



Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids



Alex Orlek^{a,b,*}, Hang Phan^{a,b}, Anna E. Sheppard^{a,b}, Michel Doumith^c, Matthew Ellington^{b,c}, Tim Peto^{a,b}, Derrick Crook^{a,b}, A. Sarah Walker^{a,b}, Neil Woodford^{b,c,1}, Muna F. Anjum^{b,d,1}, Nicole Stoesser^{a,1}

^a Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK

^b NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford, Oxford, UK

^c Antimicrobial Resistance and Healthcare Associated Infections (AMRHA) Reference Unit, National Infection Service, Public Health England, London, UK

^d Department of Bacteriology, Animal and Plant Health Agency, Addlestone, UK

ARTICLE INFO

Article history:

Received 16 December 2016

Accepted 8 March 2017

Available online 9 March 2017

Keywords:

Replicon typing

Plasmid multilocus sequence typing

MOB typing

Antibiotic resistance

Plasmid database

ABSTRACT

Plasmid typing can provide insights into the epidemiology and transmission of plasmid-mediated antibiotic resistance. The principal plasmid typing schemes are replicon typing and MOB typing, which utilize variation in replication loci and relaxase proteins respectively. Previous studies investigating the proportion of plasmids assigned a type by these schemes ('typeability') have yielded conflicting results; moreover, thousands of plasmid sequences have been added to NCBI in recent years, without consistent annotation to indicate which sequences represent complete plasmids. Here, a curated dataset of complete Enterobacteriaceae plasmids from NCBI was compiled, and used to assess the typeability and concordance of *in silico* replicon and MOB typing schemes. Concordance was assessed at hierarchical replicon type resolutions, from replicon family-level to plasmid multilocus sequence type (pMLST)-level, where available. We found that 85% and 65% of the curated plasmids could be replicon and MOB typed, respectively. Overall, plasmid size and the number of resistance genes were significant independent predictors of replicon and MOB typing success. We found some degree of non-concordance between replicon families and MOB types, which was only partly resolved when partitioning plasmids into finer-resolution groups (replicon and pMLST types). In some cases, non-concordance was attributed to ambiguous boundaries between MOB_P and MOB_Q types; in other cases, backbone mosaicism was considered a more plausible explanation. β -lactamase resistance genes tended not to show fidelity to a particular plasmid type, though some previously reported associations were supported. Overall, replicon and MOB typing schemes are likely to continue playing an important role in plasmid analysis, but their performance is constrained by the diverse and dynamic nature of plasmid genomes.

© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Plasmid genomes generally consist of a somewhat conserved 'backbone' of genes associated with functions such as replication and transfer, accompanied by variable sets of 'accessory genes'. Backbone loci have been exploited as phylogenetic markers to group plasmids (Garcillán-Barcia and de la Cruz, 2013). Accessory genes often confer adaptive traits, notably, antibiotic resistance (Partridge, 2011). Plasmid-mediated resistance dissemination is common amongst the Enterobacteriaceae family of gram-negative bacteria, which includes clinically important taxa such as *Escherichia coli* and *Klebsiella* spp.

(Iredell et al., 2016). Of particular concern is the rise in resistance to β -lactam antibiotics, frequently driven by plasmid-borne genes including extended-spectrum β -lactamase (ESBL) genes (e.g. *bla*_{CTX-M}), as well as carbapenemase genes (e.g. *bla*_{KPC}) (Livermore and Woodford, 2006; Nordmann et al., 2012). Transmission of resistance gene-carrying plasmids ('resistance plasmids') can drive the success of recipient strains (Holt et al., 2013), so it is important to understand resistance plasmid epidemiology, as well as strain epidemiology.

Plasmid typing can provide insights into resistance plasmid epidemiology (de Been et al., 2014; Leverstein-van Hall et al., 2011; Orlek et al., 2017; Pecora et al., 2015), such as whether resistance dissemination involves diverse plasmids or one dominant 'epidemic' type (Valverde et al., 2009). Some resistance genes have been associated predominantly with specific replicon types (Akiba et al., 2016; Carattoli, 2009, 2013), for example, *bla*_{OXA-48} with IncL/M (Poirel et al., 2012). However, the extent to which reported associations reflect recent local expansion of a resistance plasmid, or stable widespread associations remains unclear.

* Corresponding author at: Level 7, Microbiology, Nuffield Department of Clinical Medicine, University of Oxford, John Radcliffe Hospital, Headley Way, Oxford OX3 9DU, UK.

E-mail address: alex.orkel@wolfson.ox.ac.uk (A. Orlek).

¹ These authors contributed equally to this work.

If the latter, then plasmids stably harbouring resistance genes could potentially be targeted using sequence-specific antimicrobials (Bikard et al., 2014; Williams and Hergenrother, 2008). Other applications of plasmid typing include plasmid detection; notably, distinguishing plasmid from chromosomal contigs in short-read *de novo* assemblies (Lanza et al., 2014).

The most widely used plasmid classification schemes are replicon and MOB typing, which exploit loci encoding plasmid replication (replicons) and mobility functions (relaxases), respectively (Alvarado et al., 2012; Carattoli et al., 2005, 2014; Garcillán-Barcia et al., 2009). Replicons include various different loci, none of which are universal across plasmids (del Solar et al., 1998), whereas relaxases are thought to be universally present amongst plasmids that mobilise via the relaxase-*in-cis* mechanism (Garcillán-Barcia et al., 2011; Ramsay et al., 2016). However, relaxase homology can be distant, even amongst plasmids of the same MOB type (Garcillán-Barcia et al., 2009). Replicon types can be assigned by querying plasmids against various replicon sequences using BLASTN (Carattoli et al., 2014), whilst MOB types can be assigned using profile-based searches such as PSI-BLAST; *in silico* probes representing each relaxase family are used to query a dataset of plasmids, and thereby assign MOB types (Francia et al., 2004; Garcillán-Barcia et al., 2009; Guglielmini et al., 2011). Six probes have been used to detect relaxases of Gammaproteobacterial plasmids, whilst a further two probes have been used to detect relaxases in other taxa (Guglielmini et al., 2011). PSI-BLAST can uncover distant homology, so MOB typing provides a lower resolution but potentially more inclusive classification (Altschul et al., 1997; Garcillán-Barcia and de la Cruz, 2013).

Within the replicon typing framework, plasmids can be classified at hierarchical resolutions: for common replicon types, plasmid multi-locus sequence typing (pMLST) schemes have been devised for sub-typing (Brolund and Sandegren, 2015; Hancock et al., 2016), whilst replicon types also belong to broader replicon families. Most replicon families were originally defined according to plasmid incompatibility, which is a manifestation of the genetic relatedness of replicons (Taylor et al., 2004). *In silico* analysis of replication initiation protein genes has indicated that based on sequence similarity, the traditional incompatibility families represent phylogenetically-coherent groups (Carattoli et al., 2014). However, the Col 'family' of plasmids was defined according to the more superficial phenotype of colicin production, and it has long been known that this so-called 'family' includes phylogenetically diverse plasmids (Riley and Gordon, 1992).

A key limitation of replicon typing is that individual plasmids can contain multiple replicons, complicating classification, whereas usually just one relaxase is encoded (Garcillán-Barcia and de la Cruz, 2013). However, due to its higher resolution, replicon typing provides more detailed information on plasmid relatedness, particularly if a pMLST scheme is available (Pallecchi et al., 2007). MOB typing only types relaxase-encoding plasmids, which constitute varying proportions of total plasmids across different taxa (Guglielmini et al., 2011; Smillie et al., 2010); replicon typing is most developed for Enterobacteriaceae-associated plasmids. The proportion of plasmids that can be assigned a type using the schemes ('typeability') is likely to be influenced by biases in the sequencing datasets on which typing is conducted, as well as the bioinformatic approaches used. In 2014, *in silico* replicon typing was reported to type all publicly available clinically-relevant Enterobacteriaceae plasmids (Carattoli et al., 2014). Smillie et al. (2010) reported that around half of Gammaproteobacterial plasmids could be assigned a MOB type. In contrast, Shintani et al. (2015) found only 75% of publicly available Enterobacteriaceae plasmids could be replicon typed, and 44% of Gammaproteobacterial plasmids could be MOB typed (see Table S1 in Shintani et al., 2015); however, bioinformatic methods differed from those originally proposed.

Making accurate epidemiological inferences using plasmid typing depends on the ability of a typing scheme to assign phylogenetically-coherent types to plasmids (Belkum et al., 2007). For replicon and MOB typing schemes to reflect phylogenetic relationships accurately, they

should be concordant - that is, replicon families should nest within the broader-resolution MOB type families. Previous studies have supported concordance, with several exceptions (see Fig. 5 in Garcillán-Barcia et al., 2011). Non-concordance could result from backbone mosaicism (Hiraga et al., 1994; Kulinska et al., 2008), frequently due to recombination (Bianco and Kowalczykowski, 1997; Osborn et al., 2000). Alternatively, non-concordance may reflect fuzzy boundaries between the MOB families. Indeed, there is known to be some level of overlap between the MOBP and MOBQ families such that PSI-BLAST searches with MOBP and MOBQ probes sometimes uncover the same homologs (Garcillán-Barcia et al., 2009). However, non-concordance has not been the focus of previous studies so it remains unclear which explanation is most likely.

Re-investigating the performance of replicon and MOB typing is timely given the increasing numbers of publicly available complete plasmid sequences (Conlan et al., 2014). In this study, we curated a dataset of all complete publicly available Enterobacteriaceae plasmids to allow an up-to-date, comprehensive assessment of replicon and MOB typing schemes in terms of typeability and concordance. Concordance was assessed at hierarchical resolutions from replicon family-level to pMLST-level where available. Associations between plasmid replicon types and resistance genes were also examined.

2. Methods

2.1. Retrieving complete plasmid sequences and associated metadata from NCBI

Putative complete plasmid accessions were retrieved from the NCBI nucleotide database (<https://www.ncbi.nlm.nih.gov/nucleotide/>) on 26th August 2016, using an Entrez query with filters to exclude some incomplete or non-plasmid accessions at this stage (Supplementary methods S1). Duplicate sequences (those sharing 100% sequence identity with another retrieved sequence) were removed, preferentially retaining RefSeq over GenBank accessions (Pruitt et al., 2007), and more richly annotated over less richly annotated Genbank accessions. Biopython scripts (Cock et al., 2009) were used to retrieve information and filter accessions, including filtering-out non-coding sequences, and eliminating incomplete plasmid sequences using a regular expression search of accession title descriptions (Supplementary methods S1). Multi-locus sequence typing (MLST) using all available Enterobacteriaceae schemes (<http://pubmlst.org/data/>) was also conducted to identify and remove chromosomal sequences mis-annotated as plasmids, using BLAST as described (Larsen et al., 2012). Additional filtering involved manually examining putative plasmids at the tails of the sequence length distribution, to remove accessions thought to represent chromosomal sequences or partial plasmid sequences (Supplementary methods S1).

EDirect (Kans, 2013) was used to retrieve accession metadata, including the 'completeness' annotation which was used to guide filtering. EDirect was also used to retrieve additional metadata not used for filtering, including the sequencing technology, and the 'create date' of the accession in NCBI (Nahin, 2008). *In silico* replicon typing using the PlasmidFinder database has been designed and assessed according to a selection of clinically-relevant plasmids (Carattoli et al., 2014). Therefore, to make direct comparisons with analyses of original authors, plasmids in the quality-filtered dataset were assigned as clinical or non-clinical according to the source taxa. Specifically, genus and, where necessary, species-level information on host organism and human pathogenicity derived from the PATRIC database was used to guide assignments (Wattam et al., 2014).

2.2. Replicon typing and pMLST, MOB typing, resistance gene detection

For replicon typing, the complete plasmid sequences were queried against a locally downloaded version (retrieved 20th April 2016) of

the PlasmidFinder database (<http://www.genomicepidemiology.org/>), using recommended percentage identity and coverage thresholds of 80% and 60% respectively (Carattoli et al., 2014). Where multiple hits aligned to the same locus (defined by an overlap spanning 50% of the length of the shorter sequence), best hits were selected according to percentage identity and coverage. *In silico* pMLST was conducted for IncF, IncN, IncA/C, IncHI1, IncHI2 and IncI1 plasmids (García-Fernández et al., 2008, 2011; García-Fernández and Carattoli, 2010; Hancock et al., 2016; Phan et al., 2009; Villa et al., 2010). pMLST was conducted using locally downloaded allele databases (<http://pubmlst.org/plasmid/>) with recommended identity and coverage thresholds of 85% and 66% (Carattoli et al., 2014). For each allelic locus, the best hit was again selected according to percentage identity and coverage.

To perform MOB typing, PSI-BLAST searches (Altschul et al., 1997) were conducted using N-terminal relaxase protein sequences as queries against a database of the complete plasmid sequences, translated in all six frames. Searches were run for ≤ 14 iterations - the maximum number of iterations used by previous authors (Garcillán-Barcia et al., 2009). Initially, we used E-value thresholds in accordance with those used previously. However, we found that certain plasmids previously assigned a MOB type, were left unassigned (Supplementary methods S1, S2). As the calculation of E-values accounts for database size, the E-value associated with a given hit is inflated when BLAST searching against a larger database (Jones and Swindells, 2002). We hypothesised that this could account for the discrepancy in MOB typing. Hence, to optimise the E-value threshold, we used a set of plasmids for which MOB typing had previously been conducted and validated (Supplementary methods S1, S2). Based on this, we chose E-value thresholds as follows: MOBC, 0.001; MOBF, 0.01; MOBH, 0.01; MOBP, 1; MOBQ, 0.0001; MOBV, 0.01. After hits were produced by PSI-BLAST, no identity or coverage thresholds were applied (allowing for distant homologies to be detected); however, stringent filtering was used to select best hits (Supplementary methods S1). To investigate the robustness of our MOB typing results, we re-ran PSI-BLAST searches using a different set of six MOB query proteins representing each MOB family (Supplementary methods S1, S2).

Resistance genes were detected by BLAST querying the ResFinder database, using recommended identity and coverage thresholds of 98% and 60% respectively (Zankari et al., 2012). Best hits were selected as described above for replicon typing.

2.3. Visualisation and statistical analysis

All analyses were performed using R (<https://www.r-project.org/>). The circlize package (Gu et al., 2014) was used to create chord diagrams to visualise associations between replicon families/types/sub-types and MOB types using an edgelist as input. A given replicon type and MOB type detected on the same plasmid were represented as a single edge. An equivalent approach was used to visualise associations between replicon families/replicon subtypes/MOB types and a selection of key β -lactamase genes prevalent in our dataset.

In chord diagram visualisations of plasmid typing data, if multiple replicons of the same family/type were detected on a given plasmid, the plasmid type was considered to be the set of unique families/types detected. That is, a plasmid typed as IncFIB, IncFIC, IncFIC would be represented as IncFIB, IncFIC at the replicon type level, and as IncF at the family level. Likewise, for MOB typing, a MOBF, MOBF plasmid would be represented simply as MOBF in visualisations.

Predictors of plasmid typing success were investigated visually and using statistical analysis. Binary logistic regression was used to investigate whether plasmid size (log₁₀-transformed) and the number of detected resistance genes (all resistance genes included; values truncated at the 95th percentile to reduce outlier influence) independently predicted plasmid typeability by each scheme. Associations between replicon families/MOB types and prevalent β -lactamase genes were

investigated using Fisher's exact test, with p-values computed by Monte Carlo simulation.

3. Results and discussion

3.1. Characteristics of the curated plasmid dataset

A total of 6952 sequences representing putative Enterobacteriaceae plasmids were retrieved from the NCBI nucleotide database. Deduplicating identical sequences (2815 removed) and programmatically filtering probable non-plasmid/partial plasmid sequences (1892 removed) left 2245 accessions. Of these, 2063 met inclusion criteria (Supplementary methods S1), whilst the remaining 182 accessions were further examined to decide upon inclusion or exclusion; 148 were excluded according to Methods described above (104 not annotated as 'complete', 32 found to contain MLST loci, 12 manually filtered) leaving 2097 complete Enterobacteriaceae plasmids for analysis (Supplementary Table S1). The plasmid sequence files have been made publicly available, and could be utilized in various areas of plasmid research (Orlek et al., *in press*).

Plasmids ranged in size from 1.3 kb to 794 kb. The plasmid size distribution was log-bimodal (Supplementary Fig. S1) as reported previously (Shintani et al., 2015). The source organisms were dominated by human pathogens, especially *E. coli*, *K. pneumoniae*, and *S. enterica* (Supplementary Fig. S2). PacBio sequencing technology has clearly been a key driver of the increase in complete plasmid sequences since 2014 (Fig. 1; data prior to 2014 not shown due to a lack of annotation for sequencing technology). The first complete plasmids sequenced using Oxford Nanopore technology were added in June 2016 (Both et al., 2016) and this technology will likely drive further expansion in availability of complete plasmid sequences in future.

3.2. Assessment of database curation methods

Sequence topology annotation (circular/linear) was an unreliable indicator of complete plasmid sequences, with 57 of the 2097 curated plasmids annotated as 'linear'. At least one of these accessions represents a genuine linear plasmid (accession NC_011422) (Baker et al., 2007) but remaining accessions may represent mis-annotations, since 'linear' is set as the default value on submission (Fetchko and Kitts, 2011). Not counting accessions excluded as duplicates, 190 excluded accessions were annotated as 'circular'. Of these, 27 were excluded due to MLST allele detection (indicating an accession was likely chromosomal), and four were excluded following manual inspection. The majority of the other excluded 'circular' accessions were filtered due to being described as 'whole genome shotgun' sequences which should only be used to describe incomplete genomes (NCBI, 2014). Additional manual review of these accessions might have identified some that would warrant inclusion as complete plasmids.

3.3. Assessment of typing performance: typeability

Over 1000 complete Enterobacteriaceae plasmids have been added to NCBI in the past two years (Fig. 2A), underscoring the need to reassess the performance of plasmid typing schemes. Overall, a replicon type was detected in 1784 (85%) plasmids whilst a MOB type could be assigned to 1371 (65%) plasmids; together the schemes typed 1872 (89%) plasmids (Supplementary Table S1). When typeability was assessed over time (based on the date on which an accession became available), the proportion of plasmids replicon typed remained relatively constant, whilst the proportion of plasmids MOB typed tended to increase (Fig. 2B). This increase may reflect a bias in the size of available plasmid accessions over time (Supplementary Fig. S3); specifically, larger plasmids, which became increasingly available later on, were also more likely to be MOB typed (see Section 3.4). Previous authors reported in 2010 that around half of publicly available plasmids from

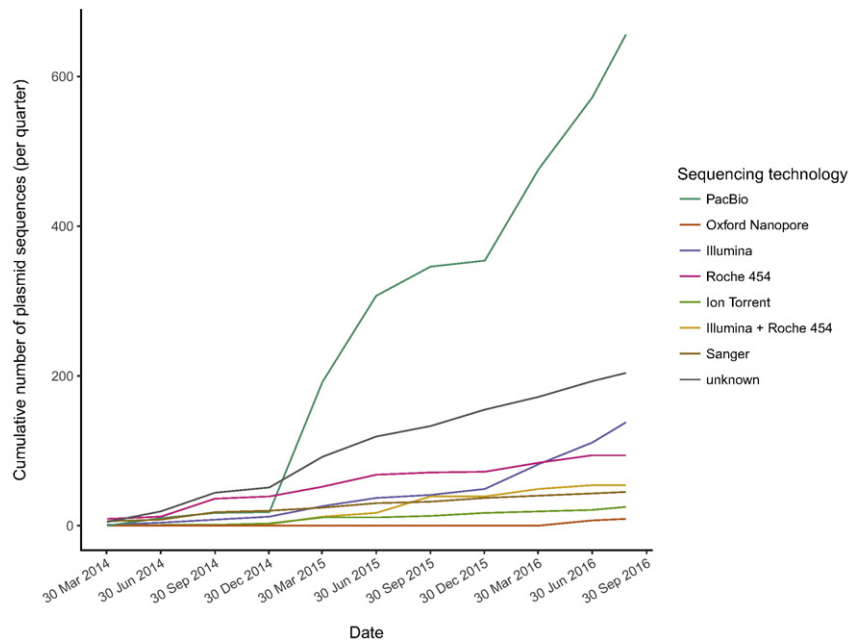


Fig. 1. Complete plasmid accessions added to NCBI since 2014. Where a hybrid short-read/long-read sequencing approach was used (as indicated by the accession metadata), this is represented as long-read only.

Gammaproteobacteria - the taxonomic class to which the Enterobacteriaceae family belongs - could be MOB typed (Smillie et al., 2010). In comparison, we were able to assign MOB types to a greater proportion of plasmids. However, the increase in proportion of MOB-typeable plasmids over time would seem to explain this discrepancy (Fig. 2B); of the plasmids in our dataset that were available in 2010, roughly half could be assigned a MOB type, in agreement with previous authors. Our finding that only 85% of plasmids could be replicon typed contrasts with results of Carattoli et al., who demonstrated 100% typing success. We wanted to determine the extent to which this discrepancy may reflect differences in the taxonomic scope of our analysis, relative to that of Carattoli et al., who used a selection of clinically-relevant Enterobacteriaceae plasmids to design and assess *in silico* replicon typing. When plasmids from non-clinical taxa were excluded from our analyses (in line

with Carattoli et al.), 1675/1818 (92%) plasmids were assigned replicon types (Table S1). The remaining disparity largely reflects incomplete classification of plasmid sequences added to NCBI in the two years since *in silico* replicon typing was devised (Supplementary Fig. S4A). Thirty-eight clinical plasmids lacking a replicon type were submitted prior to 2014, but were not included in the dataset used by Carattoli et al., perhaps due to temporary suppression of these records in NCBI (Pruitt et al., 2010).

One advantage of MOB typing lies in its ability to detect divergent plasmid backbones not detected by the replicon typing scheme (Garcillán-Barcia and de la Cruz, 2013). Supporting this, 88 plasmids were MOB typed, but not replicon typed. Interestingly, of these 88 plasmids, non-clinical plasmids were over-represented (47% versus 13% of plasmids in the whole dataset), which may reflect the fact that *in silico*

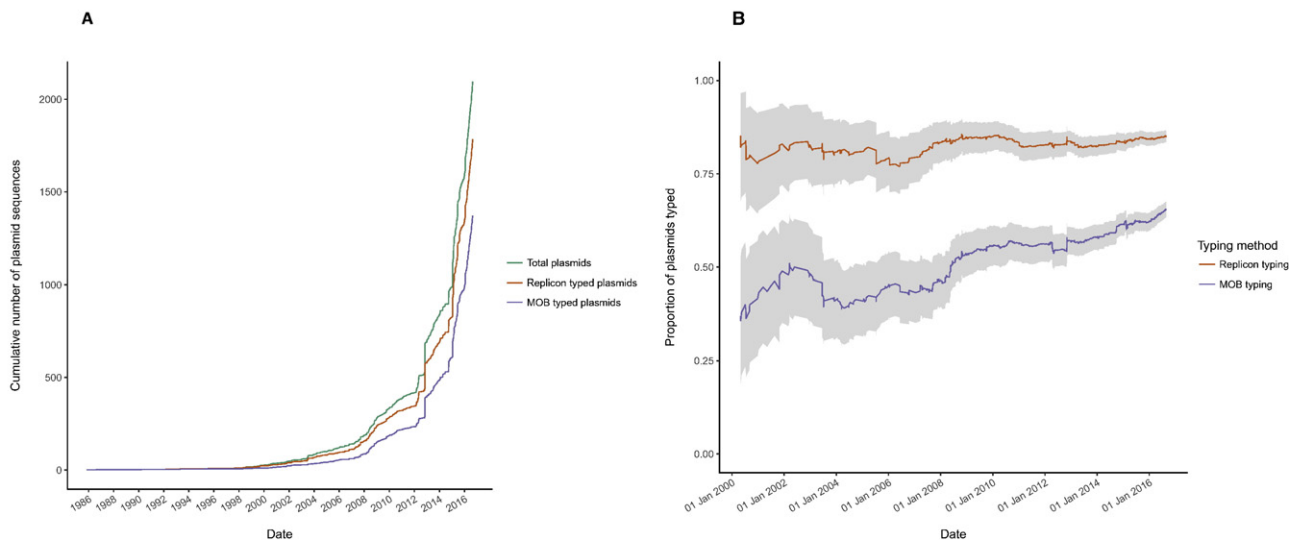


Fig. 2. (A) Cumulative number of plasmids added to NCBI from initial accession (7th November 1985) to sequence retrieval date (26th August 2016). Cumulative counts reflect all plasmids (green), as well as the subsets that were replicon typed (red) and MOB typed (blue). (B) Proportion of plasmids added to NCBI prior to a given date that could be typed by replicon typing (red) and MOB typing (blue). Grey shading around the lines represents a 95% binomial confidence interval, calculated by the Agresti-Coull method.

replicon typing has only been validated for clinical plasmids (Carattoli et al., 2014).

Relative to Shintani et al. (75% and 44% Enterobacteriaceae plasmids replicon and MOB typed respectively), we found greater typeability for both schemes. To investigate the discrepancy, our analysis pipeline was applied to the 926 Enterobacteriaceae plasmids used in the assessment by Shintani et al. (see Table S1 in Shintani et al., 2015), with E-value thresholds optimised for this dataset (Supplementary methods S1). When typing was conducted (as described in Methods), 81% and 48% of plasmids were replicon and MOB typed respectively. When filtering steps were applied (as described in Methods), 66 sequences were excluded; 83% and 48% of the remaining plasmids were replicon and MOB typed respectively (Table S1). Therefore, the discrepancy between typing performance reported here and in the study by Shintani et al. largely reflects the different bioinformatic methods used to implement the typing schemes (e.g. Shintani et al. used TBLASTN rather than PSI-BLAST to detect relaxases, and used an in-house replicon database for replicon typing). Overall, this highlights the importance of using consistent methodological approaches when undertaking comparisons across datasets.

Another aspect of typeability is the extent to which the types detected are unambiguous; more specifically, when multiple types are detected on the same plasmid, this makes it difficult to interpret the phylogenetic affiliation of the plasmid, especially when these types belong to different families of replicon/MOB type. Multi-replicon types are thought more common than multi-type MOB types, and this was also supported in our dataset. Of the 1784 plasmids replicon typed in total, 502 (28%) contained multiple replicons, of which 167 (9% of the total) represented plasmids with replicons belonging to different replicon families. In comparison, of the 1371 plasmids MOB typed in total, only 58 (4%) contained multiple relaxases, and of these, 31 (2% of the total) were plasmids with relaxases belonging to different MOB types.

Typeability of the pMLST schemes was also assessed (Supplementary Fig. S5). pMLST was conducted on plasmids belonging to an unambiguous replicon type, for which a pMLST scheme was available; overall, pMLST was conducted on 868/1784 (48%) replicon typed plasmids. Of these 868 plasmids, 82% could be assigned a known pMLST type. With the exception of the IncHI1 scheme (represented by only six plasmids), the pMLST schemes did not achieve 100% typeability. Failure to assign a pMLST type was either due to no alleles being detected, or the detected allele(s) not corresponding to a known allelic profile. Our findings suggest that plasmid backbones are more diverse than envisaged when the pMLST schemes were originally devised.

3.4. Association between plasmid characteristics and typeability

Typeability below 100% does not necessarily invalidate a classification scheme. One way to assess the usefulness of the typing schemes is by assessing their typeability in relation to plasmid characteristics that may be of most interest. Such characteristics include size (large plasmids being more likely to be conjugative, or encode accessory genes conferring important phenotypes); presence of antibiotic resistance genes; and coming from taxa with clinical relevance. As only 13% plasmids came from non-clinical taxa, clinical status was not investigated as a predictive factor in multivariate analyses. Size and number of resistance genes were univariably (Fig. 3) and multivariably associated with typeability; logistic regression analysis showed that plasmid size and number of resistance genes were significant independent predictors of plasmid typing success (Table 1). The odds of a plasmid being replicon typed were 1.49 times higher per \log_{10} kb increase in plasmid size and 1.28 times higher per additional resistance gene. The odds of a plasmid being MOB typed were 2.82 times higher per \log_{10} kb increase in plasmid size and 1.2 times higher per additional resistance gene (Table 1). Therefore, plasmids for which a replicon or MOB type could be assigned were generally larger and/or encoded more resistance genes.

Fig. 4A shows that the typeability of MOB typing varied considerably between plasmids belonging to different replicon families. The majority of Col plasmids were not assigned a MOB type. This may be linked with the small average size of Col plasmids (mean size ~6.7 kb) combined with the particularly poor typeability of MOB typing relative to replicon typing for plasmids of this size (Fig. 3).

3.5. Assessment of typing performance: concordance

The schemes showed a degree of phylogenetic concordance, with some replicon families nested entirely within a single MOB type, consistent with MOB typing being a phylogenetically broader scheme. Col plasmids are not shown at the family level, since non-concordance is to be expected; Col plasmids include phylogenetically distinct plasmids, and do not encode a single replicon type. Indeed, Col plasmids replicate by different mechanisms: in some cases, a plasmid-encoded Rep protein initiates replication; in other cases (ColE1 and ColE1-like plasmids) replication is initiated by binding of a transcript called RNAII, and inhibited by antisense RNAI (del Solar et al., 1998). Accordingly, the *in silico* replicon typing scheme targets several different Col plasmid replicon loci, including RNAI-encoding loci as well as loci encoding RepA replication

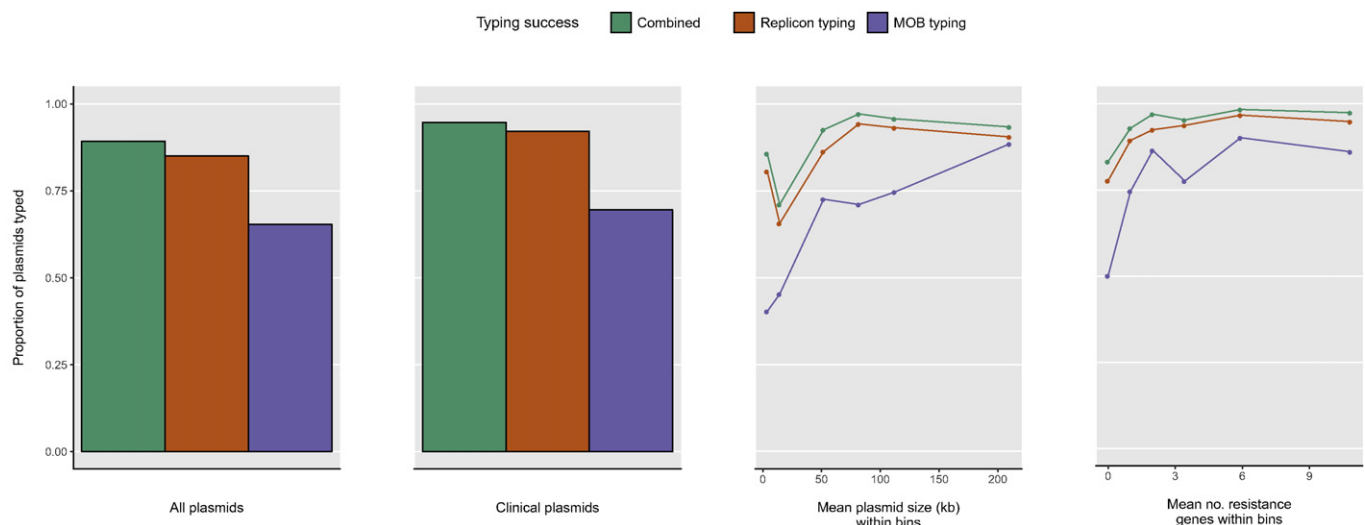


Fig. 3. Relationship between typing success and different plasmid characteristics. For plasmid size, bins are equally sized (kb) 1.3–6.1; 6.1–35; 35–68; 68–95; 95–137; 137–794. For number of resistance genes (all resistance classes included), bin ranges and sizes are: 0, $n = 1089$; 1, $n = 311$; 2, $n = 134$; 3–4, $n = 192$; 5–8, $n = 183$; 9–34, $n = 194$.

Table 1
Logistic regression analysis of plasmid typing success.

Explanatory variables	B (SE) ^a	Odds ratio (95% CI) ^b
Model of replicon typing success		
Plasmid size (log ₁₀ kb)	0.40 (0.10)***	1.49 (1.22–1.81)
Number of resistance genes	0.25 (0.04)***	1.28 (1.19–1.39)
Model of MOB typing success		
Plasmid size (log ₁₀ kb)	1.04 (0.09)***	2.82 (2.38–3.34)
Number of resistance genes	0.18 (0.02)***	1.20 (1.14–1.26)

^a B coefficients are weights associated with explanatory variables; SE is standard error.

^b Odds ratios indicate the change in odds resulting from a unit change in the value of the explanatory variable. Values >1 indicate that as the value of the explanatory variable increases, the odds of typing success increase.

*** Indicates $p < 0.0001$.

proteins (Carattoli et al., 2014); these proteins in turn belong to distinct protein families (Supplementary Table S4). Of the remaining plasmids that were investigated at the family level, IncA/C, IncF, IncH, IncQ, IncU and IncX replicon families did not show complete concordance (Fig. 4B); in cases where only a few replicon families nest outside a primary MOB type (e.g. IncA/C) non-concordance is more clearly observed by inspecting Supplementary Table S1. Replicon family–MOB type associations involving plasmids with multiple different replicon families detected are shown separately (Fig. S6); patterns of association correspond with those of the constituent single family replicons. For example, IncA/C, IncN type was associated with MOB_F, MOB_H type, consistent with Fig. 4B (IncN associated with MOB_F and IncA/C primarily associated with MOB_H). Compared with previous reports of non-concordance for just IncQ and IncP families and associated MOB types (Garcillán-Barcia et al., 2011), we find more widespread non-concordance, although we do not find non-concordance for IncP, presumably reflecting the narrower taxonomic scope of this study, in which IncP plasmids were infrequent.

There are several explanations for the patterns of non-concordance in Fig. 4B, as described previously. To investigate whether non-

concordance reflects fuzzy boundaries between MOB families, we examined PSI-BLAST hits to identify cases where different MOB query proteins aligned to the same locus on a given plasmid. This would indicate that the relaxase detected at that locus showed homology to different MOB queries, reflecting overlap between the corresponding MOB protein families; this could in turn result in non-concordance between typing schemes. There were 204 loci from 201 plasmids involved in such alignments and the MOB queries involved were all MOB_P/MOB_Q (Supplementary Table S2). This finding reiterates previous reports of overlap between MOB_P and MOB_Q families and is supported by subsequent analysis in this study with a different set of MOB queries (see Section 3.6). Accordingly, MOB_P/Q overlap could potentially explain the pattern of non-concordance observed for IncQ and IncX replicon families, which each associate with both MOB_P and MOB_Q. It could also explain the non-concordance of ColRNAI and Col(pWES) replicon types (Fig. 5A).

To gain additional insight, we further examined alignments involving different MOB queries at the same locus. If the best two hits at a locus involve different MOB queries, the assigned MOB type is presumably less reliable, compared with a situation where top hits involve the same query; therefore, MOB_P/Q overlap would seem a more plausible explanation for non-concordance in the former compared with the latter situation. Applying this reasoning to our data, we found that IncX plasmids were commonly associated with (putatively) less reliable MOB type assignments, as was one ColRNAI plasmid (accession NC_013509.1) (Supplementary Table S2). This ColRNAI plasmid, having MOB_Q and subsequently MOB_P as top hits, corresponds to the single MOB_Q-associated ColRNAI plasmid represented in Fig. 5A. In contrast, for IncQ plasmids, although there are cases where different queries align to the same locus, the top five hits at a locus consistently involve the same MOB query (Supplementary Table S2); this suggests MOB type assignments for IncQ plasmids are reliable. These findings are supported by analyses in Section 3.6.

Given our findings, for IncQ plasmids, as well as remaining replicon families not associated with MOB_P and MOB_Q, backbone mosaicism

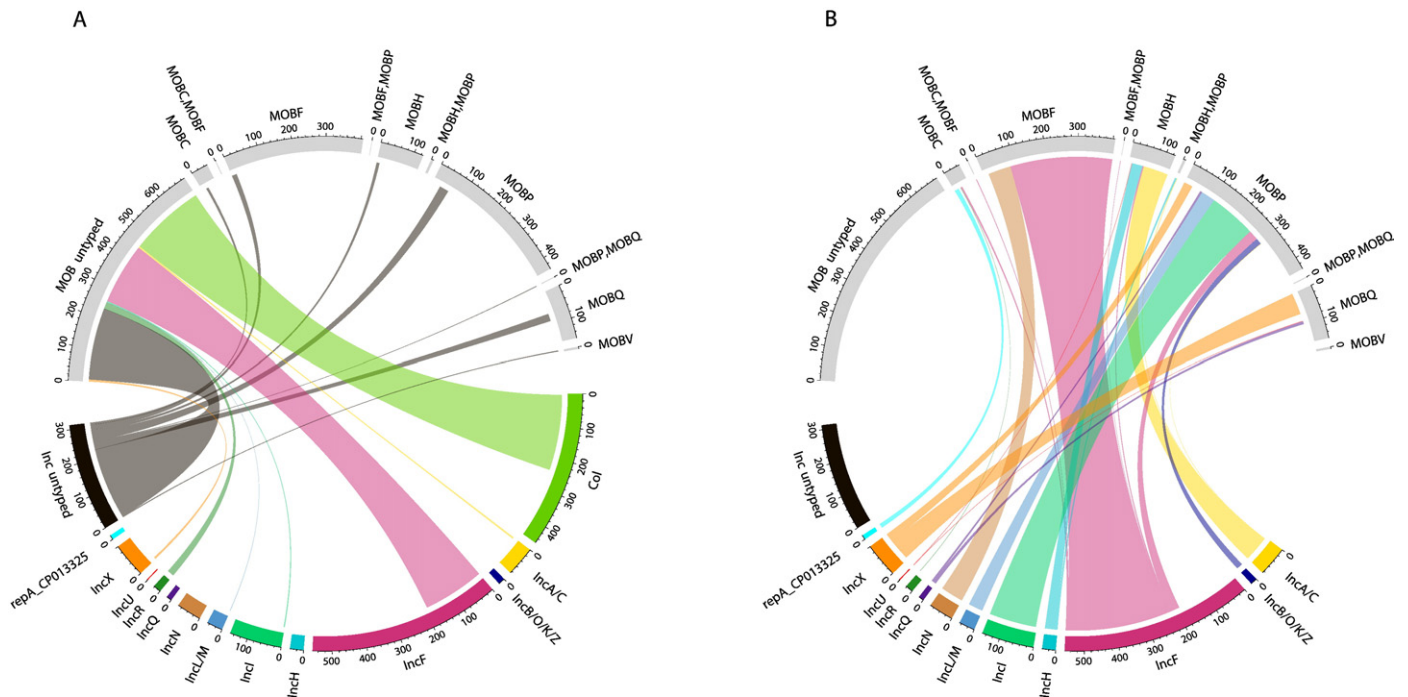


Fig. 4. Chord diagrams illustrate associations between replicon families and MOB types; circularly arranged sectors represent replicon families and MOB types, and scale bars indicate their relative sizes. Associations are indicated by intersecting chords. Replicon family sectors, coloured; MOB type sectors, grey. (A) Replicon family–MOB type associations amongst plasmids un-typed by one or both schemes. (B) Replicon family–MOB type associations amongst plasmids with both replicon and MOB type detected. Data on Col plasmids are deliberately not shown (see main text). Plasmid types are represented as the unique set of families detected on a plasmid, as described in Methods (i.e. IncF, IncF = IncF). Where multiple types of different replicon families are detected, data are not shown, but are presented in Fig. S6. Replicon family types detected in fewer than 10 plasmids are not shown except where a non-concordant pattern of association is observed (IncU) or where associated with a multi-type MOB type.

:B17. Non-concordant typing of IncF plasmids is perhaps unsurprising given the available literature on plasmid backbone mosaicism. Specifically, a study of IncFII replicons demonstrated mosaicism, with homology to the closely-related IncB, IncK and IncZ replicons (treated as IncB/O/K/Z in the Carattoli et al. *in silico* scheme), putatively due to recombination events (Praszkier et al., 1991). In our dataset, IncB/O/K/Z is associated with MOBQ (Fig. 4B) as are some IncFII replicon types (Fig. 5A), providing an explanation for IncFII non-concordance. Overall, the finding of non-concordance across hierarchical resolutions for IncF replicons indicates that non-concordance in this family may in part reflect backbone mosaicism generated by relatively recent evolutionary events.

3.6. Assessment of the robustness of MOB typing results, using alternative MOB queries

PSI-BLAST searches against the same database but using different queries may retrieve a different set of hits (Bhagwat and Aravind, 2007). Therefore, to assess the robustness of our MOB typing results, we conducted MOB typing using an alternative set of six MOB queries representing each MOB type. 54% of plasmids were MOB typed - this is lower than achieved in our original analysis, and only 6 plasmids not previously assigned a MOB type could be MOB typed using the alternative prototypes. Additional results are presented in Supplementary Figs. S8–S11. Overall, the re-analysis supports principal findings from our primary analyses which suggests that our conclusions are robust: for example, typing schemes were again found to be partially concordant, and concordance improved when plasmids were partitioned into higher resolution replicon types/sub-types.

However, there are some key differences in the assignment of plasmid types using original vs. alternative sets of MOB queries, as summarised in Supplementary Table S3. Where a plasmid is assigned a different MOB type using the original vs. alternative set of queries, this indicates that the MOB type assignment may be unreliable. Crucially, such discrepancies provide a complementary way to assess the conclusions in Section 3.5 regarding MOBQ/Q overlap as an explanation for non-concordance of the IncX replicon family and Col accession NC_013509.1. Overall, plasmids assigned a MOBQ type with original queries were more likely to be assigned a MOBQ type using alternative

queries (there are 91 plasmids for which this is the case). This underscores the ambiguous boundaries between MOBQ and MOBQ families. Of plasmids assigned a different MOB type using alternative queries, the majority belong to IncX family. When typing with alternative queries, the IncX replicon family is entirely nested within MOBQ, whereas primary analysis demonstrated non-concordance at the family-level, with the majority of IncX replicon types, except for IncX4, associated with MOBQ. One Col plasmid (NC_013509.1) was assigned a different MOB type using alternative queries; this is the same Col plasmid identified in Section 3.5 as having an unreliable MOB type. Also in support of conclusions in Section 3.5, there were no discrepancies in the typing of IncQ plasmids. Overall, this supports conclusions that MOBQ/Q overlap can complicate typing, and cause non-concordance between replicon and MOB typing schemes, as we conclude is the case for IncX non-concordance in Fig. 4B.

3.7. Associations between plasmid types and resistance genes

IncF plasmids appeared to play a key role in shuttling major β -lactamase resistance genes, including *bla*_{CTX-M-14} and *bla*_{CTX-M-15} (Fig. 6). IncA/C was strongly associated with *bla*_{CMY-2} and to a lesser extent *bla*_{NDM-1}. *bla*_{OXA-48} was associated with MOBQ/IncL/M, as previously reported (Poirel et al., 2012). Overall though, plasmid backbones were associated with a variety of resistance genes, as expected, given widespread multi-drug resistance. This generally held true even when looking at pMLST types (Supplementary Fig. S6). When associations at the replicon family/MOB type level were investigated statistically, the null hypothesis that resistance genes were distributed randomly amongst plasmid types was rejected ($p < 0.0001$). However, this result should be interpreted cautiously given the likelihood of biases in the plasmid accessions submitted to NCBI.

4. Conclusions

Retrieval and curation of a dataset of complete plasmids and associated metadata from NCBI requires bioinformatics expertise. Without such a dataset, even the most basic aspects of plasmid biology, such as the size distribution, are obscured. Overall, our experience suggests that obtaining a high-quality dataset of complete plasmids requires

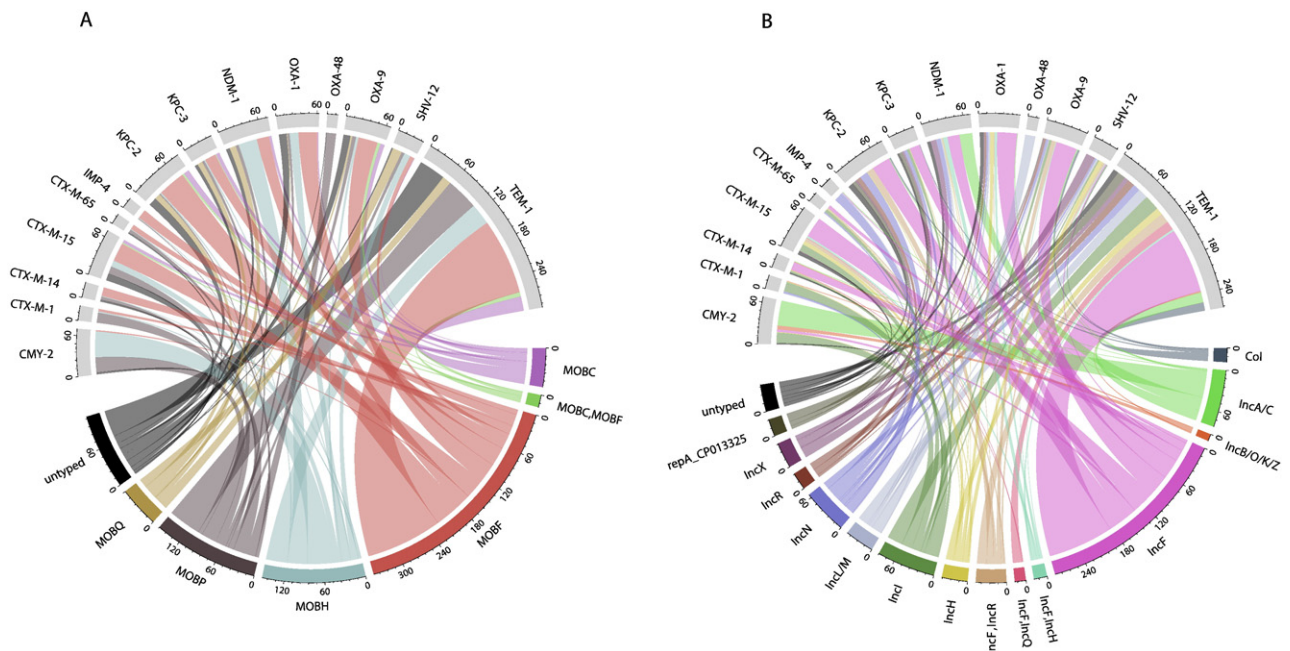


Fig. 6. Associations between β -lactamase resistance genes identified and (A) MOB types; (B) replicon families. MOB types/replicon families detected in fewer than 10 plasmids are not shown.

some degree of quality-filtering, and this is likely to become more important as more sequences are added to NCBI. However, our curation methods were guided by the annotations (and mis-annotations) observed in the retrieved dataset. Building on and fine-tuning the suggested curation methods, as publicly available databases expand, would provide a valuable resource for future research. Indeed, curated plasmid datasets have a wide range of applications beyond investigation of typing schemes (Orlek et al., *in press*).

This study demonstrated that 11% of Enterobacteriaceae plasmids available in mid-2016, could not be typed by either replicon or MOB typing schemes. As many more plasmids are sequenced, including from a wider diversity of strains (especially strains harbouring small plasmids or plasmids with few resistance genes) typing success may decrease. Overall, replicon typing covers a wider diversity of Enterobacteriaceae plasmids than MOB typing. However, it is unclear whether novel replicon probes can be found to detect currently non-typeable plasmids. Furthermore, interpreting typing success from a dataset of complete plasmid genomes does not reflect the context in which *in silico* typing tools are often used. Plasmids can be difficult to assemble accurately from reads generated by widely used short-read sequencing technology, resulting in contig assemblies of variable quality, which may impede successful typing (Clausen et al., 2016; Inouye et al., 2014).

Although previous studies have examined plasmid backbone mosaicism and the concordance of typing schemes, as well as associations between resistance genes and plasmid types, this has not been based on comprehensive bioinformatic analysis at hierarchical typing resolutions. Our findings demonstrate that plasmid typing is inherently difficult and even the relatively conserved loci used for major typing schemes exhibit phylogenetic discordance, seen most notably with the IncF replicon. In this case, partitioning plasmids into more homogenous groups using pMLST improved concordance, but only partially, perhaps reflecting mosaicism resulting from relatively recent evolutionary events.

Our study has also highlighted issues with reproducibility of typing results, in particular, MOB typing results. To conduct replicon typing, plasmids are searched against the PlasmidFinder database, which has remained relatively static in size and content since first developed, and minor changes are recorded (<https://cge.cbs.dtu.dk/services/PlasmidFinder/>); hence replicon typing results should be reproducible. On the other hand, MOB typing, has traditionally been conducted by querying known MOB proteins against a database of plasmids (and sometimes also chromosomes, if integrative conjugative elements are of interest), using profile-based methods such as PSI-BLAST. Differences in database size and content can influence MOB typing results. In practice, especially when MOB typing a handful of plasmids, standard non-iterative BLAST has been commonly used (Dziewit and Bartosik, 2014; Shintani et al., 2015); however, this approach is likely to be less powerful, since position-specific information about conservation of relaxase protein residues amongst a dataset of plasmids is not harnessed. Our curated plasmid dataset is publicly available and could be used as a basis for conducting more reproducible MOB typing (Orlek et al., *in press*), whilst harnessing information contained within the large dataset. The ConJDB database, which includes relaxase proteins (Guglielmini et al., 2013), could also be used for this purpose. Our analysis also demonstrates that the overlap between MOBp and MOBq families can complicate MOB typing; MOBp and MOBq types should therefore be treated with particular caution, and ideally, the robustness of MOBp/MOBq type assignments should be validated using different MOB queries.

We found that resistance genes tended not to show strong fidelity towards particular plasmid backbones, although the association between IncL/M and *bla*_{OXA-48} was supported. The patterns of association observed should be interpreted cautiously since they reflect biases in the NCBI database. An alternative approach might be to assess associations between plasmid backbones and resistance genes within specific geographical or temporal contexts of interest, for example by using BioSample metadata (Barrett et al., 2012; Federhen et al., 2014). If

associations between specific resistance genes and plasmid backbones hold across broad timeframes or geographies this would represent stronger evidence for genuinely stable associations. More generally, combining genomic epidemiology (plasmid typing and resistance gene detection) with BioSample metadata could also provide contextual epidemiological information for a given plasmid (i.e. whether it is an isolated case or part of an outbreak). As more plasmid sequence data become available, this kind of analysis could become increasingly useful (Chang et al., 2016; Nolte et al., 2015).

In summary, 'Ordering the mob' denotes not only the challenging pursuit of plasmid typing, but also the process of obtaining a dataset of complete plasmids from NCBI. Using our curated plasmid dataset, we demonstrate that current typing schemes fail to classify the complete diversity of plasmids, and that there is a degree of non-concordance between replicon and MOB typing schemes. In some cases, non-concordance is likely to reflect the plasticity (and consequent mosaicism) of plasmid genomes, whilst in other cases it reflects the ambiguous boundaries between MOBp and MOBq types.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.plasmid.2017.03.002>.

Acknowledgements

The research was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) [HPRU-2012-10041]. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England. NS is currently funded through an NIHR/University of Oxford Academic Clinical Lectureship. TP is an NIHR Senior Investigator.

References

- Akiba, M., Sekizuka, T., Yamashita, A., Kuroda, M., Fujii, Y., Murata, M., et al., 2016. Distribution and Relationships of Antimicrobial Resistance Determinants Among Extended-Spectrum-Cephalosporin-Resistant or Carbapenem-Resistant *Escherichia coli* Isolates From Rivers and Sewage Treatment Plants in India. 60:pp. 2972–2980. <http://dx.doi.org/10.1128/AAC.01950-15>.Address.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402. <http://dx.doi.org/10.1093/nar/25.17.3389>.
- Alvarado, A., Garcillán-Barcia, M.P., de la Cruz, F., 2012. A degenerate primer MOB typing (DPMT) method to classify gamma-proteobacterial plasmids in clinical and environmental settings. *PLoS One* 7. <http://dx.doi.org/10.1371/journal.pone.0040438>.
- Baker, S., Hardy, J., Sanderson, K.E., Quail, M., Goodhead, I., Kingsley, R.A., et al., 2007. A novel linear plasmid mediates flagellar variation in *Salmonella* Typhi. *PLoS Pathog.* 3:0605–0610. <http://dx.doi.org/10.1371/journal.ppat.0030059>.
- Barrett, T., Clark, K., Gevorgyan, R., Gorenkov, V., Gribov, E., Karsch-Mizrachi, I., et al., 2012. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* 40. <http://dx.doi.org/10.1093/nar/gkr1163>.
- de Been, M., Lanza, V.F., de Toro, M., Scharringa, J., Dohmen, W., Du, Y., et al., 2014. Dissemination of cephalosporin resistance genes between *Escherichia coli* strains from farm animals and humans by specific plasmid lineages. *PLoS Genet.* 10. <http://dx.doi.org/10.1371/journal.pgen.1004776>.
- Belkum, V.A., Tassios, P.T., Dijksoon, L., Haeggman, S., Cookson, B., 2007. Guideline for the validation and application of typing methods for use in bacterial epidemiology. *Clin. Microbiol. Infect.* 13, 1–46.
- Bhagwat, M., Aravind, L., 2007. PSI-BLAST tutorial. *Methods Mol. Biol.* 395, 177–186 (doi: 1-59745-514-8:177 [pii]).
- Bianco, P.R., Kowalczykowski, S.C., 1997. The recombination hotspot Chi is recognized by the translocating RecBCD enzyme as the single strand of DNA containing the sequence 5'-GCTGGTGG-3'. *Proc. Natl. Acad. Sci. U. S. A.* 94:6706–6711. <http://dx.doi.org/10.1073/pnas.94.13.6706>.
- Bikard, D., Euler, C.W., Jiang, W., Nussenzweig, P.M., Goldberg, G.W., Dupret, X., et al., 2014. Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. *Nat. Biotechnol.* 32:1146–1150. <http://dx.doi.org/10.1038/nbt.3043>.
- Both, A., Huang, J., Kaase, M., Hezel, J., Wertheimer, D., Fenner, I., et al., 2016. First report of *Escherichia coli* co-producing NDM-1 and OXA-232. *Diagn. Microbiol. Infect. Dis.* 86: 437–438. <http://dx.doi.org/10.1016/j.diagmicrobio.2016.09.005>.
- Brolund, A., Sandegren, L., 2015. Characterization of ESBL disseminating plasmids. *Infect. Dis.* 42:351–8. <http://dx.doi.org/10.3109/23744235.2015.1062536>.
- Carattoli, A., 2009. Resistance plasmid families in Enterobacteriaceae. *Antimicrob. Agents Chemother.* 53:2227–2238. <http://dx.doi.org/10.1128/AAC.01707-08>.
- Carattoli, A., 2013. Plasmids and the spread of resistance. *Int. J. Med. Microbiol.* 303: 298–304. <http://dx.doi.org/10.1016/j.ijmm.2013.02.001>.

- Carattoli, A., Bertini, A., Villa, L., Falbo, V., Hopkins, K.L., Threlfall, E.J., 2005. Identification of plasmids by PCR-based replicon typing. *J. Microbiol. Methods* 63:219–228. <http://dx.doi.org/10.1016/j.mimet.2005.03.018>.
- Carattoli, A., Zankari, E., Garcia-Fernández, A., Larsen, M.V., Lund, O., Villa, L., et al., 2014. In silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* 58:3895–3903. <http://dx.doi.org/10.1128/AAC.02412-14>.
- Chang, W.E., Peterson, M.W., Garay, C.D., Korves, T., 2016. Pathogen metadata platform: software for accessing and analyzing pathogen strain information. *BMC Bioinf.* 17: 379. <http://dx.doi.org/10.1186/s12859-016-1231-2>.
- Clausen, P., Zankari, E., Aarestrup, M.F., Lund, O., 2016. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *J. Antimicrob. Chemother.* 1–5. <http://dx.doi.org/10.1093/jac/dkw184>.
- Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., et al., 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423. <http://dx.doi.org/10.1093/bioinformatics/btp163>.
- Conlan, S., Thomas, P.J., Deming, C., Park, M., Lau, A.F., Dekker, J.P., et al., 2014. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci. Transl. Med.* 6 (254ra126). [10.1126/scitranslmed.3009845](http://dx.doi.org/10.1126/scitranslmed.3009845).
- Dziewit, L., Bartosik, D., 2014. Plasmids of psychrophilic and psychrotolerant bacteria and their role in adaptation to cold environments. *Front. Microbiol.* 5:1–14. <http://dx.doi.org/10.3389/fmicb.2014.00596>.
- Federhen, S., Clark, K., Barrett, T., Parkinson, H., Ostell, J., Kodama, Y., et al., 2014. Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Stand. Genomic Sci.* 9:1275–1277. <http://dx.doi.org/10.4056/sigs.4851102>.
- Fetchko, M., Kitts, A., 2011. The “nucleotide” page. GenBank Submissions Handb Available at: <https://www.ncbi.nlm.nih.gov/books/NBK293904/> (Accessed December 3, 2016).
- Francia, M.V., Varsaki, A., Garcillán-Barcia, M.P., Latorre, A., Drainas, C., De La Cruz, F., 2004. A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiol. Rev.* 28:79–100. <http://dx.doi.org/10.1016/j.femsre.2003.09.001>.
- García-Fernández, A., Carattoli, A., 2010. Plasmid double locus sequence typing for IncH2 plasmids, a subtyping scheme for the characterization of IncH2 plasmids carrying extended-spectrum β -lactamase and quinolone resistance genes. *J. Antimicrob. Chemother.* 65:1155–1161. <http://dx.doi.org/10.1093/jac/dkq101>.
- García-Fernández, A., Chiaretto, G., Bertini, A., Villa, L., Fortini, D., Ricci, A., et al., 2008. Multilocus sequence typing of IncI1 plasmids carrying extended-spectrum β -lactamases in *Escherichia coli* and salmonella of human and animal origin. *J. Antimicrob. Chemother.* 61:1229–1233. <http://dx.doi.org/10.1093/jac/dkn131>.
- García-Fernández, A., Villa, L., Moodley, A., Hasman, H., Miriagou, V., Guardabassi, L., et al., 2011. Multilocus sequence typing of IncN plasmids. *J. Antimicrob. Chemother.* 66: 1987–1991. <http://dx.doi.org/10.1093/jac/dkr225>.
- Garcillán-Barcia, M.P., de la Cruz, F., 2013. Ordering the bestiary of genetic elements transmissible by conjugation. *Mob. Genet. Elements* 3, e24263. <http://dx.doi.org/10.4161/mge.24263>.
- Garcillán-Barcia, M.P., Francia, M.V., De La Cruz, F., 2009. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.* 33: 657–687. <http://dx.doi.org/10.1111/j.1574-6976.2009.00168.x>.
- Garcillán-Barcia, M.P., Alvarado, A., De La Cruz, F., 2011. Identification of bacterial plasmids based on mobility and plasmid population biology. *FEMS Microbiol. Rev.* 35:936–956. <http://dx.doi.org/10.1111/j.1574-6976.2011.00291.x>.
- Gu, Z., Gu, L., Eils, R., Schlesner, M., Brors, B., 2014. Circlize implements and enhances circular visualization in R. *Bioinformatics* 30:2811–2812. <http://dx.doi.org/10.1093/bioinformatics/btu393>.
- Guglielmini, J., Quintais, L., Garcillán-Barcia, M.P., de la Cruz, F., Rocha, E.P.C., 2011. The repertoire of ice in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* 7. <http://dx.doi.org/10.1371/journal.pgen.1002222>.
- Guglielmini, J., De La Cruz, F., Rocha, E.P.C., 2013. Evolution of conjugation and type IV secretion systems. *Mol. Biol. Evol.* 30:315–331. <http://dx.doi.org/10.1093/molbev/mss221>.
- Hancock, S.J., Phan, M.-D., Peters, K.M., Forde, B.M., Chong, T.M., Yin, W.-F., et al., 2016. Identification of IncA/C plasmid replication and maintenance genes and development of a plasmid multi-locus sequence-typing scheme. *Antimicrob. Agents Chemother.* <http://dx.doi.org/10.1128/AAC.01740-16> (AAC01740-16).
- Hiraga, S.-I., Sugiyama, T., Itoh, T., 1994. Comparative Analysis of the Replicon Regions of Eleven ColE2-related Plasmids. 176 pp. 7233–7243.
- Holt, K.E., Thieu Nga, T.V., Thanh, D.P., Vinh, H., Kim, D.W., Vu Tra, M.P., et al., 2013. Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proc. Natl. Acad. Sci. U. S. A.* 110:17522–17527. <http://dx.doi.org/10.1073/pnas.1308632110>.
- Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M.B., Pope, B.J., Tomita, T., et al., 2014. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* 6:90. <http://dx.doi.org/10.1186/s13073-014-0090-6>.
- Iredell, J., Brown, J., Tagg, K., 2016. Antibiotic resistance in Enterobacteriaceae: mechanisms and clinical implications. *BMJ*:h6420 <http://dx.doi.org/10.1136/bmj.h6420>.
- Jones, D.T., Swindells, M.B., 2002. Getting the most from PSI-BLAST. *Trends Biochem. Sci.* 27:161–164. [http://dx.doi.org/10.1016/S0968-0004\(01\)02039-4](http://dx.doi.org/10.1016/S0968-0004(01)02039-4).
- Kans, J., 2013. Entrez direct: E-utilities on the UNIX command line. Entrez Programming Utilities Help. Entrez Program. Util. Help Available at: <http://www.ncbi.nlm.nih.gov/books/NBK179288/> (Accessed November 24, 2016).
- Kulinska, A., Czeredy, M., Hayes, F., Jagura-Burdzy, G., 2008. Genomic and functional characterization of the modular broad-host-range RA3 plasmid, the archetype of the IncU group. *Appl. Environ. Microbiol.* 74:4119–4132. <http://dx.doi.org/10.1128/AEM.00229-08>.
- Lanza, V.F., de Toro, M., Garcillán-Barcia, M.P., Mora, A., Blanco, J., Coque, T.M., et al., 2014. Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genet.* 10. <http://dx.doi.org/10.1371/journal.pgen.1004766>.
- Larsen, M.V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R.L., et al., 2012. Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.* 50:1355–1361. <http://dx.doi.org/10.1128/JCM.06094-11>.
- Leverstein-van Hall, M.A., Dierikx, C.M., Cohen Stuart, J., Voets, G.M., van den Munckhof, M.P., van Essen-Zandbergen, A., et al., 2011. Dutch patients, retail chicken meat and poultry share the same ESBL genes, plasmids and strains. *Clin. Microbiol. Infect.* 17: 873–880. <http://dx.doi.org/10.1111/j.1469-0691.2011.03497.x>.
- Livermore, D.M., Woodford, N., 2006. The β -lactamase threat in Enterobacteriaceae, *Pseudomonas* and *Acinetobacter*. *Trends Microbiol.* 14:413–420. <http://dx.doi.org/10.1016/j.tim.2006.07.008>.
- Nahin, A., 2008. Create date — new field indicates when record added to PubMed. NLM tech. Bull. Available at: https://www.nlm.nih.gov/pubs/techbull/nd08/nd08_pm_new_date_field.html (Accessed November 24, 2016).
- NCBI, 2014. What is Whole Genome Shotgun (WGS)? Whole Genome Shotgun Submissions Available at: <https://www.ncbi.nlm.nih.gov/genbank/wgs/> (Accessed December 8, 2016).
- Nolte, N., Kurzawa, N., Eils, R., Herrmann, C., 2015. MapMyFlu: visualizing spatio-temporal relationships between related influenza sequences. *Nucleic Acids Res.* 43: W547–W551. <http://dx.doi.org/10.1093/nar/gkv417>.
- Nordmann, P., Dortet, L., Poirel, L., 2012. Carbapenem resistance in enterobacteriaceae: here is the storm! *Trends Mol. Med.* 18:263–272. <http://dx.doi.org/10.1016/j.molmed.2012.03.003>.
- Orlek, A., Stoesser, N., Anjum, M.F., Doumith, M., Ellington, M., 2017. Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. *Front. Microbiol.* <http://dx.doi.org/10.3389/fmicb.2017.00182>.
- Orlek, A., Phan, H., Sheppard, A. E., Doumith, M., Ellington, M., Peto, T., et al. A curated dataset of complete Enterobacteriaceae plasmids compiled from the NCBI nucleotide database. Data in Brief (in press).
- Osborn, A.M., da Silva Tatley, F.M., Steyn, L.M., Pickup, R.W., Saunders, J.R., 2000. Mosaic plasmids and mosaic replicons: evolutionary lessons from the analysis of genetic diversity in IncFII-related replicons. *Microbiology* 146, 2267–2275.
- Pallecchi, L., Bartoloni, A., Fiorelli, C., Mantella, A., Di Maggio, T., Gamba, H., et al., 2007. Rapid dissemination and diversity of CTX-M extended-spectrum β -lactamase genes in commensal *Escherichia coli* isolates from healthy children from low-resource settings in Latin America. *Antimicrob. Agents Chemother.* 51:2720–2725. <http://dx.doi.org/10.1128/AAC.00026-07>.
- Partridge, S.R., 2011. Analysis of antibiotic resistance regions in Gram-negative bacteria. *FEMS Microbiol. Rev.* 35:820–855. <http://dx.doi.org/10.1111/j.1574-6976.2011.00277.x>.
- Pecora, N.D., Li, N., Allard, M., Li, C., Albano, E., Delaney, M., et al., 2015. Genomically informed surveillance for carbapenem-resistant enterobacteriaceae in a health care system. *MBio* 6:1–11. <http://dx.doi.org/10.1128/mBio.01030-15>.
- Phan, M.D., Kidgell, C., Nair, S., Holt, K.E., Turner, A.K., Hinds, J., et al., 2009. Variation in *Salmonella enterica* serovar typhi IncHI1 plasmids during the global spread of resistant typhoid fever. *Antimicrob. Agents Chemother.* 53:716–727. <http://dx.doi.org/10.1128/AAC.00645-08>.
- Poirel, L., Bonnin, R.A., Nordmann, P., 2012. Genetic features of the widespread plasmid coding for the carbapenemase OXA-48. *Antimicrob. Agents Chemother.* 56: 559–562. <http://dx.doi.org/10.1128/AAC.05289-11>.
- Praszkier, J., Wei, T., Siemering, K., Pittard, J., 1991. Comparative analysis of the replication regions of IncB, IncK, and IncZ plasmids. *J. Bacteriol.* 173, 2393–2397.
- Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35. <http://dx.doi.org/10.1093/nar/gkl842>.
- Pruitt, K., Brown, G., Murphy, M., 2010. RefSeq Frequently Asked Questions (FAQ). RefSeq Help Available at: https://www.ncbi.nlm.nih.gov/books/NBK50679/#RefSeqFAQ_why_are_refseq_accessions_remo (Accessed December 4, 2016).
- Ramsay, J.P., Kwong, S.M., Murphy, R.J.T., Yui Eto, K., Price, K.J., Nguyen, Q.T., et al., 2016. An updated view of plasmid conjugation and mobilization in *Staphylococcus*. *Mob. Genet. Elements* 6, e1208317. <http://dx.doi.org/10.1080/2159256X.2016.1208317>.
- Rawlings, D.E., Tietze, E., 2001. Comparative biology of IncQ and IncQ-like plasmids. *Microbiol. Mol. Biol. Rev.* 65:481–496. <http://dx.doi.org/10.1128/MMBR.65.4.481-496.2001> (table of contents).
- Riley, M., Gordon, D., 1992. A survey of Col plasmids in natural isolates of *Escherichia coli* and an investigation into the stability of Col-plasmid lineages. *J. Gen. Microbiol.* 138, 1345–1352.
- Shintani, M., Sanchez, Z.K., Kimbara, K., 2015. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Front. Microbiol.* 6:1–16. <http://dx.doi.org/10.3389/fmicb.2015.00242>.
- Smillie, C., Garcillán-Barcia, M.P., Francia, M.V., Rocha, E.P., de la Cruz, F., 2010. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* 74:434–452. <http://dx.doi.org/10.1128/MMBR.00020-10>.
- del Solar, G., Giraldo, R., Ruiz-Echevarría, M.J., Espinosa, M., Díaz-Orejas, R., 1998. Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.* 62, 434–464 (doi:1092-2172/98/\$04.0010).
- Taylor, D., Gibreel, A., Lawley, T., Tracz, D., 2004. Chapter 23: antibiotic resistance plasmids. In: Funnell, B., Phillips, G. (Eds.), *Plasmid Biology*. ASM Press, Washington, DC, pp. 473–485.
- Valverde, A., Cantón, R., Garcillán-Barcia, M.P., Novais, Á., Galán, J.C., Alvarado, A., et al., 2009. Spread of blaCTX-M-14 is driven mainly by IncK plasmids disseminated among *Escherichia coli* phylogroups A, B1, and D in Spain. *Antimicrob. Agents Chemother.* 53:5204–5212. <http://dx.doi.org/10.1128/AAC.01706-08>.

- 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. <http://dx.doi.org/10.1093/jac/dkq347>.
- Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., et al., 2014. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* 42. <http://dx.doi.org/10.1093/nar/gkt1099>.
- Williams, J.J., Hergenrother, P.J., 2008. Exposing plasmids as the Achilles' heel of drug-resistant bacteria. *Curr. Opin. Chem. Biol.* 12:389–399. <http://dx.doi.org/10.1016/j.cbpa.2008.06.015>.
- Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al., 2012. Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67:2640–2644. <http://dx.doi.org/10.1093/jac/dks261>.