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

Most bacteria exchange genetic material through Horizontal Gene Transfer (HGT). The primary vehicles for HGT are plasmids and plasmid-borne transposable elements, though their population structure and dynamics remain poorly understood. Here, we quantified genetic similarity between more than 10,000 bacterial plasmids and reconstructed a network based on their shared *k*-mer content. Using a community detection algorithm, we assigned plasmids into cliques which are highly correlated with plasmid gene content, bacterial host range, GC content, as well as replicon and mobility (MOB) type classifications. Resolving the plasmid population structure further allowed identification of candidates for yet-undescribed replicon genes. Our work provides biological insights into the dynamics of plasmids and plasmid-borne mobile elements, with the latter representing the main drivers of HGT at broad phylogenetic scales. Our results illustrate the potential of network-based analyses for the bacterial 'mobilome' and open 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 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 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 and are thus assigned 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 mutation in sequence alignments[16,17]. Network models offer an attractive alternative given they can incorporate both horizontal and vertical inheritance[18], 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[3,8,18,19]. Plasmid gene-sharing networks have proven a useful means to track AMR and virulence dissemination yielding deeper insights into HGT events[17,20,21]. However, the main drawback of previous work relying on plasmid sequence alignments is the exclusion of important non-coding elements such as non-coding RNAs, promoter regions, CRISPRs, stretches of homologous sequences, or putative, disrupted and currently unannotated genes. A more comprehensive approach could consider a plasmid network based on estimates of alignment-free sequence similarity[22]. A recently published Plasmid ATLAS tool by Jesus *et al.*[23] provides an illustration of such an approach, with a network of plasmids constructed based on pairwise genetic distances estimated using alignment-free *k*-mer matching methods implemented in Mash[24].

In this work, we have quantified the genetic similarity between more than 10,000 bacterial plasmids available on NCBI's RefSeq database and constructed a network reflecting their relatedness based on shared *k*-mer content. Applying a community detection algorithm allowed us to cluster plasmids into statistically significant cliques (complete subgraphs), and revealed a strong underlying population structure. Cliques are highly correlated with the gene content of the plasmid backbone, bacterial host and GC content, as well as replicon and MOB types. Uncovering the structure of the full plasmid population allowed for the discovery of candidates for yet-undescribed replicon genes and provided insights into broad-scale plasmid dynamics. Taken together, our results illustrate the potential of network-based analyses of plasmid sequences and open up the prospect of a natural, 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 Table 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. 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. The dominant associated terms relate to catabolic and biosynthetic processes (20.64%), 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 using a network-based approach

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 with high internal JI score.

Finding all the cliques of a network is a nondeterministic polynomial (NP)-complete problem[25]. This means that while a solution for a single clique can be quickly verified, the time required to find all possible cliques scales rapidly as the size of the network increases. In the case of our large network of plasmids, a full solution cannot be found within a reasonable timeframe given current computational limitations. As an alternative solution, a community detection algorithm OSLOM (Ordered Statistics Local Optimization Method) was implemented[26]. OSLOM detects communities (i.e. densely interconnected subgraphs) with statistical significance, meaning that they have a low probability of being encountered by chance in a random network with similar features to the plasmid network. OSLOM is well suited for this task since it can be used to analyse undirected networks with overlapping communities or hierarchical structures. In addition, OSLOM shows similar performance to other widely used methods such as Infomap or Louvain[26,27] which, unlike OSLOM, were unable to analyse this dataset due to size and memory limitations.

Despite the notable dissimilarity among plasmids, the network as a whole was too dense (network density = 0.0438) to yield a consistent performance for every OSLOM run (Figure 2 and Supplementary Figures 3 and 6). Furthermore, a large proportion of communities detected did not form cliques, and thus had to be disregarded (Figure 2A). A JI threshold was therefore introduced to increase the sparsity of the network. A range of thresholds were assessed based on the following criteria: (i) the clique to community ratio (Figure 2A), (ii) the proportion of plasmids assigned to cliques (Figure 2B), (iii) the congruence with replicon-based typing (Figure 2C), and (iv) the consistency of OSLOM performance (Figure 2 and Supplementary Figure 6). The optimum threshold was consistently obtained at a JI of 0.3. This threshold was also corroborated by an alternative data driven approach introduced by Branger *et al.*[28] called the giant component analysis. This method determines the optimal JI threshold by tracking the size of the giant component (i.e. the largest cluster) of the network, and the total number of components. In this case, the relative stability of the size of the giant component was reached at a JI threshold of 0.3 (Supplementary Figure 7). Edges with values lower than the threshold were removed from the network. The resulting sparse network is shown in Figure 3 (network density = 0.00128).

Plasmid cliques agree with current typing schemes

Analysis of the sparse network with OSLOM successfully assigned 50.21% (5371) of the plasmids into 561 cliques (Figure 1C, Figure 3, and Supplementary Figure 10). 1.64% (88) of these plasmids were assigned to multiple cliques, and were found in the densest regions of the network and at the interfaces between cliques indicating the presence of ‘chimeric plasmids’ (i.e. hybrid plasmids generated through merging of two different plasmids), large-scale transposition or recombination events, or extensive repeated transposition/recombination (Figure 1C and Figure 3). In addition, this approach covered 564 plasmids from phyla other than the Proteobacteria and Firmicutes, namely from Spirochaetes, Chlamydiae, Actinobacteria, Tenericutes, Bacteroidetes, Cyanobacteria, and Fusobacteria. Interestingly, after applying the 0.3 JI threshold, 38.01% (4066) of plasmids were separated from the network as singletons, while 10.10% (1080) shared an edge with a single plasmid. Such plasmids could not be assigned to a clique. Therefore, only 1.67% (179) of plasmids were effectively left unassigned by the algorithm.

Clique purity and Normalized Mutual Information (NMI) were used to assess the quality of clique-based classification (see Methods). These metrics were calculated for cliques comprising plasmids with identified replicon type, plasmids carrying a single identified replicon type, or plasmids with assigned MOB type. Untyped plasmids were disregarded. The observed purity scores were high (>85%) indicating the homogeneity of cliques for a particular plasmid type (Supplementary Figure 8). This was particularly the case for MOB types (purity = 0.9887) and plasmids assigned to a single replicon type (purity = 0.9522). NMI provides an entropy-based measure of the similarity between two classification systems where a score equal to one indicates identical partitioning into classes while zero means independent classification. NMI penalizes differences in the number of assignment classes which justifies the low score observed when assessing clique-based versus MOB-based typing (NMI = 0.5223). Nevertheless, high NMI scores were obtained when considering a replicon-based classification scheme (NMI = 0.9044 all types, and NMI = 0.9336 for single replicon types). It follows that plasmids with the same replicon type often fall together within the same clique. This is also supported by the high correlation between the clique membership size and the number of plasmids assigned to the corresponding replicon class (Supplementary Figure 9, R^2=0.862 for plasmids assigned to a single replicon types). However, there are exceptions where plasmids from larger replicon classes are further resolved into a few smaller evolutionary related cliques.

Candidate replicon genes recovered from cliques of untyped plasmids

The majority of plasmids with unknown replicon types formed small cliques (Supplementary Figure 10). 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[29]. A range of e-values were assessed to determine the threshold maximizing the discovery of replicon candidates while minimizing false positives (Supplementary Figure 11). The majority of plasmids were assigned to one replicon type with some plasmids having hits to a maximum of three to four different types. 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. With this in mind, a conservative e-value threshold of 0.001 was chosen which resulted in the identification of 105 candidate genes from 106 plasmid cliques (Supplementary Table 1).

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)[30]. 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 12). 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[31]. Other examples of transcriptional regulators also exist in plasmid replicon regions, while the DNA primase activity has been found on the RepB replicon protein[31]. 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[31]. 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 be useful for describing new incompatibility types.

Plasmids within cliques have a low variability in GC content and share a common bacterial host

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 13A). 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 13B). Furthermore, 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 13C and 13D). 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[32]. Indeed, the cliques showed a very high homogeneity (purity) relative to their hosts (Supplementary Figure 14), a trend which has been identified in other plasmid network reconstruction efforts[20]. 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. 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 15). 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 16). 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. 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 17) 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 bacterial 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 18.

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[33].

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 plasmid 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[34,35], was found on plasmids in all three clusters.

Discussion

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

A network-based representation of plasmid sequence similarities condenses both vertical and horizontal evolutionary histories in a similar fashion to gene-sharing networks[17,20,21], making it ideally suited for the identification of mobile genetic elements. The model employed here assigns plasmids to cliques, delineating clusters of plasmids with shared evolutionary history. This in turn allows for inference on the nature of HGT events and plasmid function. Moreover, the approach facilitates identification of new replicon gene candidates, as well as detailed investigation of the distribution of plasmid-borne genetic determinants of incompatibility, mobility, AMR, virulence, and transposon carriage. Such meta-information could be incorporated within the network framework thanks to a plethora of well-maintained bioinformatics tools, ever growing genetic databases, and gene ontology efforts to systematize gene annotation.

Jaccard index (i.e. the fraction of shared *k*-mers) was chosen as a measure of sequence similarity between pairs of plasmids due to it being a straightforward metric which considers genome sequences as a whole, embodying both point mutations and large-scale genome rearrangements. As a result, it is not biased by the ability to annotate genes, open reading frames, or other genetic elements. In addition, it is not prone to errors and biases intrinsically associated with alignment-based methods, such as: *a priori* assumptions about the sequence evolution, higher inaccuracy when comparing more dissimilar sequences, or suboptimal alignments[22]. JI can in principle provide fine-scale resolution when comparing small genomes, a characteristic common to the majority of plasmids. Conversely, JI is sensitive to varying genome sizes[24] and plasmids are known to differ more than 1000-fold in sequence length[7,36]. While differences in plasmid genome size can lead to a drop in JI score even when high proportions of *k*-mers are shared, sequence length variation did not seem to impact our structuring into cliques which comprise plasmids of different lengths (Supplementary Figure 13C and D).

Assessing the statistical significance of all resulting cliques is computationally intractable given the size of the network. Hence, the OSLOM community detection algorithm was employed to uncover cliques that are unlikely to be found in a random network. 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. This threshold was chosen to maximise compliance with replicon-based typing as well as several other criteria. The implementation of the 0.3 JI threshold somewhat allegorizes the average nucleotide identity (ANI), which was set over a decade ago at 95%, to define the species boundary for prokaryotes[37]. 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.

The strong underlying population structure we document for plasmids throughout bacteria suggests 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 even 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[38] on 26th of September 2018. The metadata accompanying each plasmid sequence was parsed from the associated GenBank files (Supplementary Table 1). 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.

Information about the taxonomic hierarchy of plasmid bacterial hosts was obtained with the *ncbi_taxonomy* module from the ETE 3 Python toolkit[39]. To determine the replicon and MOB types of plasmids included in the dataset we used the PlasmidFinder replicon database[12] and MOBtyping software[40]. The PlasmidFinder database was screened using BLAST[29] with a minimum coverage and percentage nucleotide identity of 95%. In cases where two or more replicon hits were found at overlapping positions on a plasmid, the one with 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 the Prokka[41] (version 1.13.3) and Roary[42] (version 3.12.0) pipelines run with default parameters. The identified CDSs were further associated with Gene Ontology (GO) terms[43,44] 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[45], while the GO terms for CDSs with just the HAMAP family were obtained with the hamap2go mapping table[46] (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. To do this, 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. We applied Bindash[47] to calculate the exact JI which resulted in the creation of a plasmid adjacency matrix which was used to build the network.

Implementing OSLOM community detection algorithm

OSLOM (Ordered Statistics Local Optimization Model version 2.5) was applied to identify statistically significant cliques (complete subgraphs) in the plasmid network[26]. 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, significant 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[48]. For the two clustering/classification schemes (C_1 and C_2) NMI is defined as[49]:

$$\boldsymbol{NMI(C_1, C_2)} = \frac{\boldsymbol{I(C_1, C_2)}}{\frac{\boldsymbol{[H(C_1) + H(C_2)]}}{\boldsymbol{2}}} . \qquad (1)$$

In equation (1), the mutual information, also known as the information gain and denoted as $I(C_1,C_2)$, is an information theory concept which measures the reduction of uncertainty around C_1 given knowledge about the C_2, and vice versa. It is normalized by the averaged Shannon entropy (H) between C_1 and C_2. Shannon entropy tends to be larger as the number of classes in C_1 or C_2 approach the size of the dataset in question. Consequently, the NMI is sensitive to differences in the number of classes between C_1 and 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:

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

where N is the total number of plasmids covered by a set of cliques, $C = \{ c_1, c_2, \ldots, c_i \}$ is a set of cliques in which plasmids were placed, and $T = \{ t_1, t_2, \ldots, t_j \}$ 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.

References

1. Shintani, M. & Suzuki, H. Plasmids and Their Hosts. in *DNA Traffic in the Environment* 109–133 (Springer Singapore, 2019).
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. 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).
20. 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).
21. Yamashita, A. *et al.* Characterization of Antimicrobial Resistance Dissemination across Plasmid Communities Classified by Network Analysis. *Pathogens* **3**, 356–376 (2014).
22. Zielezinski, A., Vinga, S., Almeida, J. & Karlowski, W. M. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* **18**, 1–17 (2017).
23. Jesus, T. F. *et al.* Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. *Nucleic Acids Res.* **47**, D188–D194 (2019).
24. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 1–14 (2016).
25. Karp, R. M. Reducibility among Combinatorial Problems. in *Complexity of Computer Computations* 85–103 (Springer US, 1972).
26. Lancichinetti, A., Radicchi, F., Ramasco, J. J. & Fortunato, S. Finding Statistically Significant Communities in Networks. *PLoS One* **6**, e18961 (2011).
27. Hric, D., Darst, R. K. & Fortunato, S. Community detection in networks: Structural communities versus ground truth. *Phys. Rev. E* **90**, 062805 (2014).
28. Branger, C. *et al.* Extended-spectrum β-lactamase-encoding genes are spreading on a wide range of Escherichia coli plasmids existing prior to the use of third-generation cephalosporins. *Microb. Genomics*

4, e000203 (2018).

29. 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).
30. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
31. 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).
32. 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).
33. Rizzatti, G., Lopetuso, L. R., Gibiino, G., Binda, C. & Gasbarrini, A. Proteobacteria: A Common Factor in Human Diseases. *Biomed Res. Int.* **2017**, 9351507 (2017).
34. 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).
35. 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).
36. 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).
37. 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).
38. 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).
39. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
40. Orlek, A. *et al.* A curated dataset of complete Enterobacteriaceae plasmids compiled from the NCBI nucleotide database. *Data Br.* **12**, 423–426 (2017).
41. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
42. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
43. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–9 (2000).
44. Carbon, S. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
45. Huang, H. *et al.* A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics* **27**, 1190–1191 (2011).
46. 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).
47. Zhao, X. BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics* **35**, 671–673 (2019).
48. Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Phys. Rep.* **659**, 1–44 (2016).
49. 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, 2003).

Acknowledgments

M.A. was supported by a PhD scholarship from BBSRC (XXX). L.v.D. and F.B. acknowledge financial support from the Newton Fund UK-China NSFC initiative (MRC grant MR/P007597/1). 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 about 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

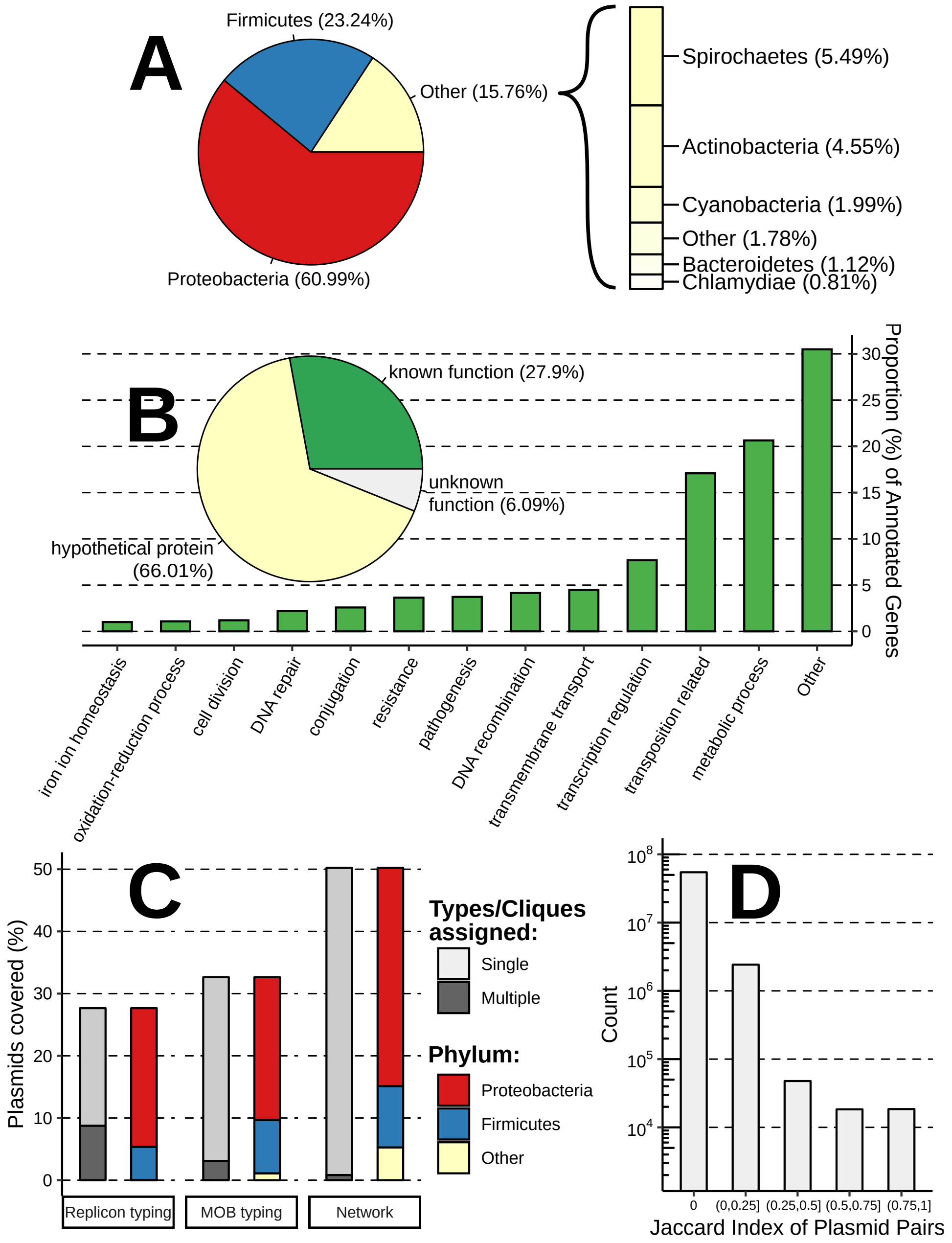

Figure 1. Summary of the dataset of complete bacterial plasmids. (A) The distribution of host phylum represented in the plasmid dataset. **(B)** Functional annotation of plasmid-borne genes. The pie chart shows the proportion of CDSs with hypothetical function as predicted by Prokka[41], and CDSs (genes) 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).

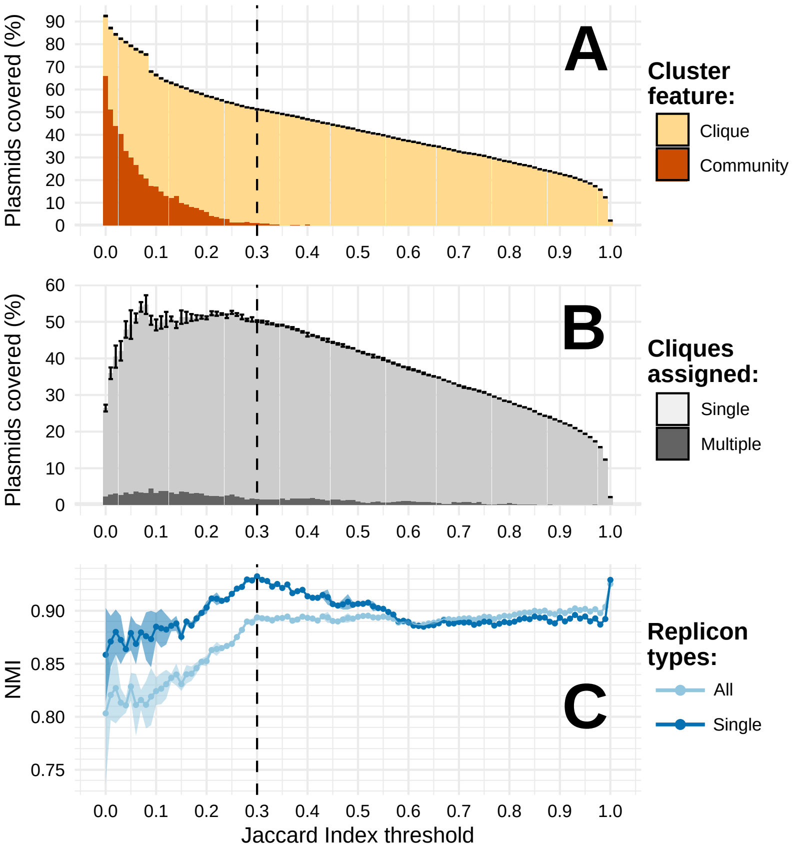

Figure 2. Searching for the optimally sparse plasmid network. A range of Jaccard Index (JI) thresholds were applied to the original plasmid network (Supplementary Figure 3) prior to OSLOM analysis. During the process, several criteria were considered: (A) clique to community ratio; (B) percentage of 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 two standard deviations of uncertainty. The dashed vertical line indicates the selected optimal JI threshold of 0.3.

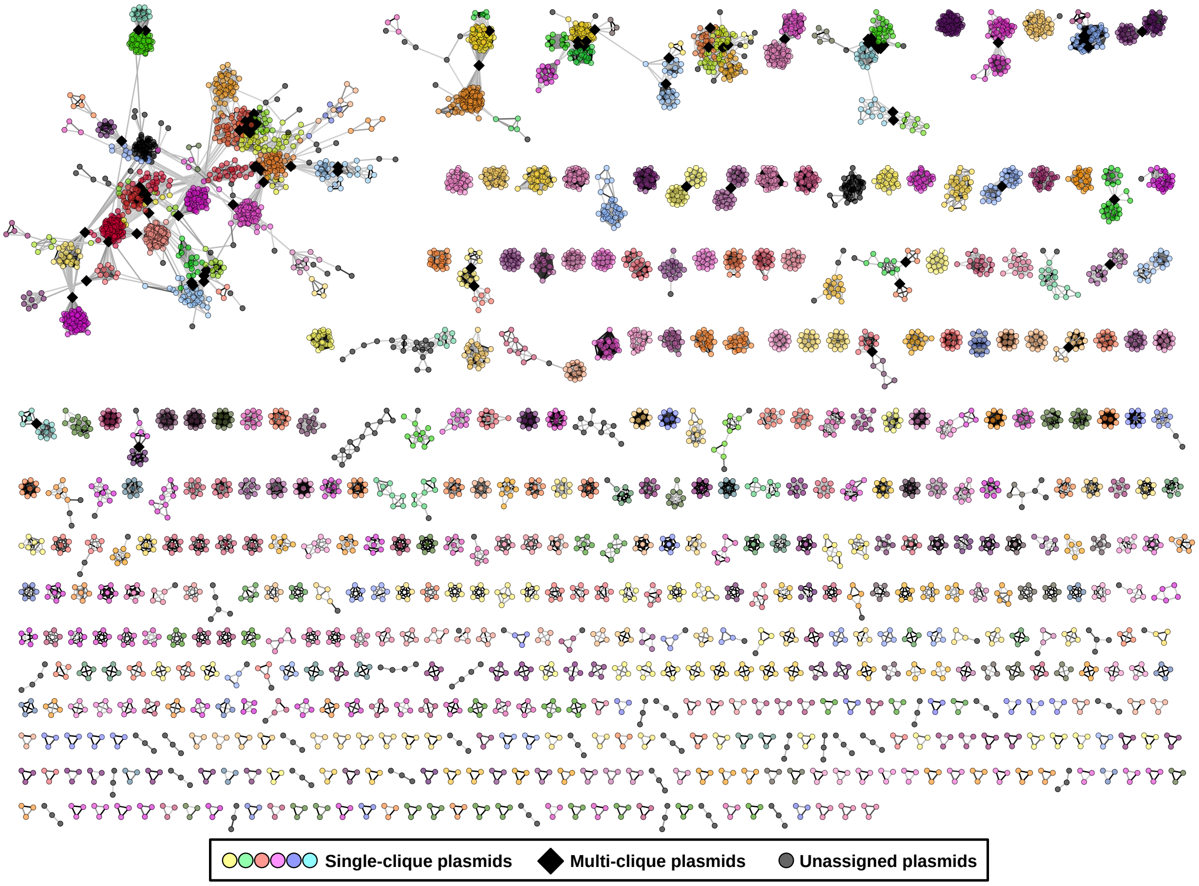

Figure 3. Sparse network of plasmids assigned to cliques by OSLOM algorithm (network density = 0.00128). The network includes 5371 plasmids (nodes) assigned into 561 cliques (connected sub-graphs). 5,008 unassigned plasmids, which formed disjoined singletons and pairs, were removed from the network.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

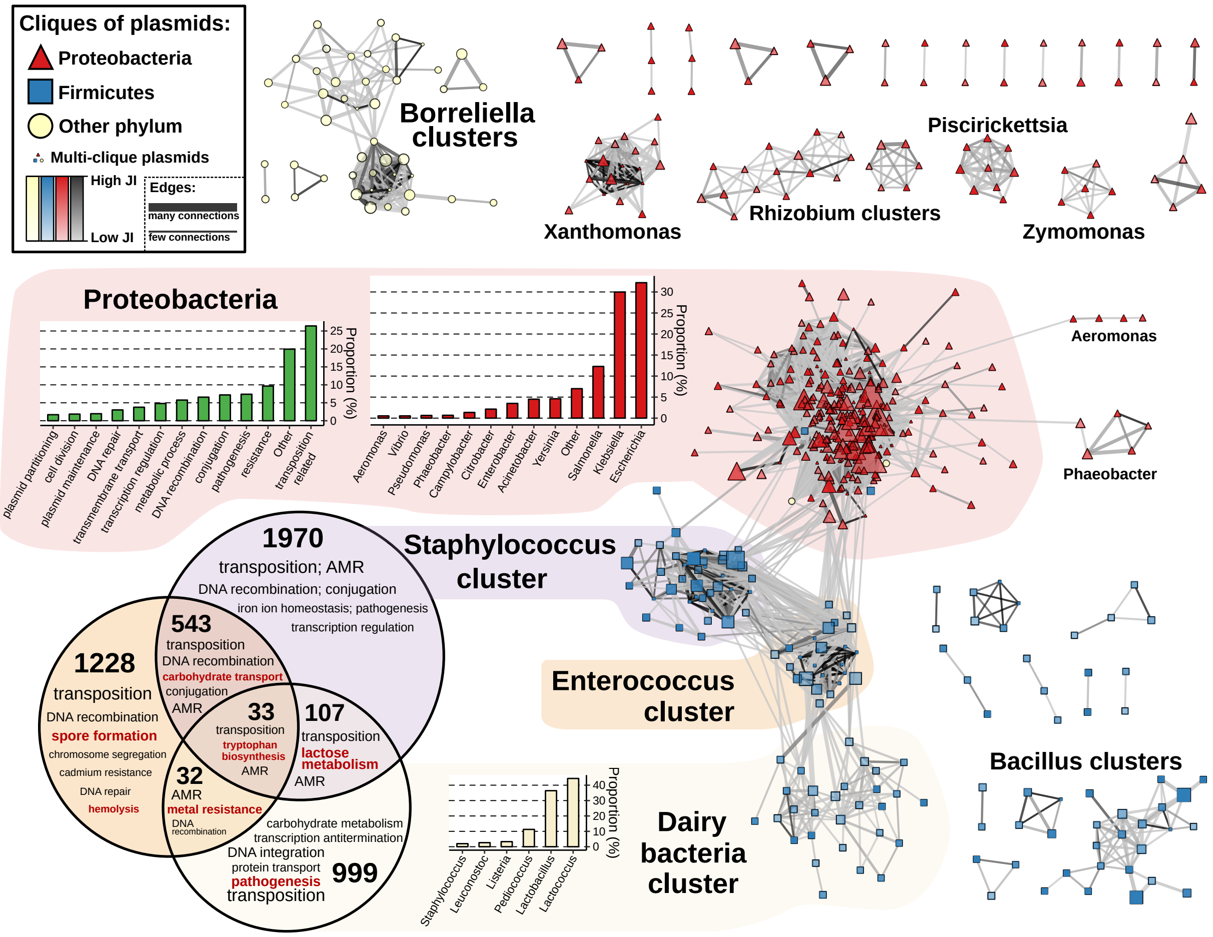

Figure 4. The network of cliques. Cliques, represented as vertices, are connected with an edge if the average Jaccard Index (JI) between plasmids of two cliques is higher than 0.01. The colour of the edges indicates the average JI while the width is proportional to the number of connections between a 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.