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[1,9], María de Toro[1,9], 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 (IRYCIS), 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 mobilome 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_{F12}/IncF plasmids were pervasive, adding just by themselves more than 350 protein families to the ST131 pangenome. Nearly 50% of the most frequent γ–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

Author Summary

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 50% of the main plasmid groups of γ–proteobacteria. MOB_{F12}/IncF plasmids are apparently the most active players in the dissemination of relevant genetic information.

epidemiological surveillance, but is restricted to individual plasmids of a few plasmid families of Enterobacteriaceae (http://pubmlst.org/plasmid/). This precludes the detection of plasmid mutations or rearrangements, as well as the identification of conjugative plasmids not represented in the pMLST database and of most mobilizable plasmids [11]. Finished plasmid/genome sequencing provides accurate and non-biased information, but is still expensive and thus seldom used specifically for plasmid analysis. Draft WGS dramatically cut down cost and analysis time. Although it allowed rapid and cheap data acquisition, WGS datasets typically result in more than a hundred contigs for a given genome, due to the short read lengths generally obtained. Genome fragmentation makes it difficult to distinguish between physical units, that is, between chromosome and plasmid sequences, as well as between different plasmids that usually coexist in bacterial cells. Several strategies can be followed to analyze WGS genome sequences, the workflow described by [12] being a typical example. There are also applications to identify plasmids in WGS sequences, such as PlasmidFinder (http://cge.cbs.dtu.dk/services/PlasmidFinder/), which identifies plasmids according to PCR-based replicon typing (PBRT) [8–10] and the subtyping scheme included in the pMLST web page (http://pubmlst.org). PlasmidFinder is limited by its inability to reconstruct the sequences of entire plasmids, underscoring the urgent need for improvement over existing tools.

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), protectins 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 sub-clones 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 (*H*30) is characterized by the presence of a *fimH*30 allele, serotype O25:H4 and a specifically conserved *gyrA/parC* allele combination that confers fluoroquinolone resistance (FQ-R). Most human infections caused by ST131 are due to isolates of the *H*30 sublineage [13,16,37–39], many of them carrying the $bla_{CTX-M-15}$ 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 *H*30-R and *H*30-Rx sublineages, respectively [13,35,37,38]. Currently, diverse O25b:H4 ST131 variants (e.g. *fimH*22, *fimH*30) or O16:H5 (e.g. *fimH*41) seem also to be widely spread [13-16,40]. Full genome sequencing of several ST131 *E. coli* genomes, most of them *H*30-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) 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 ST506 (Fig. 1) according to the cgMLST Pasteur Institute scheme (http://www.pasteur.fr/recherche/genopole/PF8/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 *H*30 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 *H*22/*H*324 sublineage (virotype D). The most distal branch to the main cluster is represented by the commensal strain SE15, a member of sublineage *H*41 identified as ST506 [16]. It does not contain any marker used for the virotype subtyping method described by Blanco et al (*afa, sat, ibeA, iroN*) [36,51,52].

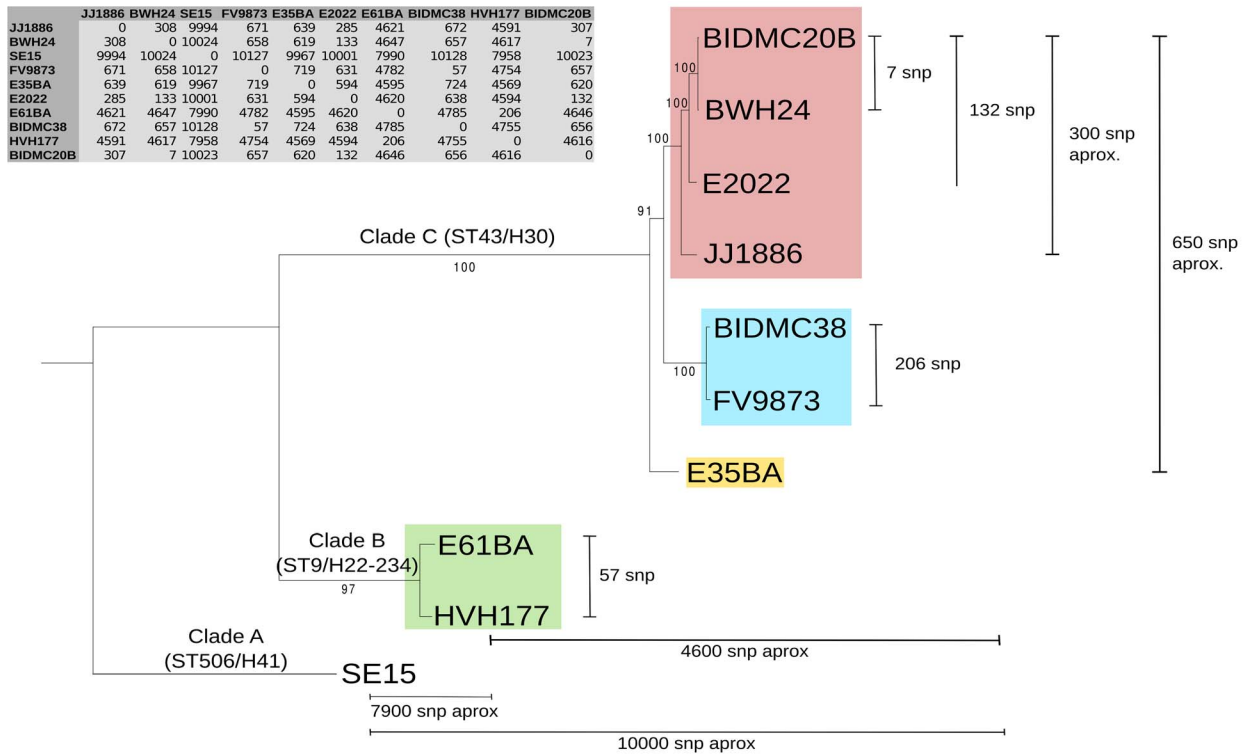|	JJ1886	BWH24	SE15	FV9873	E35BA	E2022	E61BA	BIDMC38	HVH177	BIDMC20B
JJ1886	0	308	9994	671	639	285	4621	672	4591	307
BWH24	308	0	10024	658	619	133	4647	657	4617	7
SE15	9994	10024	0	10127	9967	10001	7990	10128	7958	10023
FV9873	671	658	10127	0	719	631	4782	57	4754	657
E35BA	639	619	9967	719	0	594	4595	724	4569	620
E2022	285	133	10001	631	594	0	4620	638	4594	132
E61BA	4621	4647	7990	4782	4595	4620	0	4785	206	4646
BIDMC38	672	657	10128	57	724	638	4785	0	4755	656
HVH177	4591	4617	7958	4754	4569	4594	206	4755	0	4616
BIDMC20B	307	7	10023	657	620	132	4646	656	4616	0

Fig. 1. Phylogenetic tree of ST131 *E. coli*. The tree is based on a 3,629,034 bp core genome (3,734 orthologous genes: 90% identity and 90% coverage) and 100 bootstrapping replicates. ST131 clades are named according to [16] and further subdivided and colored according to virotypes [36]: virotype A (blue), virotype B (yellow), virotype C (pink), virotype D (green). Virotype classification is based on the presence/absence of four putative virulence factors: *afaFM955459* (encoding an Afa/Dr adhesin), *sat* (secreted autotransporter toxin, present in PAI-CFT073-pheV), *ibeA* (invasion of brain endothelium) and *iroN* (salmochelin siderophore receptor). The commensal ST131 strain SE15 was used to root the tree (virotype non typable; serotype O150 in the original publication [91] but lying within the H41 cluster in the phylogenomic study of [16]). Given SNP numbers are approximate averages of individual comparisons.
doi:10.1371/journal.pgen.1004766.g001

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 (<0.3% divergence in the core genome). There are only 650 SNPs among the genomes of cluster C lineage (i.e., <200 SNP/Mb), indicating their close phylogenetic relationship. There are <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 MOB_{F12}/IncF plasmid (pE61BA-1), a 37.7 kb MOB_{P6}/IncI2 plasmid (pE61BA-7), a 24.5 kb MOB_{C12} plasmid (pE61BA-2), a 18 kb MOB_{P11}/IncP1 plasmid (pE61BA-4), two MOB_{P5}/ColE1-like plasmids of 6.6 and 6.9 kb (pE61BA-5 and pE61BA-6, respectively) and one MOB_{Q12} 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).
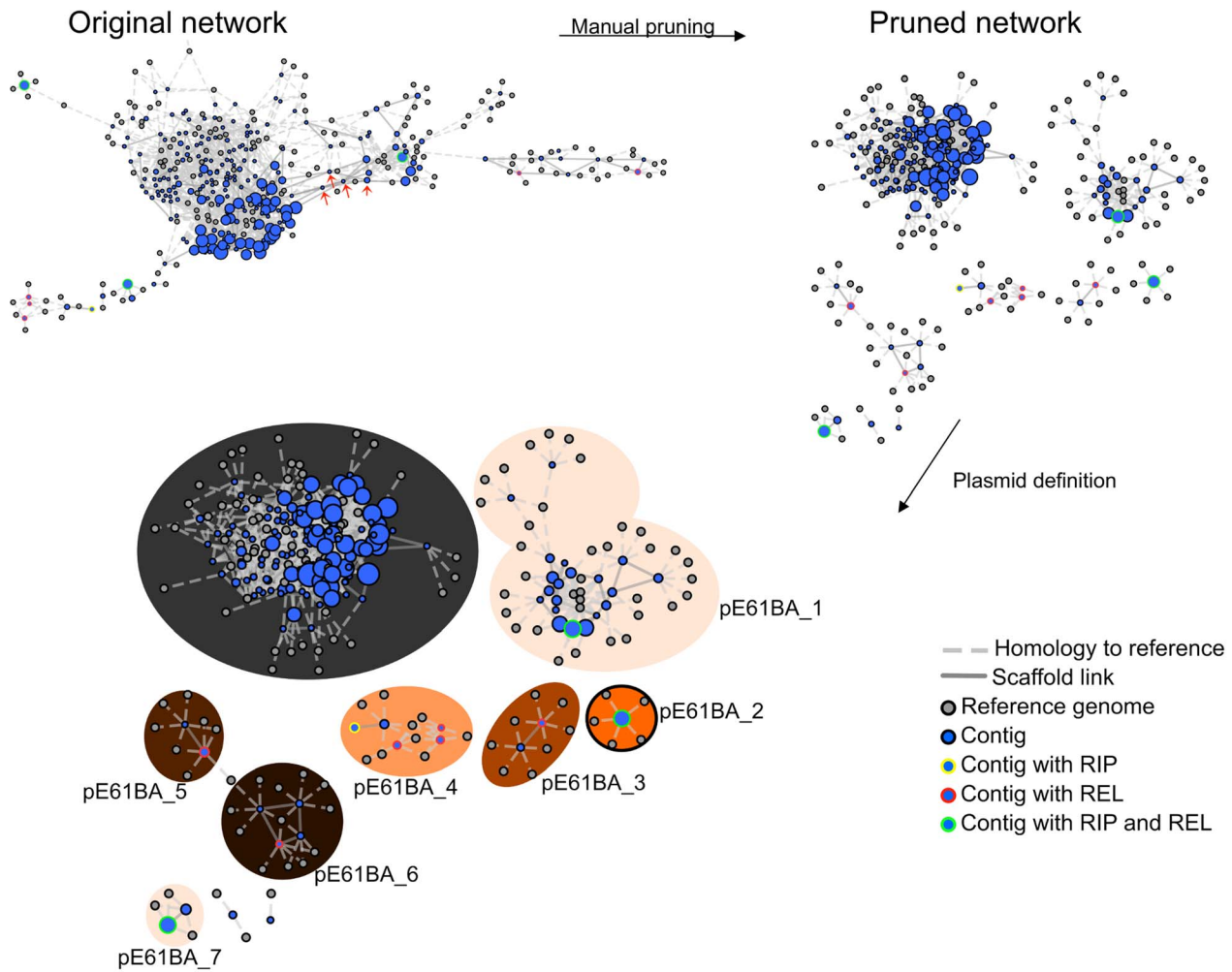
Fig. 2. PLACNET plasmid reconstruction of ST131 genome E61BA (ST9/H324/virotype D). The network contains nodes of two different colors (blue for contigs, grey for reference genomes). The size of reference nodes is always the same. The size of the contig nodes is proportional to the contig length. Besides, outlines are yellow for contigs containing RIP proteins, red for relaxases and green for both proteins. Edges are either solid (scaffold links) of dotted (homologous references). The length of the edges is arbitrarily selected by Cytoscape algorithm. In the upper left, the network output (original network) is shown, which resulted from automatic reference search, scaffold links and protein tagging rules. The original network was converted to a pruned network by eliminating contigs smaller than 200 bp and duplicating specific hubs (red arrows). Two contigs could not be assigned for lack of scaffold links: a 2,953 bp contig (putative DNA primase + lytic transglycosylase) and a 1,301 bp contig (TrbI + TraB-partial). Closed plasmids (e.g., pE61BA_2, size: 24,447 bp) are shown with a black outline in the final PLACNET network.
doi:10.1371/journal.pgen.1004766.g002

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 MOB_{F12}/IncF plasmids (11 plasmids), present in all ten sequenced ST131 genomes. Other relevant plasmid backbones belong to the MOB_P (RIP groups IncI1/K, IncI2, IncX1, IncX4, ColE1), MOB_Q (Qu, Q12) and MOB_C (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 Rep-3 RIP family. Finally, 5 plasmids corresponded to the no-MOB category. The E35BA genome (ST43/H30 virotype B) showed a MOB_{P11} relaxase within a 234 kb chromosomal contig, implying the

Table 1. Summary of plasmid content.

Genome	fimH allele	Strain virotype	MOB_F12/IncF	Phage-related/ MOB_P12/ RepFIB	MOB_P12/ IncI	MOB_P6/ IncI2	MOB_F11/IncN	MOB_P3/IncX	MOB_P11/IncP1MOB_C12 [d]	MOB_P5/ ColE1-like	MOB_Qu [d]	MOB_Q12 [d]	no-MOB [d]
FV9873	H30	A	pFV9873_5 (91.4 Kb; ΔTraI)					pFV9873_4 (33.3 Kb)			pFV9873_1 (4.1 Kb)	pFV9873_6 (5.2 Kb)	pFV9873_2 (2.2 Kb); pFV9873_3 (4.6 Kb)
BIDMC38	H30	A	pBIDMC38_5 (123 Kb)							pBIDMC38_1 (11.8 Kb)	pBIDMC38_4 (4.2 Kb)	pBIDMC38_2 (5.3 Kb)	pBIDMC38_3 (1.6 Kb)
E35BA	H30	B	p35BA_2+3 (211 Kb)				IME_E35BA (14.2 Kb)				pE35BA_1 (4.1 Kb)		
E2022	H30	C	pE2022_2 (103 Kb)		pE2022_1 (98.3 Kb)			pE2022_3 (35.0 Kb)			pE2022_4 (4.1 Kb)		pE2022_5 (2.2 Kb)
BIDMC20B	H30	C	pBIDMC20B_1 (128 Kb, ΔTraI [b])	pBIDMC20B_2 (109 Kb)									
BWH24	H30	C	pBWH24_1 (123 Kb, ΔTraI [b])	pBWH24_2 (109 Kb)		pBWH24_3 (60.3 Kb)							
JJ1886	H30	C	pJJ1886-5 (110 Kb ΔTraI)				pJJ1886-4 (55.9 Kb)			pJJ1886-3 (5.6 Kb)		pJJ1886-2 (5.2 Kb)	pJJ1886-1 (1.6 Kb)
E61BA	H324	D	pE61BA_1 (137 Kb)			pE61BA_7 (37.9 Kb)	pE61BA_4 (18.3 Kb)		pE61BA_2 (24.5 Kb)	pE61BA_5 (6.5 Kb); pE61BA_6 (6.9 Kb)		pE61BA_3 (5.5 Kb)	
HVH177	H22	D	pHVH177_1 (78.6 Kb)										
SE15	H41	Commensal	pECSF1 (122 Kb)										
Other ST131 plasmids [a]			pEK516 (64.5 Kb, ΔTraI); pEK499 (117 Kb, ΔTraI); pJIE186-2 [c] (138 Kb); pGUE-NDM (87.0 Kb)		pEK204 (93.7 Kb)		pKC394 (53.2 Kb); pKC396 (44.2 Kb); pNDM-ECS01 (41.2 Kb); pECN580 (64.9 Kb)	pJIE143 (34.3 Kb)					

[a] Plasmid references: pEK516 ([93]; EU935738); pEK499 ([93]; EU935739); pEK204 ([93]; EU935740); pJIE186-2 ([56]; NC_020271); pGUE-NDM (in [119]; JQ364967); pEK204 ([93]; EU935740); pKC394 ([120]; HM138652); pKC396 ([120], HM138653); pNDM-ECS01 (Unpublished; KJ413946); pECN580 ([121]; KF914891); pJIE143 ([94]; JN194214).
[b] pBIDMC20B_1 and pBWH24_1 plasmids lacked the REL domain of the TrwC protein.
[c] Plasmid pJIE186-2 was isolated from strain JIE186 [94], although GenBank acc. n° NC_020271 specifies it is located at EC958 strain. Strain JIE186 also contains plasmid pJIE186-1, 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
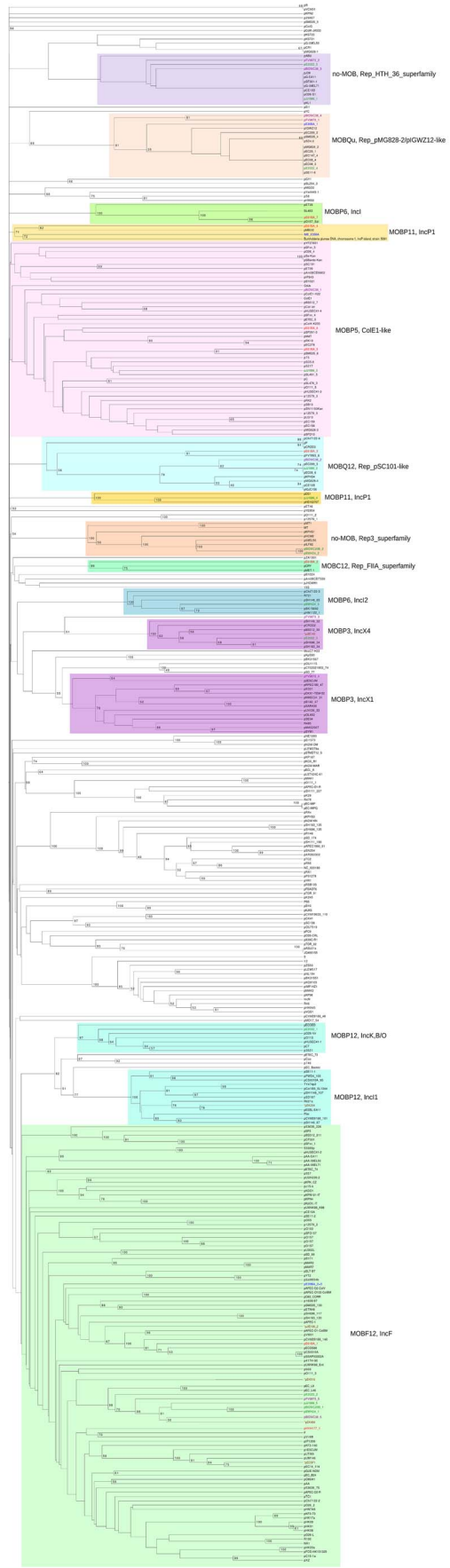
Clade C (ST43/H30), virotype A
Clade C (ST43/H30), virotype B
Clade C (ST43/H30), virotype C
Clade B (ST9/H22-234), virotype D
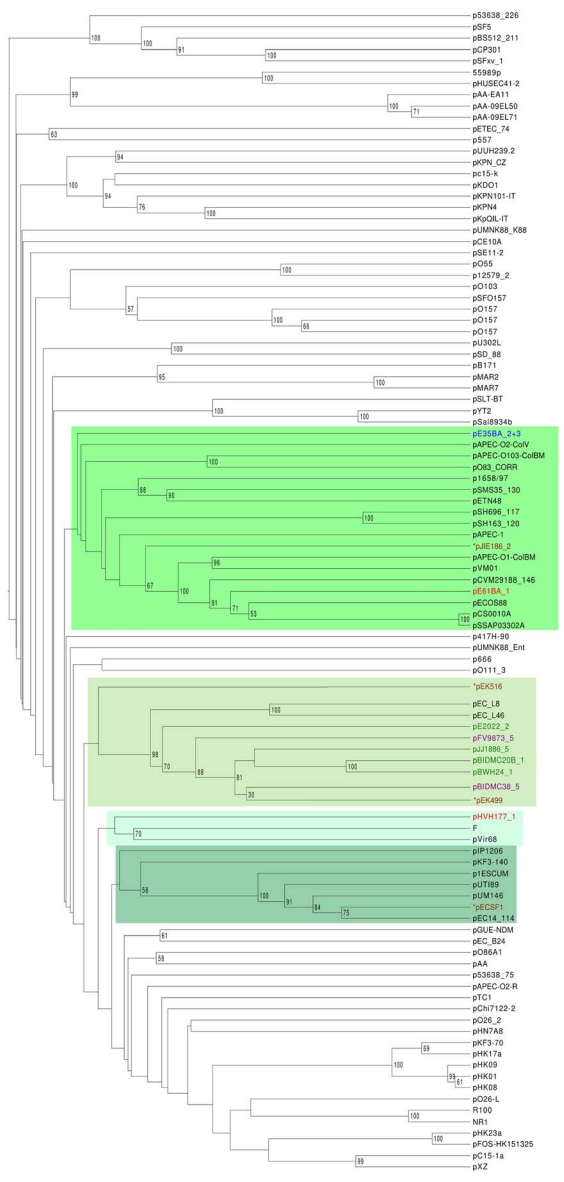* Virotype not determined

MOBF12, IncF

Fig. 4. 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 MOB_{F12}/IncF branch, with new background colors highlighting plasmid groups within this branch that are mentioned in the text.
doi:10.1371/journal.pgen.1004766.g004

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 MOB_{F12}/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 MOB_{F12}/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 MOB_{F12}/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 pHVH177_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 pJIE186-2, 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, iutA, cvaBC* and *sitC* and the cassette *ompT-hlyF-mig14*, eventually also linked to *estABCDE* [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 ST43/*fimH*30 lineage from USA. 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, pEK499 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 pHVH177_1 is not similar to any reference outside the backbone genes (S11C Fig.), the group IV plasmid pECSF1 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 MOB_{F12}/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 ST43/*H*30 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-typical 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 pLF82, isolated from a patient with Crohn's disease [58], the STEC plasmid p09EL50 [5], the *Salmonella* plasmid pHCM2 [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 MOB_{P12}/IncK plasmid pE2022_1 is most similar to pCT [61]. pE2022_1 contains a $bla_{CTX-M-14}$ gene identical to that in pCT, a plasmid carrying $bla_{CTX-M-14}$ that is globally spread among humans and animals and particularly prevalent in clinical
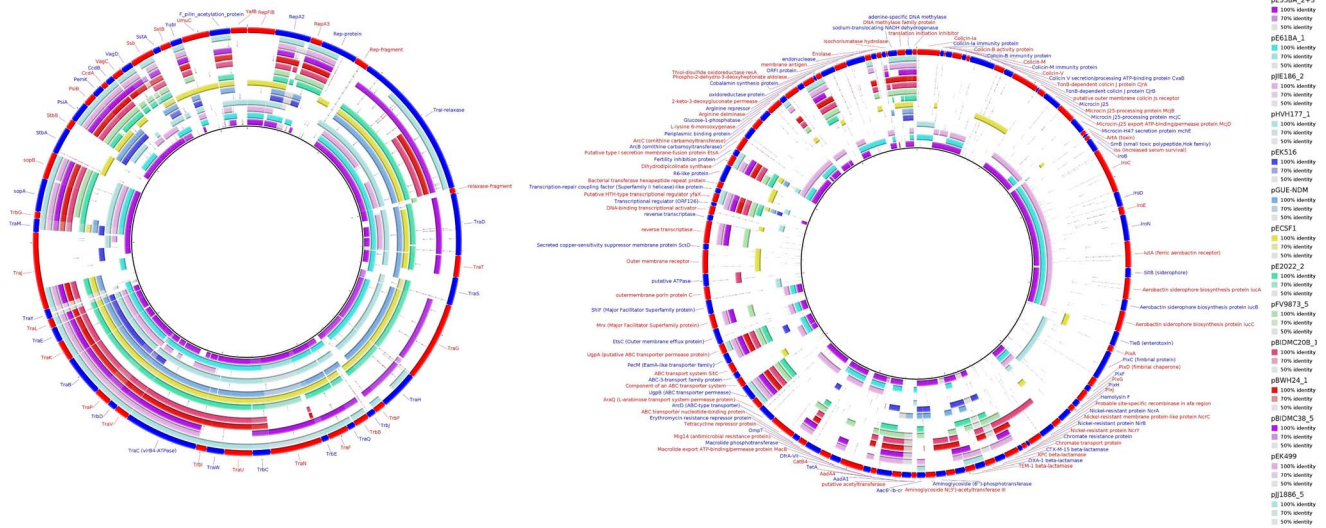
Fig. 5. MOB$_{F12}$/IncF plasmid analysis. Protein cluster analysis was performed with kClust software (parameters: 30% identity, 50% coverage) on the set of 14 plasmids shown in Table 4. Plasmid pGUE-NDM [119] was excluded from this comparison since it is only distantly related to the others (see dendrogram 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

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).

MOB$_{P6}$ is a large plasmid family, as can be observed in the phylogenetic tree of the MOB$_{P6}$ relaxase family (S14A Fig.). The two ST131 MOB$_{P6}$/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 agona* plasmid SL483 and the enterohemorrhagic *E. coli* (EHEC) plasmid pO157_Sal [65]. They were recovered from isolates identified as *H*30_virotype C from the USA and *H*324_virotype D from Spain, respectively.

Another ST131 important group is MOB$_{P3}$/IncX, composed by three ST131 plasmids (pFV9873_4 and pE2022_3, as well as the reference pJIE143). Plasmids pFV9873_4 and pE2022_3, obtained from different *H*30 subgroups, are rather different between them and belong to different plasmid groups (IncX1 and IncX4), as shown their relatively distant positions in the MOB$_{P3}$ 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_34 and the ST131 reference plasmid pJIE143 all over its sequence length (S15C Fig.).

Two plasmids and the IME belong to the MOB$_{P11}$/IncP1 family, as shown in the MOB$_{P11}$ 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 (GeneBank Acc. N° KF701335). These two plasmids thus represent new additions to the ST131 plasmidome (see S16B Fig.). The IME_E35BA is a 14.2 kb insertion within a 234 kb chromosomal contig. S16C Fig. shows some detail on the genetic structure of the IME and its insertion site in the ST131 core genome.

The last potentially conjugative plasmid is the 24.5 kb MOB$_{C12}$ 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_FIIA 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 MOB$_{P5}$/ColE1-like plasmids (11.8, 6.9, 6.6 and 5.6 kb). Three of the plasmids are relatively different, as judged by the MOB$_{P5}$ phylogenetic tree of S18A Fig. The large plasmid (pBIDMC38_1) contains a type II restriction-modification system (Cfr10I) and is almost identical to the ST131 reference pJJ1886_3 (S18B Fig.). The two MOB$_{P5}$/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 MOB$_{Qu}$ plasmids of around 4.1 kb (S19A Fig.), which populate all *H*30 subgroups (two of virotype A, one of virotype B and one of virotype C). Nothing remarkable could be distinguished it their genetic constitution, besides a common MOB region and a pIGWZ12 -like Rep protein

Table 2. Resistance genes and virulence determinants in MOB$_{F12}$ plasmids.

Plasmid	Strain Virotype	Size (Kb)	RIP (FAB formula)[a]	Antibiotic resistance genes[b]	Virulence genes[c]
pFV9873_5	A	91.4	RepFIIA-RepFIA (F2:A1:B-)	bla$_{CTX-M-15}$, bla$_{OXA-1}$, sul1, aadA5, aac(6')-lb-cr, dfrA17, catB4, tet(A), mphA	None-detected
pBIDMC38_5	A	123	RepFIIA-RepFIA (F2:A1:B-)	bla$_{TEM-1}$, sul1, aadA5, tet(A), mphA	traT, finO
pE2022_2	C	103	RepFIIA-RepFIA (F2:A1:B-)	bla$_{CTX-M-15}$, bla$_{TEM-1}$, bla$_{OXA-1}$, aac(6')-lb-cr-like[d], catB4, tet(A)	traT, finO
pBIDMC20B_1	C	128	RepFIIA-RepFIA (F2:A1:B-)	bla$_{KPC-3}$, sul1, sul2, strA, strB, aadA5, dfrA17, tet(A)	finO
pBWH24_1	C	123	RepFIIA-RepFIA (F2:A1:B-)	bla$_{KPC-3}$, sul1, sul2, strA, strB, aadA5, dfrA17, tet(A)	finO
pJJ1886-5	C	110	RepFIIA-RepFIA (F2:A1:B-)	bla$_{TEM-1}$, bla$_{OXA-1}$, aac(6')-lb-cr	None-detected
pEK499	-	117	RepFIIA-RepFIA (F2:A1:B-)	bla$_{CTX-M-15}$, bla$_{TEM-1}$, bla$_{OXA-1}$, sul1, aadA5, aac(6')-lb-cr, dfrA17[e], tet(A), mphA	None-detected
pEK516	-	64.5	RepFIIA (F2:A-:B-)	bla$_{CTX-M-15}$, bla$_{TEM-1}$, bla$_{OXA-1}$, aac(6')-lb-cr, aac(3')-IIa, catB4, tet(A)	None-detected
pGUE-NDM	-	87.0	RepFIIA (F2:A-:B-)	bla$_{NDM-1}$, bla$_{OXA-1}$, ble$_{MBL}$, sul1, aac(6')-lb-cr, aac(3')-II, aadA2, dfrA12, ΔcatB4	traT
pE61BA_1	D	137	RepFIIA-RepFIB (F2:A-:B1)	bla$_{TEM-1}$, tet(A)	ompT, iss, iroBCDEN, iucABCD, cvaBC, traT, etsABC, mig14-hlyF-finO
pECSF1	Commensal	122	RepFIIA-RepFIB (F29:A-:B10)	None-detected	traT, finO
pE35BA_2+3	B	211	RepFIA-RepFIB (F-:A2:B1)	bla$_{CTX-M-15}$, bla$_{TEM-1}$, bla$_{OXA-1}$, sul2, strA, strB, aac(6')-lb-cr, aac(3')-IIa, dfrA14, catB4, tet(A)	ompT, iss, iroBCDEN, iucABCD, iutA, cvaBC, sitC, traT, etsABC, mig14-hlyF-finO
pJIE186-2	-	138	RepFIB (F-:A-:B1)	None-detected	ompT, iss, iroBCDEN, iucABCD, iutA, cvaBC, etsABC, mig14-hlyF-finO
pHVH177_1	D	78.6	RepFIB (F-:A-:B31)	None-detected	traT, etsABC, mig14-hlyF-finO

[a]FAB formula according to http://pubmlst.org/plasmid/classification scheme [9].
[b]According to the ARG-annot database (>90% amino acid identity) [http://en.mediterranee-infection.com].
[c]According to our in-house database (>90% amino acid identity).
[d]aac(6')-lb-cr-like presents the Glu72Gly additional mutation.
[e]In 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 MOB$_{Q12}$ plasmids (around 5.2 kb) are also represented in H30 (two of virotype A, one of virotype C) and H22 (one of virotype D). They contain RIP and REL proteins but, as in the case of the MOB$_{Qu}$ plasmids, no phenotype could be pointed out (S20 Fig.). MOB$_{Qu}$ and MOB$_{Q12}$ 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 (pFV9873_3) was unique and unrelated to any reference. A detailed analysis of MOBF11/IncN plasmid family is detailed in S22 Fig.

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 constellations in a starry sky, allows user-friendly apprehension of that genome constitution. We applied PLACNET in this work to analyze the plasmidome of *E. coli* ST131 genomes, but it has been shown to work also for a series of prototypic bacterial genera with different GC content and genome architecture, such as *Salmonella*, *Klebsiella*, *Agrobacterium*, *Staphylococcus* or *Bacillus*. As an example, the PLACNET representation of the genome of *Staphylococcus aureus* strain 118 (ST772) (GenBank acc number AJGE00000000) is shown in S23 Fig. PLACNET scope of

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 (IncI2), P11 (IncP), P12 (IncI/K/BO), Q12 (Rep_pSC101-like), Qu (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 pathovars [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/*H*30 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 *H*30 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 $bla_{CTX-M-15}$, mainly associated with F2 plasmids, is considered a key event in the selection of the ST131 cluster C/*H*30 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 bla_{KPC} of bla_{NDM} genes predominant in *Klebsiella* (http://www.cdc.gov/drugresistance/threat-report-2013/) [81]. Beyond F2 plasmids, the presence of other plasmid groups (N, I1/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, MOB$_Q$, 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 MOB$_{F12}$/IncF, MOB$_H$/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 [16]. 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 Been *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 virulencegenes, 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 background 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 *fimH*30, one *fimH*324), USA (three *fimH*30), Australia (one *fimH*30), Denmark (one *fimH*22) and Japan (one *fimH*41). The *fimH*30 strains from Spain were CTX-M-15 producers and belonged to the *H*30-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; *ibeA*, 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

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

Strain	Accession[a]	Location	Collection date	Isolation source	Plasmid name (Accession number)[b]	Reference
HVH-177	PRJNA186205	Denmark	2003	Blood	pHVH177_1	PRJNA186413
BIDMC20B	PRJNA202031	USA	-	Urine	pBIDMC20B_1and _2	PRJNA202876
BWH24	PRJNA201983	USA	-	-	pBWH24_1 to _3	PRJNA202876
BIDMC38	PRJNA202050	USA	2012	-	pBIDMC38_1 to _5	PRJNA202876
FV9873	PRJEB6262	Spain	2007	Urine	pFV9873_1 to _6	This study
E35BA	PRJEB6262	Spain	2008	Urine	pE35BA_1 to _3, IME_E35BA	This study
E2022	PRJEB6262	Spain	2006	Urine	pE2022_1 to _5	This study
E61BA	PRJEB6262	Spain	2008	Abscess	pE61BA_1 to _7	This study
SE15	AP009378	Japan	-	Feces	pECSF1 (AP009379)	[91]
JJ1886	NC_022648.1	USA	-	Urine	pJJ1886_1 to _5 (NC_022649; NC_022650; NC_022651; NC_022661; NC_022662)	[90]

[a]PRJNA and PRJEB6262 accession numbers correspond to SRA datasets. AP009378 and NC_022648.1 correspond to finished genomes.
[b]Plasmids derived from this study are named according to Table 1.
doi:10.1371/journal.pgen.1004766.t003

these "virotype" designations and "ST131 strain designation" in other studies that also used capital letters to distinguish among ST131 clonal variants [16]. Other genome datasets were taken either from Bioproject NCBI database (https://www.ncbi.nlm.nih.gov/bioproject/), or from NCBI genomes database (*E. coli* JJ1886 [90] and *E. coli* SE15 [91]). In addition, fully sequenced plasmids pEK499, pEK204, pEK516, pJIE186-2 and pJIE143, previously found in other ST131 isolates [56,92–94], were used for plasmid comparisons. Information about all genomes is detailed in Table 3.

DNA sequencing

Total DNA from *E. coli* ST131 strains FV9873, E35BA, E2022 and E61BA 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 to reference sequences 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 4. Genomes assembled in this study.

Strain	ID	N° Libraries	Read length	N°Contigs	Total bp	N50	Kmer
HVH177	SRS399685	2	101	81	5035548	242711	83
BIDMC20B	SRS420795	3	101	115	5311918	209342	91
BWH24	SRS420803	2	101	114	5369063	192138	83
BIDMC38	SRS420798	4	101	135	5226831	190988	91
FV9873	ERS450218	3	71	262	5160060	153685	55
E35BA	ERS450219	3	71	419	5243070	159702	57
E2022	ERS450220	2	71	346	5296607	159635	53
E61BA	ERS450221	3	71	246	5168482	198396	57

doi:10.1371/journal.pgen.1004766.t004

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 $n+1$:

$$T_{n+1} = \frac{\sum_{i=1}^{n} S_i}{n} \cdot (0.85 + 2n)$$

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
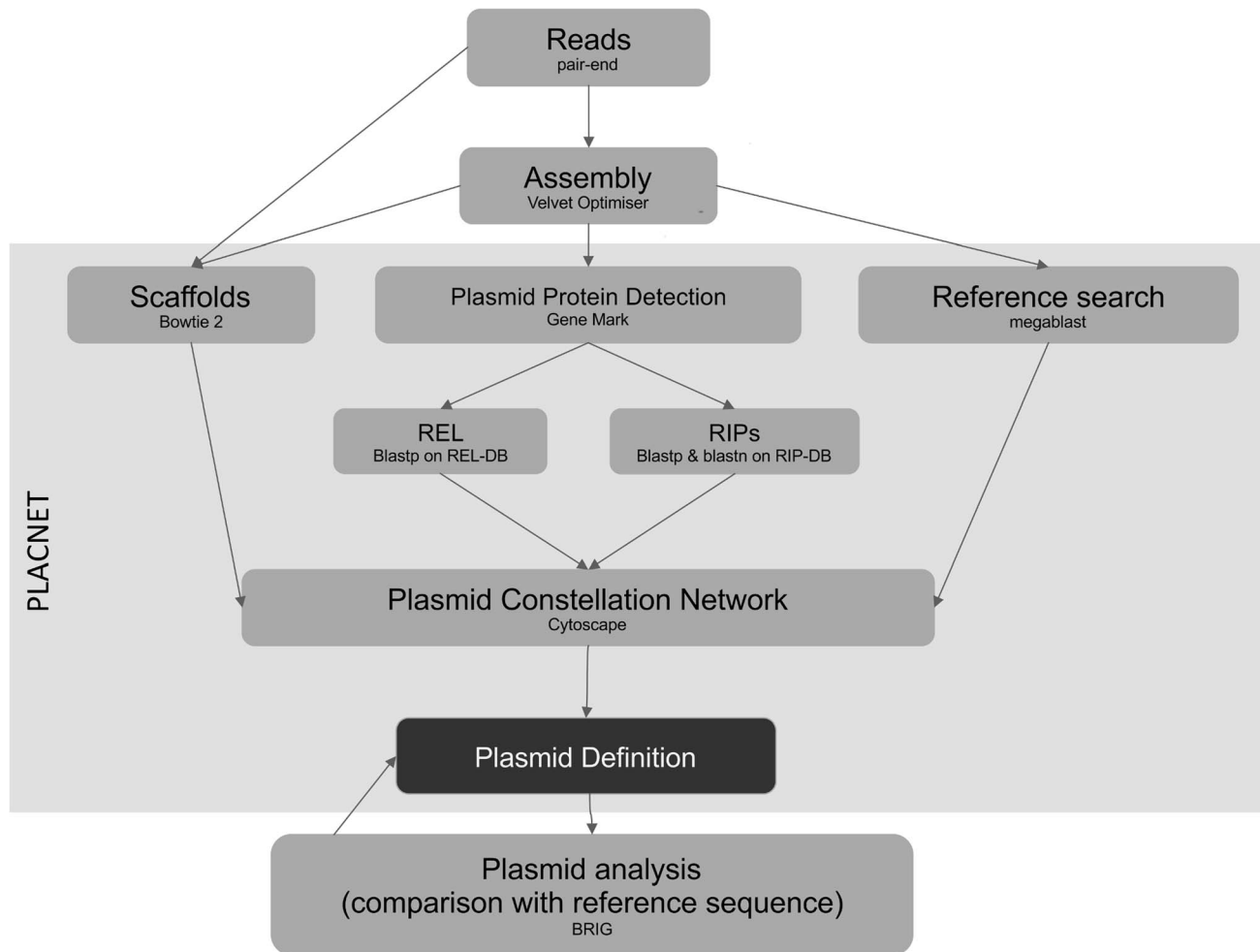
Fig. 6. PLACNET flow diagram. The diagram represents the PLACNET workflow to analyze an Illumina bacterial genome dataset. It can be separated in two sub-process: network delineation and plasmid analysis. Network delineation consists on contig assembly, determination of scaffold interactions, reference 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

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

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/*H*30/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/*H*30/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/*H*30/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/*H*324/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 transglycosilase, while another (1,301 bp) encodes for TrbI and TraB partial proteins.
(PDF)

S5 Fig PLACNET reconstruction for BIDMC20B genome (*E. coli* ST131/*H*30/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/*H*30/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/*H*30/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/*H*324/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/*H*324/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/*H*30/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_2+3 (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 pJIE186-2 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 111kb plasmid

pECOH89 as reference. **S12B**: Phylogenetic analysis of RepFIB family of RIP proteins. RaxML software (v.7.2.8) was used to infer the Maximum Likelihood tree and MEGA5.2.2 to represent the result. Bootstrap values for 100 replicates are indicated. The tree was rooted with the RepFIB protein of the IncN plasmid N3.
(PDF)

S13 Fig BRIG comparative analysis of MOBP12/IncI-complex. **S13A**: The IncI1 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 pChi7122_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 p2ESCUM 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 IncP1 plasmids using pJJ1886_4 as a reference. **S16C**: Comparison of JJ1886 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 MOBQu plasmids. **S19A**: Phylogenetic tree of MOBQu REL proteins, calculated as in Fig SF12B. The tree was rooted with the MOBQu2 subfamily. ST131 plasmids are colored in red. Different color backgrounds are used to represent MOBQu1, where ST131 MOBQu plasmids are located, and MOBQu2 branches. **S19B**: BRIG comparative analysis of MOBQu 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_RSF1010 (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: PRJNA82607). 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 MOBH12 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 Cytoscape 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 blued 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 MOB_{F12}/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)

Author Contributions

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

References

1. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW (2012) Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 13: 601–612.

2. Johnson TJ, Nolan LK (2009) Pathogenomics of the virulence plasmids of Escherichia coli. Microbiol Mol Biol Rev 73: 750–774.

3. Carattoli A (2011) Plasmids in Gram negatives: molecular typing of resistance plasmids. Int J Med Microbiol 301: 654–658.

4. Carattoli A (2013) Plasmids and the spread of resistance. Int J Med Microbiol 303: 298–304.

5. Ahmed SA, Awosika J, Baldwin C, Bishop-Lilly KA, Biswas B, et al. (2012) Genomic comparison of Escherichia coli O104:H4 isolates from 2009 and 2011 reveals plasmid, and prophage heterogeneity, including shiga toxin encoding phage stx2. PLoS One 7: e48228.

6. Boerlin P, Chen S, Colbourne JK, Johnson R, De Grandis S, et al. (1998) Evolution of enterohemorrhagic Escherichia coli hemolysin plasmids and the locus for enterocyte effacement in shiga toxin-producing E. coli. Infect Immun 66: 2553–2561.

7. Couturier M, Bex F, Bergquist PL, Maas WK (1988) Identification and classification of bacterial plasmids. Microbiol Rev 52: 375–395.

8. Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, et al. (2005) Identification of plasmids by PCR-based replicon typing. J Microbiol Methods 63: 219–228.

9. Villa L, Garcia-Fernandez A, Fortini D, Carattoli A (2010) Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. J Antimicrob Chemother 65: 2518–2529.

10. Garcia-Fernandez A, Fortini D, Veldman K, Mevius D, Carattoli A (2009) Characterization of plasmids harbouring qnrS1, qnrB2 and qnrB19 genes in Salmonella. J Antimicrob Chemother 63: 274–281.

11. Alvarado A, Garcillan-Barcia MP, de la Cruz F (2012) A Degenerate Primer MOB Typing (DPMT) Method to Classify Gamma-Proteobacterial Plasmids in Clinical and Environmental Settings. PLoS One 7: e40438.

12. Edwards DJ, Holt KE (2013) Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microb Inform Exp 3: 2.

13. Price LB, Johnson JR, Aziz M, Clabots C, Johnston B, et al. (2013) The epidemic of extended-spectrum-beta-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. MBio 4: e00377–00313.

14. Johnson JR, Clermont O, Johnston B, Clabots C, Tchesnokova V, et al. (2014) Rapid and specific detection, molecular epidemiology, and experimental virulence of the O16 subgroup within Escherichia coli sequence type 131. J Clin Microbiol.

15. Blanc V, Leflon-Guibout V, Blanco J, Haenni M, Madec JY, et al. (2014) Prevalence of day-care centre children (France) with faecal CTX-M-producing Escherichia coli comprising O25b:H4 and O16:H5 ST131 strains. J Antimicrob Chemother.

16. Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, et al. (2014) Global dissemination of a multidrug resistant Escherichia coli clone. Proc Natl Acad Sci U S A.

17. Coque TM, Novais A, Carattoli A, Poirel L, Pitout J, et al. (2008) Dissemination of clonally related Escherichia coli strains expressing extended-spectrum beta-lactamase CTX-M-15. Emerg Infect Dis 14: 195–200.

18. Nicolas-Chanoine MH, Bertrand X, Madec JY (2014) Escherichia coli ST131, an Intriguing Clonal Group. Clin Microbiol Rev 27: 543–574.

19. Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, et al. (2008) Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. J Antimicrob Chemother 61: 273–281.

20. Leflon-Guibout V, Blanco J, Amaqdouf K, Mora A, Guize L, et al. (2008) Absence of CTX-M enzymes but high prevalence of clones, including clone ST131, among fecal Escherichia coli isolates from healthy subjects living in the area of Paris, France. J Clin Microbiol 46: 3900–3905.

21. Platell JL, Johnson JR, Cobbold RN, Trott DJ (2011) Multidrug-resistant extraintestinal pathogenic Escherichia coli of sequence type ST131 in animals and foods. Vet Microbiol 153: 99–108.

22. Albrechtova K, Dolejska M, Cizek A, Tausova D, Klimes J, et al. (2012) Dogs of nomadic pastoralists in northern Kenya are reservoirs of plasmid-mediated cephalosporin- and quinolone-resistant Escherichia coli, including pandemic clone B2-O25-ST131. Antimicrob Agents Chemother 56: 4013–4017.

23. Hernandez J, Bonnedahl J, Eliasson I, Wallensten A, Comstedt P, et al. (2010) Globally disseminated human pathogenic Escherichia coli of O25b-ST131 clone, harbouring blaCTX-M-15, found in Glaucous-winged gull at remote Commander Islands, Russia. Environ Microbiol Rep 2: 329–332.

24. Pallecchi L, Bartoloni A, Fiorelli C, Mantella A, Di Maggio T, et al. (2007) Rapid dissemination and diversity of CTX-M extended-spectrum beta-lactamase genes in commensal Escherichia coli isolates from healthy children from low-resource settings in Latin America. Antimicrob Agents Chemother 51: 2720–2725.

25. Mora A, Herrera A, Mamani R, Lopez C, Alonso MP, et al. (2010) Recent emergence of clonal group O25b:K1:H4-B2-ST131 ibeA strains among Escherichia coli poultry isolates, including CTX-M-9-producing strains, and comparison with clinical human isolates. Appl Environ Microbiol 76: 6991–6997.

26. Dhanji H, Murphy NM, Akhigbe C, Doumith M, Hope R, et al. (2011) Isolation of fluoroquinolone-resistant O25b:H4-ST131 Escherichia coli with CTX-M-14 extended-spectrum beta-lactamase from UK river water. J Antimicrob Chemother 66: 512–516.

27. Colomer-Lluch M, Mora A, Lopez C, Mamani R, Dahbi G, et al. (2013) Detection of quinolone-resistant Escherichia coli isolates belonging to clonal groups O25b:H4-B2-ST131 and O25b:H4-D-ST69 in raw sewage and river water in Barcelona, Spain. J Antimicrob Chemother 68: 758–765.

28. Gibreel TM, Dodgson AR, Cheesbrough J, Bolton FJ, Fox AJ, et al. (2012) High metabolic potential may contribute to the success of ST131 uropathogenic Escherichia coli. J Clin Microbiol 50: 3202–3207.

29. Novais A, Pires J, Ferreira H, Costa L, Montenegro C, et al. (2012) Characterization of globally spread Escherichia coli ST131 isolates (1991 to 2010). Antimicrob Agents Chemother 56: 3973–3976.

30. Blanco J, Mora A, Mamani R, Lopez C, Blanco M, et al. (2011) National survey of Escherichia coli causing extraintestinal infections reveals the spread of drug-resistant clonal groups O25b:H4-B2-ST131, O15:H1-D-ST393 and CGA-D-ST69 with high virulence gene content in Spain. J Antimicrob Chemother 66: 2011–2021.

31. Johnson JR, Kuskowski MA, Gajewski A, Sahm DF, Karlowsky JA (2004) Virulence characteristics and phylogenetic background of multidrug-resistant and antimicrobial-susceptible clinical isolates of Escherichia coli from across the United States, 2000-2001. J Infect Dis 190: 1739–1744.

32. Schubert S, Darlu P, Clermont O, Wieser A, Magistro G, et al. (2009) Role of intraspecies recombination in the spread of pathogenicity islands within the Escherichia coli species. PLoS Pathog 5: e1000257.

33. Wirth T, Falush D, Lan R, Colles F, Mensa P, et al. (2006) Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol 60: 1136–1151.

34. Johnson JR, Nicolas-Chanoine MH, DebRoy C, Castanheira M, Robicsek A, et al. (2012) Comparison of Escherichia coli ST131 pulsotypes, by epidemiologic traits, 1967–2009. Emerg Infect Dis 18: 598–607.

35. Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL, et al. (2013) Abrupt emergence of a single dominant multidrug-resistant strain of Escherichia coli. J Infect Dis 207: 919–928.

36. Blanco J, Mora A, Mamani R, Lopez C, Blanco M, et al. (2013) Four main virotypes among extended-spectrum-beta-lactamase-producing isolates of Escherichia coli O25b:H4-B2-ST131: bacterial, epidemiological, and clinical characteristics. J Clin Microbiol 51: 3358–3367.

37. Peirano G, Pitout JD (2014) Fluoroquinolone resistant Escherichia coli ST131 causing bloodstream infections in a centralized Canadian region: the rapid emergence of H30-Rx sublineage. Antimicrob Agents Chemother.

38. Colpan A, Johnston B, Porter S, Clabots C, Anway R, et al. (2013) Escherichia coli sequence type 131 (ST131) subclone H30 as an emergent multidrug-resistant pathogen among US veterans. Clin Infect Dis 57: 1256–1265.

39. Banerjee R, Robicsek A, Kuskowski MA, Porter S, Johnston BD, et al. (2013) Molecular epidemiology of Escherichia coli sequence type 131 and Its H30 and H30-Rx subclones among extended-spectrum-beta-lactamase-positive and -negative E. coli clinical isolates from the Chicago Region, 2007 to 2010. Antimicrob Agents Chemother 57: 6385–6388.

40. Dahbi G, Mora A, Lopez C, Alonso MP, Mamani R, et al. (2013) Emergence of new variants of ST131 clonal group among extraintestinal pathogenic Escherichia coli producing extended-spectrum beta-lactamases. Int J Antimicrob Agents 42: 347–351.

41. Avasthi TS, Kumar N, Baddam R, Hussain A, Nandanwar N, et al. (2011) Genome of multidrug-resistant uropathogenic Escherichia coli strain NA114 from India. J Bacteriol 193: 4272–4273.

42. Totsika M, Beatson SA, Sarkar S, Phan MD, Petty NK, et al. (2011) Insights into a Multidrug Resistant Escherichia coli Pathogen of the Globally Disseminated ST131 Lineage: Genome Analysis and Virulence Mechanisms. PLoS One 6: e26578.

43. Clark G, Paszkiewicz K, Hale J, Weston V, Constantinidou C, et al. (2012) Genomic analysis uncovers a phenotypically diverse but genetically homogeneous Escherichia coli ST131 clone circulating in unrelated urinary tract infections. J Antimicrob Chemother 67: 868–877.

44. Lavigne JP, Vergunst AC, Goret L, Sotto A, Combescure C, et al. (2012) Virulence potential and genomic mapping of the worldwide clone Escherichia coli ST131. PLoS One 7: e34294.

45. Kunne C, Billion A, Mshana SE, Schmiedel J, Domann E, et al. (2012) Complete sequences of plasmids from the hemolytic-uremic syndrome-associated Escherichia coli strain HUSEC41. J Bacteriol 194: 532–533.

46. Grad YH, Godfrey P, Cerqueira GC, Mariani-Kurkdjian P, Gouali M, et al. (2013) Comparative genomics of recent Shiga toxin-producing Escherichia coli O104:H4: short-term evolution of an emerging pathogen. MBio 4: e00452–00412.

47. Brunder W, Schmidt H, Frosch M, Karch H (1999) The large plasmids of Shiga-toxin-producing Escherichia coli (STEC) are highly variable genetic elements. Microbiology 145 (Pt 5): 1005–1014.

48. Pallen MJ, Wren BW (2007) Bacterial pathogenomics. Nature 449: 835–842.

49. Keen EC (2012) Paradigms of pathogenesis: targeting the mobile genetic elements of disease. Front Cell Infect Microbiol 2: 161.

50. Woodford N, Turton JF, Livermore DM (2011) Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. FEMS Microbiol Rev 35: 736–755.

51. Blanco M, Alonso MP, Nicolas-Chanoine MH, Dahbi G, Mora A, et al. (2009) Molecular epidemiology of Escherichia coli producing extended-spectrum {beta}-lactamases in Lugo (Spain): dissemination of clone O25b:H4-ST131 producing CTX-M-15. J Antimicrob Chemother 63: 1135–1141.

52. Coelho A, Mora A, Mamani R, Lopez C, Gonzalez-Lopez JJ, et al. (2011) Spread of Escherichia coli O25b:H4-B2-ST131 producing CTX-M-15 and SHV-12 with high virulence gene content in Barcelona (Spain). J Antimicrob Chemother 66: 517–526.

53. Garcillan-Barcia MP, Alvarado A, de la Cruz F (2011) Identification of bacterial plasmids based on mobility and plasmid population biology. FEMS Microbiol Rev 35: 936–956.

54. Hauser M, Mayer CE, Soding J (2013) kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics 14: 248.

55. Peigne C, Bidet P, Mahjoub-Messai F, Plainvert C, Barbe V, et al. (2009) The plasmid of Escherichia coli strain S88 (O45:K1:H7) that causes neonatal meningitis is closely related to avian pathogenic E. coli plasmids and is associated with high-level bacteremia in a neonatal rat meningitis model. Infect Immun 77: 2272–2284.

56. Zong Z (2013) Complete sequence of pJIE186-2, a plasmid carrying multiple virulence factors from a sequence type 131 Escherichia coli O25 strain. Antimicrob Agents Chemother 57: 597–600.

57. Falgenhauer L, Yao Y, Fritzenwanker M, Schmiedel J, Imirzalioglu C, et al. (2014) Complete Genome Sequence of Phage-Like Plasmid pECOH89, Encoding CTX-M-15. Genome Announc 2.

58. Miquel S, Peyretaillade E, Claret L, de Vallee A, Dossat C, et al. (2010) Complete genome sequence of Crohn's disease-associated adherent-invasive E. coli strain LF82. PLoS One 5.

59. Kidgell C, Pickard D, Wain J, James K, Diem Nga LT, et al. (2002) Characterisation and distribution of a cryptic Salmonella typhi plasmid pHCM2. Plasmid 47: 159–171.

60. Kim M, Kim S, Ryu S (2012) Complete genome sequence of bacteriophage SSU5 specific for Salmonella enterica serovar Typhimurium rough strains. J Virol 86: 10894.

61. Cottell JL, Webber MA, Coldham NG, Taylor DL, Cerdeno-Tarraga AM, et al. (2011) Complete sequence and molecular epidemiology of IncK epidemic plasmid encoding blaCTX-M-14. Emerg Infect Dis 17: 645–652.

62. Valverde A, Canton R, Garcillan-Barcia MP, Novais A, Galan JC, et al. (2009) Spread of bla(CTX-M-14) is driven mainly by IncK plasmids disseminated among Escherichia coli phylogroups A, B1, and D in Spain. Antimicrob Agents Chemother 53: 5204–5212.

63. Kim SR, Komano T (1992) Nucleotide sequence of the R721 shufflon. J Bacteriol 174: 7053–7058.

64. Mellata M, Maddux JT, Nam T, Thomson N, Hauser H, et al. (2012) New insights into the bacterial fitness-associated mechanisms revealed by the characterization of large plasmids of an avian pathogenic E. coli. PLoS One 7: e29481.

65. Wang P, Xiong Y, Lan R, Ye C, Wang H, et al. (2011) pO157_Sal, a novel conjugative plasmid detected in outbreak isolates of Escherichia coli O157:H7. J Clin Microbiol 49: 1594–1597.

66. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, et al. (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 5: e1000344.

67. Brown CJ, Sen D, Yano H, Bauer ML, Rogers LM, et al. (2013) Diverse broad-host-range plasmids from freshwater carry few accessory genes. Appl Environ Microbiol 79: 7684–7695.

68. Smajs D, Micenkova L, Smarda J, Vrba M, Sevcikova A, et al. (2010) Bacteriocin synthesis in uropathogenic and commensal Escherichia coli: colicin E1 is a potential virulence factor. BMC Microbiol 10: 288.

69. Baquero F, Coque TM (2011) Multilevel population genetics in antibiotic resistance. FEMS Microbiol Rev 35: 705–706.

70. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, et al. (2008) The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol 190: 6881–6893.

71. Hazen TH, Sahl JW, Fraser CM, Donnenberg MS, Scheutz F, et al. (2013) Refining the pathovar paradigm via phylogenomics of the attaching and effacing Escherichia coli. Proc Natl Acad Sci U S A 110: 12810–12815.

72. Skippington E, Ragan MA (2011) Lateral genetic transfer and the construction of genetic exchange communities. FEMS Microbiol Rev.

73. Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, et al. (2009) Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. Proc Natl Acad Sci U S A 106: 17939–17944.

74. Croucher NJ, Harris SR, Grad YH, Hanage WP (2013) Bacterial genomes in epidemiology—present and future. Philos Trans R Soc Lond B Biol Sci 368: 20120202.

75. Johnson TJ, Logue CM, Johnson JR, Kuskowski MA, Sherwood JS, et al. (2012) Associations Between Multidrug Resistance, Plasmid Content, and Virulence Potential Among Extraintestinal Pathogenic and Commensal Escherichia coli from Humans and Poultry. Foodborne Pathog Dis 9: 37–46.

76. Polz MF, Alm EJ, Hanage WP (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet 29: 170–175.

77. Reeves PR, Liu B, Zhou Z, Li D, Guo D, et al. (2011) Rates of mutation and host transmission for an Escherichia coli clone over 3 years. PLoS One 6: e26907.

78. Sandegren L, Linkevicius M, Lytsy B, Melhus A, Andersson DI (2012) Transfer of an Escherichia coli ST131 multiresistance cassette has created a Klebsiella pneumoniae-specific plasmid associated with a major nosocomial outbreak. J Antimicrob Chemother 67: 74–83.

79. Chen L, Hu H, Chavda KD, Zhao S, Liu R, et al. (2014) Complete Sequence of a KPC-Producing IncN Multidrug-Resistant Plasmid from an Epidemic Escherichia coli Sequence Type 131 Strain in China. Antimicrob Agents Chemother 58: 2422–2425.

80. Partridge SR, Zong Z, Iredell JR (2011) Recombination in IS26 and Tn2 in the evolution of multiresistance regions carrying blaCTX-M-15 on conjugative IncF plasmids from Escherichia coli. Antimicrob Agents Chemother 55: 4971–4978.

81. O'Hara JA, Hu F, Ahn C, Nelson J, Rivera JI, et al. (2014) Molecular Epidemiology of KPC-Producing Escherichia coli: Occurrence of ST131-fimH30 Subclone Harboring pKpQIL-like IncFIIk Plasmid. Antimicrob Agents Chemother.

82. Johnson TJ, Wannemuehler YM, Johnson SJ, Logue CM, White DG, et al. (2007) Plasmid replicon typing of commensal and pathogenic Escherichia coli isolates. Appl Environ Microbiol 73: 1976–1983.

83. Chen L, Chavda KD, Al Laham N, Melano RG, Jacobs MR, et al. (2013) Complete nucleotide sequence of a blaKPC-harboring IncI2 plasmid and its dissemination in New Jersey and New York hospitals. Antimicrob Agents Chemother 57: 5019–5025.

84. Johnson TJ, Lang KS (2012) IncA/C plasmids: An emerging threat to human and animal health. Mob Genet Elements 2: 55–58.

85. Kim YA, Qureshi ZA, Adams-Haduch JM, Park YS, Shutt KA, et al. (2012) Features of infections due to Klebsiella pneumoniae carbapenemase-producing Escherichia coli: emergence of sequence type 131. Clin Infect Dis 55: 224–231.

86. Muniesa M, Colomer-Lluch M, Jofre J (2013) Potential impact of environmental bacteriophages in spreading antibiotic resistance genes. Future Microbiol 8: 739–751.

87. Hoyland-Kroghsbo NM, Maerkedahl RB, Svenningsen SL (2013) A quorum-sensing-induced bacteriophage defense mechanism. MBio 4: e00362–00312.

88. Schimke RT (1984) Gene amplification in cultured animal cells. Cell 37: 705–713.

89. Baquero F, Tedim AP, Coque TM (2013) Antibiotic resistance shaping multi-level population biology of bacteria. Front Microbiol 4: 15.

90. Andersen PS, Stegger M, Aziz M, Contente-Cuomo T, Gibbons HS, et al. (2013) Complete Genome Sequence of the Epidemic and Highly Virulent CTX-M-15-Producing H30-Rx Subclone of Escherichia coli ST131. Genome Announc 1.

91. Toh H, Oshima K, Toyoda A, Ogura Y, Ooka T, et al. (2010) Complete genome sequence of the wild-type commensal Escherichia coli strain SE15, belonging to phylogenetic group B2. J Bacteriol 192: 1165–1166.

92. Boyd DA, Tyler S, Christianson S, McGeer A, Muller MP, et al. (2004) Complete nucleotide sequence of a 92-kilobase plasmid harboring the CTX-M-15 extended-spectrum beta-lactamase involved in an outbreak in long-term-care facilities in Toronto, Canada. Antimicrob Agents Chemother 48: 3758–3764.

93. Woodford N, Carattoli A, Karisik E, Underwood A, Ellington MJ, et al. (2009) Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major Escherichia coli lineages from the United Kingdom, all belonging to the international O25:H4-ST131 clone. Antimicrob Agents Chemother 53: 4472–4482.

94. Partridge SR, Ellem JA, Tetu SG, Zong Z, Paulsen IT, et al. (2011) Complete sequence of pJIE143, a pir-type plasmid carrying ISEcp1-blaCTX-M-15 from an Escherichia coli ST131 isolate. Antimicrob Agents Chemother 55: 5933–5935.

95. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.

96. Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5: e11147.

97. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

98. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.

99. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829.

100. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.

101. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

102. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

103. Garcillan-Barcia MP, de la Cruz F (2013) Ordering the bestiary of genetic elements transmissible by conjugation. Mob Genet Elements 3: e24263.

104. del Solar G, Giraldo R, Ruiz-Echevarria MJ, Espinosa M, Diaz-Orejas R (1998) Replication and control of circular bacterial plasmids. Microbiol Mol Biol Rev 62: 434–464.

105. Zheng J, Peng D, Ruan L, Sun M (2013) Evolution and dynamics of megaplasmids with genome sizes larger than 100 kb in the Bacillus cereus group. BMC Evol Biol 13: 262.

106. Osborn AM, da Silva Tatley FM, Steyn LM, Pickup RW, Saunders JR (2000) Mosaic plasmids and mosaic replicons: evolutionary lessons from the analysis of genetic diversity in IncFII-related replicons. Microbiology 146 (Pt 9): 2267–2275.

107. Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EP, de la Cruz F (2010) Mobility of plasmids. Microbiol Mol Biol Rev 74: 434–452.

108. Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29: 2607–2618.

109. Guglielmini J, Quintais L, Garcillan-Barcia MP, de la Cruz F, Rocha EP (2011) The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. PLoS Genet 7: e1002222.

110. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7: e1002195.

111. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431–432.

112. Zhou Y, Call DR, Broschat SL (2013) Using protein clusters from whole proteomes to construct and augment a dendrogram. Adv Bioinformatics 2013: 191586.

113. Tekaia F, Lazcano A, Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. Genome Res 9: 550–557.

114. Tekaia F, Yeramian E (2005) Genome trees from conservation profiles. PLoS Comput Biol 1: e75.

115. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. (2013) vegan: Community Ecology Package.

116. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289–290.

117. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12: 402.

118. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M (2009) ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25: 1968–1969.

119. Bonnin RA, Poirel L, Carattoli A, Nordmann P (2012) Characterization of an IncFII plasmid encoding NDM-1 from Escherichia coli ST131. PLoS One 7: e34752.

120. Cullik A, Pfeifer Y, Prager R, von Baum H, Witte W (2010) A novel IS26 structure surrounds blaCTX-M genes in different plasmids from German clinical Escherichia coli isolates. J Med Microbiol 59: 580–587.

121. Chen L, Chavda KD, Melano RG, Jacobs MR, Koll B, et al. (2014) Comparative Genomic Analysis of KPC-Encoding pKpQIL-Like Plasmids and Their Distribution in New Jersey and New York Hospitals. Antimicrob Agents Chemother 58: 2871–2877.