Plasmid Flux in *Escherichia coli* ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences

Val F. Lanza\(^1\)*, María de Toro\(^1\)*, M. Pilar Garcillán-Barcia\(^1\), Azucena Mora\(^2\), Jorge Blanco\(^2\), Teresa M. Coque\(^3\)\(^4\)\(^5\), Fernando de la Cruz\(^1\)*

\(^1\)Departamento de Biología Molecular (Universidad de Cantabria) and Instituto de Biomedicina y Biotecnología de Cantabria IBBTEC (UC-SODERCAN-CSIC), Santander, Spain, \(^2\)Laboratorio de Referencia de *E. coli* (LREC), Departamento de Microbiología y Parasitología, Facultad de Veterinaria, Universidad de Santiago de Compostela, Lugo, Spain, \(^3\)Departamento de Microbiología, Hospital Universitario Ramón y Cajal, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCS), Madrid, Spain, \(^4\)Unidad de Resistencia a Antibióticos y Virulencia Bacteriana asociada al Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain, \(^5\)Centros de Investigación Biomédica en Red de Epidemiología y Salud Pública, (CIBER-ESP), Madrid, Spain

**Abstract**

Bacterial whole genome sequence (WGS) methods are rapidly overtaking classical sequence analysis. Many bacterial sequencing projects focus on mobiome changes, since macroevolutionary events, such as the acquisition or loss of mobile genetic elements, mainly plasmids, play essential roles in adaptive evolution. Existing WGS analysis protocols do not assort contigs between plasmids and the main chromosome, thus hampering full analysis of plasmid sequences. We developed a method (called plasmid constellation networks or PLACNET) that identifies, visualizes and analyzes plasmids in WGS projects by creating a network of contig interactions, thus allowing comprehensive plasmid analysis within WGS datasets. The workflow of the method is based on three types of data: assembly information (including scaffold links and coverage), comparison to reference sequences and plasmid-diagnostic sequence features. The resulting network is pruned by expert analysis, to eliminate confounding data, and implemented in a Cytoscape-based graphic representation. To demonstrate PLACNET sensitivity and efficacy, the plasmidome of the *Escherichia coli* lineage ST131 was analyzed. ST131 is a globally spread clonal group of extraintestinal pathogenic *E. coli* (ExPEC), comprising different sublineages with ability to acquire and spread antibiotic resistance and virulence genes via plasmids. Results show that plasmids flux in the evolution of this lineage, which is wide open for plasmid exchange. MOB\(_{64}\)/IncF plasmids were pervasive, adding just by themselves more than 350 protein families to the ST131 pan genome. Nearly 50% of the most frequent \(\gamma\)-proteobacterial plasmid groups were found to be present in our limited sample of ten analyzed ST131 genomes, which represent the main ST131 sublineages.

**Introduction**

Clinical microbiology is being transformed by whole genome sequencing (WGS) [1]. A case in point is *Escherichia coli*: there were 1,618 *E. coli* projects submitted to NCBI compared to just 68 complete genomes by year 2013. Within the realms of clinical and environmental microbiology, plasmid analysis is increasingly used to track the dissemination of genes encoding virulence, resistance to antibiotics, heavy metals and biocides [2–4] and, to a lesser extent, to analyze differences in the adaptive evolution of certain clonal backgrounds [5,6]. Hybridization with specific probes [7], amplification of plasmid replication initiator proteins (RIP) [8–10], and relaxases (REL) [11] allow preliminary identification of plasmid families. In addition, plasmid MLST (pMLST) is used for
Plasmids are difficult to analyze in WGS datasets, due to the fragmented nature of the obtained sequences. We developed a method, called PLACNET, which greatly facilitates this analysis. As an example, we analyzed the plasmidome of E. coli ST131, an ExPEC clonal group involved in human urinary tract infections and septicemia. Relevant variation within this clone (e.g., antibiotic resistance and virulence) is frequently caused by the acquisition and loss of plasmids and other mobile genetic elements. Nevertheless, our knowledge of the ST131 plasmidome is limited to a few antibiotic resistance plasmids and to identification of replicons from known plasmid groups. PLACNET analysis extends the number of sequenced plasmids in ST131, which can be used for comparative genomics, from 11 to 50. The ST131 plasmidome is seemingly huge, encompassing roughly main plasmid groups of γ-proteobacteria. Plasmids are apparently the most active players in the dissemination of relevant genetic information.

E. coli ST131 is a successful high-risk clonal complex of pandemic distribution, able to cause extraintestinal infections in humans [13-18]. The increasing recovery of ST131 isolates from hospitalized and non-hospitalized individuals and, more recently, from companion and foodborne animals [17,19–25], sewage and main rivers of large European cities [26,27] highlights the rapid spread and local adaptation to different habitats of this lineage. ST131 is characterized by high metabolic potential [28] and a variable number of virulence factors, including adhesins, siderophores, toxins, polysaccharide coats (capsules and lipopolysaccharides), proteases and invasins [19,29,30], mostly acquired by recombination and by the interplay of mobile genetic elements (MGEs) [16]. Such traits, which are common among different lineages of the E. coli B2 phylogroup [31,32], enable strains to colonize mucosal surfaces, invade tissues, foil defence mechanisms and yield injurious inflammatory responses in the host. E. coli populations identified as ST131 by the widely used ‘Achtman scheme’ of multilocus sequence typing (MLST) [33] (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli), split in diverse clusters or subclones on the basis of genomic profile, serotype, content of virulence factors, antibiotic susceptibility pattern and the presence of certain fimH alleles [21,29,34–36]. The most prevalent ST131 clonal sublineage (H30) is characterized by the presence of a fimH30 allele, serotype O25:H4 and a specifically conserved gyrA/parC allele combination that confers fluoroquinolone resistance (EQ-R). Most human infections caused by ST131 are due to isolates of the H30 sublineage [13,16,37–39], many of them carrying the blaCTX-M-15 gene which is responsible for resistance to third generation of cephalosporins. Some authors suggested differences between CTX-M-15 and non-CTX-M-15 producers, referred to as H30-R and H30-Rx sublineages, respectively [13,35,37,38]. Currently, diverse O25:H4 ST131 variants (e.g. fimH22, fimH30) or O16:H5 (e.g. fimH41) seem also to be widely spread [13-16,40]. Full genome sequencing of several ST131 E. coli genomes, most of them H30-Rx variants [16,41–44], revealed further differences among strains, mainly chromosomal SNPs, indels and plasmid variations [16,43,44]. Heterogeneity of MGEs has been reported in other relevant E. coli clones, mainly Shiga-toxin producing E. coli (STEC) producing E. coli (STEC) as O157:H7, O104:H4 or O26:H11 [5,6,45–47], often associated with ecological diversification of E. coli populations that can influence host-pathogen interactions [48,49]. Recently, International and European organisations including European Food Safety Agency, EFSA; European Centre for Disease Control, ECDC; Food Drug Administration, FDA; Centre for Diseases Control, CDC; and national food safety authorities underscored the need to identify clonal variants with enhanced transmissibility or pathogenicity as well as to infer the evolutionary history of pathogens of interest in Public Health (http://www.efsa.europa.eu/en/events/event/140616.htm). Because relevant adaptive traits are plasmid located, there is an urgent need to consider MGEs in population genetic studies.

In this work we describe PLACNET, a method to reconstruct plasmids from WGS datasets, and its application to the comprehensive analysis of bacterial plasmidomes. As a specific example, we describe the ST131 plasmidome and discuss its possible impact in the diversification of this clinically important lineage. PLACNET allows the identification of plasmids currently circulating among E. coli and other enterobacterial species that may be underestimated, thus providing a useful tool to approach comprehensive plasmid population genetic studies.

Results

Phylogeny of E. coli ST131 genomes

We analyzed ten E. coli genomes, classified as ST131 according to the Achtman scheme (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli), which branch in three main clusters identified as ST43, ST9 and ST306 (Fig. 1) according to the cgMLST Pasteur Institute scheme (http://www.pasteur.fr/recherche/genopole/ PFS/mlst/EColi.html). The use of these two schemes is widely accepted in epidemiology [50] and increasingly used for E. coli typing. The ST43 branch contains isolates of the H30 lineage, which split in three subclusters (four strains of virotype C, two of virotype A, one of virotype B). The ST9 branch corresponds to isolates of the H22/H324 sublineage (virotype D). The most distal branch to the main cluster is represented by the commensal strain SE15, a member of sublineage H41 identified as ST306 [16]. It does not contain any marker used for the virotype subtyping method described by Blanco et al (afa, sat, ubeA, irtA) [36,51,52].
Thus, the sample analyzed in this work includes representatives of all ST131 branches described to date [13,16]. The core genome of the 10 strains encompasses 3.6 Mb (Fig. 1 inset). As can be seen, the phylogenetic tree of ST131 genomes can be rooted at the commensal strain SE15. It should be noted, however, that SE15 is not necessarily the ancestor of the pathogenic lineages, as inferred by recent evidence [16]. The divergence of SE15 from the other ST131 strains is of about 3,000 SNP/Mb, a measure of the depth of the ST131 phylogenetic branch ([187x240], [195x230]0.3% divergence in the core genome). There are only 650 SNPs among the genomes of cluster C lineage (i.e., [117x219], [125x219]200 SNP/Mb), indicating their close phylogenetic relationship. There are [165x209], [172x209]300 SNPs within a given virotype.

The average distance between clades A and B is of about 4,600 SNPs (i.e., 1,300 SNP/Mb).

Plasmid reconstruction in E. coli ST131 genomes

The PLACNET protocol was used as explained in Materials and Methods. We proceeded with plasmid reconstruction, as exemplified in Fig. 2 for the reconstruction of the E61BA genome (ST9/H324/virotype D). When we applied the rules for reference homology, scaffold links and plasmid protein tagging, the E61BA network shown as “original network” was produced. Obviously, this network was not neat enough to allow plasmid reconstruction. Expert pruning of the network consisted on several steps. First, contigs smaller than 200 bp were eliminated. Second, hubs were identified (see arrows in the original network of Fig. 2), duplicated and assigned to separate disjoint connected components. Scaffold links and coverage information, as well as score values of conflict edges, were used to decide on valid component assignment. Inspection of the coding potential of hubs usually showed them to correspond to ISs, transposons or other known repeated elements (as shown in S9 and S10 Figs.). As a result, a pruned network was reconstructed as shown in Fig. 2. Differential coloring of disjoint connected components in the pruned network thus displayed the final network of plasmids (as contig constellations). In PLACNET Cytoscape representation, most plasmids can be identified by their RIP and/or REL proteins. Thus, the reconstructed E61BA genome contains seven plasmids: a 134 kb MOBF12/IncF plasmid (pE61BA-1), a 37.7 kb MOBP6/IncI2 plasmid (pE61BA-7), a 24.5 kb MOBC12 plasmid (pE61BA-2), a 18 kb MOBP11/IncP1 plasmid (pE61BA-4), two MOBP5/ColE1-like plasmids of 6.6 and 6.9 kb (pE61BA-5 and pE61BA-6, respectively) and one MOBQ12 5.0 kb plasmid (pE61BA-3). Only plasmid pE61BA-2 could be closed, the remaining contained at least two contigs. Thus, their reported sizes are minimum sizes, since they might include small repeated sequences that were taken out of the analysis during network pruning. Two contigs remained as “not assigned” to any physical unit in this particular genome because they did not show any reference or scaffold link that bind them to other contigs: a 2,953 bp contig (containing a putative DNA primase and a lytic transglycosylase) and a 1,301 bp contig (containing two conjugation-related genes: trbI and a partial traB gene).

Thus, the sample analyzed in this work includes representatives of all ST131 branches described to date [13,16]. The core genome of the 10 strains encompasses 3.6 Mb (Fig. 1 inset). As can be seen, the phylogenetic tree of ST131 genomes can be rooted at the commensal strain SE15. It should be noted, however, that SE15 is not necessarily the ancestor of the pathogenic lineages, as inferred by recent evidence [16]. The divergence of SE15 from the other ST131 strains is of about 3,000 SNP/Mb, a measure of the depth of the ST131 phylogenetic branch ([187x240], [195x230]0.3% divergence in the core genome). There are only 650 SNPs among the genomes of cluster C lineage (i.e., [117x219], [125x219]200 SNP/Mb), indicating their close phylogenetic relationship. There are [165x209], [172x209]300 SNPs within a given virotype.

The average distance between clades A and B is of about 4,600 SNPs (i.e., 1,300 SNP/Mb).

Plasmid reconstruction in E. coli ST131 genomes

The PLACNET protocol was used as explained in Materials and Methods. We proceeded with plasmid reconstruction, as exemplified in Fig. 2 for the reconstruction of the E61BA genome (ST9/H324/virotype D). When we applied the rules for reference homology, scaffold links and plasmid protein tagging, the E61BA network shown as “original network” was produced. Obviously, this network was not neat enough to allow plasmid reconstruction. Expert pruning of the network consisted on several steps. First, contigs smaller than 200 bp were eliminated. Second, hubs were identified (see arrows in the original network of Fig. 2), duplicated and assigned to separate disjoint connected components. Scaffold links and coverage information, as well as score values of conflict edges, were used to decide on valid component assignment. Inspection of the coding potential of hubs usually showed them to correspond to ISs, transposons or other known repeated elements (as shown in S9 and S10 Figs.). As a result, a pruned network was reconstructed as shown in Fig. 2. Differential coloring of disjoint connected components in the pruned network thus displayed the final network of plasmids (as contig constellations). In PLACNET Cytoscape representation, most plasmids can be identified by their RIP and/or REL proteins. Thus, the reconstructed E61BA genome contains seven plasmids: a 134 kb MOBF12/IncF plasmid (pE61BA-1), a 37.7 kb MOBP6/IncI2 plasmid (pE61BA-7), a 24.5 kb MOBC12 plasmid (pE61BA-2), a 18 kb MOBP11/IncP1 plasmid (pE61BA-4), two MOBP5/ColE1-like plasmids of 6.6 and 6.9 kb (pE61BA-5 and pE61BA-6, respectively) and one MOBQ12 5.0 kb plasmid (pE61BA-3). Only plasmid pE61BA-2 could be closed, the remaining contained at least two contigs. Thus, their reported sizes are minimum sizes, since they might include small repeated sequences that were taken out of the analysis during network pruning. Two contigs remained as “not assigned” to any physical unit in this particular genome because they did not show any reference or scaffold link that bind them to other contigs: a 2,953 bp contig (containing a putative DNA primase and a lytic transglycosylase) and a 1,301 bp contig (containing two conjugation-related genes: trbI and a partial traB gene).
The same procedure was applied to the three other strains sequenced for this work as well as to the four genomes obtained from public DBs as Illumina reads. The plasmid content of the four strains sequenced in this work was confirmed by the analysis of S1-digested genomic DNA profiles by PFGE. This analysis fully confirmed the presence of plasmids of similar size to those identified by PLACNET (S2 Text and S4 Table). In the case of strain E35BA, in which PLACNET identified two IncF plasmids that could not be separated (totaling 211 kb), S1-PFGE identified two plasmids of 140 kb and 75 kb. As a result of PLACNET analysis, we obtained the plasmid constellation networks shown in S1 to S8 Figs. A summary of the results, i.e., the reconstructed plasmids, is shown in Table 1, which includes also the plasmids of the ST131 reference strains JJ1886 and SE15. As can be seen, the number of plasmids in the ST131 genomes is variable, even from strains belonging to the same ST131 sublineage, ranging from just one plasmid in HVH177 (clade B/ST9/fimH22) or SE15 (clade A/ST506/fimH41) to seven plasmids in E61BA (clade B/ST9/fimH22), to give an average of 4 plasmids per genome. There is not a single plasmid group that appears specific of a particular sublineage. S1 Table contains the complete list of contigs assigned to each plasmid or chromosome.

**Overall plasmid diversity is visualized in plasmid dendrograms.** Overall, the ten ST131 genomes analyzed contain 39 plasmids (including one potential ICE), which can be assorted by their relative sizes and MOB groups [53], as shown in Table 1. The most conspicuous group was that of MOBF12/IncF plasmids (11 plasmids), present in all ten sequenced ST131 genomes. Other relevant plasmid backbones belong to the MOBP (RIP groups IncI1/K, IncI2, IncX1, IncX4, ColE1), MOBQ (Q12) and MOBC (C12) REL families. The non-F plasmids comprise a total of 20 plasmids belonging to eight plasmid groups. Two plasmids were phage-like and belong to the Rep3 RIP family. Finally, 5 plasmids corresponded to the no-MOB category. The E35BA genome (ST43/H30 virotype B) showed a MOBF11 relaxase within a 234 kb chromosomal contig, implying the
Table 1. Summary of plasmid content.

<table>
<thead>
<tr>
<th>Genome</th>
<th>$fimH$ allele</th>
<th>Strain type</th>
<th>MOB$_{B1}$/IncF</th>
<th>Phage-related/RepFIB</th>
<th>MOB$_{B2}$/IncN</th>
<th>MOB$_{B3}$/IncX</th>
<th>MOB$<em>{B1}$/IncP1MOB$</em>{C12}$</th>
<th>MOB$_{B2}$</th>
<th>MOB$_{Q12}$</th>
<th>no-MOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>FV9873</td>
<td>H$_{30}$</td>
<td>A</td>
<td>pFV9873$_{5.5}$</td>
<td>(91.4 Kb, ΔTraI)</td>
<td>pFV9873$_{4}$</td>
<td>(33.3 Kb)</td>
<td>pFV9873$_{1}$ (4.1 Kb)</td>
<td>pFV9873$_{6}$ (5.2 Kb)</td>
<td>pFV9873$<em>{2}$ (2.2 Kb); pFV9873$</em>{3}$ (46 Kb)</td>
<td></td>
</tr>
<tr>
<td>BIDMC38</td>
<td>H$_{30}$</td>
<td>A</td>
<td>pBIDMC38$_{5}$</td>
<td>(123 Kb)</td>
<td></td>
<td></td>
<td>pBIDMC38$_{1}$ (11.8 Kb)</td>
<td>pBIDMC38$_{4}$ (4.2 Kb)</td>
<td>pBIDMC38$_{2}$ (5.3 Kb)</td>
<td>pBIDMC38$_{3}$ (1.6 Kb)</td>
</tr>
<tr>
<td>BWH24</td>
<td>H$_{30}$</td>
<td>C</td>
<td>pBWH24$_{1}$</td>
<td>(123 Kb, ΔTraI$^b$)</td>
<td>pBWH24$_{2}$</td>
<td>(98.3 Kb)</td>
<td>pBWH24$_{1}$ (14.2 Kb)</td>
<td>pBWH24$_{2}$ (35.0 Kb)</td>
<td>pBWH24$_{3}$ (41.1 Kb)</td>
<td></td>
</tr>
<tr>
<td>JJ1886</td>
<td>H$_{30}$</td>
<td>C</td>
<td>pJJ1886-5</td>
<td>(110 Kb, ΔTraI$^a$)</td>
<td>pJJ1886-4</td>
<td>(55.9 Kb)</td>
<td>pJJ1886-3 (5.6 Kb)</td>
<td>pJJ1886-2 (5.2 Kb)</td>
<td>pJJ1886-1 (1.6 Kb)</td>
<td></td>
</tr>
<tr>
<td>E618A</td>
<td>H$_{32}$</td>
<td>D</td>
<td>pE618A$_{1}$</td>
<td>(137 Kb)</td>
<td>pE618A$_{4}$</td>
<td>(18.3 Kb)</td>
<td>pE618A$_{2}$ (24.5 Kb)</td>
<td>pE618A$<em>{5}$ (6.5 Kb); pE618A$</em>{6}$ (6.9 Kb)</td>
<td>pE618A$_{3}$ (5.5 Kb)</td>
<td></td>
</tr>
<tr>
<td>HVH177</td>
<td>H$_{22}$</td>
<td>D</td>
<td>pHVH177-1</td>
<td>(78.6 Kb)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE15</td>
<td>H$_{41}$</td>
<td>Commensal</td>
<td>pECSF 1</td>
<td>(122 Kb)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other ST131 plasmids$^a$

- pEK516 (64.5 Kb, ΔTra); pEK499 (117 Kb, ΔTra); pJE186-2 (138 Kb); pGUE-NDM (87.0 Kb)
- pEK204 (93.7 Kb)
- pKC394 (53.2 Kb); pKC396 (44.2 Kb); pNDM-EC501 (41.2 Kb); pECN580 (64.9 Kb)
- pJIE143 (34.3 Kb)

$^a$Plasmid references: pEK516 [93]; EU935739; pEK499 [93]; EU935739; pJE186-2 [56]; NC_020271; pGUE-NDM (in [119]; JQ364967); pEK204 [93]; EU935740; pKC394 [120]; HM138652; pKC396 [120], HM138653; pNDM-EC501 (Unpublished; KJ413946); pECN580 [121]; KF914987; pJE143 (94); JN194214).

$^b$pBIDMC20B$_{1}$ and pBWH24$_{1}$ plasmids lacked the REL domain of the TrwC protein.

$^c$Plasmid pJE186-2 was isolated from strain JIE186 [94], although GenBank acc. n. NC_020271 specifies it is located at EC958 strain. Strain JIE186 also contains plasmid pJE186-1, not included in this study as it is not available at public DBs.

$^d$No correlation with RIP typing methods.

doi:10.1371/journal.pgen.1004766.t001
**Fig. 3. Hierarchical clustering dendrogram of ST131 plasmids.** The UPGMA dendrogram was based on protein cluster analysis using 60% sequence identity and 80% coverage. Plasmid names are colored according to their clade, taking into account ST, *fimH* allele and virotype, following the color code shown at the upper right. The five plasmid names in black correspond to previously sequenced plasmids from ST131 strains. Different color backgrounds are shown to emphasize branches of related plasmids. To the right of the dendrogram, four columns show, respectively, plasmid size, MOB type, RIP type and Inc type.

doi:10.1371/journal.pgen.1004766.g003
PLACNET Analysis of ST131 Genomes

Clade C (ST43/H30), virotype A
Clade C (ST43/H30), virotype B
Clade C (ST43/H30), virotype C
Clade B (ST9/22-234), virotype D
* Virotype not determined

MOBF12, IncF
presence of an ICE (integrative and conjugative element). Detailed inspection of this contig identified a 14.2 kb IME (integrative and mobilizable element) (see below and Suppl. Mat.). Once plasmids were identified by PLACNET and contigs assorted, the next step in plasmid analysis consisted in the construction of a dendrogram that summarizes plasmid gene content, allowing a visualization of relatedness between individual plasmids. The dendrogram of the 44 plasmids found in ST131 genomes (39 described in this report plus five plasmids already published) is shown in Fig. 3. The figure shows how the plasmids divide into branches that coincide with backbone MOB groups. There are 14 plasmid groups, according to the dendrogram, shown in the figure by different color backgrounds. Since each dendrogram group links related plasmids, they can be now analyzed individually, by comparing them either among themselves (Fig. 3) or with selected references (Fig. 4).

Analysis of IncF plasmids. Representatives of the largest plasmid group, formed by 15 MOBP12/IncF plasmids, were found in each of the analyzed ST131 strains (Table 1). Included in this set are four plasmids lacking MOB and tra regions but containing RIP and other backbone genes related to IncF plasmids. The dendrogram of Fig. 3 clearly indicates that these plasmids belong to the IncF plasmid family, underscoring the usefulness of this step in plasmid reconstruction. Judging from their position in the dendrogram, it seems IncF plasmid genes are scrambled as if the precise constitution of each individual IncF plasmid could not be predicted at all for isolates of each specific ST131 cluster or phylotype. This is even clearer in Fig. 4. In this figure, ST131 plasmids are represented together with the reference sequences that were used for PLACNET reconstruction and analysis. Besides, BRIG comparison of IncF plasmids (Fig. 5) reveals high heterogeneity between them, with not a single completely conserved gene (confirmed by the fact that not a single plasmid gene was found to belong to the ST131 core genome). KClust software [54] at 30% identity and 50% coverage was used to group all proteins coded by IncF plasmids into 354 reference clusters. Manual curation was used to classify these 354 clusters in three groups (Fig. 5): (i) plasmid backbone (i.e., conjugation, RIP and maintenance genes) and metabolic protein genes, (ii) antibiotic resistance and virulence genes, and (iii) other protein genes such as ISs, transposases or hypothetical proteins. Conjugation proteins represent 30 of the 53 backbone proteins and constitute the most conserved set. As mentioned above, 4/14 plasmids do not keep a complete backbone. Table 2 contains the functional annotation of the IncF plasmids. As shown, 11/15 of the MOBP12/IncF plasmids contain antibiotic resistance genes (conferring resistance to beta-lactams, in all cases, but also to sulfonamides, aminoglycosides, trimethoprim, chloramphenicol, tetracycline and macrolides, in some of them). In addition, nine of the ten antibiotic resistance-plasmids confer a multidrug-resistance (MDR) phenotype (equal or more than four antibiotic families). Besides, 10/15 MOBP12/IncF plasmids presented putativevirulence genes (Table 2). As previously noted, there was an apparent trade-off between antibiotic resistance and virulence, genes coding for these adaptive traits being located in different plasmids [2]. Finally, a DNA modification gene (adenine-specific DNA methylase) was conserved in all IncF plasmids except in pHVI177_1.

The ST131 IncF plasmids belong to four different branches of the dendrogram, as shown in Fig. 4 inset. Group I includes four plasmids similar to the well-known virulence plasmids pAPEC-ColV like (also called pS88-like), which are commonly detected among avian pathogenic E. coli (APEC) [2,55]. A suitable reference is the ST131 plasmid pHJE186-2, coming from a ST131 strain previously recovered in Australia in 2006 [56]. As shown in S11A Fig., group I IncF plasmids share two large homologous regions: a 70 kb region containing virulence genes iss, iroBCDEN, iucABCD, intA, cvwBC and sitG and the cassette ompT-bhyF-mig14, eventually also linked to estABCD [55] and a 40 kb region containing the tra region and other backbone genes. Group II contains 10 MDR plasmids, 8 of which are ST131 plasmids with characteristic F2:A1:B- replicons and multiple antibiotic resistancegenes. A suitable reference is the ST131 plasmid pJJ1886-5, coming from a ST131 strain pJJ1886-5, coming from a ST131 strain previously recovered in Australia in 2006 [56]. As shown in S11B Fig., group II IncF plasmids share most of their genomes. It should be noted that three of these plasmids (pFV9873_5, pEK999 and pEK516) have extensive deletions within their tra regions, as seen in the figure. Groups III and IV are just represented by one plasmid each. None of them contains antibiotic resistance genes and they are poor in virulence genes. While group III plasmid pHVII177_1 is not similar to any reference outside the backbone genes (S11C Fig.), the group IV plasmid pEC5F1 is extensively similar to various large E. coli plasmids, (S11D Fig.). A more comprehensive comparison of F plasmids recovered from ST131 with previously described F-like plasmids is given in Suppl. Mat.

Analysis of other ST131 plasmid groups. Besides MOBP12/IncF plasmids, 28 other plasmids were represented in ST131 isolates (Table 1). Ten were large, presumably conjugative plasmids (>18 kb), while 17 were small plasmids (<12 kb) and there was one IME.

Among the large plasmids, a most remarkable branch is composed by two almost identical 109 kb plasmids (pBIDMC20B_2 and pBWH24_2) present in two ST131/H30 isolates of virotype C, for which only RepFIB (Rep3-superfamily) and the maintenance protein ParB could be identified as plasmid backbone genes. No conjugative genes were detected. On the other hand, they code for an integrase protein and several phage-like proteins. The plasmids are highly similar to pECOH89, recently recovered from a CTX-M-15 producer E. coli isolate from Germany [57]. Closest reference hits were the adherent invasive E. coli (AIEC) plasmid pLFS2, isolated from a patient with Crohn’s disease [58], the STEC plasmid p09EL50 [5], the Salmonella plasmid p9CM2 [59] and the Salmonella bacteriophage SSU5 [60]. These are all cryptic plasmids isolated from pathogenic enterobacteria that have been barely analyzed and thus are poorly annotated. The possibility arises of these elements being similar to lysogenic phages that are stably maintained as plasmids, analogous to phage P1 [57]. S12 Fig. compares this branch of related plasmids, using plasmid pECOH89 as a reference. As can be seen, both ST131 plasmids share most of their sequences with this 111 kb plasmid, including several phage-like protein genes. Significantly, none of the cryptic plasmids described in this study or those mentioned in the references, except pECOH89, harbor a resistance gene.

The 98.3 kb MOBP12/IncK plasmid pE2022_1 is most similar to pCT [61]. pE2022_1 contains a blacTX-M-14 gene identical to that in pCT, a plasmid carrying blacTX-M-14 that is globally spread among humans and animals and particularly prevalent in clinical
Two ST131 MOBP6/IncI2 plasmids are rather different, as judged only distantly related in its backbone genes to environmental different sizes (56 and 18 kb, respectively). Plasmid pE61BA_4 is obtained from different isolates identified as E. coli SL483 and the enterohemorrhagic E. coli (EHEC) plasmid pO157_Sal [65]. They were recovered from isolates form Spain [61, 62]. The backbone of plasmid pE2022_1 is homologous to that of the reference IncI1 plasmid pEK204. These plasmids are described in S13. Fig. Despite their different Inc names, IncI1, IncK and IncB/O have similar backbones, belonging to different branches of the IncI complex (an analogous case to IncF plasmids).

MOBp6 is a large plasmid family, as can be observed in the phylogenetic tree of the MOBp6 relaxase family (S14A Fig.). The two ST131 MOBp6/IncI2 plasmids are rather different, as judged by the distant positions of their REL. Plasmid pBWH24_3 (60.3 kb) is similar to the IncI2 prototype R721 [63], but most similar to the APEC plasmid pChi7122_3 [64]. In turn, pE61BA_7 (37.9 kb) is most similar to Salmonella enterica plasmid SL483 and the enterohemorrhagic E. coli (EHEC) plasmid pO157_Sal [65]. They were recovered from isolates identified as H30_virotype C from the USA and H324_virotype D from Spain, respectively.

Another ST131 important group is MOBp12/IncX, composed by three ST131 plasmids (pFV9873_4 and pE2022_3, as well as the reference pJJ1886_4). Plasmids pFV9873_4 and pE2022_3, obtained from different H30 subgroups, are rather different between them and belong to different plasmid groups (IncX1 and IncX4), as shown their relatively distant positions in the MOBp12 phylogenetic tree of S15A Fig., even if showing similar sizes of about 34 kb. Their coding capacity is shown in the BRIG phylogenetic tree of S15A Fig., even if showing similar sizes of about 34 kb. Their coding capacity is shown in the BRIG representations shown in S15B and S15C Fig. The IncX1 plasmid pFV9873_4 is most similar to the EC plasmid p2ESCUM [66], although genetic similarity is constrained to their backbone genes, occupying about half of the reference plasmid sequences. Conversely, the IncX4 plasmid pE2022_3 is most similar to pSH696_3 and the ST131 reference plasmid pJJ1886_4 over its sequence length (S15C Fig.).

Two plasmids and the IME belong to the MOBp11/IncP1 family, as shown in the MOBp11 relaxase phylogenetic tree of S16A Fig. Plasmids pJJ1886_4 and pE61BA_4 showed widely different sizes (56 and 18 kb, respectively). Plasmid pE61BA_4 is only distantly related in its backbone genes to environmental plasmid pMBUI2, isolated from an uncultured bacterium [67]. Plasmid pJJ1886_4, on the other hand, is similar to the E. coli plasmid pHS102707 (GenBank Acc. No. KF701335). These two plasmids thus represent new additions to the ST131 plasmidome (see S16B Fig.).

The last potentially conjugative plasmid is the 24.5 kb MOBp12 plasmid pE61BA_2, also not closely related to any reference, as seen in the REL phylogenetic tree of S17A Fig. This plasmid resulted in a single contig, so it could be closed. The closest homolog is the Yersinia pestis plasmid pCRY, with which it shares all backbone genes. Its RIP belongs to the Rep_FI superfamily, although this plasmid group is not represented in the classical PBRT method [8]. It does not contain any gene with known adaptive function, except a protease and a putative secreted thermonuclease. As in the case above, this plasmid represents a new addition to the ST131 plasmidome (S17B Fig.), its relaxase being only 70% identical to its closest homolog, the pCRY relaxase.

Among the small plasmids, the first group is composed by four MOBp7/ColE1-like plasmids (11.8, 6.9, 6.6 and 5.6 kb). Three of the plasmids are relatively different, as judged by the MOBp7 phylogenetic tree of S18A Fig. The large plasmid (pBIDMIC38_1) contains a type II restriction-modification system (Cfr10I) and is almost identical to the ST131 reference pJJ1886_4 (S18B Fig.). The two MOBp7/ColE1 plasmids of strain E61BA (plasmids pE61BA_5 and pE61BA_6) contain colicin ColE and ColK genes, respectively (S18C and S18D Fig.). Colicins are considered both as virulence factors as well as traits that influence bacterial fitness and survival in the presence of competitors [68].

Besides, there were four almost identical MOBp20 plasmids of around 4.1 kb (S19A Fig.), which populate all H30 subgroups (two of virotype A, one of virotype B and one of virotype C). Nothing remarkable could be distinguished in their genetic constitution, besides a common MOB region and a pGWZ12-like Rep protein.

Fig. 5. MOBp12/IncF plasmid analysis. Protein cluster analysis was performed with kClust software (parameters: 30% identity, 50% coverage) on the set of 14 plasmids shown in Table 4. Plasmid pGUE-NDM [119] was excluded from this comparison since it is only distantly related to the others (see dendrogram in Fig. 5). A total of 354 protein clusters were obtained and annotated versus the NCBI protein database (Blastp). Manual inspection was carried out to classify the reference proteins of each cluster into one of these three groups (comparative analysis shown with BRIG): (i) Backbone and metabolic proteins (panel A); (ii) Virulence and Antibiotic resistance proteins (panel B); and (iii) ISs and hypothetical proteins (not shown). doi:10.1371/journal.pgen.1004766.g005
MOBQ12 plasmids have received little attention because the phenotype could be pointed out (S20 Fig.). MOB Qu and proteins but, as in the case of the MOBQu plasmids, no

and H22 (one of virotype D). They contain RIP and REL are also represented in other 5.0 kb). They all contain distinguishable Rep proteins MOB cryptic plasmids (four of them were 1.5 kb long and the

Discussion

There are two aspects of this work that will focus the discussion. On one side, the applicability, usefulness and limitations of PLACNET will be discussed. On the other, the plasmidome of E. coli ST131 genomes that were reconstructed by PLACNET will be analyzed as an example of the applicability of the method. Analysis of the individual reconstructed plasmids, meant for plasmid specialists, is expanded in S1 Text.

Bacterial genomes and plasmid reconstructions

Most bacterial genomes contain more than one physical unit of DNA. Besides the main chromosome, some bacteria contain additional chromosomes and most contain plasmids. We propose that PLACNET can be used as a new method to analyze bacterial genomes. It allows the assignment of chromosomes and plasmids as separate physical units within a genome. Visual representation of the network in Cytoscape, in which plasmids appear as separate physical units within a genome. Visual representation of the network in Cytoscape, in which plasmids appear as

Table 2. Resistance genes and virulence determinants in MOBF12 plasmids.

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Strain Virotype</th>
<th>Size (Kb)</th>
<th>RIP (FAB formula)*</th>
<th>Antibiotic resistance genesb</th>
<th>Virulence genesc</th>
</tr>
</thead>
<tbody>
<tr>
<td>pFV9073_5</td>
<td>A</td>
<td>91.4</td>
<td>RepFIIA-RepFIA (F2A:1:B-)</td>
<td>blaCTX-M-15, blaOXA-1, sul1, adaA5, aac(6')-Ib-cr, dfrA17, catB4, tet(A), mhpA</td>
<td>None-detected</td>
</tr>
<tr>
<td>p8IDMC38_5</td>
<td>A</td>
<td>123</td>
<td>RepFIIA-RepFIA (F2A:1:B-)</td>
<td>blaTEM-1, sul1, adaA5, tet(A), mhpA</td>
<td>traT, tetO</td>
</tr>
<tr>
<td>pE2022_2</td>
<td>C</td>
<td>103</td>
<td>RepFIIA-RepFIA (F2A:1:B-)</td>
<td>blaCTX-M-15, blaOXA-1, sul1, adaA5, aac(6')-Ib-cr like, tet(A)</td>
<td>traT, tetO</td>
</tr>
<tr>
<td>p8IDMC208_1</td>
<td>C</td>
<td>128</td>
<td>RepFIIA-RepFIA (F2A:1:B-)</td>
<td>blaOXA-1, sul1, sul2, strA, strB, adaA5, dfrA17, tet(A)</td>
<td>tetO</td>
</tr>
<tr>
<td>p8WH24_1</td>
<td>C</td>
<td>123</td>
<td>RepFIIA-RepFIA (F2A:1:B-)</td>
<td>blaOXA-1, sul1, sul2, strA, strB, adaA5, dfrA17, tet(A)</td>
<td>tetO</td>
</tr>
<tr>
<td>pJJ1886-5</td>
<td>D</td>
<td>110</td>
<td>RepFIIA-RepFIA (F2A:1:B-)</td>
<td>blaTEM-1, blaOXA-1, aac(6')-Ib-cr</td>
<td>None-detected</td>
</tr>
<tr>
<td>pEK999</td>
<td>-</td>
<td>117</td>
<td>RepFIIA-RepFIA (F2A:1:B-)</td>
<td>blaCTX-M-15, blaOXA-1, sul1, adaA5, aac(6')-Ib-cr, dfrA17, tet(A), mhpA</td>
<td>None-detected</td>
</tr>
<tr>
<td>pEK516</td>
<td>-</td>
<td>64.5</td>
<td>RepFIIA (F2A:1:B-)</td>
<td>blaCTX-M-15, blaOXA-1, sul1, adaA5, aac(6')-Ib-cr, dfrA17, tet(A), mhpA</td>
<td>None-detected</td>
</tr>
<tr>
<td>pGUE-NDM</td>
<td>-</td>
<td>87.0</td>
<td>RepFIIA (F2A:1:B-)</td>
<td>blaOXA-1, blaOXA-1, blaOXA-1, sul1, adaA5, aac(6')-Ib-cr, dfrA12, dfrA14, dfrB4</td>
<td>traT</td>
</tr>
</tbody>
</table>
| pE61BA_1   | D               | 137       | RepFIIA-RepFIB (F2A:1:B1) | blaTEM-1, tet(A) | ompT, iss, iroBCDEN, iucABCD, cvaBC, traT, etsABC, mig14-hly-f
| pECSF1     | Commensal       | 122       | RepFIIA-RepFIB (F2A:1:B1) | None-detected | traT, tetO       |
| pE35BA_2+3 | B               | 211       | RepFIIA-RepFIB (F2A:1:B1) | blaCTX-M-15, blaOXA-1, sul2, strA, strB, aac(6')-Ib-cr, aac(3')-Ila, dfrA14, catB4, tet(A) | traT, etsABC, mig14-hly-f
| pVHV177_1  | D               | 78.6      | RepFIB (F2A:1:B31)    | None-detected | traT, etsABC, mig14-hly-f

*FAB formula according to http://pubmlst.org/plasmid/classification-scheme [9].

1According to the ARG-annot database (>90% amino acid identity) [http://en.mediterranee-infection.com].

2According to our in-house database (>90% amino acid identity).

3aac(6')-Ib-cr like presents the Glu72Gly additional mutation.

4In the original paper [93] dfrA7 is reported, instead of dfrA17. However, inspection of its amino acid sequence indicates it is a DfrA17 protein.

doi:10.1371/journal.pgen.1004766.t002

(S19B Fig.). Four very similar MOBQ12 plasmids (around 5.2 kb) are also represented in H30 (two of vireotype A, one of vireotype C) and H22 (one of vireotype D). They contain RIP and REL proteins but, as in the case of the MOBQ12 plasmids, no phenotype could be pointed out (S20 Fig.). MOBQ12 and MOBQ12 plasmids have received little attention because they are cryptic and remain unnoticed in most typing schemes. The present ST131 plasmidome analysis suggests they can be surprisingly abundant in E. coli. Finally, there were five no-MOB cryptic plasmids (four of them were 1.5 kb long and the other 5.0 kb). They all contain distinguishable Rep proteins (Rep_HTH_36 superfamily), without assignment in the PBRT method. Four of them are almost identical among themselves (S21 Fig.), while the fifth (pFV9073_3) was unique and unrelated to any reference. A detailed analysis of MOBF11/IncN plasmid family is detailed in S22 Fig.
application also includes multi-chromosome bacteria like *Vibrio* or *Brucella*, where it correctly predicts both chromosomes present in these species. One *Vibrio cholerae* Pacini 1854 genome (Bioproject ID: PRJEB2215) is shown in S24 Fig, as an example. Once contigs belonging to each plasmid are defined, classical plasmid analysis ensues, as explained in the Results section. Contigs selected as part of a single plasmid are taken together and its overall proteome used to build a clustering dendrogram with reference plasmids present in the network. The dendrogram tree gathers plasmids according to the number of homologous proteins they share, providing an indication on prototype plasmids closely related with the query plasmid. There are two issues in PLACNET analysis that require additional work and for which additional improvement can be expected:

### Unassigned contigs and reference sets.

After plasmid reconstruction, occasionally, one or a few contigs may remain unassigned. In the set of ST131 genomes analyzed in this work, there were only two unassigned contigs (>200 bp), both in E61BA (Fig. 2). The fact that only two unassigned contigs appeared in the analysis of eight *E. coli* genomes suggests that this is not a quantitatively important problem. As could be expected, unassigned contigs are more frequent in genomes for which there are fewer references available. The lack of a suitable reference set results in poor quality clustering and an increased fraction of contigs without references. It is obvious from the preceding discussion that bacterial taxons for which not enough references exist will be more problematic for plasmid reconstruction. Thus, any such project should start by the generation of a sufficiently ample set of plasmid references. In this respect, *E. coli* constitutes probably the best choice, due to the large reference set available.

### Repeat sequences and difficult plasmids.

Usually PLACNET works well because contigs belonging to individual plasmids pair with different selected references and thus cluster in disjoint connected components in the Cytoscape representation after a single pruning step. The pruning step consists in identifying the bridging contigs (network hubs) as repeat sequences (RS). Two sets of evidence were used: (i) homology to known ISs or transposons, and (ii) existence of three or more scaffold links. Contigs fulfilling these two criteria were assumed to be in fact repeated in the connected network. Thus, the pruning operation consisted of duplicating the alluded nodes and splitting their scaffold links. In the tested set of *E. coli* genomes that were used to validate PLACNET (the set of eight ST131 genomes analyzed here, the 32 genome set analyzed by de Been et al. [2014] < submitted to Plos Genet together with this work >, plus another set of 10 other ESBL genomes obtained from clinical strains of bioprojects PRJNA186205 and PRJNA202876), there was only one case in which RS pruning operation was not sufficient to obtain disjoint components. It was the case of genome E35BA, where two coexisting MOB_{F12}/IncF plasmids (pE35BA_{2} and pE35BA_{3}) could be inferred, but repeated pruning did not result in disjoint components. The evidence for the existence of two plasmids was the finding of two sets of contigs containing REL and other plasmid backbone genes. PLACNET failed in discriminating both plasmids probably because network links to references were interlocking, since several PLACNET-selected references established best hits to different components of each set. Besides, the assembly program could not distinguish among parts of both sequences and considered them as RSs. Closely related plasmids that coexist in a given cell poses the most serious problem we encountered in the application of PLACNET.

The ST131 plasmidome

HGT plays a critical role in shaping bacterial lineages, especially those of multi-environment opportunistic pathogens. Comprehensive characterization of plasmidomes has been impeded by methodological limitations, although they are essential for multilevel population genetics analysis, an approach necessary to explain selection and diversification of bacterial populations and to understand the reservoir dynamics of antibiotic resistance and virulence genes [69]. The application of PLACNET to ST131 genomes allowed the detection of emerging plasmid variants, important for the evolutionary history of this ExPEC lineage, which constitutes an outstanding example of a “high risk clonal complex”, a concept increasingly important in Public Health [69].

### Plasmidome description.

We describe a remarkable heterogeneity of plasmids among the *E. coli* ST131 genomes analyzed, with the identification of 39 plasmids to add to the 11 plasmids in the ST131 lineage already sequenced (Table 1). Interestingly, these plasmids encompass 8 out of 17 main MOB plasmid groups found in the whole class of γ-proteobacteria [11,53], namely F12 (IncF), P3 (IncX), P5 (ColE1), P6 (IncE2), P11 (IncP), P12 (IncI/KB), Q12 (Rep_pSC101-like), Q9 (Rep_pMG828-2/IGWZ12-like), several of them undetectable by PBRT. Besides conjugative or mobilizable plasmids, there were other plasmids, lacking REL, but identifiable by their RIP proteins. An in-depth analysis of each plasmid group identified in this study has been diverted to a Supplementary Discussion (though exciting for clinical epidemiology or plasmid biology, it is outside the mainstream goal of this work). Our findings substantially enlarge the repertoire of plasmids identified among *E. coli* ST131 isolates, which now reflect a genome widely open for plasmid infection. It is of note that this scenario has also been described for *E. coli* clones of different pathogens [47,70–73]. Comparative genomics of the few *E. coli* lineages comprehensively analyzed to date suggests that this species is a generalist able to colonize and infect humans. It also suggests that phages and plasmids make an important contribution to specialization by accessorizing the genome with new adaptive traits and tools that modify genome structure and, eventually, by modifying transcriptional regulation [70,71,74,75]. It should also be emphasized that almost all available studies on ST131, included this one, focused on strains involved in the spread of antibiotic resistance genes, which constitute, undoubtedly, a biased fraction of the ST131 plasmidome and thus preclude an accurate view of its evolutionary history [76] (see also below).

### Plasmids and *E. coli* diversification.

Specific ExPEC lineages have scarcely been analyzed in the context of multilevel population genetics with the exception of punctual cases involving clonally unrelated isolates [70]. A recent phylogenomic analysis of 95 ST131 isolates from different geographical areas identified the same three clusters studied in the present work [16]. This analysis concluded that point-mutations and recombination events associated with diverse MGEs, including prophages and genomic islands, determined the diversification of this ExPEC lineage. However, the diversity of plasmids was only inferred by searching for incompatibility regions based on PBRT schemes. The role of plasmids in genomic versatility were not further analyzed [16]. Even though our study analyzed a smaller number of isolates, some observations can be drawn about the role of plasmids in the diversification of the ST131 lineage.

The rate of mutagenesis of *E. coli* has been roughly estimated in one mutation per genome per year [77]. Although this number is no doubt controversial, such study and those addressing the role of recombination in the ST131 lineage add context to understand its evolution as represented in Fig. 1. Compared to the limited
sequence divergence among the ST131 core genomes (the ST43/ H30 branch includes just about 600 SNPs), plasmids represent a very active fraction of ST131 adaptive evolution, as can be concluded from the analysis of Table 1. Such plasmid variability suggests that independent plasmid acquisitions and losses frequently occur within and between ST131 sublineages. Within the H30 cluster, the presence of antibiotic resistance plasmids is remarkable, specially the identification of structurally similar F2:A1:B- plasmids carrying genes conferring antibiotic resistance, since early acquisition of blaoXK, mainly associated with F2 plasmids, is considered a key event in the selection of the ST131 cluster C/H30 subclone [13,16,17]. The modular structure of F2 plasmids, containing multiple copies of ISs, facilitates gene rearrangements and the interchange of antibiotic resistance platforms linked to resistance to first line antibiotics between plasmids of the same and different families. This notion, inferred from our present analysis, has already been proposed [17,78–80] and is of great concern nowadays because of the increasing risk of encountering E. coli isolates carrying blaKPC of blaoXK genes predominant in Klebsiella (http://www.cdc.gov/drugresistance/threat-report-2013/) [81]. Beyond F2 plasmids, the presence of other plasmid groups (N, H1/K/BO, I2, A/C, X) carrying antibiotic resistance genes is observed at variable rates in this and other studies, clearly influenced by local ecology. Most of these antibiotic resistant non-F plasmids occur only rarely in E. coli isolates [82]. This could be due to an intrinsic lack of fitness of these plasmids in E. coli under natural conditions. Alternatively, they could represent cryptic indigenous plasmids now identifiable because of the acquisition of antibiotic-resistance cassettes. Nevertheless, the acquisition of mosaic regions carrying multiple antibiotic resistance genes by broad host plasmids (e.g. N, I2, A/C) increases the risk to spread resistance to first line antibiotics to different bacterial species in and outside hospitals [79,83]. Besides antibiotic resistance plasmids, an outstanding finding of this work was the frequent detection other plasmid groups, generally considered cryptic, that are clearly underrepresented in previous ST131 studies, as ColE1, MOBQ, and phage-like Rep3 plasmids. All of them are highly heterogeneous plasmid groups able to acquire adaptive traits or contribute to the mobilization of other plasmids (See supplementary dataset for details).

Members of the ST131 plasmidome such as MOB12/IncF, MOBII/IncA/C and Rep3/phage-like plasmids can also shape the E. coli chromosome by facilitating mobilization in trans of genetic islands or integrating new genetic material [32,70,84]. Interestingly, recombination of large chromosomal regions occurring at the sites of insertion of either prophages or transferable genomic islands seems to have contributed to the split of the ST131 lineage in different clusters [15]. Although experimental studies on ST131 did not yet associate plasmids with genome structure, the hypothesis is plausible taking into account the frequency at which such events occur in other B2 E. coli populations.

On a more general note, our study identifies antibiotic resistance plasmids, which are favored in high density environments, such as the human gastrointestinal tract, and under antibiotic selective pressure, that are predominant in hospitals [85] together with phages (or phage-like cryptic plasmids), apparently predominant in low density environments [86,87], and other cryptic plasmids (frequently very small and devoid of any possible adaptive gene). This mixed constitution, which is difficult to understand on purely selective grounds, highlights potential roles of plasmids in the context of multilevel selection, a recurrent issue in evolutionary biology. The plasmid flux in ST131 strains occurs while disseminating genes coding for resistance to extended spectrum beta-lactamases (ESBL). The results presented here find a complement in the study of de Beutin et al. (accompanying paper), where the authors document the dissemination of ESBL-carrying epidemic plasmids from animal to human clonally unrelated E. coli lineages. Thus, many plasmids appear in a clonal lineage, and many lineages can be infected by a single predominant plasmid. These conceptual notions have relevance in Public Health as they deal with the hierarchical units of selection that contribute to increase the population size of antibiotic resistance genes in human and animal pathogens [69,88,89].

In summary, our study reveals the utility of PLACNET in multilevel population genetics analysis, critical to understand the evolutionary processes and dynamics of both bacterial and plasmid lineages. Its application to E. coli ST131 allowed us to infer the roles of plasmids in the dissemination of globally spread antibiotic resistance and virulence genes, some of them being underrepresented in Genbank. It is probable that these plasmids are critically relevant to understand the adaptive evolution of E. coli populations and their bacterial exchange communities. Armed with this new tool for plasmid analysis, future scrutiny of a larger number of significant strains will allow us to understand the interplay among different plasmid associations that often appear in bacterial pathogens.

**Conclusion**

The evolutionary processes of main bacterial pathogens are often discussed in the context of lineage-associated acquisition of a specific virulence gene set. The present study demonstrates how E. coli ST131 strains, even when they are practically identical in their core genomes, contain a striking variety of different plasmids. Many of them remain unnoticed, since they are apparently cryptic. Prevalent plasmids, such as IncFs, undergo frequent recombination, continuously resulting in novel gene repertoires. Our results shed light on the role of plasmids in E. coli ST131 evolution. Horizontal transmission of plasmids that carry not only antibiotic resistance and virulence genes, but also other poorly analyzed functions (metabolic genes, colicins and as yet cryptic functions) is common in the ST131 plasmidome and results in frequent and rapid adaptive changes. Arrival to these conclusions has been made possible by the application of PLACNET, a plasmid reconstruction method for WGS datasets.

**Materials and Methods**

**Epidemiological plasmidome analysis of bacteria and plasmids**

Comprehensive plasmidome analysis was carried out for 10 E. coli ST131 genomes, representing main ST131 sublineages described to date [13,16]. They include strains coming from Spain (three fimH30, one fimH24), USA (three fimH30), Australia (one fimH30), Denmark (one fimH22) and Japan (one fimH41). The fimH30 strains from Spain were CTX-M-15 producers and belonged to the H30-Rx sublineage (additionally, one strain was also CTX-M-14), while those collected in the USA were KPC-2 producers. The four strains from Spain represent predominant ST131 variants on the basis of PFGE patterns and the presence/absence of four putative virulence markers (afaFM955459, encoding an Afa/Dr adhesion; sat, secreted autotransporter toxin; tdh, invasion of brain endothelium; and iroN, salmochelin siderophore receptor) [36,51,52] and sequenced for this work.

The ST131 isolates studied represent epidemic variants exhibiting particular combination of putative virulence traits and were previously designed as distinct “virotypes” by capital letters A to D [36]. It should be noted that no correlation exists between
The ST131 core genome was defined as the collection of genes present in the ten ST131 genomes analyzed, with more than 90% similarity and 90% coverage. CD-HIT-EST [95] was used to cluster genes. A homemade Perl script was created to parse the cluster genes. A homemade Perl script was created to parse the cluster genes.

DNA sequencing

Total DNA from E. coli ST131 strains FV9873, E358A, E2022 and E618A was extracted with QIAmp DNA Mini Kit (Qiagen). DNA concentration was measured with Nanodrop 2000 (Thermo Scientific) and Qubit 2.0 Fluorometer (Life Technologies). 1.0 μg DNA was sonicated (20 cycles of 30 s at 4°C, low intensity) with Bioruptor Next Generation (Diagenode). Sample quality was checked in a Bioanalyzer 2100 (Agilent Technologies). DNA samples were preconditioned for sequencing by using the TruSeq DNA Sample Preparation Kit (Illumina) and quantified with Step One Plus Real-Time PCR System (Applied Biosystems). Flow-cells were prepared with TruSeq PE Cluster Kit v5-CS-GA (Illumina). Sequencing was carried out using a standard 2×71 base protocol (300-400 bp insert size) in a Genome Analyzer IIX (Illumina, San Diego, CA) at the sequencing facility of the University of Cantabria. The main statistics of the eight sequence datasets analyzed are shown in Table 4.

Phylogenetic analysis of the ST131 core genome

The ST131 core genome was defined as the collection of genes present in the ten ST131 genomes analyzed, with more than 90% similarity and 90% coverage. CD-HIT-EST [95] was used to cluster genes. A homemade Perl script was created to parse the cluster and define the core genome set. All core genes were concatenated and aligned with progressive Mauve [96]. A tabular list of SNPs was extracted from the Mauve alignment by applying the SNP export tool of Mauve GUI. The tabular list of polymorphic sites was parsed by a homemade script. A given position was counted as an SNP if it varied between two given sequences. The number of SNPs was added for each pair of strains to give the final SNP count. Polymorphic sites with gaps were removed from the SNP count matrix. The Mauve alignment was curated by trimAl [97]. RAxML [98] was used to build the core genome phylogenetic tree, using 100 replicates for bootstrap determination.

Plasmid Constellation Networks (PLACNET)

PLACNET was developed to associate contigs with specific physical DNA units in WGS experiments. Networks are powerful models that allow visualization and analysis of sequence information. PLACNET networks are composed of two types of nodes (contigs and reference genomes) and two types of edges (similarity and scaffold links). Commonly, network layout algorithms simulate repulsion forces between nodes and attraction forces by the edges that link two nodes. Thus, node distribution in the network will depend on the intensity of forces that define the edges. In such network model, a plasmid will be represented by a connected component (a set of linked nodes) or, in other words, a constellation of contigs. Different physical units (plasmids and chromosomes) should be represented by disjoint connected components (separate constellations). The workflow (Fig. 6) involves the following steps:

Assembly. Velvet assembly software [99] and its script VelvetOptimiser.pl were used to determine the best assembly and to scan the optimum parameters. Velvet provides also coverage information for each contig, which adds useful information for network interpretation.

Scaffold links. Although assembly programs perform scaffolding between contigs, when the assignation is ambiguous, contigs remain unbound. A method based in the mapping tool Bowtie 2 [100] was used to find all possible scaffold links. All reads were mapped using the contigs as references, using default parameters of Bowtie 2, with the option to report all possible hits. The output file was converted to SAM format [101] to give the potential adjacency information for each contig. We considered as potential PLACNET scaffold links those which comply with two rules: (i) the contigs were paired at their extremities, themselves defined as twice the read length, and (ii) the number of pair-end reads linking those two contigs had to be higher than one third of the mean of the total pair-end reads that scaffold all

Table 3. Human E. coli ST131 genomes analyzed in this work.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Accession*</th>
<th>Location</th>
<th>Collection date</th>
<th>Isolation source</th>
<th>Plasmid name (Accession number)*</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>HVH-177</td>
<td>PRJNA186205</td>
<td>Denmark</td>
<td>2003</td>
<td>Blood</td>
<td>pHVH177_1</td>
<td>PRJNA186413</td>
</tr>
<tr>
<td>BIDMC20B</td>
<td>PRJNA202031</td>
<td>USA</td>
<td>-</td>
<td>Urine</td>
<td>pBIDMC20B_1 and _2</td>
<td>PRJNA202876</td>
</tr>
<tr>
<td>BWH24</td>
<td>PRJNA201983</td>
<td>USA</td>
<td>-</td>
<td>-</td>
<td>pBWH24_1 to _3</td>
<td>PRJNA202876</td>
</tr>
<tr>
<td>BIDMC38</td>
<td>PRJNA202050</td>
<td>USA</td>
<td>2012</td>
<td>-</td>
<td>pBIDMC38_1 to _5</td>
<td>PRJNA202876</td>
</tr>
<tr>
<td>FV9873</td>
<td>PRIEB6262</td>
<td>Spain</td>
<td>2007</td>
<td>Urine</td>
<td>pFV9873_1 to _6</td>
<td>This study</td>
</tr>
<tr>
<td>E358A</td>
<td>PRIEB6262</td>
<td>Spain</td>
<td>2008</td>
<td>Urine</td>
<td>pE358A_1 to _3, IME_E358A</td>
<td>This study</td>
</tr>
<tr>
<td>E2022</td>
<td>PRIEB6262</td>
<td>Spain</td>
<td>2006</td>
<td>Urine</td>
<td>pE2022_1 to _5</td>
<td>This study</td>
</tr>
<tr>
<td>E618A</td>
<td>PRIEB6262</td>
<td>Spain</td>
<td>2008</td>
<td>Abscess</td>
<td>pE618A_1 to _7</td>
<td>This study</td>
</tr>
<tr>
<td>SE15</td>
<td>AP009378</td>
<td>Japan</td>
<td>-</td>
<td>Feces</td>
<td>pEC5F1 (AP009379)</td>
<td>[91]</td>
</tr>
<tr>
<td>JJ1886</td>
<td>NC_022648.1</td>
<td>USA</td>
<td>-</td>
<td>Urine</td>
<td>pJJ1886_1 to _5 (NC_022649; NC_022650; NC_022651; NC_022656; NC_022661; NC_022662)</td>
<td>[90]</td>
</tr>
</tbody>
</table>

*PRJNA and PRIEB6262 accession numbers correspond to SRA datasets. AP009378 and NC_022648.1 correspond to finished genomes.

The ST131 clonal variants [16]. Other genome datasets were taken from Bioproject NCBI database (https://www.ncbi.nlm.nih.gov/bioproject/), or from NCBI genomes database (E. coli [J]1886 [90] and E. coli SE15 [91]). In addition, fully sequenced plasmids pEK499, pEK204, pEK516, pHE186-2 and pHE143, previously found in other ST131 isolates [56,92–94], were used for plasmid comparisons. Information about all genomes is detailed in Table 3.
contigs. This procedure was implemented as an in-house Perl script.

**Reference search.** At least for bacteria widely covered by sequencing projects, most contigs in any new sequence are similar to one or more previously published sequences (reference sequences). Our initial hypothesis was that, for any physical DNA unit, its contigs will “BLAST” a related set of reference sequences. Thus, in the PLACNET network representation, they will cluster around the homologous references. The more DNA databases grow, the closer the references will be to the query sequence and the better PLACNET will work. A homemade BLAST [102] database was constructed from the NCBI genomes database by joining all sequences contained in [ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/] and [ftp://ftp.ncbi.nlm.nih.gov/genomes/Plasmids/]. The version used in this work was from March 7th 2013 and contains 6,432 genomes (plasmids and chromosomes). Megablast search of all contigs was carried out against the homemade BLAST-genome database with the objective of selecting a few best matches for network construction. Due to the different length of each contig, fixed thresholds by e-value or score cannot be chosen. Since the score is not a normalized parameter, and varies depending on sequence length, hits were selected by applying a dynamic threshold, based on the number of homologous sequences and the score of each hit. If the threshold is defined as 85% + 2n of the mean of the n previous sequence scores, and n is the ranking position of sequence i retrieved by megablast with score S_i, then T_{n+1} is the threshold for sequence n+1:

\[ T_{n+1} = \frac{\sum_{i=1}^{n} S_i}{n} \times (0.85 + 2n) \]

All reference sequences above the threshold were taken as nodes in the PLACNET representation.

**Protein prediction of replication initiator proteins (RIP) and relaxases (REL).** Some genes are indicative of a plasmid sequence. Among them we selected REL, key proteins in the conjugation process [53,103] and RIPS, key proteins in the replication of most plasmids [104]. Although not all plasmids contain a RIP and/or REL, their presence in a contig is diagnostic of a plasmid (or ICE) sequence. Some plasmids have more than one RIP (i.e. IncF family plasmids) [9,105,106] but plasmids rarely have more than one REL [107]. ORF prediction was carried out by GeneMark [108], which optimizes predictions based on GC content of DNA. The heuristic prediction implemented in this software is especially useful to predict ORFs in plasmid-containing genomes because it takes each contig individually and selects the best prediction model case by case.

To implement specific search protocols for REL and RIPS, three homemade databases (DB) were developed. A REL database (REL-DB) was constructed according to [53,109]. Similarly, RIP-DB was constructed from all RIPS annotated in UniProt database. RIP sequences were clustered by CD-HIT [95], using 40% identity as a threshold. Next, a Hidden Markov Model was built from each cluster by hmmer3 [110]. Finally, the HMM profiles were used in a search against UniProt. The HMM search and initial dataset were joined in one database (RIP-DB). An additional step was necessary to classify RIPS according to the widely used plasmid classification protocol PBRT [8]. A homemade nucleotide database (INC-DB) was created to identify PBRT types in silico using blastn. Finally, the relevant ORFs (REL and RIPS) identified to specific contigs by using REL-DB and RIP-DB
were incorporated to the network as tags. All relevant steps in network construction were implemented as a Perl script available at the following web page: http://placnet.sourceforge.net/.

**Plasmid constellations.** As explained above, each plasmid is represented in PLACNET by a connected component (a constellation). Thus, different physical units (plasmids and chromosomes) should be represented by disjoint (unlinked) connected components. Cytoscape software [111] was used to visualize and analyze plasmid constellations, which incorporate all the information (similarity to reference sequences, scaffold links, and protein tags) in a single network. Node attributes such as contig size, coverage and reference description, are added to the network. At this stage, network pruning is needed to resolve individual plasmids as disjoint components. When a genome has a number of repeated sequences (e.g., insertion sequences (ISs) or transposons), or two very similar plasmids, the assembly process outputs those sequences as contigs with multiple scaffold links. In the network context, they represent hubs, that is, nodes with a high number of connections. This makes the network very dense and complicates the analysis of network connected components. In PLACNET, contigs smaller than 200 bp were directly eliminated from the analysis. Hubs were examined by blastx against protein databases (i.e. UniProtKN or NCBI nr). If there was identity to any transposase gene, the hub was duplicated, and scaffold links were partitioned among them, to maximize the number of disjoint components. Contigs that remain unbound are classified as “unassigned sequence” in the contig assignation table.

**Plasmid definition, dendrograms and cluster analysis.** The final steps in plasmid reconstruction involve the definition and verification of each plasmid. This is an iterative process, as shown in Fig. 6. First, each contig was assigned to a putative plasmid (or chromosome) based on visualization of disjoint connected components in the Cytoscape representation. Assignments take into consideration additional types of evidence like the presence of REL and/or RIPS, size of the putative plasmid compared to reference plasmids and coverage of each contig (contigs belonging to the same plasmid must have similar coverage). Taking into account the information provided by related genomes within the same sequencing project can also be helpful (same sequences cluster around the same references). In

---

Fig. 6. PLACNET flow diagram. The diagram represents the PLACNET workflow to analyze an Illumina bacterial genome dataset. It can be separated in two sub-process: network delineation and plasmid analysis. Network delineation consists on contig assembly, determination of scaffold interactions, reference search of homologous genomes and plasmid protein prediction. Plasmid analysis basically consists in the construction of a dendrogram of plasmid protein profiles, which identifies the most relevant reference sequences, followed by plasmid cluster analysis, which compares query plasmids with its closest references. Plasmid analysis is a feedback process that helps to resolve uncertainties and results in a final definition of plasmid and chromosome content. doi:10.1371/journal.pgen.1004766.g006
this respect, PLACNET is more robust in multi-strain collections. The performance of PLACNET was validated by testing a number of previously sequenced and annotated E. coli genomes (S2 Table). The ART software (Huang et al., 2012) was used to simulate pair-end Illumina reads from those genomes, which were then analyzed by PLACNET as explained above. Results are shown in S2 Text, S3 Table and S25-S34 Figs.

After PLACNET has defined the plasmids carried in the relevant genomes, the next step in plasmidome analysis is to build a dendrogram that produces a hierarchical clustering of plasmid proteomes similar to those described in [112–114]. CD-HIT (thresholds: 70% identity and 80% coverage) was used for clustering references and query plasmids. Based on the output file, a presence/absence table (present or absence of each protein cluster in each plasmid) was built. Each table row represents a plasmid protein profile. Raup-Crick distance method, implemented in vegan package for R software [115], was used to calculate the distance matrix of plasmid protein profiles. The Ape package [116] was used to calculate the dendrogram bootstrapping confidence value. Finally, a hierarchical clustering dendrogram was built using the UPGMA algorithm.

Putative plasmids and references belonging to the same dendrogram branch were compared using BRIG [117] or Abacas [118]. While BRIG is not sensitive to contig arrangement, Abacas can be used to order contigs according to a given reference. With these tools, the curator is able to visualize the correspondence between reconstructed plasmids and references, or can go back to dendrograms or PLACNET in the search for missing or extra contigs. This iterative mode of analysis is represented in Fig. 6 by the backward arrow linking plasmid cluster analysis with plasmid definition.

Plasmids were mainly classified according to their REL in MOB families, as described by [53]. Classical Inc families are also given when typing them by in silico PBRT was possible. Plasmids that could not be classified one way or the other were termed no-MOB by exclusion.

Supporting Information

S1 Fig  PLACNET reconstruction for FV9873 genome (E. coli ST131/H30/virotype A). A total of 262 contigs were classified in chromosome and six plasmids. The black line surrounding the pFV9873_1 plasmid (4.1 bp) indicates that it is a closed plasmid. There are not unassigned contigs. (PDF)

S2 Fig  PLACNET reconstruction for E35BA genome (E. coli ST131/H30/virotype B). One 14.2 kb MOBP11 Integrative Mobilizable Element (IME) was detected in the chromosome. The pE35BA_1 plasmid (4.1 kb) is closed. There was a conflict of separation between two IncF plasmids (pE35BA_2 and pE35BA_3, total size: 211 kb). The annotation of this particular genome is limited by the quality of the assembly (many small and non-scaffolded contigs). This is the network with the highest number of contigs in our study (total: 419). No contigs remained unassigned. (PDF)

S3 Fig  PLACNET reconstruction for E2022 genome (E. coli ST131/H30/virotype C). A total of five plasmids, two of them as closed plasmids, and the chromosome were obtained in a 346-contig network. No contigs remained unassigned. (PDF)

S4 Fig  PLACNET reconstruction for E61BA genome (E. coli ST131/H324/virotype D). Seven different plasmids were obtained in the analysis. The pE61BA_2 plasmid (24.5 kb), containing a single contig, was closed. Two contigs remained unassigned, as no scaffold links were detected for them. One of them (2,953 bp) encodes for a putative DNA primase and a lytic transglycosylase, while another (1,301 bp) encodes for TraB and TraB partial proteins. (PDF)

S5 Fig  PLACNET reconstruction for BIDMC20B genome (E. coli ST131/H30/virotype C). A total of 115 contigs were fully assigned to the chromosome and two plasmids. Plasmid pBIDMC20B_2 (109 kb) appeared as a single contig that could be closed. (PDF)

S6 Fig  PLACNET reconstruction for BIDMC38 genome (E. coli ST131/H30/virotype A). Five plasmids were detected. Three small plasmids (1.6, 4.2 and 5.3 kb) are closed. No contigs remained unassigned. (PDF)

S7 Fig  PLACNET reconstruction for BWH24 genome (E. coli ST131/H30/virotype C). Three plasmids were detected, only one of them as a closed plasmid (pBWH24_2, 109 kb). No contigs remained unassigned. (PDF)

S8 Fig  PLACNET reconstruction for HVH177 genome (E. coli ST131/H324/virotype D). Only one plasmid (pHVH177_1, 78.6 kb) composed by three contigs was detected in the HVH177 genome. No contigs remained unassigned. (PDF)

S9 Fig  Contig coverage in E61BA genome (E. coli ST131/H324/virotype D). Contigs belonging to each plasmid are shown in different colors, according to the code below the histogram. Plasmid copy numbers are inferred from their contig coverage. Average coverage of chromosomal contigs was 56. All contigs with >2X average are named according to their predicted gene products. When there is more than one contig, annotations are separated by colons, if adequate. (PDF)

S10 Fig  Contig coverage in E35BA genome (E. coli ST131/H30/virotype B). Contigs belonging to each plasmid are shown in different colors, according to the code below the histogram. Plasmid copy numbers are inferred from their contig coverage. Average coverage of chromosomal contigs was 55. All contigs with >2X average were named according to their predicted gene products. If more than one contig coincided within the same coverage section, annotations of the individual contigs were separated by colons. As is shown in the histogram, contigs corresponging to the MOBF12/IncF plasmids pE35BA_243 (in green color) show similar coverage than the chromosome, which indicates that both IncF plasmids have the same copy number than the chromosome. (PDF)

S11 Fig  BRIG comparative analysis of MOBF12/IncF plasmids. These plasmids were subdivided in four groups according to Fig. 5 inset. S11A: Group I. Plasmid pJE186-2 is used as the reference for the BRIG comparison. S11B: Group II. Plasmid pJJ1886-5 is the inner reference. S11C: Plasmid F is used as a reference. S11D: Plasmid pECSF1 is used as a reference. (PDF)

S12 Fig  Comparative analysis of phage-related/RepFIB plasmids. S12A: BRIG representation using the 111 kb plasmid
S13 Fig  BRIG comparative analysis of MOBP12/IncI-complex. S13A: The IncI plasmid pEK204 is used as inner ring in the BRIG analysis. S13B: Plasmid pCT [62] was used as reference. (PDF)

S14 Fig  Comparative analysis of MOBP6/IncI2 plasmids. S14A: Phylogenetic tree of MOBP6 REL proteins, calculated as in S12B Fig. The tree was rooted with MOBP6 REL of Plasmid2 from *Nitrosomonas eutropha* C91. S14B: BRIG comparative analysis of pBWH24_3 plasmid, using pChT1222_3 as reference. S14C: BRIG comparative analysis of pE61BA_7 plasmid, using pO157_Sal as inner reference. (PDF)

S15 Fig  Comparative analysis of MOBP3/IncX plasmids. S15A: Phylogenetic tree of MOBP3 REL proteins, calculated as in S12B Fig. The tree was rooted with VirD2_pSD25 (MOBP2 subfamily). ST131 plasmids are shown in red. IncX subgroups are indicated in different color backgrounds. S15B: BRIG comparative analysis of IncX1 plasmids using pESCUM as a reference. S15C: BRIG comparative analysis of IncX4-like plasmids using pSH696_34 as reference. (PDF)

S16 Fig  Comparative analysis of MOBP11/IncP plasmids. S16A: Phylogenetic tree of MOBP11 REL proteins, calculated as in S12B Fig. The tree was rooted with NikB_R64 (MOBP12 subfamily). ST131 plasmids are colored in red. Two clearly separated groups are colored. S16B: BRIG comparative analysis of IncP plasmids using pJJ1986_4 as a reference. S16C: Comparison of pJJ1886 and E35BA genomes, showing the genetic map of the inserted IME_E35BA, and its homology to *Bukholderia glumae* Incp island. The figure was drawn with EasyFig [75]. Specific genes are specifically colored according to the code in the lower part of the figure. (PDF)

S17 Fig  Comparative analysis of MOBC12 plasmid. S17A: Phylogenetic tree of MOBC12 REL proteins, calculated as in S12B Fig. The tree was rooted with MobC_CloDF13 (MOBC11 subfamily). S17B: BRIG comparative analysis of MOBC12 plasmids using pCry as a reference. (PDF)

S18 Fig  Comparative analysis of MOBP5/ColE1-like plasmids. S18A: Phylogenetic tree of MOBP5 REL proteins, calculated as in Fig SF12B. S18B, C and D: BRIG comparative analysis of MOBP5 plasmids using ColE1 (SF18B), pJJ1886_3 (SF18C) and pColK-K235 (SF18D) as references. (PDF)

S19 Fig  Comparative analysis of MOBQ12 plasmids. S19A: Phylogenetic tree of MOBQ12 REL proteins, calculated as in Fig SF12B. The tree was rooted with the MOBQ12 subfamily. ST131 plasmids are colored in red. Different color backgrounds are used to represent MOBQ12, where ST131 MOBQ12 plasmids are located, and MOBQ12 branches. S19B: BRIG comparative analysis of MOBQ12 plasmids using pSE11-6 as a reference. (PDF)

S20 Fig  Comparative analysis of MOBQ12 plasmids. S20A: Phylogenetic tree of MOBQ12 REL proteins, calculated as in Fig SF12B. The tree was rooted with MobA_RSFI010 (MOBQ11 subfamily). S20B: BRIG comparative analysis of MOBQ12 plasmids using pCE10B as a reference. (PDF)

S21 Fig  Comparative analysis of small no-MOB plasmids. BRIG comparative analysis of no-MOB plasmids using pCE10D as a reference. (PDF)

S22 Fig  Comparative analysis of MOBF11/IncN plasmids. S22A: Phylogenetic tree of MOBF11 REL proteins, calculated as in Fig SF12B. The tree was rooted with R388 (MOBF11/IncW). IncN1 and IncN2 subgroups are indicated with different background colors. S22B: BRIG comparative analysis of IncN1 plasmids, using R46 as a reference. S22C: BRIG comparative analysis of IncN2 plasmids using p271A as a reference. (PDF)

S23 Fig  PLACNET reconstruction of the genome of *Staphylococcus aureus* strain 118 (ST772) [ID: PRJNA62607]. Assembly data: Number of libraries: 1; read length: 75 bp; number of contigs: 73; total bp: 2,798,022 bp; N50: 224673 and Kmer: 73. One 12,819 bp plasmid was identified and reconstructed. No REL or RIP proteins were detected. (PDF)

S24 Fig  PLACNET reconstruction for a genome of *Vibrio cholerae* Pacini 1854 [ID: PRJEB2215]. Assembly data: Number of libraries: 1; read length: 75 bp; number of contigs: 149; total bp: 4,022,287 bp; N50: 180089 and Kmer: 65. The two *V. cholerae* chromosomes were fully reconstructed. Chromosome I (2,992,142 bp in our study) was reported to be about 3 million bp and encodes most essential functions [95]. As shown in Fig S24, it harbors a MOBHI2 relaxase, identical to that of the integrative and conjugative element (ICE) of the SXT/R391 family [96]. Chromosome II is smaller (1,034,286 bp in our study). Finally, a 26.3 kb plasmid containing a RIP protein (87% identity with *E.coli* pABU plasmid [97]) could be reconstructed. (PDF)

S25 Fig  Cytoscope representation of the reconstructed *E. coli* JJ1886 genome. The network was constructed and codes used (in this and following figures) as explained in Fig. 6. The pruned network (Step 1) was obtained after deleting 19 contigs smaller than 200 bp. (PDF)

S26 Fig  Definition (Step 2) of *E. coli* JJ1886 plasmids p1 to p4. These plasmids contain a RIP and/or REL protein and appeared as single contigs. The inset Table shows some properties of relevant nodes. The nodes are represented in the network surrounded by circles of the same color than the background color in the Table. (PDF)

S27 Fig  Resolution of hubs and definition of plasmid p5 (Step 3). Hub nodes that were duplicated are shown in the inset Table and indicated by red arrows in the Cytoscape network. After hub duplication, the IncF plasmid p5 is shown to contain the 12 contigs surrounded by a red circle. (PDF)

S28 Fig  BRIG comparison of the reconstructed plasmid p5 with the reference plasmid pJJ1886_5. The reference plasmid (inner ring) is compared to the p5 reconstructed plasmid (purple ring).
Outer black and white ring sectors represent pJJ1886.5 gene annotations.

PDF

S29 Fig  Inverse BRIG comparison of reconstructed plasmid p5 vs pJJ1886.5 reference plasmid. The reconstructed plasmid is placed here as the reference inner ring (thin black circle line) to which the reference plasmid (purple ring) and the reconstructed p5 contigs (outer blue and red ring) are compared.

PDF

S30 Fig  Cytoscape representation of the reconstructed E. coli SE15 genome. The network was constructed and codes used as explained in Fig. 6. The pruned network (Step 1) was obtained after deleting 16 contigs smaller than 200 bp.

PDF

S31 Fig  Plasmid definition [PLACNET steps 2 and 3] of E. coli SE15 genome. Three particular nodes in the pruned network (surrounded by the red circle) were scrutinized due to their loose connection to the chromosome. As shown in the inset Table (red background files), a blastx comparison indicates they correspond to “typical” E. coli chromosomal segments, so were finally assigned to the chromosome. Three other nodes (surrounded by a green circle in the left Cytoscape representation) corresponded to hubs (green background in the inset) and were thus duplicated. The reconstructed plasmid (p1) in the final network is surrounded by a purple ring.

PDF

S32 Fig  BRIG comparison of SE15 E. coli genome reconstructed IncF plasmid (p1) with the reference plasmid pECSF1. The reference plasmid (inner ring) is compared to the IncF reconstructed plasmid (purple ring). Outer black and white ring sectors represent pECSF1 gene annotations. Three regions (5,709 bp in total) were missing from the p1 reconstruction.

PDF

S33 Fig  Cytoscape representation of the reconstructed genome of E. coli strain MG1655 containing plasmids pEC958 and R46. The network was constructed and codes used as explained in Fig. 6. The pruned network was obtained after deleting 25 contigs smaller than 200 bp and duplicating 2 hubs (surrounded by a red circle). Plasmid p1 is the reconstructed R46 while p2 is the reconstructed pEC958. Nodes surrounded by a blue circle, and described by the blue background files in the inset Table, could not be assigned. See text for further details.

PDF

S34 Fig  Final Cytoscape representation of reconstructed H0407 E. coli genome. The network was constructed as explained in Fig. 6. The pruned network was obtained after deleting 44 contigs smaller than 200 bp. Plasmids p1 and p2, represented by single contigs, are surrounded by red and blue circles, respectively. A single contig (surrounded in an intense blue circle) remained isolated from other genetic units. Nevertheless, blastx analysis demonstrate it correspond to chromosomal background, as shown in the inset Table. Nodes described by the grey background files in the inset Table correspond to one node assigned to the chromosome, plus seven hubs, which were duplicated. The green circle surrounds 52 contigs, adding 119,735 bp, which represent plasmids p3 and p4, the two IncF plasmids that PLACNET was not able to resolve (see text for further details). The red arrow indicates a node containing REL, RIP and backbone genes in common to both IncF plasmids.

PDF

S35 Fig  Hierarchical clustering dendrogram of ST131 plasmids and relevant references. The left dendrogram shows the complete tree, with references. Dendrogram construction and color codes are as in Fig. 4. The right dendrogram expands the MOBp12/IncF branch, with new background colors highlighting plasmid groups within this branch that are mentioned in the text.

TIF

S1 Table  Complete list of contigs assigned to each plasmid or chromosome.

XLS

S2 Table  Ten E. coli genomes analyzed as examples of PLACNET performance.

PDF

S3 Table  PLACNET performance on a set of ten E. coli genomes.

PDF

S4 Table  Molecular size of the plasmids estimated by S1-PFGE and PLACNET reconstruction in the four strains sequenced in this study.

PDF

S1 Text  The plasmidome of E. coli ST131.

PDF

S2 Text  PLACNET validation by reconstruction of finalized genomes.

PDF

Acknowledgments

The authors thank Carlos Revilla for expert technical assistance.

Author Contributions

Conceived and designed the experiments: VFL FdlC. Performed the experiments: VFL MtT AM MPGB. Analyzed the data: MtT VFL TMC FdlC MPGB. Contributed reagents/materials/analysis tools: JB. Wrote the paper: MtT VFL AM JB TMC FdlC MPGB.

References


8. Detection of quinolone-resistant Escherichia coli isolates belonging to clonal


4. Rapid and specific detection, molecular epidemiology, and experimental

3. Very high prevalence of multidrug-resistant clones O25b:H4-B2-ST131 and

2. Detection of AmpC β-lactamase-encoding CTX-M-15 in a Multidrug


19 December 2014 | Volume 10 | Issue 12 | e1004766