterior distribution of r/m values that were selected using ABC. Black lines represent the standard deviation of r/m estimates of the posterior distribution. Asterisks (*) denote species for which simulated recombination rate lacked accuracy based on the analysis of summary statistics.


Vibrio splendidus
Streptococcus mutans
Enterococcus faecalis
Streptococcus mitis
Bifidobacterium longum
*Haemophilus influenzae
Burkholderia pseudomallei
Streptococcus pyogenes
Streptococcus oralis
Bifidobacterium breve
Vibrio cyclitrophicus
Vibrio parahaemolyticus
Campylobacter jejuni
Streptococcus pneumoniae
Acinetobacter pittii
Helicobacter pylori
Bacillus velezensis
Klebsiella variicola
Burkholderia ubonensis
Neisseria meningitidis
Bacillus thuringiensis
Cronobacter sakazakii
Vibrio breoganii
Bacillus wiedmannii
Lactobacillus plantarum
Enterobacter hormaechei
Vibrio vulnificus
Lactobacillus paracasei
Acinetobacter baumannii
Bacteroides fragilis
Vibrio cholerae
Burkholderia territorii
Burkholderia multivorans
Leptospira interrogans
Neisseria gonorrhoeae
Fusobacterium nucleatum
Streptococcus suis
Streptococcus sanguinis
Lactobacillus iners
Bifidobacterium adolescentis
Rhodococcus hoagii
Prevotella intermedia
Klebsiella quasipneumoniae
Lactobacillus salivarius
Lactobacillus casei
Pseudomonas syringae
Salmonella enterica
Burkholderia stagnalis
Stenotrophomonas maltophilia
Burkholderia vietnamiensis
Serratia marcescens
Enterobacter cloacae
Bifidobacterium bifidum
Aeromonas veronii
Bacillus toyonensis
Streptococcus dysgalactiae
Vibrio lentus
Mycobacterium avium
Elizabethkingia anophelis
Streptococcus thermophilus
Pseudomonas aeruginosa
Staphylococcus pseudintermedius
Streptococcus equi
Lactobacillus reuteri
Leptospira santarosai
Bacillus megaterium
*Methanosarcina mazei
Enterobacter kobei
Enterococcus faecium
Campylobacter coli
Mycobacteroides abscessus
Citrobacter freundii
Clostridium perfringens
Acinetobacter nosocomialis
Yersinia pseudotuberculosis
Streptococcus agalactiae
Bordetella bronchiseptica
Lactobacillus johnsonii
Corynebacterium ulcerans
Bacillus cereus
Rhizobium leguminosarum
Photobacterium phosphoreum
Staphylococcus hominis
Bacillus anthracis
Lactobacillus helveticus
Escherichia coli
Bacillus mycoides
Flavobacterium psychrophilum
Lactobacillus rhamnosus
Actinomyces naeslundii
Corynebacterium diphtheriae
Mycobacterium intracellulare
*Streptococcus salivarius
Staphylococcus epidermidis
Corynebacterium glutamicum
Burkholderia gladioli
Mycobacteroides chelonae
Vibrio anguillarum
Staphylococcus chromogenes
Bacillus amyloliquefaciens
Lactobacillus kurkeei
Burkholderia cenocepacia
Staphylococcus simulans
Paenibacillus odorifer
Bacillus safensis
Streptococcus anginosus
Lactobacillus brevis
Clostridium botulinum
Xanthomonas citri
Aeromonas hydrophila
Burkholderia cepacia
Mycobacterium fortuitum
Bacillus subtilis
Pantoea ananatis
Pseudomonas amygdali
Leptospira kirschneri
Lactobacillus sakei
Klebsiella pneumoniae
Mycoplasma gallisepticum
Staphylococcus cohnii
Lactobacillus crispatus
Staphylococcus haemolyticus
Lactobacillus fermentum
Staphylococcus equorum
Bacillus pumilus
Enterococcus hirae
Fusobacterium necrophorum
Staphylococcus aureus
Lactobacillus delbrueckii
Pseudomonas putida
Staphylococcus argenteus
Mycobacterium kansasii
Enterobacter asburiae
Actinomyces oris
Klebsiella aerogenes
Campylobacter fetus
Klebsiella michiganensis
Klebsiella oxytoca
Pseudomonas chlororaphis
Yersinia enterocolitica
Bacillus licheniformis
Chlamydia trachomatis
Staphylococcus sciuri
Staphylococcus xylosus
Bacillus pseudomycoides
Leuconostoc mesenteroides
Xanthomonas oryzae
Achromobacter xylosoxidans
Lactococcus lactis
Chlamydia psittaci
Ralstonia solanacearum
Mycoplasma mycoides
Yersinia kristensenii
Vibrio alginolyticus
Mycobacterium tuberculosis
Brucella abortus
Staphylococcus warneri
Yersinia intermedia
Brucella melitensis
Chlamydia abortus
Corynebacterium pseudotuberculosis
Staphylococcus saprophyticus
Propionibacterium acnes
0
10
r/m
20
30

Recombination rates are thought to be shaped by multiple factors, traits, and life conditions. In turn, recombination is predicted to impact the evolution of bacterial genomes. Here we attempted to identify which factors correlate to the evolution of recombination rate in bacteria. We first compared recombination rate estimates with genomic traits intrinsic to our datasets. We found no significant correlation between *r/m* and the number of core genes, the length of the core genome, the average nucleotide pairwise identity across core genomes, or the average GC-content across core genomes (Supp. Fig. 4, C, B, D, F). We did however find a positive correlation between *r/m* and the number of genomes in each species population (Supp. Fig. 4, A: Spearman's *Rho*=0.46, $P<10^{-9}$) and a negative correlation between *r/m* and the standard deviation of nucleotide pairwise identity (Supp. Fig. 4, F: Spearman's *Rho*=-0.40, $P<10^{-5}$). These results suggest that recombination rates may be underestimated for species with fewer sequenced genomes and that species composed of more structured populations present lower *r/m* estimates. Using metabolic and structural data available from JGI's GOLD database (97) (Supp. Table 1E) we found no significant difference in the median *r/m* values for motile vs. nonmotile species (*n*=48 and *n*=68), sporulating vs. nonsporulating (*n*=9 and *n*=85), autotrophic vs. heterotrophic (*n*=22 and *n*=42), Gram negative vs. Gram positive (*n*=67 and *n*=79), and anaerobic, aerobic, and facultatively aerobic (*n*=22, *n*=49, *n*=55) species (Wilcoxon rank-sum test, *P*>0.05; Supp. Fig. 5, A-D & F). However, the median difference in *r/m* for bacteria described as free-living (*n*=103, *r/m* = 3.89) vs. those described as living as obligate intracellular bacteria (*n*=4, *r/m* = 0.59) was significant (Wilcoxon rank-sum test: *P*=0.03) which is consistent with the idea that intracellular species participate in less DNA exchanges due to their lifestyle (Supp. Fig. 5, E). Next, we analyzed the potential link between gene content and metabolic functions of bacteria and recombination rates. We tested correlations between inferred *r/m* and

the proportion of various gene functions inferred from COG (clusters of orthologous genes) and CAzy (Carbohydrate-Active Enzymes) categories identified with *eggNOG* across each reference genome for all bacterial species in this study (98–100). No statistically significant correlations between the proportion of genes annotated with specific gene epithets and recombination rate were observed after correcting for multiple testing (Supp. Table 4). Next, because bacteriophages are known vectors capable of promoting DNA transfers across bacteria and their abundance in lysogenic bacteria may potentially drive higher recombination frequencies (101) prophage abundance was assessed for each reference genome using *geNomad* (102) (n=159 as three species could not be analyzed). Both the abundance of prophages as well as the number of genes annotated by *geNomad* as phage were found to be statistically unrelated to variations in *r/m* across species (Spearman's *Rho*=0.95, *P*=0.63) (Supp. Fig. 6, Supp. Table 1G). Finally, to determine whether the ecosystem from which the bacterial strains were predominantly collected influenced recombination rate, the sampling origin (collection data) of each strain was acquired from GenBank when available and summarized into 30 discrete categories (Supp. Table 3). A positive correlation was found between the average *r/m* for a species and the number of strains denoted as "human derived" (Spearman's *Rho*=0.35, Holm's adjusted *P*=0.002). However, as mentioned previously, we observed a correlation between the average *r/m* per species and the number of genomes used in the analysis (Spearman's *Rho*=0.46, Holm's adjusted *P*=0.001) indicating that the correlation may be biased by the fact that more sequenced genomes are available for bacterial species which are associated with humans (32–34).

II.4.4 **Robustness of homologous recombination rate estimates**

Across studies, the inference of recombination rate of bacterial species has often yielded inconsistent estimates depending on the method that was used but also depending on the dataset

under study (1–3). An extreme example of these inconsistencies is *Salmonella enterica*, which has been found to be clonal (*r/m*=0) in one study (103) but was inferred as one of the most recombining bacterial species in another one (*r/m*=30) (1). It is unclear whether these varying estimates are the result of methodological issues or whether they reflect biological variations of recombination rates across samplings of populations of the same species. This is particularly problematic because the borders of bacterial species are often defined inconsistently across studies and because new genome sequences are actively being added to databases (4).

To test for the impact of population sampling on our estimates of recombination rates, we conducted several analyses. First, we tested whether variation in strain diversity affected our estimates of *r/m*. For this analysis, each species containing >25 genomes (*n*=82) were subsampled in 100 replicates where 10 genomes were randomly removed from each replicate (sample size = *n*-10). After removing 10 random genomes, the three summary statistics (*h/m*, LD_{fit} and π) were recalculated from each replicate. A new posterior distribution was then generated from each set of summary statistics using ABC to estimate *r/m* for each replicate. Estimates of *r/m* obtained for the 100 random subpopulations were then compared to those estimated for the whole species population to determine how subsampling bias and strain variation in sampling impacts *r/m* estimates for the species. We observed that *r/m* estimates for the 100 replicates of *n*-10 random genomes sampled from the parent population were typically nearly identical to the *r/m* estimates of the parent population for most species (Spearman's *Rho*= 0.85, $P<10^{-16}$) (Supp. Fig. 7). Nine species (11%) had an average *r/m* across their subsampled populations which varied from the species *r/m* by more than 5 (Supp. Table 1J), but, overall, these results indicate that most of our *r/m* estimates are robust to subsampling.

Secondly, we tested to what extent sampling biases and the method used for dataset construction impacted our estimates. We conducted a large resampling analysis of the genomes of *Escherichia coli* whose core genome was built in a previous study using a different method (3). We chose *E. coli* because of the large number of genomes available for this species and because its population structure has been clearly characterized. We conducted a phylogenomic analysis of this species and identified the major clades of *E. coli*: A, B1, B2, D and E (see Methods). We then randomly subsampled 15 genomes of each clade five times and inferred recombination rates for each subsample (n=25) with our ABC approach. Importantly, the core genome of this dataset was built using a different approach and includes genomes from different strains compared to this study. Despite the difference in the nature of the dataset, we inferred similar recombination rates across these samples (mean r/m = 2.53) (Supp. Table 1B) compared to the estimate of *E. coli* predicted independently in this study (r/m=3.68) (Supp. Table 1A). Although we observed some variation across samplings, most estimates of r/m were overall consistent across phylogroups (mean r/m = 2.53±0.79), indicating that our method is robust to population samplings (Supp. Table 1B). In addition, we did not observe substantial differences in recombination rate estimates across the five phylogroups, suggesting that sampling bias does not strongly affect our estimates and that population structure does not significantly impact recombination rate, at least in *E. coli*.

We then compared our estimates of recombination rates to estimates generated by one of the most popular tools designed to infer r/m in bacteria: *ClonalFrameML* (45). As opposed to our approach, *ClonalFrameML* relies solely on phylogenetic incongruencies to infer recombination rates. We used the same dataset of 25 samplings across *E. coli* phylogroups and generated r/m estimates with this approach (note that five of the samples could not be processed by

ClonalFrameML (see Methods, Supp. Table 1B)). Overall, inferred *r/m* values for the subpopulations tended to be higher than our estimated rates (our method: mean *r/m* = 2.53±0.79; *ClonalFrameML*: mean *r/m* = 4.16±1.38). The standard deviation of these estimates was nearly double, suggesting that *ClonalFrameML* estimates are less robust to strain-level variations than our method (Supp. Fig. 8, Supp. Table 1B). Next, we tested *ClonalFrameML* against 84 bacterial species from our dataset and found that this method yielded similar recombination rates compared to our approach for species with lower *r/m* estimates (*r/m*<5). However, we found that *ClonalFrameML* rarely estimated *r/m* values over 5 (*n*=7 species); the average rate estimated for across all 84 species was *r/m*=2.0±1.8 with *ClonalFrameML* and *r/m*=5.6±5.3 with our method (Supp. Fig. 9, Supp. Table 1C).

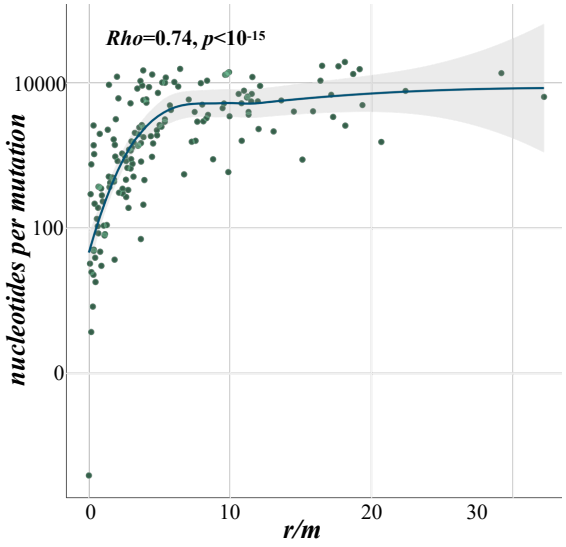
Altogether, these different analyses support the accuracy and the consistency of our approach. Although some variations are observed across samplings of a given species, estimates appear rather robust to genome sampling, especially compared to the large spread of *r/m* estimates generated across all bacterial species (see below). On average, our estimates yielded higher recombination rates than another method (*ClonalFrameML*), but as mentioned above, previous methods (including *ClonalFrameML*) are theoretically incapable of detecting recombination events that do not leave a direct signal of recombination (e.g., single alleles) and can face difficulties recognizing multiple recombination events; they tend therefore to underestimate recombination rates.

II.4.5 **Inference of overall nucleotide exchange in bacterial recombination**

The results presented thus far were reported as effective rates of recombination (*r/m*), a metric used by most approaches since it attempts to quantify the number of allelic variants exchanged by recombination. Effective rates of recombination are usually inferred because

recombination events between two identical DNA strands do not leave any genomic signature. In addition, the transfer of allelic variants does not systematically leave a signature of recombination in the genomes. Indeed, transferred alleles do not always generate patterns of homoplasies and do not necessarily impact linkage disequilibrium. As opposed to most approaches, our method can estimate an overall rate of recombination, which we defined as the total number of nucleotides that have been exchanged by homologous recombination relative to the number of alleles introduced by mutation regardless of the presence and number of allelic variants on the recombined fragments. The average number of nucleotides exchanged by homologous recombination to mutations varied from 0.003bp in *Staphylococcus saprophyticus* to 19,606bp in *Bifidobacterium breve* with an average of 3,896±1,263 across species (Supp. Table 1A). Note that this metric does not represent the number of nucleotides exchanged for a single recombination event, but, rather, the sum of all nucleotides exchanged per substitution introduced by mutation. Interestingly, we found that the number of nucleotides exchanged increased exponentially with effective recombination *(r/m)* until it reaches a plateau at $r/m \sim 5$. This suggests that at higher recombination rates, signal saturation is occurring when measuring effective recombination rate: the same alleles are being exchanged via homologous recombination multiple times through the species' evolution (Fig. 3). It further suggests that—as expected—our ability to estimate accurate effective rates of recombination declines substantially for species with high recombination rates ($r/m >>> 5$). This last observation agrees with the fact that the frequency of homoplasies is positively corelated with r/m for $r/m < 5$ and that the estimates of *ClonalFrameML* rarely exceed 5 ($r/m < 5$) (Supp. Fig. 3A, Supp. Fig. 9).

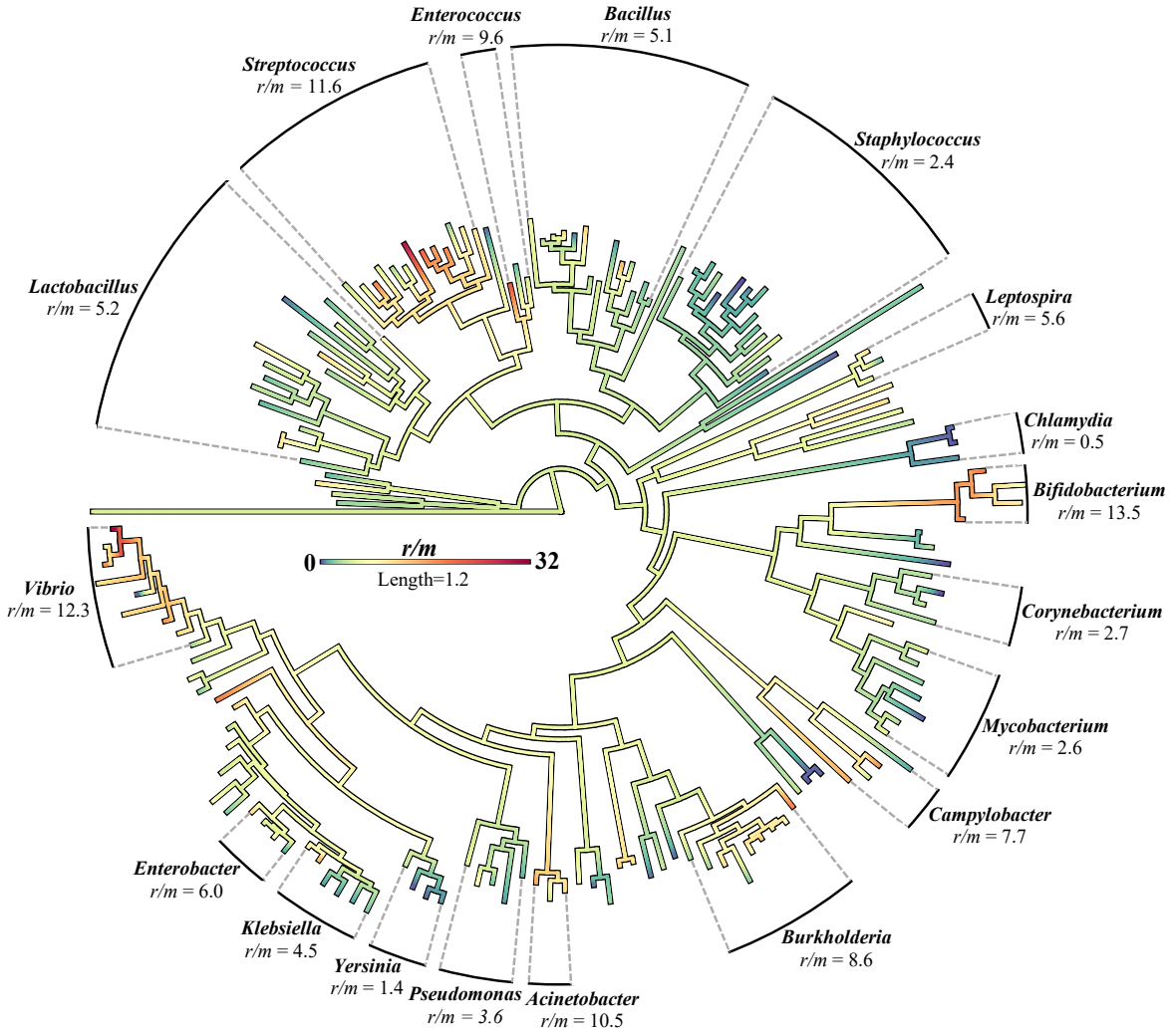
Figure II—3. Absolute number of nucleotides exchanged by recombination per mutation relative to recombination rate (*r/m*) for 162 bacteria and one archaeal species (Spearman's *Rho*=0.74, *P*<10^{-15}).



II.4.6 Evolution of recombination rate across the bacterial tree

Although recombination is a key mechanism impacting the evolution of bacteria, very little is known about the factors driving recombination rate. The evolution of recombination rate—as a trait—has also not been studied in detail thus far. Here, we conducted a phylogenomic study to reconstruct the evolution of recombination rate along the tree of bacteria. A phylogenetic tree of 162 bacterial species and one archaeon was reconstructed using 13 universal genes shared by all species. We used a maximum likelihood approach implemented in *Phytools* (104) to reconstruct the evolution of recombination rate across the internal nodes of the tree (Fig. 4 and see Methods).

Figure II—4. Evolution of homologous recombination rate (*r/m*) across bacteria. The phylogeny of the 162 bacterial and one archaeon was generated using a concatenated amino-acid alignment of 13 universal single-copy core orthologs identified in all reference strains by *eggNOG-mapper* (*v5*) (98)**. The maximum likelihood tree was built using RAxML and 1000 bootstrap replicates** (105)**. The tree was rooted using the archaeon *Methanosarcina mazei*. Ancestral state reconstruction of the of homologous recombination rate was inferred using a maximum likelihood approach with the fastAnc function of the R package Phytools** (104)**. The gradient of colors across the tree ranges from *r/m*=0 (blue) to *r/m*=32 (red) and represents the inferred rates of recombination across branches of the tree. Genera comprising three or more species are labelled and their average recombination rate is represented on the tree.**

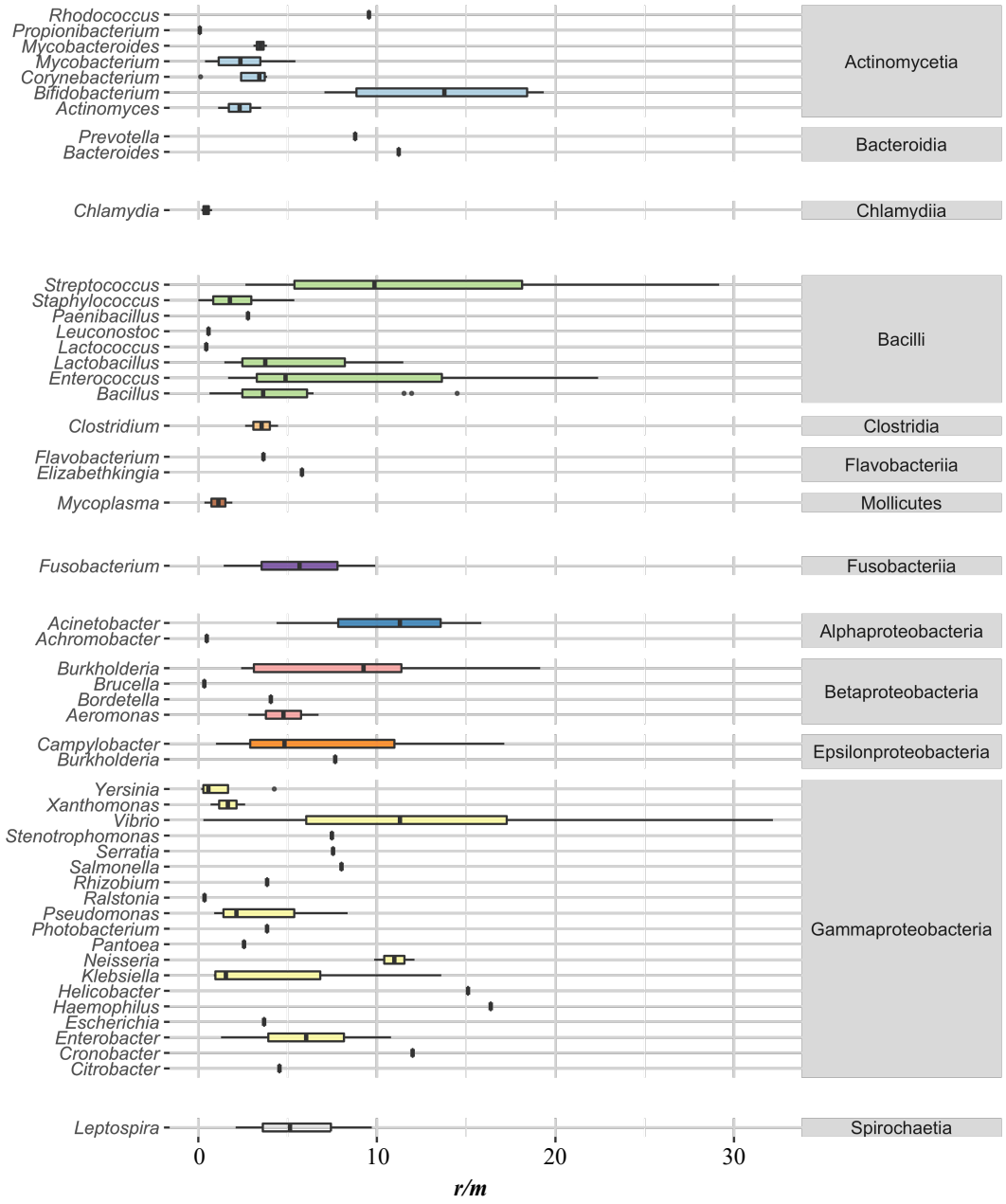


Interestingly, we found that recombination rate is a rather slowly evolving trait that appears conserved within several genera. Indeed, we observed that recombination rate varies significantly across bacterial genera (Kruskal-Wallis test, $P<10^{-3}$, genera were included in this test when composed of five species or more). Of the nine genera with ≥5 species, *Streptococcus* and *Staphylococcus* significantly differed from the distribution of *r/m* for all other genera

(Wilcoxon tests, $P<0.01$, Bonferroni correction). The recombination rate of *Staphylococcus* was consistently and significantly lower across the species of this genus (median r/m=1.8) relative to all other genera (r/m=4.0). In contrast, the species of the genus *Streptococcus* displayed consistently higher recombination rates (median r/m=9.9) which is significantly above the recombination rate of the other genera (median r/m=3.8) (Fig. 5). Among the genera composed of only three or four species, *Chlamydia* displays the lowest average recombination rate (r/m = 0.45±0.3, n=3) and *Bifidobacterium* was inferred as the most recombinogenic (r/m = 13.5±0.45, n=4). We did not observe any evidence that recombination rate was conserved over longer phylogenetic distances (i.e., above the genus level).

Although recombination rates appear somewhat conserved in most genera, several genera are composed of species whose recombination rate varies sharply from their overall genus average: *Enterococcus* (r/m=9.6±11.2, n=3) and *Vibrio* (r/m=12.3±9.5, n=9) present the most variation in r/m across their respective genus. For instance, *Vibrio alginolyticus* (r/m=0.28) has a much lower rate of recombination relative to closely related taxa such as *Vibrio splendidus (*r/m = 32.2*)*. These results indicate that, although recombination rate is—overall—a slowly evolving trait, it can also evolve much faster in some lineages.

Figure II—5. Recombination rate *r/m* across bacterial genera for 162 bacterial species. Genera are grouped by phylogenetic class. The middle bar of each box indicates the median value; the left and right vertical edges represent the first and the third quartile, respectively. The central horizontal lines indicate the data range, with a maximal distance of 1.5 interquartile ranges (*i.e.*, the distance between the first and third quartile).


Rhodococcus
Propionibacterium
Mycobacteroides
Mycobacterium
Corynebacterium
Bifidobacterium
Actinomyces
Actinomycetia
Prevotella
Bacteroides
Bacteroidia
Chlamydia
Chlamydiia
Streptococcus
Staphylococcus
Paenibacillus
Leuconostoc
Lactococcus
Lactobacillus
Enterococcus
Bacillus
Bacilli
Clostridium
Clostridia
Flavobacterium
Elizabethkingia
Flavobacteriia
Mycoplasma
Mollicutes
Fusobacterium
Fusobacteriia
Acinetobacter
Achromobacter
Alphaproteobacteria
Burkholderia
Brucella
Bordetella
Aeromonas
Betaproteobacteria
Campylobacter
Burkholderia
Epsilonproteobacteria
Yersinia
Xanthomonas
Vibrio
Stenotrophomonas
Serratia
Salmonella
Rhizobium
Ralstonia
Pseudomonas
Photobacterium
Pantoea
Neisseria
Klebsiella
Helicobacter
Haemophilus
Escherichia
Enterobacter
Cronobacter
Citrobacter
Gammaproteobacteria
Leptospira
Spirochaetia
0
10
20
30
r/m

II.5 **Discussion**

Estimating recombination rates in bacteria has proven a difficult endeavor as homologous recombination leaves complex and sparse genomic signatures. Consequently, inconsistent rates have often been inferred for the same species across studies. Multiple computational tools have been developed to estimate homologous recombination rates in bacteria and they typically rely on direct evidence from a single type of signatures of recombination (87). However, recombination events do not always leave a straightforward signature of recombination; for example, when an allele is exchanged multiple times, or when an allele is found in a single strain. Therefore, most tools designed to infer recombination rates likely underestimate true recombination rates by incorrectly inferring some recombination events as mutational events. To circumvent these caveats, our approach leverages an ABC framework to infer recombination rates and integrates multiple signatures of recombination (*i.e.,* homoplasies, linkage disequilibrium, and polymorphisms). By relying on simulations and ABC, our approach theoretically allows us to account for the complex scenarios where homologous recombination would otherwise not be detected. As predicted, we did observe that our estimates of *r/m* were on average higher when compared to previously published estimates and to those obtained with *ClonalFrameML* (1, 2). Importantly, our approach is also capable of estimating recombination rates as a function of total nucleotides exchanged per mutation event rather than an effective rate of recombination expressed as the number of polymorphisms exchanged per mutation.

The performance of our approach was evaluated using different simulations and resampling approaches. Overall, our results were robust but appeared to lack precision for simulated datasets presenting very few polymorphisms and little evidence of recombination. However, lack of precision in these cases is expected as virtually every method will not perform

well when there is a near total lack of genomic signal to analyze homologous recombination. Nevertheless, recombination rates inferred by our method in these scenarios was, overall, in the vicinity of the real value (Supp. Fig. 1). Most importantly, we tested the robustness of our approach using multiple resampling analyses and resampling biases and results showed consistent inference of recombination rates for the same species (Supp. Fig. 7).

Although we found that our estimates of *r/m* were within range of those inferred by *ClonalFrameML* for several samples of *E. coli* genomes (Supp. Fig. 8) (45), our estimates were substantially different across species. When comparing our results obtained to *r/m* values obtained with *ClonalFrameML* for other bacterial species we found that very few *ClonalFrameML* estimates exceeded *r/m*~5 (Supp. Fig. 9). This is coherent with our observation that the genomic signal left by recombination begins to saturate at *r/m*~5 (Fig. 3) and reflects the fact the same allele is likely exchanged multiple times above this rate. This is also supported by the fact that the homoplasy to mutation ratio is linearly related to *r/m* values until *r/m*~5 (or, *h/m* >1) (Supp. Fig. 3A). Overall, this indicates that approaches relying on direct evidence of site exchange may be unable to reliably detect recombination rates greater than 5. The inability to accurately infer recombination rates when recombination is high is a well-known limitation to the tools that rely on the direct estimate of recombination signal as previously noted (45).

The rates of homologous recombination estimated across phylogroups of *E. coli* was within ranges of those typically reported in prior studies (*r/m* = 1.7-3.0) (3, 106). Though less data exists for *r/m* comparison between studies for other bacterial species, estimates from our study differed substantially from previous estimates but are typically within the range of those reported in previous analyses (our estimates occupy a smaller range of *r/m* values). In their 2009 study, Vos & Didelot used *ClonalFrame* to estimate recombination rates using MLST data

across 48 species (1). This study found *r/m* rates to vary by over three orders of magnitude: from *r/m* = 0.02 (*Leptospira interrogans*) to *r/m* = 63.6 (*Flavobacterium psychrophilu*m). A more recent multi-species analysis by González-Torres et al. (2019) focused on 54 bacterial and archaeal species with a shared total of 338 genomes found rates to vary from *r/m* = 0.00 (*Frankia* genus) to *r/m* = 973.8 (*Burkholderia pseudomallei MSHR3*) by computing *r/m* values from genomic sites which were found to be exchanged by homologous recombination by at least three programs designed to detect recombination events (2). By comparing the recombination rates of species included in these studies and ours (Supp. Fig. 11), we found that these analyses inferred very low recombination rates for several species, which were inferred to recombine at higher rates in our study (see Supp. Table 1D). As stated previously, the actual rates of homologous recombination in bacteria are likely underestimated by tools relying on direct inference of recombination signal. Moreover, this study relied on an average sample size of six genomes per species and we observed that using smaller genomic samples can lead to strongly underestimate *r/m*. Our study represents the largest analysis of recombination rates from complete genomes to date with 162 bacterial species (8,706 genomes) and estimated rates varied from *r/m* = 0.003 (*Staphylococcus saprophyticus*) to *r/m* = 32.3 (*Vibrio splendidus*) (Fig. 2, Supp. Table 1A). Overall, our results encompassed a narrower range or *r/m* values and differed significantly from estimates derived in previous studies (Supp. Fig. 11, Supp. Table 1B & 1E).

We found that GC-content is not correlated to predicted *r/m* values as opposed to what was reported in eukaryotes (107, 108). Previous studies have hypothesized that GC-content may be partially driven by homologous recombination through biased gene conversion (109, 110). Alternatively, it has been suggested that higher rates of recombination would drive higher GC-content by enhancing the effectiveness of selection (111). Our results do not support an overall

correlation between GC-content and recombination rate across species (Supp. Fig. 4F). However, our analysis does not allow us to conclude whether these effects are driving variations of GC-content across genomic regions in the same species as suggested by previous results in *E. coli* (3, 111, 112). Answering this question would require analyzing the variation of recombination rates along bacterial chromosomes.

The forces driving recombination rate and its evolution remain largely unknown in bacteria, and few studies have explored this question (1, 2, 113). Here, we observed that recombination rate was rather conserved across several genera, indicating that homologous recombination appears as a relatively slowly evolving trait. However, some genera presented large differences in recombination rate—such as the genus *Vibrio.* These sharp variations in recombination rate between related species such as *V. splendidus* (*r/m* = 32) and *V. alginolyticus (r/m* = 0.3*)* may be attributed to differences in ecological niche. Perhaps these differences may also be explained by sampling biases where the sequenced genomes of *V. splendidus* which, in this study, were almost entirely represented by strains isolated from filtered seawater (65/68 genomes; core genome pairwise identity = 97.5%) whereas *Vibrio alginolyticus* was represented by strains associated with a variety of saltwater organisms (including fish, mollusk, coral, sponge, and kelp), ocean soils, and in human patients (average core genome pairwise identity = 94.0%) (Supp Table 1A, Supp. Table 3A). Though few obligate intracellular organisms were included in this study, we estimated a much lower rate of homologous recombination (*r/m*<1) for obligate intracellular bacteria such as *Chlamydia spp*. which was also reported in previous studies (Supp. Fig. 5E) (60). Intracellular bacteria are expected to present lower rates of recombination since they are much less likely to encounter conspecific individuals in nature. Additionally, we observed that the number of human-associated strains within a species

population correlated positively with *r/m* and it is possible that many of them are human pathogens due to the sequencing bias towards bacteria which impact human health (Supp. Table 3). However, we did not observe a significant correlation between pathogenicity—as inferred from literature analysis such as (114)—and recombination rate (Supp. Fig. 10, Supp. Table 1F) (107, 115–126).

Despite decades of work, the analysis of recombination in bacteria remains a challenging task (87). This study offers an integrative approach to estimate recombination rates in bacteria which is not limited to the analysis of direct signatures or recombination. By leveraging this method, we report estimates of recombination rates for the largest dataset of bacteria to date under a unified framework. Albeit computationally expensive, our approach appears to yield robust and consistent results. However, much work remains to be conducted to uncover which factors are shaping the patterns and the rates of recombination across bacteria and how homologous recombination varies across the genome.

II.6 Methods

II.6.1 Genome acquisition, Core Genome Assembly, and Phylogeny

All genomes used in this study were downloaded from NCBI's GenBank for each bacterial and archaeal species presenting 15 or more fully assembled genomes. The original dataset was composed of 333 bacterial species comprised of 83,532 assembled genomes. The threshold of 15 or more genomes for each species used in this study was established as in (8). To verify assembly completeness of the genomes, Hidden Markov Model profiles of 45 universal bacterial and archaeal protein markers were detected using HMMER as in (127–129). Genomes were conserved in our analysis when all protein markers were identified. These universal orthologs were then aligned with MAFFT and concatenated (130). Pairwise sequence identity of

the concatenate was estimated and strains with identical nucleotide identities across all marker genes were considered duplicates and a single genome of each duplicate was randomly conserved for the analysis. For each species, the assembly containing the fewest contigs and the most predicted coding regions was chosen as the reference genome for the phylogenomic analysis (see below).

We built the core genome for each species using *CoreCruncher* with the stringent option and *USEARCH* (global) using an 80% protein identity cutoff (131, 132). The core genome was defined as the set of orthologous genes shared by over 90% of the genomes of a given species. The *CoreCruncher* workflow aligns core genes with *Muscle* (133)at the amino acid level, reverse translates these alignments, and then concatenates all core genes to create a single aligned core genome concatenate per genome. Species borders were then refined as in Diop et al. 2022: genomes were considered as part of the same species when *i*) sharing an average nucleotide identity (ANI) over 94% and *ii*) no interruption of gene flow across genomes of the same species using the *ConSpeciFix* approach. Genomes that did not meet these conditions were excluded from the dataset.

Due to the removal of misclassified and low-quality genomes, several species were entirely excluded from our dataset when their number fell below the threshold of 15 genomes set for this analysis. Furthermore, some species presented too many assembled genomes for tree building and recombination analysis and these species were randomly sampled down to 100 genomes (apart from *Bacillus thuringiensis* for which we completed the analysis at 174 genomes before establishment of the 100-genome threshold). A final dataset of 7,451 unique genomes across 162 bacterial species and one archaeal species was used for this analysis (Supp. Table 2).

A strain phylogeny was then generated from the core genome nucleotide concatenate of each species using RAxML v8 with a GTR+gamma model (134).

II.6.2 Forward-in-Time Simulation with Homologous Recombination

Evolution of the core genomes was conducted with the forward-in-time simulator *CoreSimul* (96). The core genome alignments for each species were used to generate a random ancestral core genome sequence with the same length as the core genome concatenate and the same GC content. This ancestral genome was then evolved *in silico* with *CoreSimul* following the population structure (i.e., topology of the phylogenetic tree) of the species and using an identical transition/transversion ratio and substitution rate as observed from the true core genome alignments and phylogeny of each species (96). Simulations were conducted using a wide array of parameters: *i*) average recombination rate *Rho* (from 0 to 20 recombination events per mutation), *ii*) average mutation rate *theta* (rescaled from 0 to 100% of the branch lengths of the phylogenetic tree) and iii) average tract lengths of recombination *delta* (from 0 to 1,000bp). These ranges of parameters were based on estimates from diverse studies (3, 8, 96, 135). Combinations of these three parameters were used to simulate core genome evolution with *CoreSimul*; for each branch of the tree, two genomes were randomly selected, and recombination events and mutations events were drawn from a Poisson distribution of mean *Rho* and *Theta*, respectively and each recombination tract length was generated from a geometric distribution of mean *delta* as in (96, 136). This represents a total of ~500,000 simulations per species analyzed. For each simulation, the number of polymorphisms exchanged during each recombination event (*nu*) was recorded. Finally, the effective recombination rate (*r/m*) was expressed as $\frac{r}{m} = \frac{nu \times delta \times rho}{theta}$ (Fig.1, A-C) (136).

To determine which simulations best recapitulated the recombination rate of the real species, several metrics were calculated from the core genomes and phylogenies from *both* the real datasets and the populations of simulated core genomes that resulted from each of the 500,000 simulations: π (*i.e.* average nucleotide diversity), *h/m* (*i.e.* the ratio of homoplasies to non-homoplasies) as in (8), and LD_{fit}, where LD_{fit} represents the decay of genetic linkage (r^2) across distances of pairwise polymorphic sites. Briefly, a non-linear model was fit to the decay of linkage relative to nucleotide distances for the true core genome concatenate. For each simulated core genome, the pattern of linkage was compared to the pattern of linkage observed in the true core genome using the root mean square deviation (RMSE). Finally, we used these three summary statistics to compare each of the 500,000 simulations to the three statistics observed in the true core genome using Approximate Bayesian Computation (ABC) with the R package *abc* using the loclinear method (137). We set a tolerance threshold of 0.01% to generate the posterior distribution or *r/m* values that best recaptured the summary statistics of the true core genome of each species and thus, represents the most likely rate of recombination rate simulated (Fig.1, D). Real species summary statistics and simulated summary statistics from the posterior distribution as well as the inferred *r/m* rate are listed in Supp. Table 1A. The scripts developed for this analysis has been made available on GitHub as the pipeline *recABC* (https://github.com/lbobay/recABC). Methods and additional analyses of the results of *recABC* are included in the Supplementary Text.

II.7 Supplementary Materials

II.7.1 Supplementary Text

II.7.1.1 Methodology Validation, and Exploration of Bias

To assess the fidelity of *ABC* in choosing simulated populations that best recaptured genomic characteristics of the real species, the median of the posterior distribution of each simulated summary statistic—*h/m*, LD_{fit} and π, respectively—was compared to the true statistic value for the species. Outliers were considered species for which the Cook's Distance difference in real vs. simulated statistics influenced the expected regression slope of one with a cutoff of $4/n$. For the statistic *h/m*, the species *Haemophilus influenzae*, *Helicobacter pylori*, *Lactobacillus sakei*, *Methanosarcina mazei*, *Streptococcus equi*, *Streptococcus mitis*, *Streptococcus oralis, and Vibrio breoganii* had predicted values found to be influential to the regression. For the statistic π the median simulated posterior values from the species *Bacillus pumilus, Campylobacter fetus, Corynebacterium diphtheriae, Lactobacillus reuteri, Pseudomonas syringae, Streptococcus oralis, Streptococcus salivarius, Vibrio alginolyticus, and Vibrio lentus* were found to be influential. We were unable to infer LD_{fit} for 13 species due to lack of polymorphisms (all genomes were highly similar). Of the remaining, *Fusobacterium nucleatum, Haemophilus influenzae, Methanosarcina mazei, Mycobacteroides abscessus, Stenotrophomonas maltophilia, Streptococcus salivarius,* and *Vibrio cholerae* were found to have simulated values of LD_{fit} which varied strongly from the expected values. Notably, no species were found to have deviations in all three summary statistic values. Supplementary Table 1A shows the posterior estimates of *r/m* with asterisk(s) to highlight the species for which there was lower confidence in the recapitulation of one or more summary statistics which may indicate that *r/m* estimates for these indicated species are less robust. A scatter plot of the real vs. simulated summary statistics was generated for each of the three summary statistics (*h/m*, LD_{fit}, and π) with the respective species outliers removed for each plot (Supp. Fig. 2 A-F). We also verified that the central tendency of the prior and the posterior distributions of *r/m* values were clearly different from one another for

most species, indicating that simulation abundance around a particular recombination rate was not a determinant factor in identifying the most probable recombination rate for a given species.

II.7.1.2 **Analyses of recombination rates**

To determine the impact of population structure on recombination rate estimates, we analyzed the core genome built for 400 genomes of *E. coli* and its corresponding phylogenetic tree (data from (128)). We randomly selected five samples of 15 genomes from each of the five major phylogroups of *E. coli* (A, B1, B2, D and E). Phylogroups were identified from the tree and from the classification in (128). The core genome of each sample of 15 genomes was then extracted from the main concatenate, and a phylogenetic tree was built using RAxML v8 using a GTR+ Gamma model (134). For each sample (n=25) the core genome as well as the reconstructed phylogeny were used to estimate homologous recombination rates by both our method and using *ClonalFrameML* (v1.12) (45) (five of the n=25 samples did not complete the run) (Supp. Fig. 8, Supp. Table 1B). *ClonalFrameML* was also used to estimate the recombination rates for 84 species from our dataset (Supp. Fig. 9, Supp. Table 1C).

The evolution of recombination rate along the tree of bacteria was conducted using a maximum likelihood approach. The tree of all the species of our dataset was generated using 13 universal single-copy COGs found in all reference strains of the species dataset using *eggNOG-mapper* (v5) (COG0533, COG0495, COG0202, COG0099, COG0197, COG0094, COG0097, COG0096, COG0092, COG0018, COG0522, COG0480, COG0088) (98, 99). The amino-acid sequences of the 13 universal COGs were aligned using Mafft v7.49 and concatenated (130). The concatenated alignment was used to generate a multi-species phylogeny using raxmlHPC-Hybrid-SSE3 with the PROTGAMMAWAG model of evolution with 1,000 bootstrap replicates (105). The tree was then used to reconstruct the ancestral state of recombination rate at each

internal node of the phylogenetic tree using the continuous mapping fastAnc function of the R package *Phytools* (104) (Fig. 4).

eggNOG-mapper (v5) was used to annotate the reference genome of each species to determine how COG and CAzy annotated gene content and genes related to recombination correlated to species level differences in recombination rate (98–100). The number of genes defined by a specific COG or CAzy epithet were summed and divided by the total number of genes annotated in the reference species. All Spearman correlations between gene annotation epithets and average recombination rate are listed in Supplementary Table 4.

Species phenotypic and metabolic traits including gram stain, motility, sporulation, energy metabolism, oxygen tolerance, and general lifestyle were compared with recombination rate using data from the JGI's (Joint Genome Institute) Genomes Online Database (GOLD) when available (Supp. Table 1E) (97). Species trait groupings were then compared to recombination rates (Supp. Fig. 5). Using data from Bartlett et al. (2022), species were classified into pathogens, non-pathogens, and putative pathogens (Supp. Fig. 11, Supp. Table 1G) (114). When data was not available for the species from this study, evidence of pathogenicity was inferred from a separate literary analysis as denoted in Supplementary Table 1G (114–121, 123–126). Additionally, all genomes in this study with collection data available in their GenBank feature format (.gff) files were classified into discrete categories (ex: genomes originating from various human infections were classified as "human" etc.) to compare the number of genomes within a species with a specific epithet to the recombination rate of that species (Supp. Table 3). Finally, the tool *geNomad* was used to identify putative prophages in the reference strain used for each species in this analysis and both the number of genes annotated as viral as well as the total

number of putative prophages identified for each species were compared to species *r/m* values (102) (Supp. Table 1H, Supp. Fig. 6).

II.7.2 Supplementary Figures

Figure S II—1. Assessment of our ABC approach. XY-Plot comparing the accuracy of our ABC pipeline in predicting *r/m* from simulations where *r/m* is known. For each species, 50,000 simulations were discarded from the prior distribution and 5,000 of those simulated datasets with known recombination rates were randomly chosen. We predicted *r/m* from the remaining pool of 450,000 simulations (which *did not* include the test population of the 5,000 known simulations) with a tolerance of 0.01%. The *x*-axis shows the average of the known *r/m* estimates across the 5,000 simulations of each species and the *y*-axis is the average of each posterior dataset generated by ABC (*n*=45) averaged over each of the 5,000 query datasets where *r/m* was known. Vertical error bars denote the standard deviation for the average over each population (n=5,000) of predicted recombination rates (*r/m*).

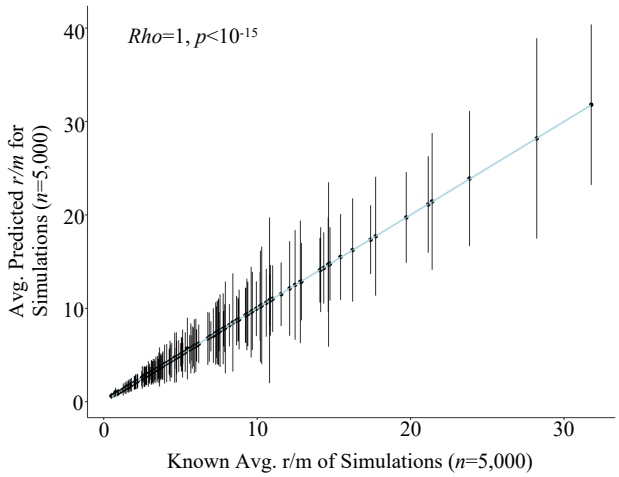


Figure S II—2. Plot of the average of the posterior distribution of simulated summary statistic values vs. real summary statistic values *before* outlier removal (A-C) and *after* outlier removal (D-F) for *h/m* (red), π (green), and LD_{fit} (yellow) calculated from the species alignment. Spearman's correlation coefficients and *P*-values are indicated above each graph.

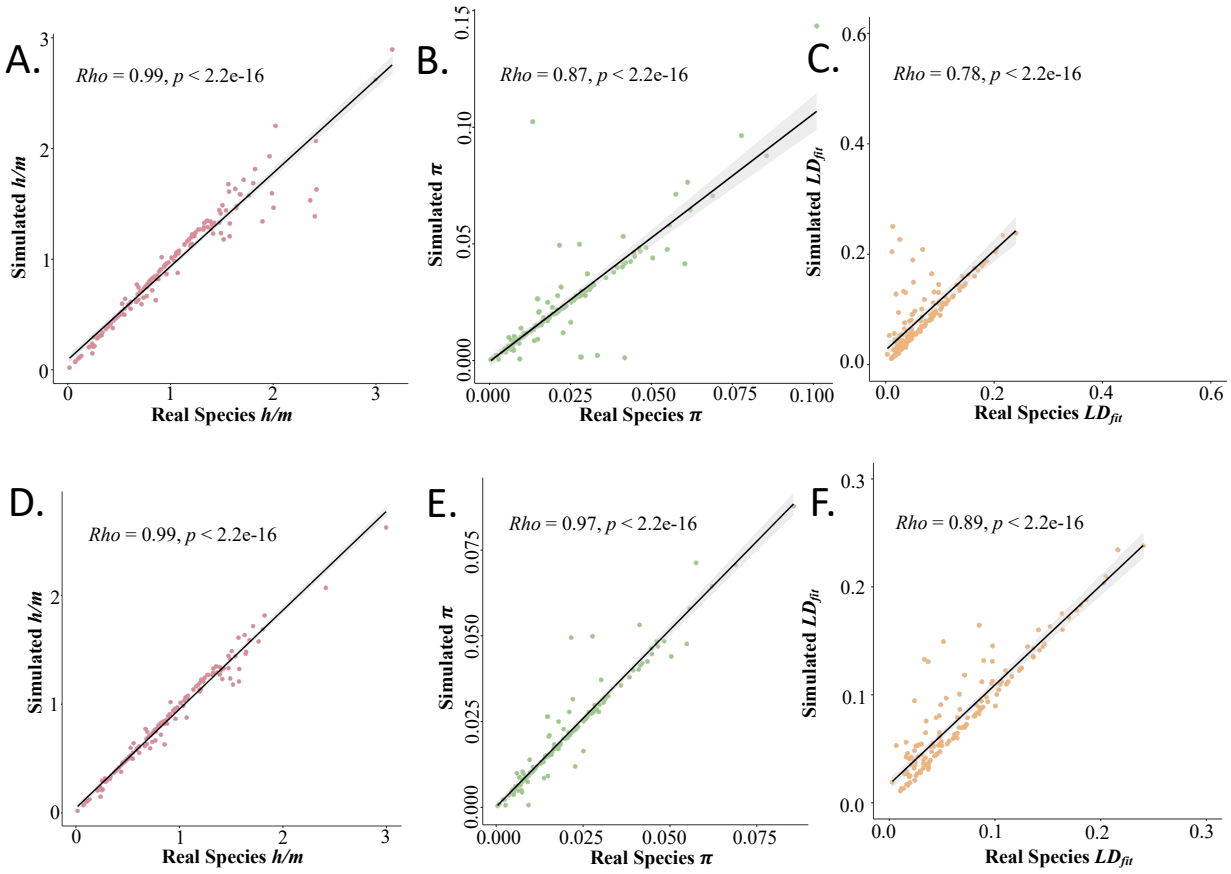


Figure S II—3. Recombination rate estimates (r/m) for each species relative to the summary statistics of each species (real dataset) A) h/m, B) LDfit, and C) π. Spearman's correlation coefficient and P-values are listed in the upper left corner of each plot.

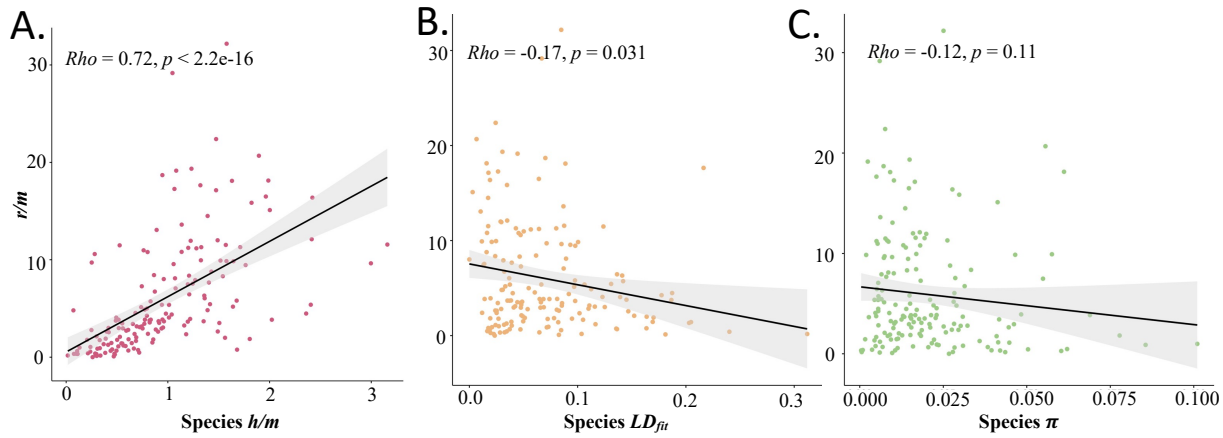


Figure S II—4. Recombination rate estimates relative to various genomic characteristics of the dataset (A-F). Each graph compares recombination rate (*r/m*) *vs*. A) the number of strains included in each species analysis for 162 species with between 15 and 100 strains, B) the length of the core genome (b.p.) for the 163 species in this analysis, C) the number of core genes for the 163 species in this analysis. the average core genome nucleotide pairwise identity for the 163 species in this analysis, E) the standard deviation of the average nucleotide pairwise identity across the core genomes of each species, and F) the average GC content across the core genome for the 163 species in this analysis. Spearman's correlation coefficient and *P*-values are listed in the upper left corner of each plot.

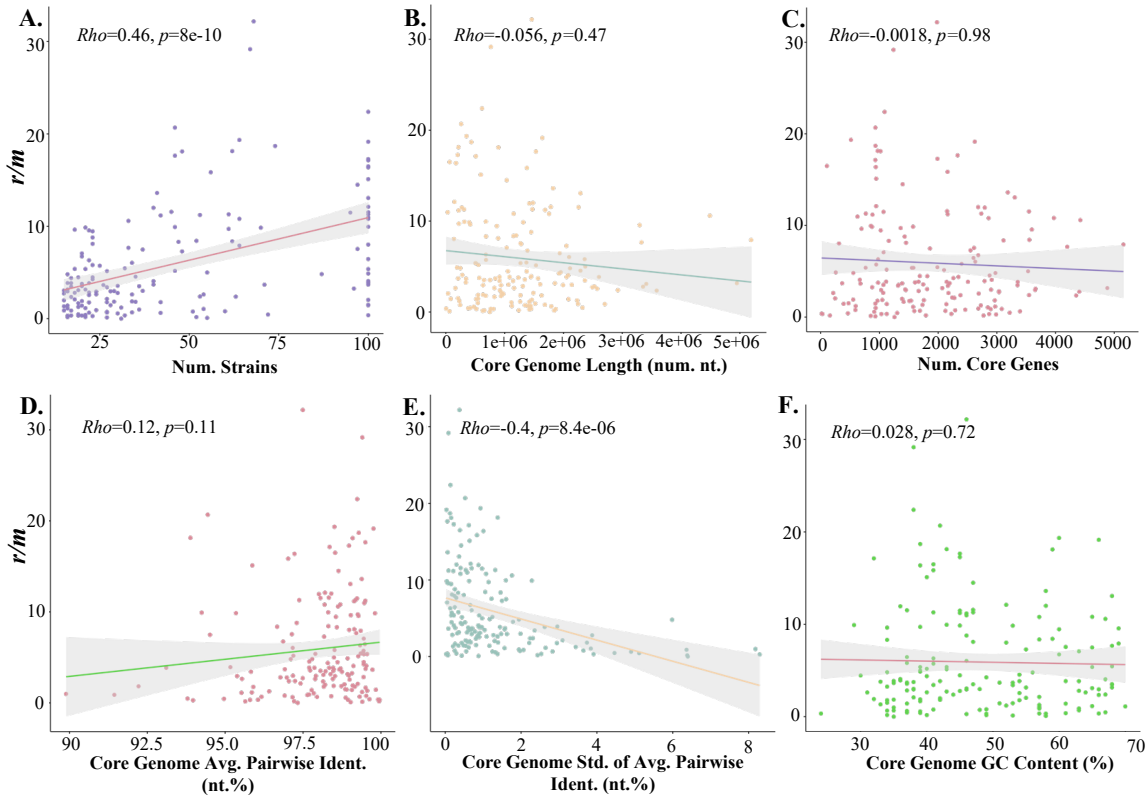


Figure S II—5. Comparison of recombination rate estimates (*r/m*) to metabolic, environmental, and physiological traits predicted from the JGI GOLD database (97) **(Supplementary Table 1: Tab F). Comparisons were made between *r/m* across A) autotrophic and heterotrophic species, B) gram– and gram+ species, C) aerobic, anaerobic, and facultatively aerobic species, D) motile and nonmotile species, E) free-living and obligately intracellular species, and F) sporulating and nonsporulating species. The number of species with data available in JGI GOLD for each comparison is denoted below the category label in each plot. Only the median difference in recombination between free-living and obligate intracellular species was found to be significant (Wilcoxon rank-sum test: *P*=0.03)** (97)**.**

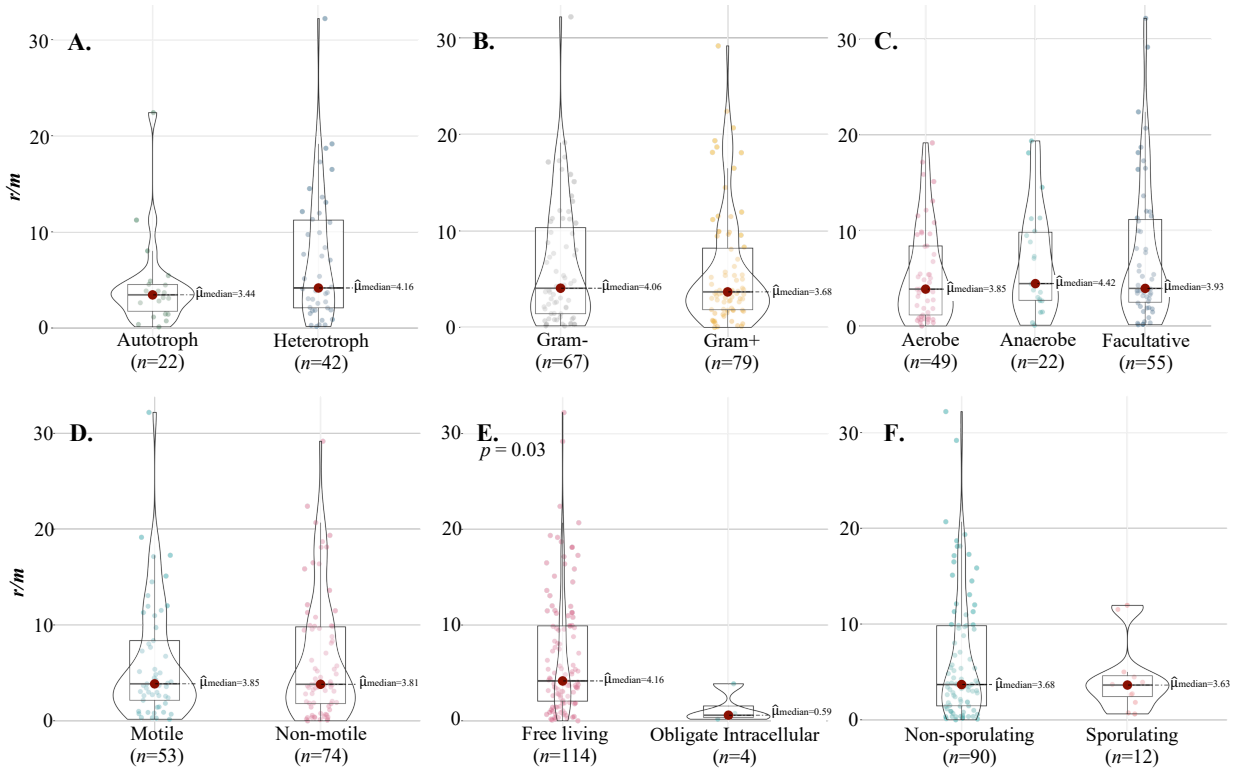


Figure S II—6. Comparison between recombination rate estimates (*r/m*) and the number of integrated viral sequences (prophages). The program *geNomad* (37) was used to identify the number of putative viral coding sequences (A) and the putative complete prophages (B) in each reference sequence from the 159 species in our dataset.

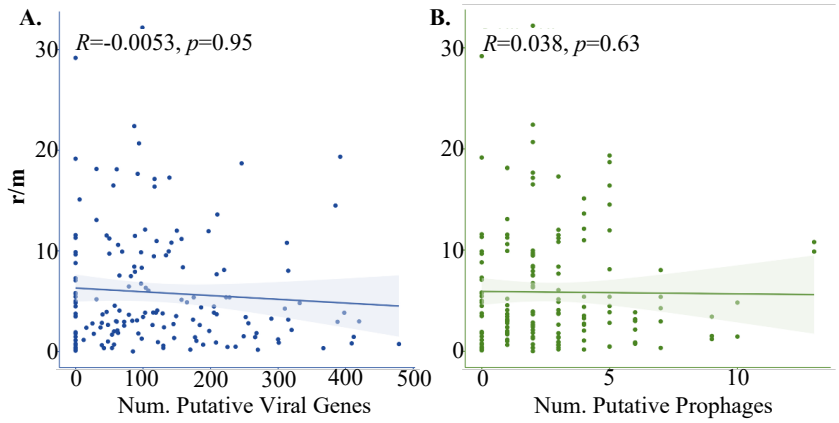


Figure S II—7. Robustness of *r/m* estimates to genome subsampling. The *y*-axis represents the average recombination rate for 100 random subsamples of *x*-10 genomes per species (where *x* is the total number of genomes present for a given species in the analysis) and *x*-axis represents the recombination rate estimate inferred for that species (*n*=101 species with ≥25 strains).

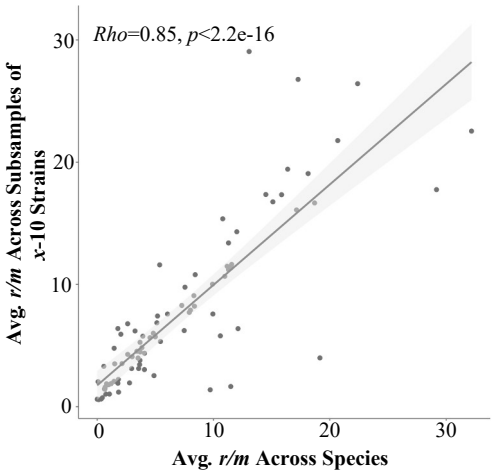


Figure S II—8. Correlation between recombination rates (*r/m*) predicted by this study (*y*-axis) and those predicted by *ClonalFrameML* (*x*-axis) (8) across different samples of *Escherichia coli*'s genomes (*n*=20). The black line on the plot represents *y=x*.

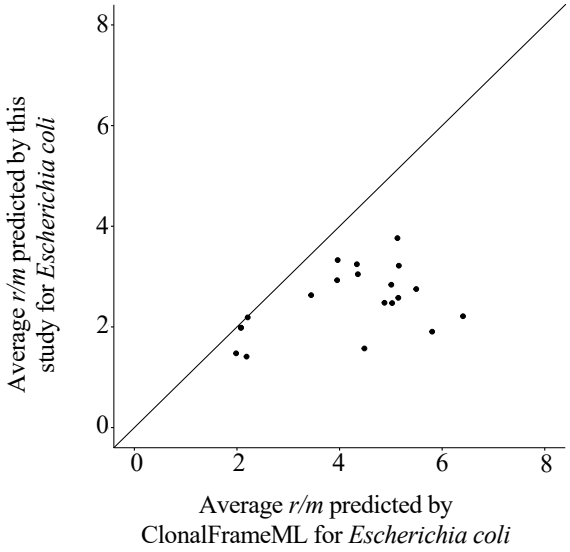


Figure S II—9. Recombination rates (*r/m)* estimated with our ABC approach (*y*-axis) relative to recombination rates estimated by *ClonalFrameML* (*x*-axis) (45) **for 84 bacterial species. The black line on the plot represents *y=x*.**

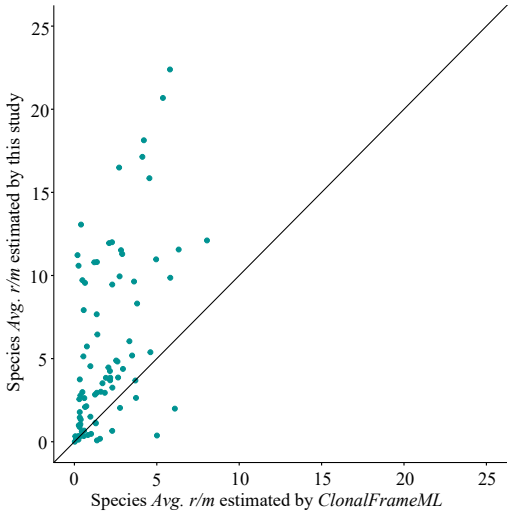


Figure S II—10. Comparison of recombination rates between pathogens, non-pathogens, and putative pathogens with a Kruskal-Wallis test. Data compiled in Bartlett et al. (2022) was used to infer pathogen status (114)**. When species in our dataset were not present in their dataset, a separate search was conducted on google scholar to determine whether publications supporting the bacteria's classification as a pathogen could be inferred** (114–121, 123–126)**. Tab F of Supplementary Table 1 contains the table as in Bartlett et al. as well as additional citation information appended by this study.**

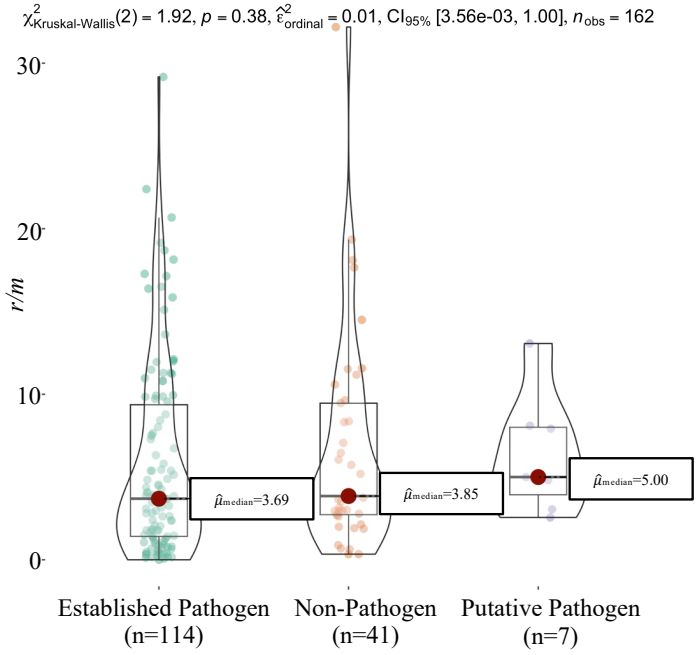
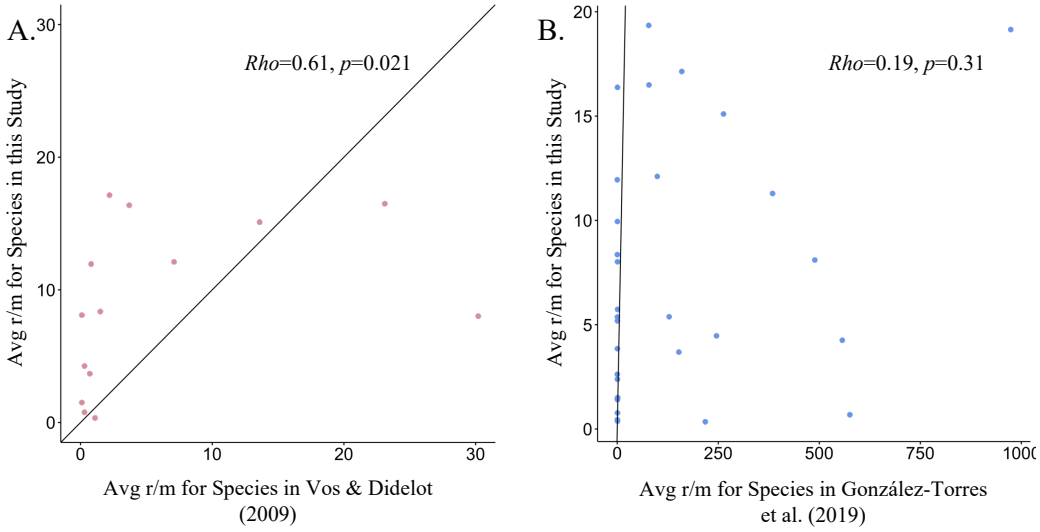


Figure S II—11. Comparison between recombination rate estimates (*r/m*) from this study and estimates from two other studies. A) Recombination rates estimated from MLST data using *ClonalFrame* (1), where 14 species were common to our study. B) Recombination rates estimated using whole genome content in (2) where 29 species were shared with our study. The black line on each graph denotes *y=x*. Spearman's correlation coefficients and *P*-values are shown in the upper portion of the graphs.



II.7.3 Supplementary Table Legends

DISCLAIMER: The supplementary tables are too large to be included in this document and so only their legends are included below. These tables will be included with the peer-reviewed manuscripts after their publication.

Table S II—1. Comprehensive table including datasets produced by and relevant to this study. Tab A) List of all 162 bacterial and one archaeal species (corresponding to 7,451 complete genomes) analyzed in this study where asterisks in column 2 indicate lack of confidence in simulating one or more summary statistic (*,,***). The taxonomic and genomic information included is as follows: column 3: core gene number, column 4: number of strains, column 5: core genome length (bp), column 6: average core genome GC-content (%), column 7: average core genome pairwise sequence identity, column 8: standard deviation of average core genome pairwise identity, column 9: combined GenBank ID and assembly accession number for the reference genome chosen for each species, column 10: the total number of genes in the reference genome, column 11: the total genome length of the reference genome (bp), column 12: the number of assembly contigs in the reference genome, column 13-16: real species values for the three summary statistics: theta, π, LDfit, and h/m, column 17: the total number of simulations conducted for each species, column 18: the number of simulations in the ABC posterior dataset with a tolerance of 0.01%, column 19-36: and the average, median, and standard deviation of the branch length coefficient, r/m, h/m, π, LDfit, and delta x rho x theta (i.e. the recombination rate expressed as the total number of nucleotides exchange via recombination relative to mutation) generated from the ABC posterior dataset for each species. Tab B) List of recombination rate estimates (r/m) computed across different replicated of E. coli**

phylogroups by both our method (column 2) and ClonalFrameML (column 3) (8). The combined averages and standard deviations of r/m computed by both methods are shown below the table. Tab C) List of 84 species with average r/m estimated by both ClonalFrameML (column 2) (45) **and our method (column 3) as well as the absolute difference in r/m between estimates (column 4). The average and standard deviation of the three columns is shown at the bottom of the table. Tab D) A comparison of homologous recombination rate estimated for species in this study (column 2) to those which were estimated for the same species in Vos & Didelot (2009)** (1) **(column 3) and Gonzalez-torres et al. (2019)** (2) **(column 1). The table includes the species name the r/m estimate reported in Gonzalez-torres et al. (2019), the r/m estimate reported in this study, and the r/m estimate reported by Vos & Didelot (2009). 'NA' denotes species for which no r/m data was available in the study. Tab E) List of ecological and metabolic data for 162 bacterial and one archaeal species used in this analysis imported from the Joint Genome Institute's Genomes Online Database (GOLD)** (97)**. The columns for oxygen tolerance (column 15: "O2"), energy metabolism (column 17: "energy"), gram staining specificity (column 18: "gram"), cell shape category (column 31: "cell_shape"), cell motility (column 33: "cell_motility"), and whether the species was capable of sporulation (column 35: "sporulation") were compared to recombination rate estimates as reported in Supplementary Figure 4. Tab F) A List of 162 bacteria in this study, their taxonomic classifiers, and their classification as putative pathogens or pathogens of humans as reported in Barlett et al., 2022. Data not included in Barlett el al. (2022) is denoted with 'NA' for the pathogenicity column (column 8) and characterization as a pathogen was separately investigated in this analysis and its status was listed in column 11 and**

supporting citation in column 12 (114–121, 123–126)**. Data from this study was compared to recombination rate to assess putative relationship between human pathogenicity and recombination rate as reported in Supplementary Figure 5. Tab G) Tabulated results of prophage prediction on each reference genome from 159 species. Table contains species name (column 1), the average recombination rate for the species (column 2), the number of complete putative prophages identified by geNomad (column 3) and the number of putative viral coding regions identified by geNomad (column 4)** (102)**. Tab H) Table of data from Supplementary Figure 7. The table contains the species name from which the simulated dataset originated (column 1), the average of the known recombination rate across the 5,000 randomly chosen simulations (column 2), the predicted average recombination rate across the average of each posterior dataset generated by ABC (n=5,000) (column 3), and the standard deviation of the predicted recombination rate based on the posterior distribution generated by ABC (n=5,000) (column 4). Tab I) Assessment of our ABC approach by predicting recombination rate from a set of 5,000 simulations where recombination rate was known for each population. Each row denotes i) the species from which the simulated dataset originated, ii) the Pearson's correlation coefficient and iii) the P-value of the Pearson correlation between the average of all predicted r/m values of the simulated dataset and the known average r/m across the simulated dataset (Table1 I), iv--vi) the average known summary statistic values (h/m, π, LDfit) across the simulated dataset (n=5,000), vii) the average predicted r/m across all posterior populations for the 5,000 simulated datasets (n=5,000), viii-x) and the average predicted summary statistics (h/m, π, LDfit) across all posterior populations for the 5,000 simulated datasets (n=5,000). Tab J) Results of the subsampling analysis. Ten genomes were randomly discarded for each**

species and recombination rate was re-estimated (100 replicates). Included are the species names, the average r/m estimated for the species, the average r/m estimated across the 100 replicates of subsamples of n-10 genomes and the difference in r/m between the two estimates. The final column contains asterisks denoting species whose r/m estimated from subpopulations varied by more than five from the species population.

Table S II—2. List of genome IDs included in this analysis for each species. File contains the species used in this study (column 1), the strain chosen as the reference sequence (column 2), and a list of the combined Genbank ID and assembly ID for each strain in the species analysis (column 3).

Table S II—3. Tab A) Collection data for 7,654 bacterial genomes as listed in the GenBank Feature Format (GFF) files. The table includes the species names, their concatenated GenBank genome IDs and assembly accession numbers, GenBank feature collection information, and coded category summarizing the GenBank collection information. Tab B) Spearman correlations between *r/m* estimates and the number of genomes within a species from each environmental category based on their available GenBank collection data. *P*-values were adjusted for multiple with Holm adjustment.

Table S II—4. Comparison between recombination rate estimates (r/m) and the percentage of genes in a COG or CAzy category (98–100)**. Tab A: The table includes the Spearman's correlation coefficient and p-value for the comparison between r/m and the percentage of genes in each COG or CAzy category as annotated by *eggNOG*** (98–100)**. Tab B: The table**

contains the species name, its average r/m values, and the percentage of genes in each COG or CAzy category as annotated by *eggNOG* (98–100)**.**

II.8 Associated Contents

II.8.1 Ethics Approval and Consent to Participate

Not applicable.

II.8.2 Consent for Publication

Not applicable.

II.8.3 Availability of Data and Materials

All data used in this analysis was downloaded from NCBI's Genbank public genomic repository. GenBank accession numbers and assembly IDs for all analyzed genomes are detailed in Supplementary Table 2. The pipeline *recABC* used to generate recombination rates is available at (https://github.com/lbobay/recABC). The datasets generated in this study (the core genomes, phylogenetic trees and the summary statistics of all the simulations for all species) are available on *Kaggle* at https://www.kaggle.com/datasets/ellistorr/bacterial-rm

II.8.4 Competing interests

The authors declare that they have no competing interests.

II.8.5 Funding Information

This study was supported by the National Institutes of Health grant R01GM132137 awarded to LMB and by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0021110 awarded to ELT.

II.8.6 Funding Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

II.8.7 **Authors' contributions**

E.L.T., A.D., C.B., and L-M. B. designed, performed research, and analyzed data in this publication. E.L.T. and L-M.B. wrote this paper and A.D., and C.B. reviewed and edited it.

II.8.8 **Acknowledgments**

We would like to thank Shaw Kagawa, LeRayah Neely-Brown and Cory Schantz for testing the early version of *recABC*. We are thankful to Daniel Schrider for providing advice and Kasie Raymann and Gavin Douglas for reading the manuscript. We would like to dedicate this manuscript to Matthew Miller.

II.9 **References**

See REFERENCES on page 130.

CHAPTER III: HOMOLOGOUS RECOMBINATION SHAPES THE ARCHITECTURE AND EVOLUTION OF BACTERIAL GENOMES

Ellis L. Torrance[1], Awa Diop[2], Louis-Marie Bobay[1,2]

[1] Biology Department, University of North Carolina at Greensboro, Greensboro, NC 27412 USA

[2] Department of Biological Sciences, North Carolina State University, Raleigh, NC

III.1 Abstract

Homologous recombination is a significant evolutionary force that varies tremendously across bacterial species. However, how the landscape of homologous recombination varies across genes and within individual genomes has only been studied in a few species. Here, we used Approximate Bayesian Computation to estimate the recombination rate along the genomes of 145 bacterial species and one archaeon. Our results show that homologous recombination is a key force shaping many aspects of bacterial genome architecture and its evolution. We find that recombination rates vary greatly along genomes and that these patterns are not random. The genomic landscape of recombination presents several key signatures: rates are highest near the origin of replication in most species, patterns of recombination appear symmetrical in both replichores and most species have genomic hotpots of recombination. Furthermore, many closely related species share conserved landscapes of recombination across orthologs indicating that recombination landscapes are conserved over significant evolutionary distances. We show clear evidence that recombination drives the evolution of GC-content through increasing the effectiveness of selection and not through biased gene conversion, thereby solving an ongoing debate. Finally, we show that the rate of recombination varies across gene function and that

many hotspots of recombination are associated with adaptive and transferable regions often encoding genes involved in pathogenicity.

III.2 Introduction

Homologous recombination is a major force shaping genome evolution across bacteria. This mechanism promotes the exchange of alleles between homologous sequences akin to the process of gene conversion in Eukaryotes. The rates of recombination vary extensively across species, but several studies have shown that recombination rates vary across genomic regions as well (14, 138). In addition to the overall fluctuations of recombination rates, some genomic regions present very low rates of recombination (*i.e.*, coldspots) while others show particularly high rates of recombination (*i.e.*, hotspots). The mechanisms shaping these variations in recombination rate along the genome are not known. Some regions might be more recombinogenic by presenting an easier access for incoming DNA or by containing sequence motifs that stimulate recombination (*e.g.*, Chi motifs). Alternatively, selection may be shaping the patterns of recombination across genomic regions. It has been hypothesized that hotspots of recombination play an important role in the ability of bacteria to adapt to selective pressures from their environment. However, the patterns of homologous recombination have only been characterized in the genome of a few species. Thus, it remains largely unknow how variations in homologous recombination contribute to shape the evolution and adaptation of bacterial genomes. In addition, we ignore to what extent these genomic landscapes of recombination are conserved after species diverge from one another.

In Eukaryotes, the genomic landscapes of recombination have been extensively studied and hotspots of recombination have been found to be associated with disease phenotypes and adaptation (139). In bacteria however, variations of recombination rate across the genome have

not been thoroughly described (81, 83, 107). Previous studies have shown that signatures of recombination are lower in core genes with housekeeping functions and higher in genes associated with virulence and defense functions (2, 81, 83, 107). Interestingly, hotspots of recombination have been observed to be flanking mobile elements such as SCC (Staphylococcal cassette chromosome) which is associated with methicillin resistance in *Staphylococcus aureus* (MRSA) (81, 107, 140). In addition, hotspots of recombination have also been found near the origin of replication (Ori) in *S. aureus*, but this trend has not been observed in other species (81, 107). Moreover, genes flanking mobile elements and clusters of accessory genes have been estimated to recombine twice more frequently than non-flanking genes across bacterial species (13).

Bacterial chromosomes present various levels of organization, and those can impose mechanistic and selective constraints on recombination. Replication proceeds bidirectionally starting at the Ori (69). In circular chromosomes, the two replichores present similar lengths and progress synchronously from Ori to the terminus of replication (Ter). A mutation accumulation study has shown that mutation frequency follows a wave-like pattern, symmetrical in the two replichores, which has been hypothesized to be the result of replication timing and its interruption (141). Because DNA synthesis proceeds necessarily from 5' to 3', one strand is synthetized continuously in the same direction as the replication fork (*i.e.*, the leading strand), while the other strand is synthetized discontinuously (*i.e.*, the lagging strand). Thus, the initiation of DNA synthesis is delayed for the lagging strand and its template strand remains single-stranded for longer periods of time. Due to the higher mutagenic nature of single-stranded DNA, it has been suggested that this asymmetry in DNA replication cause higher mutation rates on the lagging strand relative to the leading strand (142). However, it has been shown that this

asymmetry could in fact be driven by the stronger selective constraints acting on the leading strand, due to the higher prevalence of essential and highly expressed genes on this strand. Finally, genes and operons are often observed as clusters of accessory genes such as pathogenicity islands, which are often horizontally exchanged between strains (69, 143). The insertion of these transferred clusters of accessory genes can be mediated by non-homologous recombination, site-specific recombination, or by homologous recombination at flanking core genes (144–146).

Here, we leveraged a new approach to estimate the rates of homologous recombination along bacterial genomes. We adapted our tool *recABC* (see Chapter 2) to estimate the recombination rate of thousands of individual core genes across the genomes of 145 bacterial species and one archaeon. We compared the rates of recombination in the context of their chromosomal location and functions to determine whether homologous recombination rate varies with any appreciable patterns across species. Using a robust dataset of >200k core genes, we observed that homologous recombination varies extensively across the genome of most bacteria. We further found evidence that these variations are linked to gene function and chromosome structure, indicating that selection is likely shaping patterns of recombination. We detected the presence of many hotspots and coldspots of recombination across species. We did not identify any strand-specific biases in homologous recombination or any strong indication that recombination rate is elevated in genes directly flanking accessory regions. However, we did observe a significant relationship between homologous recombination and GC content as well as signatures of selection.

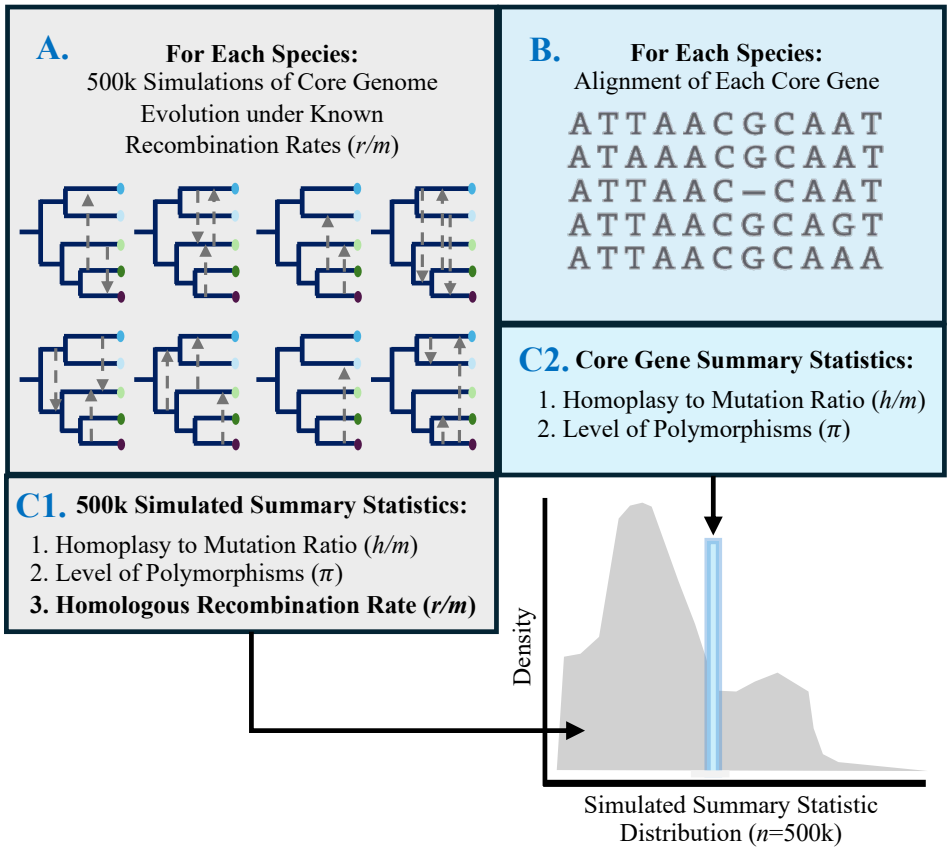
III.3 Results

III.3.1 Homologous Recombination Rate Varies across Bacterial Core Genes

In a recent study, we developed and applied a novel approach to estimate recombination rate across bacterial using Approximate Bayesian Computation (ABC) (see Chapter 2). Briefly, our tool generates forward-in-time simulations of genome populations evolved under various rates of recombination (n=500k simulations per species) and our approach selects the simulations that best recapitulates genomic signatures of recombination observed in the real genomes through comparison of summary statistics with ABC. We utilized the summary statistics from individual core gene alignments to estimate recombination rate for that gene from the simulated population of sequences (see Fig. 1). For each core gene, we measured the effective rate of recombination r/m, which represents the number of times alleles has been exchanged by recombination (r) relative to the number of alleles introduced by mutation (m). Using this approach, we inferred the recombination rate of each core gene across all the core genomes (*i.e.*, the set of genes shared by nearly all the members of a species) of our dataset: 162 bacterial species and one archaeon. Each species was composed of between 15-100 non-redundant genomes and genomes were classified to a species using both the average nucleotide identity (ANI) and the patterns of gene flow (4). The core genes whose summary statistics could not be robustly used for inference of recombination rate (*i.e.*, that fell outside of the distribution of simulated summary statistics) were excluded from the analysis (see Methods). We further excluded the species whose core genome was highly reduced (<200 core genes) after excluding those genes. After applying these criteria, our final dataset was composed of a total of 145 bacterial and one archaeal species, for which recombination rates of individual core genes could be confidently inferred. We found that the average rate of recombination across core genes

within each species was tightly correlated to the species *r/m* previously inferred using the whole core genome (Spearman's *Rho*=0.91, $P<10^{-15}$: Supp. Fig. 1). The number of core genes for which a robust estimate of *r/m* could be inferred varied from a minimum of 218 for *Salmonella enterica* to a maximum of 4,161 for *Burkholderia gladioli.* In total, the *r/m* values were estimated for 208,686 core genes. The data summarizing the size of the core gene dataset, the distribution of *r/m* and summary statistic values for species included in this analysis are presented Supp. Table 1.

Figure III—1. Description of the method used to infer rates of homologous recombination (*recABC*) across the core genome for 145 bacteria and one archaeal species in this study. A) For each species, the core genome is aligned and concatenated and a phylogeny is built from the alignment of the core genome. Using the nucleotide length and GC content of the core genome alignment from the real species, a single ancestral genome randomly generated to initiate the ancestral genome of each simulation. This ancestral genome is then evolved in a forward-in-time simulation following the phylogenetic topology of the real species under varied recombination rates (*Rho*) and recombination tract lengths (*delta*) using *CoreSimul* (96). The corresponding effective recombination rate *r/m* is also computed during the simulations. A total of 500,000 simulations are generated for each species. B) An alignment is generated for each core gene within each species. C1) Two summary statistics are computed: *i*) the ratio of homoplasy to mutation (*h/m*), and *ii*) the average nucleotide diversity (*Pi*) computed for the core genome alignment of the real species. C2) The same summary statistics (*h/m, pi*) are calculated for each of the 500k simulations. D) Approximate Bayesian Computation (ABC) is then used to compare the summary statistics from the real species gene data to the distribution of the same statistics generated by simulation under known recombination rates (the prior distribution in grey). The simulations with statistics which most closely match the summary statistics from the real species are selected using ABC (the posterior distribution in blue) with a tolerance threshold of 0.01% (*n*=50). The average rate of recombination of the posterior distribution is then used as an estimate of recombination rate for each core gene in each species.



D. **Approximate Bayesian Computation** (ABC) used to compare real and simulated summary statistics (1-2) to infer the most probable rate of recombination (*r/m*) for each core gene.

III.3.2 **Homologous Recombination Rate Variation by Gene Function**

We first tested whether rates of homologous recombination vary across gene functional categories (*i.e.,* clusters of orthologous genes (COG)), all genes were classified into COG categories using *eggNOG* (*n*=208,686 core genes) (98, 99). Overall, a significant difference in *r/m* values across COG categories was detected (Kruskal-Wallis, $P<10^{-22}$) and recombination rates were then compared for each relevant functional category independently (18 COGs in total) using a Wilcoxon Test with Benjamini-Hochberg adjustment (Supp. Fig 2A). Significantly lower *r/m* values were observed for functional categories coding for central cellular functions: *i*) cell cycle control, cell division, and chromosome partitioning (COG category D, $P<10^{-5}$), *ii*) translation, ribosomal structure, and biogenesis (COG category J, $P<10^{-76}$), and *iii*) transcription (COG category K, $P<0.05$). In contrast, significantly higher *r/m* values were observed for genes encoding more diverse functional categories: *i*) energy production and conversion (COG category C, $P<10^{-4}$), *ii*) amino acid transport and metabolism (COG category E, $P=10^{-12}$), *iii*) carbohydrate transport and metabolism (COG category G, $P= P<10^{-5}$), *iv*) coenzyme metabolism and transport (COG category H, $P<10^{-3}$), *v*) cell wall/membrane/envelope biogenesis (COG category M, $P<10^{-6}$), *vii*) inorganic ion transport and metabolism (COG category P, $P<10^{-7}$), *viii*) signal transduction mechanisms (COG category T, $P<10^{-3}$) and *ix*) defense mechanisms (COG category V, $P<10^{-12}$).

To determine to what extent recombination rate varied across gene functions at the species level, the same test was performed within each species individually (*n*=146). As expected, few significant relationships were observed, which is likely due to the decrease in statistical power. Nevertheless, a significantly lower recombination rate (Benjamini-Hochberg adjusted $P<0.05$) was observed for genes belonging to the COG category J for 43 species

(translation, ribosomal structure, and biogenesis, 29% of species) and K for 7 species (transcription, 5% of species). In contrast, a significantly higher rate of recombination was observed for genes belonging to the COG category E for 11 species (amino acid transport and metabolism, 8% of the species). All other categories had less than ≤5 species with significant trends. A bar graph showing the number of species which was significant for each COG group is shown in Supp. Fig. 2B.

III.3.3 Homologous Recombination Rate in Genes Flanking Clusters of Accessory Genes

Accessory genes (*i.e*., genes that are typically found in a single of in a few genomes) are often transferred by Horizontal Gene Transfer (HGT) and the insertion of these sequences can be mediated by homologous recombination and other mechanisms (7). In prior studies, the rate of homologous recombination has been estimated to be higher in the regions flanking horizontally transferred accessory genes (13, 84, 112). However, only Oliveira *et al*. (2017) analyzed this trend over multiple species. Here, the authors quantified recombination rates by calculating both the number of estimated recombination events and the amount of phylogenetic incongruencies in core genes flanking accessory gene regions (13). To test whether we observed similar trends in our dataset, we defined accessory gene clusters as strings of ≥5 consecutive accessory genes and we compared *r/m* estimates of core genes flanking these regions versus *r/m* estimated of the non-flanking core genes (*i.e.* core genes not directly located next to accessory gene regions) for each species (Supp. Table 2). We found that the majority of species had an increase in recombination rate in flanking core genes relative to non-flanking core genes (98, 68%). However, the increase was only significant for nine species (*Bacillus megaterium, B. wiedmannii, B. cereus, Staphylococcus warneri, S. equorum, Lactobacillus kunkeei, Burkholderia multivorans*, *B.*

vietnamiensis, Bifidobacterium adolescentis) (Benjamini-Hochberg adjusted $P<0.05$) and no species were found to have a statistically significant decrease. The increase in *r/m* in the core genes flanking accessory regions was overall relatively small: we measured an average increase in *r/m* across species of 0.47 ± 0.64.

III.3.4 Homologous Recombination Rate, GC-content, and Selection

Previous studies have reported a positive relationship between the rate of recombination and the GC-content of gene sequences. This result has been interpreted as evidence that recombination can enhance genomic GC-content *i*) by a mechanistic bias during the recombination process (the biased gene conversion hypothesis) (147) or *ii*) by increasing the effectiveness of selection (the selection hypothesis) (135, 148). We compared our estimates of recombination rates to the average GC-content estimated across all the sequences of each core gene using a Spearman's correlation test with Benjamini-Hochberg *P*-value adjustment for all 146 species (Supp. Table 2). As reported in previous studies, we observed that the relationship between *r/m* and GC-content was positive in most species ($Rho>0$, $n=119$, (82%). The relationship between *r/m* and GC% was significant for 83 (57%) of these species, among which nine had a significant negative correlation (6%) and the remaining 74 showed a significant positive correlation (50%). The distribution of correlation coefficients across significant species comparisons is shown in Supp. Fig. 3 (Average $Rho=0.08\pm0.1$).

To determine whether a relationship was observed between *r/m* and the impact of selection, we estimated the ratio of non-synonymous to synonymous substitution rates (*dN/dS*) using PAML for each core gene alignment and for each species (see Methods). Species were included in this analysis if *r/m* and *dN/dS* could be both estimated for at least 200 core genes ($n=142$). We found a significant correlation between *r/m* and *dN/dS* for 73 species (51%)

(Spearman's Benjamini-Hochberg corrected *P*-Value<0.05). Of those that were significant, 14 showed a positive correlation and the remaining 81% (*n*=59) presented a negative correlation between *r/m* and *dN/dS*. These results indicate that genes with higher recombination rates are evolving under stronger selective pressures. Separately, we observed that *r/m* and *dN* were significantly corelated for 114 species with a positive significant relationship for 67% (*n*=79) and *r/m* vs. *dS* were significantly correlated for 123 species with a positive significant relationship for 80% of species (*Rho*>0.0, *n*=98). These results further indicate that the correlation between *r/m* and *dN/dS* is not solely driven by *dN* or *dS*. Results for each of these tests comparing r/m to *dN*, *dS*, and *dN/dS* across core genes for each species are available in Supp. Table 3.

We further tested whether a relationship existed between *dN/dS* and GC-content across the core genes of the 142 species (Supp. Table 4). Indeed, we observed a significant relationship between *dN/dS* and GC% across 113 species (77%) (Benjamini-Hochberg adjusted *P*<0.05). Of these, 68% of species had a significant negative relationship between *dN/dS* and GC% (*n*=77, *Rho*<0.0) and the remaining 32% (*n*=36) had a significant positive relationship. Species with a significant positive relationship tended to be species with high GC content such as *Burkholderia sp.* and *Mycobacterium sp.*

III.3.5 **Homologous Recombination Rate and DNA Strand Bias**

To determine whether the correlations between *r/m* and GC-content were related to strand bias— and to evaluate the hypothesis that homologous recombination should be more prevalent on the lagging strand due to increased prevalence of head-on collision between replication and transcriptional machinery and subsequent strand repair through homologous recombination (149)—we compared the *r/m* between genes present on the leading and lagging strands using a Wilcoxon Test with Benjamini-Hochberg *P*-value adjustment for *n*=102 species (see

supplementary methods for determination of leading vs. lagging strand). We found little evidence supporting a difference in recombination rate between the leading and lagging strands. Of the 102 species assessed, only seven species presented a significant difference in recombination rate across core genes between strands (Supp. Table 5). Of those seven, four species had a significant increase in their lagging strand *r/m* (*Pseudomonas chlororaphis*, *Lactobacillus fermentum*, *Staphylococcus saprophyticus*, and *L. kunkeei*) and three species presented the reverse trend (*Yersinia intermedia*, *Serratia marscescens*, and *Lactobacillus salivarus*). Notably, this difference in *r/m* between strands was small (Difference in *r/m* average between leading and lagging strands ≤ 1) for all species. Overall, *r/m* is very similar between the core genes of the leading and lagging strands across bacterial species.

III.3.6 **Evolution of the Genomic Landscape of Recombination**

Here we tested whether the patterns of recombination rates were conserved between species following speciation and divergence. We compared the genomic patterns of recombination between all pairs of species within each genus, which represented 109 unique pairs of species. For each pair, recombination rates were compared between shared orthologs to determine whether the patterns of recombination rate were a conserved trait between closely related species (Supp. Table 6). We found a positive significant correlation between the recombination rates of the shared orthologs for 36 species pairs (Spearman's Benjamini-Hochberg adjusted $P<0.05$), indicating that genomic landscapes of recombination are somewhat conserved between related species. Conservation in recombination rate was then compared to pairwise divergence between species of the pair. We found that the species with evidence of conservation in *r/m* across shared orthologs had lower pairwise divergence (average A.A. pairwise divergence = 0.25 ± 0.18) versus those without evidence of conservation in *r/m* (average

A.A. pairwise divergence = 0.37±0.30) (Supp. Figure 4). This result further supports that the genomic patterns of recombination are conserved over short evolutionary distance.

III.3.7 Genomic Landscapes of Homologous Recombination

Bacterial chromosomes are highly organized entities, and these constrains can shape recombination rates across genomes. We first tested whether recombination rates varied between the origin of replication origin (Ori) and the terminus (Ter) for the 102 species with circular chromosomes for which Ori and Ter could be confidently identified based on GC-skew (see Methods) (Supp. Table 7). We found that *r/m* was higher near Ori for 67% of species (*n*=66). Correlations between *r/m* and distance to Ori were statistically significant for 36 species (Spearman's Benjamini-Hochberg adjusted $P < 0.05$). Here, the majority (*n*=26, 72%) displayed higher rates of recombination near Ori whereas the inverse (*r/m* is lower near the Ori) was observed as significant in 10 species (28%). These results indicate that bacteria present an overall bias of increased recombination near Ori.

We further tested whether recombination rates were symmetrical across both replichores. Core gene sets were arbitrarily bisected at the Ori-Ter axis into "right" and "left" replichores. For each replichore, we correlated the rate of recombination of the core genes relative to their positions in the Ori-Ter axis. Interestingly, we found that the absolute value of Spearman's *Rho* between both replichores were very similar between the two replichores of most species (*Rho*=0.4, $P<10^{-4}$) supporting the evidence of symmetry in the patterns of recombination rates in the two replichores (Supp. Fig. 5).

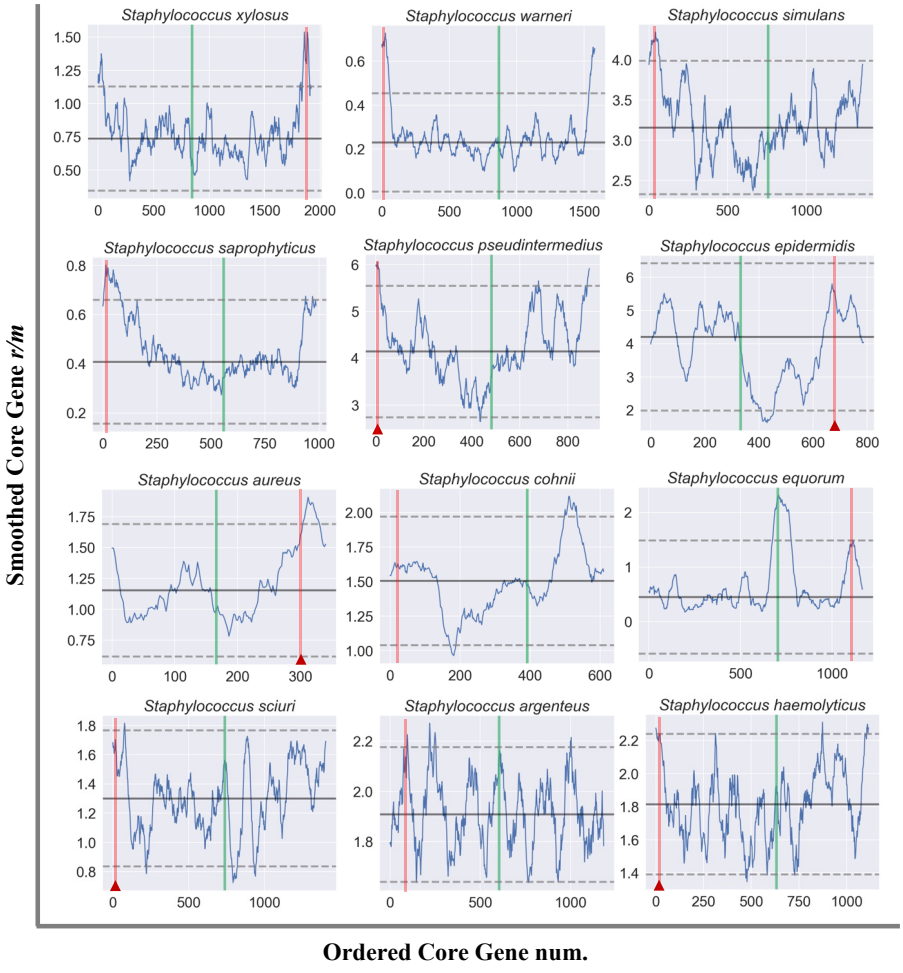
Previous studies have found that recombination rates vary extensively across bacterial chromosomes in species such as *Staphylococcus aureus*, *Streptococcus pyogenes*, and *Campylobacter jejuni* with little visible pattern for most species except for anomalous spikes, or

"hotspots" of recombination (81, 83, 107, 112). Hotspots of recombination are of particular interest as they have previously been found to correspond to the presence of genes associated with potentially adaptive traits such as virulence or antibiotic resistance (81, 107). As mentioned above, and in accordance with other studies, we observed that rates of recombination are significantly elevated in genes associated with adaptive traits such as cellular defense among others (see Supp. Fig. 2) (2, 107). To visualize the patterns of recombination along bacterial genomes, we plotted the average of recombination rates across the core genome of each species (n=102) using a sliding window (starting at the origin of replication (x=0) and the relative location of the terminus plotted as a line in green) (Supplementary Folder 1). We defined hotspots and coldspots (regions of anomalously low recombination) as regions where local *r/m* differed by more than two standard deviations from the mean across the genome. Overall, we detected hotspots and coldspots (regions of anomalously low recombination) across species with most species having at least one hot- and coldspot as defined by this metric. In accordance with prior studies; coldspots primarily contained housekeeping genes, whereas hotspots were enriched in genes that are potentially adaptive such as those associated with metabolic or virulence traits (107).

We found some general patterns in the shape of *r/m* variation across the chromosome of many species. We noted the presence of *i)* an upward parabola-like trend where recombination rates were highest near Ori and lowest near Ter for 16 species (Supp. Fig. 6A) and *ii*) a downward parabola-like trend where the inverse pattern was observed in 5 species (Supp. Fig. 6B). Overall, these patterns support our previous results that the variations of recombination rates along the genomes are somewhat symmetrical in both replichores, *i.e*., symmetrical relative to the Ori-Ter axis. Interestingly, although the number of estimable core genes varied across

species, we observed similar genomic landscapes of recombination rates across the species of some genera (*e.g*. *Lactobacillus* (Supp. Fig. 11) and *Staphylococcus* (Fig. 2)).

Figure III—2. The shape of recombination rate variation across *Staphylococcus* species. Each graph represents the smoothed average *r/m* across estimable core genes in a sliding-window of 50 with a step of 2. The approximate location of the Ori is at *x*=0 and the approximate location of the Ter is demarcated by a green line in each plot. The solid horizontal line represents the mean *r/m* across estimable core genes and the dashed lines are the *r/m* values two standard-deviations from the mean. The red line on each plot represents the location of OrfX (a SCC integration site) and the red triangle indicates the presence of SCCmec integration at this site.



III.3.8 **Overview of Clinically Relevant Genes in Hotspots of Recombination**

Outside of *Staphylococcus* several species had clinically relevant gene sets present in their respective hotspots of recombination. In *B. subtillis*, the hotspot near Ter is associated with genes annotated as belonging to the *yngABC* operon, which is associated with lipid metabolism, biofilm formation, and anaerobic growth (150) (Supp. Fig. 7). Several of these genes were found to have *r/m* rates at approximately 10 times the genome average (ex: *yngB r/m*=12.9, *yngI*, *yngL*, and *yngK r/m*>10) contributing heavily to the spike in *r/m* average in a region with an otherwise low recombination rate. Importantly, the gene *MurM* (*r/m*=9.3, gene 198 in Supp. Fig. 8) is present in a hotspot in *Streptococcus pyogenes* which is linked to penicillin resistance and peptidoglycan formation as well as a regulator of the stringent response pathway in *S. pneumoniae* (151). The second hotspot in *S. pyogenes* houses a lactose specific PTS (phosphotransferase system) (*r/m*~19, gene 684 in Supp. Fig. 8) which has been found to play a role in Group A *Streptococcus* (GAS) virulence in mice (152). In *S. mutans*, the single hotspot was associated with pyrimidine biosynthesis (*pyrK, pyrD, pyrF, pyrE*: *r/m*~20), the upregulation of which is associated with acid-tolerance in *S. mutans* cultures (153) and may contribute to their role in generating dental carries. In *Yersinia enterolitica* the main hotspot was associated with the gene *YadG* (r/*m* =2.7) which is a hypothesized ATP-binding protein of an ABC transporter system and was found to be associated with granuloma (*i.e.*, aggregate of host immune cells) formation in *Y. pseudotuberculosis* (154) (Supp. Fig. 9). *Pseudomonas aeruginosa* had a spike in recombination rate at a region identified as likely being the PA1272 operon (*Cob(I)alamin adenosyltransferase*) which is involved with Vitamin B12 synthesis and plays a role in bacterial persistence within the host as well as disease outcome (Supp. Fig. 10) (155, 156). Here, *CobO* had an *r/m*=18.5 and the other associated Cob proteins were found to be accessory genes and so

did not have estimable *r/m* rates (157). Notably, many genes within this hotspot were excluded from the analysis as their summary statistics fell outside of the distribution of simulated summary statistics for *P. aeruginosa* indicating the true rate of recombination may be underestimated in this region. In *Pseudomonas chlororaphis*, hotspots proximal to Ter are associated with genes for molybdenum metabolism (*MoeA*, *MoaB r/m*~2.5) which is also potentially associated with activation of anaerobic growth (158, 159) (Supp. Fig. 10). The hotspot of recombination in *Corynebacterium diptheriae* was associated with ion transport such as the *czcD* gene (*r/m*=14) which is hypothesized to aid in avoidance of macrophage induced zinc toxicity in human infection (160) (Supp. Fig. 13). Though not all hotspots of recombination were analyzed for all species in this analysis (*n*=146), all gene annotations and corresponding recombination rate data are available on Kaggle (www.kaggle.com/datasets/ellistorr/bacteria-gene-rm).

III.4 **Discussion**

Though homologous recombination rate is expected to vary at the genomic scale, its variation across bacterial genomes has, to our knowledge, been examined in only 10 species and often across relatively few genomes and genomic sites (80, 81, 83, 112, 161). In this study, we estimated the recombination rate (*r/m*) for individual core genes (*n*=208,686 total) across 145 bacterial species and one archaeon using an ABC framework (see Chapter 2). We found that estimations of recombination rate averaged across individual core genes were nearly identical to those inferred for the entire core genome in our previous analysis (Supp. Fig. 1) indicating that core gene estimates of *r/m* by this method seem robust. As observed in prior studies, we noted a statistically significant decrease in *r/m* for genes associated with conserved housekeeping functions such as those coding for transcriptional and cellular replication machinery (Supp. Fig.

2A) (2, 107). The inverse was found for genes related to metabolism, signaling, and virulence among others. Though different metrics were used to assign both gene functionality and homologous recombination rate in prior analyses, this was found to be true for virulence associated genes in only three species in (107) (*Escherichia coli*, *Neisseria meningitidis*, and *Staphylococcus aureus*) (Supp. Fig. 2A). High recombination rate in (107) was most frequently found to be associated with genes involved in modulation of the cell surface and we observed similar trends, but also high levels of recombination in genes involved in cellular metabolism and transport. Within our study, these trends were less statistically salient at the species level—presumably due to decrease in statistical power (Supp. Fig. 2B).

For species with a circular chromosomal and where Ori and Ter could be confidently inferred (n=102), we found that *r/m* was higher at the Ori than the Ter for 67% of species and a statistically significant decrease in *r/m* with increasing distance from the Ori was noted in at least one replichore for 36 species (35%). The inverse (lower *r/m* near the Ori) was observed as significant in 10 species. This trend is most visually observable in Supp. Fig. 6A where we found 16 species had an "upward parabolic" shape to their genomic landscape of recombination and Supp. Fig. 6B where the inverse—a downward parabola-like trend—was observed for five species. Notably, we found that for most bacteria, recombination is higher near Ori, which may be due to increased exposure of this region to the recombination machinery during chromosomal replication. Indeed, during replication, Ori-proximal DNA is present in two or more copies relative to Ter-proximal regions, and this likely offers more opportunities for recombination near Ori. We compared the statistical relationship between *r/m* and distance from the Ori separately for each replichore and found that, for most species, the recombination landscape across the two replichores was quite symmetrical (Supp. Fig. 5). Studies have found the mutational load in

bacteria is also symmetrical in both replichores and varies with replication timing (141, 162). Thus, it is likely that replication timing similarly impacts the variation in homologous recombination rates across the bacterial chromosome.

Interestingly, within some genera (*e.g.*, *Staphylococcus* (Fig. 2) and *Lactobacillus* (Supp. Fig. 11)) several species displayed similar landscapes of recombination. A previous study found evidence that recombination rate may be conserved across orthologs of closely related species (n=3 species pairs (107)) which may contribute to some of the similarity we observed in *r/m* across genomic landscapes. Across 109 species pairs (within the same genus) we found that 36 had statistically significant correlation in *r/m* across shared orthologs (Supp. Table 7). Comparing the pairwise divergence between significant and non-significant groups (Supp. Fig. 4), we found that the significant group were composed of pairs of species that were more related to one another providing further evidence that recombination rate across shared orthologs is conserved across closely related species. It is unclear, however, whether this trend is driven by the conserved synteny of these species and an overall conservation of the recombination landscape or whether homologous recombination rate is similar in genes with similar functions and similar selective pressures.

Increased substitution rate has been observed in the genes of the lagging strand of DNA relative to those in the leading strand and some authors have suggested that this may be the result of asymmetrical mutation rate (163), whereas others have shown evidence that this is due to the higher prevalence of genes evolving under stronger purifying selective pressures on the leading strand (164). Head-on collisions between the replication machinery and the transcription apparatus are thought be more common on the lagging rather than leading strand, and this is thought to have a detrimental impact on gene expression and possibly replication (164). In

contrast to these asymmetrical patterns of substitution rates, we observed statistical difference in the *r/m* values across genes present in the leading strand relative to those found on the lagging strand for only seven species (Supp. Table 6).

Though the base composition of the genome varies across prokaryotes, the relative GC-content across the genome is conserved at the genus and phylum level (165). Variations in GC-content have been proposed to be shaped by selective forces (*e.g.*, for translation efficiency) or neutral processes such as mutational biases or biased gene conversion, or a combination of both (165). Specifically, GC-biased gene conversion (gBGC) is a force shaping the base composition of Eukaryotic genomes whereby mismatches introduced during recombination events are preferentially repaired into G's and C's rather than A's and T's. Therefore, in Eukaryotes, regions of high GC-content are expected to be regions of high recombination and vice versa (166). GC-biased gene conversion is expected to be a neutral process that perhaps counteracts the mutational bias that is universally biased towards A and T (167). However, whether gBGC plays a neutral role in the base composition of Prokaryotic genomes is a subject of debate because higher recombination rates are also expected to enhance the effectiveness of selection through Hill-Robertson effects (147, 165). In accordance with a previous analysis (107) we did observe a positive correlation between *r/m* and GC-content across most species (Supp. Fig 3) which is expected under the gBGC model and a selective model. However, we additionally found a negative correlation between *r/m* and signatures of selection (*dN/dS*) (Supp. Table 4) in most species, and *dN/dS* and GC-content were also correlated (Supp. Table 5). This result is consequential because it shows that GC-content is higher when recombination rate is higher, but *only* when a stronger signature of selection is also observed. In contrast, genes with higher recombination rates did not show an increased GC-content when increased selection was not

observed. Thus, it shows that gBGC does not play a significant role in the evolution of bacterial genomes and that, conversely, homologous recombination shapes GC-content by increasing the effectiveness of selection.

The presence of recombination hotspots (*i.e*., regions with high relative recombination rate) across the genomes of bacteria is thought to be associated with adaptive genes, including those associated with virulence and pathogenicity because these genes frequently arise from horizontal transmission. Since the transmission and integration of these elements is expected to be partially mediated by homologous recombination we expect to observe higher recombination rates, or hotspots, across the genome in relation to some of these regions (83, 161). We observed that the number and frequency of hotspots varied extensively across bacterial chromosomes of different species as has been observed for ten species in (107). Although few of these genes within these regions have been reported in previous analyses, we did observe that *ksgA* in *S. pyogenes* was similarly elevated in both (107) and our analysis. Further, for the genus *Staphylococcus*, we observed a strong trend of a decreasing *r/m* with increasing distance from the Ori and hotspots of *r/m* proximal to the Ori, and this had been observed in two prior studies on *S. aureus* (81, 107). Here the authors note the presence of *OrfX* (a SPOUT-methyltransferase homolog) in recombination hotspots which acts as an integration site for clinically relevant MGEs such as SCCmec (a Staphylococcal Cassette Chromosome carrying genes implicated in methicillin resistance). Interestingly, we found that this integration site is always associated with the recombination hotspot proximal to Ori, which is conserved in all 12 *Staphylococcus* species in our analysis (Fig. 2; *OrfX* highlighted in red). Furthermore, a MGE was found to be integrated at this site in all but one species, and an SCCmec-like element (*i.e.,* an MGE encoding a *MecA* (*PBP2a*) gene specifically associated with high-level methicillin resistance (168)) was found

present at this site in five species (Fig. 2; red triangle denotes presence of SCCmec). Due to the relative conservation of *OrfX* in this highly recombining region proximal to Ori across *Staphylococcus* species, it is likely that an MGE such as SCCmec could be efficiently transferred between species in this genus. Thus, it is likely within the capacity of all *Staphylococcus* species to develop methicillin resistance through inter-species transfers of elements such as SCCmec. Furthermore, this indicates that a hotspot of recombination, which is associated with adaptation, has been conserved for a long period of time.

Finally, it is expected the homologous recombination plays a role in horizontal gene transfer (13). In fact, prior studies have estimated homologous recombination rates to be elevated in regions flanking mobile genetic elements (13, 84). These mobile regions are associated with tracts of accessory genes and so, in this analysis, we compared recombination rates of core genes flanking regions containing at least five consecutive accessory genes to all other core genes. Though we found that flanking core genes had a slight elevation in recombination rate across all species, we observed this elevation to be statistically significant in only nine. However, the differences between studies may be due to differences in defining accessory regions or recombination rate. Perhaps as well, the findings may be more nuanced in that different types of mobile elements do not rely on homologous recombination for integration. For instance, many mobile elements like most temperate bacteriophages encode their own site-specific integrase, and these enzymes are not expected to leave a signal of recombination.

Overall, our results show that homologous recombination plays a major role in shaping the architecture and the evolution of bacterial genomes. Recombination rate is highly variable across the bacterial genome and varies in tandem with both as gene functional roles and chromosomal structure. It further contributes to genome plasticity. A limitation of our

methodology is that we were not able to estimate recombination rates for all genes because some displayed parameters that fell outside of the distributions of simulated range (Supp. Table 1). Thus, the number hotspots of recombination may be underestimated for some species. Nevertheless, our results provide a rather complete picture showing how recombination rate is associated with many aspects of the architecture and evolution of bacterial genomes: gene function, replication timing, the origin and terminus of replication, GC-content, selection, and gene transfers.

III.5 Supplementary Materials

III.5.1 Supplementary Methods

III.5.1.1 Data Assembly

The bacterial species used in this analysis were defined in a previous study as having ≥15 sequenced non-identical genomes available on Refseq (see Torrance *et al.* (2024): *unpublished* manuscript in Chapter 2). In this study, the assembly quality of the genomes within each species was ensured by checking that each genome contained the expected number of universal genes as in (129). Further, the genomes within each species were also redefined in accordance with methods proposed by Diop *et al*. (2022) by ensuring the genomes adhered to their given species definition by both *i)* ANI (≥ 94% pairwise identity across the core genome) and *ii)* gene flow analysis (*i.e.* strains were only included in the analysis if they were determined to be engaging in gene flow with other members of the species) (4). For the 162 bacterial species and one archaeon that contained ≥15 species, orthologs were inferred using *CoreCruncher* with default parameters and orthologous genes were defined as core genes when present in at least 90% of the genomes of a species (131). The core genes were aligned using Mafft (130). One reference genome was chosen for each species by having, first, the most complete genome assembly (i.e., the least

number of contigs) and second, the highest number of predicted coding sequences. Accessory genes were defined as all the other genes that were not defined as core present in the reference genome of each species. A core genome phylogeny was generated from the core genome concatenate for each species using *RAxML* (134).

III.5.1.2 **Estimation of recombination rates**

A set of 500,000 forward-in-time simulations with varied rates of homologous recombination (*r/m*) was generated using each species tree and each core genome concatenate using *CoreSimul* through the *recABC* pipeline as in Torrance *et al.,* (2024) (see *unpublished* manuscript in Chapter 2). For each species, the recombination rate of each core gene was estimated by comparing the signature of recombination between each core gene alignment and the simulated sequences using ABC. The recombination rate of each core gene was estimated by the recombination rate used to evolve the simulated sequences that most closely and robustly matched the signatures of recombination to the gene. Here, the summary statistics π (i.e., levels of polymorphisms) and *h/m* (*i.e.* the ratio of homoplasic to nonhomoplasic alleles) were calculated for each core gene within each species. Then, using ABC, the summary statistics for each core gene were compared to the summary statistics generated from each simulated species population with a tolerance of 0.01% to determine the most probable rate of recombination for each gene.

As expected, not all summary statistics for genes fell within the distribution of simulated summary statistics of the simulated sequences. These genes were removed from further analysis when the summary statistic for a given gene varied from the average of its nearest simulated sequence set by more than ± 0.1 for *h/m or* ± 0.01 for π. This threshold was determined by comparing the graphs and correlations (*Spearman's rho*) of each simulated summary statistic *vs*.

real summary statistic (*h/m* and π) for each species. Here, we found that removal of genes outside these tolerance levels yielded the highest number of estimable genes where the simulated and real value summary statistic was closest to equivalent (x=y, or *Spearman's rho* $\geq$0.97). The number of genes that were excluded from further analysis by this metric are listed in Supplementary Table 1. Furthermore, some core genes had no polymorphisms and thus *r/m* could not be estimated (*i.e., m*=0) and those were also excluded from the analysis. These genes are denoted as a *r/m* of "NA" in column 2 of the datasets available on Kaggle (www.kaggle.com/datasets/ellistorr/bacteria-gene-rm). Species with fewer than 200 inferred core genes for which *r/m* could be predicted were excluded from further analysis (*n*=10). Thus, our final dataset was composed of 146 species. The species included in this analysis and the total number of core genes and accessory genes, as well as the number of core genes which had no polymorphisms or had summary statistics which fell outside of the range of simulated summary statistics, are detailed in Supplementary Table 1. The average rate of recombination estimated across all genes was compared to the *r/m* values estimated on the core genome concatenate in Torrance et al. (2024) (see *unpublished* manuscript in Chapter 2) in Supp. Fig. 1. In depth data for each species detailing the summary statistics, *r/m* estimation of each core gene, and other gene characteristics amassed in this study is available at Kaggle (www.kaggle.com/datasets/ellistorr/bacteria-gene-rm).

III.5.1.3 **Identification of Ori and Ter**

The origin (Ori) and terminus (Ter) of replication were identified using the cumulative GC (CGC) skew in the reference genome for each species as in (74). We build a graph for each species by plotting the CGC skew using a 10kb sliding window. Species were only included in this analysis if they had a single clear maximum (corresponding to approximate Ter location) and

a single clear minimum (corresponding to approximate Ori location) (*n*=110 species) (74). In other words, Ori and Ter were not inferred for the species presenting additional local maxima or minima. Species with known linear chromosomes or multiple chromosomes were identified from (64) and excluded from the analyses of recombination symmetry. Using the 102 species with clearly defined Ori and Ter locations, the recombination rate of each core gene was plotted relative to the absolute distance from Ori and Ter to determine whether recombination rate varied on both replichores of each species. These species were also used to compare *r/m* variations between leading and lagging strand genes. Leading and lagging strands were determined using the map of GC skew where core genes present on the strand with increasing GC skew with increasing distance from the Ori and a positive orientation were determined to be leading strand genes while those with a negative orientation in this region were determining to be lagging strand genes. For the segment of chromosome with a decreasing GC-skew with increasing distance from the Ori, the positively oriented genes were lagging strand and the negatively oriented genes were leading strand. Variations of recombination rate along the core genome of each species were computed using a sliding window of 50 core genes and a step of two genes. The plots are ordered by Ori start location (Supplementary Folder 1). Hotspots and coldspots of recombination were defined as genome locations where *r/m* differed by more than two standard deviations from the average *r/m* of the species.

III.5.1.4 **Other Gene Analyses**

To determine whether recombination rate varied significantly across gene functions, the reference genome of all species (*n*=146) was annotated using *EggNOG* (98) and recombination rate was compared across 18 COG (clusters of orthologous genes) categories (208,686 core genes) (99). For this analysis, the COG categories A (RNA processing and modification) and B

(chromatin structure and dynamics) were excluded because these categories typically only correspond to Eukaryotic systems. Category S was excluded because it corresponds to groups of orthologs without known functions. Further, genes were only compared if they were annotated as belonging to a single COG category and genes that were not assigned to a category were excluded. For each species, a Wilcoxon test with Benjamini-Hochberg *P*-value adjustment was conducted where *r/m* for the genes in each COG category was compared to *r/m* of the genes classified in all the other COG categories. Additionally, a Wilcoxon test where all species gene data was pooled to compare *r/m* of across the genes in each COG category to the genes in all other COG categories was performed. We also used the same dataset of 146 species to determine whether *r/m* varied with any appreciable pattern in genes flanking accessory gene clusters to those not flanking accessory regions. To do this, we defined accessory regions and regions of the genome containing at least five consecutive accessory genes. We then compared the recombination rate of the core genes flanking these regions to the recombination rate of non-flanking core genes using a Wilcoxon test with Benjamini-Hochberg adjustment for each species.

III.5.2 **Supplementary Figures**

Figure S III—1. Species' core genome *r/m* values from Torrance *et al.* (2024) (Chapter 2) are highly similar to the average *r/m* values across genes for the same species (Spearman's *Rho*=0.91, *P*<10^{-15}).

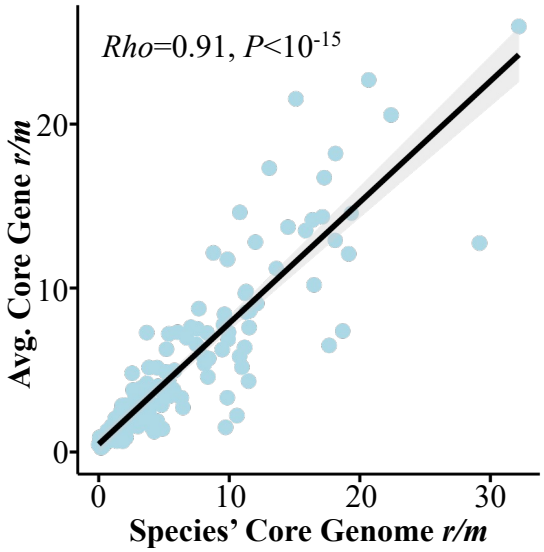
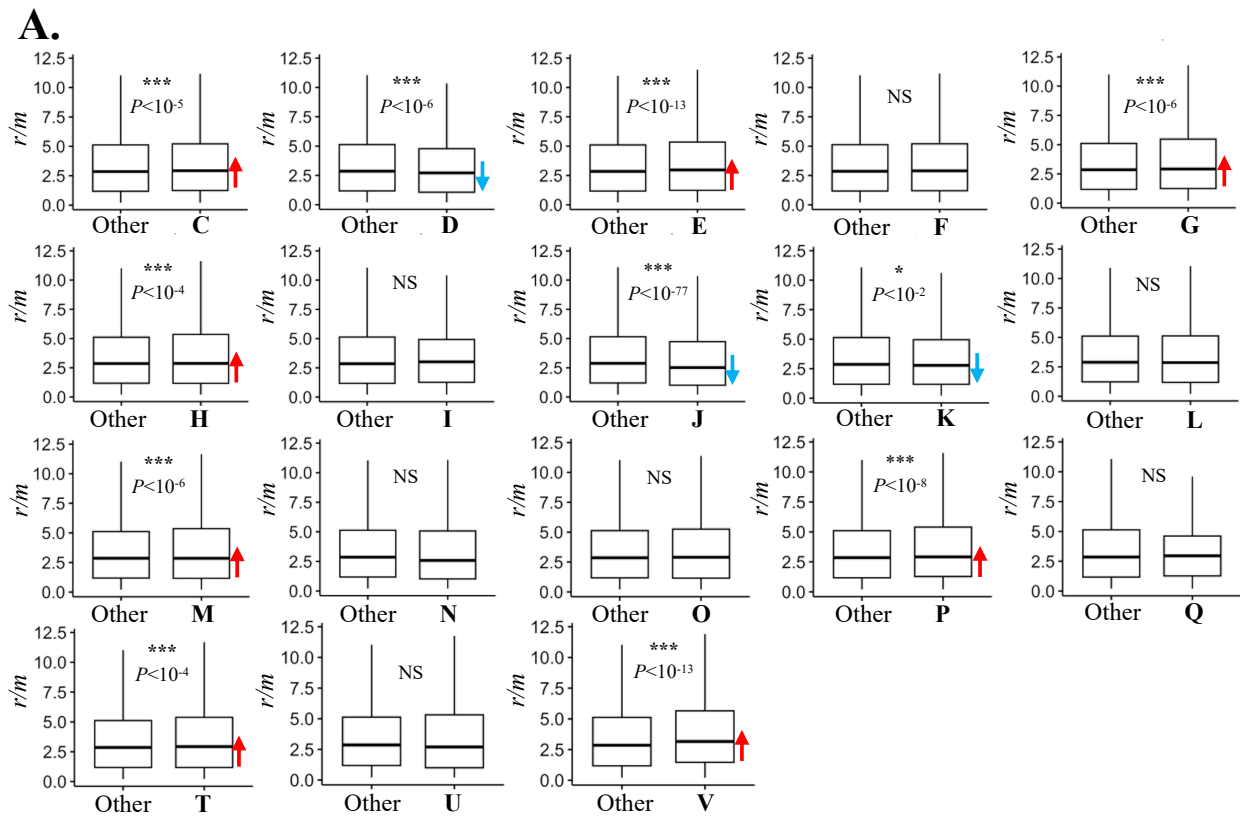


Figure S III—2. The variation in *r/m* across gene functional categories. A) Boxplots show the variation in *r/m* between each COG (Conserved Ortholog Group (cite)) category and all other genes ("Other"). On each plot, the significance level and Benjamini-Hochberg adjusted *P*-value for each Wilcoxon test is shown. Red (higher *r/m*) and blue (lower *r/m*) arrows are shown for each COG category that had a significant difference between it and all other genes. A key listing the description for each COG category is shown in the bottom left. B) A bar graph shows the number of species which had a significant difference in *r/m* between the COG category and all other genes (Wilcoxon Test, Benjamini-Hochberg adjusted *P*-value<0.05).



COG Categories

- **C**: Energy Production and Conversion
- **D**: Cell Cycle Control, Cell Division, and Chromosome Partitioning
- **E**: Amino Acid Transport and Metabolism
- **F**: Nucleotide Transport and Metabolism
- **G**: Carbohydrate Transport and Metabolism
- **H**: Coenzyme Transport and Metabolism
- **I**: Lipid Transport and Metabolism
- **J**: Translation, Ribosomal Structure, and Biogenesis
- **K**: Transcription
- **L**: Replication, Recombination, and Repair
- **M**: Cell Wall/Membrane/Envelope Biogenesis
- **N**: Cell Motility
- **O**: Posttranslational Modification, Protein Turnover, and Chaperones
- **P**: Inorganic Ion Transport and Metabolism
- **Q**: Biosynthesis, Transport, and Catabolism of Secondary Metabolites
- **T**: Signal Transduction Mechanisms
- **U**: Intracellular Trafficking, Secretion, and Vesicular Transport
- **V**: Defense Mechanisms

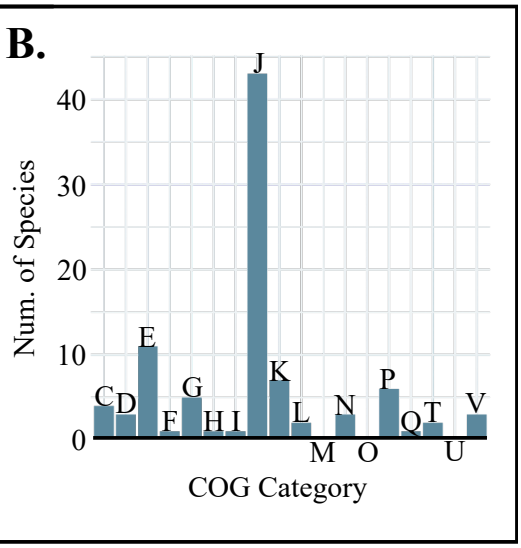


Figure S III—3. Histogram of Spearman's *Rho* values for species which had a significant correlation between *r/m* value and GC% across their genes. The blue dashed line shows the average *Rho* value (Average *Rho*=0.08±0.1).

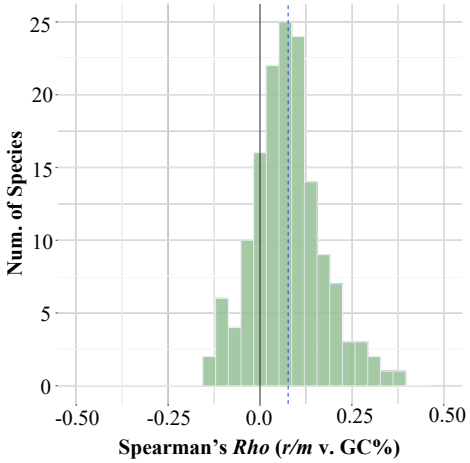


Figure S III—4. Boxplot comparison of the pairwise divergence (A.A.) values for species pairs (*n*=109 species pairs) which had significant correlation in *r/m* values across shared orthologs ("Significant", *n*=36 species pairs) and those that did not ("Non-Significant", *n*=73).

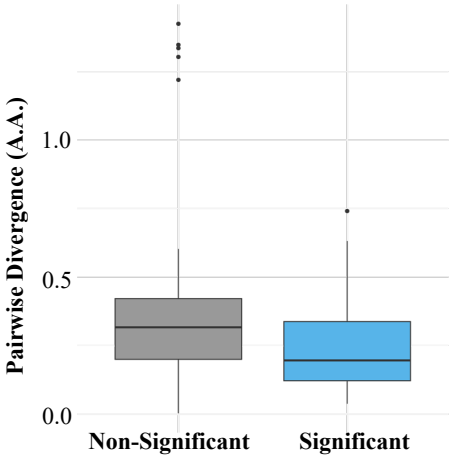


Figure S III—5. Correlation between Spearman's *Rho* values from the correlation of *r/m* vs. distance from Ori in replichore 1 and replichore 2 of *n*=102 species with circular chromosomes.

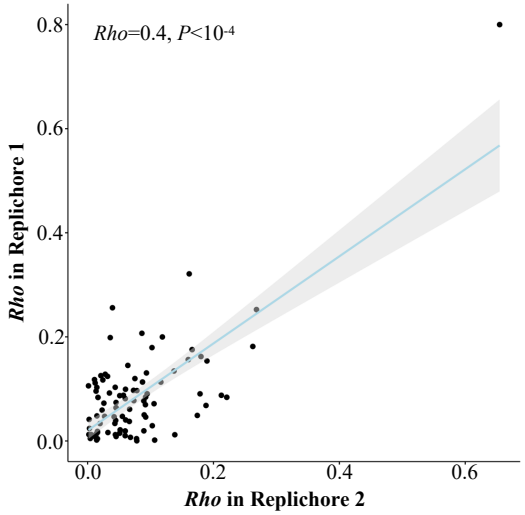


Figure S III—6. The shape of recombination rate across bacterial genomes. A) Some bacterial species (*n*=16) displayed an upward "parabola-like" shape in recombination rate variation across their genomes. B) Some bacterial species (*n*=5) displayed a downward "parabola-like" shape in recombination rate variation across their genomes. These graphs display the smoothed average of a sliding window of 50 estimable core genes with a step of 2 where *x* is the core gene number in order of its appearance with increasing distance from the Ori. For each graph, *x*=0 demarcates the Ori and the green line demarcates the Ter. The black horizontal line denotes the average *r/m* value of all estimable core genes, and the dashed lines denote two standard-deviations from the mean *r/m* in both directions.

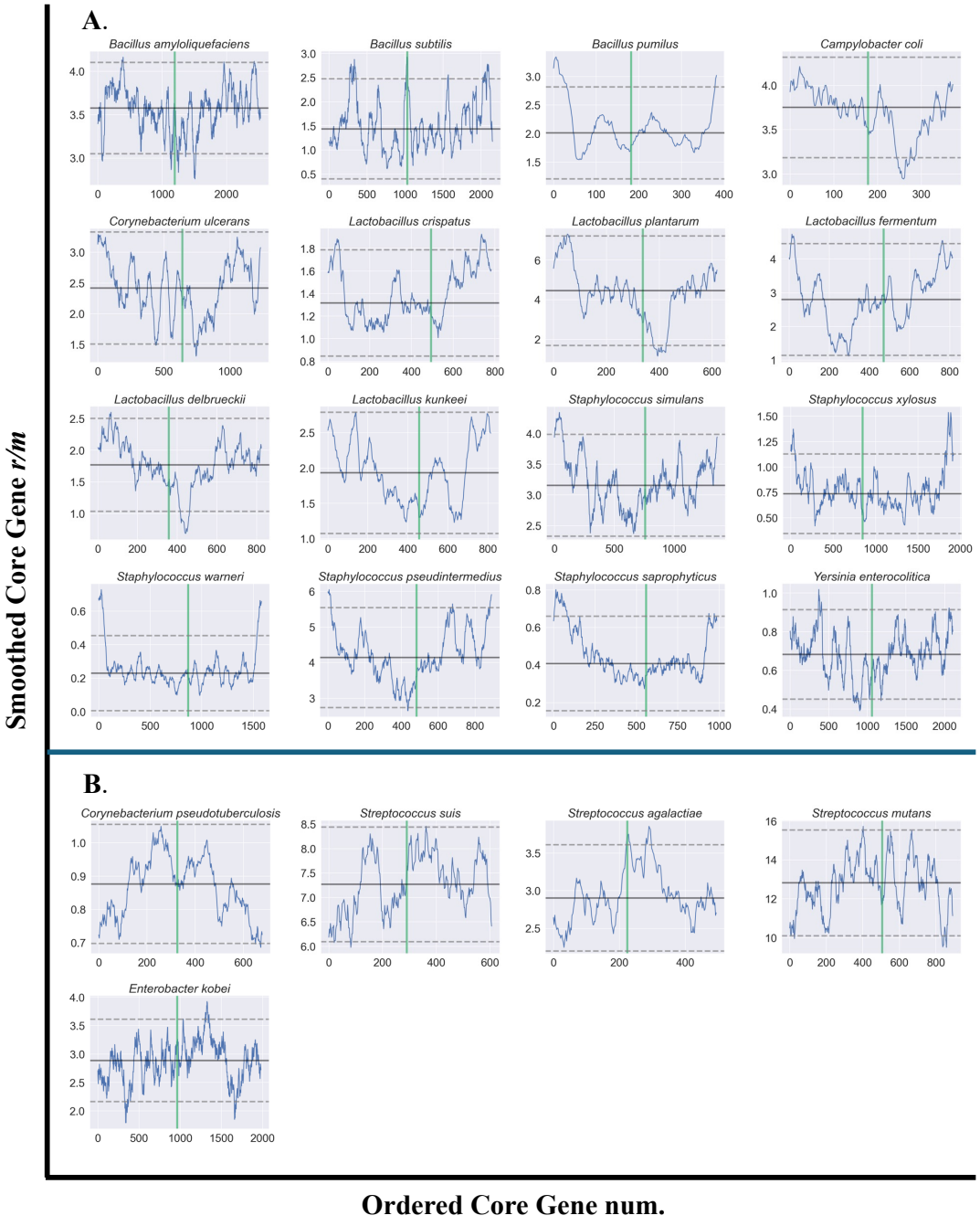

A.
Bacillus amyloliquefaciens
Bacillus subtilis
Bacillus pumilus
Campylobacter coli
Corynebacterium ulcerans
Lactobacillus crispatus
Lactobacillus plantarum
Lactobacillus fermentum
Lactobacillus delbrueckii
Lactobacillus kunkeei
Staphylococcus simulans
Staphylococcus xylosus
Staphylococcus warneri
Staphylococcus pseudintermedius
Staphylococcus saprophyticus
Yersinia enterocolitica
B.
Corynebacterium pseudotuberculosis
Streptococcus suis
Streptococcus agalactiae
Streptococcus mutans
Enterobacter kobei
Smoothed Core Gene r/m
Ordered Core Gene num.

Figure S III—7. The shape of recombination rate across *Bacillus* genomes (*n*=10). These graphs display the smoothed average across 50 estimable core genes with a step of 2 where *x* is the core gene number in order of its appearance with increasing distance from the Ori. For each graph, *x*=0 demarcates the Ori and the green line demarcates the Ter. The black horizontal line denotes the average *r/m* value of all estimable core genes, and the dashed lines denote two standard-deviations from the mean *r/m*.

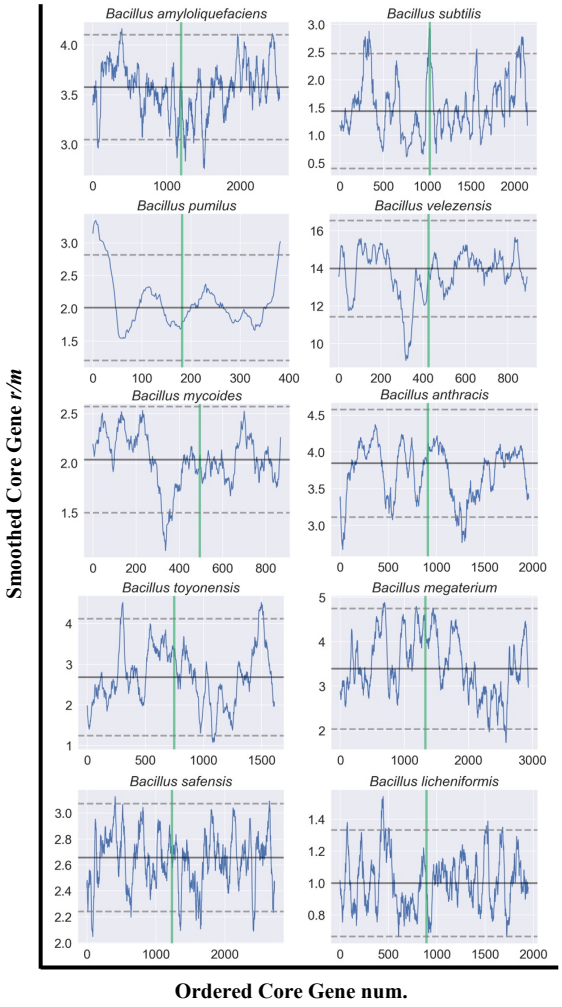


Figure S III—8. The shape of recombination rate across *Streptococcus* genomes (*n*=11). These graphs display the smoothed average across 50 estimable core genes with a step of 2 where *x* is the core gene number in order of its appearance with increasing distance from the Ori. For each graph, *x*=0 demarcates the Ori and the green line demarcates the Ter. The black horizontal line denotes the average *r/m* value of all estimable core genes, and the dashed lines denote two standard-deviations from the mean *r/m*.

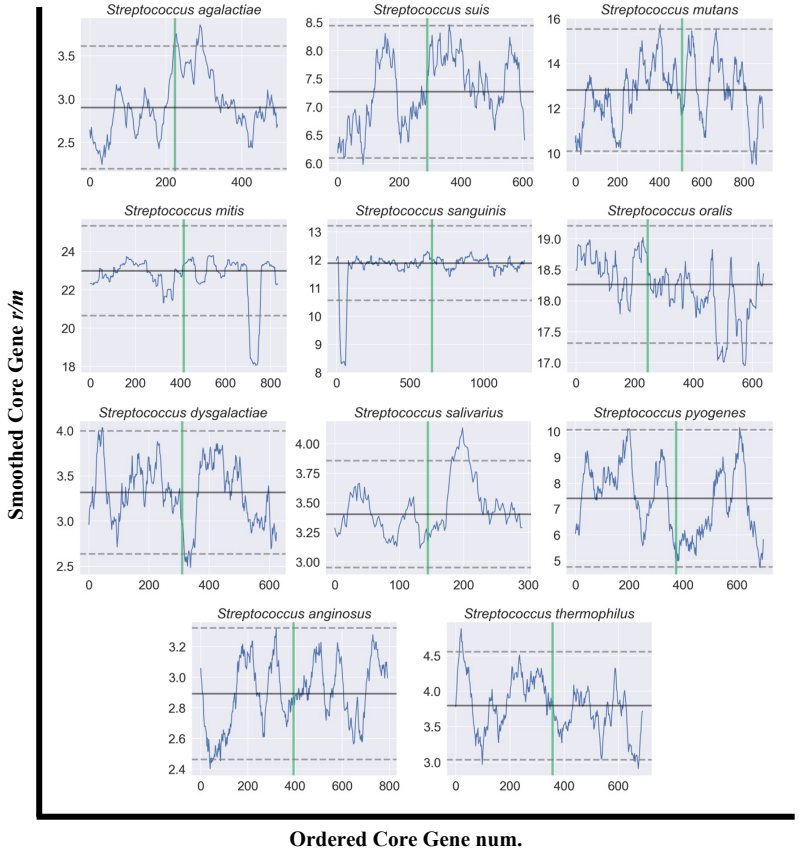


Figure S III—9. The shape of recombination rate across *Yersinia* genomes (*n*=4). These graphs display the smoothed average across 50 estimable core genes with a step of 2 where *x* is the core gene number in order of its appearance with increasing distance from the Ori. For each graph, *x*=0 demarcates the Ori and the green line demarcates the Ter. The black horizontal line denotes the average *r/m* value of all estimable core genes, and the dashed lines denote two standard-deviations from the mean *r/m*.

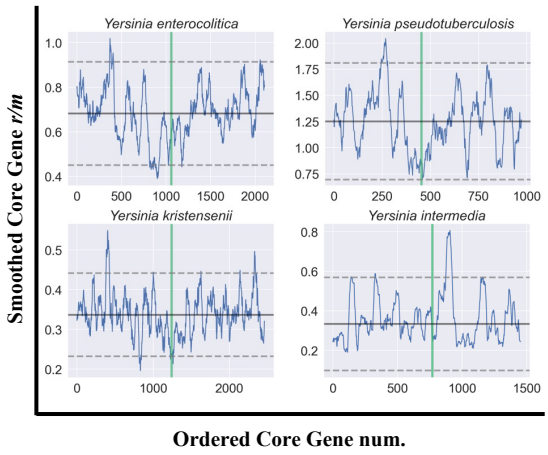


Figure S III—10. The shape of recombination rate across *Pseudomonas* genomes (*n*=5). These graphs display the smoothed average across 50 estimable core genes with a step of 2 where *x* is the core gene number in order of its appearance with increasing distance from the Ori. For each graph, *x*=0 demarcates the Ori and the green line demarcates the Ter. The black horizontal line denotes the average *r/m* value of all estimable core genes, and the dashed lines denote two standard-deviations from the mean *r/m*.

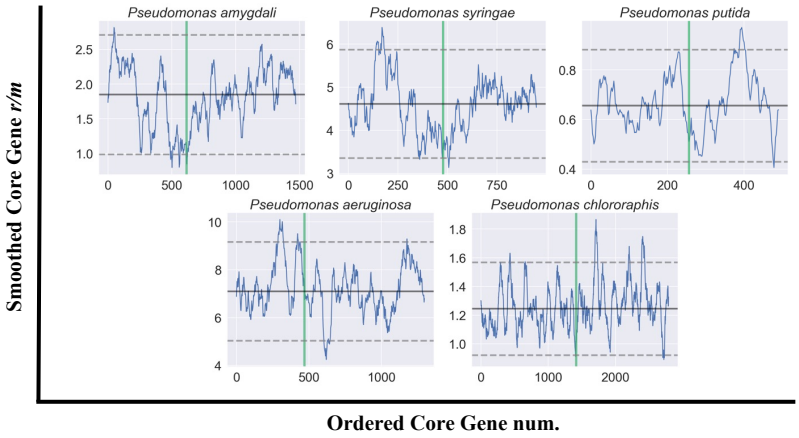


Figure S III—11. The shape of recombination rate across *Lactobacillus* genomes (*n*=14). These graphs display the smoothed average across 50 estimable core genes with a step of 2 where *x* is the core gene number in order of its appearance with increasing distance from the Ori. For each graph, *x*=0 demarcates the Ori and the green line demarcates the Ter. The black horizontal line denotes the average *r/m* value of all estimable core genes, and the dashed lines denote two standard-deviations from the mean *r/m*.

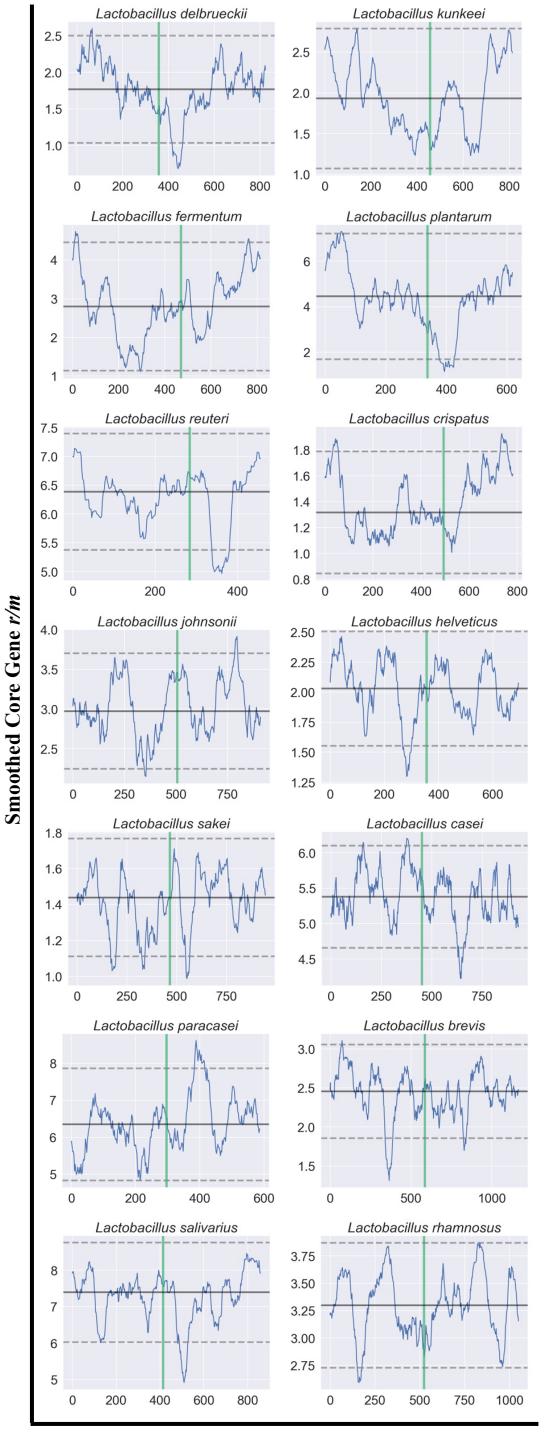

Lactobacillus delbrueckii
Lactobacillus kunkeei
Lactobacillus fermentum
Lactobacillus plantarum
Lactobacillus reuteri
Lactobacillus crispatus
Lactobacillus johnsonii
Lactobacillus helveticus
Lactobacillus sakei
Lactobacillus casei
Lactobacillus paracasei
Lactobacillus brevis
Lactobacillus salivarius
Lactobacillus rhamnosus
Smoothed Core Gene r/m
Ordered Core Gene num.

Figure S III—12. The shape of recombination rate across *Klebsiella* genomes (*n*=4). These graphs display the smoothed average across 50 estimable core genes with a step of 2 where *x* is the core gene number in order of its appearance with increasing distance from the Ori. For each graph, *x*=0 demarcates the Ori and the green line demarcates the Ter. The black horizontal line denotes the average *r/m* value of all estimable core genes, and the dashed lines denote two standard-deviations from the mean *r/m*.

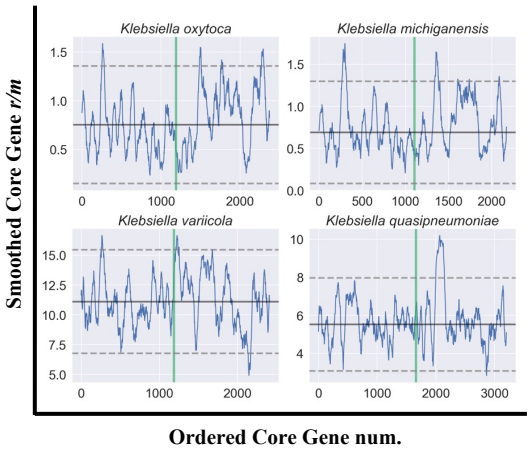
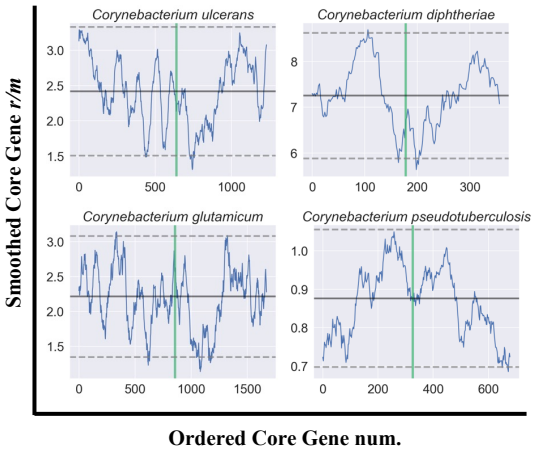


Figure S III—13. The shape of recombination rate across *Corynebacterium* genomes (*n*=4). These graphs display the smoothed average across 50 estimable core genes with a step of 2 where *x* is the core gene number in order of its appearance with increasing distance from the Ori. For each graph, *x*=0 demarcates the Ori and the green line demarcates the Ter. The black horizontal line denotes the average *r/m* value of all estimable core genes, and the dashed lines denote two standard-deviations from the mean *r/m*.



III.5.3 **Supplementary Table Legends**

DISCLAIMER: The supplementary tables are too large to be included in this document and so only their legends are included below. The tables will be included in the peer-reviewed manuscript once publication occurs.

Table S III—2. Summary of gene data for each species analyzed in this study (*n*=146). The table columns are as follows: Col 1) Species name, Col 2) Number of core genes (total), Col 3) Number of accessory genes, Col 4) Number of core genes without polymorphisms (*r/m* not estimable), Col 5) Number of core genes outside of simulation distribution (*r/m* not estimable), Col 6) Number of core genes used in this analysis (*r/m* is estimable), Col 7) Species' core genome *r/m* (as per Torrance *et al*. (2024)), Col 8) Average *r/m* across estimable core genes, Col 9) Median *r/m* across estimable core genes, Col 10) Standard deviation of *r/m* across estimable core genes, Col 11) Average simulated *h/m* across estimable core genes, Col 12) Median simulated *h/m* across estimable core genes, Col 13) Standard deviation of simulated *h/m* across estimable core genes, Col 14) Average simulated π across estimable core genes, Col 15) Median simulated π across estimable core genes, Col 16) Standard deviation of simulated π across estimable core genes, Col 17) Average real *h/m* across estimable core genes, Col 18) Median real *h/m* across estimable core genes, Col 19) Standard deviation of real *h/m* across estimable core genes, Col 20) Average real π across estimable core genes, Col 21) Median real π across estimable core genes, and Col 22) Standard deviation of real π across estimable core genes.

Table S III—2. Wilcoxon's test comparison of *r/m* for core genes flanking accessory regions ("Flanking") vs. *r/m* of core genes not flanking accessory regions ("Non-Flanking") across

all species *n*=146. Significant Benjamini-Hochberg adjusted *P*-values are highlighted in red. The median *r/m* of each group and difference in *r/m* between each group is also provided.

Table S III—3. Spearman's correlation test results for comparison between *r/m* and GC% values across core genes for each species (*n*=146). Significant Benjamini-Hochberg adjusted *P*-values are highlighted in red.

Table S III—4. Spearman's correlation test results for comparison between *r/m* and *dN/dS* (Tab 1), *dN* (Tab 2), and *dS* (Tab 3) values across core genes for each species (*n*=142).

Table S III—5. Spearman's correlation test results for comparison between GC% and dN/dS values across core genes for each species (*n*=142). Significant Benjamini-Hochberg adjusted *P*-values are highlighted in red.

Table S III—6. Wilcoxon Test results for comparison of *r/m* between leading and lagging strand genes for each species (*n*=102). Significant Benjamini-Hochberg adjusted *P*-values are highlighted in red.

Table S III—7. Spearman's correlation test results for comparison of *r/m* values across the shared orthologs of related species (*n*=109 species pairs). Significant Benjamini-Hochberg adjusted *P*-values are highlighted in red. The table also contains the pairwise divergence (A.A.) between each species pair.

Table S III—8. Spearman's correlation test results for comparison between r/m and distance from the Ori (Ori-Ter) and Ter (Ter-Ori) (*i.e.* both replichores) for *n*=102 species with circular chromosomes. Significant Benjamini-Hochberg adjusted *P*-values are highlighted in red.

III.6 Associated Contents

III.6.1 Ethics Approval and Consent to Participate

Not applicable.

III.6.2 Consent for Publication

Not applicable.

III.6.3 Availability of Data and Materials

All data used in this analysis was downloaded from NCBI's Genbank public genomic repository. GenBank accession numbers and assembly IDs for all analyzed genomes are detailed in Supplementary Table 2 of Chapter 2. The pipeline *recABC* used to generate recombination rates is available at (https://github.com/lbobay/recABC). The datasets generated in this study (the core genomes, phylogenetic trees and the summary statistics of all the simulations for all species) are available on *Kaggle* at https://www.kaggle.com/datasets/ellistorr/bacterial-rm. The individual gene recombination rates, associated summary statistics, and other gene parameters used in this study are available at https://www.kaggle.com/datasets/ellistorr/Bacteria-Gene-rm.

III.6.4 Competing interests

The authors declare that they have no competing interests.

III.6.5 Funding Information

This study was supported by the National Institutes of Health grant R01GM132137 awarded to LMB and by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0021110 awarded to ELT.

III.6.6 **Funding Disclaimer**

III.6.7 **Authors' contributions**

E.L.T., A.D., and L-M. B. designed, performed research, and analyzed data in this publication. E.L.T. and L-M.B. wrote this paper and A.D. reviewed and edited it.

III.6.8 Acknowledgments

We would like to thank Ophelia Adams, Daniel Schrider and Kasie Raymann for providing advice and expertise.

III.7 References

See REFERENCES on page 130.

CHAPTER IV: CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Prokaryotes are the most abundant lifeforms on this planet and yet much remains unknown regarding the processes that govern their species and genomic evolution. The overarching goal of this dissertation was to discern and then explore the variation in homologous recombination (HR) rates across bacterial species and across genomic positions. Specifically, I was interested in *i*) determining whether polymorphisms imparted by HR were more impactful to genome diversification than mutation alone and whether this was true for most species, *ii*) whether recombination was a fast or slow evolving trait and whether there was conservation of the trait within bacterial lineages and, *iii*) if recombination rate varied with any patterns across bacterial genes and genomes.

The first step in accomplishing these aims was leveraging a methodological framework based on Approximate Bayesian Computation (ABC), as in Chapter 2. This computational pipeline – now aptly named *recABC* – allowed me to infer bacterial recombination rate (*r/m*) which is a ratio of the polymorphisms imparted by recombination relative to mutation alone. I did this for 162 bacterial species and one archaeon and found that HR rate varied extensively across species. Furthermore, I found that, for most species, recombination was more impactful to bacterial evolution than mutation alone. Mapping the variation in the trait of recombination rate across species revealed a hereto unforeseen pattern of conservation in recombination rate amongst several genera. This finding indicates that homologous recombination rate is a somewhat conserved trait amongst some lineages while it sharply changes in others. I also found, as seen in previous studies, that recombination rate was significantly lower amongst obligate intracellular bacteria. Interestingly, no correlations were observed between recombination rate

and metabolic capabilities, virulence phenotypes, prophage content, and genomic variations. I also used the tool ClonalFrameML (CFML) (45) to compare similarities between results achieved with *recABC* and found that CFML rarely inferred recombination rates of greater than five. In fact, *r/m*=5 was where I observed that the homoplasy to mutation ratio was no longer directly related to *r/m* and that signal saturation of recombination began to occur.

In Chapter 2, I analyzed several species traits and found no correlation to recombination rate variation across species. However, there are still many traits left to be examined which are theorized to vary with recombination rate. For instance, HR across species should also be compared to the presence and/or absence of different subtypes of recombination machinery, genes conveying species competence (i.e., the ability to express proteins necessary for transformation (see Ch 1: Introduction)), or restriction modification systems (2, 169). Analysis such as these have been made simpler by my estimation of recombination rate for many species and finding traits that correlate with these rates could reveal much about the factors driving recombination rate variations across species and the evolution of HR in bacteria.

Unfortunately, due to the stringency with which we defined our genomic datasets, I was only able to estimate recombination rate for a single archaeon. Ideally, a greater effort to sequence and assemble archaeal genomes from the environment would enable the determination of whether homologous recombination was as great of an evolutionary driver for archaeal evolution as it is for bacteria. Further, though I attempted to observe differences in *r/m* between bacterial species and ecological subtypes; genomes relevant to human and agricultural disease were overrepresented in our genetic sampling relative to environmental strains. This made it hard to determine whether any differences in patterns of recombination exist between species associated with different environments, metabolic capabilities, and lifestyles. Thus, these

analyses should be ongoing as sequencing data for prokaryotes becomes more widely available. Continuing work on this topic should unveil whether the impact or frequency of recombination varies between archaea and bacteria and whether any phenotypic, ecological, or environmental factors reveal why patterns in recombination rate variation exist across species and genera.

Comparison of HR rates obtained with *recABC* to predecessor software revealed that *recABC* is much more adept at inferring recombination rate. Especially in instances of high recombination which leads to signal saturation at recombination rates of >5. However, there are also several assumptions of the software which could be improved upon with time and advancements in computational capabilities. For example, recombination simulations assume a constant rate or recombination along the branches of the phylogeny and that all genomes have the same probability of recombining with one another. Reality is obviously much more complex and there are likely population and genomic barriers to recombination which we are presently unable to simulate in this analysis. I look forward to seeing improvements in this methodology and inference of bacterial genetic exchange as scientific and computational advancements in the study of Prokaryotic population genetics occur.

Though *recABC* appears to be much more accurate at inferring homologous recombination rates, it also requires more time for analysis and is more computationally expensive than most predecessor software. For example, though one can use a standard laptop to run the simulations sequentially, the lower estimation for time to complete 500,000 simulations on a relatively small set of genomes (15 strains) would be ~15 days. When launching the simulations on a computing cluster, I batch-launched the simulations in sets of 1,000 to speed up the process. However, doing this is not very user-friendly, most biologists are not trained to use computing clusters or do not have access to one, and many computing clusters have strict run-

time limitations. Therefore, future work will entail parallelizing the simulation portion of *recABC* and adding code to let users easily control the number of simulations launched. It is my hope that perhaps much of the data generated from our extensive analysis of bacterial species and genomes could be used to train a machine learning algorithm to speed up inference of recombination rates.

The final aim of this dissertation was addressed in Chapter 3 and involved determining how recombination rate varied across bacterial chromosomes. I found that recombination rate varied by gene functionality in a somewhat predictable pattern: core genes associated with conserved 'housekeeping' functions tended to have lower recombination rates and genes associated with virulence and metabolism tended to have higher recombination rates. Further, gene recombination rate varied extensively across the chromosome and showed patterns in some species that correlated to replication, pathogenicity islands, and accessory gene location. Specifically, I found a very conserved genomic landscape of recombination across the species of the genus *Staphylococcus.* This finding indicates that genome-wide patterns of recombination rates may be a conserved trait in some lineages as was observed for genera in our analysis of the evolution of recombination rate across bacteria (see Chapter 2 Results).

One interesting direction for future work in this domain is the exploration of whether the landscape of recombination is shaped by genomic evolution and structure or rather genomic structure itself is shaped by recombination. This question could be resolved by comparing the landscape of recombination across closely related species. Indeed, we found that recombination rate was somewhat conserved across orthologs in related species. However, it is not known whether the conservation in recombination rate is syntenic or whether the conservation relates to the gene's functional role. To determine this, I would need to determine whether conserved tracts

of recombination (such as the region of high recombination rate near the Ori in *Staphylococcus*) contained similar gene sets in conserved order. Once this is accomplished, I could also work to determine whether recombination rate varies in tandem with evidence of positive selection for individual genes or across the genomic landscape. As recombination is expected to increase the power of selection, I would expect the interplay of recombination and selection to be tightly coupled.

In conclusion, my dissertation work has allowed the estimation of homologous recombination rate for many bacterial species and genomes. Overall, this work has provided insight on the diversity in recombination and the commonality of DNA exchange across species — establishing a "first-step" in understanding how recombination shapes bacterial evolution, genomes, adaptation, and population diversity. Future work in this field will enable the discernment of the interplay of selective processes and recombination on the evolution of Prokaryotic species and genomes, thereby allowing us to better predict disease emergence and pathogenicity, as well as employ Prokaryotes more effectively in agricultural, biotechnological, and medical settings.

REFERENCES

1. M. Vos, X. Didelot, A comparison of homologous recombination rates in bacteria and archaea. *ISME Journal* **3**, 199–208 (2009).
2. P. González-Torres, F. Rodríguez-Mateos, J. Antón, T. Gabaldón, Impact of homologous recombination on the evolution of prokaryotic core genomes. *mBio* **10** (2019).
3. L. M. Bobay, C. C. Traverse, H. Ochman, Impermanence of bacterial clones. *Proc Natl Acad Sci U S A* **112**, 8893–8900 (2015).
4. A. Diop, E. L. Torrance, C. M. Stott, L. M. Bobay, Gene flow and introgression are pervasive forces shaping the evolution of bacterial species. *Genome Biol* **23** (2022).
5. S. H. Kung, A. C. Retchless, J. Y. Kwan, R. P. P. Almeida, Effects of DNA Size on Transformation and Recombination Efficiencies in Xylella fastidiosa. *Appl Environ Microbiol* **79**, 1712–1717 (2013).
6. X. Didelot, G. Méric, D. Falush, A. E. Darling, Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics* **13** (2012).
7. C. M. Thomas, K. M. Nielsen, Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 711–721 (2005).
8. L.-M. Bobay, H. Ochman, Biological Species Are Universal across Life's Domains. *Genome Biol Evol* **9**, 491–501 (2017).
9. I. Matic, C. Rayssiguier, M. Radman, Interspecies gene exchange in bacteria: The role of SOS and mismatch repair systems in evolution of species. *Cell* **80**, 507–515 (1995).

10. J. Iranzo, Y. I. Wolf, E. V. Koonin, I. Sela, Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. *Nat Commun* **10** (2019).

11. J. C. Dunning Hotopp, Horizontal gene transfer between bacteria and animals. *Trends in Genetics* **27**, 157–163 (2011).

12. M. Touchon, J. A. Moura de Sousa, E. P. Rocha, Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr Opin Microbiol* **38**, 66–73 (2017).

13. P. H. Oliveira, M. Touchon, J. Cury, E. P. C. Rocha, The chromosomal organization of horizontal gene transfer in bacteria. *Nat Commun* **8** (2017).

14. A. P. Steinberg, M. Lin, E. Kussell, Core genes can have higher recombination rates than accessory genes within global microbial populations. *Elife* **11** (2022).

15. M. Vos, Why do bacteria engage in homologous recombination? *Trends Microbiol* **17**, 226–232 (2009).

16. E. P. C. Rocha, E. Cornet, B. Michel, Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet* **1**, 0247–0259 (2005).

17. H. Marti, R. J. Suchland, D. D. Rockey, The Impact of Lateral Gene Transfer in Chlamydia. *Front Cell Infect Microbiol* **12** (2022).

18. M. S. Dillingham, S. C. Kowalczykowski, RecBCD Enzyme and the Repair of Double-Stranded DNA Breaks. *Microbiology and Molecular Biology Reviews* **72**, 642–671 (2008).

19. G. R. Smith, Homologous Recombination in Procaryotes. *Microbiol Rev* **52**, 1–28 (1988).

20. P. Dabert, S. D. Ehrlich, A. Gruss, Chi sequence protects against RecBCD degradation of DNA in vivo. *Proceedings of the National Academy of Sciences* **89**, 12073–12077 (1992).

21. K. C. Cheng, G. R. Smith, Recombinational hotspot activity of chi-like sequences. *J Mol Biol* **180**, 371–377 (1984).

22. A. Buton, L. M. Bobay, Evolution of chi motifs in proteobacteria. *G3: Genes, Genomes, Genetics* **11** (2021).

23. L. R. Bazemore, E. Folta-Stogniew, M. Takahashi, C. M. Radding, RecA tests homology at both pairing and strand exchange. *Biochemistry* **94**, 11863–11868 (1997).

24. T.-R. Hahns, S. West, P. Howard-Flanders, RecA-mediated Strand Exchange Reactions between Duplex DNA Molecules Containing Damaged Bases, Deletions, and Insertions. *J Biol Chem* **263**, 7431–7436 (1988).

25. J. Ghanam, *et al.*, DNA in extracellular vesicles: from evolution to its current application in health and disease. *Cell Biosci* **12**, 37 (2022).

26. C. Molina-Santiago, P. Bernal, Nanotube-mediated plasmid transfer as a natural alternative for the improvement of industrially relevant bacteria. *Microb Biotechnol* **16**, 706–708 (2023).

27. J. Lederberg, E. L. Tatum, Gene Recombination in Escherichia coli. *Nature* **158**, 558–558 (1946).

28. L. S. Frost, "Genetics, Genomics | Conjugation, Bacterial" in *Encyclopedia of Microbiology*, 3rd Ed., M. Schaechter, Ed. (Academic Press, 2009), pp. 517–531.

29. M. Shintani, Z. K. Sanchez, K. Kimbara, Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Front Microbiol* **6** (2015).

30. F. Delavat, R. Miyazaki, N. Carraro, N. Pradervand, J. R. van der Meer, The hidden life of integrative and conjugative elements. *FEMS Microbiol Rev* **41**, 512–537 (2017).

31. J. S. Parkinson, Classic spotlight: The discovery of bacterial transduction. *J Bacteriol* **198**, 2899–2900 (2016).

32. Y. N. Chiang, J. R. Penadés, J. Chen, Genetic transduction by phages and chromosomal islands: The new and noncanonical. *PLoS Pathog* **15** (2019).

33. M. Morse, E. M. Lederberg, J. Lederberg, Transduction in Escherichia Coli K-12. *Genetics* **41**, 142–156 (1956).

34. N. D. Zinder, J. Lederberg, Genetic Exchange in Salmonella. *J Bacteriol* **64**, 679–699 (1952).

35. H. Schmieger, Short Communication Packaging Signals for Phage P22 on the Chromosome of Salmonella typhimurium. *Molec Gen Genet* **187**, 516–518 (1982).

36. L. M. Bobay, M. Touchon, E. P. C. Rocha, Pervasive domestication of defective prophages by bacteria. *Proc Natl Acad Sci U S A* **111**, 12127–12132 (2014).

37. R. J. Redfield, S. M. Soucy, Evolution of Bacterial Gene Transfer Agents. *Front Microbiol* **9** (2018).

38. C. Johnston, B. Martin, G. Fichant, P. Polard, J. P. Claverys, Bacterial transformation: Distribution, shared mechanisms and divergent control. *Nat Rev Microbiol* **12**, 181–196 (2014).

39. J. M. Solomon, A. D. Grossman, Who's competent and when: regulation of natural genetic competence in bacteria. *Trends in Genetics* **12**, 150–155 (1996).

40. C. H. G. Johnston, *et al.*, The RecA-directed recombination pathway of natural transformation initiates at chromosomal replication forks in the pneumococcus. *Proceedings of the National Academy of Sciences* **120** (2023).

41. A. Karnholz, *et al.*, Functional and topological characterization of novel components of the comB DNA transformation competence system in Helicobacter pylori. *J Bacteriol* **188**, 882–893 (2006).

42. M. Prudhomme, L. Attaiech, G. Sanchez, B. Martin, J.-P. Claverys, Antibiotic Stress Induces Genetic Transformability in the Human Pathogen Streptococcus pneumoniae. *Science (1979)* **313**, 89–92 (2006).

43. H. Steinmoen, E. Knutsen, L. S. Håvarstein, Induction of natural competence in Streptococcus pneumoniae triggers lysis and DNA release from a subfraction of the cell population. *Proceedings of the National Academy of Sciences* **99**, 7681–7689 (2002).

44. D. S. Guttman, D. E. Dykhuizen, Clonal Divergence in Escherichia coli as a Result of Recombination, Not Mutation. *Science (1979)* **266**, 1380–1383 (1994).

45. X. Didelot, D. J. Wilson, ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput Biol* **11** (2015).

46. X. Didelot, D. Falush, Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266 (2007).

47. N. J. Croucher, *et al.*, Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15 (2015).

48. R. Mostowy, *et al.*, Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol Biol Evol* **34**, 1167–1182 (2017).

49. M. Lin, E. Kussell, Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat Methods* **16**, 199–204 (2019).

50. S. Krishnan, *et al.*, Rhometa: Population recombination rate estimation from metagenomic read datasets. *PLoS Genet* **19** (2023).

51. G. A. T. McVean, *et al.*, The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. *Science (1979)* **304**, 581–584 (2004).

52. J. P. Spence, Y. S. Song, Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Sci Adv* **5** (2019).

53. A. R. Rogers, How Population Growth Affects Linkage Disequilibrium. *Genetics* **197**, 1329–1341 (2014).

54. M. Touchon, *et al.*, Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* **5** (2009).

55. L. M. Bobay, H. Ochman, The evolution of bacterial genome architecture. *Front Genet* **8** (2017).

56. D. F. Lato, G. B. Golding, The Location of Substitutions and Bacterial Genome Arrangements. *Genome Biol Evol* **13** (2021).

57. A. E. Shikov, I. A. Savina, A. A. Nizhnikov, K. S. Antonets, Recombination in Bacterial Genomes: Evolutionary Trends. *Toxins (Basel)* **15**, 568 (2023).

58. M. Badawi, I. Giraud, F. Vavre, P. Gre'Ve, R. Cordaux, Signs of neutralization in a redundant gene involved in homologous recombination in Wolbachia Endosymbionts. *Genome Biol Evol* **6**, 2654–2664 (2014).

59. Z. Zhang, *et al.*, Human-to-human transmission of Chlamydia psittaci in China, 2020: an epidemiological and aetiological investigation. *Lancet Microbe* **3**, e512–e520 (2022).

60. S. J. Joseph, X. Didelot, K. Gandhi, D. Dean, T. D. Read, Interplay of recombination and selection in the genomes of Chlamydia trachomatis. *Biol Direct* **6** (2011).

61. T. Wirth, *et al.*, Sex and virulence in Escherichia coli: An evolutionary perspective. *Mol Microbiol* **60**, 1136–1151 (2006).

62. C. Fraser, W. P. Hanage, B. G. Spratt, Recombination and the nature of bacterial speciation. *Science (1979)* **315**, 476–480 (2007).

63. Y. Sun, H. Luo, Homologous Recombination in Core Genomes Facilitates Marine Bacterial Adaptation. *Appl Environ Microbiol* **84** (2018).

64. M. Y. Galperin, Linear chromosomes in bacteria: No straight edge advantage? *Environ Microbiol* **9**, 1357–1362 (2007).

65. J.-N. Volff, J. Altenbuchner, A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol Lett* **186**, 143–150 (2006).

66. M. J. Szafran, *et al.*, Spatial rearrangement of the Streptomyces venezuelae linear chromosome during sporogenic development. *Nat Commun* **12** (2021).

67. M. Touchon, E. P. C. Rocha, Coevolution of the organization and structure of prokaryotic genomes. *Cold Spring Harb Perspect Biol* **8** (2016).

68. N. P. Robinson, *et al.*, Identification of Two Origins of Replication in the Single Chromosome of the Archaeon Sulfolobus solfataricus. *Cell* **116**, 25–38 (2004).

69. E. P. C. Rocha, The organization of the bacterial genome. *Annu Rev Genet* **42**, 211–233 (2008).

70. A. H. Syeda, J. U. Dimude, O. Skovgaard, C. J. Rudolph, Too Much of a Good Thing: How Ectopic DNA Replication Affects Bacterial Replication Dynamics. *Front Microbiol* **11** (2020).

71. W. M. Huang, *et al.*, Linear chromosome-generating system of agrobacterium tumefaciens C58: Protelomerase generates and protects hairpin ends. *Journal of Biological Chemistry* **287**, 25551–25563 (2012).

72. E. Couturier, E. P. C. Rocha, Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol Microbiol* **59**, 1506–1518 (2006).

73. M. J. Mclean, K. H. Wolfe, K. M. Devine, "Base Composition Skews, Replication Orientation, and Gene Orientation in 12 Prokaryote Genomes" (1998).

74. A. Grigoriev, Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* **26**, 2286–2290 (1998).

75. C. H. Kuo, N. A. Moran, H. Ochman, The consequences of genetic drift for bacterial genome complexity. *Genome Res* **19**, 1450–1454 (2009).

76. J. P. McCutcheon, N. A. Moran, Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **10**, 13–26 (2012).

77. H. L. Hendrickson, D. Barbeau, R. Ceschin, J. G. Lawrence, Chromosome architecture constrains horizontal gene transfer in bacteria. *PLoS Genet* **14** (2018).

78. H. N. Lim, Y. Lee, R. Hussein, Fundamental relationship between operon organization and gene expression. *Proc Natl Acad Sci U S A* **108**, 10626–10631 (2011).

79. S. M. Soucy, J. Huang, J. P. Gogarten, Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**, 472–482 (2015).

80. K. Yahara, *et al.*, The landscape of realized homologous recombination in pathogenic bacteria. *Mol Biol Evol* **33**, 456–471 (2016).

81. R. G. Everitt, *et al.*, Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. *Nat Commun* **5** (2014).

82. N. Malekian, A. A. Agrawal, T. U. Berendonk, A. Al-Fatlawi, M. Schroeder, A genome-wide scan of wastewater E. coli for genes under positive selection: focusing on mechanisms of antibiotic resistance. *Sci Rep* **12** (2022).

83. K. Yahara, X. Didelot, M. A. Ansari, S. K. Sheppard, D. Falush, Efficient inference of recombination hot regions in bacterial genomes. *Mol Biol Evol* **31**, 1593–1605 (2014).

84. N. J. Croucher, *et al.*, Rapid pneumococcal evolution in response to clinical interventions. *Science (1979)* **331**, 430–434 (2011).

85. J. M. Smith, N. H. Smith, M. O'rourke, B. G. Spratt, How clonal are bacteria? *Proc. Natl. Acad. Sci. USA* **90**, 4384–4388 (1993).

86. E. P. C. Rocha, Neutral theory, microbial practice: Challenges in bacterial population genetics. *Mol Biol Evol* **35**, 1338–1347 (2018).

87. A. E. Shikov, Y. V. Malovichko, A. A. Nizhnikov, K. S. Antonets, Current Methods for Recombination Detection in Bacteria. *Int J Mol Sci* **23**, 6257 (2022).

88. P. Marttinen, *et al.*, Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res* **40** (2012).

89. D. P. Martin, *et al.*, RDP5: A computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol* **7** (2021).

90. M. Lin, E. Kussell, Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat Methods* **16**, 199–204 (2019).

91. D. J. Wilson, *et al.*, Rapid evolution and the importance of recombination to the gastroenteric pathogen Campylobacter jejuni. *Mol Biol Evol* **26**, 385–397 (2009).

92. D. P. Martin, P. Lemey, D. Posada, Analysing recombination in nucleotide sequences. *Mol Ecol Resour* **11**, 943–955 (2011).

93. F. J. Medina-Aguayo, X. Didelot, R. G. Everitt, Speeding up Inference of Homologous Recombination in Bacteria. *Bayesian Anal* **1** (2023).

94. P. Marttinen, *et al.*, Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res* **40** (2012).

95. L. M. Bobay, B. S. H. Ellis, H. Ochman, ConSpeciFix: Classifying prokaryotic species based on gene flow. *Bioinformatics* **34**, 3738–3740 (2018).

96. L. M. Bobay, CoreSimul: A forward-in-time simulator of genome evolution for prokaryotes modeling homologous recombination. *BMC Bioinformatics* **21** (2020).

97. S. Mukherjee, *et al.*, Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. *Nucleic Acids Res* **51**, D957–D963 (2023).

98. J. Huerta-Cepas, *et al.*, Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol* **34**, 2115–2122 (2017).

99. M. Y. Galperin, *et al.*, COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* **49**, D274–D281 (2021).

100. E. Drula, *et al.*, The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res* **50**, D571–D577 (2022).

101. A. Kuzminov, Homologous Recombination—Experimental Systems, Analysis, and Significance. *EcoSal Plus* **4** (2011).

102. A. P. Camargo, *et al.*, You can move, but you can't hide: identification of mobile genetic elements with geNomad https:/doi.org/10.1101/2023.03.05.531206.

103. Z. Zhou, *et al.*, Transient darwinian selection in salmonella enterica serovar paratyphi a during 450 years of global spread of enteric fever. *Proc Natl Acad Sci U S A* **111**, 12199–12204 (2014).

104. L. J. Revell, phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* **3**, 217–223 (2012).

105. A. Stamatakis, RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).

106. P. D. Dixit, T. Y. Pang, F. W. Studier, S. Maslov, Recombinant transfer in the basic genome of Escherichia coli. *Proc Natl Acad Sci U S A* **112**, 9070–9075 (2015).

107. K. Yahara, *et al.*, The landscape of realized homologous recombination in pathogenic bacteria. *Mol Biol Evol* **33**, 456–471 (2016).

108. D. A. Kiktev, Z. Sheng, K. S. Lobachev, T. D. Petes, GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **115**, E7109–E7118 (2018).

109. F. Lassalle, *et al.*, GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. *PLoS Genet* **11**, 1–20 (2015).

110. J. L. Weissman, W. F. Fagan, P. L. F. Johnson, Linking high GC content to the repair of double strand breaks in prokaryotic genomes. *PLoS Genet* **15** (2019).

111. R. Raghavan, Y. D. Kelkar, H. Ochman, A selective force favoring increased G+C content in bacterial genes. *Proc Natl Acad Sci U S A* **109**, 14504–14507 (2012).

112. M. Touchon, *et al.*, Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* **5** (2009).

113. M. Vos, Why do bacteria engage in homologous recombination? *Trends Microbiol* **17**, 226–232 (2009).

114. A. Bartlett, D. Padfield, L. Lear, R. Bendall, M. Vos, A comprehensive list of bacterial pathogens infecting humans. *Microbiology (United Kingdom)* **168** (2022).

115. M. B. Bocchi, *et al.*, A rare case of bacillus megaterium soft tissues infection. *Acta Biomedica* **91**, 1–5 (2020).

116. B. T. Tierney, *et al.*, Multidrug-resistant Acinetobacter pittii is adapting to and exhibiting potential succession aboard the International Space Station. *Microbiome* **10** (2022).

117. H. Liu, V. Prajapati, S. Prajapati, H. Bais, J. Lu, Comparative Genome Analysis of Bacillus amyloliquefaciens Focusing on Phylogenomics, Functional Traits, and Prevalence of Antimicrobial and Virulence Genes. *Front Genet* **12** (2021).

118. V. C. Scoffone, *et al.*, Burkholderia cenocepacia infections in cystic fibrosis patients: Drug resistance and therapeutic approaches. *Front Microbiol* **8** (2017).

119. S. Simoni, *et al.*, The Emerging Nosocomial Pathogen Klebsiella michiganensis : Genetic Analysis of a KPC-3 Producing Strain Isolated from Venus Clam . *Microbiol Spectr* **11** (2023).

120. C. Tommasi, *et al.*, Diagnostic difficulties of Lactobacillus casei bacteraemia in immunocompetent patients: A case report. *J Med Case Rep* **2** (2008).

121. J. Chery, D. Dvoskin, F. P. Morato, B. Fahoum, Lactobacillus fermentum, a pathogen in documented cholecystitis. *Int J Surg Case Rep* **4**, 662–664 (2013).

122. S. Giuliano, *et al.*, “Lactococcus lactis blood products contamination resulting in fatal human case: insights from a forensic case” (2023).

123. L. Victoria, A. Gupta, J. L. Gómez, J. Robledo, Mycobacterium abscessus complex: A Review of Recent Developments in an Emerging Pathogen. *Front Cell Infect Microbiol* **11** (2021).

124. A. C. Büchler, *et al.*, Mycobacterium chelonae infection identified by metagenomic next-generation sequencing as the probable cause of acute contained rupture of a biological composite graft— a case report. *Int J Mol Sci* **23** (2022).

125. P. da Silva Campana, *et al.*, Rhodococcus hoagii bloodstream infection in an allogeneic hematopoietic stem cell transplantation patient: Case report and review of literature. *IDCases* **20** (2020).

126. P. Xu, *et al.*, Genome of the opportunistic pathogen Streptococcus sanguinis. *J Bacteriol* **189**, 3166–3175 (2007).

127. S. R. Eddy, Accelerated profile HMM searches. *PLoS Comput Biol* **7** (2011).

128. L. M. Bobay, H. Ochman, Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol Biol* **18** (2018).

129. K. Raymann, C. Brochier-Armanet, S. Gribaldo, The two-domain tree of life is linked to a new root for the Archaea. *Proc Natl Acad Sci U S A* **112**, 6670–6675 (2015).

130. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**, 772–780 (2013).

131. C. D. Harris, E. L. Torrance, K. Raymann, L. M. Bobay, Corecruncher: Fast and robust construction of core genomes in large prokaryotic data sets. *Mol Biol Evol* **38**, 727–734 (2021).

132. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

133. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).

134. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

135. L. M. Bobay, H. Ochman, Impact of Recombination on the Base Composition of Bacteria and Archaea. *Mol Biol Evol* **34**, 2627–2636 (2017).

136. X. Didelot, D. Lawson, A. Darling, D. Falush, Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186**, 1435–1449 (2010).

137. K. Csilléry, O. François, M. G. B. Blum, Abc: An R package for approximate Bayesian computation (ABC). *Methods Ecol Evol* **3**, 475–479 (2012).

138. K. Yahara, *et al.*, Genome-wide survey of mutual homologous recombination in a highly sexual bacterial species. *Genome Biol Evol* **4**, 628–640 (2012).

139. J. Hey, What's so hot about recombination hotspots? *PLoS Biol* **2** (2004).

140. Y. Uehara, Current Status of Staphylococcal Cassette Chromosome mec (SCCmec). *Antibiotics* **11** (2022).

141. M. M. Dillon, W. Sung, M. Lynch, V. S. Cooper, Periodic Variation of Mutation Rates in Bacterial Genomes Associated with Replication Timing. *mBio* **9** (2018).

142. S. Million-Weaver, *et al.*, An underlying mechanism for the increased mutagenesis of lagging-strand genes in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences* **112** (2015).

143. M. Fondi, G. Emiliani, R. Fani, Origin and evolution of operons and metabolic pathways. *Res Microbiol* **160**, 502–512 (2009).

144. B. J. Arnold, I.-T. Huang, W. P. Hanage, Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol* **20**, 206–218 (2022).

145. N. L. Craig, THE MECHANISM OF CONSERVATIVE SITE-SPECIFIC RECOMBINATION. *Annu Rev Genet* **22**, 77–105 (1988).

146. S. Austin, M. Ziese, N. Sternberg, A novel role for site-specific recombination in maintenance of bacterial replicons. *Cell* **25**, 729–736 (1981).

147. F. Lassalle, *et al.*, GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. *PLoS Genet* **11**, e1004941 (2015).

148. J. Bohlin, V. Eldholm, J. H. O. Pettersson, O. Brynildsrud, L. Snipen, The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics* **18**, 151 (2017).

149. B. Michel, *et al.*, Rescue of arrested replication forks by homologous recombination. *Proceedings of the National Academy of Sciences* **98**, 8181–8188 (2001).

150. C.-H. Wu, *et al.*, Bacillus subtilis YngB contributes to wall teichoic acid glucosylation and glycolipid formation during anaerobic growth. *Journal of Biological Chemistry* **296**, 100384 (2021).

151. S. D. Aggarwal, *et al.*, A molecular link between cell wall biosynthesis, translation fidelity, and stringent response in *Streptococcus pneumoniae*. *Proceedings of the National Academy of Sciences* **118** (2021).

152. K. Gera, T. Le, R. Jamin, Z. Eichenbaum, K. S. McIver, The Phosphoenolpyruvate Phosphotransferase System in Group A Streptococcus Acts To Reduce Streptolysin S

Activity and Lesion Severity during Soft Tissue Infection. *Infect Immun* **82**, 1192–1204 (2014).

153. J. L. Baker, *et al.*, Transcriptional profile of glucose-shocked and acid-adapted strains of *Streptococcus mutans*. *Mol Oral Microbiol* **30**, 496–517 (2015).

154. M. Aswal, A. Garg, N. Singhal, M. Kumar, Comparative in-silico proteomic analysis discerns potential granuloma proteins of Yersinia pseudotuberculosis. *Sci Rep* **10**, 3036 (2020).

155. L. Balabanova, L. Averianova, M. Marchenok, O. Son, L. Tekutyeva, Microbial and Genetic Resources for Cobalamin (Vitamin B12) Biosynthesis: From Ecosystems to Industrial Biotechnology. *Int J Mol Sci* **22**, 4522 (2021).

156. C. A. Rowley, M. M. Kendall, To B12 or not to B12: Five questions on the role of cobalamin in host-microbial interactions. *PLoS Pathog* **15**, e1007479 (2019).

157. D. G. Glanville, *et al.*, A High-Throughput Method for Identifying Novel Genes That Influence Metabolic Pathways Reveals New Iron and Heme Regulation in Pseudomonas aeruginosa. *mSystems* **6** (2021).

158. L. Zamorano, *et al.*, The Pseudomonas aeruginosa CreBC Two-Component System Plays a Major Role in the Response to β-Lactams, Fitness, Biofilm Growth, and Global Regulation. *Antimicrob Agents Chemother* **58**, 5084–5095 (2014).

159. S. F. Li, J. A. DeMoss, Promoter region of the nar operon of Escherichia coli: nucleotide sequence and transcription initiation signals. *J Bacteriol* **169**, 4614–4620 (1987).

160. E. D. Peng, M. P. Schmitt, Identification of zinc and Zur-regulated genes in Corynebacterium diphtheriae. *PLoS One* **14**, e0221711 (2019).

161. X. Didelot, G. Méric, D. Falush, A. E. Darling, “Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli” (2012).

162. B. A. Niccum, H. Lee, W. MohammedIsmail, H. Tang, P. L. Foster, The Symmetrical Wave Pattern of Base-Pair Substitution Rates across the Escherichia coli Chromosome Has Multiple Causes. *mBio* **10** (2019).

163. I. J. Fijalkowska, P. Jonczyk, M. M. Tkaczyk, M. Bialoskorska, R. M. Schaaper, Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. *Proceedings of the National Academy of Sciences* **95**, 10020–10025 (1998).

164. S. Paul, S. Million-Weaver, S. Chattopadhyay, E. Sokurenko, H. Merrikh, Accelerated gene evolution through replication–transcription conflicts. *Nature* **495**, 512–515 (2013).

165. E. R. Reichenberger, G. Rosen, U. Hershberg, R. Hershberg, Prokaryotic Nucleotide Composition Is Shaped by Both Phylogeny and the Environment. *Genome Biol Evol* **7**, 1380–1389 (2015).

166. S. M. Fullerton, A. Bernardo Carvalho, A. G. Clark, Local Rates of Recombination Are Positively Correlated with GC Content in the Human Genome. *Mol Biol Evol* **18**, 1139–1142 (2001).

167. R. Hershberg, D. A. Petrov, Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLoS Genet* **6**, e1001115 (2010).

168. C. L. C. Wielders, A. C. Fluit, S. Brisse, J. Verhoef, F. J. Schmitz, *mecA* Gene Is Widely Disseminated in *Staphylococcus aureus* Population. *J Clin Microbiol* **40**, 3970–3975 (2002).

169. P. H. Oliveira, M. Touchon, E. P. C. Rocha, Regulation of genetic flux between bacteria by restriction–modification systems. *Proceedings of the National Academy of Sciences* **113**, 5658–5663 (2016).