Large-scale network analysis captures biological features of bacterial plasmids

Mislav Acman[1], Lucy van Dorp[1], Joanne M. Santini[2] and Francois Balloux[1]

[1] UCL Genetics Institute, University College London, Gower Street, London WC1E 6BT, UK

[2] Institute of Structural & Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK

Correspondence: Mislav Acman: mislav.acman.17@ucl.ac.uk and Francois Balloux: f.balloux@ucl.ac.uk

Keywords: plasmid classification, Jaccard index, community detection algorithm, horizontal gene transfer, mobile genetic elements, antimicrobial resistance

Abstract

Many bacteria can exchange genetic material through horizontal gene transfer (HGT) mediated by plasmids and plasmid-borne transposable elements. Here, we study the population structure and dynamics of over 10,000 bacterial plasmids, by quantifying their genetic similarities and reconstructing a network based on their shared *k*-mer content. We use a community detection algorithm to assign plasmids into cliques, which correlate with plasmid gene content, bacterial host range, GC content, and existing classifications based on replicon and mobility (MOB) types. Further analysis of plasmid population structure allows us to uncover candidates for yet undescribed replicon genes and to identify transposable elements as the main drivers of HGT at broad phylogenetic scales. Our work illustrates the potential of network-based analyses of the bacterial 'mobilome' and opens up the prospect of a natural, exhaustive classification framework for bacterial plasmids.

Introduction

Plasmids are extra-chromosomal DNA molecules found across all three Domains of Life. In bacteria, they are considered one of the main mediators of horizontal gene transfer (HGT) through the processes of conjugation and transformation[1–3]. Plasmids generally harbour non-essential genes that can modulate the fitness of their bacterial host. Some prominent examples include toxin-antitoxin systems, virulence factors, metabolic pathways, antibiotic biosynthesis, metal resistance and antimicrobial resistance (AMR) genes. These accessory genes can be located on transposable elements involved in lateral gene transfer across genomes and can thus lead to a highly mosaic structure of plasmid genomes[4]. The mix of vertical and horizontal inheritance of plasmids, together with exchanges of plasmid-borne genes, generates complex dynamics that are difficult to capture with classical population genetics tools and make it challenging to classify plasmids within a coherent universal framework.

Currently, there are two well-established plasmid classification schemes which attempt to bin plasmids according to their propagation mechanisms, while indirectly capturing some features of the plasmid backbone. The first scheme is based on replicon types[5] and the second on mobility (MOB) groups[6]. Replicon-based typing relies on relatively conserved genes of the replicon region which encode the plasmid replication and partitioning machinery[5]. Plasmids with matching replication or partitioning systems cannot stably coexist within the same cell. Conversely, MOB typing is used to classify self-transmissible and mobilizable plasmids into six MOB types[6]. The MOB typing scheme relies on the conserved N-terminal sequence of the relaxase, a site-specific DNA endonuclease which binds to the origin of transfer (*oriT*) cleaving at the *nic* site and is essential for plasmid conjugation.

Despite being widely used and informative, these typing schemes only work within a limited taxonomic range[7–9]. Replicon typing is dependent on the availability of prior experimental evidence and remains restricted to culturable bacteria from the family *Enterobacteriaceae* and several well-studied genera of gram-positive bacteria[1,10–12]. Furthermore, this approach can lead to ambiguous classification, even for experimentally validated replicons, as recently demonstrated by the discovery of compatible plasmids assigned to the same replicon type, which led to the further subdivision of the IncK type into IncK1 and IncK2[13], and IncA/C type into IncA and IncC[14]. In addition, plasmids can carry genes from more than one replication machinery leading to assignment to multiple replicon types, further reducing interpretability[7,8]. MOB typing schemes generate fewer multiple assignments and can cover a potentially wider taxonomic range, however they are not applicable to the classification of non-mobilizable plasmids. These two typing schemes have inspired several *in silico* classification tools, such as PlasmidFinder[12], the plasmid MultiLocus Sequence Typing (MLST) database, and MOB-suite[15]. However, all of these tools intrinsically rely on the completeness of their reference sequence databases, which typically lack representatives from understudied and/or unculturable bacterial hosts.

As bacterial plasmids undergo extensive recombination and HGT, their evolutionary history is not well captured by phylogenetic trees, which are designed for the analysis of point mutations in sequence alignments[16,17]. Network models offer an attractive alternative given they can incorporate both horizontal and vertical inheritance[18,19], and can deal with point mutations as well as structural variants. Networks have gained much attention in the past decade as an alternative method for studying prokaryotic evolution, including plasmids[3,8,18–20]. Plasmid gene-sharing networks have proven a useful means to track AMR and virulence dissemination yielding deeper insights

59 into HGT events[17,21,22]. However, the main drawback of previous work relying on plasmid sequence alignments
60 is the exclusion of important non-coding elements such as non-coding RNAs, promoter regions, CRISPRs,
61 stretches of homologous sequences, or putative, disrupted and currently unannotated genes. A more
62 comprehensive approach could consider a plasmid network based on estimates of alignment-free sequence
63 similarity[23]. Alignment-free genetic distance methods are becoming established tools for the analysis of large
64 genomic datasets, and their usefulness has been validated in both prokaryotes and eukaryotes[19,23–26]. A recently
65 published Plasmid ATLAS tool by Jesus *et al.*[27] provides an illustration of such an approach, with a network of
66 plasmids constructed based on pairwise genetic distances estimated using alignment-free *k*-mer matching methods
67 implemented in Mash[28].

68 In this work, we quantify the genetic similarity between more than 10,000 bacterial plasmids available on NCBI's
69 RefSeq database and construct a network which reflects their relatedness based on shared *k*-mer content. Applying
70 a community detection algorithm allows us to cluster plasmids with high genetic similarity into cliques (complete
71 subgraphs) revealing a strong underlying population structure. We find cliques to be highly correlated with the
72 gene content of the plasmid backbone, bacterial host and GC content, as well as replicon and MOB types.
73 Uncovering the structure of the full plasmid population further allows the discovery of candidates for yet-
74 undescribed replicon genes and provides insight into broad-scale plasmid dynamics. Taken together, our results
75 illustrate the potential of network-based analyses of plasmid sequences and opens up the prospect of a natural,
76 exhaustive classification framework for bacterial plasmids.

Results

A dataset of complete bacterial plasmids

A dataset of complete bacterial plasmids was assembled comprising 10,696 sequences found in bacteria from 22 phyla and over 400 genera (Supplementary Data 1, Figure 1A, and Supplementary Figure 1). The composition of plasmid hosts reflects current research interests, with the Proteobacteria and Firmicutes phyla together representing over 84% of plasmid sequences. Though, the dataset includes plasmids from a diversity of bacterial hosts, with 66 plasmids from unknown bacterial families, 14 from uncultured bacteria and 37 samples from *candidatus* species (Supplementary Data 1). In total, 510,463 different Coding Sequences (CDSs) were identified in the plasmid dataset. 66.01% of the CDSs were predicted to encode a hypothetical protein, 27.9% had a known product with Gene Ontology (GO) biological process annotation, with the remaining 6.09% encoded a known protein product with unknown biological function (Figure 1B). There are 3,328,916 bacterial genes available in the RefSeq database (NCBI Gene Statistics accessed on June 19th 2019), meaning that roughly one in twenty of the currently known bacterial genes are plasmid-borne. The GO biological processes associated with plasmid CDSs are diverse. After accounting for multiple occurrences of annotated CDSs in the dataset, the dominant associated GO terms relate to catabolic and biosynthetic processes (20.64% relative to total number of annotated CDSs), transposon mobility (17.09%) and positive and negative regulation of transcription (7.70%). Replicon-based typing classified 27.66% of the plasmids into 163 different replicon types (Figure 1C and Supplementary Figure 2). However, 31.67% of these classified plasmids were assigned to multiple replicon types. MOB typing was more comprehensive, successfully classifying 32.63% of the plasmids into six MOB types of which 9.48% were assigned to multiple types (Figure 1C). Unsurprisingly, classification by these two methods performed best for well-studied plasmids of the phyla Proteobacteria and Firmicutes.

Uncovering the population structure of plasmids

We constructed a network based on the plasmid pairwise sequence similarities. This represents a weighted, undirected network with plasmids (vertices) connected by edges indicating similarity (Supplementary Figure 3). Similarity was scored using the exact Jaccard index (JI), defined as the size of the intersection divided by the size of the union of two sets of *k*-mers. Plasmid pairs which shared less than 100 *k*-mers were considered to have a JI equal to zero. This cut-off value was implemented since the majority of CDSs found on plasmids have lengths greater than 100bp, thus only a fraction of the functional genome is common between plasmids with low shared *k*-mer count (Supplementary Figures 4 and 5). The majority of plasmid pairs shared little to no similarity (Figure 1D). 6.14% (657) of the plasmids were singletons, whilst 3.31% (354) were connected to only one other plasmid, illustrating the high levels of diversity across bacterial plasmid genomes. It follows that plasmids with more *k*-mers in common are more likely to share the same functional genetic elements and hence participate in similar biological processes falling within the same host niche (Supplementary Figure 5). Such plasmids are presumed to form cliques within the network with higher internal JI score. The objective is then to identify cliques which contain plasmids with markedly higher similarity between themselves, relative to their immediate network neighbourhood.

113 Listing all cliques of our large plasmid network and assessing their internal similarity is computationally
114 intractable with current tools[29]. A solution for a single clique can be quickly verified, but the time required to
115 process all possible cliques scales rapidly as the size of the network increases. As an alternative solution, a
116 stochastic community detection algorithm OSLOM (Ordered Statistics Local Optimization Method) was
117 implemented[30]. OSLOM detects communities (i.e. densely interconnected subgraphs) with statistical significance,
118 meaning that they have a low probability of being encountered by chance in a random network with similar
119 features to the plasmid network. OSLOM is well suited for this task since it can be used to analyse undirected
120 networks with overlapping communities or hierarchical structures. In addition, OSLOM shows similar
121 performance to other widely used methods such as Infomap or Louvain[30,31] which, unlike OSLOM, were unable
122 to analyse this dataset due to computational limitations. To validate the results from the stochastic clique
123 assignment, all communities of size three or more detected by OSLOM were assessed for their completeness (i.e.
124 whether they form cliques) against the original plasmid network (Supplementary Figure 3).

125 Despite the notable dissimilarity among plasmids, the original network was too dense (network density = 0.0438)
126 to yield a consistent performance for every OSLOM run (Figure 2 and Supplementary Figures 3 and 6).
127 Furthermore, a large proportion of communities detected did not form cliques and would have to be disregarded
128 (Figure 2A). A JI threshold was introduced to increase the sparsity of the network and to upweight more similar
129 plasmids, thus optimizing the performance of OSLOM. A range of thresholds were assessed based on the
130 following criteria: (i) the clique to community ratio (Figure 2A), (ii) the proportion of plasmids assigned to cliques
131 (Figure 2B), (iii) the congruence with replicon-based typing (Figure 2C), and (iv) the consistency of OSLOM
132 performance (Figure 2 and Supplementary Figure 6). The optimum threshold was consistently obtained at a JI of
133 0.3. The resulting sparse network is shown in Figure 3 (network density = 0.00128).

134 The OSLOM-guided clique detection algorithm offers flexibility and identifies cliques of plasmids with a wide
135 range of internal similarity scores (Supplementary Figure 7). We assessed the importance of considering pairwise
136 JI distances as a continuous variable by reanalysing the dataset with the Bron-Kerbosch Max-clique algorithm[32],
137 implemented in the graph-tool Python library[33]. The Bron-Kerbosch algorithm is computationally highly
138 effective, but the pairwise distances between plasmids are treated as binary values defined by the given threshold.
139 Applied across a range of JI thresholds, the Max-clique approach systematically identifies a very large number of
140 cliques (Supplementary Figure 8A), assigns a large proportion of plasmids to multiple cliques (Supplementary
141 Figure 8B) and leads to a low correlation between resulting cliques and plasmid replicon types (Supplementary
142 Figure 8C).

143 Plasmid cliques agree with current typing schemes

144 Analysis of the sparse network with OSLOM successfully assigned 50.21% (5371) of the plasmids into 561
145 cliques of size three or more (Figure 1C, Figure 3, and Supplementary Figure 14). Only 1.64% (88) of these
146 plasmids were assigned to multiple cliques, and these were found in the densest regions of the network and at the
147 interfaces between cliques indicating the presence of 'chimeric plasmids' (i.e. hybrid plasmids generated through
148 merging of two different plasmids), large-scale transposition or recombination events, or extensive repeated
149 transposition/recombination (Figure 1C and Figure 3). In addition, this approach covered 564 plasmids from phyla
150 other than the Proteobacteria and Firmicutes, namely from Spirochaetes, Chlamydiae, Actinobacteria,

151 Tenericutes, Bacteroidetes, Cyanobacteria, and Fusobacteria. Interestingly, after applying the 0.3 JI threshold,
152 38.01% (4066/10696) of plasmids that could not be assigned to cliques of size three or more, were separated from
153 the network as singletons, while 10.10% (1080) shared an edge with a single plasmid. Therefore, only 1.67%
154 (179) of plasmids were effectively left unassigned. Nonetheless, due to the apparent lack of shared genetic signal,
155 plasmid singletons and pairs were not considered in subsequent analyses. To assess the extent to which 'mobile
156 elements' shared between plasmids affect the classification into cliques, we repeated the clique assignment
157 analyses after having removed all accessory CDSs (29,913) associated with transposition, pathogenesis, or
158 resistance (Supplementary Figure 9). Pruning these genes did not markedly affect the assignment of plasmids into
159 cliques, which gives support to the genetic signal being driven by the genetic similarity of plasmid backbones
160 rather than shared mobile genetic elements.

161 Clique purity and Normalized Mutual Information (NMI) were used to assess the quality of clique-based
162 classification (see Methods). These metrics were calculated for cliques comprising plasmids with identified
163 replicon type, plasmids carrying a single identified replicon type, or plasmids with assigned MOB type. Untyped
164 plasmids were disregarded. The observed purity scores were high (>85%) indicating the homogeneity of cliques
165 for a particular plasmid type (Supplementary Figure 10). This was particularly the case for MOB types (purity =
166 0.9887) and plasmids assigned to a single replicon type (purity = 0.9522). NMI provides an entropy-based measure
167 of the similarity between two classification systems where a score equal to one indicates identical partitioning
168 into classes while zero means independent classification. NMI penalizes differences in the number of assignment
169 classes which justifies the low score observed when assessing clique-based versus MOB-based typing (NMI =
170 0.5223). Nevertheless, high NMI scores were obtained when considering a replicon-based classification scheme
171 (NMI = 0.9044 all types, and NMI = 0.9336 for single replicon types). It follows that plasmids with the same
172 replicon type often fall together within the same clique. This is also supported by the high correlation between the
173 clique membership size and the number of plasmids assigned to the corresponding replicon class (Supplementary
174 Figure 11, R^2=0.862 for plasmids assigned to a single replicon types).

175 There are exceptions where plasmids from larger replicon classes are further resolved into a few smaller
176 evolutionary related cliques. One such example is provided by the 22 'broad-host-range' IncP plasmids which
177 have been split into three cliques (14, 118 and 332) (Supplementary Figure 12, Supplementary Data 1). While
178 plasmids within these cliques share notably high JI similarity, the similarities between cliques remain low. This
179 is especially true for clique 332 and 14, for which between-clique similarity is zero. Interestingly, plasmids from
180 clique 332 have been exclusively associated with Gammaproteobacteria, while the ones from cliques 118 and 14
181 are mostly found in hosts from the Betaproteobacteria class. This arrangement of IncP into multiple cliques with
182 a more constrained host-range is in line with previous findings of weaker incompatibilities in IncP[34] and the
183 existence of multiple genetically distinct IncP sub-lineages whose backbone is coadapted to their host[35]. Another
184 example of a genetically heterogeneous replicon type is provided by IncY and p0111 plasmids collected from *E.*
185 *coli* strains which fall into three cliques (119, 230 and 372) (Supplementary Figure 13). Clique 119 and 372 cluster
186 IncY and p0111 plasmids respectively, with a single, possibly misplaced IncFIB plasmid. Conversely, clique 230
187 comprises both IncY and p0111 plasmids, with a remarkably related genetic backbone. The latter result raises
188 questions on the distinctiveness of IncY and p0111 plasmid types.

Candidate replicon genes recovered from untyped plasmids

The majority of plasmids with unknown replicon types formed small cliques (Supplementary Figure **14**). In fact, 81.02% of the smallest cliques (carrying three to five plasmids) contain exclusively untyped plasmids. Together with the aforementioned singletons and lone plasmid pairs, this trend highlights the many understudied and underrepresented plasmids in sequence databases. Accordingly, the next objective was to investigate the genetic content of untyped cliques to determine candidate replicon genes and further traits of biological relevance.

In total, there are 388 cliques with no assigned replicon types. As the cliques tend to be homogeneous for a replicon type, only the core genes (i.e. genes occurring on all plasmids of a particular clique) found on untyped cliques were considered. Core genes were translated into protein sequences and screened against the translated PlasmidFinder database using TBLASTN[36]. A range of e-values were assessed to determine the threshold maximizing the discovery of replicon candidates while minimizing false positives (Supplementary Figure 15). The majority of plasmids were assigned to one replicon type with some plasmids having hits to a maximum of three to four different types. Accordingly, the optimal e-value threshold was selected when the total number of core gene hits started to diverge from the number of untyped cliques covered. A conservative e-value threshold of 0.001 was chosen which resulted in the identification of 105 candidate genes from 106 plasmid cliques. The accession numbers and positions of candidate genes are listed in the Supplementary Data 1 (*Candidate_replicon_gene* column) for all carrier plasmids.

To verify the plausibility of the identified gene candidates, HMMER (version 3.2.1) was used to scan amino acid sequences for known protein domain families found in the Pfam database (version 32.0)[37]. 166 families, with e-values lower than 0.001, were identified on 97 protein sequences and were most commonly associated with replication initiation (Supplementary Figure 16). Moreover, the majority of functions associated with the discovered protein families relate to plasmid replicon proteins. For example, domains with helix-turn-helix motifs are important for DNA binding of replicon proteins and allow some proteins to regulate their own transcription[38]. Other examples of transcriptional regulators also exist in plasmid replicon regions, while DNA primase activity has been found on the RepB replicon protein[38]. Interestingly, replicon proteins involved in rolling-circle replication (a mechanism of plasmid replication) share some of their motifs with proteins involved in plasmid transfer and mobilization[38]. This could explain why some of the discovered domain families are linked to plasmid mobilization. On the whole, the candidate replicon genes are highly specific to a particular clique of plasmids and should assist description of new incompatibility types.

Cliques exhibit common GC content and bacterial hosts

The unprocessed plasmid network exhibited a pronounced structure in terms of the plasmid nucleotide composition, measured by GC content (Supplementary Figure 3). This trend was also reflected in the clique composition (Supplementary Figure 17A). Within a clique, the standard deviation of GC content rarely exceeds 0.02 and is weakly correlated with the clique size (R^2 = 0.0155) (Supplementary Figure 17B). Moreover, a significant difference in GC content is often found between cliques. Analysis of variance (ANOVA), followed by a Tukey test, found that 85.3% of the time the GC content between two cliques differs significantly (adjusted *p*-value < 0.001). In contrast, the sequence lengths of plasmids within a clique are more variable, but are also not

strongly correlated with clique size (R^2 = 0.029) (Supplementary Figure 17C and 17D). Similarly, a Tukey test showed that a significant difference in plasmid length between cliques is observed less than 34% of the time (adjusted p-value < 0.001).

Plasmid GC content has been shown to be strongly correlated to the base composition of the bacterial host's chromosome[39]. Indeed, the cliques showed a very high homogeneity (purity) relative to their hosts (Supplementary Figure 18), a trend which has been identified in other plasmid network reconstruction efforts[21]. At higher taxonomic levels, cliques have near perfect purity scores (>0.99). The purity score slightly decreases at the level of the plasmid host family, reaching a value of 0.807 at the species level. Therefore, plasmids with high genetic similarity rarely transcend the level of the bacterial genus, which suggests a limited host range for the vast majority of plasmids. However, these results need to be carefully considered due to inherent biases in the dataset, especially in terms of the predominance of well-studied taxa. Overall, the plasmid cliques show a strong intrinsic propensity towards confined GC content and are found in a limited range of bacterial hosts.

Plasmids within cliques have uniform gene content

The gene content of cliques was assessed for all genes occurring five or more times in the dataset. This threshold was chosen to facilitate computation, and to adequately characterize more prevalent genes. In total, 15,851 out of 35,883 (44.17%) of the assessed genes were 'core' genes, meaning they had a within-clique frequency equal to one, suggesting an overall uniformity of gene content in cliques (Supplementary Figure 19). Furthermore, 6,577 (18.33%) of the genes were 'private'. Private genes are those found in only one clique, with a frequency of one, and their relatively high abundance in the dataset suggests the uniqueness of some cliques with respect to their gene content. However, there is an inherent bias. Plasmids within larger cliques tend to be more dissimilar and share proportionally fewer genes (Supplementary Figure 20). This pattern can in part be explained by the broader gene content of large cliques and the high sequence similarity required for same-gene clustering (95%) within the default implementation of the Prokka-Roary annotation pipeline. 31.94% of cliques containing five or more plasmids were found to have one to 10 core genes. However, cliques exhibited a wide range in the number of core genes with 7.74% of cliques carrying over 100 shared genes. Interestingly, 13.55% (42) of cliques had no core genes which could also be an artefact of the gene annotation pipeline sensitivity or poor-quality assemblies. For instance, plasmids from 19 cliques carried no recognized genes from the pool of 35,883 assessed genes. Functionally, core genes were found to be more often associated with various metabolic processes, transcription regulation and transmembrane transport (Supplementary Figure 21) when compared to the overall distribution of GO terms, shown in Figure 1B. Similarly, fewer core genes were involved in transposon movement, pathogenesis, and resistance.

Inferring horizontal gene transfer through clique interactions

Gene content was also considered in the context of clique structure and interconnectedness. To do so, the original network of plasmids (Supplementary Figure 3) was rearranged such that: (i) plasmids assigned to the same clique were clustered under a single vertex; (ii) plasmids assigned to multiple cliques were left as solitary vertices anchoring the cliques; (iii) unassigned plasmids were removed. The resulting network is shown in Figure 4. As highlighted earlier, large cliques generally show lower internal similarity compared to the smaller ones. It is

important to note that an arbitrary JI threshold of 0.01 was introduced in Figure 4 to assist visual interpretation, but the unfiltered version of the network is provided in Supplementary Figure 22.

The clustering of cliques in Figure 4 shows high concordance with the phylogenetic hierarchy of the bacterial hosts. On a global scale, there are four large interconnected clusters (three corresponding to cliques from the phylum Firmicutes and one from the Proteobacteria), eight disjointed clusters, and a dozen singled-out triplets and pairs. The clique clusters mostly contain plasmids from a specific genus with some minor deviations – hence the cluster naming. The only two exceptions are the large and diverse Proteobacteria cluster which harbours plasmids mainly from the genera *Escherichia*, *Klebsiella*, and *Salmonella*, and the Dairy bacteria. The majority of genes identified in these four large clusters were those functionally involved in transposition. Specifically, 26.4% of the genes in the Proteobacteria cluster were transposition related. In addition, 9.66% of the genes in the Proteobacteria were involved in some form of AMR or metal resistance, and 7.38% in pathogenesis, which may reflect the high number of pathogens found in this phylum[40].

The core and shared gene content of the three Firmicutes clusters (*Staphylococcus*, *Enterococcus* and Dairy) was also assessed (Figure 4, Venn diagram). Gene sharing was most common between the clusters associated with *Staphylococcus* and *Enterococcus* potentially indicating a high frequency of HGT between them, and the least between the *Staphylococcus* and Dairy bacteria cluster. Analysing the content of these shared genes provides insight into both plasmid function and dynamics, such as the identification of HGT events. For example, the same lactose metabolism genes were found in both *Staphylococcus* and Dairy bacteria plasmids. Also, the *trpF* gene, involved in tryptophan biosynthesis and previously associated with the Tn*3000* and Tn*125* transposable elements[41,42], was found on plasmids in all three clusters. In contrast to these, the more disjoint clusters of plasmid cliques observed for other genera may be driven by the species' ecology and life history, which may lead to limited opportunities for contact between lineages. Such an explanation seems plausible for strict pathogens with restricted host range, such as *Xanthomonas* or *Borrellelia*. Conversely, for lineages with a wider environmental niche like *Bacillus*, the lower connectivity between cliques may be due to intrinsic genetic factors leading to lower between-plasmid recombination and/or transposition rates.

288 # Discussion

289 Using alignment-free sequence similarity comparison and subsequent network analysis, we uncovered strong
290 population structure in bacterial plasmids. This approach, applied to a comprehensive set of complete bacterial
291 plasmids, yielded a network in which over half of the plasmids were classified into cliques. This is a significant
292 improvement in coverage over existing plasmid typing methods. Additionally, the cliques capture biologically
293 meaningful information. For example, plasmids assigned to the same clique show good accord with replicon and
294 MOB typing schemes, high homogeneity in terms of their respective bacterial hosts, and similar GC and gene
295 content.

296 A network-based representation of plasmid sequence similarities condenses both vertical and horizontal
297 evolutionary histories in a similar fashion to gene-sharing networks[17,21,22], making it ideally suited for the
298 identification of mobile genetic elements. The cliques we recovered delineate clusters of plasmids with shared
299 evolutionary history. This in turn allows for inference on the nature of HGT events and plasmid function.
300 Moreover, the approach facilitates identification of new replicon gene candidates, as well as detailed investigation
301 of the distribution of plasmid-borne genetic determinants of incompatibility, mobility, AMR, virulence, and
302 transposon carriage. Such meta-information could be incorporated within the network framework thanks to a
303 plethora of well-maintained bioinformatics tools, ever growing genetic databases, and gene ontology efforts to
304 systematize gene annotation.

305 The strong, host-constrained population structure we document for the majority of bacterial plasmids points to
306 transposable genetic elements as the main drivers of HGT in bacteria. In such a case, plasmids primarily act as
307 vehicles for the transfer of genetic material between different bacterial taxa and are eventually lost, while
308 transposons are more successful at maintaining themselves by relocating onto a host-compatible plasmid or the
309 chromosome. Such dynamics could explain the relative uniformity of plasmid cliques in their host range, gene
310 and GC% content, as well as the excess of transposable elements shared between cliques of different taxa. This
311 likely extends to so-called 'broad-host-range' plasmids such as IncP, whose representatives in our dataset fell into
312 three genetically distinct cliques associated to different host species.

313 Jaccard index (i.e. the fraction of shared *k*-mers) was chosen as a measure of sequence similarity between pairs
314 of plasmids due to it being a straightforward metric which considers genome sequences as a whole, embodying
315 both point mutations and large-scale genome rearrangements. As a result, it is not biased by the ability to annotate
316 genes, open reading frames, or other genetic elements. In addition, it is not prone to errors and biases intrinsically
317 associated with alignment-based methods, such as: *a priori* assumptions about the sequence evolution, higher
318 inaccuracy when comparing more dissimilar sequences, or suboptimal alignments[23]. JI can in principle provide
319 fine-scale resolution when comparing small genomes, a characteristic common to the majority of plasmids.
320 Conversely, JI is sensitive to varying genome sizes[28] and plasmids are known to differ more than 1000-fold in
321 sequence length[7,43]. While differences in plasmid genome size can lead to a drop in JI score even when high
322 proportions of *k*-mers are shared, sequence length variation did not seem to impact our structuring into cliques
323 which comprise plasmids of different lengths (Supplementary Figure 17C and 17D).

Assessing the statistical significance of all cliques is computationally intractable given the size of the network. Hence, OSLOM community detection algorithm was employed to uncover cliques of plasmids with high genetic similarity. In an effort to optimize the performance of the OSLOM algorithm and maximize the number of biologically meaningful cliques, all edges with a JI value below 0.3 were removed from the network prior to the analysis. This threshold was chosen to maximise compliance with replicon-based typing as well as several other criteria. Whilst our classification of plasmids into cliques is fairly robust to this exact JI threshold, we appreciate that a 0.3 JI threshold remains somewhat arbitrary. This being said, any taxonomy based on sequence similarity will be partly subjective. As such, our 0.3 JI threshold is comparable in its subjectivity to the 95% average nucleotide identity (ANI) which was set over a decade ago and is routinely used to define species boundaries in prokaryotes[44]. However, depending on the question pursued, enforcing a strict JI threshold may not be necessary, and it could be left to plasmid sequences in the network to solely inform the cut-offs. Some boundaries are likely to be blurrier than others, largely reflecting the extensive variation of genetic inheritance in different bacterial hosts.

Our results suggest it should be possible to devise a 'natural', global sequence-based classification scheme for bacterial plasmids. This being said, our findings do not diminish the relevance of replicon and MOB typing schemes, rather they build upon these prior classification schemes and may extend them to plasmids from understudied and uncultured bacteria. Beyond just plasmid classification, our network-based approach also has potential to infer key features of plasmid groupings. Indeed, plasmid clique assignment can be completely automated and inspection of any particular area of the network facilitates biological inference about plasmid dynamics and their biological features within various groups of bacterial hosts.

Methods

Assembling a dataset of complete bacterial plasmids

A dataset of complete plasmids was downloaded from NCBI's RefSeq release repository[45] on 26th of September 2018. The metadata accompanying each plasmid sequence was parsed from the associated GenBank files. The resulting dataset was then systematically curated to include only those plasmids sequenced from a bacterial host and with a sequence description which implies a complete plasmid sequence (regular expression term used: `"plasmid.*complete sequence"`). This is a simpler, but similar approach to a previously reported curation effort by Orlek and collegues[9]. Nevertheless, a large portion of unsuitable entries, such as gene sequences, partial plasmid genomes, whole genomes, non-bacterial sequences and other poorly annotated sequences, were removed. The final dataset included 10,696 complete bacterial plasmids as listed with full metadata in Supplementary Data 1.

Information about the taxonomic hierarchy of plasmid bacterial hosts was obtained with the *ncbi_taxonomy* module from the ETE 3 Python toolkit[46]. To determine the replicon and MOB types of plasmids included in the dataset we used the PlasmidFinder replicon database (version: 2018-09-04)[12] and MOBtyping software[47]. The PlasmidFinder database was screened using nucleotide BLAST[36] with a minimum coverage and percentage identity of 95%. In cases where two or more replicon hits were found at overlapping positions on a plasmid, the one with the higher percentage identity was retained. For determining the plasmid MOB type, MOBtyping software was used with the recommended settings of 14 PSI-BLAST iterations.

Plasmid CDSs were annotated using Prokka[48] (version 1.13.3) and Roary[49] (version 3.12.0) pipelines run with default parameters. The identified CDSs were further associated with Gene Ontology (GO) terms[50,51] to facilitate downstream gene content analysis. Since Prokka uses a variety of databases to annotate identified CDSs, different resources have been used to append the corresponding GO terms. For example, GO terms for CDSs with a known protein product have been obtained using Uniprot's 'Retrieve/ID Mapping' tool[52], while the GO terms for CDSs with just the HAMAP family were obtained with the hamap2go mapping table[53] (version date: 2019/05/04). CDSs annotated with the ISfinder database were given GO terms GO:0070893 and GO:0004803 in order to associate them with transposition. Similarly, CDS annotated with Aragorn, MinCED, and BARRGD were given GO:0006412, GO:0099048, and GO0046677 terms respectively.

Assessing similarity between pairs of plasmids

The exact Jaccard index (JI) was used as a measure of similarity between all possible plasmid pairs. Each plasmid sequence was converted to a set of 21 bp *k*-mers. The JI was then calculated as the fraction of shared *k*-mers between two sets. JI thus takes a value between 0 and 1, where 1 indicates 100% *k*-mer similarity, and 0 indicates no *k*-mers shared. This allows balanced comparison of diverse plasmid genomes and universality. Also, JI does not weight k-mers based on their abundance, like the popular D_2^* and D_2^S statistics[54], which would exacerbate the inherent sampling biases towards well-studied species to the dataset. We applied Bindash[55] to calculate the exact JI which resulted in the creation of a plasmid adjacency matrix which was used to build the network. All networks presented here have been explored and visualized using the Cytoscape software[56].

Implementing OSLOM community detection algorithm

OSLOM (Ordered Statistics Local Optimization Model version 2.5) was applied to identify cliques (complete subgraphs) with high internal JI similarity in the plasmid network[30]. OSLOM aims to identify highly cohesive clusters of vertices (communities) which may or may not be cliques (complete subgraphs). The statistical significance of a cluster is measured as the probability of finding the cluster in a configuration model which is designed to build random networks while preserving the degrees (number of neighbours) of each vertex. The method locally optimizes the statistical significance with respect to vertices directly neighbouring a particular cluster. In brief, OSLOM starts by randomly choosing vertices from a network which are regarded as clusters of size one. These small clusters alongside their neighbouring vertices are assessed. Vertices are scored based on their connection strength with a particular cluster and are either added or removed from the cluster. The process continues until the entire network is covered. Due to the stochastic nature of the algorithm, this network assessment goes through many iterations after which the frequently emerging significant clusters (i.e. communities) are kept. The algorithm then proceeds to assess the clusters of the next hierarchical level; vertices belonging to the significant clusters are condensed into super-vertices with weighted edges connecting them. The process of cluster assessment is repeated at higher hierarchical levels until no more significant clusters are recovered.

OSLOM was executed for an undirected and weighted network with the following parameters:

```
oslom_undir -w -t 0.05 -r 50 -cp 0 -singlet -hr 0 -seed 1
```

Clusters were considered significant if their *p*-value was lower than 0.05 (`-t 0.05`). The number of iterations required before the recovery of significant clusters was set to 50 during the search for the optimally sparse network (`-r 50`), and 250 for the final network analysis after the introduction of the 0.3 JI threshold (`-r 250`). After the iteration process, OSLOM considers merging similar significant clusters if the significance of their union is high enough. This feature can potentially yield less cliques and was suppressed with the coverage parameter set to zero (`-cp 0`) thus forcing OSLOM to opt for the biggest and most significant cluster from a set of similar clusters. In addition, OSLOM tries to place all vertices of a network in clusters which is also unfavourable for clique recovery and was suppressed with option `-singlet`. Lastly, cliques can only be recovered at the first hierarchical level. Therefore, the OSLOM analysis of the higher hierarchical levels was disregarded (`-hr 0`).

As mentioned earlier, OSLOM is a non-deterministic algorithm and the initial single-vertex clusters are chosen at random. While looking for the optimally sparse network, five OSLOM runs were executed to assess every JI threshold and were given seeds for a random number generator (`-seed`) of 1, 5, 42, 93, and 212. The final network analysis was performed with a seed equal to 42, after which only cliques were considered**,** with non-complete communities disregarded.

Scoring normalized mutual information (NMI) and purity

The compliance of cliques with replicon and MOB typing schemes was assessed by measuring the Normalized Mutual Information (NMI) and purity between them. NMI is a commonly used method to assess the performance of clustering algorithms[57]. For the two clustering/classification schemes ($\mathbf{C_1}$ and $\mathbf{C_2}$) NMI is defined as[58]:

$$\mathrm{NMI}(\mathrm{C}_1, \mathrm{C}_2) = \frac{\mathrm{I}(\mathbf{C_1}, \mathbf{C_2})}{\frac{[\mathrm{H}(\mathbf{C_1}) + \mathrm{H}(\mathbf{C_2})]}{2}} . \quad (1)$$

In equation (1), the mutual information, also known as the information gain and denoted as $\mathrm{I}(\mathbf{C_1},\mathbf{C_2})$, is an information theory concept which measures the reduction of uncertainty around $\mathbf{C_1}$ given knowledge about the $\mathbf{C_2}$, and vice versa. It is normalized by the averaged Shannon entropy (H) between $\mathbf{C_1}$ and $\mathbf{C_2}$. Shannon entropy tends to be larger as the number of classes in $\mathbf{C_1}$ or $\mathbf{C_2}$ approach the size of the dataset in question. Consequently, the NMI is sensitive to differences in the number of classes between $\mathbf{C_1}$ and $\mathbf{C_2}$, and to extensively fragmented classifications. The NMI equals one if the two classifications yield identical partitioning of the dataset, whereas a value of zero indicates complete incoherence. The NMI was measured using the *R* package *NMI* (version 2.0; https://CRAN.R-project.org/package=NMI). During the assessment, plasmids which were not classified by replication or MOB typing schemes were disregarded.

Purity was used to estimate the homogeneity of cliques for replicon or MOB types, and plasmid host taxa. For a set of cliques **C**, and a plasmid typing scheme **T**, purity is defined as:

$$\mathrm{purity}(\mathbf{C}, \mathbf{T}) = \frac{1}{N} \sum_{c_i \in \mathbf{C}} \max_{t_j \in \mathbf{T}} |c_i \cap t_j| \quad (2)$$

where N is the total number of plasmids covered by a set of cliques, $\mathbf{C} = \{ c_1, c_2, \ldots, c_i \}$ is a set of cliques in which plasmids were placed, and $\mathbf{T} = \{ t_1, t_2, \ldots, t_j \}$ are the types associated with plasmids. Similar to NMI, the purity scores a value between 0 and 1 with high purity indicating high homogeneity of classes in the dataset for a given set of plasmid types. The purity was only assessed for cliques which contain at least one typed plasmid. Untyped plasmids found within the assessed cliques were disregarded.

The full MOB, incompatibility types and clique assignments for each plasmid are provided in Supplementary Data 1.

Data availability

The sequences that make up the dataset analysed during the current study are available in the NCBI's RefSeq repository, ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plasmid/. The accession numbers of the sequences are listed in Supplementary Data 1. The source data underlying all but one figures are provided as a Source Data file.

Code availability

All code used in this research is available at https://github.com/macman123/plasmid_network_analysis.

References

1. Shintani, M. & Suzuki, H. Plasmids and Their Hosts. in *DNA Traffic in the Environment* 109–133 (Springer Singapore, 2019). doi:10.1007/978-981-13-3411-5_6
2. Von Wintersdorff, C. J. H. *et al.* Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* **7**, (2016).
3. Halary, S., Leigh, J. W., Cheaib, B., Lopez, P. & Bapteste, E. Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl. Acad. Sci.* **107**, 127–132 (2010).
4. Stokes, H. W. & Gillings, M. R. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS Microbiol. Rev.* **35**, 790–819 (2011).
5. Carattoli, A. *et al.* Identification of plasmids by PCR-based replicon typing. *J. Microbiol. Methods* **63**, 219–228 (2005).
6. Garcillán-Barcia, M. P., Francia, M. V. & de La Cruz, F. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.* **33**, 657–687 (2009).
7. Shintani, M., Sanchez, Z. K. & Kimbara, K. Genomics of microbial plasmids: Classification and identification based on replication and transfer systems and host taxonomy. *Front. Microbiol.* **6**, 1–16 (2015).
8. Orlek, A. *et al.* Plasmid classification in an era of whole-genome sequencing: Application in studies of antibiotic resistance epidemiology. *Frontiers in Microbiology* **8**, 1–10 (2017).
9. Orlek, A. *et al.* Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. *Plasmid* **91**, 42–52 (2017).
10. Lozano, C. *et al.* Expansion of a plasmid classification system for Gram-positive bacteria and determination of the diversity of plasmids in Staphylococcus aureus strains of human, animal, and food origins. *Appl. Environ. Microbiol.* **78**, 5948–55 (2012).
11. Jensen, L. B. *et al.* A classification system for plasmids from enterococci and other Gram-positive bacteria. *J. Microbiol. Methods* **80**, 25–43 (2010).
12. Carattoli, A. *et al.* In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. **58**, 3895–3903 (2014).
13. Rozwandowicz, M. *et al.* Plasmids of Distinct IncK Lineages Show Compatible Phenotypes. *Antimicrob. Agents Chemother.* **61**, e01954-16 (2017).
14. Ambrose, S. J., Harmer, C. J. & Hall, R. M. Compatibility and entry exclusion of IncA and IncC plasmids revisited: IncA and IncC plasmids are compatible. *Plasmid* **96–97**, 7–12 (2018).
15. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genomics* **4**, (2018).
16. Bapteste, E. *et al.* Prokaryotic evolution and the tree of life are two different things. *Biol. Direct* **4**, 34 (2009).
17. Brilli, M. *et al.* Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. *BMC Bioinformatics* **9**, 551 (2008).
18. Corel, E., Lopez, P., Méheust, R. & Bapteste, E. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends Microbiol.* **24**, 224–237 (2016).
19. Bernard, G., Greenfield, P., Ragan, M. A. & Chan, C. X. k -mer Similarity, Networks of Microbial Genomes, and Taxonomic Rank . *mSystems* **3**, (2018).
20. Dagan, T., Artzy-Randrup, Y. & Martin, W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10039–44 (2008).
21. Tamminen, M., Virta, M., Fani, R. & Fondi, M. Large-Scale Analysis of Plasmid Relationships through Gene-Sharing Networks. *Mol. Biol. Evol.* **29**, 1225–1240 (2012).
22. Yamashita, A. *et al.* Characterization of Antimicrobial Resistance Dissemination across Plasmid Communities Classified by Network Analysis. *Pathogens* **3**, 356–376 (2014).
23. Zielezinski, A., Vinga, S., Almeida, J. & Karlowski, W. M. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* **18**, 1–17 (2017).
24. Bernard, G. *et al.* Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Brief. Bioinform.* **20**, 426–435 (2019).
25. Ren, J. *et al.* Alignment-Free Sequence Analysis and Applications. *Annu. Rev. Biomed. Data Sci.* **1**, 93–114 (2018).
26. Zielezinski, A. *et al.* Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* **20**, (2019).
27. Jesus, T. F. *et al.* Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. *Nucleic Acids Res.* **47**, D188–D194 (2019).
28. Ondov, B. D. *et al.* Mash : fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 1–14 (2016). doi:10.1186/s13059-016-0997-x

29. Karp, R. M. Reducibility among Combinatorial Problems. in *Complexity of Computer Computations* 85–103 (Springer US, 1972). doi:10.1007/978-1-4684-2001-2_9
30. Lancichinetti, A., Radicchi, F., Ramasco, J. J. & Fortunato, S. Finding Statistically Significant Communities in Networks. *PLoS One* **6**, e18961 (2011).
31. Hric, D., Darst, R. K. & Fortunato, S. Community detection in networks: Structural communities versus ground truth. *Phys. Rev. E* **90**, 062805 (2014).
32. Bron, C. & Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph [H]. *Commun. ACM* **16**, 575–577 (1973).
33. Peixoto, T. P. The graph-tool python library. *figshare* (2014). doi:10.6084/m9.figshare.1164194
34. Chikami, G. K., Guiney, D. G., Schmidhauser, T. J. & Helinski, D. R. Comparison of 10 IncP plasmids: homology in the regions involved in plasmid replication. *J. Bacteriol.* **162**, 656–60 (1985).
35. Norberg, P., Bergström, M., Jethava, V., Dubhashi, D. & Hermansson, M. The IncP-1 plasmid backbone adapts to different host bacterial species and evolves through homologous recombination. *Nat. Commun.* **2**, (2011).
36. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
37. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
38. del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M. & Díaz-Orejas, R. Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.* **62**, 434–64 (1998).
39. Nishida, H. Comparative Analyses of Base Compositions, DNA Sizes, and Dinucleotide Frequency Profiles in Archaeal and Bacterial Chromosomes and Plasmids. *Int. J. Evol. Biol.* **2012**, 1–5 (2012).
40. Rizzatti, G., Lopetuso, L. R., Gibiino, G., Binda, C. & Gasbarrini, A. Proteobacteria: A Common Factor in Human Diseases. *Biomed Res. Int.* **2017**, 9351507 (2017).
41. Hu, H. *et al.* Novel plasmid and its variant harboring both a bla(NDM-1) gene and type IV secretion system in clinical isolates of Acinetobacter lwoffii. *Antimicrob. Agents Chemother.* **56**, 1698–702 (2012).
42. Campos, J. C. *et al.* Characterization of Tn3000, a Transposon Responsible for blaNDM-1 Dissemination among Enterobacteriaceae in Brazil, Nepal, Morocco, and India. *Antimicrob. Agents Chemother.* **59**, 7387–95 (2015).
43. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–52 (2010).
44. Klappenbach, J. A. *et al.* DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* **57**, 81–91 (2007).
45. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-45 (2016).
46. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
47. Orlek, A. *et al.* A curated dataset of complete Enterobacteriaceae plasmids compiled from the NCBI nucleotide database. *Data Br.* **12**, 423–426 (2017).
48. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
49. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
50. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–9 (2000).
51. Carbon, S. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
52. Huang, H. *et al.* A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics* **27**, 1190–1191 (2011).
53. Lima, T. *et al.* HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res.* **37**, D471–D478 (2009).
54. Reinert, G., Chew, D., Sun, F. & Waterman, M. S. Alignment-free sequence comparison (I): Statistics and power. *J. Comput. Biol.* **16**, 1615–1634 (2009).
55. Zhao, X. BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics* **35**, 671–673 (2019).
56. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
57. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Phys. Rep.* **659**, 1–44 (2016).
58. Fred, A. L. N. & Jain, A. K. Robust data clustering. in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* **2**, II-128-II–133 (IEEE Comput. Soc,

561 2003).

562

Acknowledgments

M.A. was supported by a PhD scholarship from University College London. L.v.D. and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (MRC grant MR/P007597/1) and a Wellcome Institutional Strategic Support Fund (ISSF3) – AI in Healthcare (19RX03). F.B. additionally acknowledges support from the BBSRC GCRF scheme and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. The funders had no role in study design, data collection, interpretation of results, or the decision to submit the work for publication. Lastly, we would like to thank Prof Chris Barnes for providing advice on the network analyses.

Contributions

M.A., and F.B. conceived the project and designed the experiments. M.A. performed all the analyses under the guidance of L.v.D and F.B. J.M.S advised on plasmid biology. M.A., Lv.D and F.B. take responsibility for the accuracy and availability of the results. L.v.D. provided moral support to M.A. M.A. wrote the paper with contributions from L.v.D and F.B. All authors read and commented on successive drafts and all approved the content of the final version.

Competing Interests

The authors declare no competing financial interests.

Figures

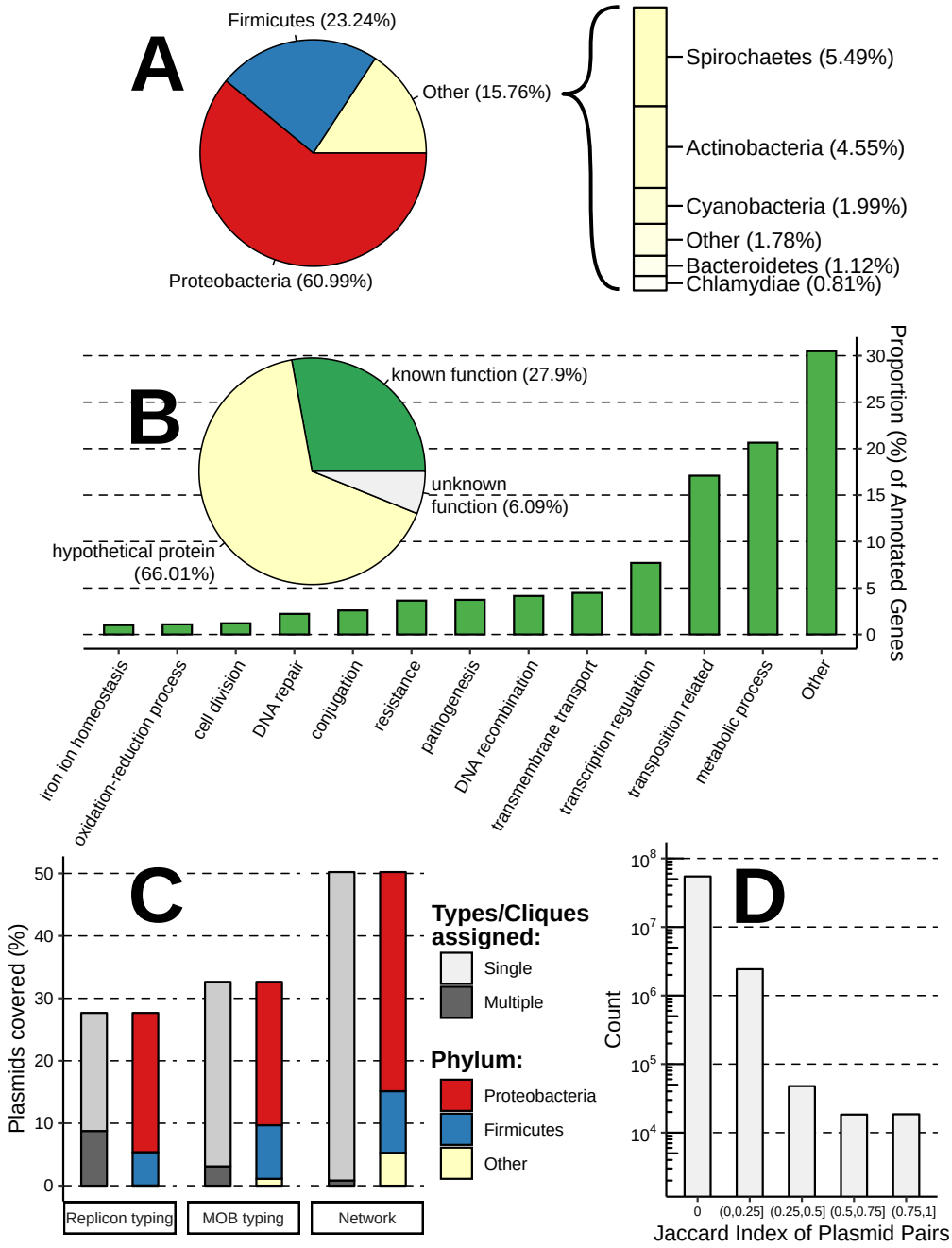

Figure 1. Summary of the dataset of complete bacterial plasmids. (A) The distribution of host phylum represented in the plasmid dataset. (B) Functional annotation of plasmid-borne genes. The pie chart shows the proportion of unique CDSs with hypothetical function as predicted by Prokka[48], and CDSs (genes) with known/unknown biological function based on GO annotation. The bar chart provides the most common biological functions associated with plasmid-borne genes also considering the respective frequency of these genes on plasmid genomes. (C) The percentage of plasmids covered by the three classification methods: replicon and MOB typing schemes, and clique assignment. (D) The distribution of pairwise plasmid similarities (Jaccard Index).

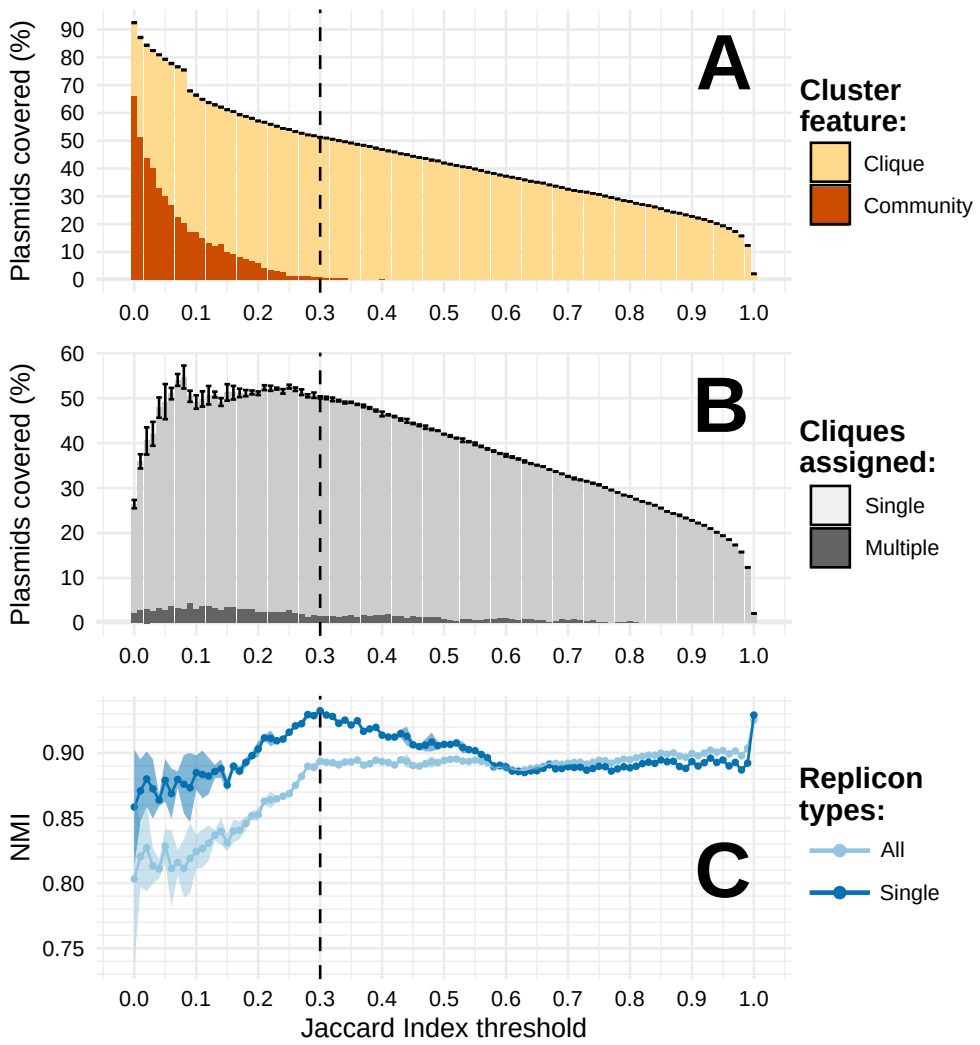

Figure 2. Optimization of OSLOM performance. A range of Jaccard Index (JI) thresholds were applied to the original plasmid network (Supplementary Figure 3) with edges below a particular threshold being removed prior to OSLOM analysis. During the process, several criteria were considered: (A) clique to community ratio; (B) percentage of plasmids covered by the cliques; (C) the congruence with replicon typing measured by NMI score. NMI was calculated for all cliques containing plasmids assigned to a single or multiple replicon types (legend: All) and just to a single replicon type (legend: Single). Error bars (A and B) and light-coloured shading (C) provide +/- 2 standard deviations (SD) of uncertainty. Standard deviation around every value on the *y*-axis across all JI thresholds assessed (points and bars) was calculated based on results of n=5 iterations of OSLOM software (see Methods). The dashed vertical line indicates the selected optimal JI threshold of 0.3.

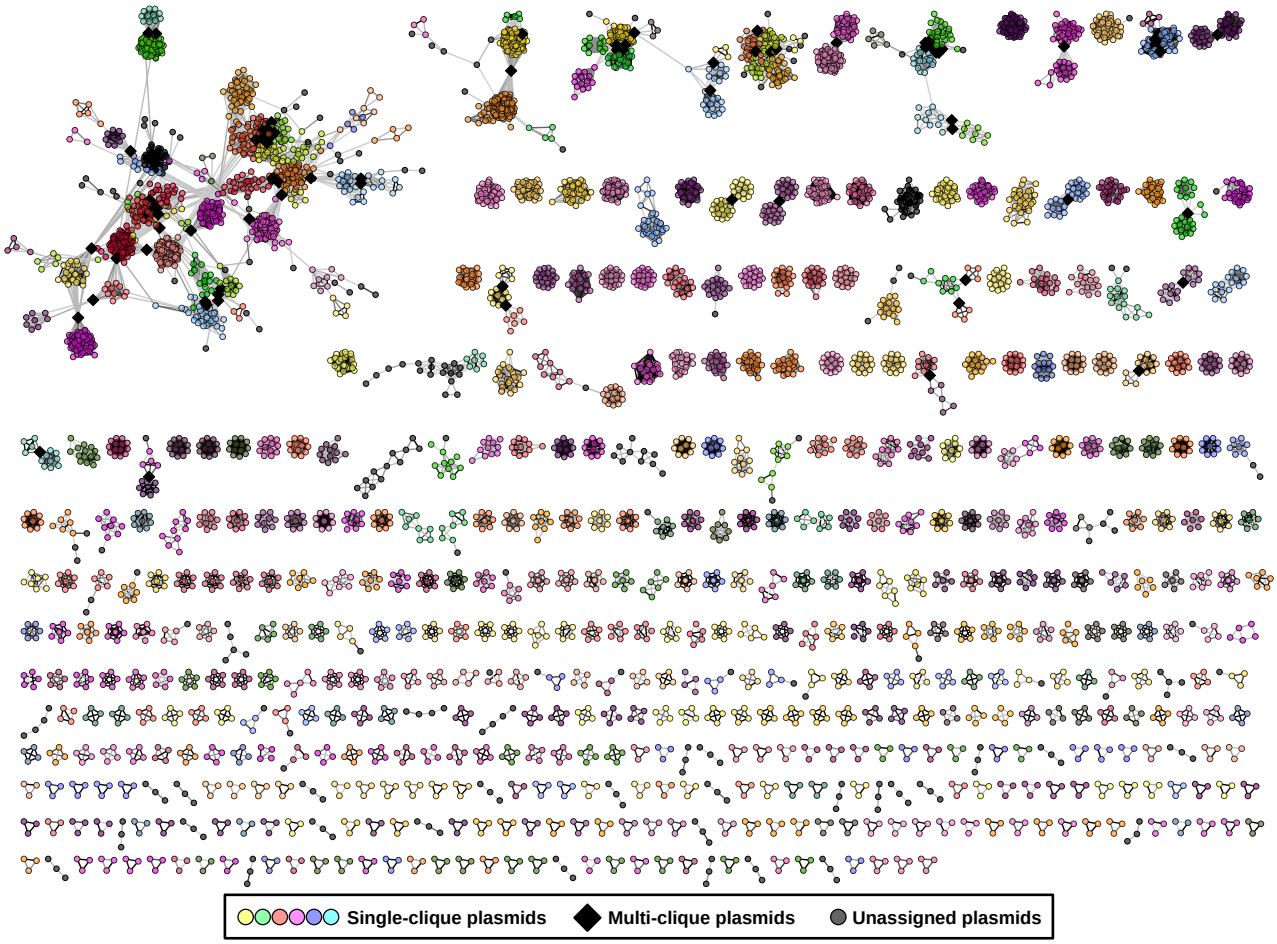

Figure 3. Sparse network of plasmids assigned to cliques by OSLOM algorithm (network density = 0.00128). The network includes 5371 plasmids (nodes) assigned into 561 cliques (complete sub-graphs). The completeness of identified cliques was evaluated based on the original network (Supplementary Figure 3). 5,008 unassigned plasmids, which formed disjoined singletons and pairs, were removed from the network. Coloured nodes indicate plasmids assigned to a single clique.

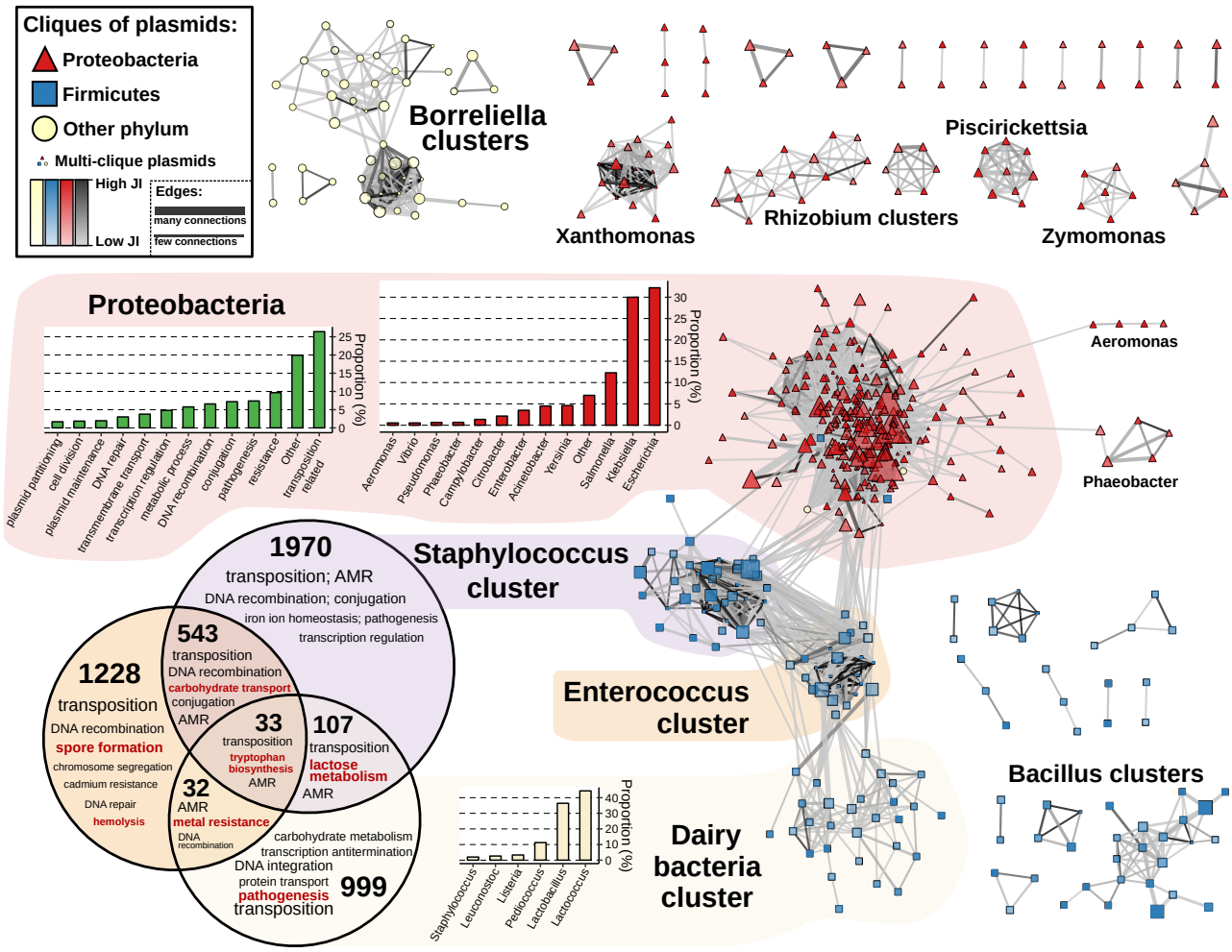

Figure 4. The network of cliques. Cliques, represented as vertices, are connected with an edge if the average Jaccard Index (JI) between plasmids of two cliques is higher than 0.01. The colour of the edges indicates the average JI while the width is proportional to the number of connections between a pair of cliques. The shape and colour of the cliques indicates the phylum of the predominant bacterial host. The size and the transparency are proportional to the clique size and the internal JI respectively. The cliques form multiple clusters which have been named based on the genus of the bacterial host characteristic for a particular cluster. There are two exceptions – the Proteobacteria and the Dairy (Lactic) cluster whose respective genera distributions have been provided. The most common GO biological functions of the genes found on plasmids of Proteobacteria, *Staphylococcus*, *Enterococcus* and Dairy clusters were further assessed. During the assessment, the respective frequencies of the genes were considered. In case of Proteobacteria, the bar chart distribution of the biological functions is provided. The shared and core gene content of *Staphylococcus*, *Enterococcus* and Dairy clusters is presented in the Venn diagram with the numbers in the diagram indicating the number of core and shared genes.