

Genomic network analysis of an environmental and livestock IncF plasmid population

William Matlock¹, Kevin K. Chau¹, Manal AbuOun², Emma Stubberfield², Leanne Barker¹, James Kavanagh¹, Hayleah Pickford¹, Daniel Gilson², Richard P. Smith², H. Soon Gweon^{3,4}, Sarah J. Hoosdally¹, Jeremy Swann¹, Robert Sebra⁵, Mark J. Bailey³, Timothy E. A. Peto^{1,6,7}, Derrick W. Crook^{1,6,7}, Muna F. Anjum², Daniel S. Read³, A. Sarah Walker^{1,6,7*}/Nicole Stoesser^{1,6,7*}/Liam P. Shaw^{1*}, and on behalf of the REHAB consortium[†]

*contribution considered equal

[†] see acknowledgements

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

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

Keywords. plasmid genomics; *Enterobacteriaceae*; environmental bacteria; antimicrobial resistance (AMR); network analysis; One Health

Abstract

IncF plasmids are diverse and of great clinical significance, often carrying genes conferring antimicrobial resistance (AMR) such as extended-spectrum β -lactamases, particularly in *Enterobacteriaceae*. Organising this plasmid diversity is challenging, and current knowledge is largely based on plasmids from clinical settings. Here, we present a network community analysis of a large survey of IncF plasmids from environmental (influent, effluent, and upstream/downstream waterways surrounding wastewater treatment works) and livestock settings. We use a tractable and scalable methodology to examine the relationship between plasmid metadata and network communities. This reveals how niche (sampling compartment and host genera) partition and shape plasmid diversity. We also perform pangenome-style analyses on network communities. We show that such communities define unique combinations of core genes, with limited overlap. Building plasmid phylogenies based on alignments of these core genes, we demonstrate that plasmid accessory function is closely linked to core gene content. Taken together, our results suggest that stable IncF plasmid backbone structures can persist in environmental settings while allowing dramatic variation in accessory gene content that may be linked to niche adaptation. The recent association of IncF plasmids with AMR likely reflects their suitability for rapid niche adaptation.

Introduction

Environmental (non-clinical, non-human) populations of *Enterobacteriaceae* may act as a genetic reservoir for antimicrobial resistance (AMR). This includes livestock [1-5] and water-borne [6] resistance. Frequent horizontal gene transfer (HGT) in *Enterobacteriaceae* populations results in a large and open pangenome, enabling the wide-spread transmission of the genes conferring AMR [7-9]. This includes AMR transmission between humans and the environment and vice versa [10]. However, evidence for this transmission is often context and sequence type (ST)-specific, with broader transmission patterns less conclusive [10, 11]. IncF plasmids are a diverse group of *Enterobacteriaceae*-associated replicons which mediate the transfer of AMR genes. They have also been observed in other Proteobacteria, such as *Aeromonadaceae* and *Comamonadaceae* [12]. In

particular, their involvement in the dissemination of genes encoding extended-spectrum β -lactamases (ESBLs), such as *bla*_{CTX-M-15}, is of major clinical concern [13, 14], and almost 40% of plasmid-borne carbapenemases are carried on IncF plasmids [15]. IncF plasmids are often low copy-number and conjugative [16]. Further, recent database analysis suggests IncF alleles are carried in over 50% of multireplicon plasmids [12].

Previous studies of IncF plasmids have often focussed on clinically relevant isolates, often only those encoding ESBLs [15]. Further, they have been limited to studies with smaller sample sizes. Here we analyse hundreds of IncF plasmids drawn from a survey of environmental diversity in *Enterobacteriaceae*, sampled in 2017 from a region of south-central England, UK [17]. Sampling was from livestock (cattle, pig, sheep), and from influent, effluent, and upstream/downstream waterways surrounding wastewater treatment works (collectively termed WwTWs). Potential seasonal variation was accounted for by sampling over three time-points at each site. This provided a high-quality dataset of $n=726$ plasmids for characterising natural plasmid populations.

Frequent co-integration and HGT events mean plasmid evolution cannot be described with a phylogenetic tree. Instead, networks based on sequence similarity can be used [18]. In such networks, nodes represent plasmids, and edges are weighted by a metric on the plasmid sequences. This captures both vertical and horizontal evolution at the cost of not providing a most recent common ancestor. Communities are a topological property of networks. They are defined as subsets of nodes with dense intra-connections, but sparse inter-connections [19]. In our analyses, they represented groups of similar plasmid sequences. Detecting these structures gives a coarse-grained view of the plasmid population. Previous efforts have often focussed on the relationship between network features used in plasmid classification schemes, such as replicon presence, MOB-type, or predicted mobility [20-23]. Further, studies have often focussed on curated selections from online databases [20, 22-24]. It is yet to be seen if similar community structure is present in large-scale natural populations. Additionally, it is important to develop fast and scalable methods for analysis of large and diverse WGS datasets. Here we aimed to provide a framework applicable to such studies.

Results.

A natural population of complete plasmids with IncF replicons.

We recovered $n=726$ circularised plasmids containing an IncF replicon (see Table S6) from a large dataset of high-quality *Enterobacteriaceae* genomes, obtained by hybrid assembly using both short-read (Illumina, 150bp paired end) and long-read (PacBio or Nanopore) sequencing of cultured isolates [17]. These isolates were collected over three time-points in 2017 from a region of south-central England, UK. Sampling was from 14 livestock farms (4 pig, 5 cattle, 5 sheep) and from waterways (influent, effluent and rivers) surrounding five WwTWs. Of the livestock plasmids, 120 were from pigs, 137 were from cattle and 150 were from sheep. The remaining 319 plasmids were from WwTWs.

IncF plasmids were found across all four genera in the dataset: *Citrobacter* (53 *C. freundii*), *Enterobacter* (67: 65 *E. cloacae*, 2 untyped *Enterobacter* sp.), *Escherichia* (471 *E. coli*), and *Klebsiella* (135: 61 *K. oxytoca*, 67 *K. pneumoniae*, 7 untyped *Klebsiella* sp.). Livestock plasmids mostly came from *Escherichia* (392/407), whereas WwTW plasmids had a more uniform distribution over all four genera in line with the greater diversity of genera in WwTW isolates (Fig. 1a). Our plasmids originated from $n=558$ host *Enterobacteriaceae* isolates, with a majority of chromosomes circularised (431/558).

Plasmids ranged in length from approximately 20kb to 480kb (Fig. 1b). A majority of plasmids were predicted to be conjugative (516/726), with a smaller number predicted to be mobilisable (39/726) or non-mobilisable (171/726) (see Materials and Methods). Plasmid length was linked to mobility, with conjugative plasmids generally larger than mobilisable and non-mobilisable plasmids (Kruskal-Wallis test [$H=36.7$, $p=1.08\text{e-}8$] followed by Dunn test with Holm adjusted p -value [Conj—Mob: $Z=3.45$, p -value= $1.14\text{e-}03$; Conj—Non-Mob: $Z=5.39$, p -value= $2.07\text{e-}07$; Mob—Non-Mob: $Z=-0.54$, p -value= 0.59]). We found 25 different replicons across all plasmids, including 11 in unspecified gene clusters, present in 62 different combinations or 'replicon haplotypes' (Table S1). 28 replicon

haplotypes appeared only once in the sample. Plasmids carried between 1 and 5 replicons, with a majority carrying 2 (318/726) or 3 (209/726). Plasmid length was positively associated with number of replicons carried (one-way ANOVA test [$F(4, 721)=5.64, p\text{-value}=1.8\text{e-}4$] followed by Tukey's HSD). All plasmids contained at least one IncF replicon: IncFIA (147), IncFIB (460), IncFII (574) and IncFIIA (370). Of the remaining replicons, IncII was most common (28), and was always found with an IncFII replicon. We observed different replicon co-occurrence patterns (Fig. 1c), with individual IncF replicons associated with different non-IncF replicons. For instance, IncU and IncN replicons were only found with IncFIB and IncFII respectively. Overall, these co-occurrence patterns corroborate previously observed patterns of frequent IncF association with replicons such as IncII, IncX and IncR [12].

IncF plasmids tended to be AT-rich relative to their host chromosomes. This trend has been widely reported before [25, 26]. However, we found that relative AT-richness significantly varied between host genus (one-way ANOVA test [$F(3, 561)=111, p\text{-value}<2\text{e-}16$] followed by Tukey's HSD), independently of average host GC-content, with *Klebsiella* plasmids having a greater relative AT-richness than other *Enterobacteriaceae* plasmids (Fig. 1d).

Detecting communities in plasmid k-mer networks.

Plasmid sequence distances were calculated using MASH, a *k*-mer based distance estimation [27] (see Materials and Methods), and these distances used as weighted edges in a plasmid network. The output MASH edge list is presented in Table S7. Communities were detected using the Louvain algorithm, which optimises the modularity of the networks, and is a weighted community detection algorithm, meaning it also accounts for the MASH distances [19]. To effectively detect communities, we reduced the density of the network by thresholding the edges (i.e. by 'sparsification'). Fig. 2a-b shows how the number of identified communities and percentage of plasmids covered changed as the edge (i.e. MASH distance) threshold was varied. Generally, the Louvain algorithm became more consistent in coverage as we sparsified. To ensure the communities represented potential sub-populations, we only considered those with at least 10 plasmids. The large drop in community coverage seen at

threshold=0.175 (Fig. 2b) was due to the break-up of a large connected component (analogous plots for communities with at least 3 plasmid members are shown in Fig. S1a-b).

The application of different MASH distance thresholds revealed different community structures within the network. Fig. 2c shows the kernel density estimates (KDEs) of MASH distances stratified by sampling compartment, with an overall range of [0, 0.602], highlighting that livestock plasmids (median=0.152) were generally more similar to each other than WwTW plasmids (median=0.258) and suggesting that plasmid diversity was higher in WwTW isolates. A distance threshold low enough to reveal the livestock sub-network structure could destroy the structure of the WwTW sub-network, so for this study, we selected a threshold=0.05, which revealed the structure of livestock plasmids at the expense of some WwTW structure break-up. Note that this threshold is far lower than those used in previous plasmid network analyses of global plasmid diversity [21-23]; our sample was smaller and restricted to a broad plasmid family so required more severe sparsification to reveal communities. At this level, the network's largest connected component (LCC) had 201 nodes with 182 connected components in total (Fig. 2d). It also had the highest number of communities (13) containing at least 10 plasmids (Fig. 2a), and coverage of over 50% (Fig. 2b). There were 99 singleton plasmids, consistent with high levels of diversity in the population. A visualisation of the network at this threshold with the 13 communities coloured is presented in Fig. 3.

Community metadata analysis

To evaluate the relationship between the node metadata labels and the network, two entropic measures were considered: homogeneity (h) and completeness (c) (both range from 0-1; see Materials and Methods). Homogeneity measures the distribution of labels given a community, with an ideal community containing a single label: a high homogeneity means that plasmids with similar sequences tend to have similar metadata labels. Conversely, completeness measures the distribution of communities given a label: a high completeness means that instances of a label tend to fall within a single community. Importantly, both homogeneity and completeness are independent of community

size, the number of communities, and the number of metadata labels. This makes the approach robust to uneven sampling strategies, such as the disproportionate number of *E. coli* isolates in our sample.

Each plasmid was assigned a set of metadata labels, consisting of a sampling compartment (livestock type [pig, cattle, sheep] or WwTW-association [influent, effluent, upstream, downstream]), a host genus (*Citrobacter*, *Enterobacter*, *Escherichia* or *Klebsiella*), and a time-point (1, 2 or 3).

Homogeneity (Table 1) and completeness (Table 2) were averaged over 100 runs of the Louvain algorithm. Despite the number of communities remaining consistent, some variation in the measures arose from minor changes in community boundaries.

Homogeneity scores showed that sampling compartment shaped plasmid similarity. At the coarsest resolution there was high homogeneity considering livestock versus WwTW ($h=0.715$; Table 1), meaning that plasmid communities were largely distinct between livestock and WwTW settings. This metadata partition is projected on the network in Fig. 4a. However, homogeneity was lower when comparing different livestock types (pig, cattle, sheep) ($h=0.591$) and even more so when comparing different farms ($h=0.403$), meaning that there was a loss of structure at these levels and plasmid communities were not well segregated by individual farm. Homogeneity was also low if plasmids were stratified by individual WwTWs ($h=0.467$). However, homogeneity increased for influent/upstream versus effluent/downstream compartments ($h=0.550$) indicating some differences in plasmids before and after WwTW treatment. Overall, plasmids from WwTWs were weakly structured by wastewater catchment.

Completeness scores highlighted higher WwTW diversity compared to lower livestock diversity. For the binary livestock or WwTW label plasmid communities scored a low completeness (Table 2; $c=0.199$), which changed little when stratified over the individual WwTWs ($c=0.238$), indicating a uniform distribution of WwTW labels over the plasmid communities and high diversity. Based on our MASH distance KDEs (Fig. 2c), we would expect livestock plasmids to have higher completeness scores than WwTW plasmids due to the lower levels of diversity; as anticipated, when stratifying the

livestock metadata, completeness scores increased ($c=0.332$ and $c=0.400$). This indicated plasmids from the same farm were more likely to be found in the same community.

Host genus also played an important factor in partitioning plasmid diversity. The homogeneity scores were very high, implying a significant genetic partition by host (Table 1; $h=0.888$). This metadata partition is displayed in Fig. 4b. The lower completeness suggested a moderate level of diversity across all *Enterobacteriaceae* plasmids (Table 2; $c=0.307$). There was a very weak time-point effect found in the network (Tables 1, 2; $h=0.050$ and $c=0.023$). Under a one-tailed permutation test, all metadata label configurations except time-point had a zero p -value for homogeneity and completeness (Table S2; see Materials and Methods), indicating that overall, there was a significant association between niche (sampling compartment and host genus) and plasmid population structure.

Community pangenomes

To explore the genetic structure of the communities we considered the set of all represented genes within a community, known as the pangenome (see Materials and Methods). Plasmids had a median of 35 annotated genes (range=4-112). Genes conferring AMR were found in 17% (122/726) of plasmids; this included 33 plasmids carrying ESBLs (9 pig, 8 cattle and 16 WwTw), with 4 carrying *bla*_{CTX-M-15} (all WwTw). IncF plasmids in isolates cultured from pigs were disproportionately associated with AMR genes (45/109 [41%] AMR plasmids).

Core genes with well-conserved synteny comprise the plasmid ‘backbone’ [18], which often controls essential replication and mobility functions. Genes with accessory function, such as AMR genes, are inserted into the backbone. For the 13 IncF plasmid communities identified in this study using the 0.05 threshold above (see Fig. 3), we found a median of 13 core genes (range=0-88) (Table 3). Each community possessed a unique combination of core genes, and pairs of communities shared a median of 0 core genes between them (range=0-21) (Table S3). The communities had a median of 463 accessory genes (range=151-790), sharing a median of 312 accessory genes (range=99-570) (Table S4). Pairs of communities sharing a higher number of genes tended to have a higher sum of individual

genes ($r=0.820$, $t=12.505$, $p\text{-value}<2.2e-16$), indicating overlap between larger pangenomes. Within a plasmid community, a lower mean MASH value indicates greater overall sequence similarity; as anticipated therefore, we found a lower mean MASH distance was associated with more core genes ($r=-0.615$, $t=-2.586$, $p\text{-value}=0.025$) and a lower total number of genes in the pangenome ($r=0.654$, $t=2.865$, $p\text{-value}=0.015$).

For an example community of 30 IncF plasmids from isolates from sheep farms, we produced a neighbour-joining phylogeny based on 64/384 core genes (Fig. 5). The tree accounts for homologous recombination, with events detected in 11/30 leaf nodes and 21 internal nodes, consistent with a high number of exchange events affecting this plasmid community. The median tract length was 156bp (range=2bp-2249bp). Annotation of the phylogeny with the 316 accessory genes for this community revealed that accessory gene presence aligned almost identically with the core gene phylogeny, suggesting that the evolution of the plasmid backbone is highly linked to accessory function. All host genera for this plasmid community were diverse *E. coli*, with 13 known STs present, consistent with widespread horizontal transfer of the plasmids from this community. Within this community, no plasmids carried AMR genes. Core genome phylogenies for other plasmid communities also showed a strong link between accessory gene presence and backbone contents (Figs. S2-S12).

Discussion

We have analysed plasmid communities using distance-free genomic networks to explore diversity within a large, natural population of IncF plasmids from four *Enterobacteriaceae* genera (*Citrobacter*, *Enterobacter*, *Escherichia* and *Klebsiella*). These IncF plasmids contained a diversity of replicons (plasmids contained 21 other replicons, forming 62 unique combinations) and we resolved plasmids into communities (12 communities of ≥ 10 plasmids). We found that 15% of IncF plasmids contained at least one AMR gene, and 5% carried an ESBL. This underlines that non-clinical plasmid populations can also carry AMR genes, and that WwTW environment and livestock niches are part of an AMR network for *Enterobacteriaceae* [2, 10].

Our network analysis revealed IncF plasmids were well partitioned by sampling compartment, with distinct communities isolated to WwTWs or livestock; however, there were also clear instances of sharing events between, for example, specific farm locations. There was also moderate partitioning by specific livestock species: pig, cattle and sheep. Additionally, there was a difference in plasmids before and after WwTW treatment. Sampling compartment also influenced diversity, with a higher diversity in WwTW-associated plasmids than livestock plasmids. This is probably because both river and wastewater catchments integrate a large number of human, livestock (farmed and wild) and environmental sources. Further, they also experience higher rates of inflow and outflow than farm-specific environments. The analysis also revealed a significant partition by host genera. Despite IncF plasmids ranging over all *Enterobacteriaceae* genera, it suggested some genus-specific adaptations. Notably, the extent of plasmid-host AT-richness relative to the host chromosome varied depending on the genus. It remains to be seen how such observed differences relate to plasmid function. However, this may be related to the livestock-WwTW partition, since our livestock plasmids were predominantly hosted by *E. coli*. We did not detect an effect of sampling time-point. This is may be because our time-points were too close and sample size too small to capture any significant evolution, or it may indicate that time of year is not a strong factor in determining community structure. It would be interesting to see how plasmids from clinical samples relate to those from our samples within the network, especially if pre-WwTW plasmids are considered as a proxy for human gut microbiomes.

Pangenome analysis of the inferred plasmid communities revealed that core gene content was mostly unique to communities. Further, they were strongly related to accessory function. Taken with the above results, we propose that sampling compartment and host greatly influence the function of plasmids. This includes AMR presence, with pigs, and hence *Escherichia*, carrying a disproportionate burden in our sample. The pangenomes for communities varied greatly in the number of core genes, with one community having zero. This may be because the threshold was not severe enough to segregate this particular community into uniquely similar groups. It also may result from how Panaroo (see Materials and Methods) corrects annotation errors, splitting gene clusters too readily. Generally, more genetically similar communities possessed a greater number of core genes and

smaller pangenome. Our results for IncF plasmid communities are in line with a recent study of the wider prokaryotic plasmidome which concluded that clusters of plasmids contain common genomic backbones [23].

Our study has several limitations. One important limitation, which applies more widely to network approaches which cluster or partition diversity, is that thresholding of the network is highly subjective and dataset dependent. Trade-offs are required to reveal the intermediate structures of the network whilst maintaining good community detection performance. We determined a threshold by considering MASH distance distributions and component evolution alongside Louvain output diagnostics. When diversity varies greatly between sampling compartments, a single threshold is unlikely to be globally optimal. In these cases, it is probably best to focus on sub-populations of interest. Despite only considering several hundred nodes here, our methodology is scalable to far larger studies. Originally, the Louvain algorithm had runtime complexity $O(e)$, where e is the number of edges in the network. This has since been improved to $O(v \log k)$, where v is the number of nodes and k is the average node degree [28]. Further, recent efforts have parallelised the Louvain algorithm to networks with billions of edges, though this approach was not necessary here [29]. Finally, our dataset is limited to the four *Enterobacteriaceae* genera under study and conclusions may not reflect the wider diversity of IncF plasmids beyond these genera.

In conclusion, our study adds to the growing literature on distance-free networks to characterise and partition plasmid diversity, introducing a scalable framework to quantify the relationship between network structure and plasmid metadata by identifying network communities. Overall, our approach represented a high-resolution strategy for summarising similarities and differences within plasmid populations, leveraging the advantages of having complete plasmid sequences and analysing these in the context of associated metadata. For IncF plasmids we were able to show the distinct, local effects of sampling compartment on plasmid structure and population, but also identify evidence for sharing of plasmids between bacterial lineages, farms and WwTW-associated contexts, with relevance for the "One Health"-associated study of mobile genetic elements and AMR genes. As long-read sequencing

302 costs fall, and increasingly large numbers of plasmids can be characterised, future work applying this
 303 method will contribute to better understanding plasmid populations, estimating transfer rates of
 304 important AMR genes and MGEs between potential reservoirs, and identifying hotspots of
 305 selection/transfer that might be amenable to intervention.

306

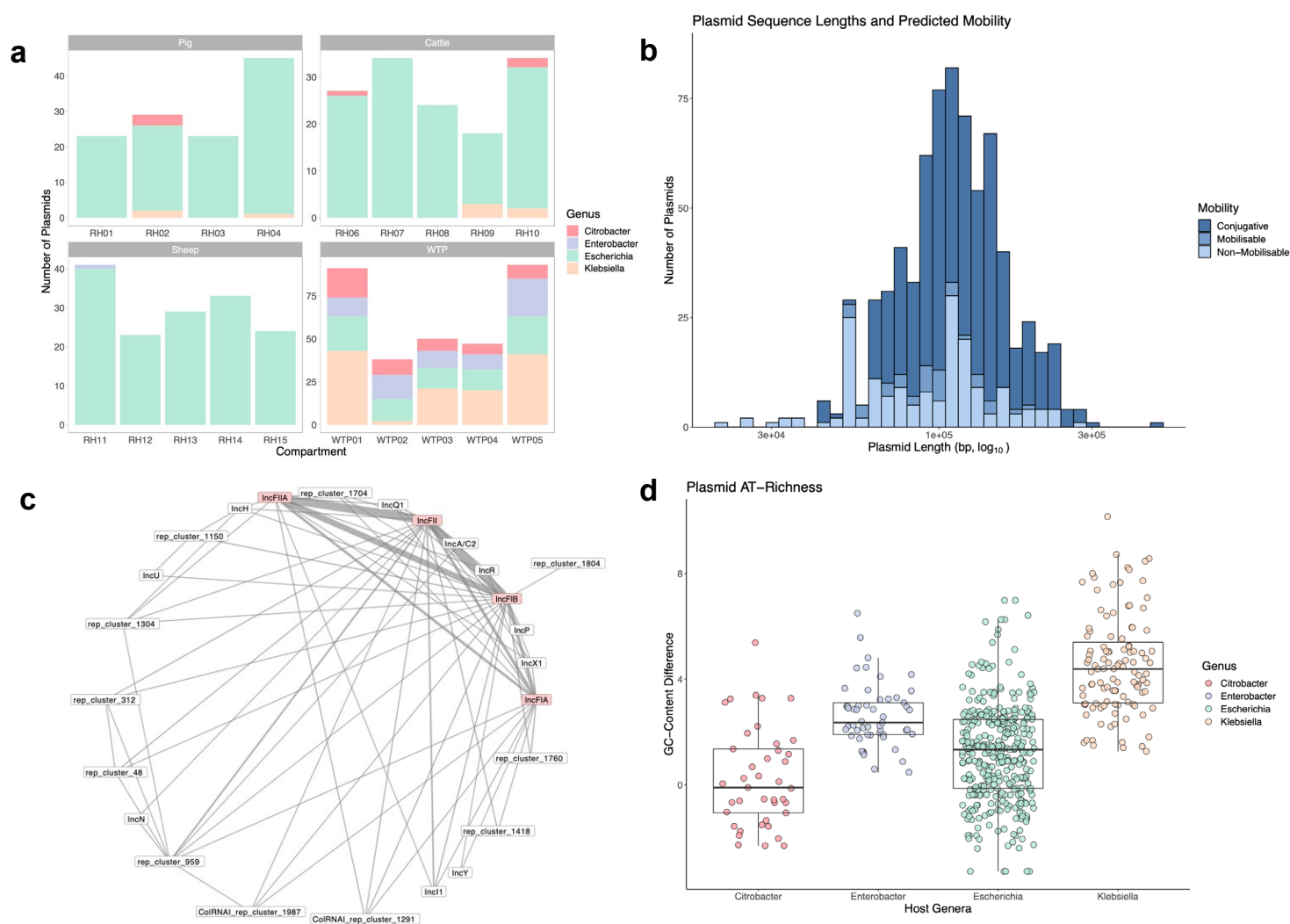


Figure 1. Overview of plasmid population. (a) Plasmid host genera distribution by compartment. (b) Distribution of plasmid sequence lengths with predicted mobilities. (c) Graph representing the association between replicon alleles. IncF nodes are coloured pink. Line weight is proportional to frequency of association in the sample. (d) Plasmid GC-content subtracted from host chromosome GC-content. A value greater than zero indicates the plasmid is AT-richer than the host. Only plasmids with circularised host chromosomes were used (565/726).

307

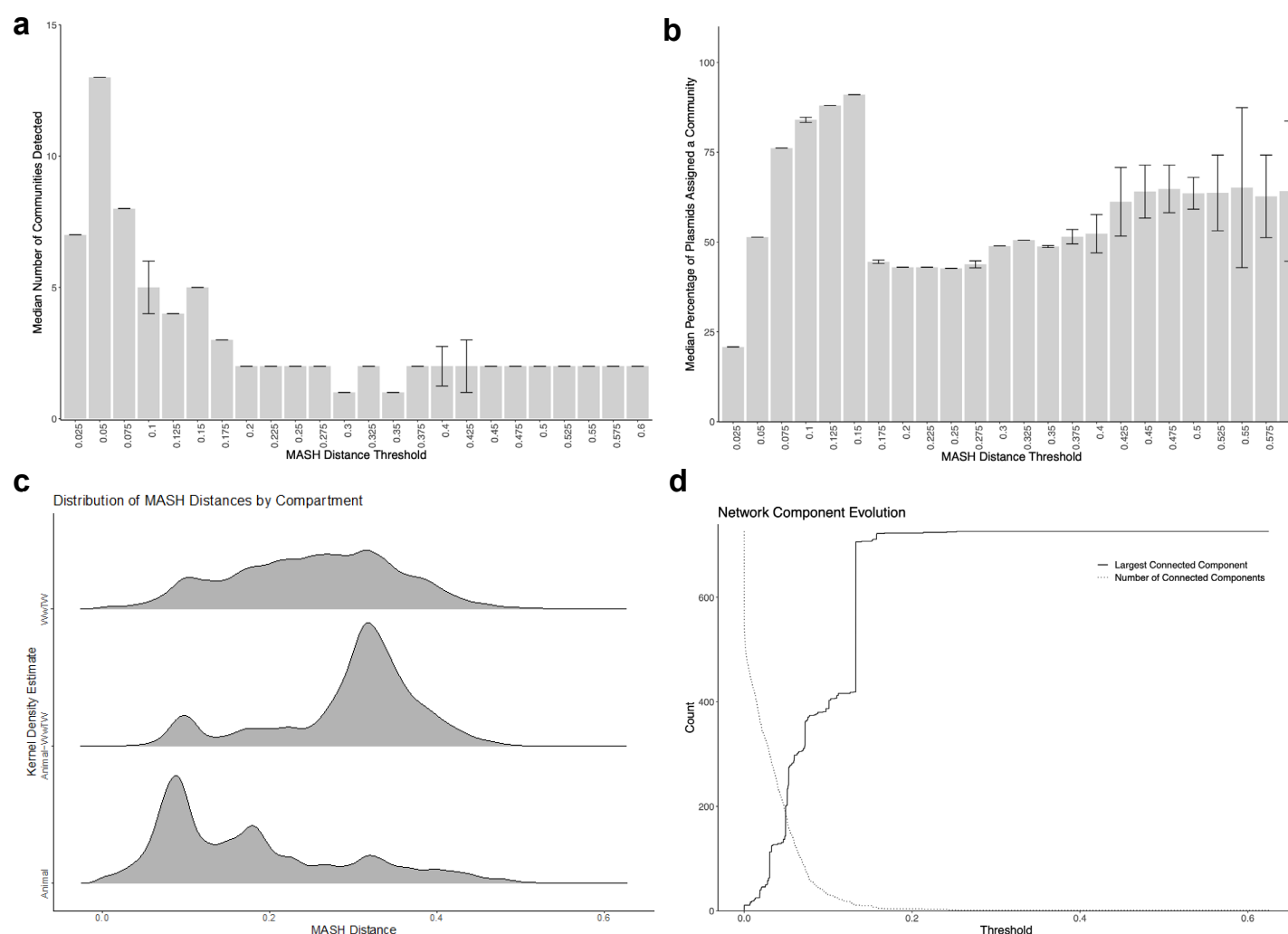
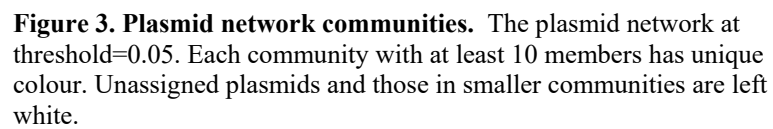


Figure 2. Thresholding the plasmid network. (a) Number of communities (at least 10 nodes) detected over a varying MASH threshold. Median and IQR bars shown. (b) Community coverage of the network over a varying MASH threshold. Median and IQR bars shown. (c) Gaussian kernel density estimates of MASH distances stratified by compartment. Bandwidth calculated by Silverman's 'rule of thumb'. (d) Evolution of the largest connected component and number of components over a varying MASH threshold.



309

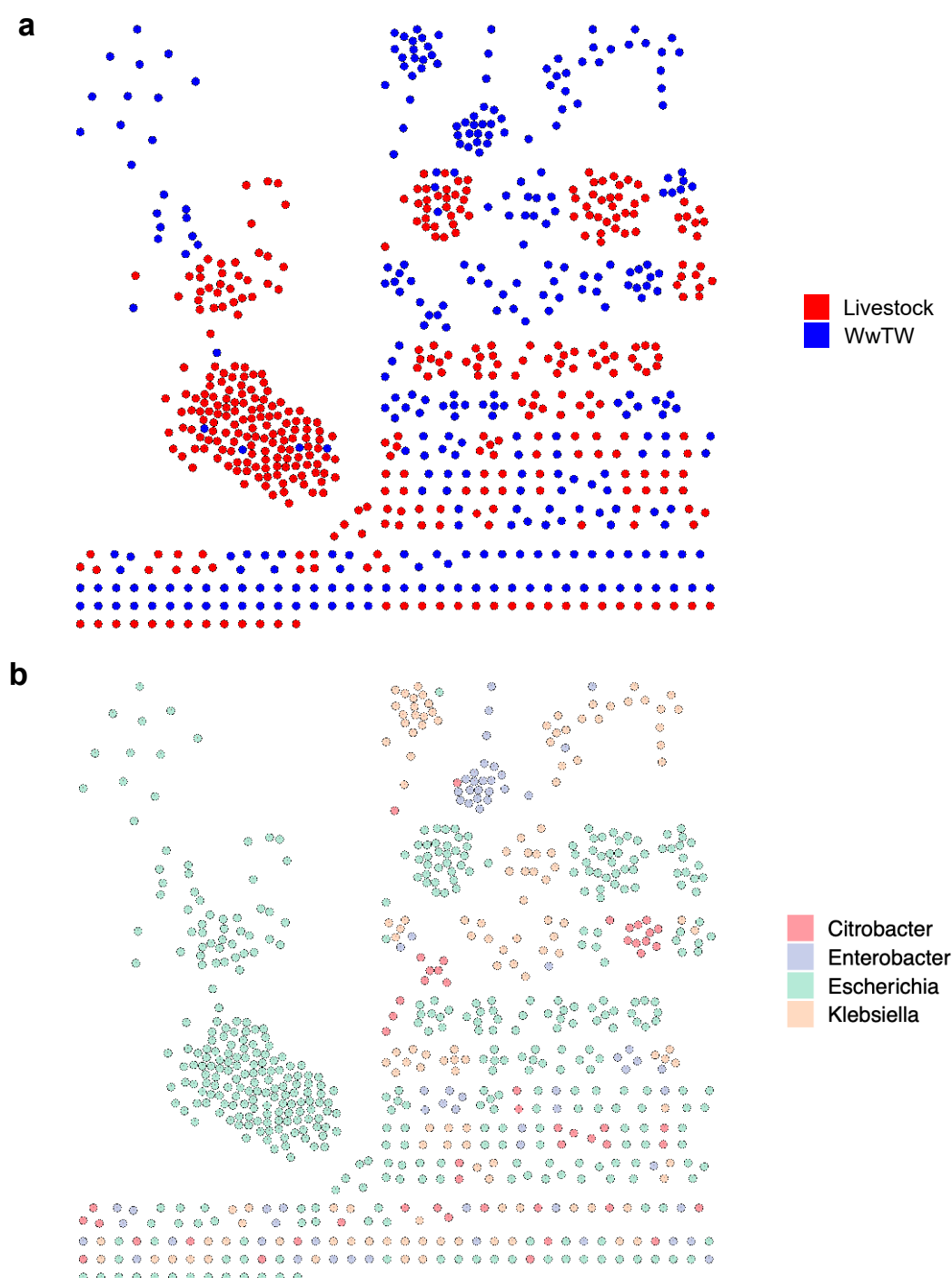


Figure 4. Plasmid network coloured by metadata. All nodes are coloured, not just those in our detected 13 communities of at least 10 members. (a) Partition by livestock or WwTW sampling compartment. (b) Partition by plasmid host genera.

310

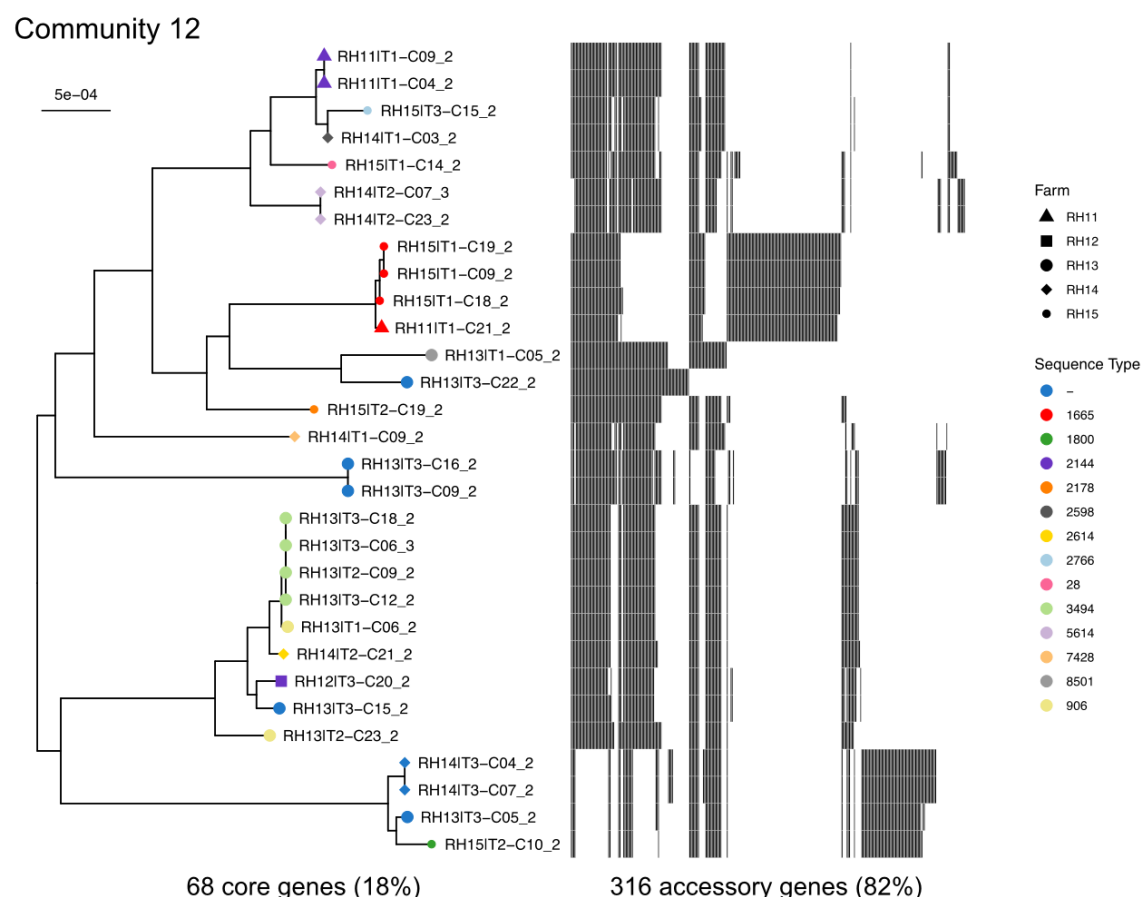


Figure 5. Community core gene phylogeny. A neighbour-joining tree based on alignments of the 68 core genes. A heatmap of the 316 accessory genes is also shown. Node colour represents host sequence type and node shape represents farm. Unknown STs are labelled by '-'. Branch lengths have been corrected for homologous recombination.

311

Table 1. Community metadata homogeneity.							
	Mean \pm sd Homogeneity						
Median \pm IQR Communities with at least 10 Plasmids	Livestock, WwTW	Pig, Cattle, Sheep, WwTW	14 Livestock Farms, WwTW	Livestock, 5 WwTWs	Livestock, Upstream/ Influent, Downstream/ Effluent	Host Genera	Time- point
13 \pm 0	0.715 \pm 0.002	0.591 \pm 0.008	0.403 \pm 0.005	0.467 \pm 0.012	0.550 \pm 0.009	0.888 \pm 0.000	0.050 \pm 0.001

Table 1. Community metadata homogeneity. Homogeneity score averages over 100 runs of the Louvain algorithm for all 13 communities.

Table 2. Community metadata completeness.							
	Mean \pm sd Completeness						
Median \pm IQR Communities with at least 10 Plasmids	Livestock, WwTW	Pig, Cattle, Sheep, WwTW	14 Livestock Farms, WwTW	Livestock, 5 WwTWs	Livestock, Upstream/ Influent, Downstream/ Effluent	Host Genera	Time- point
13 \pm 0	0.199 \pm 0.002	0.332 \pm 0.001	0.400 \pm 0.001	0.238 \pm 0.002	0.211 \pm 0.002	0.307 \pm 0.002	0.023 \pm 0.000

Table 2. Community metadata completeness. Completeness score averages over 100 runs of the Louvain algorithm for all 13 communities.

312

Table 3. Community pangenome results.								
Community	Nodes	Edges	MASH mean	Core Genes	Soft Core Genes	Shell Genes	Cloud Genes	Total Genes
1	61	1252	0.0284	3	6	172	263	444
2	82	1817	0.0314	4	18	138	364	524
3	46	325	0.0355	35	8	86	369	498
4	12	21	0.0382	2	0	290	129	421
5	14	23	0.0383	2	0	225	260	487
6	21	111	0.0366	13	6	354	430	803
7	34	263	0.0344	2	1	278	359	640
8	23	135	0.0222	27	1	142	362	532
9	12	34	0.0344	18	0	364	324	706
10	13	37	0.0233	0	0	309	38	347
11	15	55	0.0194	62	0	116	35	213
12	30	391	0.0242	68	3	126	187	384
13	12	45	0.0223	88	0	195	48	331

Table 3. Community pangenomes. Characteristics of each of the 13 communities, including number of nodes, edges and MASH mean (mean weight of all edges), and gene counts at each level of the pangenome: core genes, soft core genes, shell genes and cloud genes are those found in [100, 99], (99, 95], (95, 15], and (15, 0] percent of plasmids respectively.

Materials and Methods

Plasmids and corresponding host isolates were sampled and sequenced on behalf of the REHAB project in 2017, which aimed to characterise the non-clinical, non-human *Enterobacteriaceae* microbiome in south-central England, with a focus on better understanding antimicrobial resistance (AMR) spread. Specifically, livestock (pig farms, cattle farms and sheep farms) and wastewater treatment work environments (WwTWs; influent, effluent, upstream and downstream waterways) were sampled. To account for seasonal variation, sampling occurred at three discrete time-points (TPs): January-April 2017 (TP1), June-July 2017 (TP2), October-November 2017 (TP3). All the plasmids presented have at least one IncF replicon (classified by with MOB-typer, see below). In total, we present $n=726$ plasmids originated from $n=558$ isolates. This comprises a subset of the entire REHAB dataset, which overall contains $n=2,293$ circularised plasmids recovered from $n=828$ isolates. This dataset is described in more detail [17].

Livestock. Four pig farms (RH01-04), five cattle (RH06-10) and five sheep farms (RH11-15) were selected for sampling over all three TPs. All participating farmers provided written consent for participation. Specific details on farm recruitment and sampling procedure can be found in [17] and Anjum et al. (paper in preparation).

Wastewater treatment works (WwTWs) environment. Five WwTWs (WTP01-05) were selected based on a number of criteria, including; geographic location within the region, wastewater treatment configuration, wastewater population equivalent (PE) served, consented flow, and the accessibility of the effluent receiving river for sampling both upstream and downstream. The chosen WwTWs and their details are shown in Table S5. Sampling took place over all three TPs. Specific details are provided in [17].

DNA sequencing. The isolates were selected for sequencing to represent diversity within the four major genera (*Citrobacter*, *Enterobacter*, *Escherichia* and *Klebsiella*) in each niche, including the use

of third-generation cephalosporin resistance to identify a subset of multi-drug resistant isolates within each genus. Sequencing involved either PacBio SMRT ($n=293$) or Oxford Nanopore Technologies (ONT) ($n=268$) methodologies. Specific details are provided in [17].

Genome assembly, assignment and typing. We used the hybrid assembly and sequencing methods described in our pilot study [30] to produce high-quality *Enterobacteriaceae* genomes from short and long reads. We assigned species and sequence type (ST) from assembled genomes using mlst (version 2.16.43) [31]. Further details on validation are provided in [17].

Plasmid assembly. We used the hybrid assembly and sequencing methods described in a pilot study [30] to produce high-quality *Enterobacteriaceae* genomes from short and long reads. In short, we used Unicycler (version 0.4.7) [32] with ‘normal’ mode, --min_component_size 500, --min_dead_end_size 500, and otherwise default parameters. From these, we selected $n=726$ plasmids which contained an IncF replicon after classification with MOB-typer (see below). We searched all plasmids against PLSDB (version 2020-03-04) [33] which contains 20,668 complete published plasmids, using mash screen [34] and keeping the top hit. All plasmids had a match.

Mobility typing. We used MOB-typer from MOB-suite (version 2.0.0) [35]. We clustered plasmids using MOB-cluster and assigned replicon types with MOB-typer, both part of the MOB-suite. MOB-cluster uses single linkage clustering with a cutoff of a mash distance of 0.05 (corresponding to 95% ANI). A recent large-scale study [12] showed MOB-typer to have a higher correct classification rate than the widely used PlasmidFinder [36].

Plasmid distance estimation. Distances between the complete plasmid sequences was calculated using MASH (version 2.2) [27]. MASH reduces sequences to a fixed-length MinHash sketch, which is used to estimate the Jaccard index. This measures extent of k -mer sharing between plasmids. The representative sketch is far shorter than the original sequence, making distance calculations efficient over large datasets. A k -mer length of 13 and a sketch size of 5000 was used. All other settings were

default. Using MASH considerably reduces distance computation time from exact k -mer profile methods, whilst maintaining good performance.

Louvain community detection. The Louvain algorithm detects communities by optimising the modularity by iterative expectation-maximisation (EM) [19]. This aims to maximise the density of edges within communities against edges between communities. The algorithm was implemented using the python-louvain (version 0.14) Python module.

Community metadata analysis. Homogeneity (h) and completeness (c) are dual conditional entropy-based measures [37]. They are independent of clustering algorithm, dataset size, number of label-types, number of communities and community sizes. This means they are appropriate for uneven metadata distributions. A community partition satisfies homogeneity ($h = 1$) if all members have the same metadata label-type. Suppose we have 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

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

where

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

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 communities and the entropy of the communities, respectively $H(M|C) = 0$ when the community partition coincides with the metadata partition, and no new information is added. A community partition satisfies completeness ($c = 1$) if all

instances of a metadata label-type are assigned the same community. Completeness is defined dually by

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

The measures were calculated using the scikit-learn (version 0.22.2) Python module [38].

Permutation test. We first calculated a Louvain partition for the network and selected all nodes in communities with at least 10 members. Homogeneity and completeness score medians were used from Table 1 and Table 2. The partition labels were then randomly permuted 1,000 times. For each permutation, the homogeneity and completeness scores were calculated. These were then used to calculate a right-tailed *p*-value. The results are shown in Table S2.

Plasmid annotation and pangenome analysis. Plasmids were annotated using Prokka (version 1.14.6) [39]. Pangenome analysis used Panaroo (version 1.2.2) [40]. Core genes, soft core genes, shell genes and cloud genes are those found in [100, 99], (99, 95], (95, 15], and (15, 0] percent of sequences respectively. Within the pangenome, core genes are typically defined as those shared by $\geq 99\%$ of constituent plasmids. However, since no plasmid community in this study had >100 members, core genes were strictly shared by 100%. AMR annotations used Abricate (version 0.9.8) [41] with the NCBI AMRFinder Plus database [42] with a threshold of 90% sequence identity and 90% coverage.

Community phylogeny. Alignment of core genes used Clustal Omega (version 1.2.4) [43], and ClonalFrameML (version 1.2) [44] was used to adjust for homologous recombination. We used ggtree (version 3.11) [45] to visualise the phylogeny.

Data visualization. All figures were made in using the R package ggplot2 (version 3.3.0) [46], except for the network figures (1c, 3 and 4a-b), which were made using Cytoscape (version 3.8.0) [47]. Cytoscape was also used to calculate some network descriptive statistics.

Code and data availability. Plasmid sequence data, metadata (Table S6) and MASH edge list (Table S7) output are available in a figshare collection (<https://doi.org/10.6084/m9.figshare.c.5066684.v1>). Further details on computing methods can be found in the GitHub repository for the paper (<https://github.com/wtmatlock/plasmid-network-analysis>). This includes scripts for calculating the LCC and NCCs, Louvain performance diagnostics, and running the permutation test. Other data can be found in [17].

Acknowledgements. The REHAB consortium is represented by the following: AbuOun M, Anjum MF, Bailey MJ, Brett H, Bowes M, Chau KK, Crook DW, de Maio N, Duggett N, Wilson DJ, Gilson D, Gweon HS, Hubbard A, Hoosdally SJ, Matlock W, Kavanagh J, Jones H, Peto TEA, Read DS, Sebra R, Shaw LP, Sheppard AE, Smith R, Stubberfield E, Stoesser N, Swann J, Walker AS, Woodford N. Also, thanks to Fowler P for his comments on the draft.

Author contributions. Author contributions under the CRediT system were as follows:

Conceptualisation: WM, NS, MA, DS, MJB, DWC, LPS, ASW

Methodology: WM, LPS

Software: WM

Validation: WM, KKC, LB, HP, LPS

Formal analysis: WM

Investigation: KKC, MA, ES, JK, HP, LB, RS, DSR, HSG, NS, RS

Resources: MA, MFA, HSG, DSR, RS, JS, NS, TEAP, MJB, ASW, RS

Data curation: WM, LPS, DSR, MA, NS, ES, DG

Writing – original draft: WM

Writing – review and editing: All authors

445 Visualisation: WM

446 Supervision: LPS, NS, ASW, DWC

447 Project administration: NS, DSR, SH, MFA

448 Funding acquisition: NS, DWC, MJB, DSR, MFA, ASW, TEAP

449

450 **Competing interest declarations.** The authors declare no competing interests.

451

452 **Funding.** This work was funded by the Antimicrobial Resistance Cross-council Initiative supported
453 by the seven research councils [grant NE/N019989/1]. Crook, George, Peto, Sheppard, Stoesser, and

454 Walker are supported by the National Institute for Health Research Health Protection Research Unit

455 (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at the University of

456 Oxford in partnership with Public Health England (PHE) [grant HPRU-2012–10041 and

457 NIHR200915]. Walker, Crook, and Peto are also supported by the NIHR Oxford Biomedical

458 Research Centre. Walker is an NIHR Senior Investigator. The computational aspects of this research

459 were funded from the NIHR Oxford BRC with additional support from a Wellcome Trust Core

460 Award Grant [grant 203141/Z/16/Z]. The views expressed are those of the authors and not necessarily

461 those of the NHS, the NIHR, the Department of Health or Public Health England. Matlock is

462 supported by a scholarship from the Medical Research Foundation National PhD Training Programme

463 in Antimicrobial Resistance Research (MRF-145-0004-TPG-AVISO).

464

References

1. Thanner, S., D. Drissner, and F. Walsh, Antimicrobial resistance in agriculture. MBio, 2016. 7(2).
2. Wyres, K.L. and K.E. Holt, Klebsiella pneumoniae as a key trafficker of drug resistance genes from environmental to clinically important bacteria. Current opinion in microbiology, 2018. 45: p. 131-139.
3. Collis, R.M., et al., Extended-Spectrum Beta-Lactamase-Producing Enterobacteriaceae in Dairy Farm Environments: A New Zealand Perspective. Foodborne pathogens and disease, 2019. 16(1): p. 5-22.
4. Velasova, M., et al., Detection of extended - spectrum β - lactam, AmpC and carbapenem resistance in Enterobacteriaceae in beef cattle in Great Britain in 2015. Journal of applied microbiology, 2019. 126(4): p. 1081-1095.
5. AbuOun, M., et al., Characterizing Antimicrobial Resistant Escherichia coli and Associated Risk Factors in a Cross-Sectional Study of Pig Farms in Great Britain. Frontiers in Microbiology, 2020. 11.
6. Bartley, P.S., et al., Antibiotic resistance in Enterobacteriaceae from surface waters in urban Brazil highlights the risks of poor sanitation. The American journal of tropical medicine and hygiene, 2019. 100(6): p. 1369-1377.
7. Decano, A.G. and T. Downing, An Escherichia coli ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. Scientific reports, 2019. 9(1): p. 1-13.
8. Passarelli - Araujo, H., et al., Genomic analysis unveils important aspects of population structure, virulence, and antimicrobial resistance in Klebsiella aerogenes. The FEBS journal, 2019. 286(19): p. 3797-3810.

9. Nakamura, K., et al., Differential dynamics and impacts of prophages and plasmids on the pangenome and virulence factor repertoires of Shiga toxin-producing *Escherichia coli* O145: H28. *Microbial Genomics*, 2020: p. mgen000323.
10. Woolhouse, M., et al., Antimicrobial resistance in humans, livestock and the wider environment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2015. 370(1670): p. 20140083.
11. Allcock, S., et al., Antimicrobial resistance in human populations: challenges and opportunities. *Global health, epidemiology and genomics*, 2017. 2.
12. Douarre, P.-E., et al., Analysis of COMPASS, a new comprehensive plasmid database revealed prevalence of multireplicon and extensive diversity of IncF plasmids. *Frontiers in Microbiology*, 2020. 11: p. 483.
13. Agyekum, A., et al., blaCTX-M-15 carried by IncF-type plasmids is the dominant ESBL gene in *Escherichia coli* and *Klebsiella pneumoniae* at a hospital in Ghana. *Diagnostic microbiology and infectious disease*, 2016. 84(4): p. 328-333.
14. Irrgang, A., et al., CTX-M-15-producing *E. coli* isolates from food products in Germany are mainly associated with an IncF-type plasmid and belong to two predominant clonal *E. coli* lineages. *Frontiers in microbiology*, 2017. 8: p. 2318.
15. Kopotsa, K., J. Osei Sekyere, and N.M. Mbelle, Plasmid evolution in carbapenemase - producing *Enterobacteriaceae*: a review. *Annals of the New York Academy of Sciences*, 2019. 1457(1): p. 61-91.
16. Rozwandowicz, M., et al., Plasmids carrying antimicrobial resistance genes in *Enterobacteriaceae*. *Journal of Antimicrobial Chemotherapy*, 2018. 73(5): p. 1121-1137.
17. Shaw, L.P., et al., Niche and local geography shape the pangenome of wastewater- and livestock-associated *Enterobacteriaceae*. *bioRxiv*, 2020: p. 2020.07.23.215756.

513 18. Orlek, A., et al., Plasmid classification in an era of whole-genome sequencing:
514 application in studies of antibiotic resistance epidemiology. *Frontiers in microbiology*, 2017.
515 8: p. 182.

516 19. Blondel, V.D., et al., Fast unfolding of communities in large networks. *Journal of*
517 *statistical mechanics: theory and experiment*, 2008. 2008(10): p. P10008.

518 20. Yamashita, A., T. Sekizuka, and M. Kuroda, Characterization of antimicrobial
519 resistance dissemination across plasmid communities classified by network analysis.
520 *Pathogens*, 2014. 3(2): p. 356-376.

521 21. Branger, C., et al., Extended-spectrum β -lactamase-encoding genes are spreading on a
522 wide range of *Escherichia coli* plasmids existing prior to the use of third-generation
523 cephalosporins. *Microbial genomics*, 2018. 4(9).

524 22. Acman, M., et al., Large-scale network analysis captures biological features of
525 bacterial plasmids. *bioRxiv*, 2019: p. 785212.

526 23. Redondo-Salvo, S., et al., Pathways for horizontal gene transfer in bacteria revealed
527 by a global map of their plasmids. *Nature Communications*, 2020. 11(1): p. 1-13.

528 24. Jesus, T.F., et al., Plasmid ATLAS: plasmid visual analytics and identification in
529 high-throughput sequencing data. *Nucleic acids research*, 2019. 47(D1): p. D188-D194.

530 25. Almpanis, A., et al., Correlation between bacterial G+ C content, genome size and the
531 G+ C content of associated plasmids and bacteriophages. *Microbial genomics*, 2018. 4(4).

532 26. Dietel, A.-K., et al., Selective advantages favour high genomic AT-contents in
533 intracellular elements. *PLoS genetics*, 2019. 15(4): p. e1007778.

534 27. Ondov, B.D., et al., Mash: fast genome and metagenome distance estimation using
535 MinHash. *Genome biology*, 2016. 17(1): p. 132.

536 28. Traag, V.A., Faster unfolding of communities: Speeding up the Louvain algorithm.
537 *Physical Review E*, 2015. 92(3): p. 032801.

538 29. Que, X., et al. Scalable community detection with the louvain algorithm. in 2015
539 IEEE International Parallel and Distributed Processing Symposium. 2015. IEEE.

540 30. De Maio, N., et al., Comparison of long-read sequencing technologies in the hybrid
541 assembly of complex bacterial genomes. *Microbial genomics*, 2019. 5(9).

542 31. Seeman, T., mlst.

543 32. Wick, R.R., et al., Unicycler: resolving bacterial genome assemblies from short and
544 long sequencing reads. *PLoS computational biology*, 2017. 13(6): p. e1005595.

545 33. Galata, V., et al., PLSDB: a resource of complete bacterial plasmids. *Nucleic acids*
546 *research*, 2019. 47(D1): p. D195-D202.

547 34. Ondov, B.D., et al., Mash Screen: high-throughput sequence containment estimation
548 for genome discovery. *Genome biology*, 2019. 20(1): p. 232.

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

551 36. Carattoli, A., et al., In silico detection and typing of plasmids using PlasmidFinder
552 and plasmid multilocus sequence typing. *Antimicrobial agents and chemotherapy*, 2014.
553 58(7): p. 3895-3903.

554 37. Rosenberg, A. and J. Hirschberg. V-measure: A conditional entropy-based external
555 cluster evaluation measure. in *Proceedings of the 2007 joint conference on empirical methods*
556 *in natural language processing and computational natural language learning (EMNLP-*
557 *CoNLL)*. 2007.

558 38. Pedregosa, F., et al., Scikit-learn: Machine learning in Python. *the Journal of machine*
559 *Learning research*, 2011. 12: p. 2825-2830.

560 39. Seemann, T., Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 2014.
561 30(14): p. 2068-2069.

40. Tonkin-Hill, G., et al., Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline. *BioRxiv*, 2020.
41. Seemann, T., Abricate: mass screening of contigs for antimicrobial and virulence genes. Department of Microbiology and Immunology, The University of Melbourne, Melbourne, Australia. Available online: <https://github.com/tseemann/abricate> (accessed on 28 February 2019), 2018.
42. Feldgarden, M., et al., Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrobial agents and chemotherapy*, 2019. 63(11): p. e00483-19.
43. Sievers, F. and D.G. Higgins, Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, 2018. 27(1): p. 135-145.
44. Didelot, X. and D.J. Wilson, ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS computational biology*, 2015. 11(2).
45. Yu, G., et al., ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 2017. 8(1): p. 28-36.
46. Wickham, H., ggplot2: elegant graphics for data analysis. 2016: Springer.
47. Shannon, P., et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 2003. 13(11): p. 2498-2504.