Enterobacterales plasmid sharing amongst human bloodstream infections, livestock, wastewater, and waterway niches in Oxfordshire, UK

William Matlock*[,1], Samuel Lipworth*[,1,2], Kevin K. Chau[1], Manal Abu Oun[3], Leanne Barker[1], James Kavanagh[1], Monique Andersson[2], Sarah Oakley[2], Marcus Morgan[2], Derrick W. Crook[1,2,4,5], Daniel S. Read[6], Muna Anjum[3], Liam P. Shaw[#,7]/Nicole Stoesser[#,1,2,4,5], REHAB Consortium

[*] Contributed equally as first authors

[#] Contributed equally as senior authors

[1] Nuffield Department of Medicine, University of Oxford, Oxford, UK

[2] Oxford University Hospitals NHS Foundation Trust, Oxford, UK

[3] Animal and Plant Health Agency, Addlestone, UK

[4] NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, UK

[5] NIHR Biomedical Research Centre, Oxford, UK

[6] Centre for Ecology and Hydrology, Wallingford, UK

[7] Department of Biology, University of Oxford, Oxford, UK

Correspondence: William Matlock (william.matlock@ndm.ox.ac.uk) and Nicole Stoesser (nicole.stoesser@ndm.ox.ac.uk)

Keywords: plasmid, *Enterobacterales*, antimicrobial resistance (AMR), bloodstream infections (BSI), One Health, genomic epidemiology, *Escherichia coli*

Abstract

Plasmids enable the dissemination of antimicrobial resistance (AMR) in common *Enterobacterales* pathogens, representing a major public health challenge. However, the extent of plasmid sharing and evolution between *Enterobacterales* causing human infections and other niches remains unclear, including the emergence of resistance plasmids. Dense, unselected sampling is highly relevant to developing our understanding of plasmid epidemiology and designing appropriate interventions to limit the emergence and dissemination of plasmid-associated AMR. We established a geographically and temporally restricted collection of human bloodstream infection (BSI)-associated, livestock-associated (cattle, pig, poultry, and sheep faeces, farm soils) and wastewater treatment work (WwTW)-associated (influent, effluent, waterways upstream/downstream of effluent outlets) *Enterobacterales*. Isolates were collected between 2008-2020 from sites <60km apart in Oxfordshire, UK. Pangenome analysis of plasmid clusters revealed shared "backbones", with phylogenies suggesting an intertwined ecology where well-conserved plasmid backbones carry diverse accessory functions, including AMR genes. Many plasmid "backbones" were seen across species and niches, raising the possibility that plasmid movement between these followed by rapid accessory gene change could be relatively common. Overall, the signature of identical plasmid sharing is likely to be a highly transient one, implying that plasmid movement might be occurring at greater rates than previously estimated, raising a challenge for future genomic One Health studies.

Funding

This study was funded by the Antimicrobial Resistance Cross-council Initiative supported by the seven research councils and the NIHR, UK.

Introduction

Enterobacterales are found both in human niches (e.g., hospital patients(1,2) and wastewater(3)) and non-human niches (e.g., livestock-associated(4,5) and waterways(6)). In recent decades, widespread carriage of antimicrobial resistance (AMR) genes has complicated the treatment of *Enterobacterales* infections(7,8). The dissemination of AMR genes between *Enterobacterales* occurs in a 'Russian-doll'-style hierarchy of nested, mobilisable genetic structures(9): genes not only move between bacterial hosts on mobilisable or conjugative plasmids but can also be transferred within and between plasmids and chromosomes by smaller mobile genetic elements (MGEs) such as insertion sequences(10,11). Despite gene gain/loss events, many plasmids have been shown to have a persistent structure encoding replication and transfer machinery(12,13).

Many plasmids can transfer between species and are seen across different niches(14) but the extent to which they are shared between human and non-human niches remains poorly understood. Previous studies investigating this topic have often been limited in size given the genetic diversity in these niches(15,16), and/or restricted to single species(17) or drug-resistant isolates(18), or are systematic studies, pooling geographically/temporally disparate samples (19,20). Further, fragmented genome assemblies in many cases make recovering complete plasmids, and other MGEs, impossible(21).

Instances of cross-niche transfer of plasmids are well-described, but the frequency of such events is poorly characterised. There are multiple instances where AMR genes have emerged from non-human niches and subsequently become major clinical problems in human *Enterobacterales* infections, highlighting the relevance of inter-niche transfer in AMR gene dissemination (e.g., $bla_{CTX\text{-}M}$, *mcr-1*(22) and $bla_{\text{NDM-1}}$(23)). In general, environmental bacteria

are believed to be the original source of AMR genes that eventually become prevalent in clinical settings after transfer into clinical pathogens. However, we know little about natural rates of inter-niche transfer beyond these high-profile examples. It remains unclear how plasmids evolve within natural populations, meaning we understand little about the wider context in which AMR genes emerge and disseminate.

To explore *Enterobacterales* plasmid diversity and sharing across niches in a geographically and temporally restricted context, we studied hybrid assemblies (i.e., using both long and short reads) of large *Enterobacterales* isolate collections in Oxfordshire, UK, from (i) human bloodstream infections (BSI; 2008-2018), (ii) livestock-associated sources (faeces from cattle, pigs, poultry, sheep; surrounding environmental soils; [all 2017 except poultry 2019-2020], and (iii) wastewater treatment work (WwTW)-associated sources (influent, effluent, waterways upstream/downstream of effluent outlets; Oxfordshire, 2017).

Results

Our dataset of *n*=3,697 plasmids from *n*=1,458 isolates (Figure 1a, Table 1) contained bacteria from human bloodstream infections (BSI; *n*=1,880 plasmids from *n*=738 isolates), livestock-associated sources (cattle, pig, poultry, and sheep faeces, soils surrounding livestock farms; *n*=1,155 plasmids from *n*=512 isolates), and from wastewater treatment works (WwTW)-associated sources (influent, effluent, waterways upstream/downstream of effluent outlets; *n*=662 plasmids from *n*=208 isolates). All sampling sites were <60km apart (Figure 1b) and timeframes overlapped (2008-2020; Figure 1c). Isolates had a median 2 plasmids (IQR=1-4, range=0-16). Major *Enterobacterales* genera represented included: *n*=1,044 *Escherichia*, *n*=211 *Klebsiella*, *n*=125 *Citrobacter*, and *n*=63 *Enterobacter*.

Table 1. Isolate niche breakdown.			
Niche	**Sample type(s)**	**No. isolates**	**No. plasmids**
Bloodstream infections (BSI)	Community, nosocomial, other healthcare associated infections	738	1,880
Livestock-associated	Cattle faeces	133	215
	Sheep faeces	113	286
	Pig faeces	104	352
	Poultry faeces	34	112
	Soil surrounding livestock farms	128	190
Wastewater treatment work (WwTW)-associated	Influent	88	313
	Upstream waterways	25	60
	Effluent/downstream waterways	95	289
Total		**1,458**	**3,697**

Table 2. Isolate genus breakdown.						
Niche	**Isolate genus**					**Total**
	Citrobacter	***Enterobacter***	***Escherichia***	***Klebsiella***	**Other**	
Bloodstream infections (BSI)	6	11	547	161	13	**738**
Livestock-associated	54	10	433	14	1	**512**
Wastewater treatment work (WwTW)-associated	65	42	64	37	0	**208**
Total	**125**	**63**	**1,044**	**212**	**14**	**1,458**

Sampling niche was strongly associated with isolate genus (Fisher's test, p-value<0.001; Table 2). *Klebsiella* isolates were disproportionately derived from BSI versus other niches (76% [161/212] versus 51% [738/1,458]). *Citrobacter* and *Enterobacter* were disproportionately derived from WwTW-associated versus other niches (52% [65/125] and 67% [42/63] versus 14% [208/1,458]). Chromosomal Mash trees (see Materials and Methods) for the two most common species in the dataset, *E. coli* (72% [1,044/1,458]; see Appendix 1 Figure 1) and *K. pneumoniae* (11% [163/1,458]; Appendix 1 Figure 2) demonstrated intermixing of human and non-human isolates within clades, consistent with species-lineages not being structured by niche.

We contextualised our plasmids within known plasmid diversity using 'plasmid taxonomic units' (PTUs; using COPLA, see Materials and Methods), designed to be equivalent to a plasmid 'species'. We found 32% (1,193/3,697) of plasmids were unclassified, highlighting the substantial plasmid diversity within this geographically restricted dataset. In total, we

found *n*=67 known PTUs, containing a median 9 plasmids (IQR=4-30, range=1-556), with the largest PTU-F_E (556/2,504), corresponding to F-type *Escherichia* plasmids.

Near-identical plasmid sharing observed between human and livestock-associated *Enterobacterales*

We screened for near-identical plasmids shared across isolates by grouping those with a low Mash distance (d<0.0001) and highly similar lengths (longest plasmid ≤1% longer than shorter plasmids; note that this near-identical threshold becomes an identical threshold for extremely small plasmids; see Materials and Methods). We found *n*=225 near-identical groups of ≥2 members, recruiting 19% (712/3,697) plasmids. Bootstrapping accumulation curves for near-identical plasmid groups and singletons per the number of isolates (ACs; see Materials and Methods), we revealed a highly 'open' accumulation (Heap's parameter γ=0.97, Appendix 1 Figure 3) suggesting further isolate sampling would detect more unique plasmids approximately linearly. Restricted to BSI/livestock-associated isolates alone, we found similar curves for both niches (BSI γ=0.98, livestock-associated γ=0.94), suggesting they had similar levels of plasmid diversity.

Near-identical pairs of plasmids were most common, representing 71% (159/225) of groups (group size IQR=2-3, range=2-32). Plasmid members of near-identical groups represented multiple bacterial host STs (25% [56/225]), species (4% [9/225]), and genera (4% [9/225]), consistent with plasmids capable of inter-lineage/species/genus transfer. Further, 8% (17/225) of near-identical groups contained plasmids found across human BSIs and ≥1 other sampling niche (livestock-associated/WwTW-associated), suggesting inter-niche transfer (i.e., 'cross-niche groups'; Figure 2a). Within cross-niche groups, *n*=3/17 contained plasmids from multiple bacterial species (Figure 2b), and most consisted of conjugative plasmids (*n*=5/17

conjugative, *n*=9/17 mobilisable, *n*=3/17 non-mobilisable; Figure 2c). AMR genes were carried by plasmids in *n*=6/17 cross-niche groups (Figure 2d), with *n*=5/6 of these groups containing ≥1 beta-lactamase protein encoding gene.

Sharing between BSI and livestock-associated isolates was supported by 8/17 cross-niche groups (*n*=45 plasmids). Of these, *n*=3/8 groups contained BSI/sheep plasmids: one group contained mobilisable Col-type plasmids, the remaining two groups contained conjugative FIB-type plasmids, of which one group contained plasmids carrying the AMR genes *aph(3'')-Ib*, *aph(6)-Id*, bla_{TEM-1}, *dfrA5*, *sul2*, and the other group contained plasmids carrying the MDR efflux pump protein *robA* (see Materials and Methods). A further *n*=2/8 groups contained BSI/pig mobilisable Col-type plasmids, of which one group other carried the AMR genes *aph(3'')-Ib*, *aph(6)-Id*, *dfrA14*, and *sul2*. Lastly, *n*=1/8 groups contained BSI/poultry non-mobilisable Col-type plasmids, *n*=1/8 contained BSI/pig/poultry/influent non-mobilisable Col-type plasmids, and *n*=1/8 contained BSI/cattle/pig/poultry/influent mobilisable Col-type plasmids.

Plasmid clustering reveals a diverse but intertwined population structure across niches

Near-identical plasmids shared across niches are a likely signature of recent transfer events, but we also wanted to examine the wider plasmid population structure. We therefore agnostically clustered all plasmids based on alignment-free sequence similarity (clusters were groups of *n*≥3 plasmids; see Materials and Methods and Appendix 1 Figures 4-5). We defined *n*=247 plasmid clusters with median 5 members (IQR=3-10, range=3-123) recruiting 71% (2,627/3,697) of the plasmids. The remainder were either singletons (i.e., single, unconnected plasmids; 19% [718/3,697]) or doubletons (i.e., pairs of connected plasmids; 10% [352/3,697]). By bootstrapping *b*=1,000 ACs for plasmid clusters, doubletons, and

singletons found against number of isolates sampled (Appendix 1 Figure 6; see Materials and Methods), we estimated that the rarefaction curve had a Heap's parameter γ=0.75, suggesting further isolate sampling would likely detect more plasmid diversity and clusters.

Of the plasmid clusters, *n*=69/247 (28%) plasmid clusters had ≥10 members, representing 50% (1,832/3,697) of all plasmids (Figure 3a). 122/247 (49%) clusters contained BSI plasmids and plasmids from ≥1 other niche. This included 73/247 (30%) of clusters with both BSI and livestock-associated plasmids, representing *n*=38 unique plasmid replicon haplotypes (i.e., combinations of replication proteins) of which only 24% (9/38) were Col-type plasmids, which are often well-conserved and carry few genes(24). 72/247 (29%) of clusters contained both BSI, and influent/effluent/downstream plasmids, reflecting a route of *Enterobacterales* dissemination into waterways. In contrast, only 18/247 (7%) of clusters contained both BSI and upstream waterway plasmids, of which most (13/18 [72%]) also contained influent/effluent/downstream plasmids.

Overall, plasmid clusters scored high homogeneity (*h*) but low completeness (*c*) with respect to biological and ecological characteristics (non-putative PTUs [*h*=0.99, *c*=0.66]; replicon haplotype [*h*=0.92, *c*=0.69]; bacterial host ST [*h*=0.84, *c*=0.14] in Figure 3b; predicted mobility [*h*=0.93, *c*=0.20] in Figure 3c). This indicated that clustered plasmids often had similar characteristics, but the same characteristics were often observed in multiple clusters. When scoring plasmid clusters against broad sampling niche (BSI, livestock-associated or WwTW-associated; Figure 3a), homogeneity was low (*h*=0.12, *c*=0.61), indicating mixed clusters. The imperfect homogeneity is to be anticipated as replicon haplotypes and mobilities can vary within plasmid families, and plasmid families can have diverse host ranges(14).

Plasmids carrying AMR genes were found in 21% (52/247) of the plasmid clusters (i.e., 'antimicrobial resistance gene (ARG)-carrying clusters'), representing *n*=550 plasmids (Figure 3d). Of the ARG-carrying clusters, 92% (48/52) contained at least one beta-lactamase-carrying plasmid (*n*=437 plasmids in total). AMR genes were present in a median proportion 67% of ARG-carrying cluster members (IQR=28-100%, range=3-100%). This highlights that AMR genes are not necessarily widespread on genetically similar plasmids and can be potentially acquired multiple different times through the activity of smaller MGEs (e.g., transposons) or recombination. For example, cluster 12 was a group of *n*=42 conjugative, PTU-F_E plasmids found in BSI, wastewater, and waterways. Of these, 31% (13/42) carried the AMR gene bla_{TEM-1}, and in a range of genetic contexts: *n*=9/13 bla_{TEM-1} genes were found within Tn*3* and *n*=4/13 were carried without a transposase, of which *n*=2/4 were found with the additional AMR genes *aph(6)-Id*, *aph(3'')-Ib*, and *sul2*. AMR genes were disproportionately carried by F-type plasmids (61% [337/550] ARG-carrying cluster plasmids versus 34% [891/2627] of the total clustered plasmids), underlining the known role of F-type plasmids in AMR gene dissemination(13).

The beta-lactamase $bla_{\text{TEM-1}}$ was the most common AMR gene detected (8% of total AMR gene annotations [424/5402]; see Materials and Methods). In terms of sequence length (bp), plasmids made up 3.1% of the overall dataset but 13.8% of the $bla_{\text{TEM-1}}$ -carrying proportion. Of the plasmid clusters, 16% (39/247) carried $bla_{\text{TEM-1}}$, and of these 9 clusters were seen in human BSI and at least one other niche. Plasmid clusters either variably or always carrying $bla_{\text{TEM-1}}$ were strongly associated with BSI ($p<0.01$, Chi-squared test $X^2=8.19$, 33/161 of BSI clusters containing $bla_{\text{TEM-1}}$ vs. 5/86 for non-BSI clusters) and carried a higher number of other AMR genes ($p<0.01$, Wilcoxon text of $bla_{\text{TEM-1}}$-plasmid clusters vs. others; see Appendix 1 Figure 7).

An intertwined ecology of plasmids across human and livestock-associated niches

Plasmids can change their genetic content, particularly when subject to new selective pressures(25,26). Many plasmids have a structure with a 'backbone' of conserved core genes and a 'cargo' of variable accessory genes(12,13,27). We wanted to explore evidence for cross-niche plasmids with minimal mutational evolution in a shared backbone (compatible with ~years of evolutionary separation) but variable accessory gene repertoires.

We first conducted a pangenome-style analysis (see Materials and Methods) on the *n*=69/247 plasmid clusters with ≥10 members. For each cluster, we determined "core" (genes found in ≥95% of plasmids) and "accessory" gene repertoires (found in <95% of plasmids). Within clusters, we found median 9 core genes (IQR=4-53, range=0-219), and median 9 accessory genes (IQR=3-145, range=0-801) (Figure 3e). Core genes comprised a median proportion 42.2% of the total pangenome sizes (IQR=20.9-66.7%). At an individual plasmid level, core genes shared by a cluster comprised a median proportion 62.5% of each plasmid's gene repertoire (IQR=37.4-83.3%; Figure 3e). Putatively conjugative plasmids carried a significantly higher proportion of accessory genes in their repertoires than mobilisable/non-mobilisable plasmids (Kruskal-Wallis test [$H(2)=193.01$, p-value<0.001] followed by Dunn's test).

Using multiple sequence alignments of the core genes within each cluster, we produced maximum likelihood phylogenies (see Supplementary File 1 and Materials and Methods). For this step, we only considered the *n*=62/69 clusters where each plasmid had ≥1 core gene. With the *n*=27/62 clusters that contained both BSI and livestock-associated plasmids, we measured the phylogenetic signal for plasmid sampling niche using Fritz and Purvis' *D* (see

Supplementary File 2 and Materials and Methods). The analysis indicated that the evolutionary history of plasmid clusters is neither strictly segregated by sampling niche nor completely intermixed, but something intermediate.

Alongside the core gene phylogenies, we generated gene repertoire heatmaps (example cluster 2 in Figure 4a-b; all clusters and heatmaps in Supplementary File 1). By visualising the genes in a consensus synteny order (see Materials and Methods), the putative backbone within each plasmid cluster is shown alongside its accessory gene and transposase repertoire. This highlights how plasmids might gain/lose accessory functions within a persistent backbone. Log-transformed linear regression revealed a significant relationship between Jaccard distance of accessory genes presence against core gene cophenetic distance (y=0.080log(x)+0.978, R^2=0.47, F(1,52988)=4.75e4, p-value <0.001; see Appendix 1 Figure 8 and Materials and Methods).

Plasmid dissemination between human and livestock-associated niches is not structured by bacterial host

Alongside vertical inheritance, conjugative and mobilisable plasmids are capable of inter-host transfer, crossing between bacterial lineages, species, up to phyla(14). Phylogenetic analysis can determine whether plasmid evolution between BSI and livestock-associated niches is driven by host clonal expansion or other means, as well as allow us to explore the early emergence of AMR gene carrying plasmids.

As a detailed example, we evaluated the largest plasmid cluster containing both human and livestock-associated plasmids (cluster 2, n=100 members). All plasmids carried at least one F-type replicon and were all putatively conjugative, with 75% (75/100) and 25% (25/100)

assigned PTU-F_E and a putative PTU, respectively. Further, 48% (48/100) plasmids carried bla_{TEM-1}, and 51% (51/100) carried >1 AMR gene. All host chromosomes were *E. coli* except OX-BSI-481_2 (*S. enterica* ST 2998; hereon omitted from the analysis). The n=99 *E. coli* isolates represented six phylogroups: A (5/99), B1 (18/99), B2 (52/99), C (14/99), D (7/99), and G (3/99; see Materials and Methods).

Figure 4b-c shows the plasmid core gene phylogeny (T_{plasmid}) and the *E. coli* host core gene phylogeny ($T_{\text{chromosome}}$). The *E. coli* phylogeny was structured by six clades corresponding to the six phylogroups (see Materials and Methods). We found low congruence between the plasmid core-gene phylogeny and the chromosomal core-gene phylogeny as seen in the central 'tanglegram' (i.e., lines connecting pairs of plasmid and chromosome tips from the same isolate). Additionally, we calculated a Robinson-Foulds distance $RF(T_{\text{plasmid}}, T_{\text{chromosome}})$=162, reflecting a high number of structural differences between the phylogenies (see Materials and Methods). There was some evidence of plasmid structuring by niche (Fritz and Purvis' D=0.24; see Materials and Methods).

Within the plasmid phylogeny, there was a clade of n=44 plasmids (support 100%; circled in grey in Figure 4b) containing both BSI and livestock-associated plasmids, which were within median 4 core gene SNPs of each other (IQR=2-8, range=0-59). Estimating plasmid evolution at an approximate rate of one SNP per year (see Materials and Methods) would give a median time to most recent common ancestor of the backbone at approximately 4 years prior to sampling, consistent with recent movement between human and livestock-associated niches. This plasmid clade was mainly present in phylogroup B2 (20/44), but also A (3/44), B1 (9/44), C (8/44), and D (4/44), suggesting plasmid movement. Further, 77% (34/44) of plasmids within the clade carried bla_{TEM-1} (BSI: 25/34, Livestock-associated: 8/34,

WwTW-associated: 1/34), and 82% (36/44) carried ≥1 AMR gene, highlighting the role of plasmids in cross-niche dissemination of AMR.

To examine the evolution of entire plasmid sequences within the clade, we represented all *n*=44 plasmids as a 'pangraph' (Figure 4d; see Materials and Methods). Briefly, pangraph converts input sequences into a consensus graph, where each sequence is a path along a set of homologous sequence alignments i.e., 'blocks', which in series form 'pancontigs'. Filtering for 'core blocks' (i.e., those found in ≥95% plasmids), we found 4 pancontigs (40 blocks total), with the longest 98,269bp (total length 125,369bp), indicating a putative plasmid backbone (Figure 4e). Then, filtering for 'accessory blocks' (i.e., those found in <95% plasmids), we found 18 pancontigs (39 blocks total), with median length 2,380bp (total length 63,753bp), forming the accessory gene repertoire (Figure 4f). Core and accessory pancontigs contained (22%; 57/261) and (78%; 204/261) of gene annotations, respectively, of which over half encoded hypothetical proteins (51%; 134/261; see Supplementary File 5 and Materials and Methods). Core annotations included replication (*repB*) and conjugation (*finO*, *traI*, *traM*) proteins, whereas accessory gene annotations included antimicrobial resistance (*bcr*, *blaTEM, tetA, tetR*) and mercury resistance (*merA*, *merC*, *merP*, *merT*) proteins. Transposase/insertion sequence annotations were disproportionately found in accessory pancontigs (88%; 38/43) versus core pancontigs (12%; 5/43). This points to a persistent plasmid backbone structure with loss/gain events at particular 'hotspots' as well as rearrangements.

Discussion

Sharing of plasmids between different niches is normally focused on those carrying AMR genes that are of particular current clinical concern, such as extended-spectrum beta-

lactamase (ESBL) or carbapenemase genes, meaning we lack information on the vast 'denominator' of background plasmid sharing, and on the dissemination of other AMR genes which are now widespread in clinical isolates and from which important insights might be gained. By analysing a dataset of *n*=3,697 systematically collected *Enterobacterales* plasmids sampled from human BSI, livestock- and WwTW-associated sources in a geographically and temporally restricted context, we found evidence supporting significant plasmid dissemination across niches, putting those which carry AMR genes of current major clinical concern into context. We found 225 instances of shared, near-identical plasmid groups, 25% of which were found across multiple bacterial STs, 4% across multiple bacterial species, and 8% in both human BSI and ≥1 non-BSI niche. Beyond this near-identical sharing, we analysed 'clusters' of plasmids and found that 73/247 clusters contained plasmids seen in both human BSIs and other contexts. Approximately a fifth (52/247) of plasmid clusters contained plasmids carrying AMR genes (*n*=550 plasmids). Our results suggest the need for broad, unselected, and detailed sampling frames to fully understand plasmid diversity and evolution, and to evaluate the "One Health" risk of AMR associated with plasmid-sharing across niches.

Whilst many plasmid clusters were strongly structured by host phylogeny and isolate source, some plasmids from human BSIs were highly genetically related to those in other niches, including livestock. However, not all of these carried AMR genes. Our results highlight the potential routes for transfer that exist through similar plasmids. However, recovering these instances of putative sharing is a sampling challenge. Accumulation curve analyses suggested increasing the size of our dataset would have led to further near-identical matches at an approximately linear rate, meaning even a dataset of this size captures only a small fraction of the true extent of plasmid sharing between human clinical and other non-human/clinical

niches. This presents a challenge for designing appropriately powered studies. Had we only sampled *n*=100 livestock-associated isolates (i.e., around 20% of our actual sample), there was only a 39% chance that we would have detected ≥5 matches with BSI plasmids (Appendix 1 Figure 9).

Understanding the evolutionary history, distribution, and epidemiology of well-known genes in environmental plasmids may offer insights into the future trajectories of more recently emerged genes. For example, the first plasmid-encoded beta-lactamase to be described was $bla_{\text{TEM-1}}$, identified in 1965 in an *E. coli* isolate in Greece(28) and now widely prevalent in *Enterobacterales*(29). $bla_{\text{TEM-1}}$ has a narrow spectrum of activity and is now less clinically concerning than newer genes which mediate broad-spectrum resistance, but in our dataset $bla_{\text{TEM-1}}$ was strongly associated with plasmid clusters seen in BSI and with the carriage of other AMR genes. $bla_{\text{TEM-1}}$ may continue to play an important role in the spread of AMR-carrying plasmids which can transfer recently emerged genes, and similarities in its association with plasmids and other smaller transposable mobile genetic elements may reflect the future trajectory of other AMR genes of more recent clinical concern such as ESBLs and carbapenemases.

Given that plasmids observed in BSI isolates represent small proportion of human *Enterobacterales* diversity, many more sharing events may occur in the human gut(30) which we only sampled incompletely using wastewater influent as a proxy. The human colon contains around 10^{14} bacteria(31), with large ranges of *Enterobacteriaceae* abundance. Further, even small numbers of across-niche sharing events, such as transfer events of important AMR genes from species-to-species or niche-to-niche, may have significant clinical implications, as has been seen with several important AMR genes globally. Future

366 studies need to carefully consider the limitations of sampling frames in detecting any genetic
367 overlap, given both substantial diversity and the effects of niches and geography(11,16).
368
369 By examining plasmid relatedness compared to bacterial host relatedness in *E. coli*, we
370 demonstrated that plasmids seen across different niches are not necessarily associated with
371 clonal lineages. Using a pangenome-style analysis, we showed that plasmids can share sets of
372 near-identical core genes alongside diverse accessory gene repertoires. While plasmids with
373 more distantly related core genes tended to have dissimilar accessory gene content, plasmids
374 with more closely related core genes shared a wide range of accessory gene content. This
375 would be consistent with a hypothesis of persistent 'backbone' structures gaining and losing
376 accessory functions as they move between hosts and niches. We suggest that this mode of
377 transfer might be worth considering. Evolutionary models for plasmids which can
378 accommodate well-conserved backbone evolution alongside accessory structural changes and
379 gain/loss events are urgently needed. Estimating plasmid evolutionary rates remains a
380 challenge, with little known about appropriate values for mutation rates in plasmids, and even
381 less for non-mutational processes such as gene gain/loss.
382
383 Our study had several limitations. Our non-BSI isolates were not as temporally varied as the
384 BSI isolates, meaning we could not fully explore temporal evolution. Although we evaluated
385 four bacterial genera, 72% (1,044/1,458) of our sequenced isolates were *E. coli*, and so our
386 analyses and findings are particularly focused on this species. Additionally, we did not
387 sample livestock-associated niches densely enough to explore individual livestock types
388 (cattle/pig/poultry/sheep) sharing plasmids with BSI isolates (see Appendix 1 Figure 9).
389 Isolate-based methodologies are limited in evaluating the true diversity of the niches
390 sampled; composite approaches including metagenomics might shed additional insight in

391 future studies. Further, the exact source of an isolate is poorly defined for
392 wastewater/waterway isolates as they act as a confluence of multiple sources, although they
393 represent important niches in their own right. We only analysed plasmids from complete
394 genomes i.e., where the chromosome and all plasmids were circularised, meaning we
395 disregarded ~23% and ~33% of BSI and non-BSI assemblies, respectively. The exclusive use
396 of complete assemblies was to ensure full plasmid sequences could be examined in their full
397 genomic context. We only focused on plasmids as horizontally transmissible elements here;
398 detailed study of other smaller mobile genetic elements across-niches would represent
399 interesting future work. We have also investigated a limited subset of *Enterobacterales*:
400 plasmid sharing likely extends to other bacterial hosts not investigated here. Lastly, our
401 isolate culture methods for livestock-associated samples may not have been as sensitive for
402 the identification of *Klebsiella* spp. as for other *Enterobacterales* such as *Escherichia*, as we
403 did not use enrichment and selective culture on Simmons citrate agar with inositol(32). This
404 limited our ability to study the epidemiology of livestock *Klebsiella* plasmids.
405
406 In conclusion, this study presents to our knowledge the largest evaluation of systematically
407 collected *Enterobacterales* plasmids across human and non-human niches within a
408 geographically and temporally restricted context. Plasmids can clearly disseminate between
409 niches, although this dynamic likely varies by cluster; the overall number of near-identical
410 plasmid groups identified across niches consistent with recent transfer events was 8%
411 (17/225) and influenced by sample size. We demonstrate a likely intertwined ecology of
412 plasmids across human and non-human niches, where different plasmid clusters are variably
413 but incompletely structured and putative ‘backbone’ plasmid structures can rapidly gain and
414 lose accessory genes following cross-niche spread. Future “One Health” studies require dense

and unselected sampling, and complete/near-complete plasmid reconstruction, to appropriately understand plasmid epidemiology across niches.

Materials and Methods

Livestock-associated isolates

n=247 *Enterobacterales* isolates from farm-proximate soils and poultry faeces (*n*=19 farms; *n*=5 cattle, *n*=4 pig, *n*=5 poultry, *n*=5 sheep) were collected and sequenced for this study in 2017-2020. DNA extraction and sequencing was performed as in Shaw *et al.*, 2021(11). Genomes were hybrid assemblies reconstructed using Unicycler(33) (v. 0.4.4; default hybrid assembly parameters except min_component_size 500 and --min_dead_end_size 500). Only complete assemblies (plasmids and chromosomes) were considered (*n*=162/247).

BSI isolates

Sequenced Human BSI Enterobacterales isolates from patients presenting to *n*=4 hospitals within Oxfordshire, UK, September 2008-December 2018, as described in Lipworth *et al*., 2021(34) were also included. Although all patients were sampled in Oxfordshire, a total of *n*=505/738 patients resided in Oxfordshire, *n*=133/738 in surrounding counties, and *n*=100/738 had location information omitted. Only complete assemblies (*n*=738/953 total assembled) were considered.

Other livestock-associated and WwTW-associated isolates

Enterobacterales isolates from faeces from the *n*=14 non-poultry farms and wastewater influent, effluent, and waterways upstream/downstream of effluent outlets surrounding *n*=5 WwTWs, across 3 seasonal timepoints in 2017 (as in (11)) were included. Only complete assemblies (*n*=558/827 total assembled) were considered.

Taxonomic assignment

Chromosome sequence types (STs) were determined with mlst(35) (v. 2.19.0; PubMLST database(36)). For the *n*=11/1,458 chromosomes which could not be typed with mlst, species were determined with the PubMLST 'species ID' web-tool(37), for which all had a support=100, except for *L. nimipressuralis* (support=83). Of these, *n*=5/11 were from BSI, *n*=4/11 from livestock, and *n*=2/11 from effluent/downstream of WwTWs. From the BSI isolates, we also included n=2 *Aeromonas* spp., a non-*Enterobacterales* genus from the wider *Gammaproteobacteria* class.

Chromosome trees

Trees for *E. coli* and *K. pnemoniae* chromosomes were produced using Mashtree(38) on 'accurate' mode (--mindepth 0 --numcpus 12).

PTU classification

Plasmids were assigned a plasmid taxonomic unit (PTU) using COPLA(39) (default parameters except -t circular, -k Bacteria, -p Pseudomonadota, -c Gammaproteobacteria, and -o Enterobacterales)(14). COPLA compares query plasmids to a database of PTU reference plasmids, assigning a PTU when both (i) the ANI>0.7 along 50% of the length of the smallest plasmid in the comparison, and (ii) a graph-neighbouring condition to existing PTU clusters is satisfied. The COPLA reference database contains over 10,000 curated, non-redundant plasmids retrieved from the 84th NCBI RefSeq database in 2017(40). We contextualised our plasmids within known plasmid diversity using COPLA to determine each plasmid's 'plasmid taxonomic unit' (PTU; see Materials and Methods), which is designed to be equivalent to a 'species' concept for plasmids(39). Briefly, COPLA classifies query plasmids based on average nucleotide identity (ANI) against a non-redundant reference plasmid database where most plasmids have been assigned to a reference PTU[40]. Within our sample, 64% (2,369/3,697) plasmids were assigned a PTU and 4% (135/3,697) a putative PTU (i.e., the query plasmid was clustered with 3 unclassified reference plasmids). This is consistent

with a previous COPLA analysis of 1,000 *Enterobacterales* plasmids which found that 63% were classified into a PTU (39). The remaining 32% (1,193/3,697) of plasmids were unclassified (i.e., connected set with less than 4 plasmids) highlighting the previously unsampled plasmid diversity within our dataset. In total, we found *n*=67 known PTUs, containing a median 9 plasmids (IQR=4-30, range=1-556), where the largest assigned PTU (556/2,504) was PTU-F_E, corresponding to F-type *Escherichia* plasmids(13,24). The proportion of unclassified plasmids was higher in environmental/livestock samples (33%; 385/1,155) versus BSI samples (26%; 485/1,880), emphasising the underrepresentation of non-human plasmids in reference plasmid databases.

Plasmid annotation

All plasmids were annotated with Prokka(41) (v. 1.14.5) with default parameters. For replicon typing, Abricate(42) (v. 1.0.0) was used with the PlasmidFinder(43), ISfinder(44), and BacMet(45) databases with default parameters and output filtered for 80% minimum coverage. For annotating AMR genes, NCBI Antimicrobial Resistance Gene Finder (AMRFinderPlus)(46) (v. 3.10.18) was used with default parameters. For putative plasmid mobilities, we used MOB-typer from MOB-suite(47) (v. 3.03) with default parameters, which predicts mobility based on annotations of MOB-typer predicts mobility based on of annotations of relaxase (*mob*), mating pair formation (MPF) complex, and *oriT* genes. Briefly, a plasmid is putatively labelled conjugative if it has both relaxase and MPF, mobilisable if it has either relaxase or *oriT* but no MPF, and non-mobilisable if it has no relaxase and *oriT*.

Near-identical plasmid screening

Groups of near-identical plasmids were detected as connected components in a plasmid-plasmid network with Mash distance(48) (v. 2.3; default parameters except sketch size -s 1000000) weighted edges, at a threshold $d<0.0001$. Briefly, Mash distance estimates an

evolutionary distance on a reduced-length MinHash sketch of the sequences. Since Mash is a probabilistic estimate of evolutionary distance, we confirmed the probability of seeing any of our pairwise Mash distances in the near-identical groups by chance was 0. For whole genomes, Mash distance has a strong positive correlation with ANI (49). We also required the shortest plasmid to be within 1% length (bp) of the longest plasmid, to account for assembly errors. Network analysis was performed using the igraph(50) library (v. 1.2.7) in R.

The stringency of a *k*-mer-based distance threshold for near-identical plasmid clustering is equivalent to a threshold on the Jaccard index (i.e., rearranging the mash distance calculation ($d = \frac{-1}{k}\ln(\frac{2j}{1+j})$ with d=10^{-4} and k=21 gives a Jaccard index threshold of j=0.9958). The effect of this threshold varies with plasmid size: at very small plasmid sizes, clusters contain only identical plasmids because the presence of a single SNP means plasmids are placed in different clusters. For example, two 1,552bp plasmids with a single SNP (e.g., RHB03-C05_6 and RHB02-C22_6) will have a mash distance of d=5.0x10^{-4} (>10^{-4} threshold). In contrast, at length=150kb a single SNP (not at the start/end of the plasmid) would lead to d=5.6x10^{-6} (<< 10^{-4} threshold); even two 150kbp plasmids with ~30 SNPs would have d≈2x10^{-4} (>10^{-4} threshold) and so be split into near-identical plasmids. Our analysis of plasmid sharing is therefore maximally conservative at small plasmid sizes but remains highly conservative for large plasmids.

Accumulation and rarefaction curves

To generate an accumulation curve, isolates were sampled without replacement in a random order. For each isolate, the new plasmid diversity was recorded. For Appendix 1 Figure 3, we recorded the number of new near-identical plasmid groups and singletons. For Appendix 1 Figure 9, we recorded the number of near-identical matches with BSI plasmids from only environmental/livestock isolates. For Appendix 1 Figure 6, we recorded the number of new clusters, doubletons, and singletons. A bootstrapped average of b=1,000 accumulation curves

514 was plotted for the rarefaction curve. The bootstraps were also used to estimate Heap's
515 parameter (γ) by fitting a linear regression to log-log transformed data using standard R
516 libraries. For $\gamma<0$, it is possible to sample the entire diversity, and for $1>\gamma>0$, the diversity
517 will increase with every additional sample(51).

518 **Plasmid similarity**

519 Plasmid Jaccard index (*JI*) was calculated using Mash(48) (v. 2.3; default parameters except
520 sketch size -s 1000000). The Jaccard index (*JI*), given by

$$JI(A,B) = \frac{|A \cup B|}{|A \cap B|}$$

521

522 where *A*, *B* are the sets of *k*-mers of plasmids *a*, *b*, respectively. This measures extent of *k*-
523 mer sharing between plasmids, range=0-1, where 1 indicates an identical *k*-mer repertoire.
524 Since the sketch size was larger than the plasmid lengths (except for one plasmid in the
525 dataset, OX-ENV-67_2, which was larger than 1Mbp at 1,310,597bp and was not clustered;
526 the next smallest was OX-WTW-80_2 at 394,284bp), the calculated Jaccard indices were
527 almost always exact.

528 **Plasmid network and clustering**

529 The determination of the plasmid-plasmid network, threshold, and clusters could be achieved
530 with several alternative methodologies. Plasmid networks have previously been constructed
531 by full sequence alignments(52), annotated genes(53), and alignment-free Mash
532 distances(13,54,55). We chose to use the Jaccard index of entire plasmid *21*-mer distributions
533 to capture coding sequences, their immediate contexts(56,57), and intergenic regions(58,59),
534 all of which have known importance to bacterial evolution. Further, our contained previously
535 unsampled diversity as seen by the PTU analysis, and because reference-based classifications
536 such as MOB and replicon typing schemes are known to be incongruent(60) or unreliable:
537 16% (602/3,697) of our plasmids had an unidentifiable replicon type, which is not
538 uncommon(24). The evolutionary histories of plasmids can incorporate multiple gain, loss,

and rearrangement events in addition to mutations(61), and as such, traditional measures of genetic relatedness (e.g., single nucleotide variant [SNV] thresholds) used for genomic epidemiology of whole genomes are likely less appropriate here. These similarities formed the edge-weights in a plasmid-plasmid network, which was subsequently thresholded to sparsify the network and allow the detection of clusters.

Network thresholding to some extent depends subjectively on the dataset, with trade-offs between successfully revealing the underlying structure of plasmid relationships without excessively separating relatives. We chose a data-driven threshold as adopted by Ledda *et al.*, 2018(53) for their plasmid network, which examined the evolution of connected components within the network. This ensured the threshold was chosen where the regime of connected component evolution approximately stabilises, minimising excessive network breakup. The threshold was chosen at JI=0.5, meaning that edges between plasmids with JI<0.5 were deleted from the network. From this threshold onwards, both the number of connected components and the number of singletons steadily increased at a similar rate (Appendix 1 Figure 4). This regime indicates an approximately stable non-singleton structure from JI=0.5 onwards.

We defined plasmid clusters as groups of n≥3 plasmids with high within-cluster-similarities and low between-cluster-similarities. Plasmid clusters were detected using the Louvain algorithm which optimises the network modularity by iterative expectation-maximisation(62). This aims to maximise the density of edges within clusters against edges between clusters. Though non-deterministic, the Louvain algorithm showed low variation in cluster distribution over 50 runs, consistent with reproducible segregation of plasmids in clusters (range of clusters detected: 245-247; Appendix 1 Figure 5). The algorithm was

implemented using the python-louvain (v. 0.16) Python module. Although the algorithm is non-deterministic, multiple runs demonstrated minimal variation at our chosen network threshold. Overall, these approaches add to the growing literature describing suitable methodologies for clustering plasmids.

Near-identical plasmid groups were also included in the wider cluster analysis, as many were cross-compartmental and found across bacterial hosts (see earlier, Figure 2). Of the n=194/225 groups which were clustered, 100% (194/194) had all members fall within the same plasmid cluster, with *n*=30/247 clusters containing multiple near-identical plasmid groups. Only 6% (14/247) of plasmid clusters comprised exclusively near-identical plasmid groups, suggesting that near-identical groups of plasmids often have nearby genetically related plasmids. Examining the entire PTU distribution within clusters, most contained at least one unclassified plasmid (51%; 127/247) or plasmid assigned a putative PTU (9%; 23/247). However, many clusters exclusively contained just one known PTU (42%; 105/247).

Cluster homogeneity and completeness

Homogeneity (h) and completeness (c) are dual conditional entropy-based measures, independent of cluster and metadata label distributions(63). A clustering satisfies homogeneity (h =1) if all cluster members have the same metadata label-type. Consider a network with N nodes, partitioned by a set of metadata labels, $M = \{m_i | i = 1, \dots, n\}$, and a set of communities, $C = \{c_j | j = 1, \dots, m\}$. Let $A = \{a_{ij}\}$ represent the ij^{th} entry in the contingency table of partitions. Hence, a_{ij} counts the number of nodes with label m_i in community c_j. We then say

$$\mathrm{h} = \begin{cases} 1 & \text{if } H(M,C) = 0 \\ 1 - \dfrac{H(M \mid C)}{H(M)} & \text{else} \end{cases}$$

where

$$H(M \mid C) = -\sum_{c=1}^{|C|} \sum_{m=1}^{|M|} \frac{a_{mc}}{N} \log \frac{a_{mc}}{\sum_{c=1}^{|M|} a_{mc}}$$

and

$$H(M) = -\sum_{m=1}^{|M|} \frac{\sum_{c=1}^{|C|} a_{mc}}{n} \log \frac{\sum_{c=1}^{|C|} a_{mc}}{n}$$

are the conditional entropy of the metadata given the clusters and the entropy of the clusters, respectively $H(M|C) = 0$ when the cluster partition coincides with the metadata partition, and no new information is added. A cluster partition satisfies completeness (c =1) if all instances of a metadata label-type are assigned the same cluster. Completeness is defined dually by

$$c = \begin{cases} 1 & \text{if } H(C, M) = 0 \\ 1 - \dfrac{H(C \mid M)}{H(C)} & \text{else} \end{cases}$$

The measures were calculated using the clver library (v. 0.1.1) in R.

Cluster pangenome analysis

Cluster pangenomes were generated using Panaroo(64) (v. 1.2.9) with parameters default except --clean-mode sensitive, --aligner mafft, -a core, and --core_threshold 0.95. For core gene alignments, the threshold was set at minimum 95% presence amongst clustered plasmids, whereby they were aligned using MAFFT(65) (v. 7.407) with default parameters. An identical approach was taken for the host chromosome phylogeny in Figure 4. The median length of plasmids within a cluster was positively correlated with number of core genes (R=0.85, t=13.4, p-value<2.2e-16) and total pangenome size (R=0.87, t=14.6, p-value<2.2e-16).

Plasmid core gene phylogenies

Maximum likelihood core-gene phylogenies were generated using IQ-Tree(66) (v. 2.0.6) with parameters -m GTR+F+I+G4 -keep-ident -T 2 -B 1000. The substitution model used was general time reversible (GTR) using empirical base frequencies form the alignment (F),

allowing for invariable sites (I) and variable rates of substitution (G4). We used *n*=1000 ultrafast bootstraps (B 1000; see Minh *et al.*, 2013(67)) to visually inspect larger clades for support. Briefly, 95% support approximates a 95% probability that the clade is genuine. Only the *n*=62/69 clusters (excluding 6, 8, 26,2 9,3 2, 40, and 65) where every plasmid carried at least 1 core gene were analysed. Phylogenies were primarily plotted using the R library ggtree(68).

Fritz and Purvis' *D*

Fritz and Purvis' *D* measures phylogenetic signal for binary traits(69). First, we calculate the character state changes required to observe our phylogeny (d_{obs}). To account for phylogeny size and prevalence, d_{obs} is standardised under the two null models (i) tip labels are random permuted (d_r), and (ii) tip labels are distributed under the expectation of a Brownian motion model of evolution (d_b). Then, we define

$$D = (d_{obs} - \overline{d_b})/(\overline{d_r} - \overline{d_b}).$$

Hence, for $D \approx 1$, d_{obs} follows d_r more closely, and for $D \approx 0$, d_{obs} follows d_b more closely. We calculated d_{obs} *n*=10,000 times and averaged the result, as well as calculate *p*-values for significant deviation from d_r or d_b. *D* was implemented using the R library caper(70). Fritz and Purvis' *D* is normally used for cross-species analysis so is not benchmarked for plasmids. Results for phylogenies with less than 25 tips should be viewed more conservatively due to reduced statistical power in these instances.

We considered the binary 'trait' of human or livestock-associated isolate and estimated *D* with *n*=10,000 permutations. We found 42% (11/26) clusters had $D>0.5$ (see Supplementary File 2). However, only 23% (6/26) of phylogenies were significantly different (p-value<0.05) from the conserved null model, compared to 50% (13/26) significantly different from the random null model.

Consensus gene synteny heatmaps

636 For each cluster, we first generated a list of every possible pair of genes in the pangenome.
637 Then for each plasmid, we counted the distance between these pairs, modulo the number of
638 genes in the plasmid. If a gene was absent in a plasmid, NA was used. We then calculated the
639 median of these values across all plasmids in the cluster. We then built a dendrogram from a
640 hierarchal clustering of the median distances. The order of the tip labels in the dendrogram
641 were then used as the 'consensus gene synteny'.

642 **Accessory gene distances**

643 Plasmid accessory gene distances were calculated using pairwise Jaccard distances on gene
644 presence-absences matrices. For plotting the cluster-wise plasmid core gene cophenetic
645 distance against accessory gene presence-absence Jaccard distance, only the n=26/62 clusters
646 with at least 50 accessory genes were plotted. The log-transformed linear regression of
647 Jaccard distance of accessory genes presence against core gene cophenetic distance was fitted
648 in R with standard libraries.

649 **Chromosome core gene phylogeny**

650 An identical approach was taken to the plasmid phylogenies. *E. coli* phylogroups were typed
651 using EzClermont(71) (v. 0.7.0) with default parameters. Robinson-Foulds distance was
652 calculated using the R library phangorn(72).

653 **Plasmid mutation rate**

654 Mutation rates per base pair in microbes typically arise from DNA replication and tend to be
655 below $m=10^{-9}$ per site per generation(73) or perhaps as low as 10^{-10} per site per
656 generation(74,75). For a plasmid of size L, one therefore expects L x m mutations per plasmid
657 per generation. For example, if the plasmid has $L=10^5$ then in each generation 1 in 10,000
658 plasmids will gain a mutation. The generation time of *E. coli* per day in the human gut has
659 been estimated to be between 6 and 20 generations per day(76). For large plasmids that exist
660 at a copy number of ~1, the plasmid generation time is the cell generation time. More

generally, for a plasmid copy number *p* the number of replications of the plasmid expected for a given number of cell generations *g* will be *p* x *g* (assuming that plasmid copies are simply and linearly related to the realised number of replications per cell). A crude estimate for the expected mutation rate per time period for a plasmid is therefore given by *L* x *m* x *p* x *g*. For a plasmid of *L*=100 kbp and *p*=1, assuming *m*=[0.1-1]x10^{-9} per site per generation and *g*=[6-20]x365 per year, one would expect it to accumulate ~0.5 mutations a year (between ~0.02-0.7 depending on assumptions). One obtains the same result for *L*=10 kbp and *p*=10. There is a strong inverse correlation between plasmid size and copy number. This suggests that a suitable upper bound for the expected number of mutations for a typical plasmid per year (under neutral evolution) is of the order of magnitude of 1 SNP a year. This rough 'SNPs and years' rule-of-thumb appears consistent with known empirical results. For example: 100kbp I1-type *Shigella* plasmids isolated between 2007-2010 in Vietnam were separated by at most 2 SNPs(77); 30kbp X4-type plasmids carrying *mcr-1* isolated between 2016-2018 in China were separated by most 4 SNPs(18) (analysis not shown); 63.5 kbp pOXA48-like plasmids (*n*=202) in *Klebsiella pneumoniae* collected across Europe between 2013 and 2014 as part of EUSCAPE were overwhelmingly within 2 SNPs of each other (176/202)(78); the same was true of 45.4kbp IncX3 plasmids (*n*=135) from the EUSCAPE dataset (all were within 6 SNPs of each other; see Figure 4 of that paper); and also of 113.4 kbp pKpQIL-like plasmids (*n*=91) from the EUSCAPE dataset – although a minority of these plasmids were separated by up to 20 SNPs, which seems suggestive of either ancestry before the two-year sampling frame or recombination.

Pangraph analysis

We used pangraph(79) (v. 0.5.0) to build a pangraph of the clade within plasmid cluster 2, using the --circular flag and otherwise default parameters. We removed duplicated blocks from the pangraph. We used pangraph export (--edge-minimum length 0, default parameters)

to export the graph to GFA format and then visualised this using Bandage(80).

Supplementary File 5 used Prokka annotations (see above) of the core and accessory pancontigs.

Data visualisation

Plots were primally produced using the R library ggplot2(81), with additional graphics in BioRender(82).

Data availability

Study metadata is provided in Supplementary File 3. Accessions for poultry and environmental soil isolate reads are given in Supplementary File 4, and assemblies will shortly be made available on NCBI. Accessions for reads for the BSI isolates can be found in Lipworth *et al.*, 2021(34) (BioProject PRJNA604975) and for reads and assemblies for previously assembled REHAB isolates in Shaw *et al.*, 2021(11) (BioProject PRJNA605147).

Code availability

Analysis scripts can be found in the GitHub repository https://github.com/wtmatlock/oxfordshire-overlap.

REHAB Consortium.

Manal AbuOun[2], Muna F. Anjum[2], Mark J. Bailey[3], Brett H[8], Mike J. Bowes[3], Kevin K. Chau[1], Derrick W. Crook[1,6,7], Nicola de Maio[1], Nicholas Duggett[2], Daniel J. Wilson[1,9], Daniel Gilson[2], H. Soon Gweon[3,4], Alasdair Hubbard[10], Sarah J. Hoosdally[1], William Matlock[1], James Kavanagh[1], Hannah Jones[2], Timothy E. A. Peto[1,6,7], Daniel S. Read[3], Robert Sebra[5], Liam P. Shaw[1], Anna E. Sheppard[1,6], Richard P. Smith[2], Emma Stubberfield[2], Nicole Stoesser[1,6,7], Jeremy Swann[1], A. Sarah Walker[1,6,7], Neil Woodford[11]

[1] Nuffield Department of Medicine, University of Oxford, Oxford, UK

[2] Animal and Plant Health Agency, Weybridge, Addlestone, UK

[3] UK Centre for Ecology & Hydrology, Wallingford, UK

[4] University of Reading, Reading, UK

[5] Icahn Institute of Data Science and Genomic Technology, Mt Sinai, NY, USA

[6] NIHR HPRU in healthcare-associated infection and antimicrobial resistance, University of Oxford, Oxford, UK

[7] NIHR Oxford Biomedical Research Centre, University of Oxford, Oxford, UK

[8]Thames Water Utilities, Clearwater Court, Vastern Road, Reading, UK

[9]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, UK

[10]Department of Tropical Disease Biology, Liverpool School of Tropical Medicine, Liverpool, UK

[11]Antimicrobial Resistance and Healthcare Associated Infections (AMRHAI) Reference Unit, National Infection Service, Public Health England, London, United Kingdom

Declarations of interests. The authors declare no interests.

Role of the funding source.

This work was funded by the Antimicrobial Resistance Cross-council Initiative supported by the seven research councils [grant NE/N019989/1]. The UKCEH component of the REHAB consortium was supported by the Natural Environment Research Council (NERC) [grant NE/N019660/1]. DWC, SG, TEAP and NS are supported by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare-Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with Public Health England (PHE) [grant HPRU-2012–10041 and NIHR200915]. DWC and TEAP are also supported by the NIHR Oxford Biomedical Research Centre. The computational aspects of this research were funded from the NIHR Oxford BRC with additional support from a Wellcome Trust Core Award Grant [grant 203141/Z/16/Z]. 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. WM and KKC are supported by a scholarship from the Medical Research Foundation National PhD Training Programme in Antimicrobial Resistance Research (MRF-145-0004-TPG-AVISO). NS is an Oxford Martin Fellow and a Senior NIHR BRC Oxford Fellow. LPS is a Sir Henry Wellcome Postdoctoral Fellow funded by Wellcome (grant 220422/Z/20/Z).

Bibliography

1. Linh TD, Thu NH, Shibayama K, Suzuki M, Yoshida LM, Thai PD, et al. Expansion of KPC-producing Enterobacterales in four large hospitals in Hanoi, Vietnam. J Glob Antimicrob Resist. 2021;27.
2. Kraftova L, Finianos M, Studentova V, Chudejova K, Jakubu V, Zemlickova H, et al. Evidence of an epidemic spread of KPC-producing Enterobacterales in Czech hospitals. Sci Rep. 2021;11(1).
3. Cahill N, O'Connor L, Mahon B, Varley Á, McGrath E, Ryan P, et al. Hospital effluent: A reservoir for carbapenemase-producing Enterobacterales? Science of the Total Environment. 2019;672.
4. Subramanya SH, Bairy I, Metok Y, Baral BP, Gautam D, Nayak N. Detection and characterization of ESBL-producing Enterobacteriaceae from the gut of subsistence farmers, their livestock, and the surrounding environment in rural Nepal. Sci Rep. 2021;11(1).
5. Abuoun M, Jones H, Stubberfield E, Gilson D, Shaw LP, Hubbard ATM, et al. A genomic epidemiological study shows that prevalence of antimicrobial resistance in enterobacterales is associated with the livestock host, as well as antimicrobial usage. Microb Genom. 2021;7(10).
6. Díaz-Gavidia C, Barría C, Rivas L, García P, Alvarez FP, González-Rocha G, et al. Isolation of Ciprofloxacin and Ceftazidime-Resistant Enterobacterales From Vegetables and River Water Is Strongly Associated With the Season and the Sample Type. Front Microbiol. 2021;12.
7. Buchy P, Ascioglu S, Buisson Y, Datta S, Nissen M, Tambyah PA, et al. Impact of vaccines on antimicrobial resistance. Vol. 90, International Journal of Infectious Diseases. 2020.

8. Ruppé E, Cherkaoui A, Charretier Y, Girard M, Schicklin S, Lazarevic V, et al. From genotype to antibiotic susceptibility phenotype in the order Enterobacterales: a clinical perspective. Clinical Microbiology and Infection. 2020;26(5).
9. Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A, Anson LW, et al. Nested Russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene blakpc. Antimicrob Agents Chemother. 2016;60(6).
10. Che Y, Yang Y, Xu X, Brinda K, Polz MF, Hanage WP, et al. Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. Proc Natl Acad Sci U S A. 2021;118(6).
11. Shaw LP, Chau KK, Kavanagh J, AbuOun M, Stubberfield E, Gweon HS, et al. Niche and local geography shape the pangenome of wastewater-and livestock-associated Enterobacteriaceae. Sci Adv. 2021;7(15).
12. Orlek A, Stoesser N, Anjum MF, Doumith M, Ellington MJ, Peto T, et al. Plasmid classification in an era of whole-genome sequencing: Application in studies of antibiotic resistance epidemiology. Vol. 8, Frontiers in Microbiology. 2017.
13. Matlock W, Chau KK, AbuOun M, Stubberfield E, Barker L, Kavanagh J, et al. Genomic network analysis of environmental and livestock F-type plasmid populations. ISME Journal. 2021;15(8).
14. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. Nat Commun. 2020;11(1).
15. Mounsey O, Schubert H, Findlay J, Morley K, Puddy EF, Gould VC, et al. Limited phylogenetic overlap between fluoroquinolone-resistant Escherichia coli isolated on dairy farms and those causing bacteriuria in humans living in the same geographical region. Journal of Antimicrobial Chemotherapy. 2021;76(12).

16. Hanage WP. Two health or not two health? That is the question. mBio. 2019;10(2):e00550-19.
17. Ludden C, Raven KE, Jamrozy D, Gouliouris T, Blane B, Coll F, et al. One health genomic surveillance of escherichia coli demonstrates distinct lineages and mobile genetic elements in isolates from humans versus livestock. mBio. 2019;10(1).
18. Shen C, Zhong LL, Yang Y, Doi Y, Paterson DL, Stoesser N, et al. Dynamics of mcr-1 prevalence and mcr-1-positive Escherichia coli after the cessation of colistin use as a feed additive for animals in China: a prospective cross-sectional and whole genome sequencing-based molecular epidemiological study. Lancet Microbe. 2020;1(1):e34–43.
19. Cherak Z, Loucif L, Moussi A, Rolain JM. Epidemiology of mobile colistin resistance (mcr) genes in aquatic environments. Vol. 27, Journal of Global Antimicrobial Resistance. 2021.
20. Carlos Bastidas-Caldes, Jacobus H. de Waard, María Soledad Salgado, María José Villacís, Marco Coral-Almeida, Yoshimasa Yamamoto, et al. Worldwide Prevalence of mcr-mediated Colistin-Resistance Escherichia coli in Isolates of Clinical Samples, Healthy Humans, and Livestock—A Systematic Review and Meta-Analysis. Pathogens. 2022 Jun;11(6).
21. Hilpert C, Bricheux G, Debroas D. Reconstruction of plasmids by shotgun sequencing from environmental DNA: Which bioinformatic workflow? Brief Bioinform. 2021;22(3).
22. Wang R, van Dorp L, Shaw LP, Bradley P, Wang Q, Wang X, et al. The global distribution and spread of the mobilized colistin resistance gene mcr-1. Nat Commun. 2018;9(1):1–9.

817 23. Sekizuka T, Matsui M, Yamane K, Takeuchi F, Ohnishi M, Hishinuma A, et al.
818 Complete sequencing of the bla NDM-1-positive IncA/C plasmid from Escherichia
819 coli ST38 isolate suggests a possible origin from plant pathogens. PLoS One.
820 2011;6(9).
821 24. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B, Guerra
822 B, et al. Plasmids carrying antimicrobial resistance genes in Enterobacteriaceae.
823 Journal of Antimicrobial Chemotherapy. 2018;73(5).
824 25. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á.
825 Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Vol. 19,
826 Nature Reviews Microbiology. 2021.
827 26. Pesesky MW, Tilley R, Beck DAC. Mosaic plasmids are abundant and unevenly
828 distributed across prokaryotic taxa. Plasmid. 2019;102.
829 27. Coluzzi C, Garcillán-Barcia MP, de La Cruz F, Rocha EPC. Evolution of Plasmid
830 Mobility: Origin and Fate of Conjugative and Nonconjugative Plasmids. Mol Biol
831 Evol. 2022;39(6).
832 28. Datta N, Kontomichalou P. Penicillinase synthesis controlled by infectious R factors in
833 enterobacteriaceae. Nature. 1965;208(5007).
834 29. Bush K, Bradford PA. Epidemiology of β-lactamase-producing pathogens. Vol. 33,
835 Clinical Microbiology Reviews. 2020.
836 30. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human
837 gut bacterial genome and culture collection for improved metagenomic analyses. Nat
838 Biotechnol. 2019;37(2).
839 31. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria
840 Cells in the Body. PLoS Biol. 2016;14(8).

841 32. Rodrigues C, Hauser K, Cahill N, Ligowska-Marzęta M, Centorotola G, Cornacchia A,
842 et al. High Prevalence of Klebsiella pneumoniae in European Food Products: a
843 Multicentric Study Comparing Culture and Molecular Detection Methods. Microbiol
844 Spectr. 2022;10(1).
845 33. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome
846 assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13(6).
847 34. Lipworth S, Vihta KD, Chau K, Barker L, George S, Kavanagh J, et al. Ten-year
848 longitudinal molecular epidemiology study of Escherichia coli and Klebsiella species
849 bloodstream infections in Oxfordshire, UK. Genome Med. 2021;13(1).
850 35. Torsten Seeman. mlst. https://github.com/tseemann/mlst; 2017.
851 36. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at
852 the population level. BMC Bioinformatics. 2010;11.
853 37. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics:
854 BIGSdb software, the PubMLST.org website and their applications [version 1;
855 referees: 2 approved]. Wellcome Open Res. 2018;3.
856 38. Katz L, Griswold T, Morrison S, Caravas J, Zhang S, Bakker H, et al. Mashtree: a
857 rapid comparison of whole genome sequence files. J Open Source Softw. 2019;4(44).
858 39. Redondo-Salvo S, Bartomeus-Peñalver R, Vielva L, Tagg KA, Webb HE, Fernández-
859 López R, et al. COPLA, a taxonomic classifier of plasmids. BMC Bioinformatics.
860 2021;22(1).
861 40. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): A curated
862 non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids
863 Res. 2007;35(SUPPL. 1).
864 41. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics.
865 2014;30(14).

42. Torsten Seeman. Abricate. https://github.com/tseemann/abricate; 2015.
43. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, et al. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58(7).
44. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34(Database issue).
45. Pal C, Bengtsson-Palme J, Rensing C, Kristiansson E, Larsson DGJ. BacMet: Antibacterial biocide and metal resistance genes database. Nucleic Acids Res. 2014;42(D1).
46. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11(1).
47. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom. 2018;4(8).
48. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: Fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1).
49. Figueras MJ, Beaz-Hidalgo R, Hossain MJ, Liles MR. Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis. Genome Announc. 2014;2(6).
50. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Systems. 2006;Complex Sy(1695).
51. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Vol. 11, Current Opinion in Microbiology. 2008.

52. Yamashita A, Sekizuka T, Kuroda M. Characterization of antimicrobial resistance dissemination across plasmid communities classified by network analysis. Pathogens. 2014;3(2).
53. Branger C, Ledda A, Billard-Pomares T, Doublet B, Fouteau S, Barbe V, et al. Extended-spectrum β-lactamase-encoding genes are spreading on a wide range of escherichia coli plasmids existing prior to the use of third-generation cephalosporins. Microb Genom. 2018;4(9).
54. Acman M, van Dorp L, Santini JM, Balloux F. Large-scale network analysis captures biological features of bacterial plasmids. Nat Commun. 2020;11(1).
55. Jesus TF, Ribeiro-Gonçalves B, Silva DN, Bortolaia V, Ramirez M, Carriço JA. Plasmid ATLAS: Plasmid visual analytics and identification in high-Throughput sequencing data. Nucleic Acids Res. 2019;47(D1).
56. Matlock W, Lipworth S, Constantinides B, Peto TEA, Walker AS, Crook D, et al. Flanker: a tool for comparative genomics of gene flanking regions. Microb Genom. 2021;7(9).
57. Arcilla MS, van Hattem JM, Matamoros S, Melles DC, Penders J, de Jong MD, et al. Dissemination of the mcr-1 colistin resistance gene. Lancet Infect Dis. 2016;16(2):147–9.
58. Zhi S, Li Q, Yasui Y, Edge T, Topp E, Neumann NF. Assessing host-specificity of Escherichia coli using a supervised learning logic-regression-based analysis of single nucleotide polymorphisms in intergenic regions. Mol Phylogenet Evol. 2015;92.
59. Delihas N. Intergenic regions of Borrelia plasmids contain phylogenetically conserved RNA secondary structure motifs. BMC Genomics. 2009;10.

60. Orlek A, Phan H, Sheppard AE, Doumith M, Ellington M, Peto T, et al. Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. Plasmid. 2017;91.
61. Kizny Gordon A, Phan HTT, Lipworth SI, Cheong E, Gottlieb T, George S, et al. Genomic dynamics of species and mobile genetic elements in a prolonged blaIMP-4-associated carbapenemase outbreak in an Australian hospital. Journal of Antimicrobial Chemotherapy. 2020;75(4).
62. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008;2008(10).
63. Rosenberg A, Hirschberg J. V-Measure: A conditional entropy-based external cluster evaluation measure. In: EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007.
64. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020;21(1).
65. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30(4).
66. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37(5).
67. Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 2013;30(5).

68. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8(1).
69. Fritz SA, Purvis A. Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. Conservation Biology. 2010;24(4).
70. Orme D. The caper package : comparative analysis of phylogenetics and evolution in R. R package version 05, 2. 2013;
71. Waters NR, Abram F, Brennan F, Holmes A, Pritchard L. Easy phylotyping of Escherichia coli via the EzClermont web app and command-line tool. Access Microbiol. 2020;2(9).
72. Schliep KP. phangorn: Phylogenetic analysis in R. Bioinformatics. 2011;27(4).
73. Drake JW. A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci U S A. 1991;88(16).
74. Foster PL, Lee H, Popodi E, Townes JP, Tang H. Determinants of spontaneous mutation in the bacterium Escherichia coli as revealed by whole-genome sequencing. Proc Natl Acad Sci U S A. 2015;112(44).
75. Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S, et al. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc Natl Acad Sci U S A. 2013;110(1).
76. Ghalayini M, Launay A, Bridier-Nahmias A, Clermont O, Denamur E, Lescat M, et al. Evolution of a dominant natural isolate of Escherichia coli in the human gut over the course of a year suggests a neutral evolution with reduced effective population size. Appl Environ Microbiol. 2018;84(6).

962 77. Holt KE, Nga TVT, Thanh DP, Vinh H, Kim DW, Tra MPV, et al. Tracking the
963 establishment of local endemic populations of an emergent enteric pathogen. Proc Natl
964 Acad Sci U S A. 2013;110(43).
965 78. David S, Cohen V, Reuter S, Sheppard AE, Giani T, Parkhill J, et al. Integrated
966 chromosomal and plasmid sequence analyses reveal diverse modes of carbapenemase
967 gene spread among Klebsiella pneumoniae. Proc Natl Acad Sci U S A. 2020;117(40).
968 79. Nicholas Noll MMRAN. PanGraph: scalable bacterial pan-genome graph construction.
969 bioRxiv. 2022 Feb 24;
970 80. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of de
971 novo genome assemblies. Bioinformatics. 2015;31(20).
972 81. Gómez-Rubio V. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition) . J Stat
973 Softw. 2017;77(Book Review 2).
974 82. Munday E. BioRender. N/a. 2021.

975

976

Figure 1. A diverse sample of geographically and temporally restricted *Enterobacterales* **(a)** Number of chromosomes and plasmids by niche, stratified by isolate genus. **(b)** Map of approximate, relative distances between sampling sites, coloured by niche (human bloodstream infection [BSI], livestock-associated (cattle, pig, poultry, and sheep faeces, soils nearby livestock sites), and wastewater treatment work (WwTW)-associated sources (influent, effluent, waterways upstream/downstream of effluent outlets). Number in circles indicates how many of the *n*=1,458 isolates are from that location. **(c)** Sampling timeframe for BSI and REHAB (non-BSI) isolates.

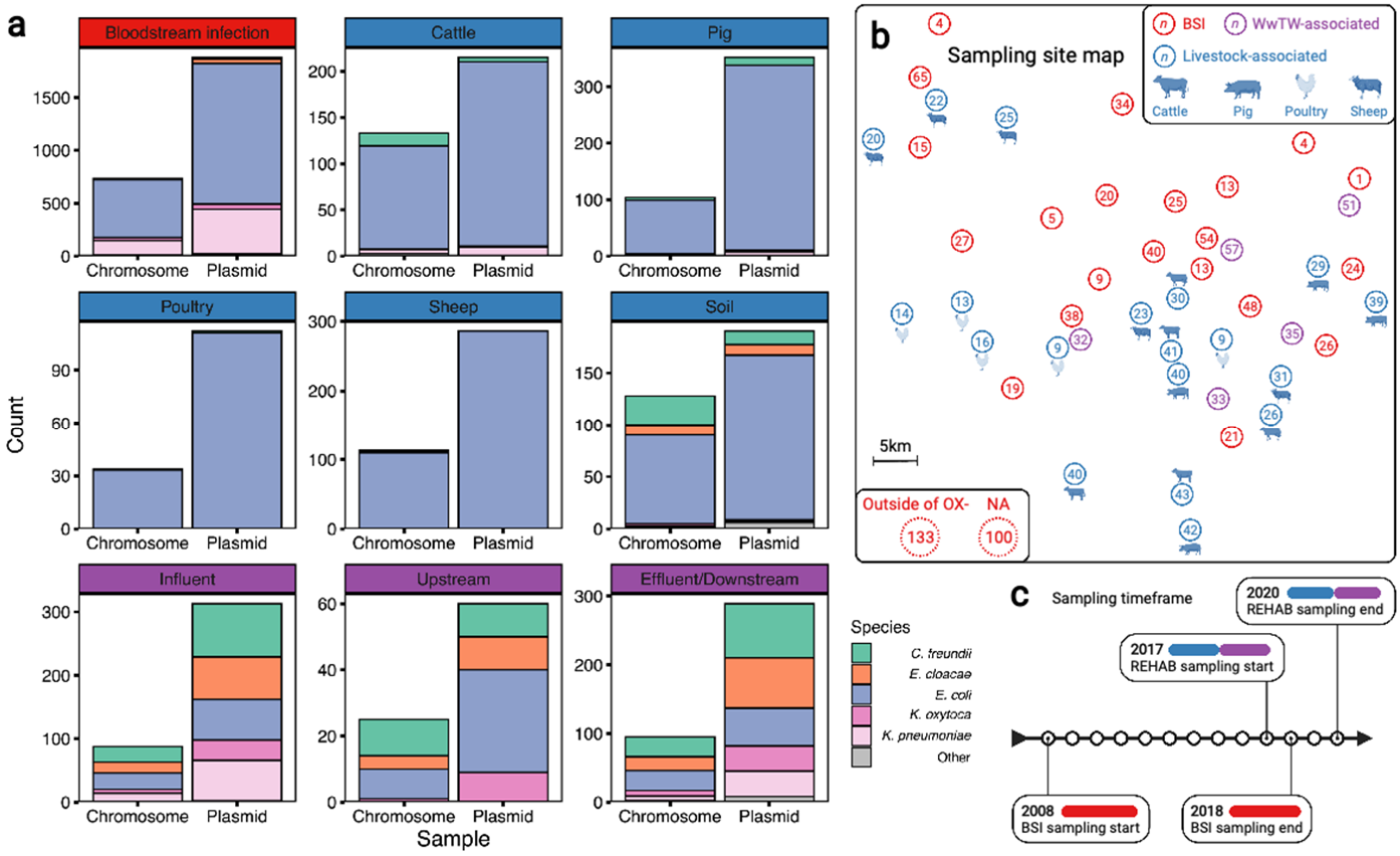
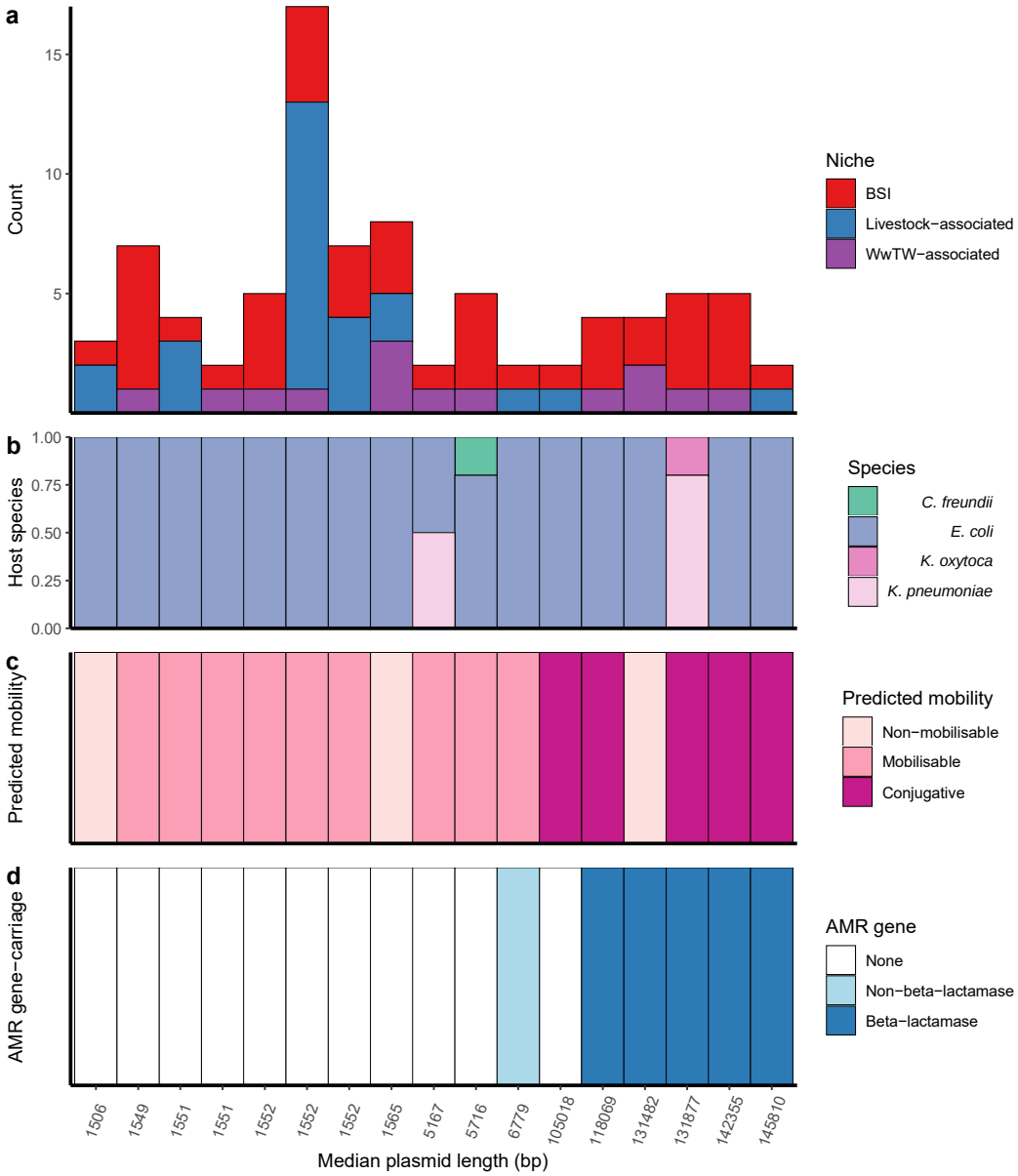


Figure 2 Cross-niche, near-identical plasmids.

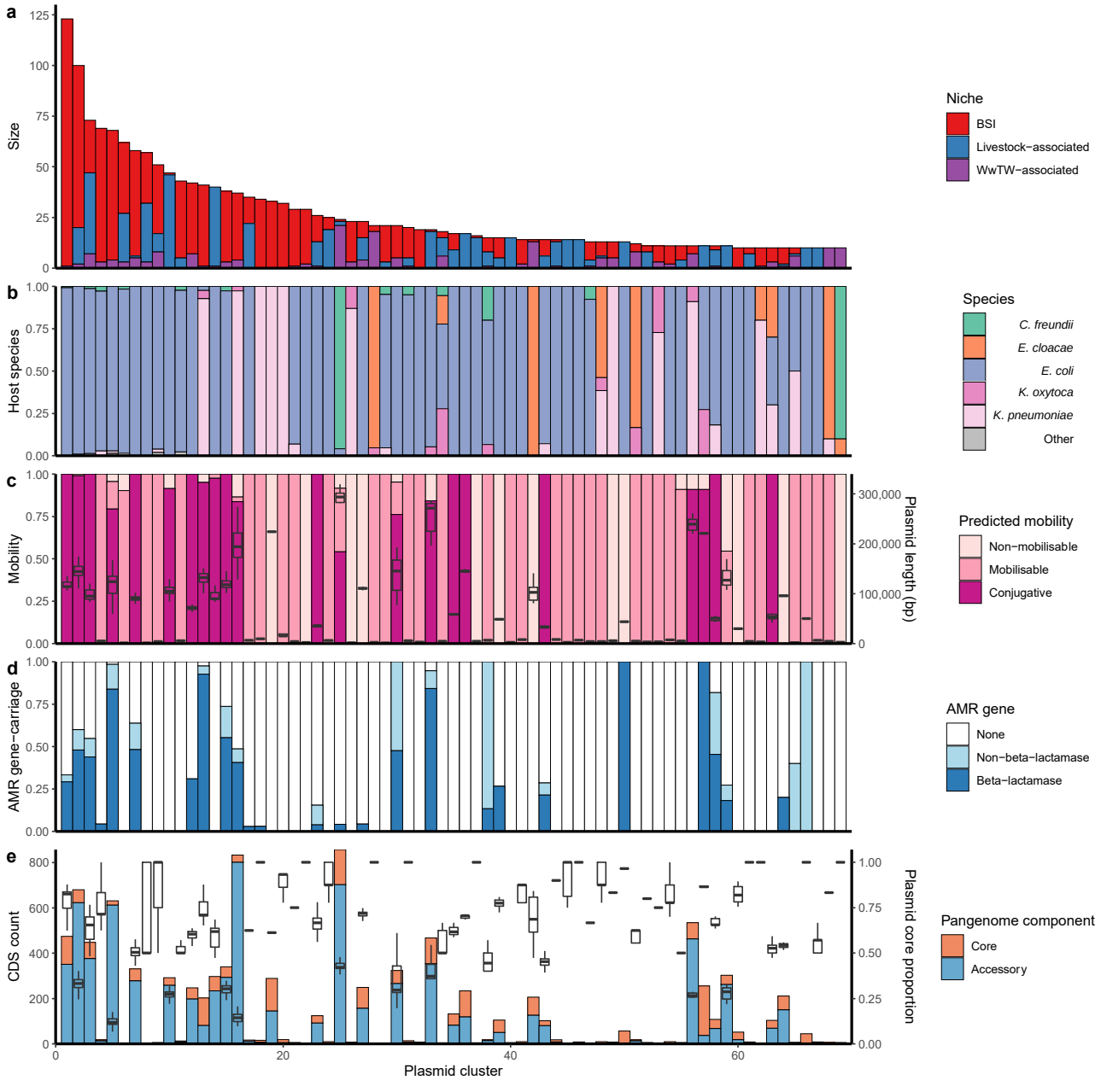
(a) Size of cross-niche, near-identical plasmid groups, coloured by niche (total *n*=84 plasmids). Median length (bp) of plasmids within groups increases from left to right. **(b)** Proportion of plasmid host species by group. **(c)** Predicted mobility of plasmid. **(d)** AMR gene carriage in plasmid. For small plasmids, the stringent distance threshold (*d*<0.0001) becomes an identical threshold, meaning that plasmids of the same length with a single SNP between them are grouped into different groups (e.g., the three groups with length=1,552bp;



992 see Materials & Methods). From left to right, the near identical groups are named in
993 Supplementary File 3 as 156, 18, 117, 210, 22, 29, 44, 19, 184, 6, 208, 139, 32, 26, 10, 192,
994 217.

Figure 3. Genetically similar plasmids share between niches

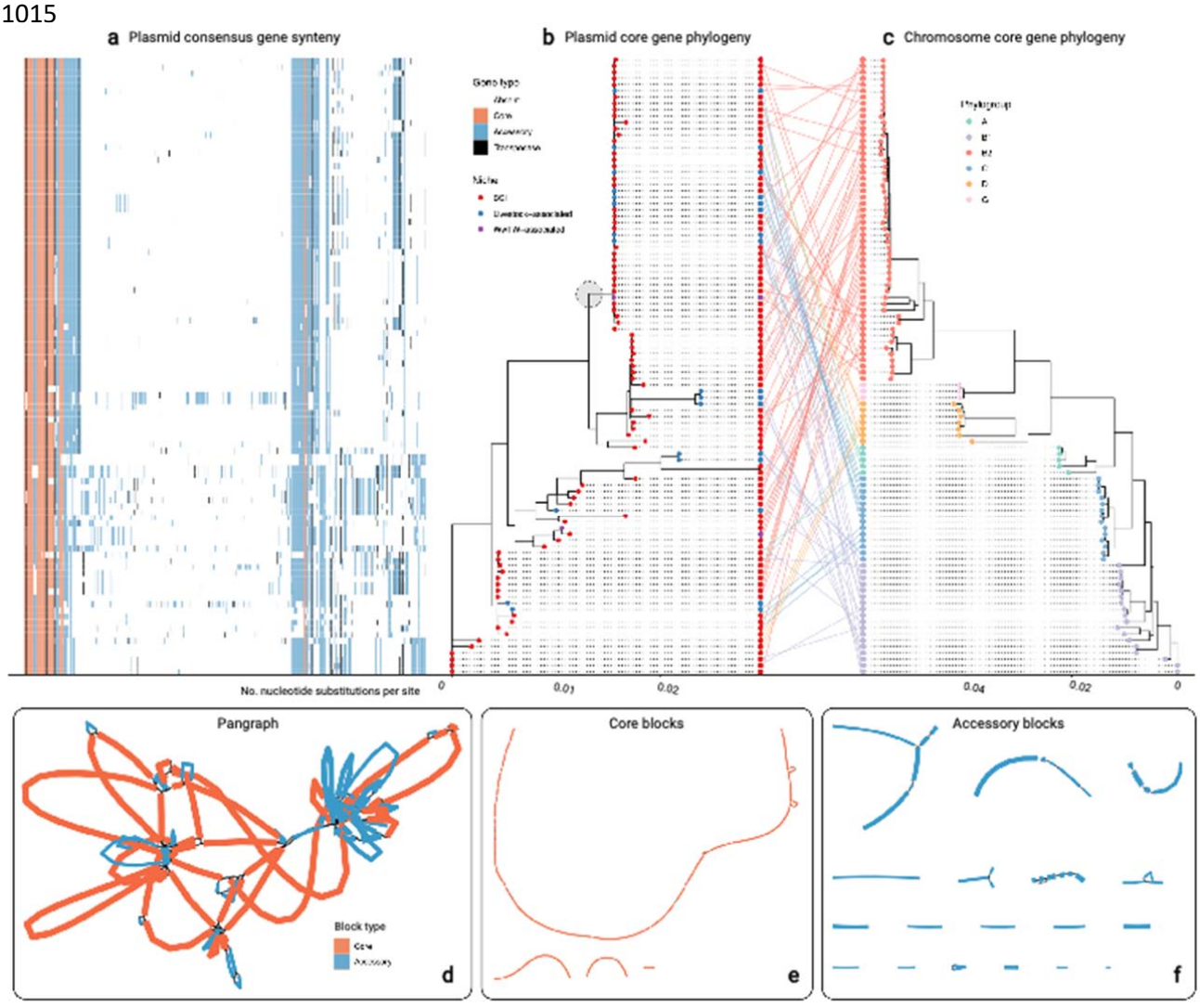
(a) Size of plasmid clusters with at least 10 members, coloured by niche. Size of clusters decreases from left to right. (**b)** Proportion of plasmid host species by cluster. **(c)** Plasmid mobility class and size: Left hand axis shows proportion of plasmids with a predicted mobility class by cluster. Right hand axis shows plasmid length boxplots by cluster. **(d)** Proportions of AMR gene carriage by cluster. **(e)** Plasmid core and accessory genomes: Left hand axis shows the count of core and accessory coding sequences (CDS) for the entire cluster as a bar chart. Right hand axis shows plasmid core gene proportions (i.e., core



1003 CDS/total CDS for each plasmid) as a boxplot.

Figure 4. Cluster 2 plasmid and host evolution

(a) Consensus gene ordering for plasmid cluster 2, coloured by gene type (total *n*=99 plasmids; *n*=1 *S. enterica* isolate omitted). Genes are coloured by core, accessory, or transposase. **(b)** Plasmid core gene phylogeny with tips coloured by sampling niche. The grey circle highlights the clade of *n*=44 plasmids which were further analysed. **(c)** Plasmid host chromosome core gene phylogeny with tips coloured by sampling niche. Plasmid and host phylogeny tips are connected in a 'tanglegram' which connects pairs of plasmids and chromosomes from the same isolate. **(d)** Visualisation of the pangraph for *n*=44 plasmids in the grey-circled clade in (b). Blocks are coloured by presence in plasmids. **(e)** Core blocks (found in at least 95% of the *n*=44 plasmids). **(f)** Accessory blocks (found in less than 95% of the *n*=44 plasmids).



Appendix 1 Figure captions

Figure 1. Mash tree for *n*=1,044 *E. coli* chromosomes. Tree tips are coloured by sampling compartment, scale is Mash distance.

Figure 2. Mash tree for *n*=163 *K. pneumoniae* chromosomes. Tree tips are coloured by sampling compartment, scale is Mash distance.

Figure 3. Accumulation curves of near-identical plasmid groups and singletons against isolate sample size. Black lines represent *b*=1000 bootstrap simulations, the red line represents their average.

Figure 4. Network evolution of largest connected component, number of connected components, and number of singletons, as edges are removed at increasing JI thresholds. The vertical red line represents the chosen threshold of *JI*=0.5.

Figure 5. Number of clusters detected within the plasmid network at increasing *JI* thresholds. Interval bars represent the IQR in cluster number at a given threshold over 50 runs of the Louvain algorithm.

Figure 6. Accumulation curves of plasmid clusters, doubletons, and singletons against isolate sample size. Black lines represent *b*=1000 bootstrap simulations, red line represents their average.

Figure 7. Plasmid clusters containing *bla*$_{\text{TEM-1}}$ carry more AMR genes. Each point is one plasmid cluster. *n*=247 clusters are shown, with panels facetted by the number of niches the plasmid cluster represented. *p*-values are from the Wilcoxon test.

Figure 8. Plasmid accessory gene presence/absence Jaccard distance against core gene cophenetic distance. Presented are data points from 27/247 clusters for which (i) all plasmids had at least 1 core gene, and (ii) the cluster contained at least 50 accessory genes. The red line is a statistically significant (p-value<2.2e-16) log-transformed linear regression.

Figure 9. Accumulation curves of near-identical plasmid matches with BSI plasmids and singletons against livestock-associated (environmental soils/livestock) isolate sample size. Black lines represent *b*=1000 bootstrap simulations, red line represents their average.

Supplementary File captions

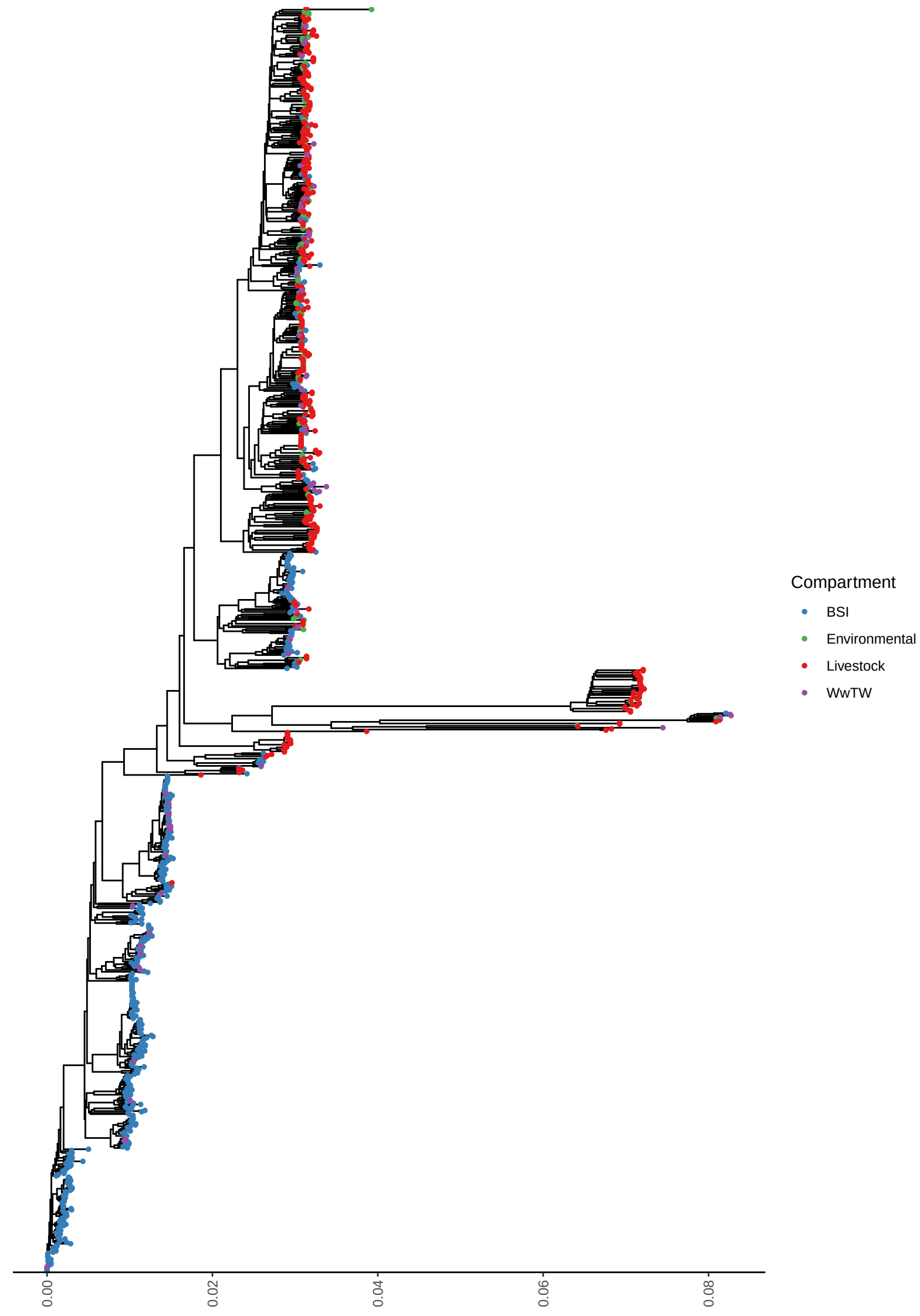
File 1. In order, this file contains (i) a metadata table for all plasmid clusters where every plasmid has at least one core gene. Presented are the cluster name, cluster size, mlst PubMLST genera of plasmid hosts, plasmid PlasmidFinder annotations, and plasmid NCBIAMRFinder annotations, and (ii) A core gene phylogeny and consensus gene synteny heatmap for all clusters in the metadata table, in cluster size decreasing order. Core gene phylogeny scales are in single nucleotide polymorphisms (SNPs).

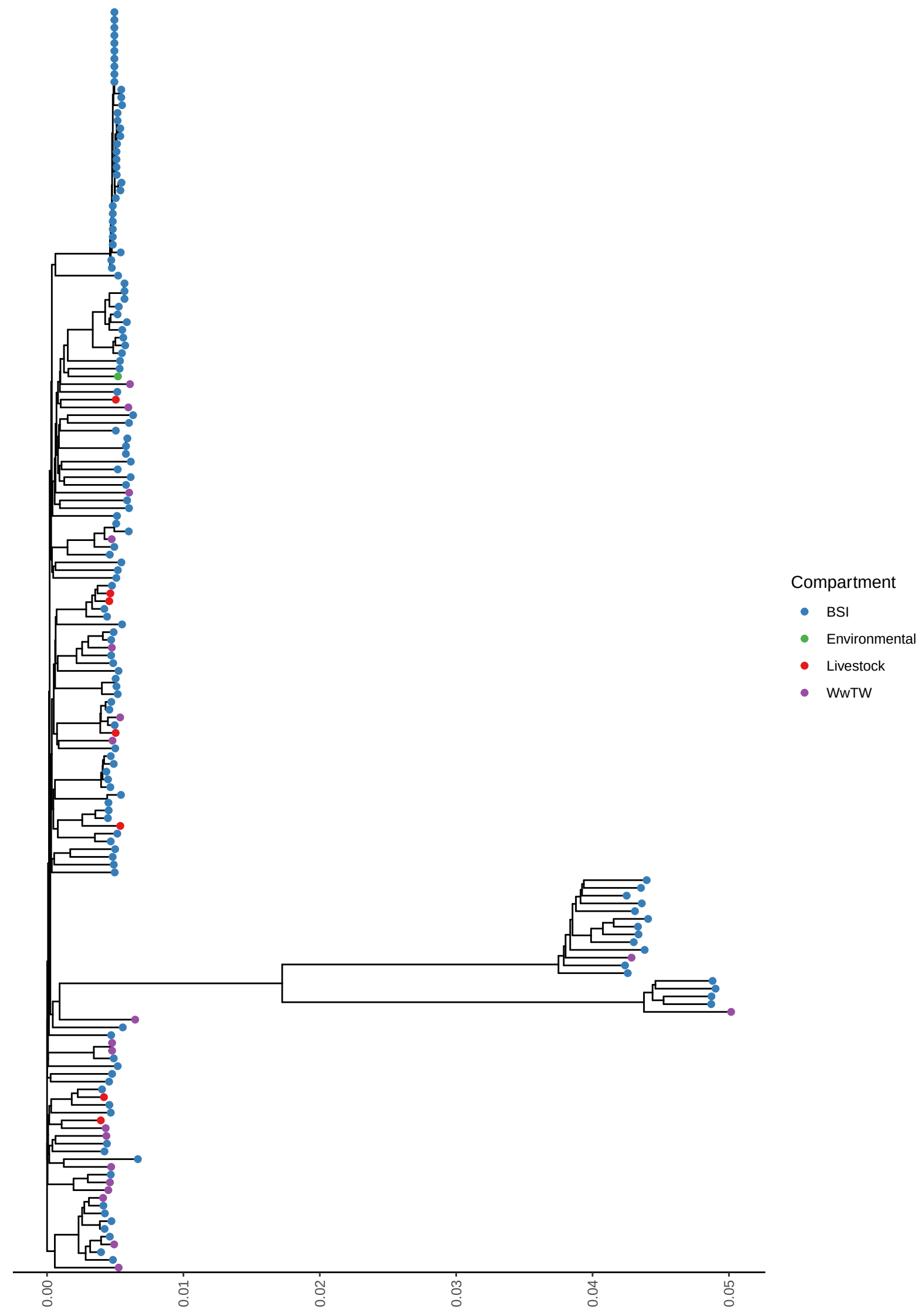
File 2. Fritz and Purvis' *D* estimates for the *n*=27/62 plasmid clusters that contained both BSI and livestock-associated plasmids.

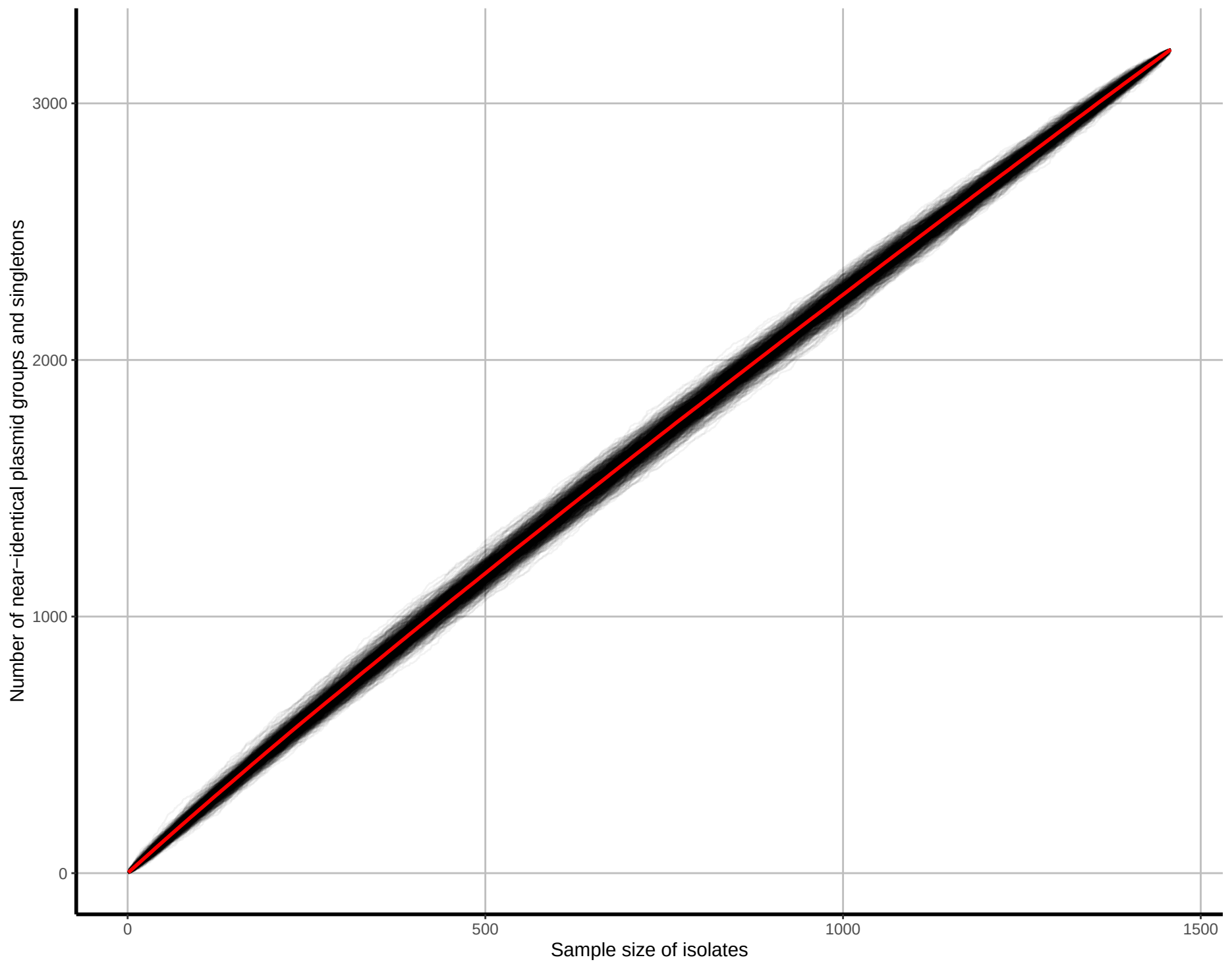
File 3. Isolate and assembly metadata.

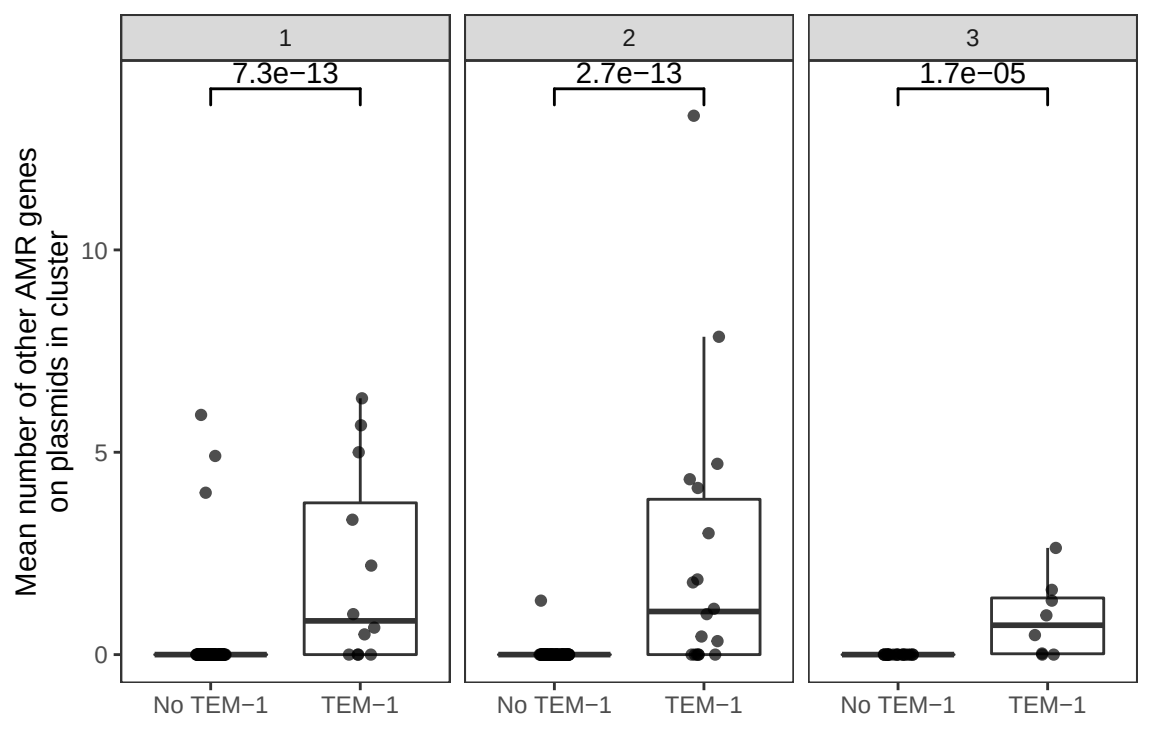
File 4. Accessions for poultry and environmental soil isolate reads assembled for this study.

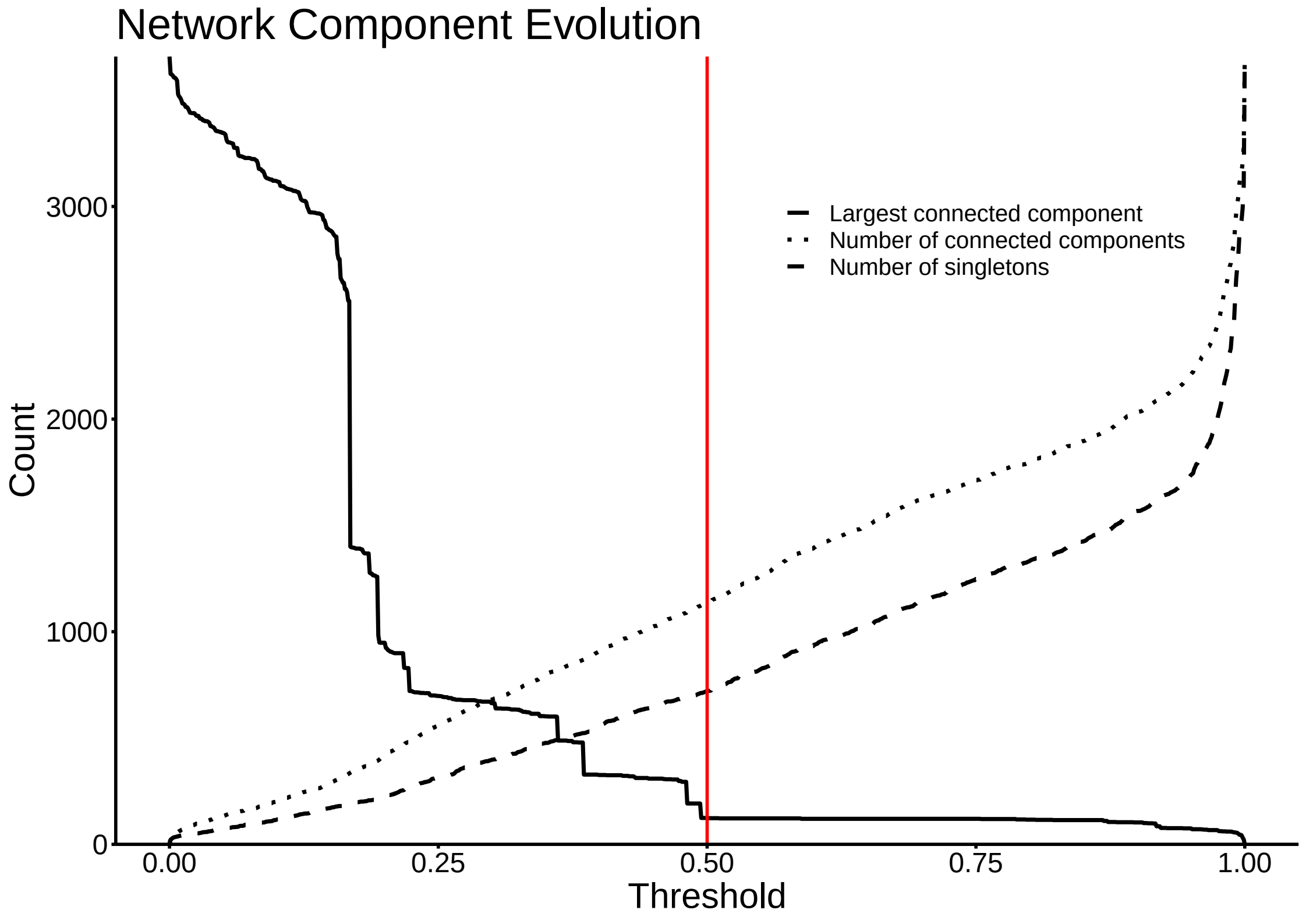
File 5. Cluster 2 core and accessory pancontig gene annotations.

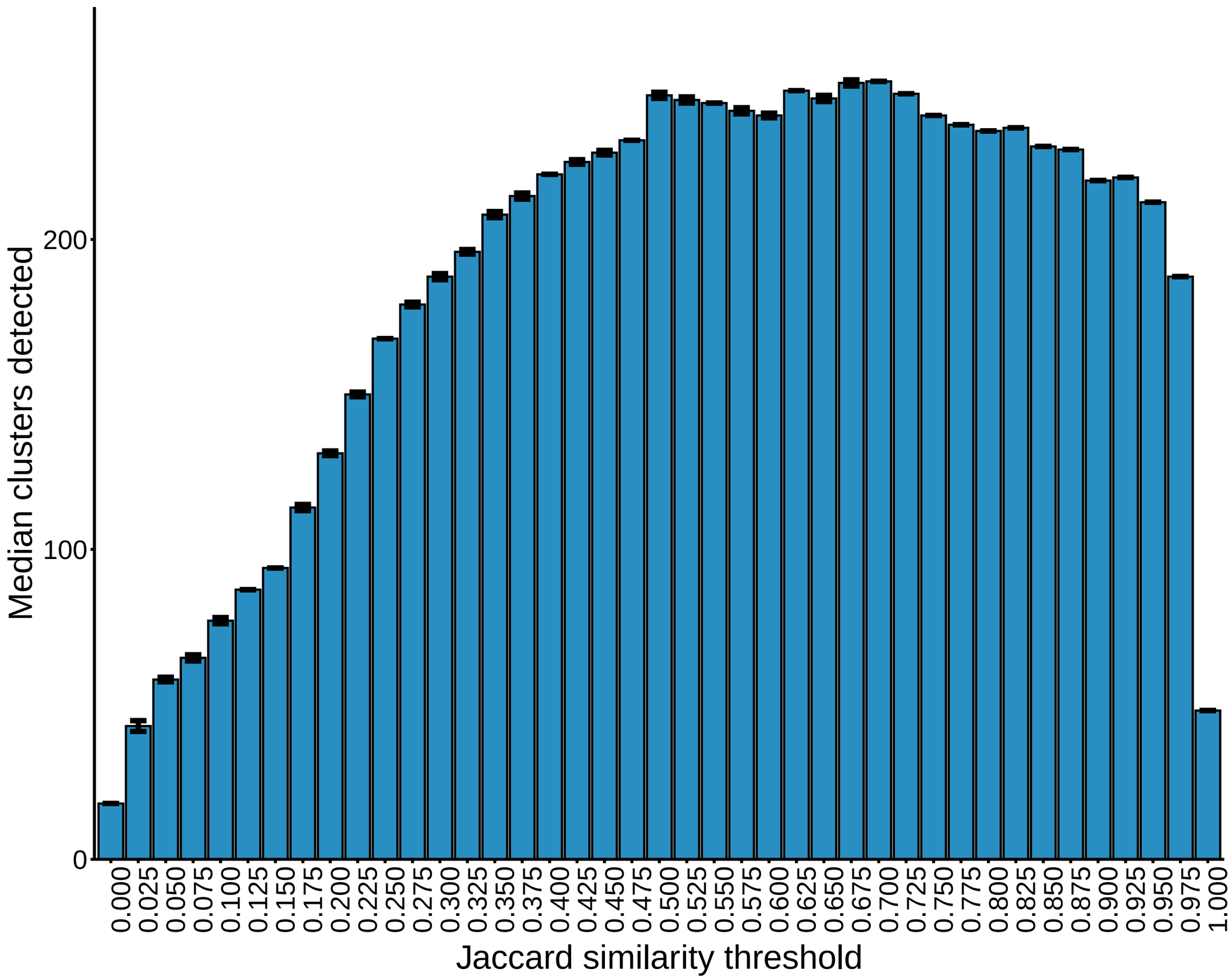

Compartment
BSI
Environmental
Livestock
WwTW
0.00
0.02
0.04
0.06
0.08

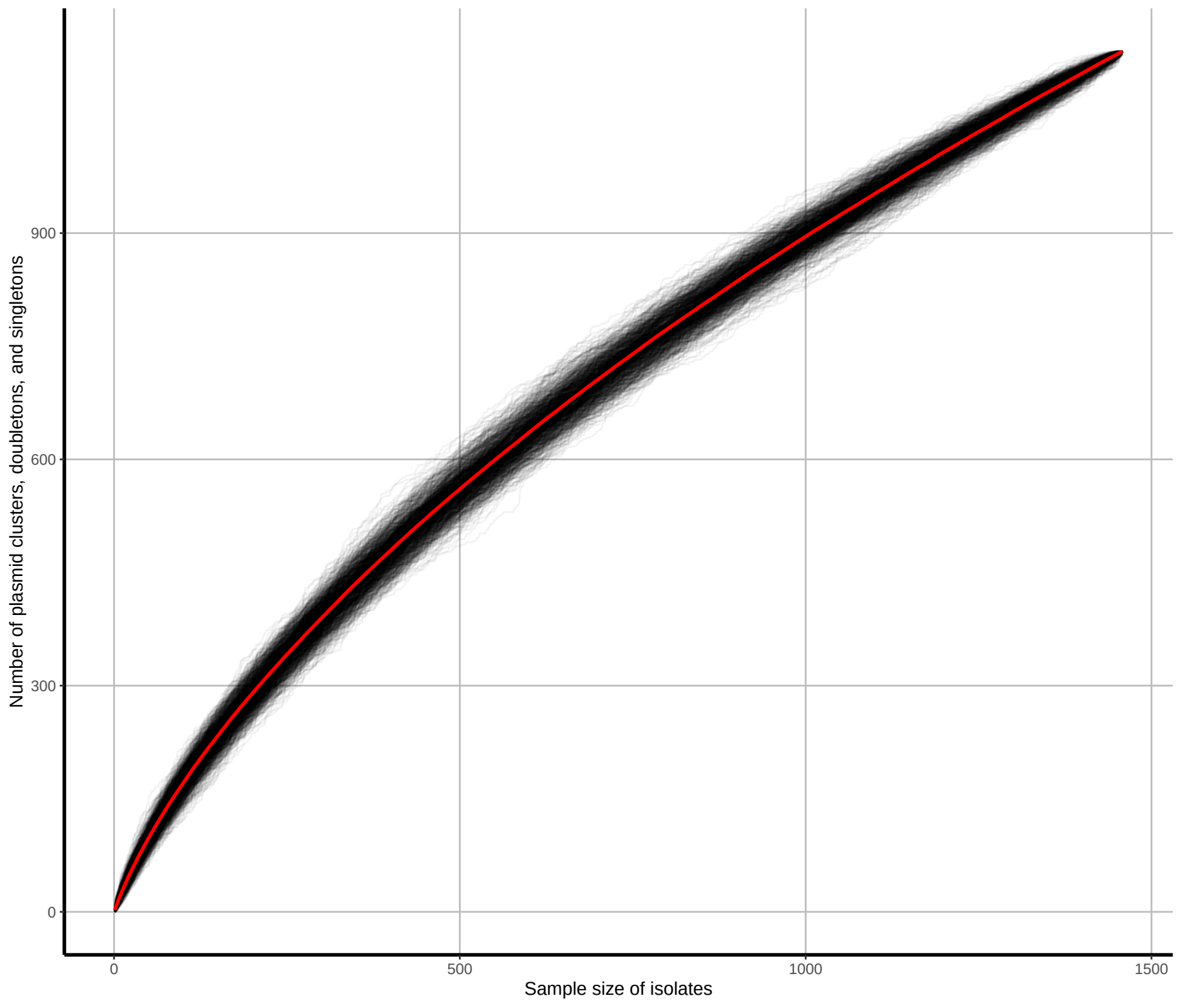

Compartment
BSI
Environmental
Livestock
WwTW
0.00
0.01
0.02
0.03
0.04
0.05

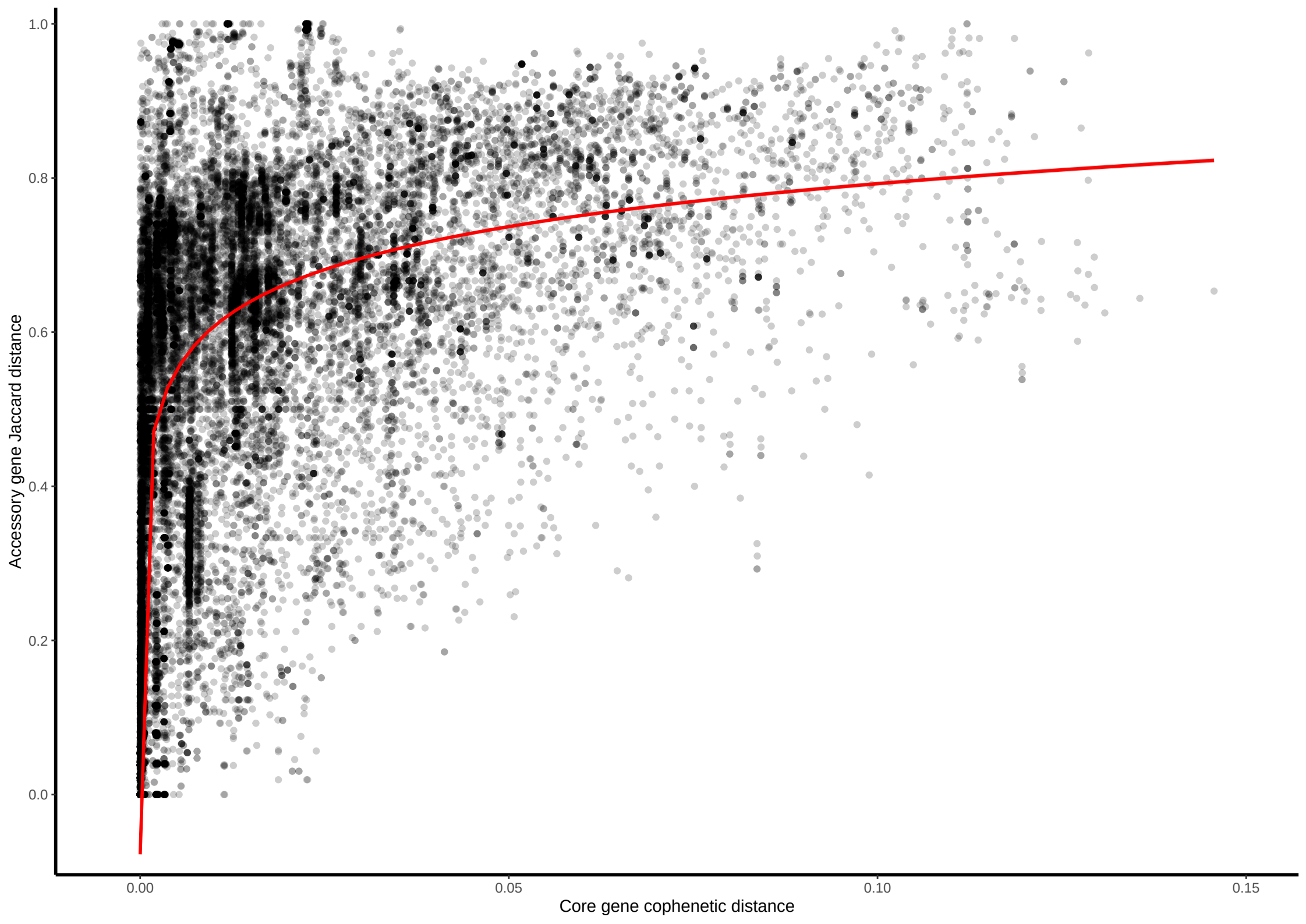

Number of near−identical plasmid groups and singletons
3000
2000
1000
0
0
500
1000
1500
Sample size of isolates

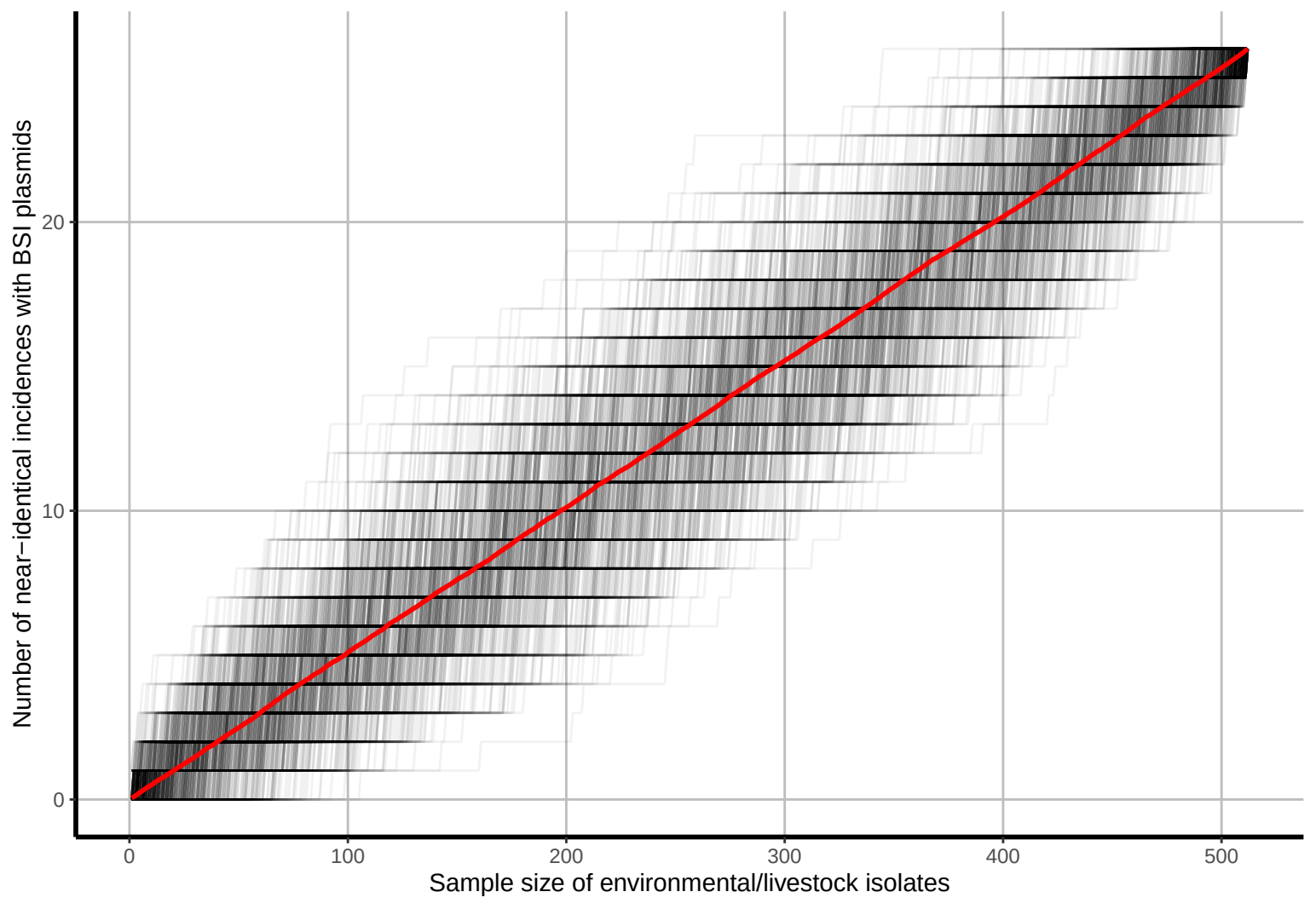

1
2
3
7.3e−13
2.7e−13
1.7e−05
Mean number of other AMR genes on plasmids in cluster
10
5
0
No TEM−1
TEM−1
No TEM−1
TEM−1
No TEM−1
TEM−1


Network Component Evolution
Largest connected component
Number of connected components
Number of singletons
Count
0
1000
2000
3000
0.00
0.25
0.50
0.75
1.00
Threshold


Median clusters detected
0
100
200
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
0.225
0.250
0.275
0.300
0.325
0.350
0.375
0.400
0.425
0.450
0.475
0.500
0.525
0.550
0.575
0.600
0.625
0.650
0.675
0.700
0.725
0.750
0.775
0.800
0.825
0.850
0.875
0.900
0.925
0.950
0.975
1.000
Jaccard similarity threshold


Number of plasmid clusters, doubletons, and singletons
900
600
300
0
0
500
1000
1500
Sample size of isolates


Accessory gene Jaccard distance
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
Core gene cophenetic distance


Number of near−identical incidences with BSI plasmids
0
10
20
0
100
200
300
400
500
Sample size of environmental/livestock isolates