PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Huisman JS, Vaughan TG, Egli A, Tschudin-Sutter S, Stadler T, Bonhoeffer S. 2022 The effect of sequencing and assembly on the inference of horizontal gene transfer on chromosomal and plasmid phylogenies. *Phil. Trans. R. Soc. B* **377**: 20210245. https://doi.org/10.1098/rstb.2021.0245

Received: 15 November 2021
Accepted: 30 March 2022

One contribution of 11 to a discussion meeting issue 'Genomic population structures of microbial pathogens'.

Subject Areas:
bioinformatics, computational biology, evolution, microbiology

Keywords:
antibiotic resistance, plasmid, phylogenetics, whole-genome sequencing, assembly, long-read sequencing

Author for correspondence:
Jana S. Huisman
e-mail: jana.huisman@env.ethz.ch

†These authors contributed equally to this study.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6080814.

THE ROYAL SOCIETY PUBLISHING

The effect of sequencing and assembly on the inference of horizontal gene transfer on chromosomal and plasmid phylogenies

Jana S. Huisman[1,2,3], Timothy G. Vaughan[2,3], Adrian Egli[4,6], Sarah Tschudin-Sutter[5,7], Tanja Stadler[2,3,†] and Sebastian Bonhoeffer[1,†]

[1]Department of Environmental Systems Science, ETH Zurich, 8092 Zurich, Switzerland
[2]Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland
[3]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[4]Division of Clinical Microbiology, University Hospital Basel, Petersgraben 4, 4031 Basel, Switzerland
[5]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Petersgraben 4, 4031 Basel, Switzerland
[6]Department of Biomedicine, University of Basel, Hebelstrasse 20, 4031 Basel, Switzerland
[7]Department of Clinical Research, University of Basel, Schanzenstrasse 55, 4031 Basel, Switzerland

JSH, 0000-0002-1782-8109; TS, 0000-0001-6431-535X

The spread of antibiotic resistance genes on plasmids is a threat to human and animal health. Phylogenies of bacteria and their plasmids contain clues regarding the frequency of plasmid transfer events, as well as the co-evolution of plasmids and their hosts. However, whole genome sequencing data from diverse ecological or clinical bacterial samples are rarely used to study plasmid phylogenies and resistance gene transfer. This is partially due to the difficulty of extracting plasmids from short-read sequencing data. Here, we use both short- and long-read sequencing data of 24 clinical extended-spectrum β-lactamase (ESBL)-producing *Escherichia coli* to estimate chromosomal and plasmid phylogenies. We compare the impact of different sequencing and assembly methodologies on these phylogenies and on the inference of horizontal gene transfer. We find that chromosomal phylogenies can be estimated robustly with all methods, whereas plasmid phylogenies have more variable topology and branch lengths across the methods used. Specifically, hybrid methods that use long reads to resolve short-read assemblies (HybridSPAdes and Unicycler) perform better than those that started from long reads during assembly graph generation (Canu). By contrast, the inference of plasmid and antibiotic resistance gene transfer using a parsimony-based criterion is mostly robust to the choice of sequencing and assembly method.

This article is part of a discussion meeting issue 'Genomic population structures of microbial pathogens'.

1. Introduction

The rapid spread of antibiotic resistance is a global threat for human and animal health. Antibiotic-resistant infections are associated with increased morbidity and mortality [1] and carry a substantial economic cost due to the use of second-line treatment options, treatment complications and longer hospital stays [2]. The spread of antibiotic resistance genes is aided by their association with mobile genetic elements that are transferred between diverse bacterial populations [3]. In Enterobacterales, an order of Gram-negative bacteria that cause both nosocomial and community-associated infections in humans, conjugative plasmids are considered the main driver of the horizontal transfer of antibiotic resistance genes [4–6].

In the epidemiology of antibiotic resistance, the distinction between chromosome- and plasmid-driven spread is important for monitoring,

transmission risk assessment and the planning of interventions [4,6]. Some resistance is spread mostly by successful bacterial lineages (or 'clones'), in association with one or several resistance plasmids [4]. A prime example is *Escherichia coli* sequence type (ST) 131 with a variety of IncF plasmids [4,7]. Other resistance genes are carried on more promiscuous plasmids, such as plasmid pOXA-48, which easily spread to various species [6]. Surveillance strategies including the accurate typing of resistance genes, plasmids and bacterial host lineages are essential to monitor plasmid-mediated spread of resistance both in hospital settings and between epidemiological compartments such as animals, humans and the environment [8–10].

Over the past 10–15 years, whole-genome sequencing (WGS) has become ever more important for molecular epidemiology, as it supplies detailed information on the presence and genetic context of resistance genes, in addition to strain and plasmid typing [11,12]. Three main sequencing methods are currently used for microbial genomics. Short-read sequencing, typically on Illumina machines, is cost-efficient and produces reads with a low error rate [12,13]. However, the short-read length is often not enough to distinguish repeat regions, leading to a fragmented assembly [13–15]. To overcome this issue, both Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (NP) have developed single molecule, long-read sequencing methods that produce reads with a median length of 10 kb [12]. Long reads allow for a more contiguous assembly, but result in a greater cost per sample, and—for NP—higher error rates, although both have been decreasing in recent years [12].

A primary reason why plasmid phylogenies have been understudied is that it is difficult to detect plasmids in short-read WGS data in an accurate and automated fashion [12,15]. Plasmids are often assembled into several different contiguous sequences (contigs), which can be identified as plasmid sequences only by the presence of plasmid-specific genes [16,17], or a coverage or GC-content that differs from the chromosome [18]. Typically long-read sequence information is needed to assemble the full plasmid sequence. Multiple studies have compared short-read, long-read and hybrid methods in their ability to reconstruct plasmid sequences and determine the location of resistance genes [12–14]. These studies generally concluded that hybrid assembly, especially combining Illumina with NP, enables accurate plasmid identification and localization of resistance genes. However, although one study speculated that the observed differences may affect phylogenetic analyses [13], none explicitly tested the effect of sequencing and assembly choices on downstream inferences. Since different phylogenetic analyses are sensitive to different errors, it is not clear which sequencing strategy has the optimal cost/benefit ratio to study the phylogenetics of bacteria and their plasmids.

Here, we used genomic data of 24 clinical extended-spectrum β-lactamase (ESBL)-producing *E. coli* to study the impact of sequencing and assembly methods on the phylogenetic inference of chromosomal and plasmid trees. We compared *de novo* assemblies based on Illumina, PacBio and NP read sets, assembled both independently and in hybrid fashion. The inferred plasmid phylogenies differed substantially across the assembly method combinations tested. However, horizontal transmission of plasmids and the associated antibiotic-resistance genes could be quantified even in the absence of long-read information.

2. Methods

(a) *Escherichia coli* isolates

The 24 ESBL-producing *E. coli* strains were previously isolated at the University Hospital Basel (UHB) and an affiliated long-term care centre, the Felix Platter Hospital (FPH), in the context of a hospital transmission study [19]. *E. coli* strains from routine diagnostics were identified as ESBL-producing strains via two different approaches: (i) *E. coli* antimicrobial resistance against third-generation cephalosporins (cefpodoxime, ceftriaxone, ceftazidime) was confirmed using a phenotypic Rosco Disk assay (Rosco, Taastrup, Denmark); and (ii) *E. coli* strains growing on ESBL Chromogenic screening agar plates (chrom ID ESBL, bioMérieux, Marcy-l'Étoile, France) were also confirmed using a phenotypic Rosco Disk assay. Antibiotic minimal inhibitory concentration (MICs) were interpreted according to EUCAST guidelines (www.eucast.org). Strains were stored at −80°C.

The set of strains contains four known transmission pairs, as well as one pair of strains that were isolated from the same patient half a year apart. The strains are representative of the known clinical diversity of ESBL-producing *E. coli* [20]. From the 24 isolates, 4 are from phylogroup B1, 11 from B2 (all ST131), 1 from C, 5 from D, and 3 from phylogroup F (see electronic supplementary material, table S2; Clermont Typing using EzClermont [21]).

(b) Library preparation and sequencing

The samples were sequenced using Illumina, PacBio and NP. DNA was extracted at the UHB, using the QIAamp® DNA mini kit (QIAGEN) with the QIAcube® robot (QIAGEN) according to the manufacturer's instructions. For Illumina sequencing, libraries were prepared with the Nextera XT library preparation kit (Illumina) and sequenced on a MiSeq machine (300 bp paired-end) at the UHB. PacBio sequencing was performed without size selection, on a PacBio Sequel machine at the Functional Genomics Center Zürich. The samples were multiplexed in three pools (barcoding), where one pool had to be resequenced because of low yield in the first round. The resulting coverage of PacBio reads was not uniform across the 24 samples (mean coverage: 106 ± 43). NP sequencing was performed following the protocol developed by Noll *et al.* [22], on a MinION at the Biozentrum Basel. In brief, libraries were prepared using the 1D ligation sequencing kit (LSK-108) and native barcoding expansion (EXP-NBD103; both from NP), without the shearing and repair step and with an increased amount of DNA. Base-calling was performed with albacore (v. 2.0.2; see http://nanoporetech.com/community) and barcode demultiplexing followed the consensus of both albacore and porechop (v. 0.2.3; see https://github.com/rrwick/Porechop).

(c) Assembly

The sequencing reads were assembled in a variety of ways, including in hybrid combinations of long- and short-read sequencing methods. An overview can be found in table 1. The Illumina reads were trimmed using Trimmomatic in paired-end mode [23]. The Illumina Nextera adapters were removed (2:30:10:8), as well as the leading and trailing three bases of each read. The quality trimming was performed using a minimal phred quality score of 20 per base. Reads with a length shorter than 36 bp were removed. FastQC (v. 0.11.7) was used for quality control on the trimmed reads [24]. Assembly of Illumina reads was performed using SPAdes (v. 3.11.1) [25], without further error correction and set to 'careful'. PacBio reads were subsetted for multiplex barcode quality above 45 (recommendation by PacBio). Both long-read methods were assembled by themselves using Canu (v. 1.7) [26]. The long and short reads were combined

Table 1. Number of samples with a given replicon, per assembly method. PB refers to PacBio reads, NP to Oxford Nanopore Technologies.

name	reads	assembly method
Illumina–SPAdes	Illumina	SPAdes
PB–Canu	PacBio	Canu
NP–Canu	Oxford Nanopore	Canu
PB–Canu–Hybrid	PacBio and Illumina	Canu + Polishing w. Illumina
PB–SPAdes–Hybrid	PacBio and Illumina	HybridSPAdes
PB–Unicycler–Hybrid	PacBio and Illumina	Unicycler
NP–Canu–Hybrid	Oxford Nanopore and Illumina	Canu + Polishing w. Illumina
NP–SPAdes–Hybrid	Oxford Nanopore and Illumina	HybridSPAdes
NP–Unicycler–Hybrid	Oxford Nanopore and Illumina	Unicycler

into hybrid assemblies, once by polishing the Canu assemblies with the Illumina reads using unicycler-polish (v. 0.4.7; based on the Pilon polishing tool [27]) and alternatively by assembling both long and short reads together using HybridSPAdes (v. 3.11.1) [28] and Unicycler (v. 0.4.7) [29]. The Canu assembler was run using a separate read correction step with high coverage settings (parameter *corOutCoverage* was set to 1000 to include short plasmids) and a subsequent trim-assemble step. HybridSPAdes assembly was performed using the basic settings. Unicycler was run in normal and 'conservative' mode, but since these assemblies did not differ substantially, we report results for the normal mode only.

(d) Assembly comparison

The assemblies were compared according to six different measures of assembly quality: first, the number of assembled contiguous sequences (# contigs); second, the N50, i.e. the value for which 50% of the assembly is contained in contigs equal to or larger than this value; and third, the error rate, i.e. the number of single nucleotide polymorphisms (# SNPs) and insertion-deletions (# indels) found in the assembly, divided by the total assembly size. The SNPs and indels were found by mapping the Illumina reads against the completed assembly, and calling errors using bcftools (v.1.7) with a quality score of 20 [30]. Fourth, the number of coding domain sequences (# CDS) found in the assembly. This information was extracted from the prokka annotations (v.1.13) [31], which in turn uses prodigal for gene prediction [32]. Fifth, the total length of the contigs on which an antibiotic resistance gene was detected. Sixth, the total length of the contigs on which a plasmid replicon was detected. Plasmid replicons and resistance genes were determined using abricate (v. 0.8.10; Torsten Seemann, https://github.com/tseemann/abricate) with the PlasmidFinder [17] and ResFinder [33] databases respectively.

(e) Chromosomal alignment

We estimated the chromosomal genomic information of a strain using core genome Multi-Locus Sequence Typing (cgMLST) [11,34]. We used chewBBACA [35] for allele calling against the Enterobase *E. coli* cgMLST scheme [36]. This scheme includes genes that are present in 95% of their *E. coli* assemblies, which currently number more than 100 000 and should thus lead to a highly stable core definition. The scheme includes 2513 'core' genes, each of which has on average 1052 known alleles [min = 20, max = 4748] (downloaded 14 Augurst 2019).

We then transformed the matrix of allele calls to one multi-FASTA per gene, and used MAFFT [37] (v. 7.313) to create a per-gene alignment. Simple concatenation of the gene sequences for each bacterial strain yielded an alignment ready for use in subsequent phylogenetic analysis. All Enterobase core genes were included in the alignment, independent of the number of samples they were present in.

(f) Plasmid alignment

We developed a pipeline to obtain a plasmid alignment from assembled WGS contigs. To extract putative plasmid sequences from the WGS assembly, we used genes known to be implicated in plasmid replication as probes. These genes are specific to a plasmid incompatibility group and are used for replicon (REP) typing against the PlasmidFinder database [17]. Once we identified all putative plasmid sequences with the same REP, these still contained regions of genome rearrangement and recombination. Thus, we annotate the putative plasmid sequences with Prokka [31], and determine the alignment for each REP using Roary [38]. All genes present in at least two samples were included in the alignment, as opposed to core genes only (set with the core definition parameter `cd = 0.08`).

(g) Phylogenetic analysis and tree comparison

The rooted chromosomal and plasmid time-trees were inferred using BEAST 2 [39]. Since the samples were closely spaced in time, the sampling dates contained little information about the timing and age of the tree. To aid the inference under the Hasegawa–Kishino–Yano (HKY) model, we constrained the mutation rate around the estimate for *E. coli* by Wielgoss *et al.* [40]. They determine a mutation rate of 8.9×10^{-11} $[4 \times 10^{-11} \ldots 14 \times 10^{-11}]$ mutations per base-pair per generation for a genome of 4.6×10^{6} bp in length. Assuming between 1000 to 10 000 generations per year, this leads to a parameter range of $0.18 \cdots 6.44$ mutations per genome per year. We captured this using a lognormal($M = 0.4$, $S = 1$) prior on the ClockRate. Here, we used mutation rates and generation time estimates from the laboratory, although the generation times in the wild are likely around 50 times longer than this [41]. We also assumed that plasmids and chromosomes have the same mutation rate, since they use the same DNA replication and repair machinery. However, a different mutation rate for the plasmids would not impact the relative comparison of the different assembly methods.

Trees can differ in their topology or branch lengths. Thus, we used three methods to compare the chromosomal and plasmid trees we obtained using different sequencing and assembly methods on the same samples. To compare the tree topology, we used CladeSetComparator from the Babel package for BEAST 2 (https://github.com/rbouckaert/Babel) [39]. This program uses the posterior of trees obtained through two separate Bayesian phylogenetic inferences, matches the clades from both

tree posteriors and reports the probability with which the clades are found in each posterior. In addition, we used the treespace R package (v. 1.1.4.1) to compare the overall separation of the inferred tree topologies in tree space (using the Kendall–Colijn distance metric between trees) [42]. To get a proxy for the combined branch length, we compared the inferred age of the trees (the tree 'height').

(h) Identifying horizontal gene transfer

Incongruence between plasmid and chromosomal phylogenies indicates a violation of the assumption of clonal inheritance, and thus points towards horizontal gene transfer. We quantified this incongruence using the Robinson–Foulds metric, which describes the distance between two phylogenetic trees *A* and *B* [43]. For rooted trees, it is calculated as the number of clades in tree *A* that are not present in *B*, added to the number of clades in tree *B* that are not present in *A*. Since this number depends on the size of the tree, we normalized by the maximal possible distance between *A* and *B* (yielding a value between 0 and 1). These statistics were calculated using the phangorn package in R [44]. For comparison against a sample of random trees, we used the `rmtree` function from the ape package (v.5.5) [45].

Horizontal gene transfer can also be assessed using a parsimony analysis: if one labels the tips of the chromosomal tree by the presence or the absence of a given plasmid or resistance gene in the whole genome assembly, the parsimony score describes the number of horizontal gene transfer events that are needed to explain this pattern of presence/absence. To put the estimated parsimony score into perspective, we compared against a null model that assumes the amount of plasmid transfer is very high, such that for each terminal branch the chance of observing the plasmid is equal to the frequency with which the plasmid is observed in the population. This was achieved by keeping the number of observed plasmid presences constant but randomizing the tips they were assigned to (1000 times). Significant departures from the null model were determined by comparing the observed parsimony to the empirical cumulative distribution function of the null model at a 0.05 significance threshold.

3. Results

(a) A comparison of sequencing and assembly methods

We set out to study the effect of sequencing and assembly methods on the inference of horizontal gene transfer in clinical *E. coli*. We used three different sequencing methods (Illumina, PacBio and Oxford Nanopore) and four single- or hybrid assembly methods on the same 24 ESBL-producing *E. coli* samples. In total, we tested nine different combinations of sequencing and assembly methods (an overview is given in table 1) and constructed both chromosomal and plasmid phylogenies for each combination (figure 1; electronic supplementary material, figures S5 and S6).

The general characteristics and quality of the assemblies differed substantially across the nine sequencing–assembly method combinations (electronic supplementary material, figure S1). As expected, the Illumina-SPAdes assemblies were most fragmented, as testified by the large amount of contigs (electronic supplementary material, figure S1*a*) and low N50 (electronic supplementary material, figure S1*b*). For long-read and hybrid methods, especially those involving PacBio reads, the distribution of the N50 statistic was also broad and often below the expected 5Mb of an *E. coli* genome. This shows that these methods struggled to assemble closed genomes, with large variability across the clinical samples. The error rates were quite low for all method combinations, with the exception of the Nanopore-only assembly (NP-Canu, electronic supplementary material, figure S1*c*). These errors also resulted in an elevated number of putative genes for NP-Canu (electronic supplementary material, figure S1*d*), likely due to spurious stop-codons and frame-shifts. When looking at the length of contigs with resistance genes (electronic supplementary material, figure S2), we found that Illumina recovered shorter contigs than all other methods. This is likely because most resistance genes in our isolates were flanked by insertion sequences with repeat regions. By contrast, the majority of contigs that carry plasmid genes were not substantially shorter than those found with other methods (with some notable exceptions for larger plasmids; see also figure 2). When long-read information was added, the assemblies were more contiguous and resistance genes could be assigned their place in the genome. Based on the combination of error profile and high N50, NP-Unicycler-Hybrid had the most desirable assembly characteristics on this dataset.

(b) The chromosomal tree can be determined equally well from all assemblies

To quantify the impact of the assembly method on the chromosomal phylogeny, we inferred nine chromosomal tree posteriors from the cgMLST alignment of each assembly. We compared these phylogenies according to both their topology and branch lengths (figure 1 and electronic supplementary material, figure S3). All methods, except NP-Canu, recovered the same distribution of tree topologies and maximum clade credibility 'consensus' trees (figure 1 and electronic supplementary material, figure S3*a*). This incongruence of the NP-Canu trees is likely due to the high error rate observed in the NP-Canu assemblies, and the resulting difficulty in locating coding domain sequences and constructing a cgMLST alignment (electronic supplementary material, figure S1*c*). Out of the 2513 *E. coli* genes probed for, we failed to detect 531 genes on average in the NP-Canu assemblies, as opposed to 7–67 genes for the other assembly methods. The NP-Canu trees were also notably shorter than those resulting from other methods (figure 1 and electronic supplementary material, figure S3*b*).

(c) Plasmid assemblies differ across the assembly methods

Since assembly methods differ in the length of the putative plasmid sequences they recover (electronic supplementary material, figure S2), we tested whether this also affects the length of plasmid alignments, and their subsequent tree inference.

To obtain a plasmid alignment from the assembled WGS contigs, we used Roary to extract the gene-by-gene alignment for each set of putative plasmid sequences [38]. These plasmid sequences were extracted from the WGS assembly by probing for genes known to be implicated in plasmid replication. This method of REP typing is specific to a plasmid incompatibility group [17], so we created separate plasmid alignments for each incompatibility group in our dataset. Some plasmids carry multiple REP genes [46] and are thus present in several separate alignments (e.g. a plasmid may be included in both IncFIA and IncFIB alignments). Table 2 lists the number of samples found to contain a given REP (the correspondence to each sample is given in electronic

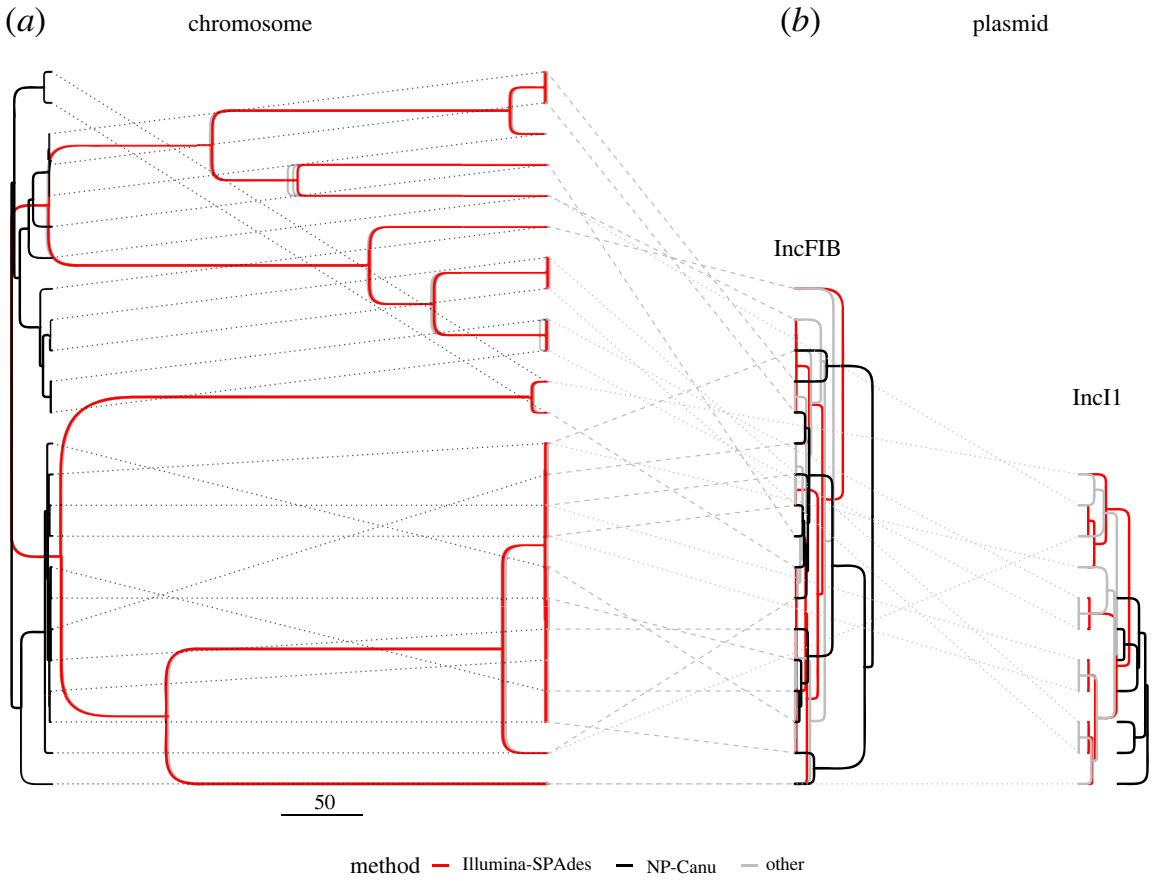

Figure 1. (*a*) Chromosomal, and (*b*) plasmid maximum clade credibility trees for 24 ESBL-positive *E. coli* isolates. The chromosomal tree encompasses all 24 samples, the plasmid trees only the subset that carried the plasmid (IncFIB: 17, IncI1: 11; with some variation across assembly method, see table 2). The colours indicate different assembly methods (see legend). The dashed and dotted lines indicate which tips are the same across the different trees. In panel (*a*), 7 phylogenies are included in the 'other' category (in grey), whereas panel (*b*) includes only the example of PB-Unicycler. Further methods are shown in electronic supplementary material, figures S5 and S6. The scale bar represents the time in years (given the assumed mutation rate of around 1.5 mutations per genome per year). (Online version in colour.)

supplementary material, figure S4). Wick *et al*. have shown that the Oxford Nanopore ligation protocol can miss small plasmids (less than 20 kb) [47], but this was not the case in our assemblies (table 2 and electronic supplementary material, figure S4). In the remainder, we focus our analysis on REPs found in at least 5 samples (Col MG828, Col156, ColRNAI, IncFIA, IncFIB AP001918, IncFIC FII, IncFII pRSB107 and IncI1) or a subset thereof (ColRNAI, IncFIA, IncFIB AP001918, IncI1) to illustrate the behaviour of representative small and large plasmid types.

For each plasmid, the number of samples it was found in (table 2 and electronic supplementary material, table S1), as well as the length of the resulting alignments (figure 2 and electronic supplementary material, figure S7) differed strongly across method combinations. When long read information was available, the methods that started from short-read assemblies (SPAdes–Hybrid and Unicycler–Hybrid) resulted in overall longer alignment lengths than those that started from long-reads during assembly graph generation (Canu and Canu–Hybrid). For the small Col plasmids (e.g. ColRNAI), NP–Canu and NP–Canu–Hybrid even showed alignments shorter than the shortest contig, which is probably because few coding domain sequences could be found (electronic supplementary material, figures S7 and S8).

For the IncF plasmids (IncFIA and IncFIB), Illumina–SPAdes showed both shorter plasmid contigs and a shorter multiple sequence alignment. Yet, for IncI1, the shorter plasmid contigs were not associated with a shorter overall alignment. This could be related to a greater plasticity of IncF plasmids in our sample, as opposed to the relatively conserved IncI1 plasmids. In general, the average assembled plasmid sequence length was not clearly associated with the total alignment length. For example, the 15 IncFIA plasmids in our sample had similar average assembled plasmid sequence lengths (except for Illumina–SPAdes), yet ranged from 44 to 273 kbp in alignment length across the assembly and alignment methods (figure 2).

(d) Plasmid trees differ across the assembly methods

The large diversity observed in the plasmid sequence alignments (figure 2) was clearly carried over to the plasmid phylogenies, both in terms of their topology (figure 1 and electronic supplementary material, figures S5, S6, S9, S11) and tree height (electronic supplementary material, figure S10). To achieve a conservative estimate of the plasmid tree (dis)similarity across alignment methods (independent of the ability to locate plasmids in the assembly), we subsetted the plasmid alignments to the tips present across all assembly methods prior to tree inference. For some plasmids (e.g. ColRNAI), many clade configurations were explored in the tree posterior, all with low clade probabilities, indicating large uncertainty in the phylogeny (electronic supplementary material, figures S9 and S11). The NP–Canu and NP–Canu–Hybrid

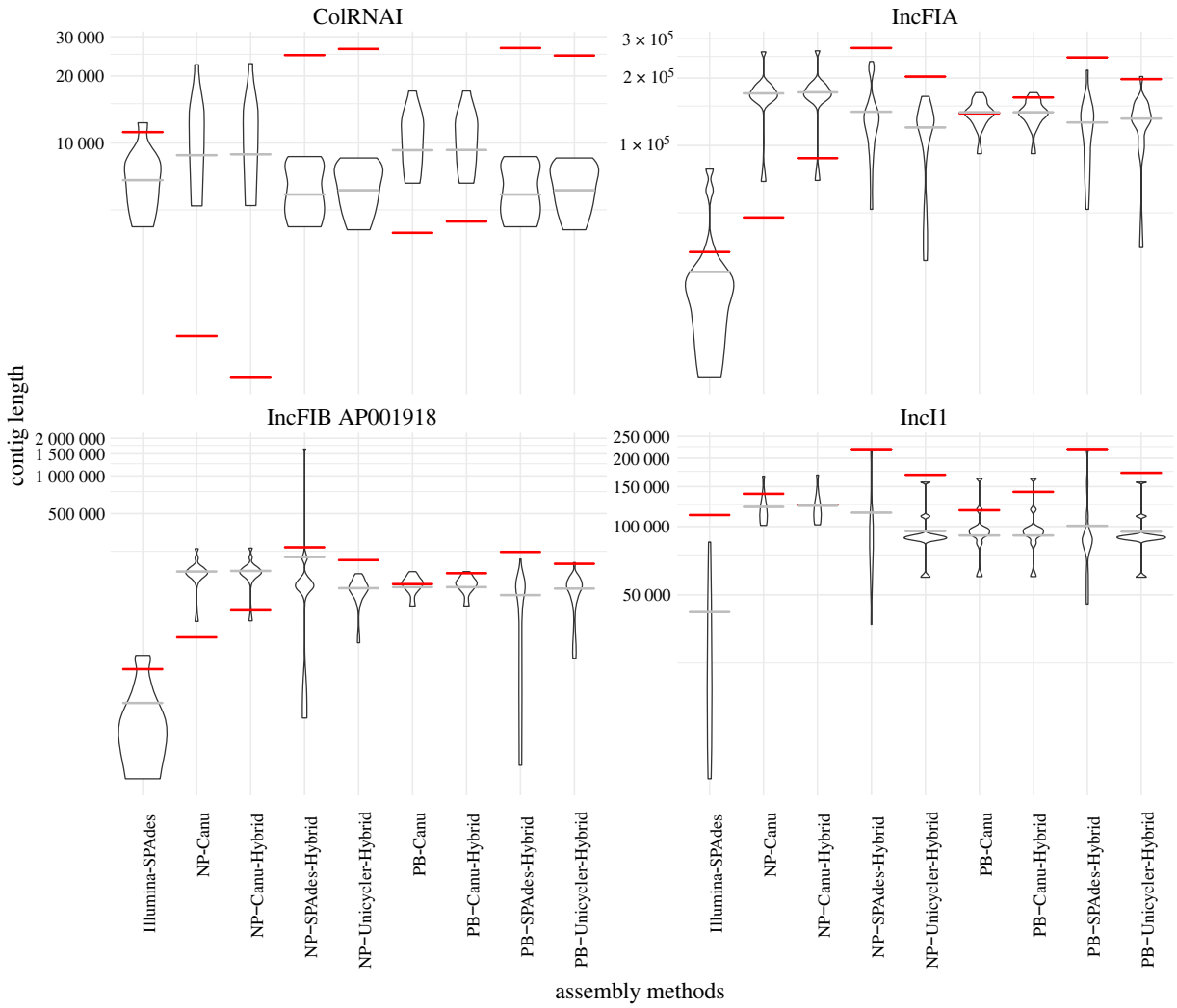

Figure 2. Plasmid contig length (violin plot, average indicated in grey) and length of the resulting alignment (red), for four selected plasmid replicons. ColRNAI was found in 8 out of 24 samples for all assembly methods, IncFIA in 15 (except NP-Canu and NP-Canu-Hybrid where it was found in 13), IncFIB AP001918 in 17 (except for NP-Canu and NP-Canu-Hybrid, where it was found in 15), and IncI1, which was found in 11 samples by Illumina-SPAdes, NP-SPAdes-Hybrid, PB-SPAdes-Hybrid and PB-Unicycler-Hybrid; in 7 by NP-Canu and NP-Canu-Hybrid; and in 10 by NP-SPAdes-Hybrid, PB-Canu and PB-Canu-Hybrid. (Online version in colour.)

methods resulted in different tree topologies from the other sequencing/assembly methods across all plasmids (electronic supplementary material, figure S9). When comparing the separation of tree topologies in tree space (using the tree space distance metric), all methods differed substantially, although also here Illumina–SPAdes, NP–Canu and NP–Canu–Hybrid resulted in trees that are the furthest removed from the other methods (but not more similar to each other; electronic supplementary material, figure S11).

In terms of tree height (electronic supplementary material, figure S10), Illumina–SPAdes, NP–Canu and NP–Canu–Hybrid showed overall lower plasmid tree heights than the other methods. All plasmid trees were substantially shorter (indicating more recent divergence) than the associated chromosomal trees (when subsetted to the plasmid-carrying taxa; compare figure 1). This is likely due to the shorter sequence length and the fixed mutation rate assumed in the phylogenetic analysis.

(e) The effect on inferred transmission patterns and rates

Incongruence between the chromosomal and plasmid phylogeny of a set of samples is an indication of possible horizontal gene transfer. This incongruence can be quantified using the normalized Robinson–Foulds distance between both types of trees. A large distance means that the compared trees differ in their topology (they describe a different set of clades), whereas a short distance indicates congruence. For each method combination, we compared the distance between all trees in the plasmid tree posterior and the single chromosomal maximum clade credibility (MCC) tree, where the latter was subsetted to the taxa that carry the plasmid (figure 3).

All plasmid trees exhibited incongruence with the chromosomal phylogeny of the plasmid-carrying taxa, but to varying degrees. For the IncF and Col replicons, Illumina–SPAdes showed the largest Robinson–Foulds distance, followed by the Canu-based alignments. Where the different methods show similar Robinson–Foulds distances (e.g. for IncI1), this can be seen as consistent signal for horizontal gene transfer of the plasmid. Yet, the larger plasmids (IncFIA, IncFIB, IncI1) all showed lower Robinson–Foulds distances to the chromosomal tree than a set of randomly generated trees with the same tips (electronic supplementary material, figure S12), indicating some amount of coevolution.

As a control we also compared the plasmid tree posteriors against a single chromosomal MCC tree (PB–Unicycler). This changed the results only slightly for the NP–Canu method (electronic supplementary material, figure S13), which indicates that the observed variation in Robinson–Foulds

Table 2. Number of samples with a given replicon, per assembly method.

name	Illumina–SPAdes	NP–Canu	NP–Canu–Hybrid	NP–SPAdes–Hybrid	NP–Unicycler–Hybrid	PB–Canu	PB–Canu–Hybrid	PB–SPAdes–Hybrid	PB–Unicycler–Hybrid
Col-MG828	9	9	9	9	9	5	5	9	9
Col-MGD2	—	2	2	—	—	—	—	—	
Col156	10	10	10	10	10	8	8	10	10
Col8282	3	3	3	3	3	—	—	3	3
ColRNAI	8	8	8	8	8	8	8	8	8
IncB	3	3	3	3	3	3	3	3	3
IncFIA	15	13	13	15	15	15	15	15	15
IncFIB AP001918	17	15	15	17	17	17	17	17	17
IncFIB-pKPHS1	3	—	—	3	3	3	3	3	3
IncFIC-FII	9	8	9	9	9	9	9	9	9
IncFII pRSB107	7	6	6	7	7	7	7	7	7
IncI1	11	7	7	11	10	10	10	11	11
IncN	—	2	2	—	—	—	—	—	—
IncR	—	2	2	—	—	—	—	—	—
IncX1	3	4	3	3	3	3	3	3	3
IncX3	3	3	3	3	3	3	3	3	3
IncX4	2	—	—	—	—	—	—	—	—
p0111	4	3	3	4	4	4	4	4	4

distances across the method combinations truly stems from differences in the plasmid tree topology (as observed in the previous section), rather than from the comparison to different chromosomal trees.

A different way to investigate horizontal gene transfer is to count the number of plasmid acquisitions or losses that are needed to explain the pattern of plasmid presence and absence at the tips of the chromosomal tree (i.e. to use the parsimony score). A low parsimony score indicates that the plasmid follows the chromosomal tree closely. To put the estimated parsimony score into perspective, we compared against a null model that assumes the amount of plasmid transfer is so high that for each terminal branch the chance of observing the plasmid is equal to the frequency with which the plasmid is observed in the entire population (i.e. no phylogenetic dependency). This was achieved by keeping the number of observed plasmid presences constant but randomizing the tips they were assigned to (1000 times). Biologically speaking, a deviation from this null model means plasmids are not continuously lost and picked up from a shared pool, independent of bacterial strain identity, but rather share some evolutionary history with their hosts.

For all plasmids, the observed parsimony scores were mid-range to low (4–7 gain or loss events on 24 tips), and for the majority this was below the mean of the parsimony distribution of the corresponding null model, indicating fewer observed host jumps than expected with free association (electronic supplementary material, figure S14). The conclusion is also in line with the results from the Robinson–Foulds comparison, which showed lower than random distances between the plasmid and chromosomal trees (electronic supplementary material, figure S12). In general, this signal may be influenced by imbalanced sampling, and in our dataset the large fraction of ST131 genomes likely dictates some of these findings. Eight plasmid REP types were present in more than 4 samples, and for five of these the majority of method combinations showed a statistically significant difference with respect to the null model (figure 4*a*; $\alpha = 0.05$). This was not corrected for multiple comparisons, since we wanted to illustrate how using one or the other assembly method would lead to differing conclusions of significance. Assembly with NP–Canu would have led to different conclusions from the majority of methods for 6 out of 8 plasmids, NP–Canu–Hybrid for 2 out of 8 and Illumina–SPAdes and NP–Unicycler–Hybrid for 1 out of 8 plasmids.

A similar parsimony analysis can be carried out for the presence/absence of antibiotic-resistance genes on the plasmid or chromosomal phylogeny. Again, most observed parsimony scores, both on the chromosomal and selected plasmid MCC trees, fell below the distribution expected under the null model of extensive horizontal gene transfer (HGT; electronic supplementary material, figure S15). Note that in contrast to the other figures we include IncFII pRSB107 here instead of IncFIB AP001918, since the latter showed very similar patterns to IncFIA.

For some genes ($bla_{\text{CTX-M-27}}$, $bla_{\text{CTX-M-1}}$, $bla_{\text{CTX-M-15}}$) and *sul1* the observed parsimony scores on the chromosomal tree were significantly lower than expected under the null model (figure 4*b*), suggesting a mostly clonal inheritance of these genes. The $bla_{\text{CTX-M-15}}$ gene also showed signs of being inherited with the IncFIA REP, whereas the observed parsimony score of the resistance genes on the IncI1 and IncFII pRSB107 trees did not significantly differ from the

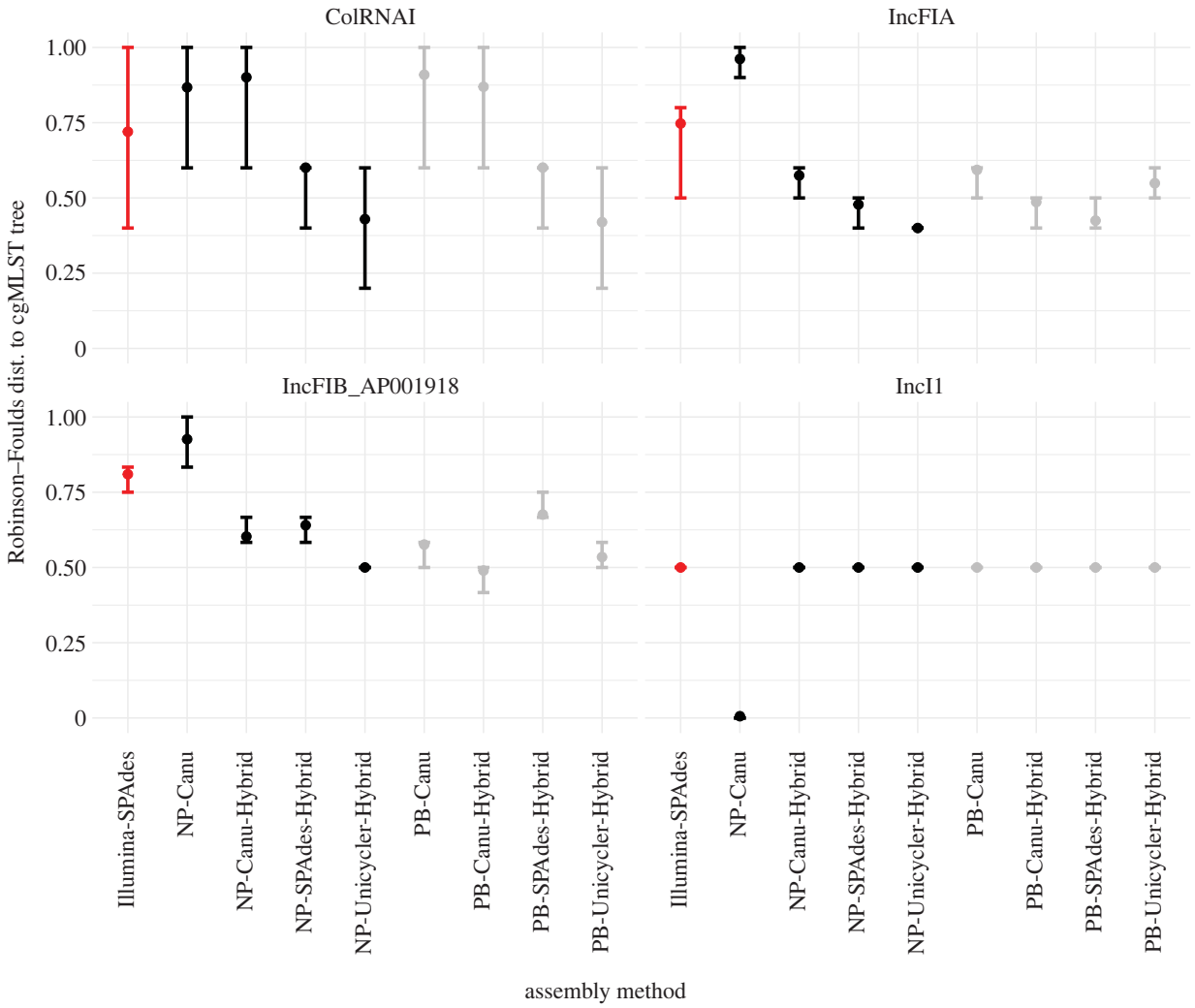

Figure 3. The normalized Robinson–Foulds distance between the chromosomal maximum clade credibility (MCC) tree and the posterior of plasmid trees, for all sequencing/assembly method combinations. Bars indicate the 95% highest posterior density interval (HPD) and are coloured by primary sequencing method (red for Illumina, black includes NP reads, grey includes PB reads). Large HPDs stem from highly divergent tree topologies in the plasmid tree posterior. Prior to tree inference, plasmid alignments were subsetted to the tips present across all assembly methods. Top left to bottom right, this resulted in trees containing 8 (ColRNAI), 13 (IncFIA), 15 (IncFIB AP001918) and 5 (IncI1) samples. (Online version in colour.)

null model. This is in agreement with the known association between $bla_{\text{CTX-M-15}}$ and certain IncF plasmids in ST131 [4,7]. The differences between plasmid trees obtained with different assembly methods translate into substantial differences in the parsimony score (electronic supplementary material, figure S16*b*). For a few gene/method combinations this also resulted in differing conclusions of significance. The gene $bla_{\text{CTX-M-1}}$ on the IncI1 plasmid tree is the most extreme example, where the assembly methods are split about the significance of the association. Notably the results obtained with Illumina–SPAdes were always the same as the majority of other methods, despite substantial differences in the plasmid tree topology.

4. Discussion

In this study, we investigated the impact of sequencing and assembly methods on the inference of chromosomal and plasmid phylogenies, as well as the downstream analysis of horizontal gene transfer. We showed that chromosomal trees can be constructed equally well from all sequencing and assembly combinations, excluding NP–Canu. Importantly, we showed that the high error rate of NP data when used alone impacts the estimated topology and height of the chromosomal tree, leading to erroneous trees. Plasmid phylogenies show much greater variability across assembly methods. Surprisingly, in this small dataset this variability had comparatively little impact on the inference of horizontal gene transfer of plasmids and antibiotic resistance genes.

In terms of assembly quality and chromosomal tree inference our results are in line with previous reports in the literature. Illumina sequencing is commonly used for bacterial phylogenetics in research and clinical diagnostics [5,9,48,49]. This is likely the most cost-effective choice when chromosomal trees are required. However, the accurate localization of resistance genes to plasmids or the chromosome requires the addition of long-read sequencing information. We confirm previous sequencing and assembly comparisons in Enterobacterales that showed Unicycler hybrid assembly of Illumina short reads with NP long-reads is well-suited for this purpose [12–14].

We showed that parsimony-based methods can be used to quantify horizontal gene transfer, quite independently of the sequencing and assembly method (except for NP–

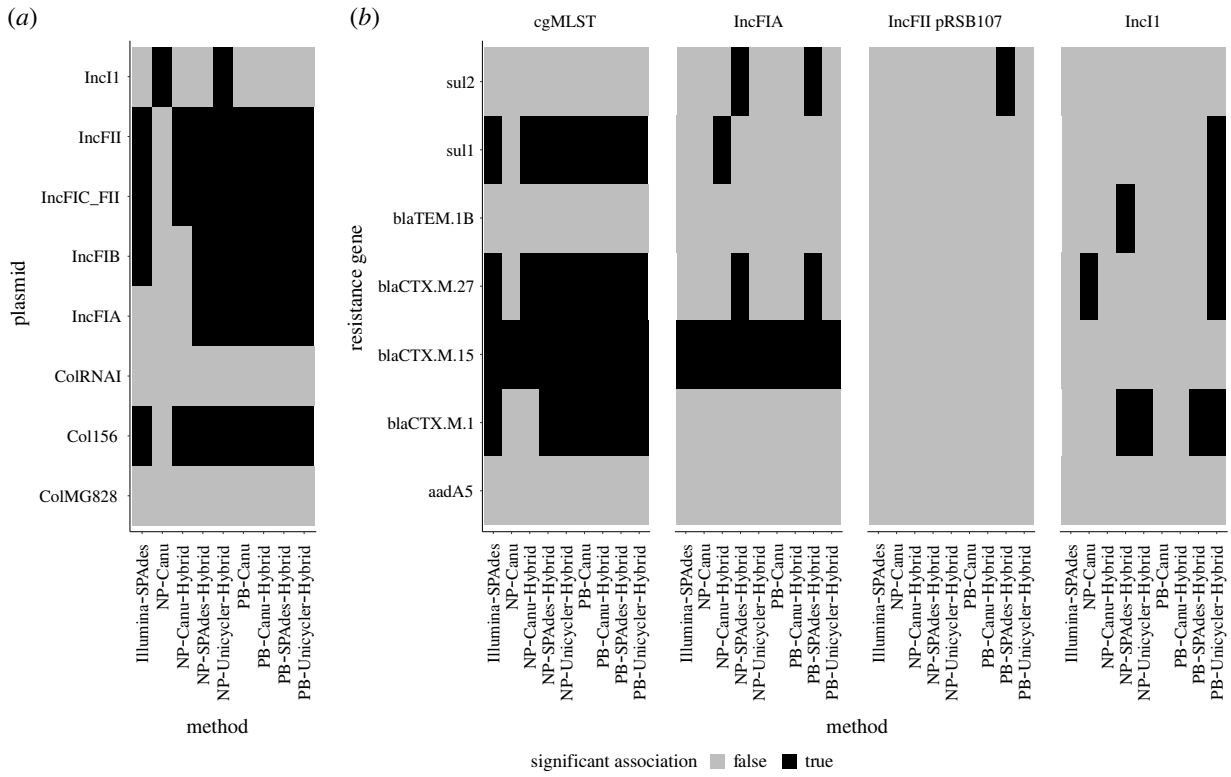

Figure 4. Significant deviations from the null model of gene presence/absence measured using the parsimony score. (*a*) Plasmid presence/absence on the chromosomal tree for different assembly methods. The number of samples with a replicon is given in table 2. (*b*) Resistance gene presence/absence on the chromosomal or selected plasmid trees (IncFIA, IncFII pRSB107, IncI1) inferred for each method.

Canu). However, researchers interested in plasmid evolution and phylogenetics are better off combining short- and long-read sequencing, and using hybrid methods that start from short-read assemblies (HybridSPAdes and Unicycler).

While comparing the phylogenetic trees resulting from different assembly methods, we did not explicitly consider the impact of recombination on our ability to reconstruct the phylogenetic relationship of our samples. We assume that the sequenced samples code for a single 'true' alignment, which should code for a single phylogeny (given a particular tree reconstruction method). This is independent of whether the inferred tree is also an unbiased account of the true phylogenetic relationship of these samples or rather summarizes average recombination rates between different subpopulations [50,51]. However, since recombination causes different parts of the genome to code for alternative genetic histories [52], it can increase the difference in trees resulting from alignments with different gene compositions. This likely contributed to the variability of the plasmid trees we found across different assembly methods.

This study has several limitations and ways in which it could be extended. First, we use a relatively small number of samples, taken from only one bacterial species and dominated by a single sequence type (ST131). Species differ greatly in both the length and number of repeats in the genome, the length and similarity of plasmids, and thus how difficult it is to resolve the full genome [29]. This could lead to an even greater spread of trees obtained by different assembly methods. In addition, the generality of biological conclusions we can draw is restricted by the limited number of samples. Using similar methods on a larger dataset would enable researchers to make broader statements about plasmid–chromosome co-evolution.

Second, our samples span much of the known phylogenetic diversity of ESBL-producing *E. coli*. This is a much greater diversity than would be expected over the course of a clinical outbreak. The inferred horizontal gene transfer thus likely occurred at evolutionary timescales, rather than in the context of the hospital in which the sequences were sampled. Nonetheless, this diversity will not be uncommon for samples collected through routine surveillance for a particular (antibiotic resistance) phenotype.

Third, it would exceed the scope of our study to include all of the (25 or more) different methods that have been developed to detect plasmids from WGS data [53]. Our method of BLASTing against the PlasmidFinder database to detect putative plasmid sequences will return a lower bound on the plasmid content for a specific strain. Arredondo-Alonso *et al.* [15] have shown that PlasmidFinder has perfect precision but less good recall, i.e. it does not manage to recover all plasmid information contained in the sample. In particular, it performs less well on assemblies with short contigs and for plasmids with unknown replicons. In our dataset, the long-read and hybrid assembly methods mostly recovered the same number of plasmids. Yet the variable length of the recovered sequences and corresponding alignment resulted in substantial differences between the inferred plasmid trees. Additional methods to detect plasmids would likely only increase the differences between plasmid phylogenies shown in this paper. However, researchers seeking to estimate diverse plasmid phylogenies may need to optimize this aspect of the pipeline.

Fourth, we have used only Roary to obtain plasmid alignments. One could envision alternatives that combine plasmid identification and alignment, such as plasmid multi-locus sequence typing (pMLST) or mapping reads to a plasmid

reference. The advantage of pMLST is that one could identify gene presence and absence directly from the raw reads, which removes the error-prone and time-intensive step of *de novo* assembly [54]. The drawback of pMLST is that typing schemes have been published only for few incompatibility groups, and yield alignments only a few genes in length. Mapping approaches can be quite powerful, but presuppose that closely related plasmids are available in public databases, or that the relevant plasmid reference sequences can be generated as part of the study (like in [6]). In addition, one must take care not to bias subsequent phylogenetic analyses by mapping diverged sequences to a single reference without taking into account non-SNP sites [55].

Lastly, one could construct trees using distance metrics that depart from the substitution models used in traditional phylogenetics. K-mer-based distances, such as Mash, are widely used for alignment-free sequence comparison and to approximate the average nucleotide identity between two samples [56]. This can provide a useful alternative to annotation-based alignments. However, the authors themselves stress that Mash is not intended for phylogenetic reconstruction, especially on highly divergent genomes or those with large differences in size [56]. Different distance metrics can also account for other forms of evolutionary signal, not captured in mutational differences between strains but rather in gene order. Synteny likely contains relevant information about the evolutionary history of a set of plasmids, especially when these are closely related and do not carry many mutational differences [22].

To conclude, we have started to analyse the effect of sequencing and assembly on downstream analyses. Such understanding is important to achieve standardization in diagnostics and comparability across studies, but also to inform studies that aim to combine genomes obtained from varying sequencing and assembly pipelines (e.g. as deposited in public databases).

Data accessibility. The code and data necessary to reproduce the figures are available from https://github.com/JSHuisman/plasmid_phylo. Assemblies and all three sequencing read sets have been deposited on NCBI, under Bioproject PRJNA554638.

The data are provided in electronic supplementary material [57].

Authors' contributions. J.S.H.: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; T.G.V.: methodology, supervision, writing—review and editing; A.E.: data curation, project administration, resources, writing—review and editing; S.T.-S.: conceptualization, data curation, funding acquisition, project administration, resources, writing—review and editing; T.S.: conceptualization, funding acquisition, methodology, resources, supervision, writing—review and editing; S.B.: conceptualization, funding acquisition, methodology, resources, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This research was supported by the Swiss National Science Foundation grant no 407240_167121 awarded to A.E., T.S. and S.B., and grant no. 407240_167060 awarded to S.T.-S. and T.S.

Acknowledgements. We thank Nicholas Noll, Richard Neher and the rest of the Neher group for their MinION sequencing efforts, as well as discussions concerning these data. We thank Rosa-Maria Vesco, Elisabeth Schultheiss, Magdalena Schneider, Christine Kissling, Clarisse Straub, and Dr Dominik Meinel from the University Hospital Basel for their technical assistance in Illumina sequencing. We further thank the Bonhoeffer and Stadler groups for helpful input and discussions, and Dr João Pires for critical reading of the manuscript.

References

1. Cassini A *et al.* The Burden of AMR Collaborative Group. 2019 Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect. Dis.* **19**, 56–66. (doi:10.1016/S1473-3099(18)30605-4)
2. Jit M, Lin Ng DH, Luangasanatip N, Sandmann F, Atkins KE, Robotham JV, Pouwels KB. 2020 Quantifying the economic cost of antibiotic resistance and the impact of related interventions: rapid methodological review, conceptual framework and recommendations for future studies. *BMC Med.* **18**, 1–14. (doi:10.1186/s12916-019-1443-1)
3. Ochman H, Lawrence JG, Groisman EA. 2000 Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304. (doi:10.1038/35012500)
4. Mathers AJ, Peirano G, Pitout JDD. 2015 The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant *Enterobacteriaceae*. *Clin. Microbiol. Rev.* **28**, 565–591. (doi:10.1128/CMR.00116-14)
5. Sheppard AE *et al.* 2016 Nested Russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene bla_{KPC}. *Antimicrob. Agents Chemother.* **60**, 3767–3778. (doi:10.1128/AAC.00464-16)
6. León-Sampedro R *et al.* 2021 Pervasive transmission of a carbapenem resistance plasmid in the gut microbiota of hospitalized patients. *Nat. Microbiol.* **6**, 606–616. (doi:10.1038/s41564-021-00879-y)
7. Stoesser N *et al.* 2016 Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *mBio* **7**, 1–15. (doi:10.1128/mBio.02162-15)
8. de Been M *et al.* 2014 Dissemination of cephalosporin resistance genes between *Escherichia coli* strains from farm animals and humans by specific plasmid lineages. *PLoS Genet.* **10**, e1004776. (doi:10.1371/journal.pgen.1004776)
9. Balloux F, Brynildsrud OB, van Dorp L, Shaw LP, Chen H, Harris KA, Wang H, Eldholm V. 2018 From theory to practice: translating whole-genome sequencing (WGS) into the clinic. *Trends Microbiol.* **26**, 1035–1048. (doi:10.1016/j.tim.2018.08.004)
10. Stadler T *et al.* 2018 Transmission of ESBL-producing *Enterobacteriaceae* and their mobile genetic elements—identification of sources by whole genome sequencing: study protocol for an observational study in Switzerland. *BMJ Open* **8**, e021823. (doi:10.1136/bmjopen-2018-021823)
11. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. 2018 Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene based approaches. *Clin. Microbiol. Infect.* **24**, 350–354. (doi:10.1016/j.cmi.2017.12.016)
12. De Maio N *et al.* 2019 Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genomics* **5**, e000294. (doi:10.1099/mgen.0.000294)
13. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017 Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genomics* **3**, e000132. (doi:10.1101/160614)
14. George S *et al.* 2017 Resolving plasmid structures in enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and MinION/illumina hybrid data assembly approaches. *Microb. Genomics* **3**, 1–8. (doi:10.1099/mgen.0.000118)
15. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. 2017 On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb. Genomics* **3**, e000128. (doi:10.1099/mgen.0.000128)
16. Lanza VF, de Toro M, Garcillán-Barcia MP, Mora A, Blanco J, Coque TM, de la Cruz F. 2014 Plasmid flux in *E. coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for

plasmid reconstruction from whole genome sequences. *PLoS Genet.* **10**, e1004766. (doi:10.1371/journal.pgen.1004766)

17. Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014 *In silico* detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903. (doi:10.1128/AAC.02412-14)
18. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. 2016 PlasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* **32**, 3380–3387. (doi:10.1093/bioinformatics/btv688)
19. Tschudin-Sutter S, Frei R, Schwahn F, Tomic M, Conzelmann M, Stranden A, Widmer AF. 2016 Prospective validation of cessation of contact precautions for extended-spectrum β-lactamase-producing *Escherichia coli*. *Emerg. Infect. Dis.* **22**, 1094–1097. (doi:10.3201/eid2206.150554)
20. Benz F *et al.* 2021 Plasmid- and strain-specific factors drive variation in ESBL-plasmid spread in vitro and in vivo. *ISME J.* **15**, 862–878. (doi:10.1038/s41396-020-00819-4)
21. Ezclermont. https://github.com/nickp60/EzClermont (accessed 10 March 2022).
22. Noll N, Urich E, Wüthrich D, Hinic V, Egli A, Neher RA. 2018 Neher: resolving structural diversity of Carbapenemase-producing gram-negative bacteria using single molecule sequencing. *BioRxiv*. (doi:10.1101/456897)
23. Bolger AM, Lohse M, Usadel B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. (doi:10.1093/bioinformatics/btu170)
24. Andrews S. 2010 FastQC: a quality control tool for high throughput sequence data.
25. Nurk S *et al.* 2013 Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **20**, 714–737. (doi:10.1089/cmb.2013.0084)
26. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017 Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736. (doi:10.1101/gr.215087.116)
27. Walker BJ *et al.* 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963. (doi:10.1371/journal.pone.0112963)
28. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. 2015 HybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015. (doi:10.1093/bioinformatics/btv688)
29. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017 Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595. (doi:10.1371/journal.pcbi.1005595)
30. Li H. 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993. (doi:10.1093/bioinformatics/btr509)
31. Seemann T. 2014 Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069. (doi:10.1093/bioinformatics/btu153)
32. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010 Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.* **11**, 119. (doi:10.1186/1471-2105-11-119)
33. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012 Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644. (doi:10.1093/jac/dks261)
34. Maiden M, *et al.* 1998 Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA* **95**, 3140–3145. (doi:10.1073/pnas.95.6.3140)
35. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Andre Carrico J. 2018 chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. *Microb. Genomics* **4**, e000166. (doi:10.1099/mgen.0.000166)
36. Enterobase. See http://enterobase.warwick.ac.uk (accessed 23 March 2018).
37. Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. (doi:10.1093/molbev/mst010)
38. Page AJ *et al.* 2015 Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693. (doi:10.1093/bioinformatics/btv421)
39. Bouckaert R, Heled J, Kühnert D, Vaughan T, Hsi Wu C, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014 BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, 1–6. (doi:10.1371/journal.pcbi.1003537)
40. Wielgoss S, Barrick JE, Tenaillon O, Cruveiller S, Chane-Woon-Ming B, Médigue C, Lenski RE, Schneider D. 2011 Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3: Genes Genomes Genet.* **1**, 183–186. (doi:10.1534/g3.111.000406)
41. Gibson B, Wilson DJ, Feil E, Eyre-Walker A. 2018 The distribution of bacterial doubling times in the wild. *Proc. R. Soc. B* **285**, 20180789. (doi:10.1098/rspb.2018.0789)
42. Jombart T, Kendall M, Almagro-Garcia J, Colijn C. 2017 Treespace: statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Resour.* **17**, 1385–1392. (doi:10.1111/1755-0998.12676)
43. Robinson DF, Foulds LR. 1981 Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)
44. Schliep KP. 2011 Phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593. (doi:10.1093/bioinformatics/btq706)
45. Paradis E, Schliep K. 2019 Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528. (doi:10.1093/bioinformatics/bty633)
46. Villa L, García-Fernández A, Fortini D, Carattoli A. 2010 Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. *J. Antimicrob. Chemother.* **65**, 2518–2529. (doi:10.1093/jac/dkq347)
47. Wick RR, Judd LM, Wyres KL, Holt KE. 2021 Recovery of small plasmid sequences via Oxford Nanopore sequencing. *Microb. Genomics* **7**, 000631. (doi:10.1101/2021.02.21.432182)
48. Holt KE *et al.* 2015 Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc. Natl Acad. Sci. USA* **112**, E3574–E3581. (doi:10.1073/pnas.1501049112)
49. Cuénod A *et al.* 2021 Whole-genome sequence-informed MALDI-TOF MS diagnostics reveal importance of *Klebsiella oxytoca* group in invasive infections: a retrospective clinical study. *Genome Med.* **13**, 150. (doi:10.1186/s13073-021-00960-5)
50. Didelot X, Falush D. 2007 Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266. (doi:10.1534/genetics.106.063305)
51. Sakoparnig T, Field C, van Nimwegen E. 2021 Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *eLife* **10**, e65366. (doi:10.7554/eLife.65366)
52. Sen D, Brown CJ, Top EM, Sullivan J. 2013 Inferring the evolutionary history of IncP-1 plasmids despite incongruence among backbone gene trees. *Mol. Biol. Evol.* **30**, 154–166. (doi:10.1093/molbev/mss210)
53. Paganini JA, Plantinga NL, Arredondo-Alonso S, Willems RJL, Schürch AC. 2021 Recovering *Escherichia coli* plasmids in the absence of long-read sequencing data. *Microorganisms* **9**, 1613. (doi:10.3390/microorganisms9081613)
54. Inouye M, Dashnow H, Raven L-a, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. 2014 SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* **6**, 1–16. (doi:10.1186/s13073-014-0090-6)
55. Bertels F, Silander OK, Pachkov M, Rainey PB, Van Nimwegen E. 2014 Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Mol. Biol. Evol.* **31**, 1077–1088. (doi:10.1093/molbev/msu088)
56. Ondov BD, Treangen TJ, Melsted Páll, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016 Mash: fast genome and metagenome distance estimation using MinHashs. *Genome Biol.* **17**, 132. (doi:10.1186/s13059-016-0997-x)
57. Huisman JS, Vaughan TG, Egli A, Tschudin-Sutter S, Stadler T, Bonhoeffer S. 2022 The effect of sequencing and assembly on the inference of horizontal gene transfer on chromosomal and plasmid phylogenies. Figshare. (doi:10.6084/m9.figshare.c.6080814)