Tracking the horizontal transfer of plasmids in *Shigella sonnei* and *Shigella flexneri* using phylogenetics

Nicola F. Müller[1,*], Sebastián Duchêne[2], Deborah A. Williamson[3], Trevor Bedford[1,4], Benjamin P. Howden[2] and Danielle J. Ingle[2]

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA
[2]Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, VIC, AUS
[3]Department of Infectious Diseases at the Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, VIC, AUS
[4]Howard Hughes Medical Institute, Seattle, WA, USA

[*]Corresponding author

Contact: nicola.felix.mueller@gmail.com

Abstract: Antimicrobial resistance (AMR) poses an ever-increasing challenge to the treatment of infections. AMR mechanisms are commonly associated with AMR genes that are carried on mobile elements, such as plasmids that can move between bacterial lineages. Here we introduce an approach that allows us to reconstruct how plasmids move between bacterial lineages. To do so, we model the co-evolution of chromosomal and plasmid DNA in a Bayesian phylogenetic network approach using a joint coalescent and plasmid transfer process. We apply this new approach to a five-year dataset of *Shigella* isolates from Melbourne, Australia. Doing so, we reconstruct the gain and loss of small plasmids, and the recent dissemination of a multidrug-resistance plasmid between *S. sonnei* and *S. flexneri* lineages in multiple independent events and through steady growth in the prevalence since 2010. This approach has a strong potential to improve our understanding of where AMR-carrying plasmids are introduced and maintained.

Keywords: Phylodynamics, Antimicrobial resistance, phylogenetic network, Bayesian phylogenetics, Bacterial phylogenetics, BEAST

Introduction

Antimicrobial resistance (AMR) in bacteria represents one of the most serious public health threats of the 21st century, with the burden of disease estimated to be over 1 million deaths per year (Murray *et al.*, 2022). Multiple reports have highlighted the urgent need for novel computational approaches to track the emergence and spread of AMR in both known and emerging bacterial pathogens (World Health Organization, 2020). AMR genes that mediate either reduced susceptibility or resistance to therapeutics are often carried on mobile elements, such as plasmids, that typically form part of the accessory genome (Rozwandowicz *et al.*, 2018; Partridge *et al.*, 2018). Plasmids can move between lineages of the same bacterial species, or between unrelated bacterial species (Partridge *et al.*, 2018). Importantly, horizontal transfer of genetic material enables bacterial populations to rapidly develop AMR, as plasmids may carry multiple genes that confer AMR to different antimicrobials (Hawkey *et al.*, 2022; Ingle *et al.*, 2021; Park *et al.*, 2018). The spread of drug-resistant plasmids within a bacterial population increases the chance of AMR genes disseminating to other bacterial species in the same ecological niche.

Shigella are a key exemplar pathogen to develop new methodological approaches to study the movement of plasmids between bacterial lineages. The core genome of *Shigella* is comprised of the chromosome and a large virulence plasmid (pINV) that is essential for infection in humans Yang *et al.* (2005). This virulence plasmid, pINV, co-evolves with the *Shigella* chromosome, with variation in size and genetic content between the

Shigella species (The *et al.*, 2016). The pINV encodes genes that facilitate several interactions with the host cell machinery enabling the bacterium to survive and replicate in the human host cells (Schroeder and Hilbi, 2008). Two species of *Shigella*, *S. sonnei* and *S. flexneri*, are responsible for the main burden of disease of Shigellosis globally (Bengtsson *et al.*, 2022). These species may have different plasmids. For example, in *S. sonnei* three other smaller plasmids have been characterized in the reference genome Ss046. These three plasmids, spA, spB and spC, are commonly found within *S. sonnei* global lineage III (Holt *et al.*, 2012). AMR determinants to streptomycin and sulfonamide are encoded on spA (Hawkey *et al.*, 2021). Additionally, *S. sonnei* and *S. flexneri* have been associated with multidrug resistant (MDR) outbreaks, particularly in men who have sex with men (MSM) (Baker *et al.*, 2015; Ingle *et al.*, 2019, 2020; Mason *et al.*, 2022), representing a major public health threat. These outbreaks have been driven by the presence of an MDR plasmid, pKSR100 (Baker *et al.*, 2015; Ingle *et al.*, 2019). Recent reports have shown the acquisition of *bla*CTX-M-27 gene mediating resistance to extended spectrum betalactams on variants of the pKSR100 plasmid driving the outbreaks of extensively drug resistant *S. sonnei* (Mason *et al.*, 2022).

The movement of plasmids is a major barrier to understanding and controlling AMR in bacterial species and new approaches are urgently required to understand these movements. The advent of whole-genome sequencing (WGS) and ongoing implementation of routine WGS of bacterial pathogens into public health laboratories means that AMR is increasingly detected using *in silico* approaches. However, to date there have been limited approaches to explore plasmid evolution in large scale population analyses. This shortfall is in part due to the genetic makeup of plasmids that complicates computational analyses (Robertson and Nash, 2018). Plasmids are usually typed using a limited number of markers based on replicon or MOB genes, due to a lack of core backbone (Partridge *et al.*, 2018; Robertson and Nash, 2018). Plasmid types may co-occur with different AMR profiles however, complete plasmid genomes are still required to confidently characterise AMR mechanisms.

WGS provides an avenue to tackle the question of how plasmids move between bacterial lineages by allowing us to infer the shared ancestral history of genomes isolated from different bacteria. The horizontal transfer of plasmids occurs between bacterial lineages and, as such, is a co-divergent process with the chromosomal DNA of these bacterial lineages. We model this process by using a "coalescent with plasmid transfer (CoalPT)" model, instead of overlaying plasmid presence or absence, as is typically done. The model, which we refer to as CoalPT, can be described as a joint coalescent and plasmid transfer process, where lineages can coalesce from present to past or undergo a plasmid transfer event, similar to how recombination is often modeled (Hudson, 1983). The model has two key parameters, the effective population size and the plasmid transfer rate, that denote the rate at which coalescent and plasmid transfer events occur. The estimated 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. 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 chromsome or plasmid trees. As such, the plasmid network denotes the co-evolutionary history of the chromosome and the plasmid is denoted by a timed phylogenetic network, in which the chromosome and plasmid trees are embedded. To perform inference under the CoalPT model, we use Markov chain Monte Carlo (MCMC) sampling to infer the timed phylogenetic network, related to the MCMC inference of reassortment (Müller *et al.*, 2020) and recombination networks (Müller *et al.*, 2022). 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 data and the network and parameter estimates.

We implemented this approach as a package for the open source software BEAST2 (Bouckaert *et al.*, 2019) to facilitate its adoption. We then use CoalPT to reconstruct the acquisition, movement and co-divergence of several plasmids within two species of *Shigella*, of different sizes, virulence and AMR potential. To do so, we use a dataset of overall 1,105 *Shigella* isolates from a five year timespan from Australia, representing one of the most comprehensive datasets of *Shigella* globally. We first infer of the rate of plasmid transfer in *S. sonnei*, showing that there were multiple events where plasmids were transferred between bacterial lineages, but also that these plasmids get lost repeatedly. We then show how modeling the co-divergence of plasmid and chromosomal DNA

enables inference of the rate of evolution of plasmids with high precision despite limited genomic information. Lastly, we show the movement of an MDR plasmid within and between two *Shigella* species. Understanding how plasmids move within and between bacterial lineages can provide insights into dissemination of plasmid-mediated AMR to inform targeted public health interventions.

Results

Smaller, non-essential plasmids are readily transferred between bacterial lineages

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) (fig 1). We assume a strict molecular clock with a different molecular clock rate and HKY+Γ_4 (Hasegawa *et al.*, 1985) substitution model for the chromosome and 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 coalescent with plasmid transfer assumes that there is no inter-lineage recombination within the chromosome or plasmids. Thus, we masked sections with evidence of recombination in the chromosome using Gubbins (Croucher *et al.*, 2015). *S. sonnei* can carry multiple plasmids in addition to the pINV, one of which (spA) contains AMR genes streptomycin and sulphonamides. The relative prevalence of the smaller plasmids was reasonably constant over the sampling period (fig 1C), with spA and spB being detected in most samples, while spC was detected only in relatively few isolates. We observed some evidence for a recent increase in the proportion of bacterial lineages carrying any of the three small plasmids, notably from around 2016 for spC and spA and from late 2010's for spB (fig 1E). However, the proportion of ancestral lineages carrying spA and spC appears to have remained largely constant (fig 1E).

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 (fig 1F). In contrast, as shown in figure 1F, we find strong evidence for the three smaller plasmids (spA-spC) being transferred between bacterial lineages. These plasmid transfer events correspond to co-infection events, subsequent exchange of plasmids within a host and subsequent onward transmission of the bacteria. We estimate large differences in the rates of plasmid transfer between pINV, spA, spB and spC (see fig S1). Importantly, the spA plasmid is inferred to have the highest transfer rate between bacterial lineages and is the only plasmid known to confer resistance, out of those considered here (see fig S1), with other plasmids displaying substantially lower rates of plasmid transfer (see fig S1).

We next computed the rate at which plasmids are being lost. 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. The virulence plasmid, which in *S. sonnei* is known to be often lost in culture (The *et al.*, 2016), but forms part of the core genome of all *Shigella* species, had the highest rate of being lost (fig 1G). The smaller plasmids were all lost at a similar rate (fig 1G), suggesting similar maintenance costs.

Accounting for the co-divergence of chromosomes and plasmids is essential for estimating rates of evolution on plasmids

Evolution of the genome of bacterial species occurs 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, sometimes by several fold (fig 2A). Single nucleotide polymorphisms (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

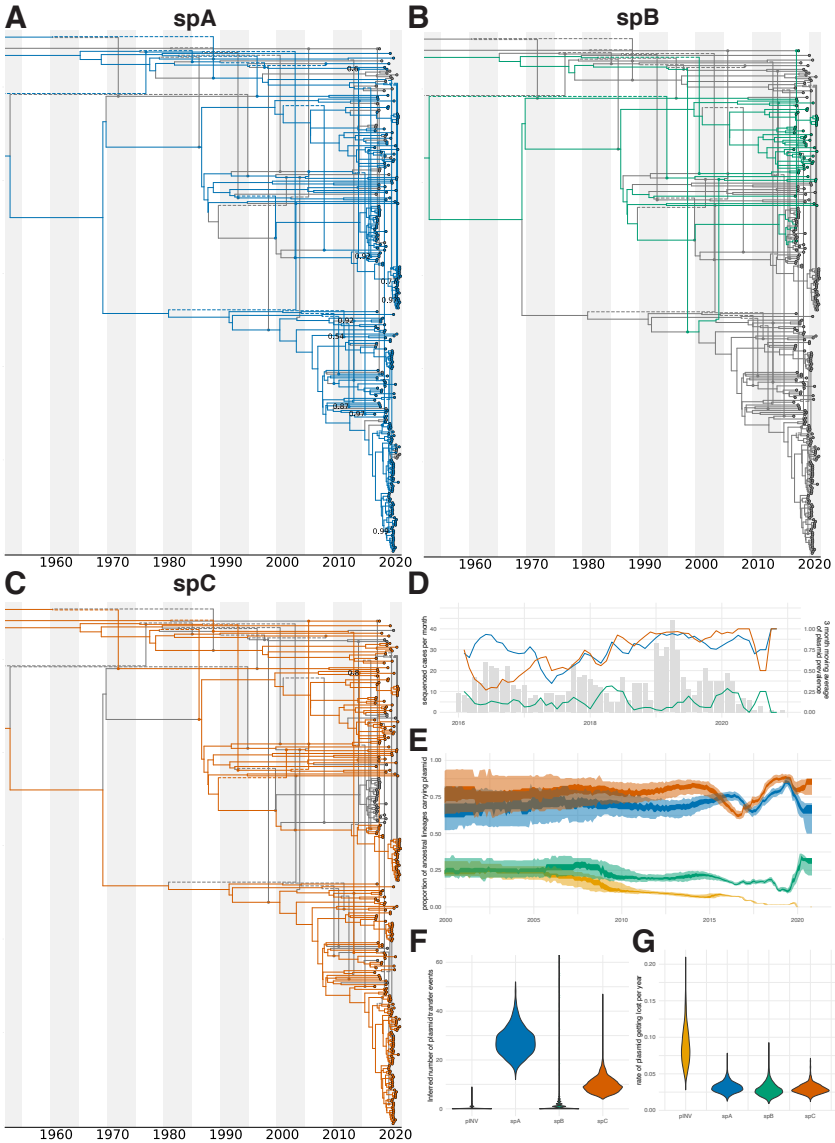

Figure 1: **Co-divergence of the core chromosome and plasmids in *Shigella sonnei*.** Maximum clade crediblity (MCC) networks (i.e. the network with the highest product of posterior node probabilities) of *Shigella sonnei* samples with the embedding of the plasmid trees for spA, spB and spC (**A-C**). Vertical lines are used to denote plasmid transfer events, where the circles denote the branch to which a plasmid was transferred. Dashed lines correspond branches from which plasmids branch off. Branches with colors correspond to those that carry a plasmid, whereas those in grey do not and tips labeled with color circles are samples for which the plasmid was available. The text denotes the posterior probability of plasmid transfer events for events with a posterior support of over 0.5. **D** Monthly number of sequenced *S. sonnei* cases between 2016 and 2020 and prevalence of three plasmids spA, spB and spC over time as a 3 month moving average. **E** Proportion of lineages in the past that contained a plasmid. **F** Posterior estimate of the number of recorded times a plasmid jumped between lineages. **G** Posterior estimate of the rate at which plasmids are getting lost.

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 larger number of SNPs due to its sheer size. As such, plasmids are less likely to contain as much information as the chromosome and may therefore be less likely to behave as measurably evolving populations (Drummond *et al.*, 2003; Biek *et al.*, 2015).

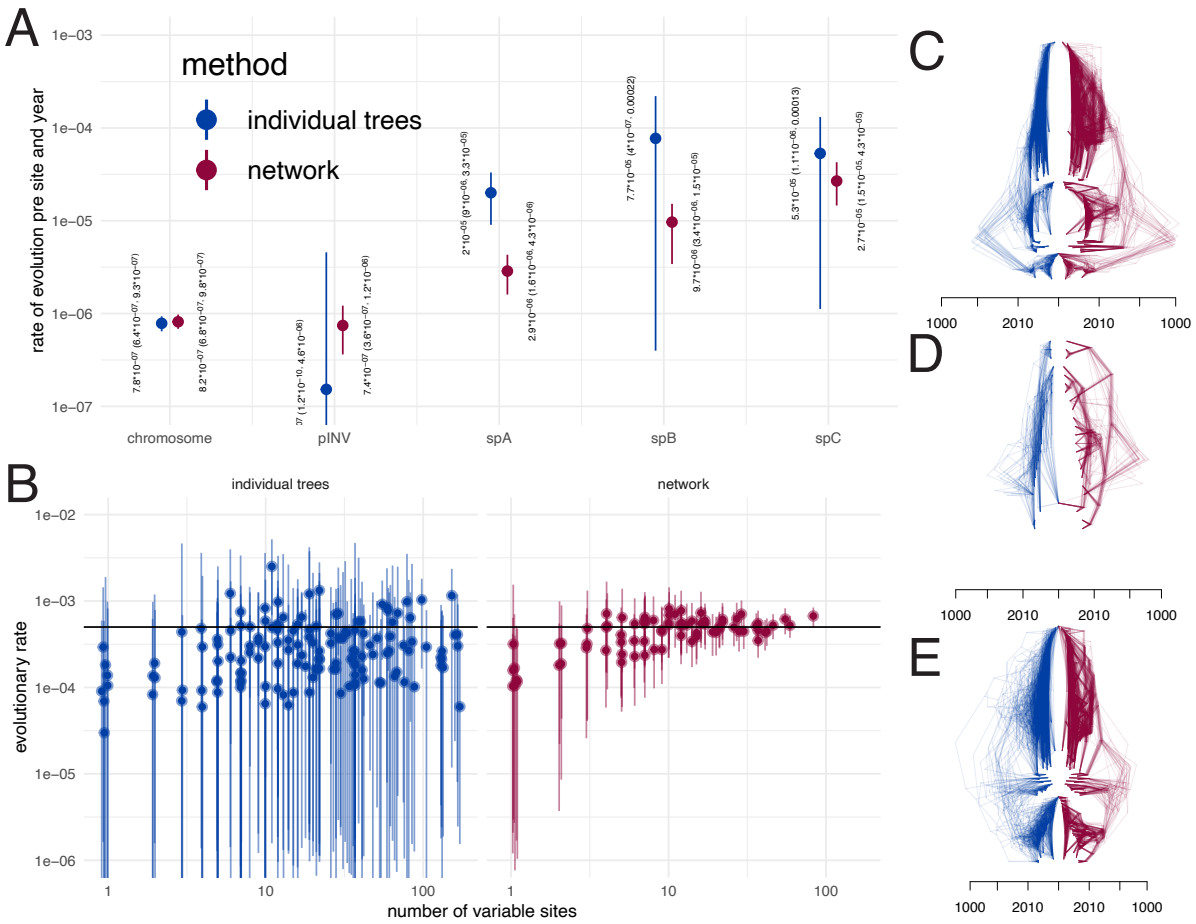

Figure 2: **Rates of evolution for plasmids and core chromosome in *Shigella sonnei*. A** Evolutionary rate posterior estimates of the the chromosomal and plasmid DNA of *S. sonnei* sequences isolated in Melbourne, Australia over several years. We compare estimates inferred by assuming an individual rate of evolution for the chromosome and plasmids to those where we explicitly model the joint evolutionary history of these lineages as a phylogenetic network. **B** Rates of evolution of plasmids using simulated datasets. 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. Density tree representation of the plasmid tree using the coalescent with plasmid transfer (red) and inferring the plasmid trees individually (blue) for spA (**C**), spB (**D**), spC (**E**). HPD: highest posterior density

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, spA, spB and spC plasmids individually. For the chromosome and the pINV, we used the SNP alignment, that only contains 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 mapping against 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 figure 2A, we found the chromosome to evolve at a rate of mean 7.8×10^{-7} subs/site/year (95% highest posterior density, HPD $6.6 - 9.4 \times 10^{-7}$), and the virulence plasmid to evolve at at rate of 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. 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 chromosome.

To further explore the impact of out approach in estimates of evolutionary rates, we compared the inferred rates for plasmids using the coalescent with plasmid transfer and individual tree inference using simulations. We simulated 50 phylogenetic networks under the coalescent with plasmids transfer with three plasmids sampled over five years. We assume that the chromosome and the three plasmids at a rate of 5×10^{-4} subs/site/unit time. The chromosome has a SNP alignment length of 8000bp, while the three plasmids had SNP alignments of 200bp, 100bp and 50bp respectively. These setting will produce approximately the same number of SNPs per unit of time as a chromosome of 4.8×10^{6}bp evolving at a rate of 8×10^{-7} subs/site unit time.

As shown in figure 2B 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 5 years is $5\,\text{years} \times 200\,\text{bp} \times 5 \times 10^{-4}$ 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 few SNPs occur (fig 2B).

This is also true for the tree topology of the plasmid trees, which is also aided by modeling the co-evolution of plasmids and core chromosome. As shown in figures 2C-E, particularly more recent topologies are resolved with higher precision.

Evidence for cross-species MDR plasmid exchange and steady growth of pKSR100 prevalence

We next investigated the movement of a multidrug resistance (MDR) plasmid that has been previously well-characterised using genomic epidemiological approaches, in 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 = 316 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 had ≥70% coverage of the pKSR100 reference were included in the alignment. We then subsampled 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 2, while estimating the rate of evolution of the MDR plasmid.

As shown in figure 3A&B, we found evidence for multiple events where the MDR plasmid jumped between bacterial lineages within species and also between species. 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. As shown in figure 3D, there is evidence for multiple introductions of plasmids into both species from each other, but also from unknown bacterial lineages. These could be other *Shigella* lineages or other from other bacterial species in the same ecological niches as has been previously reported (Duy *et al.*, 2020).

We next computed the proportion of lineages in the past that carried the plasmid pKSR100. As shown in figure 3C, 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 from when we have the most recent samples in the dataset.

Discussion

Our work presents a novel way to infer how plasmids move between bacterial lineages by using a phylogenetic network approach that explicitly models the co-divergence of plasmids with chromosomes. This represents a substantial advancement to the field of bacterial population genomics as it enables for the greater exploration of the plasmid movements within bacterial pathogens over time. In line with other research we find the co-divergence of virulence plasmid, pINV, with the chromosome of *S. sonnei* (The *et al.*, 2016), and the movement of small plasmids within the *S. sonnei* population. Further, we find evidence for multiple MDR plasmid transfer events between *S. sonnei* lineages, but also between *S. sonnei* and *S. flexneri* lineages (Baker *et al.*, 2015; Ingle *et al.*, 2019; Locke *et al.*, 2021; Mason *et al.*, 2022). Future work could explore where these plasmid are originating from to improve our understanding of how AMR genes move between species, by incorporating other bacterial species into this analyses.

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. The latter pertains to the product of the evolutionary rate and the number of sites. Outbreaks of many bacterial species, including those of *Shigella spp.* contain enough information to calibrate the molecular clock (Duchêne *et al.*, 2016). Our results show that plasmid sequence data alone are insufficient to calibrate the molecular clock, such that joint analyses of chromosome and plasmid data are essential to understand plasmid evolution.

Explicitly modeling the co-divergence of plasmids and core genomes also allows us to quantify the number of these events, the timings of introductions, the lineages where plasmids were introduced from, while also accounting for uncertainty in the genomic data. Tracking the movement of plasmids over time has been difficult, but is increasingly of interest to to better understand the epidemiology of bacterial outbreaks. As such, this provides a framework to study other bacterial populations where the plasmid dynamics are less clear. These approaches would be immediately relevant to drug-resistant plasmids, but could also be extended to virulence plasmids or where there has been reported convergence of AMR and virulence (Lam *et al.*, 2019). Such as investigating whether the expansion of a plasmid was from one introduction and subsequent expansion, or the result of repeated introductions.

Currently, we assume that each plasmid has a neutral fitness effect, meaning that lineages carrying a plasmid are assumed to be equally fit as lineages that do not. This assumption could in principle be relaxed to study the fitness benefits and costs of plasmids on a population level by modeling the fitness of a lineage as a function of the presence or absence of a plasmid using phylogenetic fitness models (Łuksza and Lässig, 2014). 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 (Müller *et al.*, 2017, 2018; Stolz *et al.*, 2022).

Finally, we showed that modeling the co-divergence of plasmid and chromosomal DNA allows 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 timescale of

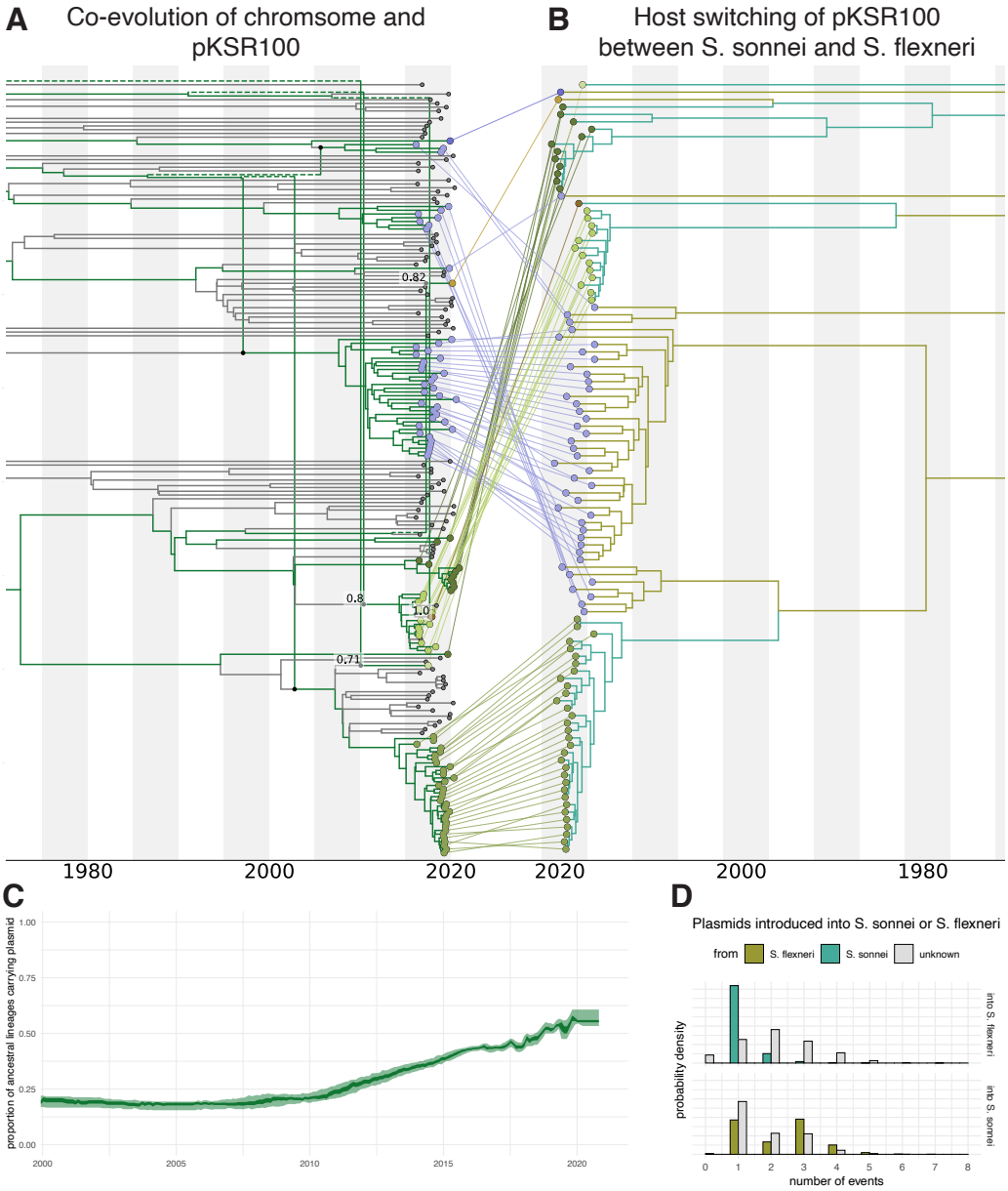

Figure 3: **Transmission of pKSR100 between *S. sonnei* and *flexneri*.** MCC network of *S. sonnei* and *flexneri* samples with the embedding of the psk100 plasmid tree **A**. The text denote the posterior support values for plasmid transfer events. **B** Plasmid tree of pKSR100 with the host species *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. **C** Proportion of ancestral lineages which carried the pKSR100 plasmid. **D** Distribution of the number of introductions of the MDR plasmid into *S. sonnei* and *S. flexneri*. These events are computed for the posterior distribution of networks with y-axis denoting the probability density for the number of events. MCC: maximum clade credibility. MDR: multidrug resistance

point mutations of epidemiological relevance. Importantly, the only source of evolutionary information that we consider are point mutations. Novel approaches that model, for example, rearrangements of genes on a plasmid could provide additional insight into the evolutionary dynamics of those plasmids, but also the movement of plasmid between lineages. Such approaches would have applications to better understand the movement of e.g. drug resistant plasmids both locally, within specific clinical settings, or internationally, such as tracking the dissemination of plasmids of interest across the globe.

Methods

Coalescent with plasmid transfer

Different bacterial lineages can exchange plasmids through different mechanisms. To model this process, we use a coalescent based model, related to the coalescent with reassortment (Müller *et al.*, 2020). In the coalescent with plasmid transfer model, we model a backwards in time process starting from sampled individuals (fig 4). 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 then 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 timing to the next plasmid transfer event is drawn from an exponential distribution with mean $\frac{1}{n*\rho}$. Upon a coalescent event, the parental lineage will carry the union of chromosomal or 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 to how reassortment is modeled in (Müller *et al.*, 2020) in that a plasmid transfer occurs relative to the chromosome and only one plasmid is transferred at a time. This is the backwards 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 happening 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 (Müller *et al.*, 2022). Importantly, the resulting phylogenetic network is not constrained to be tree based (as e.g (Didelot *et al.*, 2010; Vaughan *et al.*, 2017)) but allowed to have any possible structure one can simulated 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 a MCMC algorithm to characterize the joint posterior density. The posterior density is denoted as:

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

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 described 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 $P(\mu, \theta, \rho) = P(\mu)P(\theta)P(\rho)$.

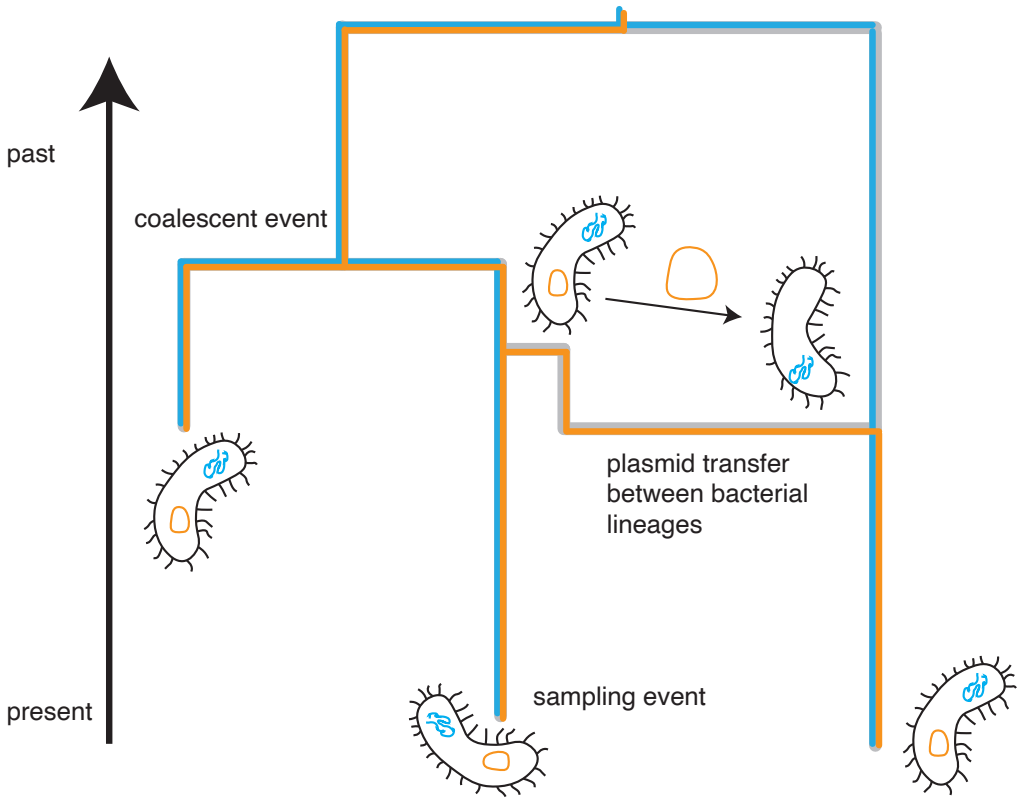

Figure 4: **Schematic representation of the coalescent with plasmid transfer model.** The coalescent with plasmid transfer models a backwards in time process where any two network lineages can coalesce (share a common ancestor). Additionally, network lineages can undergo a plasmid transfer event, modeled backwards in time as one of the plasmid lineages branching of the main branch. How rapidly two lineages share a common ancestor backwards in time is given by the effective population size and the rates of plasmid transfer denote the rate of observing plasmid transfer events backwards in time.

Network Likelihood

As we assume that there is no interlineage 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:

$$P(D|N,\mu) = \prod_{i=1}^{chromosome+nrplasmid} P(D_i|T_i,\mu), \tag{2}$$

The tree likelihood calculations use the default implementation of the tree likelihood in BEAST2 (Bouckaert *et al.*, 2019) and can use beagle (Ayres *et al.*, 2012) 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 (Bouckaert *et al.*, 2019).

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 a unstructured coalescent population model that described effective population sizes Ne over time, and ρ the per plasmid transfer rate . 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):

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

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:

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

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

$$P(event_i|L_i,\theta,\rho) = \rho. \tag{5}$$

This event contribution can be generalized to account for different rates of transfer for different plasmids by using 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 coalescent, nor any plasmid transfer event in interval i. We can therefore write:

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

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

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

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

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

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 those 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 (Müller *et al.*, 2020) and in (Müller *et al.*, 2022), but condition on only one plasmid jumping between bacterial lineages at a time. We here summarize each MCMC operator briefly:

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 (Bordewich *et al.*, 2017). Similar to Müller *et al.* (2022), 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 to change 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 analogue to the Gibbs operator in (Müller *et al.*, 2020) for reassortment 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 that the roots of the individual plasmid trees. As in Müller *et al.* (2022), we assume the plasmid transfer rate to be reduced prior to the individual plasmid trees having reached their root. As shown in Müller *et al.* (2022), this assumption does not affect parameter inferences, but can speed up inference

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 can compare the distributions of networks simulated under a set of parameters to the ones 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 figure S3, the sampled and simulated network distributions match.

We next perform 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 sampled the parameters for simulations. As shown in figures S4 and S5, 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 in 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.

Dataset

S. sonnei (n = 789) and *S. flexneri* (n = 316) 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 (Ingle *et al.*, 2020).

Alignments of the core genome were generated for the both the *S. sonnei* and *S. flexneri* isolates. The 789 *S. sonnei* were aligned to the reference *S. sonnei* chromosome Ss046 (accession number NC 007384) to call SNPs using Snippy v.4.6.0 (https://github.com/tseemann/snippy), with filtering of phage regions identified using PHASTER (Arndt *et al.*, 2016) and recombination detection undertaken with Gubbins (v2.4.1) (Croucher *et al.*, 2015). SNPsites (v2.5.1) (Page *et al.*, 2016) was used to extract the variant SNPs, resulting in a SNP alignment of 7,640. The same approach was used for the 316 *S. flexneri* isolates using the reference *S. flexneri* 2a str 301 (accession number NC 004337) resulting in a SNP alignment of 41,041.

All 789 *S. sonnei* were also aligned to the four plasmids of Ss046 using using Snippy v.4.4.5. These include 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). An alignment for each plasmid was generated for isolates which had $\geq$70% coverage of each plasmid sequence. 89 SNPs in the 46 isolates where the virulence plasmid was detected and this alignment was used in the model. For the three small plasmids of Ss046, the full alignment (including gaps and N's) was used instead of the core SNP alignment alone.

All *S. sonnei* and *S. flexneri* isolates were aligned to the MDR *S. flexneri* plasmid pKSR100 strain SF7955 (accession number LN624486, 73,047 bases). An alignment was generated for all 587 *S. sonnei* and *S. flexneri* isolates which had $\geq$70% coverage of the MDR plasmid. This MDR plasmid has been found in *S. sonnei* and *S. flexneri* lineages circulating in MSM populations since 2015 (Baker *et al.*, 2015).

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`.

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 (Bouckaert, 2010) uses and adapted version of the one implemented as part of the phangorn package (Schliep, 2011)

Funding

N.F.M. is supported by NIH NIGMS R35 GM119774. S.D. is supported by the Australian Research Council (DE190100805). 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.

Acknowledgments

The authors would like to acknowledge the members of the Microbiological Diagnostic Unit Public Health Laboratory in the Enterics and Bioinformatics sections including Dr Susan Ballad, Mary Valcanis, Jessica Barden, Dr Jake Lacey and Dr Kristy Horan.

References

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D. S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Research*, **44**(W1), W16–W21.

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., *et al.* (2012). Beagle: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic biology*, **61**(1), 170–173.

Baker, K. S., Dallman, T. J., Ashton, P. M., Day, M., Hughes, G., Crook, P. D., Gilbart, V. L., Zittermann, S., Allen, V. G., Howden, B. P., *et al.* (2015). Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: a cross-sectional study. *The Lancet infectious diseases*, **15**(8), 913–921.

Bengtsson, R. J., Simpkin, A. J., Pulford, C. V., Low, R., Rasko, D. A., Rigden, D. J., Hall, N., Barry, E. M., Tennant, S. M., and Baker, K. S. (2022). Pathogenomic analyses of Shigella isolates inform factors limiting shigellosis prevention and control across LMICs. *Nature Microbiology*, **7**(2), 251–261.

Biek, R., Pybus, O. G., Lloyd-Smith, J. O., and Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends in ecology & evolution*, **30**(6), 306–313.

Bordewich, M., Linz, S., and Semple, C. (2017). Lost in space? generalising subtree prune and regraft to spaces of phylogenetic networks. *Journal of theoretical biology*, **423**, 1–12.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., *et al.* (2019). Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology*, **15**(4), e1006650.

Bouckaert, R. R. (2010). Densitree: making sense of sets of phylogenetic trees. *Bioinformatics*, **26**(10), 1372–1373.

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., Parkhill, J., and Harris, S. R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, **43**(3), e15–e15.

Didelot, X., Lawson, D., Darling, A., and Falush, D. (2010). Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics*, **186**(4), 1435–1449.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., and Rodrigo, A. G. (2003). Measurably evolving populations. *Trends in ecology & evolution*, **18**(9), 481–488.

Duchêne, S., Holt, K. E., Weill, F.-X., Le Hello, S., Hawkey, J., Edwards, D. J., Fourment, M., and Holmes, E. C. (2016). Genome-scale rates of evolutionary change in bacteria. *Microbial genomics*, **2**(11).

Duy, P. T., Nguyen, T. N. T., Thuy, D. V., The, H. C., Alcock, F., Boinett, C., Thanh, H. N. D., Tuyen, H. T., Thwaites, G. E., Rabaa, M. A., and Baker, S. (2020). Commensal Escherichiacoli are a reservoir for the transfer of XDR plasmids into epidemic fluoroquinolone-resistant Shigella sonnei. *Nature Microbiology*, **5**(2), 256–264.

Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*, **22**(2), 160–174.

Hawkey, J., Paranagama, K., Baker, K. S., Bengtsson, R. J., Weill, F.-X., Thomson, N. R., Baker, S., Cerdeira, L., Iqbal, Z., Hunt, M., Ingle, D. J., Dallman, T. J., Jenkins, C., Williamson, D. A., and Holt, K. E. (2021). Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, Shigella sonnei. *Nature Communications*, **12**(1), 2684.

Hawkey, J., Wyres, K. L., Judd, L. M., Harshegyi, T., Blakeway, L., Wick, R. R., Jenney, A. W. J., and Holt, K. E. (2022). ESBL plasmids in Klebsiella pneumoniae: diversity, transmission and contribution to infection burden in the hospital setting. *Genome Medicine*, **14**(1), 97.

Holt, K. E., Baker, S., Weill, F.-X., Holmes, E. C., Kitchen, A., Yu, J., Sangal, V., Brown, D. J., Coia, J. E., Kim, D. W., Choi, S. Y., Kim, S. H., Silveira, W. D. d., Pickard, D. J., Farrar, J. J., Parkhill, J., Dougan, G., and Thomson, N. R. (2012). Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nature Genetics*, **44**(9), 1056–1059.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, **23**(2), 183–201.

Ingle, D. J., Easton, M., Valcanis, M., Seemann, T., Kwong, J. C., Stephens, N., Carter, G. P., Silva, A. G. d., Adamopoulos, J., Baines, S. L., Holt, K. E., Chow, E. P. F., Fairley, C. K., Chen, M. Y., Kirk, M. D., Howden, B. P., and Williamson, D. A. (2019). Co-circulation of Multidrug-resistant Shigella Among Men Who Have Sex With Men in Australia. *Clinical Infectious Diseases*, **69**(9), 1535–1544.

Ingle, D. J., Andersson, P., Valcanis, M., Barnden, J., Silva, A. G. d., Horan, K. A., Seemann, T., Easton, M., Williamson, D. A., Sherry, N. L., and Howden, B. P. (2020). Prolonged Outbreak of Multidrug-Resistant Shigella sonnei Harboring blaCTX-M-27 in Victoria, Australia. *Antimicrobial Agents and Chemotherapy*, **64**(12), e01518–20.

Ingle, D. J., Ambrose, R. L., Baines, S. L., Duchene, S., Silva, A. G. d., Lee, D. Y. J., Jones, M., Valcanis, M., Taiaroa, G., Ballard, S. A., Kirk, M. D., Howden, B. P., Pearson, J. S., and Williamson, D. A. (2021). Evolutionary dynamics of multidrug resistant Salmonella enterica serovar 4,[5],12:i:- in Australia. *Nature Communications*, **12**(1), 4786.

Lam, M. M. C., Wyres, K. L., Wick, R. R., Judd, L. M., Fostervold, A., Holt, K. E., and Löhr, I. H. (2019). Convergence of virulence and MDR in a single plasmid vector in MDR Klebsiella pneumoniae ST15. *Journal of Antimicrobial Chemotherapy*, **74**(5), 1218–1222.

Locke, R. K., Greig, D. R., Jenkins, C., Dallman, T. J., and Cowley, L. A. (2021). Acquisition and loss of CTX-M plasmids in Shigella species associated with MSM transmission in the UK. *Microbial Genomics*, **7**(8), 000644.

Łuksza, M. and Lässig, M. (2014). A predictive fitness model for influenza. *Nature*, **507**(7490), 57–61.

Mason, L., Greig, D., Cowley, L., Partridge, S., Martinez, E., Blackwell, G., Chong, C., Silva, M. D., Bengtsson, R., Draper, J., Ginn, A., Sandaradura, I., Sim, E., Iredell, J., Sintchenko, V., Ingle, D., Howden, B., Lefevre, S., Njamkepo, E., Weill, F.-X., Ceyssens, P.-J., Jenkins, C., and Baker, K. (2022). The evolution and international spread of extensively drug resistant Shigella sonnei. *research square*.

Müller, N. F., Rasmussen, D. A., and Stadler, T. (2017). The structured coalescent and its approximations. *Molecular biology and evolution*, **34**(11), 2970–2981.

Müller, N. F., Rasmussen, D., and Stadler, T. (2018). Mascot: Parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, **34**(22), 3843–3848.

Müller, N. F., Stolz, U., Dudas, G., Stadler, T., and Vaughan, T. G. (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.

Müller, N. F., Kistler, K. E., and Bedford, T. (2022). A bayesian approach to infer recombination patterns in coronaviruses. *Nature communications*, **13**(1), 1–9.

Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Aguilar, G. R., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., *et al.* (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, **399**(10325), 629–655.

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., and Harris, S. R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*, **2**(4), e000056.

Park, S. E., Pham, D. T., Boinett, C., Wong, V. K., Pak, G. D., Panzner, U., Espinoza, L. M. C., Kalckreuth, V. v., Im, J., Schütt-Gerowitt, H., Crump, J. A., Breiman, R. F., Adu-Sarkodie, Y., Owusu-Dabo, E., Rakotozandrindrainy, R., Soura, A. B., Aseffa, A., Gasmelseed, N., Keddy, K. H., May, J., Sow, A. G., Aaby, P., Biggs, H. M., Hertz, J. T., Montgomery, J. M., Cosmas, L., Olack, B., Fields, B., Sarpong, N., Razafindrabe, T. J. L., Raminosoa, T. M., Kabore, L. P., Sampo, E., Teferi, M., Yeshitela, B., Tayeb, M. A. E., Sooka, A., Meyer, C. G., Krumkamp, R., Dekker, D. M., Jaeger, A., Poppert, S., Tall, A., Niang, A., Bjerregaard-Andersen, M., Løfberg, S. V., Seo, H. J., Jeon, H. J., Deerin, J. F., Park, J., Konings, F., Ali, M., Clemens, J. D., Hughes, P., Sendagala, J. N., Vudriko, T., Downing, R., Ikumapayi, U. N., Mackenzie, G. A., Obaro, S., Argimon, S., Aanensen, D. M., Page, A., Keane, J. A., Duchene, S., Dyson, Z., Holt, K. E., Dougan, G., Marks, F., and Baker, S. (2018). The phylogeography and incidence of multi-drug resistant typhoid fever in sub-Saharan Africa. *Nature Communications*, **9**(1), 5094.

Partridge, S. R., Kwong, S. M., Firth, N., and Jensen, S. O. (2018). Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clinical Microbiology Reviews*, **31**(4).

Robertson, J. and Nash, J. H. E. (2018). MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics*, **4**(8), e000206.

Rozwandowicz, M., Brouwer, M. S. M., Fischer, J., Wagenaar, J. A., Gonzalez-Zorn, B., Guerra, B., Mevius, D. J., and Hordijk, J. (2018). Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae. *Journal of Antimicrobial Chemotherapy*, **73**(5), 1121–1137.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in r. *Bioinformatics*, **27**(4), 592–593.

Schroeder, G. N. and Hilbi, H. (2008). Molecular Pathogenesis of Shigella spp.: Controlling Host Cell Signaling, Invasion, and Death by Type III Secretion. *Clinical Microbiology Reviews*, **21**(1), 134–156.

Stolz, U., Stadler, T., Müller, N. F., and Vaughan, T. G. (2022). Joint inference of migration and reassortment patterns for viruses with segmented genomes. *Molecular biology and evolution*, **39**(1), msab342.

The, H. C., Thanh, D. P., Holt, K. E., Thomson, N. R., and Baker, S. (2016). The genomic signatures of Shigella evolution, adaptation and geographical spread. *Nature Reviews Microbiology*, **14**(4), 235–250.

Vaughan, T. G., Welch, D., Drummond, A. J., Biggs, P. J., George, T., and French, N. P. (2017). Inferring ancestral recombination graphs from bacterial genomic data. *Genetics*, **205**(2), 857–870.

World Health Organization (2020). GLASS whole-genome sequencing for surveillance of antimicrobial resistance. Technical report.

Yang, F., Yang, J., Zhang, X., Chen, L., Jiang, Y., Yan, Y., Tang, X., Wang, J., Xiong, Z., Dong, J., Xue, Y., Zhu, Y., Xu, X., Sun, L., Chen, S., Nie, H., Peng, J., Xu, J., Wang, Y., Yuan, Z., Wen, Y., Yao, Z., Shen, Y., Qiang, B., Hou, Y., Yu, J., and Jin, Q. (2005). Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic Acids Research*, **33**(19), 6445–6458.

Supplementary material

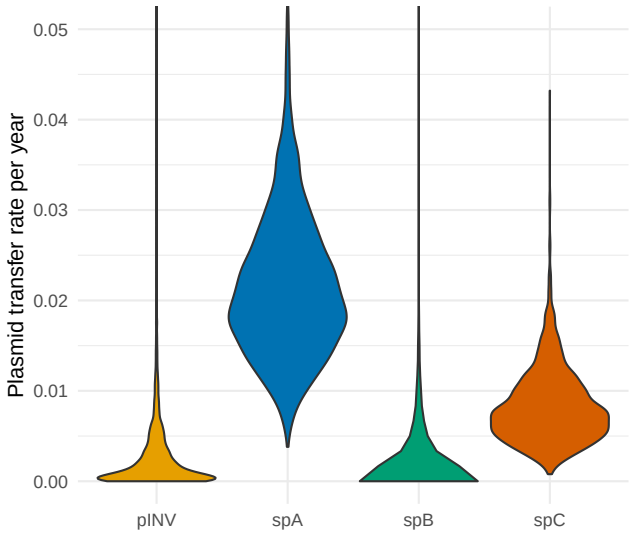

Figure S1: **Inferred rate at which plasmids are transferred per plasmid per year.** Posterior distribution of plasmid transfer rates inferred from *S. sonnei* sequence data (y-axis), for the different plasmids in the analyses (x-axis).

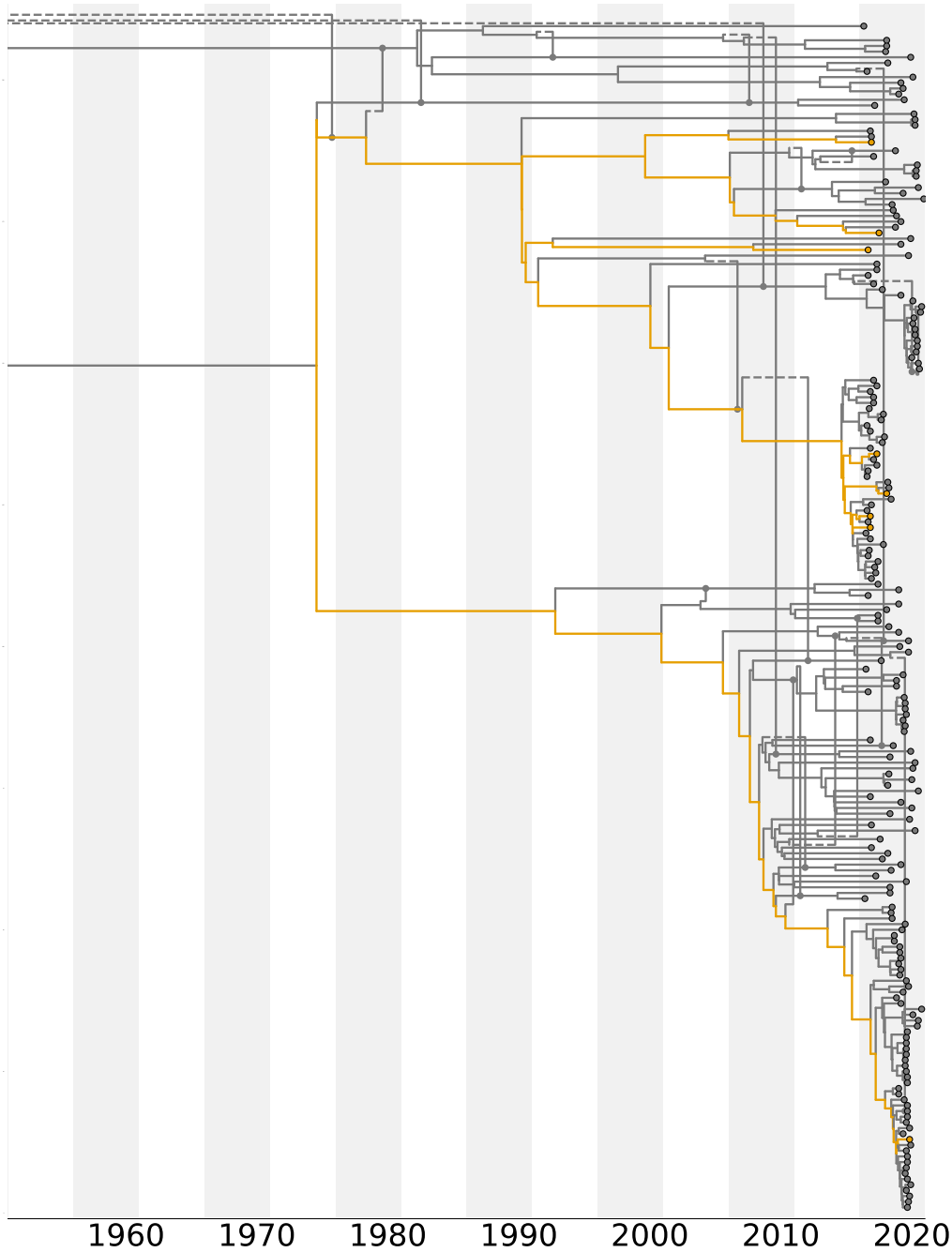

Figure S2: **Embedding of pINV plasmid tree in the coalescent with plasmid network.** Here, we show the embedding of the virulence plasmid tree within the maximum clade credibility coalescent with plasmid network. The virulence plasmid is around 200kb long and is not inferred to jump between bacterial lineages.

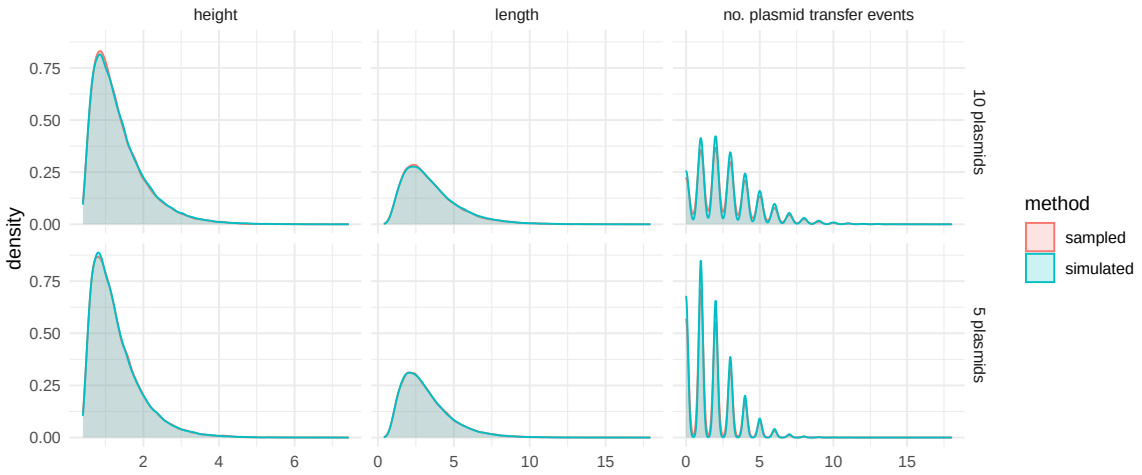

Figure S3: **Comparison of network height, length and plasmid transfer events between sampled and simulated networks.** To validate the implementation of CoalPT, we simulated networks under the CoalPT model, once with 5 plasmids and once with 10 plasmids. We then sampled phylogenetic networks under our 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.

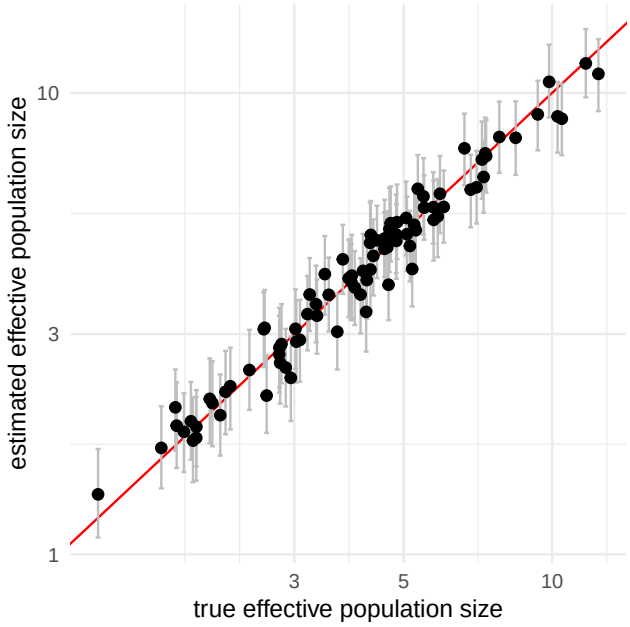

Figure S4: **Inferred effective population sizes from simulated data.** To test the performance of the coalescent with plasmid transfer, we simulated 100 networks in a well calibrated simulated study. The effective population sizes were sampled from a Lognormal distribution with M=1.4844 and S=0.5. The 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.

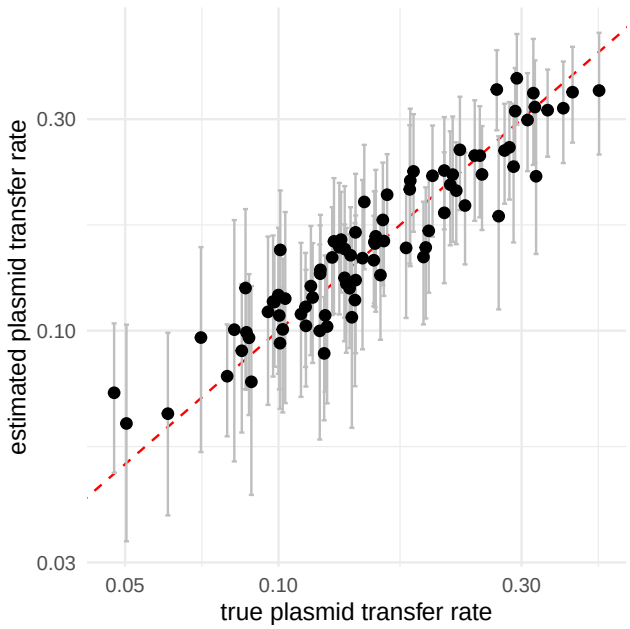

Figure S5: **Inferred plasmid transfer rates from simulated data.** Here, we show the inferred plasmid transferred rates(y-axis) compared to the true/simulated rates on the x-axis. These estimates are from the same analyses as the ones in fig S4. The point denote the median estimate and the error bars the lower 95% highest posterior density interval.