Quantifying plasmid movement in drug-resistant *Shigella* species using phylodynamic inference

Nicola F. Müller[a,b,1], **Ryan R. Wick**[c,], **Louise M. Judd**[d], **Deborah A. Williamson**[e,f], **Trevor Bedford**[b,g], **Benjamin P. Howden**[c,d], **Sebastián Duchêne**[c,h,1,2], **and Danielle J. Ingle**[c,d,1,2]

[a]Division of HIV, ID and Global Medicine, University of California San Francisco, CA, USA; [b]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA; [c]Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, VIC, AUS; [d]Centre for Pathogen Genomics, University of Melbourne (Doherty Institute); [e]Department of Infectious Diseases at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, VIC, AUS; [f]School of Medicine, University of St Andrews, Fife, Scotland; [g]Howard Hughes Medical Institute, Seattle, WA, USA; [h]EDID unit, Department of Computational Biology, Institut Pasteur, Paris, France

The 'silent pandemic' of antimicrobial resistance (AMR) represents a significant global public health threat. AMR genes in bacteria are often carried on mobile elements, such as plasmids. The horizontal movement of plasmids allows AMR genes and resistance to key therapeutics to disseminate in a population. However, the quantification of the movement of plasmids remains challenging with existing computational approaches. Here, we introduce a novel method that allows us to reconstruct and quantify the movement of plasmids in bacterial populations over time. To do so, we model chromosomal and plasmid DNA co-evolution using a joint coalescent and plasmid transfer process in a Bayesian phylogenetic network approach. This approach reconstructs differences in the evolutionary history of plasmids and chromosomes to reconstruct instances where plasmids likely move between bacterial lineages while accounting for parameter uncertainty. We apply this new approach to a five-year dataset of *Shigella*, exploring the plasmid transfer rates of five different plasmids with different AMR and virulence profiles. In doing so, we reconstruct the co-evolution of the large *Shigella* virulence plasmid with the chromosome DNA. We quantify higher plasmid transfer rates of three small plasmids that move between lineages of *Shigella sonnei*. Finally, we determine the recent dissemination of a multidrug-resistant plasmid between *S. sonnei* and *S. flexneri* lineages in multiple independent events and through steady growth in prevalence since 2010. This approach has a strong potential to improve our understanding of the evolutionary dynamics of AMR-carrying plasmids as they are introduced, circulate, and are maintained in bacterial populations.

Phylodynamics | Antimicrobial resistance | phylogenetic network | Bayesian phylogenetics | Bacterial phylogenetics |BEAST

Antimicrobial resistance (AMR) in bacteria represents one of the most serious public health threats of the 21st century (1, 2, 3). Bacterial pathogens can evolve AMR either through mutations in core genes or via the acquisition of AMR genes by horizontal gene transfer on mobile genetic elements, such as plasmids. The horizontal transfer of genetic material, including AMR genes, allows bacterial populations to rapidly adapt and evolve under changing selective pressures and ecological niches (4, 5, 6). Plasmids are typically considered to be part of the accessory genome; genes that are variably present across genomes that are drawn from the species pangenome (7, 8, 9). Some plasmids can move between lineages of the same bacterial species or between unrelated bacterial species (8). Multiple studies to date have identified plasmids as playing central roles in driving the emergence, spread, and increasing prevalence of AMR in several bacterial species, including *Klebsiella pneumoniae*, *Salmonella enterica*, *Escherichia coli*, and *Shigella* species (5, 10, 11, 12, 13). Quantifying the horizontal movement of plasmids in populations and modeling rates of transfer of plasmids in bacterial species where drug resistance is often driven by plasmids is a major barrier to our understanding of the drivers of the prevalence and dissemination of AMR. Hence, there is an unmet need for novel computational approaches to quantify the movement and spread of mobile elements in the accessory genome of bacterial pathogens (3).

Shigella is an exemplar bacterial pathogen to explore the methodological approaches to quantify the evolution and movement of plasmids. There are four species of *Shigella*, with two species, *Shigella sonnei* and *Shigella flexneri*, responsible for the global burden of Shigellosis (14, 15, 16, 17). *Shigella* is a WHO AMR priority pathogen due to increased resistance to clinical therapeutics (2). *Shigella* has been shown to carry different plasmids that vary in size and function (AMR, virulence), and with different evolutionary histories (12, 13, 14, 16). While the pINV plasmid is considered part of the core genome, plasmid content varies within and between the *Shigella* species. For example, in *S. sonnei*, three other smaller plasmids have been characterized in the reference genome Ss046. These three small plasmids, spA, spB, and spC, are commonly found within *S. sonnei* global lineage III (16). More recently, attention has focused on the movement and spread of a large multidrug-resistant plasmid (MDR; resistant to three or more antimicrobials), particularly in men who have sex with men (MSM) (12, 13, 18, 19, 20, 21). Importantly, these outbreaks have been driven by variants of the MDR plasmid, pKSR100, which mediates AMR to co-trimoxazole, azithromycin and ampicillin, moving between populations of *Shigella*, (12, 13, 19, 22, 23). The horizontal transfer of plasmids occurs between bacterial lineages and, as such, is not a co-divergent process with the chromosomal DNA of these bacterial lineages.

Here, we model this process by using the novel method of a 'coalescent with plasmid transfer' (CoalPT) model. CoalPT is an approach to reconstructing the acquisition, movement, and co-divergence of plasmids in routinely generated WGS data. Our method is implemented as a package for the open-source software BEAST2 (24) to facilitate its adoption. The plasmid transfer rate is a population level rate and a function of how often bacterial lineages are in the same location, the probability of

them exchanging plasmids (if they have one), and also the degree of selection that acts on the bacterium that picked up a new plasmid. We used empiric data from *Shigella* isolates to explore and quantify the plasmid transfer rates of several plasmids within *Shigella* that differ in size, AMR and virulence profile, and expected evolutionary history. We then show that modeling the co-divergence of plasmid and chromosomal DNA enables inference of the rate over which plasmids accrue mutations (known as the 'rate of evolution') with high precision and accuracy, despite limited genomic information.

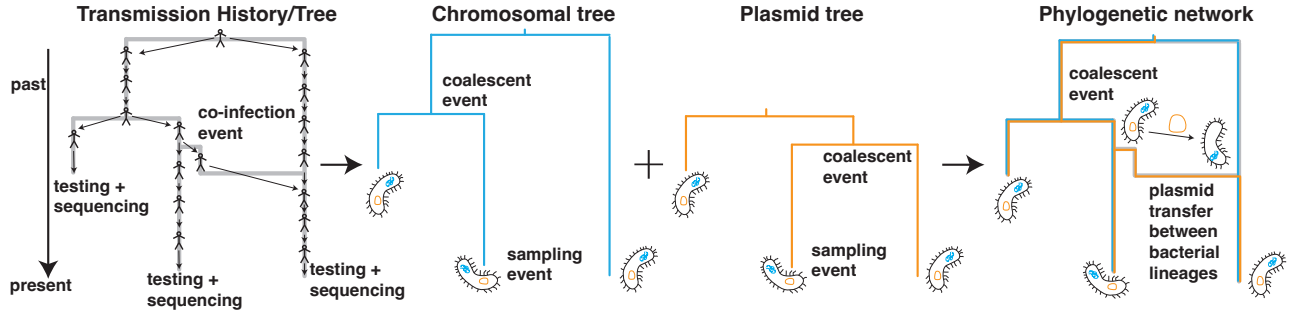


Fig. 1. Schematic representation of the coalescent with plasmid transfer model. The transmission history of bacterial lineages, including the movement of plasmid between lineages, can be described as a transmission network. From that transmission history, we can track the history of the chromosomal and plasmid DNA individually. The history of both chromosomal and plasmid DNA can be described by using a tree. The coalescent with plasmid transfer models a backward-in-time process where any two network lineages can coalesce (share a common ancestor). Additionally, network lineages can undergo a plasmid transfer event, modeled backward in time as one of the plasmid lineages branching off the main branch. How rapidly two lineages share a common ancestor backward in time is given by the effective population size, and the rates of plasmid transfer denote the rate of observing plasmid transfer events backward in time.

Results

In CoalPT we describe the movement of plasmids between bacterial lineages as a joint coalescent and plasmid transfer process (Figure 1), where lineages can coalesce from present to past or undergo a plasmid transfer event, similar to how recombination is often modeled (25). The model has two key parameters: the effective population size (Ne) and the plasmid transfer rate (ρ). These two parameters determine the rates at which coalescent and plasmid transfer events occur. The plasmid transfer events are agnostic about the precise biological mechanism under which plasmids move between bacteria. The result of the CoalPT model is a timed phylogenetic network with each lineage of the network corresponding to one or more lineages of either the chromosome or plasmid trees. As such, the co-evolutionary history of the chromosome and the plasmid is denoted the timed phylogenetic network in which the chromosome and plasmid trees are embedded. To perform inference under the CoalPT model, we use a Markov chain Monte Carlo (MCMC) sampling technique for the timed phylogenetic network that is related to the MCMC inference of re-assortment (26) and recombination networks (27). Using an MCMC approach allows us to infer the phylogenetic network, effective population sizes, plasmid transfer rates, and evolutionary parameters, all while accounting for uncertainty in the network and parameter estimates.

Plasmid transfer rates differ depending on plasmid size and function. To investigate how plasmids move between *S. sonnei* lineages, we reconstructed the joint evolutionary history of *S. sonnei* chromosomal DNA and four plasmids (pINV, spA, spB & spC) (Figure 2, Supplementary Figure S1). We further considered two additional alignments of the spA plasmid to explore the potential use of this new method in reconstructing the movement of specific mobile AMR elements within populations. This small plasmid contains up to four AMR genes that include *strA* & *B, sul2* , and sometimes *tet(A)* that were established to move between lineages. For spA, the number of reconstructed transfer events strongly depends on which part of the spA plasmid is used for the analyses (Supplementary Figure S6). For the first analyses (*entire spA*), we used the entire spA plasmid for isolates that had ≥70% coverage of the spA plasmid, regardless of AMR profile, with a focus on tracking the plasmid. In a second set of analyses we used an alignment of the four AMR genes *strA* & *B, sul2* and *tet(A)*, that included the region between *strB* and *tetA* which also encoded a transposase and *tetR* (*AMR genes only* dataset) (see Methods). Finally, we considered only the genes coding for *strA* & *B* and *sul2* and the flanking region of ∼100 bases of *strB* (*strA* & *B* + *sul* + flanking dataset) as these three genes are known to frequently move together. Supplementary Figure S2 shows the location of these genes on the spA plasmid and it highlights the regions of the spA plasmid used. We also performed long-read sequencing using ONT platforms of selected isolates (see Methods) that are members of the clade highlighted in red in S7 to determine the location of AMR genes in genomes with different coverage of the spA reference genome (Supplementary Figure S2). We made this distinction in part due to evidence that the AMR genes carried on the spA plasmid had integrated into the chromosome for some isolates for which we had lower coverage of the spA plasmid. This lower coverage suggested that these AMR genes were not being carried on the spA plasmid in these genomes (Supplementary Figure S2).

To illustrate the differences, we reconstructed a tanglegram from chromosomal and spA timed plasmid trees inferred using IQ-TREE 2 and TreeTime (28, 29, 30). As shown in Supplementary Figure S7, there is evidence for multiple rearrangements between the chromosome and plasmid trees, indicating plasmids are moving. The 70% coverage cutoff used for the *entire spA* tree, removes a group of isolates highlighted in red in Supplementary Figure S7, suggesting either that the genes present on spA changed or that the AMR genes integrated into a different plasmid or the chromosome (Supplementary Figure S2).

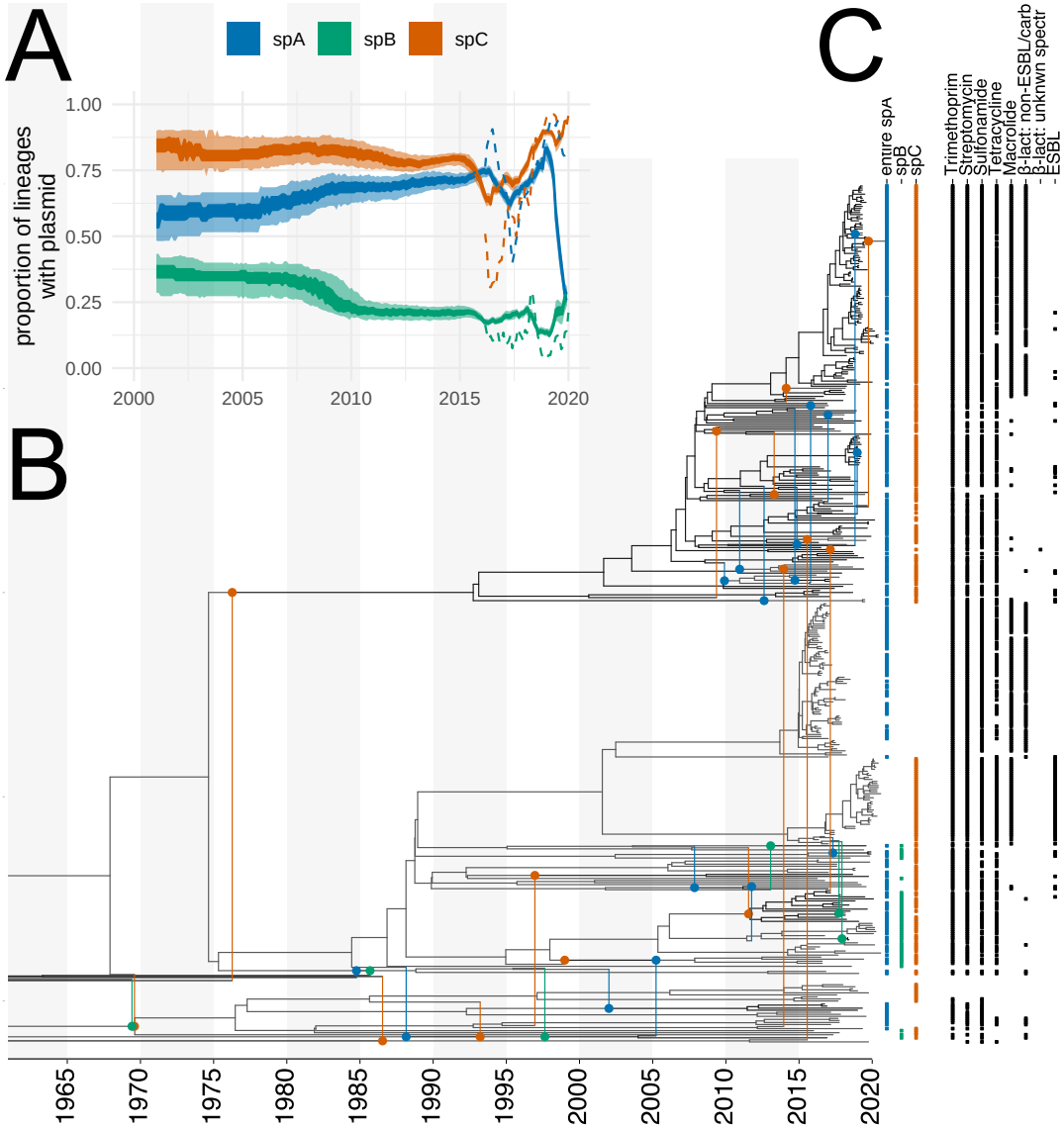


Fig. 2. Co-divergence of the core chromosome and plasmids in *Shigella sonnei*. A Proportion of lineages carrying a plasmid between 2000 and 2020. The inner shaded areas denote the 50% HPD, and the outer area is the 95% HPD. The dotted lines denote the proportion of samples with a plasmid. **B** Here, we show the maximum clade credibility (MCC) network of *Shigella sonnei* inferred using the chromosomal DNA, the virulence plasmid pINV, and the small plasmids spA, spB, and spC. Vertical lines are used to denote plasmid transfer events, where the circles denote the branch to which a plasmid was transferred. The color of the circle denotes either spA, spB, or spC, having jumped between bacterial lineages. The dashed lines correspond to branches from which plasmids branch off. The text denotes the posterior probability of plasmid transfer events for events with a posterior support of over 0.5. Branches are colored to denote separate clades originating from individual jumps of plasmids into a bacterial clade. **C** The tip labels in blue, green, and red denote if a plasmid was detected at a leave. The black dots denote the presence of antimicrobial resistance to the antimicrobials on top.

We then reconstructed the movement of the plasmids between bacterial lineages using CoalPT in three different analyses using the different spA alignments. Since CoalPT assumes that there is no inter-lineage recombination within, we masked sections with evidence of recombination in the chromosome but assumed that there is no intralineage recombination on the plasmid alignments (see Methods). We used a separate strict molecular clock and HKY+Γ_4 (31) substitution model for the chromosome and for each plasmid. We additionally assume a constant-size coalescent process and infer the effective population size and the rate of plasmid transfer, allowing each plasmid to have a different transfer rate. The prevalence of the smaller plasmids was relatively constant over the sampling period (Supplementary Figure S1C), with spA and spC being frequently detected in different lineages and genotypes of *S. sonnei* (Supplementary Figures S2, S3). In contrast, the spB plasmid was predominately found in *S. sonnei* isolates belonging to genotypes that are part of global lineage III, although it was detected in six isolates in lineage 2 (Supplementary Table 1). We found little to no support for the virulence plasmid, pINV, having been transferred between different bacterial lineages, suggesting co-divergence of the chromosome and pINV (Supplementary Figure S4). This finding is consistent with the known evolution of *Shigella* species. Importantly, the spA plasmid was inferred to have the highest transfer rate between bacterial lineages of *S. sonnei*. The spA plasmid is the only small plasmid that carried AMR determinants (Supplementary Figure S5). The other two small plasmids, spB and spC, displayed lower rates of

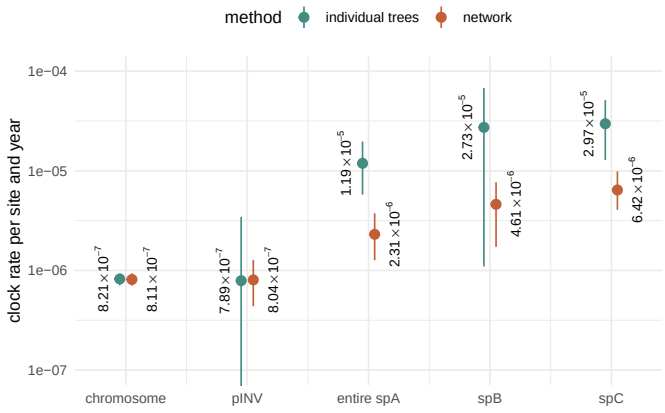


Fig. 3. Rates of evolution for plasmids and core chromosome in *Shigella sonnei*. Here, we compare estimates inferred by assuming an individual rate of evolution for the chromosome and plasmids (in green) to those where we explicitly model the joint evolutionary history of these lineages as a phylogenetic network (in orange). The posterior estimates of the evolutionary rate of the chromosomal and plasmid DNA of *S. sonnei* sequences isolated in Melbourne, Australia, over several years.

plasmid transfer (Supplementary Figure S5). These two plasmids were inferred to have high copy numbers based on relative read depth to the chromosome from Oxford Nanopore Technologies (ONT) data and were not characterized to have any direct ability to mobilize.

The number of inferred plasmid transfer events was the highest when using entire spA plasmids and the lowest when only using AMR genes. The same patterns hold when looking at the rate at which plasmids are transferred (Supplementary Figure S5 and Supplementary Figure S8) with the entire spA showing the highest transfer rate. This demonstrates the utility of this novel method for the reconstruction of the movement of plasmids but also shows the limitation of this approach, particularly for small AMR mobile elements, as there may be a lack of phylogenetic signal.

We next computed the rate at which plasmids are being lost in *S. sonnei*. We calculated the number of times a plasmid has been lost as the number of child edges (i.e., branches) in a network for which the parent branch carries a plasmid while the child branch itself does not. We then divide this number by the total length of the plasmid tree to get an estimate of the rate at which the plasmid is lost in units of plasmid loss events per unit time. We restrict this analysis to events in the last 10 years since sample collection.

The virulence plasmid, which in *S. sonnei* is known to be often lost in culture (14, 32), but it forms part of the core genome of all *Shigella* species, had the highest rate of being lost (Supplementary Figure S9). This was expected given the known loss in culture. The smaller plasmids were all lost at a lower rate relative to the pINV plasmid of *S. sonnei* (Supplementary Figure S9). These patterns may, at least in part, be driven by the plasmids not being detected or being lost in culture, but may also reflect that these smaller plasmids have limited fitness costs and, as such, are readily maintained.

Accounting for the co-divergence of chromosomes and plasmids is essential to estimating rates of evolution in plasmids. The evolution of the genome of bacterial species occurs, in part, as a result of selective pressures on the core and accessory genome. The core will likely be under strong selective constraints, while the accessory may be subject to weaker selection. Indeed, we find that plasmids tend to have higher molecular evolutionary clock rates than those of the chromosome (Supplementary Figure 3 & S10). SNPs within the bacterial chromosome have been the focus of bacterial phylodynamics to date due to enough temporal signal in the sequence data to model the population dynamics, facilitated by the bacterial chromosome being orders of magnitude larger than some plasmids. In the case of *S. sonnei*, the chromosome has approximately 22 times more nucleotides than the virulence plasmid, pINV, and between ∼570 to ∼2300 times more than spA–spC. In spite of the higher rates of evolution in plasmids, compared to the chromosome, we would still expect an alignment of SNPs in the core chromosome to have a larger number of SNPs due to its sheer size. As such, plasmids are less likely to contain as much information as the chromosome and thus be less likely to behave as measurably evolving populations (33, 34).

To illustrate how modeling the co-divergence of the chromosomal and plasmid DNA impacts inferences of the evolutionary rate, we reconstructed the phylogenetic trees of the chromosomes, virulence (pINV), spA, spB, and spC plasmids individually (Figure 3). For the chromosome and the pINV, we used SNP alignments, which only contain the SNPs, in order to reduce the size of the dataset. For spA–spC, we used the full alignments (with gaps, Ns, and both variant and invariant sites) obtained from alignment to the reference genomes (see Methods). We used the same priors and evolutionary models as for the network inference described above and then inferred the phylogenetic trees, evolutionary rates and other parameters. As shown in Supplementary Figure S10A, we found the chromosome to evolve at a rate with mean 7.8×10^{-7} subs/site/year (95% highest posterior density, HPD 6.6×10^{-7} -9.4×10^{-7}), and the virulence plasmid to evolve at a rate with mean 9.3×10^{-7} subs/site/year (95% HPD $5.8 \times 10^{-7} - 1.3 \times 10^{-6}$). The small plasmids spA–spC all evolve at substantially higher rates, with means of between 2.9×10^{-6} and 1.9×10^{-5} subs/site/year. Importantly, inferring these rates of evolution would be impossible using the plasmid alignments alone and thus require information about the co-divergence of the plasmids and the chromosome.

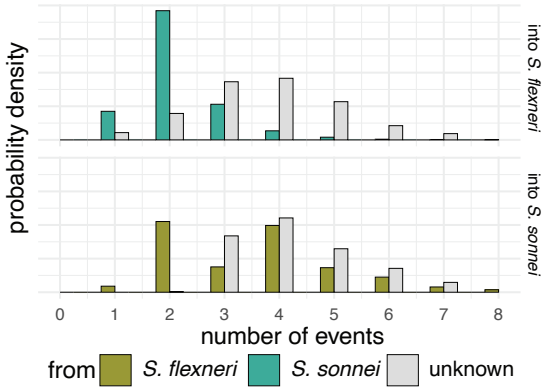


Fig. 4. Frequency of crossspecies movement of pKSR100 between *S. sonnei* and *flexneri*. Here, we show the posterior estimate of the plasmid transfer rate between bacterial lineages. The plasmid transfer rate is given per lineage per year and denotes the rate at which we expect to observe a plasmid transfer event when tracking the history of a bacterial lineage backward in time.

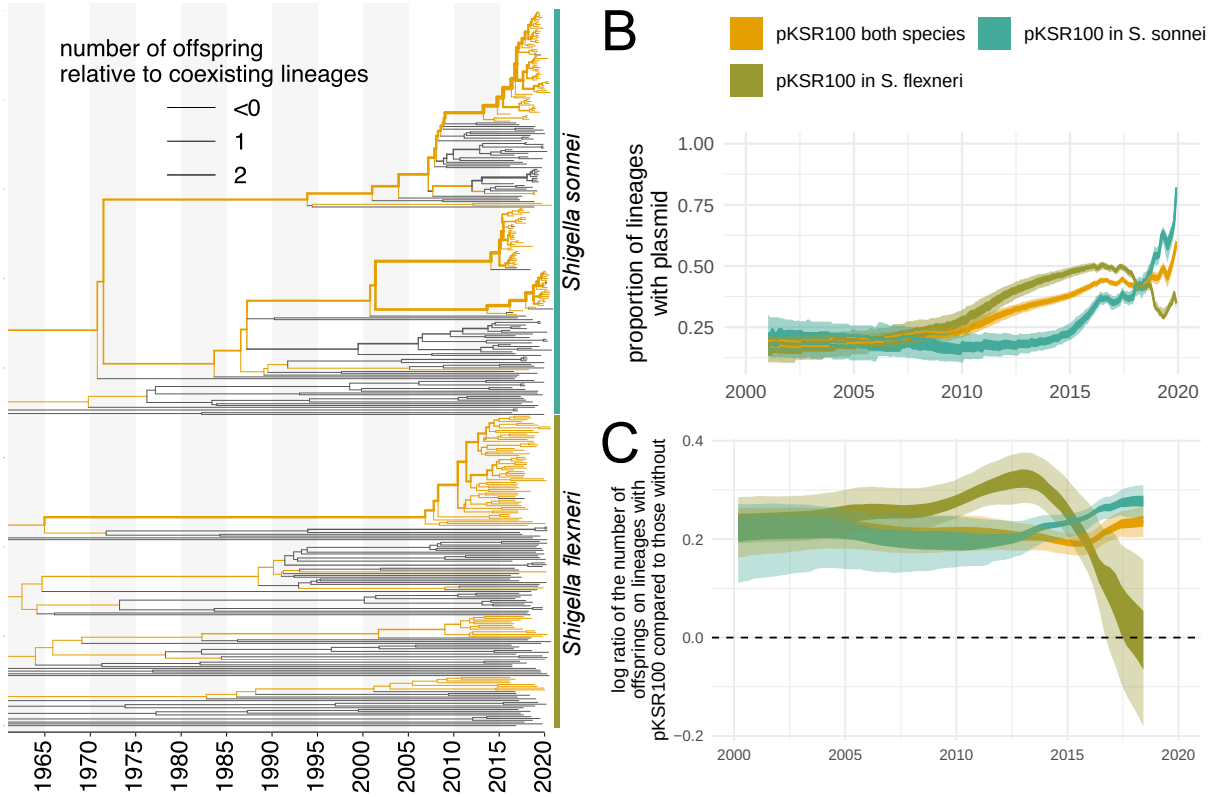


Fig. 5. Frequency of *S. sonnei* and *S. flexneri* clusters with and without pKSR100 plasmid. A Chromosomal trees of *S. sonnei* and *S. flexneri* with the presence or absence of pKSR100 mapped onto the plasmid tree. The width of the branches denotes the log standardized cluster size below a node relative to all other co-existing lineages. **B** Proportion of lineages with and without pKSR100 since 2020. The inner interval denotes the 50% HPD and the outer the 95% computed using the logged trees. **C** The average log standardized cluster sizes below nodes with pKSR100 compared to co-existing lineages without pKSR100.

To further explore the impact of our approach on estimates of evolutionary rates, we compared the inferred rates for plasmids using the coalescent with plasmid transfer and individual tree inference using simulations. As shown in Supplementary Figure S11, using tree inference only to retrieve rates of evolution will return the prior on the evolutionary rate, even for cases with relatively many SNPs, implying that the data are not sufficiently informative to drive the estimate of this parameter. The reason is that even in cases with many SNPs in total, the number of SNPs per time that one expects to occur over the sampling period of five years is only $5\,\text{years} \times 200\,\text{bp} \times 5 \times 10^{-4}$ subs/site/year $= 0.5$ SNPs for the largest plasmid. The network approach, on the other hand, is able to infer the rates of evolution of plasmids even when only a few SNPs occur (Supplementary Figure S10B) because the chromosome data act as a form of molecular clock calibration and thus there is more data available for inference.

Evidence for cross-species MDR plasmid exchange and steady growth of pKSR100 prevalence. We next investigated the movement of an MDR plasmid that has been previously well-characterized using genomic epidemiological approaches to be moving within and between lineages of two *Shigella* species, *S. sonnei* and *S. flexneri*. To do so, we compiled three alignments. We made an alignment from SNPs in the reference chromosome for both *S. sonnei* (n = 789 isolates) and *S. flexneri* (n = 297 isolates) individually (see Methods). For the MDR plasmid (pKSR100) known to circulate in both species, we aligned sequences from both species jointly. All *S. sonnei* and *S. flexneri* where isolates that had ≥80% coverage of the reference plasmid were included in the alignment and a core SNP alignment with ≥95% conservation of sites (Supplementary Figure S12)(see Methods). We then randomly sub-sampled 250 isolates equally from *S. sonnei* and *S. flexneri* that carried the pKSR100-like plasmids. The chromosomal DNA of *S. sonnei* and *S. flexneri* were assumed to be their individual trees, while all samples of the pKSR100 plasmids were assumed to be from the same trees. We next reconstructed the joint evolutionary history of the core chromosome and the MDR plasmid assuming a strict molecular clock for both the chromosome and the MDR plasmid and a HKY+Γ_4 substitution model. In order to improve the computational efficiency, we fixed the rate of evolution of the core chromosomes to be equal to the estimates in Supplementary Figure S10, while estimating the rate of evolution of the MDR plasmid.

We found evidence for multiple events where the MDR plasmid jumped between bacterial lineages within species and also between species (Figure 4, Supplementary Figure S13). These jumps between lineages were, in some cases, associated with a rapid expansion of a clade. For example, we found that the *S. sonnei* clade expanded after the introduction of an MDR plasmid into the bacterial lineage from *S. flexneri* around 2010. We next sought to distinguish introductions of the MDR plasmid into *S. sonnei* and *S. flexneri* clades by whether they likely originated from the other bacterial species or from an unknown species entirely. To do so, we followed the procedure described in *Directionality of plasmid transfer*. Additionally, we only considered plasmid transfer events that were introduced into *S. sonnei* or *S. flexneri* in the last 50 years. There is evidence for multiple introductions of plasmids into both species from each other (Figure 4), but also from unknown bacterial lineages, which could be other *Shigella* lineages or from other bacterial species in the same ecological niches, as has been previously reported (35). We next computed the proportion of lineages in the past that carried the plasmid pKSR100. As shown in Figure 5, we find a steady increase in the proportion of bacterial lineages that carry the pKSR100 plasmid. This increase is inferred to start around the year 2010 and to continue relatively steady until 2020 when we have the most recent samples in the dataset.

Discussion

Our novel inference approach represents a substantial advancement in the field of bacterial population genomics by enabling a more comprehensive understanding of the plasmid movements within bacterial pathogens over time from routinely generated WGS short-read data. To date, studies exploring the movement of specific plasmids have relied on long-read sequencing data and have been undertaken in localized settings over shorter timespans (4, 10, 36). CoalPT represents an approach that can leverage the short-read data generated from genomic surveillance efforts in public health laboratories, providing a method to explore the movement and transfer of plasmids in bacterial datasets, all while accurately accounting for uncertainty in the data.

The model behind CoalPT offers the potential for integrating geographical data through phylogeographic approaches (37, 38). Such advances will be critical in exploring, monitoring, and potentially modeling future scenarios of how drug-resistant plasmids of interest are moving in large-scale datasets of high-priority pathogens. Reconstructing the movement of plasmids over time has been difficult, but is increasingly of interest to better understand the evolutionary dynamics shaping potential plasmid-driven outbreaks. Importantly, CoalPT is tailored for tracking plasmids, with a complete reference sequence, meaning that the quality of plasmid alignments warrants special attention. Our approach of varying the coverage of the plasmid, core-SNP filter, and use of the consistency index for quantifying homoplasy (39) can help guide assessments of the impact of data quality on inference of plasmid dynamics. Explicitly modeling the co-divergence of plasmids and core genomes also allows us to quantify the number of these events, the timing of introductions, and the lineages from where plasmids originate that are later introduced into other lineages or species. As such, this framework is amenable to studying other bacterial populations where the plasmid dynamics are less clear. In line with other research, we find a high degree of co-divergence of virulence plasmid, pINV, with the chromosome of *S. sonnei* (14), and the movement of small plasmids within the *S. sonnei* population. Furthermore, we find evidence for multiple MDR plasmid transfer events between *S. sonnei* lineages, but also between *S. sonnei* and *S. flexneri* lineages (12, 13, 19, 22).

We note that our method to quantify plasmid transfer rates is not suited to tracking unknown plasmids, it is suited to exploration of specific plasmids of interest. These approaches would be immediately relevant to drug-resistant plasmids that may be driving population expansions, but could also be extended to virulence plasmids or where there has been reported convergence of AMR and virulence (40). We also demonstrated that it is not suited for tracking the movement of specific AMR genes within populations in the absence of any other phylogenetic context. Instead, our approach addresses the question of whether the expansion of a particular plasmid was due to a single introduction and subsequent expansion, or due to repeated introductions from different sources. Such insight is key to better understanding the complex evolutionary dynamics of plasmids and their role in the emergence of drug resistance and virulence. Future work could explore where plasmids carrying AMR genes are emerging and disseminating that are driving plasmid-driven outbreaks in multiple species by incorporating other bacterial species into these analytical approaches.

Modeling plasmid evolution has profound implications for calibrating their molecular clock and inferring their evolutionary rates and timescales. The main factors to consider for molecular clock calibration are sequence sampling times and the amount of information that accumulates over time, where the latter pertains to the product of the evolutionary rate and the number of sites. Our results show that plasmid sequence data alone are insufficient to calibrate the molecular clock, such that joint

analyses of chromosome and plasmid data, as in CoalPT, are essential for understanding plasmid evolution.

Our current implementation does not model potential differences in fitness between lineages that carry plasmids and those that do not. This effect could, in principle, be modeled to study the fitness benefits and costs of plasmids on a population level by treating the fitness of a lineage as a function of the presence or absence of a plasmid (41). Such analyses would be particularly interesting in the context of empirically measured fitness costs in culture. An additional insight that could be gained is how plasmids are introduced and transferred between different host types, by extending the current unstructured coalescent approach to account for population structure (38, 42, 43).

Finally, we showed that modeling the co-divergence of plasmid and chromosomal DNA allows us to reconstruct the plasmid phylogeny much more precisely. In turn, these inferences improve the accuracy with which we can unravel key evolutionary pathways, such as the timing of their introduction to a population and the timescale of point mutations of epidemiological relevance. Importantly, the only source of evolutionary information that we consider is point mutations. Novel approaches that integrate, for example, structural rearrangements of plasmids, could provide additional insight into the evolutionary dynamics of plasmids moving within and between lineages. Such approaches would have applications to better understand the movement of drug-resistant plasmids both locally, within specific clinical settings, or internationally, such as tracking the dissemination of plasmids of interest across the globe.

Materials and Methods

Coalescent with plasmid transfer. Bacterial lineages can exchange plasmids through different mechanisms. To model this process, we use a coalescent-based model related to the coalescent with re-assortment (26). In the coalescent with plasmid transfer model, we model a backward-in-time process starting from sampled individuals (Figure 1). The sampled individuals are required to have a chromosome but can have anywhere from 0 to n plasmid sequences. For a given effective population size, Ne, and plasmid transfer rate ρ, we sample the time to the next coalescent event (from present to past) from an exponential distribution with a rate of $\frac{\binom{k}{2}}{Ne}$. The time until the next plasmid transfer event is drawn from an exponential distribution with mean $\frac{1}{k*\rho}$, with ρ being the plasmid-specific transfer rate and k being the number of lineages that have both the chromosome and the plasmid. Upon a coalescent event, the parental lineage will carry the union of chromosomal and plasmid lineages of the two child lineages. Upon a plasmid transfer event, one plasmid lineage is randomly chosen to branch off into one parental lineage, whereas all other plasmid and the chromosomal lineages will follow the other parental lineage. This is different from how re-assortment is modeled in (26) in that a plasmid transfer occurs relative to the chromosome, and only one plasmid is transferred at a time. In fact, it is the backward-in-time equivalent of one plasmid being transferred between bacterial lineages at a time. The method is agnostic to how a plasmid is transferred, other than the assumption that only one plasmid is transferred at a time. However, we assume that there is no interlineage recombination within the chromosomal or plasmid DNA, although this is an assumption that could potentially be relaxed in the future by employing a similar approach to (27). Importantly, the resulting phylogenetic network is not constrained to be tree-based (as e.g (44, 45)) but allowed to have any possible structure one can simulate under the coalescent with plasmid transfer.

Posterior probability. In order to perform joint Bayesian inference of phylogenetic networks, the embedding of chromosome and plasmid trees, together with the parameters of the associated models, we use an MCMC algorithm to characterize the joint posterior density. The posterior density is denoted as:

$$P(N, \mu, \theta, \rho|D) = \frac{P(D|N, \mu)P(N|\theta, \rho)P(\mu, \theta, \rho)}{P(D)}, \quad [1]$$

where N denotes the network, μ the parameters of the substitution model, θ the coalescent model and ρ the plasmid transfer rate. The coalescent model θ can be any model that describes an effective population size over time, meaning it can describe a constant rate coalescent process (constant Ne) or parametric or non-parametric Ne dynamics. The plasmid transfer rate is currently assumed to be constant over time but can vary between different plasmids. The multiple sequence alignment, that is, the data, is denoted D. $P(D|N, \mu)$ denotes the network likelihood, $P(N|\theta, \rho)$, the network prior and $P(\mu, \theta, \rho)$ the parameter priors. As is usually done in Bayesian phylogenetics, we assume that $P(\mu, \theta, \rho) = P(\mu)P(\theta)P(\rho)$.

Network likelihood. As we assume that there is no between-lineage recombination within the chromosomal or plasmid DNA, we can simplify the network likelihood $P(D|N, \mu)$ into the tree likelihood of the chromosomal and plasmid DNA. If T_i is the tree of the chromosome or plasmid (with $i = 0$ being the chromosome tree and $i > 0$ being plasmid trees) and if D_i is either the chromosomal or plasmid alignment, we can write the network likelihood as:

$$P(D|N, \mu) = \prod_{i=0}^{chromosome+nrplasmid} P(D_i|T_i, \mu), \quad [2]$$

The tree likelihood calculations use the default implementation of the tree likelihood in BEAST2 (24) and can use beagle (46) to increase the speed of likelihood calculations. Importantly, this approach allows us to all the default substitution and clock models in BEAST2, including, for example, relaxed clock models discussed here (24).

Network prior. The network prior is denoted by $P(N|\theta, \rho)$, which is the probability of observing a network and the embedding of chromosomal and plasmid trees under the coalescent with plasmid transfer model. θ denotes an unstructured coalescent population model with population size, Ne, and per plasmid transfer rate, ρ. The network prior is the equivalent to the tree prior in phylogenetic tree analyses.

We can calculate $P(N|\theta, \rho)$ by expressing it as the product of exponential waiting times between events (i.e., plasmid transfer, coalescent, and sampling events):

$$P(N|\theta, \rho) = \prod_{i=1}^{\#events} P(event_i|L_i, \theta, \rho) \times P(interval_i|L_i, \theta, \rho), \quad [3]$$

where we define t_i to be the time of the i-th event and L_i to be the set of lineages extant immediately prior to this event. (That is, $L_i = L_t$ for $t \in [t_i - 1, t_i)$.)

Given that the coalescent process is a constant size coalescent and given the i-th event is a coalescent event, the event contribution is denoted as:

$$P(event_i|L_i,\theta,\rho) = \frac{1}{Ne(t_i)}. \quad [4]$$

If the i-th event is a plasmid transfer event and assuming a constant rate over time, the event contribution is denoted as:

$$P(event_i|L_i,\theta,\rho) = \rho. \quad [5]$$

This event contribution can be generalized to account for different rates of transfer for different plasmids by substituting ρ with the plasmid-specific rate depending on which plasmid was transferred. The interval contribution denotes the probability of not observing any event in a given interval. It can be computed as the product of not observing any coalescence nor any plasmid transfer event in the interval i. We can, therefore, write:

$$P(interval_i|L_i,\theta,\rho) = exp[-(\lambda^c + \lambda^r)(t_i - t_{i-1})], \quad [6]$$

where λ^c denotes the rate of coalescence and can be expressed as:

$$\lambda^c = \binom{|L_i|}{2} \frac{(t_i - t_{i-1})}{\int_{t_{i-1}}^{t_i} Ne(t)}, \quad [7]$$

and λ^r denotes the rate of observing a plasmid transfer event on any co-existing lineage and can be expressed as:

$$\lambda^r = \rho \sum_{l \in L_i} \mathcal{L}(l) * \begin{cases} 0, & \text{if } n_i = 1 \\ n_i, & \text{otherwise} \end{cases} \quad [8]$$

with n_i being the number of plasmids on $\mathcal{L}_i$.

MCMC algorithm for plasmid transfer networks. In order to infer the network topology, timings of individual events as well as embedding of chromosome and plasmid trees within the plasmid transfer network, we employ Markov chain Monte Carlo sampling of the networks and embedding of trees. This MCMC sampling employs operators that operate on the network topology, embedding of trees within that network, or the timings of individual events, such as coalescent or plasmid transfer events. The operators we use are similar to the ones used in (26) and in (27), but the condition on only one plasmid jumping between bacterial lineages at a time. The MCMC operators are summarized in Text S1.

***Shigella* dataset.** *S. sonnei* (n = 789) and *S. flexneri* (n = 297) isolates (excluding *S. flexneri* serotype 6 isolates) received at the Microbiological Diagnostic Unit Public Health Laboratory (MDU PHL, the bacteriology reference laboratory for the state of Victoria, Australia), between January 2016 and December 2020 were included in this study. These isolates were accompanied by year and month of collection. These isolates undergo routine WGS on Illumina NextSeq platforms using DNA extraction and sequencing protocols previously described (18). Data were collected in accordance with the Victorian Public Health and Wellbeing Act 2008. Ethical approval was received from the University of Melbourne Human Research Ethics Committee (study number 1954615.3 and reference number 2024-30320-59894-3).

Reference genomes. The complete reference sequences for *S. sonnei* strain Ss046 and *S. flexneri* 2a strain 301 were downloaded from NCBI. For *S. sonnei* strain Ss046 these included the chromosome (accession number NC 007384), the virulence plasmid, pINV (accession number NC 007385 214,396 bases), spA (accession number NC 009345 8,401 bases), spB (accession number NC 009346 5,153 bases), and spC (accession number NC 009347 2,101 bases). For *S. flexneri* these included the chromosome (accession number NC 004337) and the virulence plasmids pINV (accession number NC 004851 221,618 bases). The complete reference for the pKSR100 plasmid pKSR100 strain SF7955 (accession number LN624486, 73,047 bases) was also downloaded. This MDR plasmid has been found in *S. sonnei* and *S. flexneri* lineages circulating in MSM populations since 2015 (12).

The complete plasmid genomes were characterized *in silico* for replicon family, relaxase type, mate-pair formation type, and predicted transferability of the plasmid using mob-suite (v3.0.2) (47). This confirmed the pINV and large MDR plasmid were all IncF plasmids, consistent with previous studies (12, 14, 22). The predicted mobility of two of the small *S. sonnei* plasmids was typed as mobilizable (spA and spB), while spC was typed as non-mobilizable. The presence of known AMR genes on the plasmids was confirmed using the ISO-accredited abritAMR with the species flag for *E. coli* (48). This confirmed *bla*TEM-1, *erm(B)* and *mph(A)* on the pKSR100 plasmid, and *strAB, tet(A)* and *sul2* on spA.

SNP alignments of the chromosome. Alignments of the core genome were generated for both the *S. sonnei* and *S. flexneri* isolates. While the virulence plasmid constitutes part of the core genome for *Shigella* species, it wasn't included in the core SNP analysis as the pINV in *S. sonnei* is frequently lost in culture. The 789 *S. sonnei* were aligned to the reference *S. sonnei* chromosome Ss046 (accession number NC 007384) to call SNPs using Snippy v4.4.5 (https://github.com/tseemann/snippy), with filtering of phage regions identified using PHASTER (49) and recombination detection undertaken with Gubbins (v2.4.1) (50). SNPsites (v2.5.1) (51) was used to extract the variant SNPs, resulting in a SNP alignment of 7,793. The genotype for the *S. sonnei* isolates was also determined using the Sonneityping scheme (52).

The same approach was used for the 297 *S. flexneri* isolates using the reference *S. flexneri* 2a str 301 (accession number NC 004337), resulting in a SNP alignment of 21,311 sites.

Generation of plasmid alignments for S. sonnei and S. flexneri. All 789 *S. sonnei* were also aligned to the four plasmids of Ss046 using using Snippy v4.4.5. To determine the optimal alignments for the different plasmids from the short-read data, three approaches were explored. First, the percentage coverage of the plasmid reference by the *Shigella* isolates was determined, with different alignments generated using snippy-core for isolates that had ≥60% coverage, ≥70% coverage, ≥80% coverage and ≥90% coverage of plasmids. Given that snippy-core has a strict core threshold, variation in the core threshold in the different plasmid alignments was explored using Core-SNP-filter (https://github.com/rrwick/Core-SNP-filter)(**?**). Alignments were generated using 0.80, 0.85, 0.90, 0.95, and 1.00 core thresholds from the full snippy alignments for the large plasmids (pINV and pKSR100). To determine the optimal balance of plasmid reference coverage and core SNP threshold for plasmid alignment, the Consistency Index implemented in the R package phangorn was used (53). The Consistency index is defined as the minimum number of changes / number of changes required on the tree. A Consistency Index of 1 is if there is no homoplasy. Maximum likelihood trees were inferred for each alignment using IqTree2 (v2.1.4-beta) (54) with a model of GTR+G and 1000 bootstraps. These were used as input along with the alignment to determine the Consistency Index. The final alignment for the *S. sonnei* pINV with 57 isolates was ≥60% coverage and a core SNP alignment with ≥95% conservation, and a

Consistency Index ≥0.90. The final alignment for the pKSR100 alignment with both *S. sonnei* and *S. flexneri* isolates (n = 560) was ≥80% coverage and a core SNP alignment with ≥95% conservation, and a Consistency Index ≥0.80.

For the three small *S. sonnei* plasmids, different core thresholds and plasmid reference coverage were considered. The plasmid reference coverage threshold for spA was ≥70% coverage, while a higher threshold of ≥90% coverage was used for spB and spC based off the binomial distribution of plasmid coverage. Given the small size of these plasmids and low number of SNPs, the full alignment with all variant sites were used, effectively a core threshold of 0.00. This included the gaps and N's in the sites in the alignments. A total of 555 isolates were included in the spA alignment, 94 isolates for the spB alignment, and 572 isolates for the spC alignment. The final alignments had a Consistency Index of ≥0.85 for spA and 1 for spB and spC

To determine the potential use of this approach for tracking AMR genes as opposed to plasmids, two additional alignments of the AMR genes carried on spA were also generated for 452 *S. sonnei*. Isolates were included in both alignments if *strA* & *B* and *sul2* were confidently detected (partial hits were not considered as these were unlikely to be functional). For the first alignment snippy-core was used for the spA plasmid with masking to include the region spanning from *sul2* to *tetA*. For the second additional alignment, snippy-core was used for the spA plasmid with masking to only include from *sul2* to the end of the plasmid reference (and additional ∼100 bases flanking *strB*). For both of these, the full alignment with all variant, invariant, and masked sites was used.

Draft genome assembly and characterization of AMR . All *S. sonnei* and *S. flexneri* isolates were assembled with Unicycler v0.5.0 (55). The draft genomes were screened for known AMR determinants using abritAMR (48) with the *E. coli* species flag. The genes detected are either 'exact matches' (100% identity and 100% sequence coverage compared to the reference protein sequence) or 'close matches' (90–100% identity and 90–100% sequence coverage compared to the reference protein sequence, marked by an asterisk [*] to distinguish from exact matches). The assembly graphs from Unicycler were initially explored in isolates that had ≥90% coverage of the small plasmids spB and spC and were found to be circularised in the short-read data. This was explored further using Circular-Contig-Extractor (https://github.com/rrwick/Circular-Contig-Extractor) that takes a GFA assembly graph as input and extracts complete circular contigs.

Long read sequencing, assembly, and exploration of AMR of selected representative isolates. A total of 31 *Shigella* isolates underwent long-read sequencing on the Oxford Nanopore Technologies (ONT) platforms. The selected isolates for ONT (23 *S. sonnei* and 8 *S. flexneri*) represented the different AMR profiles, genotype (for *S. sonnei*) and tree structure *S. flexneri* The isolates were cultured overnight at 37°C Luria-Bertani (LB) Miller agar. DNA was extracted using GenFind V3 according to the manufacturer's instructions (Beckman Coulter). The SQK-NBD112.96 kit was used for sequencing libraries. Isolates were sequenced on the R10 MinION flow cells with base calling by Guppy v3.2.4. The long read data were assembled using Unicycler v0.5.0 using the hybrid approach (55). The assembly graphs from Unicycler were investigated for the location of the AMR genes of interest using Bandage (56). BRICK was used to visualize the plasmids (https://github.com/esteinig/brick).

Validation and testing. Phylogenetic networks sampled under the coalescent with plasmid transfer should describe the same distribution as those simulated under the coalescent with plasmid transfer. As such, we compare the distributions of networks simulated under a set of parameters to those sampled using MCMC under the same set of parameters (in other words to sampled under the prior). If the implementation of the MCMC is correct, the two distributions of networks should match. As shown in Supplementary Figure S14, the sampled and simulated network distributions match.

We next performed a well-calibrated simulated study, where we simulated phylogenetic networks under effective population size and plasmid transfer rates sampled from the prior. We then infer the effective population sizes, plasmid transfer rates, and phylogenetic networks using, as priors, the same distributions used to sample the parameters for simulations. As shown in Supplementary Figures S15 and S16, we can retrieve the effective population sizes and plasmid transfer rates from simulated datasets.

Directionality of plasmid transfer. In order to estimate the directionality of plasmid transfers, we first classify each network lineage that carries the information of a chromosome into either *S. sonnei* and *S. flexneri*, based on the chromosome. Each reticulation event, which corresponds to a plasmid being introduced into a new bacterial lineage, is then classified based on the chromosome assignment, telling us into which species a plasmid has been introduced. For example, a plasmid being transferred onto a network lineage with the chromosome belonging to *S. sonnei* is classified as an introduction into *S. sonnei*.

We then infer that a plasmid has originated from *S. sonnei* or *S. flexneri* if the plasmid lineage has originated from a chromosomal lineage belonging to either species or from an unknown species entirely. To do so, we follow the plasmid lineage at each reticulation event backwards in time until we reach the next coalescent event of that plasmid lineage with another plasmid lineage. If this coalescent event has a corresponding chromosomal lineage, we say the plasmid originated from the species this lineage belongs to. As we do not explicitly consider plasmids other than *S. sonnei* or *S. flexneri*, we further assume that a plasmid has originated from an unknown species if this coalescent event is more than 50 years in the past.

Cluster size comparison. To get a measure of the relative fitness of lineages with and without pKSR100, we compare the number of offspring of lineages with and without the plasmid. To do so, we first map the presence and absence of pKSR100 onto the chromosomal tree. For each lineage, we then count the total number of leaves in the cluster below that node. For each node in the tree, we then compare the number of leaves below that node, to the number of leaves below any other co-existing lineage. Next, we log-standardize the cluster sizes across these co-existing lineages and then compare all nodes with and without pKSR100. We do this once for the entire tree and once only for *S. sonnei* and *S. flexneri* lineages. We repeat this for all the logged iterations in the posterior distribution to get the 95 % highest posterior density interval across the different iterations.

1. Murray CJ, et al. (2022) Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* 399(10325):629–655.
2. Organization WH (2024) WHO Bacterial Priority Pathogens List, 2024: bacterial pathogens of public health importance to guide research, development and strategies to prevent and control antimicrobial resistance, Technical report.
3. World Health Organization (2020) GLASS whole-genome sequencing for surveillance of antimicrobial resistance, Technical report.
4. Hawkey J, et al. (2022) ESBL plasmids in Klebsiella pneumoniae: diversity, transmission and contribution to infection burden in the hospital setting. *Genome Medicine* 14(1):97.
5. Ingle DJ, et al. (2021) Evolutionary dynamics of multidrug resistant Salmonella enterica serovar 4,[5],12:i:- in Australia. *Nature Communications* 12(1):4786.
6. Park SE, et al. (2018) The phylogeography and incidence of multi-drug resistant typhoid fever in sub-Saharan Africa. *Nature Communications* 9(1):5094.
7. Rozwandowicz M, et al. (2018) Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *Journal of Antimicrobial Chemotherapy* 73(5):1121–1137.
8. Partridge SR, Kwong SM, Firth N, Jensen SO (2018) Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clinical Microbiology Reviews* 31(4).
9. Douglas GM, Shapiro BJ (2021) Genic Selection Within Prokaryotic Pangenomes. *Genome Biology and Evolution* 13(1).
10. Ledda A, et al. (2022) Hospital outbreak of carbapenem-resistant Enterobacterales associated with a bla OXA-48 plasmid carried mostly by Escherichia coli ST399. *Microbial Genomics* 8(4):000675.

11. Roberts LW, et al. (2023) Long-read sequencing reveals genomic diversity and associated plasmid movement of carbapenemase-producing bacteria in a UK hospital over 6 years. *Microbial Genomics* 9(7).
12. Baker KS, et al. (2015) Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: a cross-sectional study. *The Lancet infectious diseases* 15(8):913–921.
13. Ingle DJ, et al. (2019) Co-circulation of Multidrug-resistant Shigella Among Men Who Have Sex With Men in Australia. *Clinical Infectious Diseases* 69(9):1535–1544.
14. The HC, Thanh DP, Holt KE, Thomson NR, Baker S (2016) The genomic signatures of Shigella evolution, adaptation and geographical spread. *Nature Reviews Microbiology* 14(4):235–250.
15. Bengtsson RJ, et al. (2022) Pathogenomic analyses of Shigella isolates inform factors limiting shigellosis prevention and control across LMICs. *Nature Microbiology* 7(2):251–261.
16. Holt KE, et al. (2012) Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nature Genetics* 44(9):1056–1059.
17. Connor T, et al. (2015) Species-wide whole genome sequencing reveals historical global spread and recent local persistence in Shigella flexneri. *eLife* 4(e07335).
18. Ingle DJ, et al. (2020) Prolonged Outbreak of Multidrug-Resistant Shigella sonnei Harboring blaCTX-M-27 in Victoria, Australia. *Antimicrobial Agents and Chemotherapy* 64(12):e01518–20.
19. Mason L, et al. (2023) The evolution and international spread of extensively drug resistant Shigella sonnei. *Nature Communications* 14.
20. Charles H, et al. (2022) Outbreak of sexually transmitted, extensively drug-resistant Shigella sonnei in the UK, 2021–22: a descriptive epidemiological study. *The Lancet Infectious Diseases* 22(10):1503–1510.
21. Mason LCE, et al. (2024) The re-emergence of sexually transmissible multidrug resistant Shigella flexneri 3a, England, United Kingdom. *npj Antimicrobials and Resistance* 2(1):20.
22. Locke RK, Greig DR, Jenkins C, Dallman TJ, Cowley LA (2021) Acquisition and loss of CTX-M plasmids in Shigella species associated with MSM transmission in the UK. *Microbial Genomics* 7(8):000644.
23. Lefèvre S, et al. (2023) Rapid emergence of extensively drug-resistant Shigella sonnei in France. *Nature Communications* 14(1):462.
24. Bouckaert R, et al. (2019) Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology* 15(4):e1006650.
25. Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical population biology* 23(2):183–201.
26. Müller NF, Stolz U, Dudas G, Stadler T, Vaughan TG (2020) Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. *Proceedings of the National Academy of Sciences* 117(29):17104–17111.
27. Müller NF, Kistler KE, Bedford T (2022) A bayesian approach to infer recombination patterns in coronaviruses. *Nature communications* 13(1):1–9.
28. Scornavacca C, Zickmann F, Huson DH (2011) Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics* 27(13):i248–i256.
29. Minh BQ, et al. (2020) Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution* 37(5):1530–1534.
30. Sagulenko P, Puller V, Neher RA (2018) Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution* 4(1):vex042.
31. Hasegawa M, Kishino H, Yano Ta (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution* 22(2):160–174.
32. Miles SL, et al. (2025) A public resource of 15 genomically characterised representative strains of Shigella sonnei. *bioRxiv*.
33. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG (2003) Measurably evolving populations. *Trends in ecology & evolution* 18(9):481–488.
34. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X (2015) Measurably evolving pathogens in the genomic era. *Trends in ecology & evolution* 30(6):306–313.
35. Duy PT, et al. (2020) Commensal Escherichiacoli are a reservoir for the transfer of XDR plasmids into epidemic fluoroquinolone-resistant Shigella sonnei. *Nature Microbiology* 5(2):256–264.
36. Arredondo-Alonso S, et al. (2024) Plasmid-driven strategies for clone success in Escherichia coli. *bioRxiv* p. 2023.10.14.562336.
37. Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. *PLoS Comput Biol* 5(9):e1000520.
38. Stolz U, Stadler T, Müller NF, Vaughan TG (2022) Joint inference of migration and reassortment patterns for viruses with segmented genomes. *Molecular biology and evolution* 39(1):msab342.
39. Givnish T, Sytsma K (1997) Consistency, characters, and the likelihood of correct phylogenetic inference. *Molecular Phylogenetics and Evolution* 7(3):320–330.
40. Lam MMC, et al. (2019) Convergence of virulence and MDR in a single plasmid vector in MDR Klebsiella pneumoniae ST15. *Journal of Antimicrobial Chemotherapy* 74(5):1218–1222.
41. Łuksza M, Lässig M (2014) A predictive fitness model for influenza. *Nature* 507(7490):57–61.
42. Müller NF, Rasmussen DA, Stadler T (2017) The structured coalescent and its approximations. *Molecular biology and evolution* 34(11):2970–2981.
43. Müller NF, Rasmussen D, Stadler T (2018) Mascot: Parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics* 34(22):3843–3848.
44. Didelot X, Lawson D, Darling A, Falush D (2010) Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* 186(4):1435–1449.
45. Vaughan TG, et al. (2017) Inferring ancestral recombination graphs from bacterial genomic data. *Genetics* 205(2):857–870.
46. Ayres DL, et al. (2012) Beagle: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic biology* 61(1):170–173.
47. Robertson J, Nash JHE (2018) MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics* 4(8):e000206.
48. Sherry NL, et al. (2023) An ISO-certified genomics workflow for identification and surveillance of antimicrobial resistance. *Nature Communications* 14(1):60.
49. Arndt D, et al. (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research* 44(W1):W16–W21.
50. Croucher NJ, et al. (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* 43(3):e15–e15.
51. Page AJ, et al. (2016) SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics* 2(4):e000056.
52. Hawkey J, et al. (2021) Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, Shigella sonnei. *Nature Communications* 12(1):2684.
53. Schliep KP (2011) phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4):592–593.
54. Minh BQ, et al. (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* 37(5):1530–1534.
55. Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* 13(6).
56. Wick RR, Schultz MB, Zobel J, Holt KE (2015) Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 31(20):3350–3352.
57. Bouckaert RR (2010) Densitree: making sense of sets of phylogenetic trees. *Bioinformatics* 26(10):1372–1373.
58. Schliep KP (2011) phangorn: phylogenetic analysis in r. *Bioinformatics* 27(4):592–593.
59. Bordewich M, Linz S, Semple C (2017) Lost in space? generalising subtree prune and regraft to spaces of phylogenetic networks. *Journal of theoretical biology* 423:1–12.

Acknowledgments. N.F.M. is supported in part by NIH NIGMS R35 GM119774. S.D. is supported by the Inception program (Investissement d'Avenir grant ANR-16-CONV-0005). R.R.W is supported by the Australian Research Council (ARC) through a Discovery Early Career Researcher Award (DECRA) [DE250100677]. T.B. is an Investigator of the Howard Hughes Medical Institute. B.P.H. is supported by an NHMRC Investigator Grant (GNT1196103). D.J.I. is supported by an NHMRC Investigator Grant (GNT1195210). This work was supported by a National Health and Medical Research Council (NHMRC) Australia partnership grant (GNT1149991). The Microbiological Diagnostic Unit Public Health Laboratory is funded by the State

Government of Victoria, Australia.

Data availability. The source code for the analyses performed, such as the R scripts to recreated Figures is available here https://github.com/nicfel/Plasmids-Material. Details of the isolates included in this study are available in Supplementary Tables 1 and 2.

Code availability. The coalescent with plasmid transfer is implemented as a package to BEAST2 called CoalPT. The source code for this package is available here https://github.com/nicfel/CoalPT. The source code for the analyses performed, such as the R scripts to recreated Figures is available here https://github.com/nicfel/Plasmids-Material. The networks are plotted using an adapted version of baltic https://github.com/evogytis/baltic/. The densitree plot (57) uses an adapted version of the one implemented as part of the phangorn package (58)

Supporting Information (SI). Supplementary Appendix 1 Supplementary Appendix 2

SI Datasets. Supplementary Tables 1 and 2

Supplementary Appendix 1: Operator Descriptions

Add/remove operator. The add/remove operator adds and removes plasmid transfer events. The add-remove operator on networks is an extension of the subtree prune and regraft move for networks (59). Similar to Müller 2022 (27), we also added an adapted version to sample re-attachment under a coalescent distribution to increase acceptance probabilities.

Exchange operator. The exchange operator changes the attachment of edges in the network while keeping the network length constant.

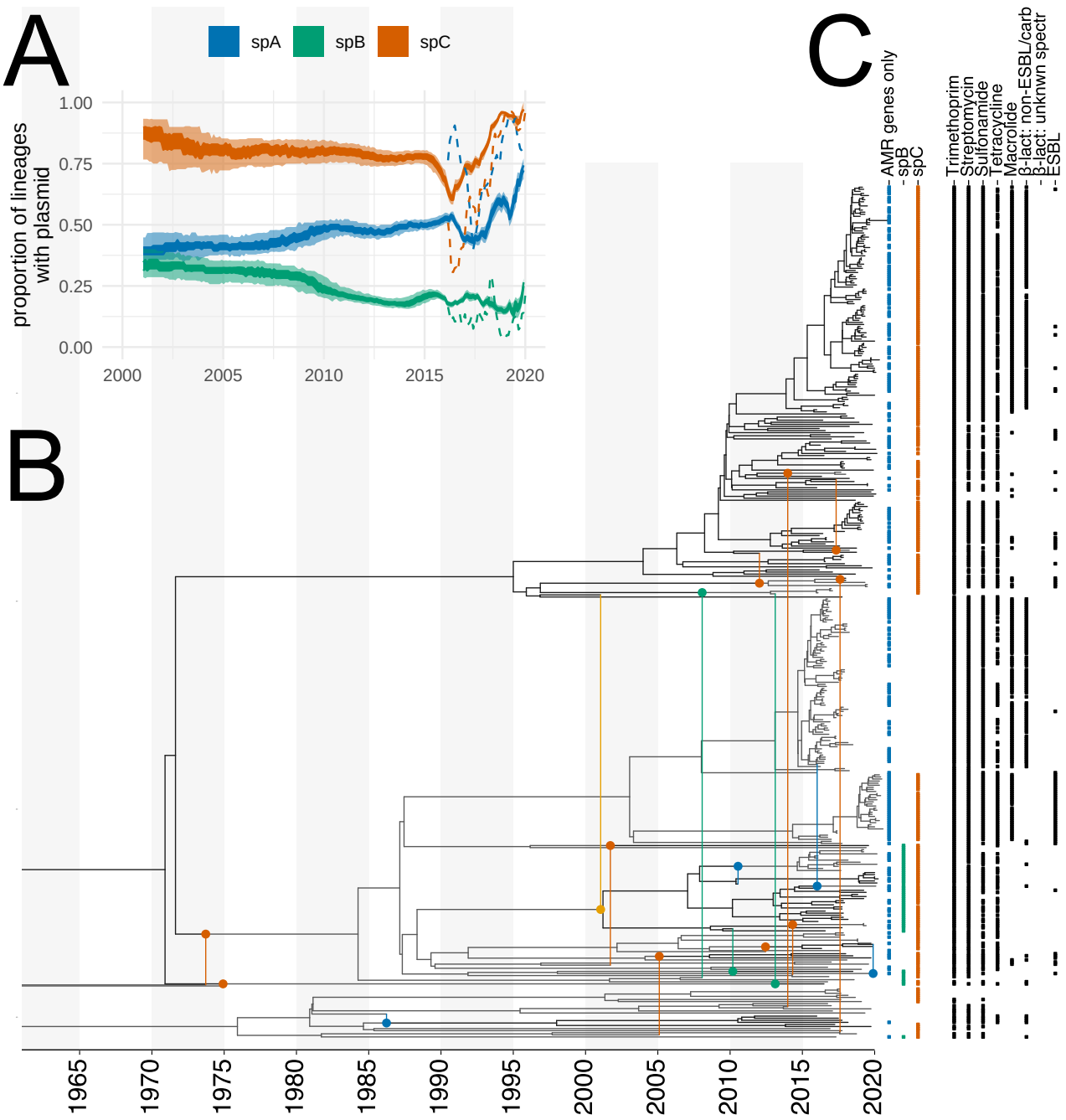
Subnetwork slide operator. The subnetwork slide operator changes the height of nodes in the network while allowing the change in the topology.

Scale operator. The scale operator scales the heights of the root node or the whole network without changing the network topology.

Gibbs operator. The Gibbs operator efficiently samples any part of the network that is older than the root of any segment of the alignment and is thus not informed by any genetic data and is the analog to the Gibbs operator in (26) for re-assortment networks.

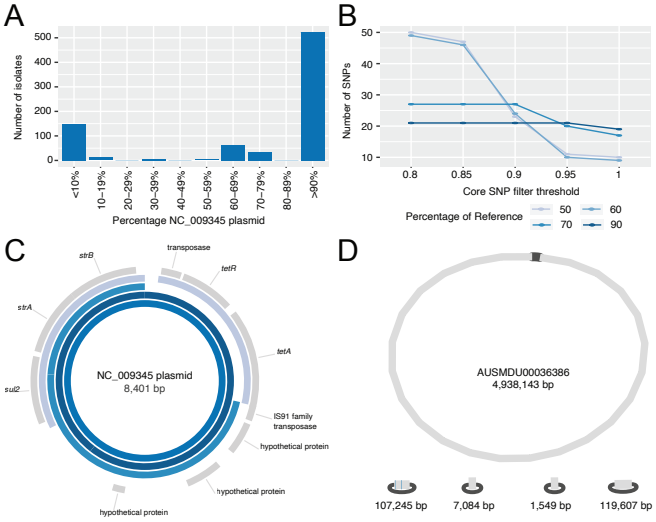
Empty edge preoperator. The empty edge preoperator augments the network with edges that do not carry any loci for the duration of a move, to allow for larger jumps in network space.

The roots of phylogenetic networks can be much more distant than the roots of the individual plasmid trees. As in (27), we assume the plasmid transfer rate to be reduced prior to the individual plasmid trees having reached their root. As shown in (27), this assumption does not affect parameter inference.

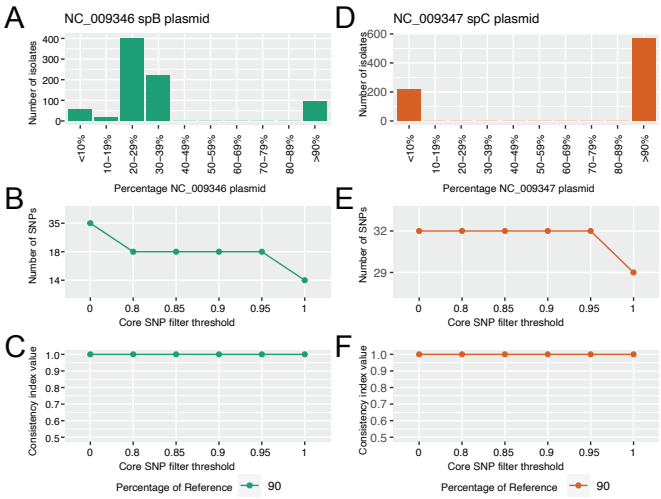


Supplementary Figure S1. Co-divergence of the core chromosome and plasmids in *Shigella sonnei.* when only using the AMR genes instead of the entire spA plasmid A Proportion of lineages carrying a plasmid between 2000 and 2020. The inner shaded areas denote the 50% HPD, and the outer area is the 95% HPD. The dotted lines denote the proportion of samples with a plasmid. **B** Here, we show the maximum clade credibility (MCC) network of *Shigella sonnei* inferred using the chromosomal DNA, the virulence plasmid pINV and the small plasmids spA, spB, and spC. Vertical lines are used to denote plasmid transfer events, where the circles denote the branch to which a plasmid was transferred. The color of the circle denotes either spA, spB, or spC, having jumped between bacterial lineages. The dashed lines correspond to branches from which plasmids branch off. The text denotes the posterior probability of plasmid transfer events for events with a posterior support of over 0.5. Branches are colored to denote separate clades originating from individual jumps of plasmids into a bacterial clade. **C** The tip labels in blue, green, and red denote if a plasmid was detected at a leave. The black dots denote the presence of antimicrobial resistance to the antimicrobials on top.

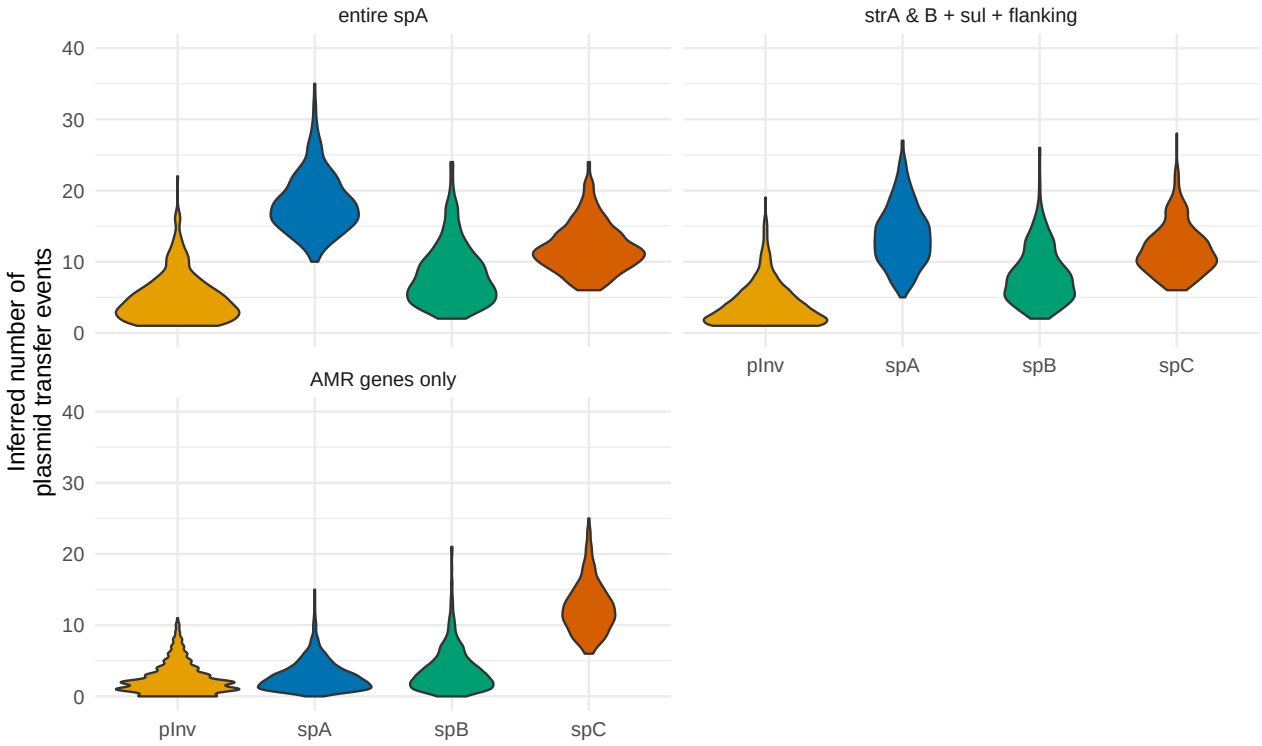
483 ## Supplementary Appendix 2: Supplementary Figures



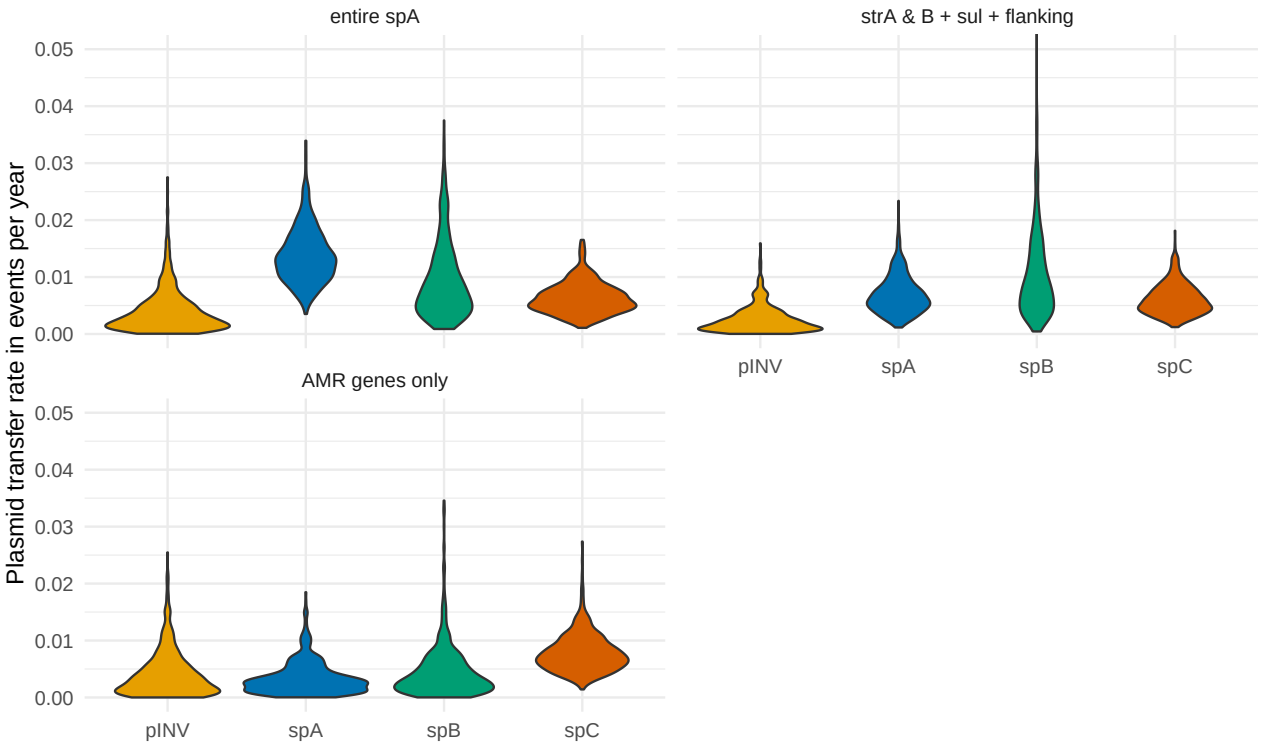
Supplementary Figure S2. Overview of spA detection in the *S. sonnei* data. Here, we show the approach for the spA alignment. Panel A) shows the percentage cover of the reference plasmid. Panel B) shows the number of SNPs detected at four different percentages of reference thresholds and different core SNP filter thresholds. Panel C) shows the visualization of contigs from ONT assemblies for representative isolates AUSMDU00029307 ($\geq$90), AUSMDU00020566 ($\geq$70) and AUSMDU00036386 ($\geq$50) at different percentage cover of the reference plasmid. Panel D) the complete genome assembly of AUSMDU00036386 which shows a single chromosome and four plasmids. The blue lines in the plasmid of 107,245 bases show the blast hits for the two AMR regions in spA.



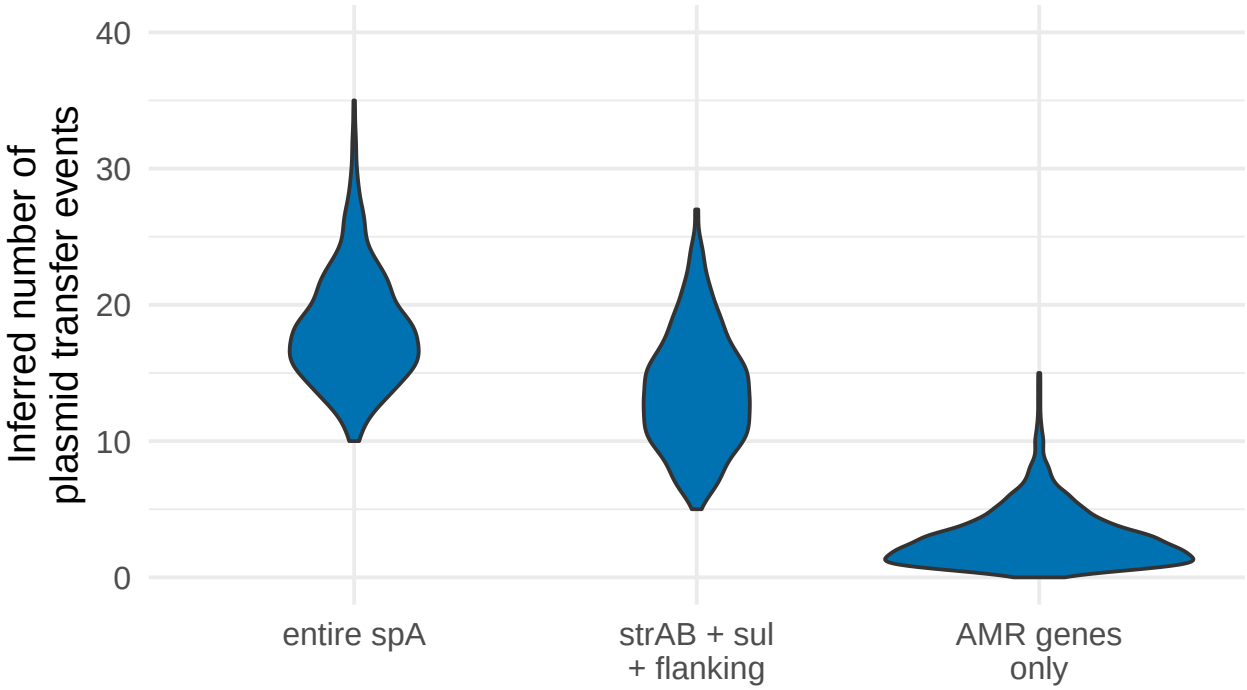
Supplementary Figure S3. Overview of spBC detection in the *S. sonnei* data. Here, we show the approach for the spB and spC alignment. Panels A) and D) show the percentage cover of the reference plasmid for spB and spC, respectively. Panels B) and E) show the number of SNPs detected in each dataset for isolates with $\geq$90 of the plasmid reference with different core SNP filter thresholds. Panels C) and F) show the consistency index (CI) at different core SNP thresholds.



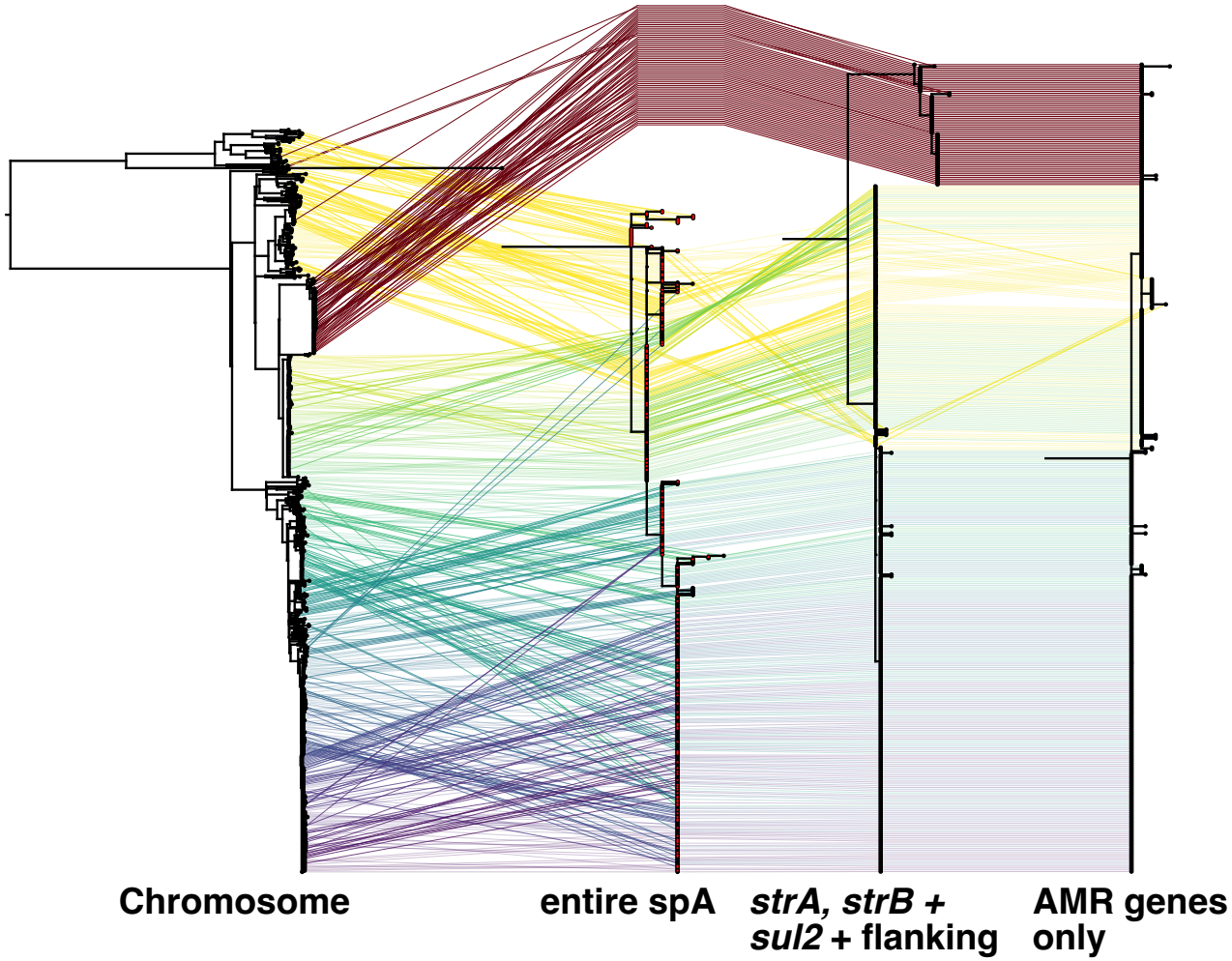
Supplementary Figure S4. Inferred number of times at which plasmids jumped between bacterial lineages. Here, we show the posterior distribution of how often a pINV (the virulence plasmid), spA, spB, and spC on the x-axis moved between *S. sonnei* lineages on the y-axis.



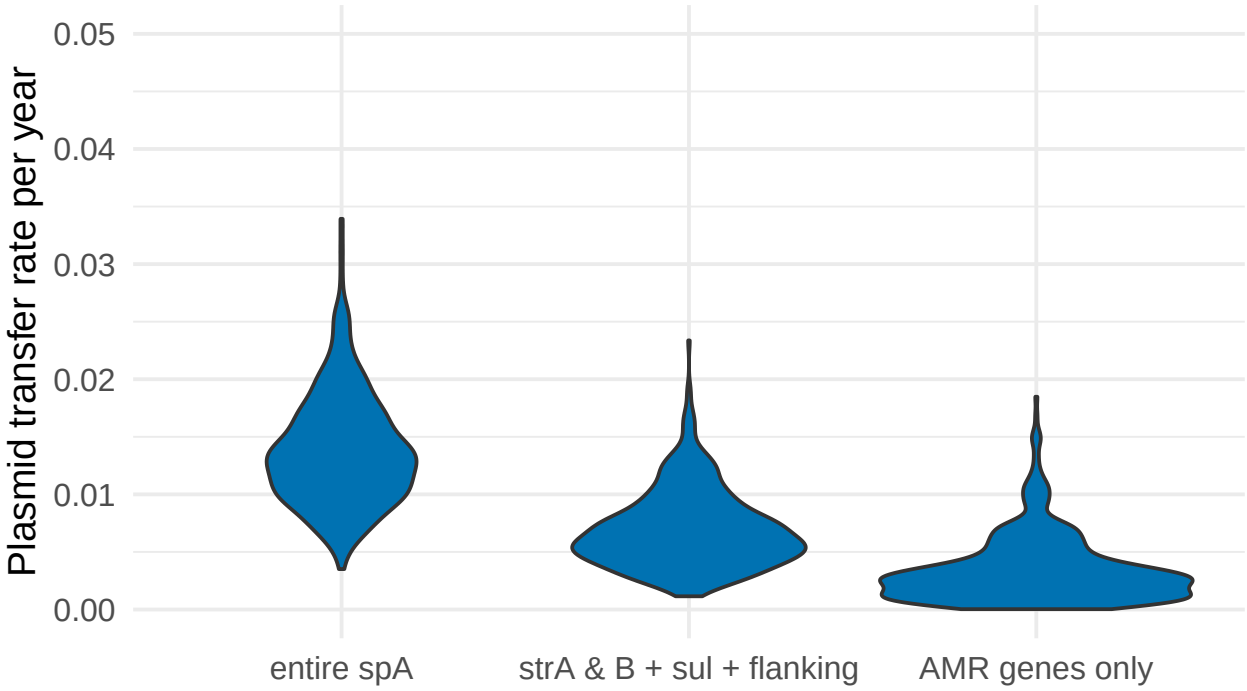
Supplementary Figure S5. Comparison of the plasmid transfer rate between spA, spB and spC for three separate analyses that use different parts of the spA plasmid. Here, we compare the posterior distribution of the plasmid transfer rate when using different parts of the spA plasmid for inference. Each violin plot is created from a different analysis using either the entire spA plasmid, the combination of four AMR genes from *sul2* to *tetA*, the three AMR genes *sul2, strA,strB* and flanking region of ∼100 bases.



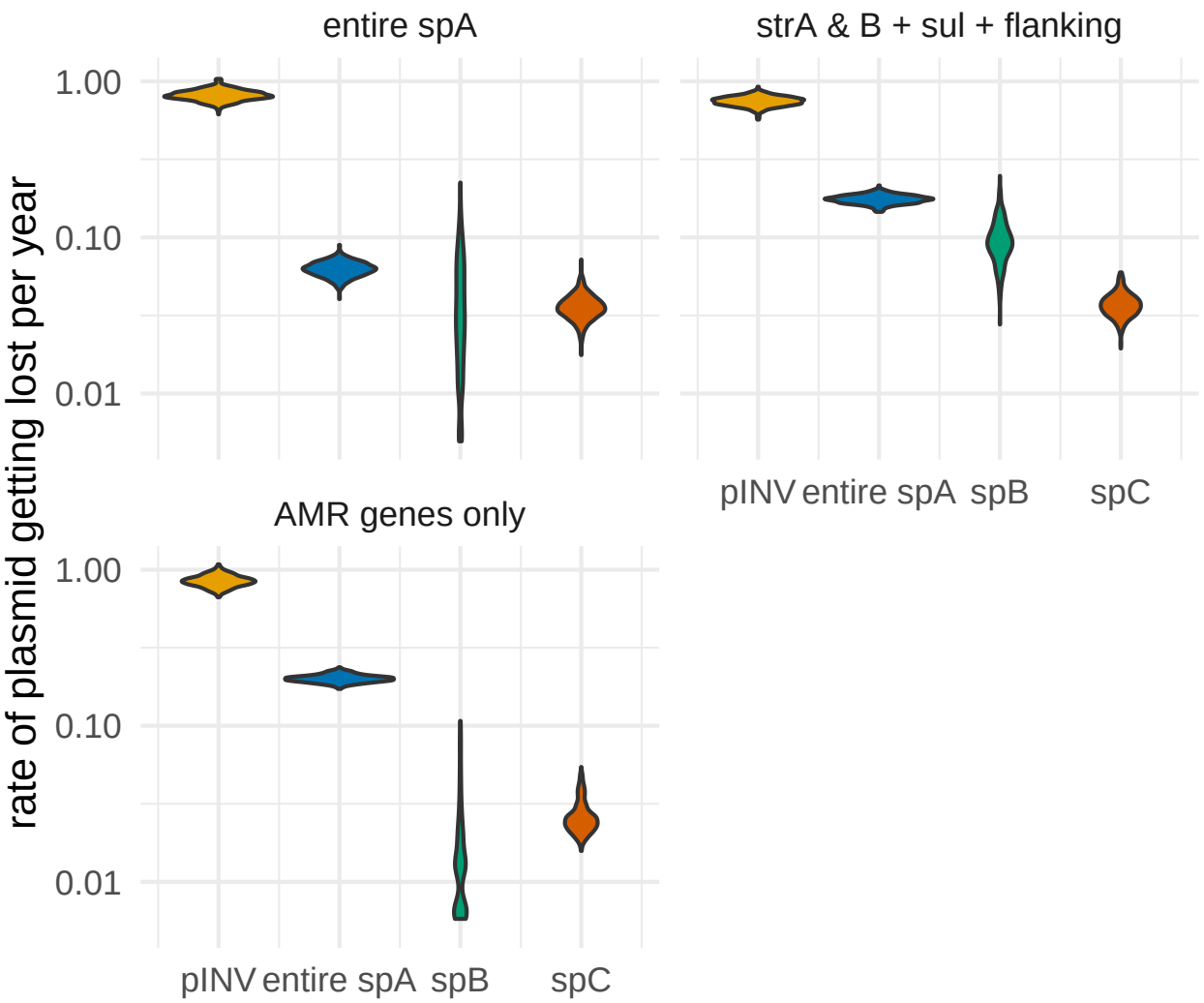
Supplementary Figure S6. Comparison of the posterior distribution of plasmid transfer events when using different parts of the spA plasmid. Here, we compare the posterior distribution of the number of plasmid transfer events when using different parts of the spA plasmid for inference. Each violin plot is created from a different analysis using either the entire spA plasmid, the combination of four AMR genes from *sul2* to *tetA*, the three AMR genes *sul2, strA,strB* and flanking region of ∼100 bases.



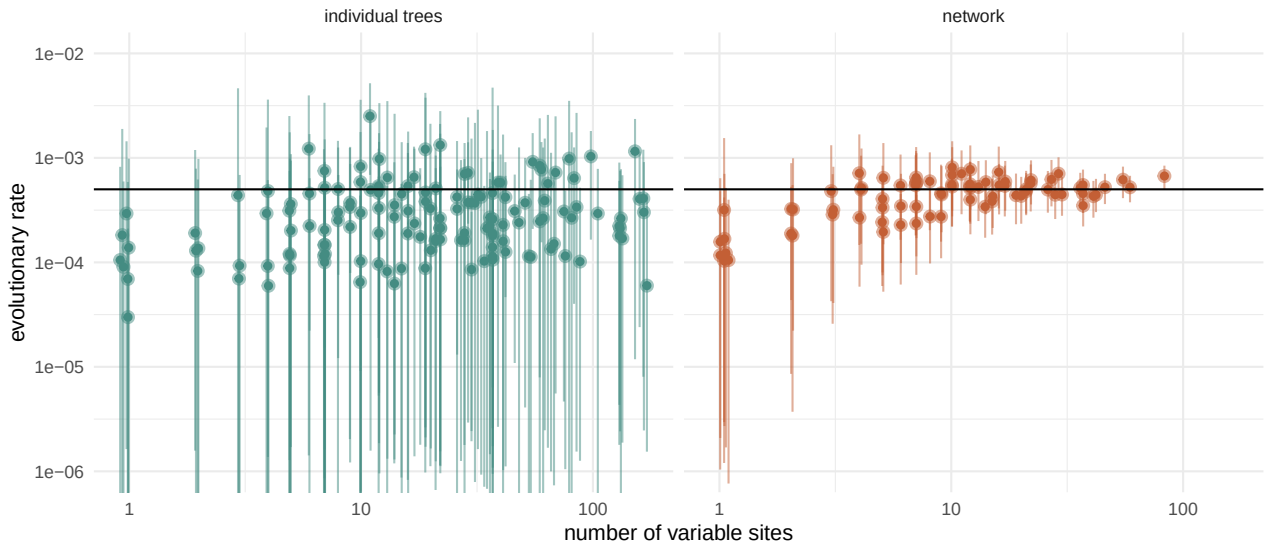
Supplementary Figure S7. Tanglegrams of the chromosome, and reconstructed trees when using different parts of the spA plasmid. Here, we show a tangle-gram of four trees. The leftmost tree represents the chromosomal data, and the next tree the corresponding spA sequences with more than 70% coverage of the reference genome. When using the 70% coverage threshold, the clade denoted by the red line is removed from the dataset.



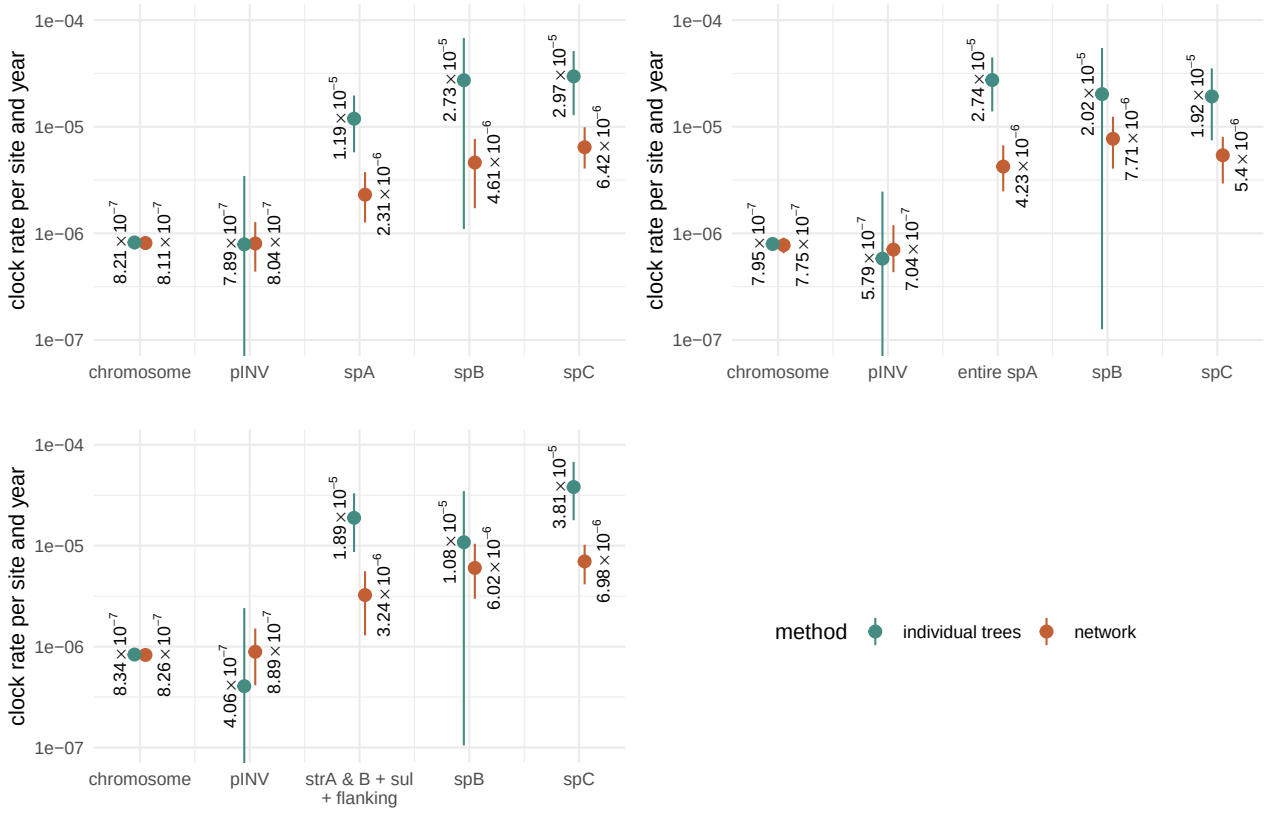
Supplementary Figure S8. Comparison of the plasmid transfer rate when using different parts of the spA plasmid. Here, we compare the posterior distribution of the plasmid transfer rate focusing only on spA and comparing between using different parts of the spA plasmid for inference. Each violin plot is created from a different analysis using either the entire spA plasmid, the combination of four AMR genes from *sul2* to *tetA*, the three AMR genes *sul2, strA,strB* and flanking region of ∼100 bases.



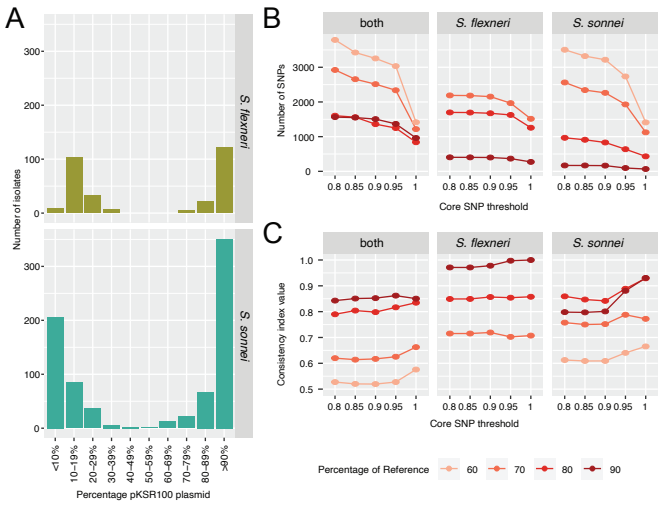
Supplementary Figure S9. Rate at which the different plasmids are being lost by *S. sonnei* lineages. We compute the rate at which plasmids are being lost as the number of events where a plasmid was lost divided by the tree length of that plasmid. We assume a plasmid was lost on edges where the parent node carried a plasmid while the child node did not.



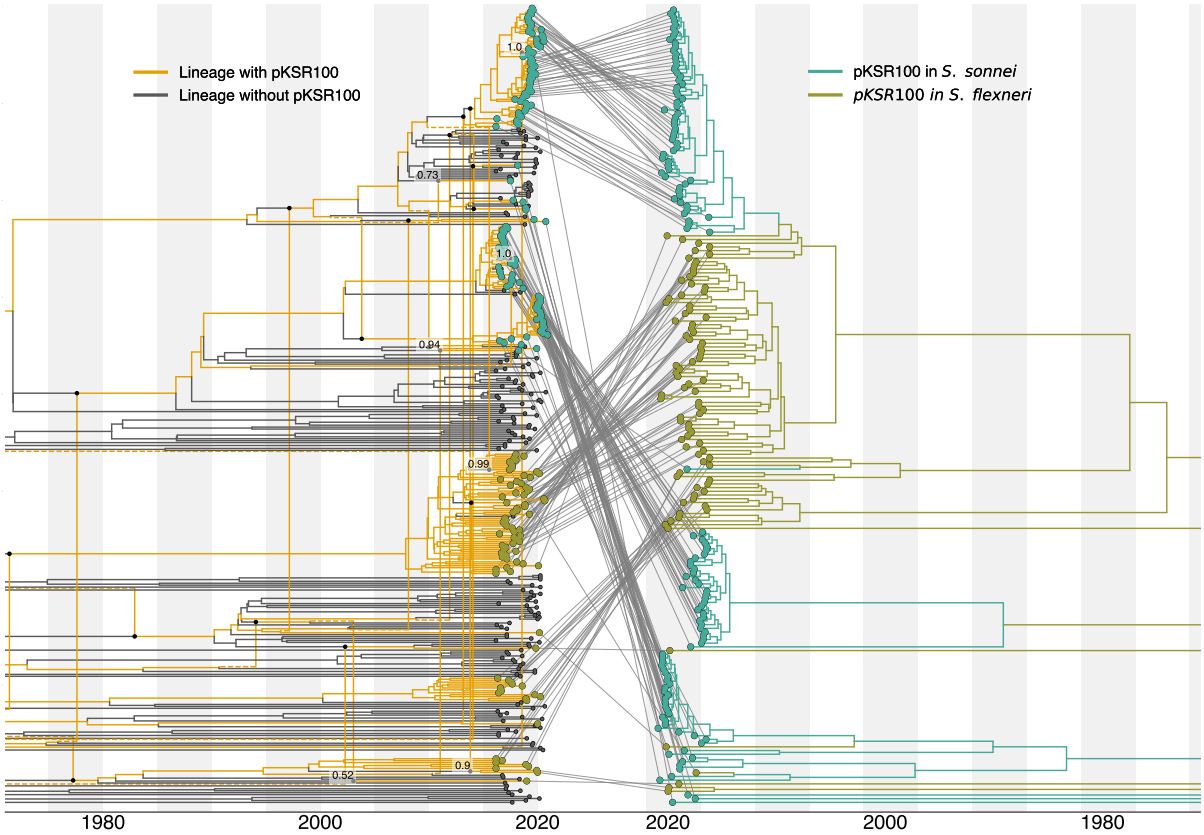
Supplementary Figure S10. Rates of evolution estimated for plasmid when modeling and when not modeling the joint history with the chromosomal DNA We simulated 50 phylogenetic networks under the coalescent with plasmid transfer with three plasmids sampled over five years. We assume that the chromosome and the three plasmids evolved at a rate of 5×10^{-4} subs/site/unit time. The chromosome has an SNP alignment length of 8000bp, while the three plasmids had SNP alignments of 200bp, 100bp, and 50bp, respectively. These settings will produce approximately the same number of SNPs per unit of time as a chromosome of 4.8 Mbp evolving at a rate of 8×10^{-7} subs/site unit time. On the y-axis, we show the inferred evolutionary rates with the error bars denoting the 95% HPD and the point denoting the mean estimates. The x-axis is the number of variable sites in the alignment.



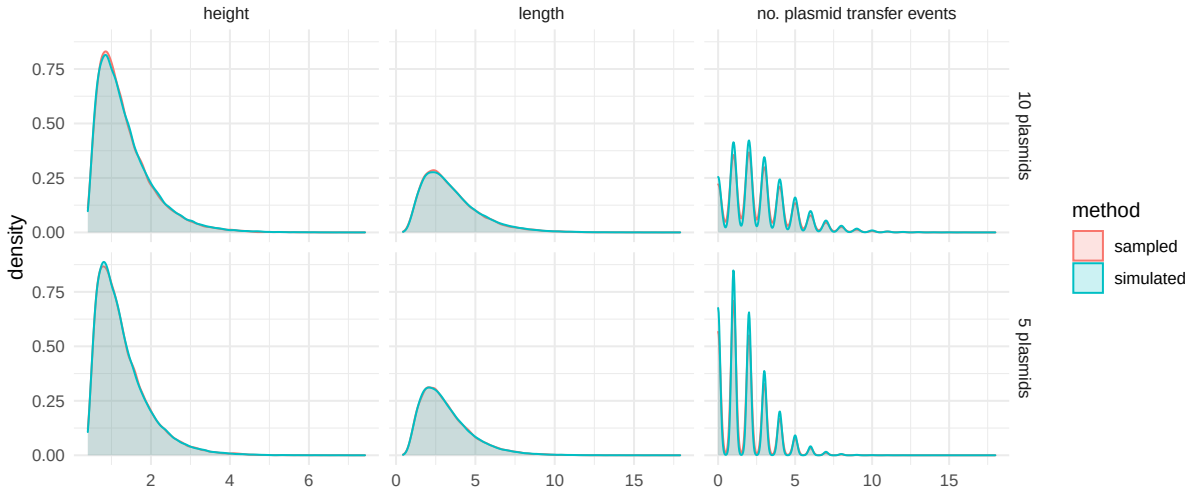
Supplementary Figure S11. Comparison of the clock rates of chromosomal DNA and plasmid DNA. Here, we compare the posterior distribution of the number of plasmid transfer events when using different parts of the spA plasmid for inference. Each violin plot is created from a different analysis using either the entire spA plasmid, the combination of four AMR genes from *sul2* to *tetA*, the three AMR genes *sul2, strA,strB* and flanking region of ∼100 bases.



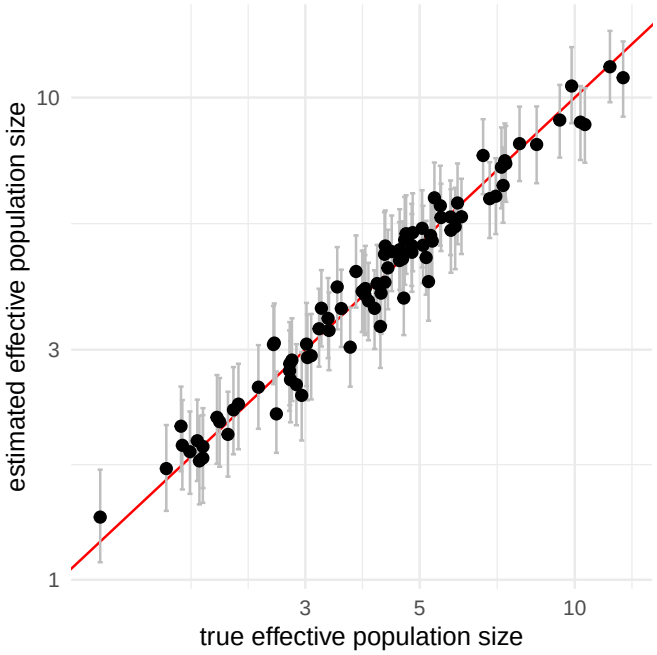
Supplementary Figure S12. Overview of detection of MDR plasmid pKSR100 in the *S. sonnei* and *S. flexneri* datasets. Here, we show the approach for the MDR plasmid pKSR100 alignments. Panels A) shows the percentage cover of the reference pKSR100 plasmid for *S. sonnei* and *S. flexneri*. Panel B) shows the number of SNPs detected in each dataset for isolates with four different thresholds for the coverage of the plasmid reference with different core SNP filter thresholds. Panels C) show the consistency index (CI) of four different thresholds for the coverage of the plasmid reference with different core SNP filter thresholds for three datasets.



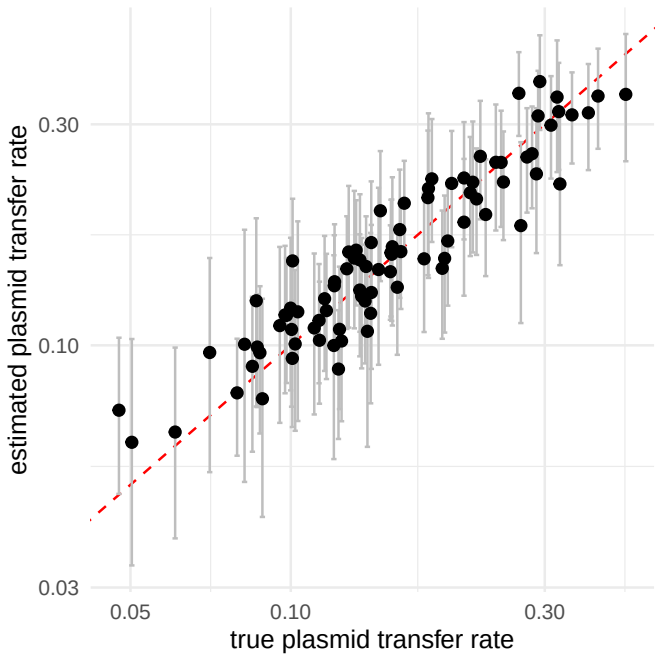
Supplementary Figure S13. Transmission of pKSR100 between *S. sonnei* and *flexneri*. MCC network of *S. sonnei* and *flexneri* samples with the embedding of the pKSR100 plasmid tree **A**. The text denotes the posterior support values for plasmid transfer events. **B** Plasmid tree of pKSR100 with the host species *S. sonnei* or *S. flexneri* mapped onto the tree. The different colors of the tips show clusters of sequences that are the result of separate introductions of the MDR plasmid. MCC: maximum clade credibility. MDR: multidrug resistance



Supplementary Figure S14. Comparison of network height, length, and plasmid transfer events between sampled and simulated networks. To validate the implementation of CoalPT, we simulated networks under the CoalPT model, once with 5 plasmids and once with 10 plasmids. We then sampled phylogenetic networks under our implementation of the CoalPT model in BEAST2 under the prior (i.e., without any sequence information). As shown here, the summary statistics between networks simulated and sampled (using MCMC) under CoalPT match.



Supplementary Figure S15. Inferred effective population sizes from simulated data. To test the performance of the coalescent with plasmid transfer, we simulated 100 networks in a well-calibrated simulated study. The effective population sizes were sampled from a Lognormal distribution with M=1.4844 and S=0.5. The plasmid transfer rates were sampled from a Lognormal distribution with M=-1.7344 and S=0.5. We then simulated genomic sequences for the core genome and 3 plasmids under the Jukes-Cantor Model. Last, we inferred the phylogenetic network, effective population sizes and plasmid transfer rates from these sequences using the above lognormal distributions as priors on the Ne and plasmid transfer rates. Here, we show the inferred Ne sizes (y-axis) compared to simulated Ne (x-axis). The point denote the median estimate and the error bars the lower 95% highest posterior density interval.



Supplementary Figure S16. Inferred plasmid transfer rates from simulated data. Here, we show the inferred plasmid transferred rates(y-axis) compared to the true/simulated rates on the x-axis. These estimates are from the same analyses as the ones in fig S15. The point denote the median estimate and the error bars the lower 95% highest posterior density interval.