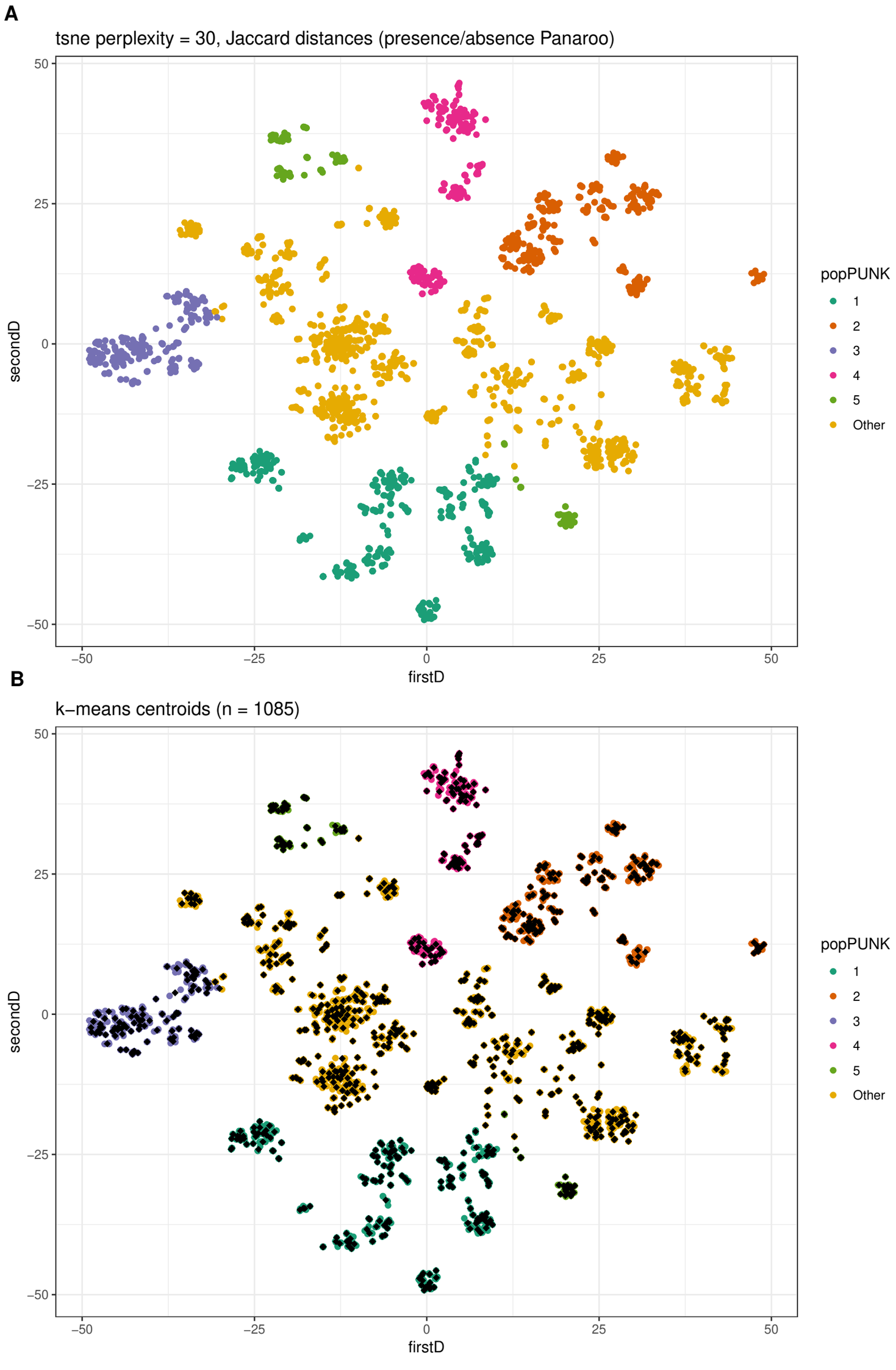

A
tsne perplexity = 30, Jaccard distances (presence/absence Panaroo)
secondD
firstD
50
25
0
−25
−50
popPUNK
1
2
3
4
5
Other
B
k−means centroids (n = 1085)
secondD
firstD
50
25
0
−25
−50
popPUNK
1
2
3
4
5
Other

Figure S1. Selection of 1,085 ExPEC isolates for long-read sequencing based on their accessory genome diversity. A) Panaroo was used to infer the pangenome of the isolates included in the NORM collection (3,245) using short-read sequencing data. To represent the accessory genome diversity inherent in the collection, the presence/absence matrix of genes was embedded into 2D using t-sne (perplexity = 30) and isolates were coloured according to their PopPUNK group [75]. B) To select 1,085 isolates for long-read sequencing, the k-means algorithm was used to locate 1,085 centroids in the 2D embedding. The isolate closest (Euclidean distance) to each centroid was selected for long-read sequencing.

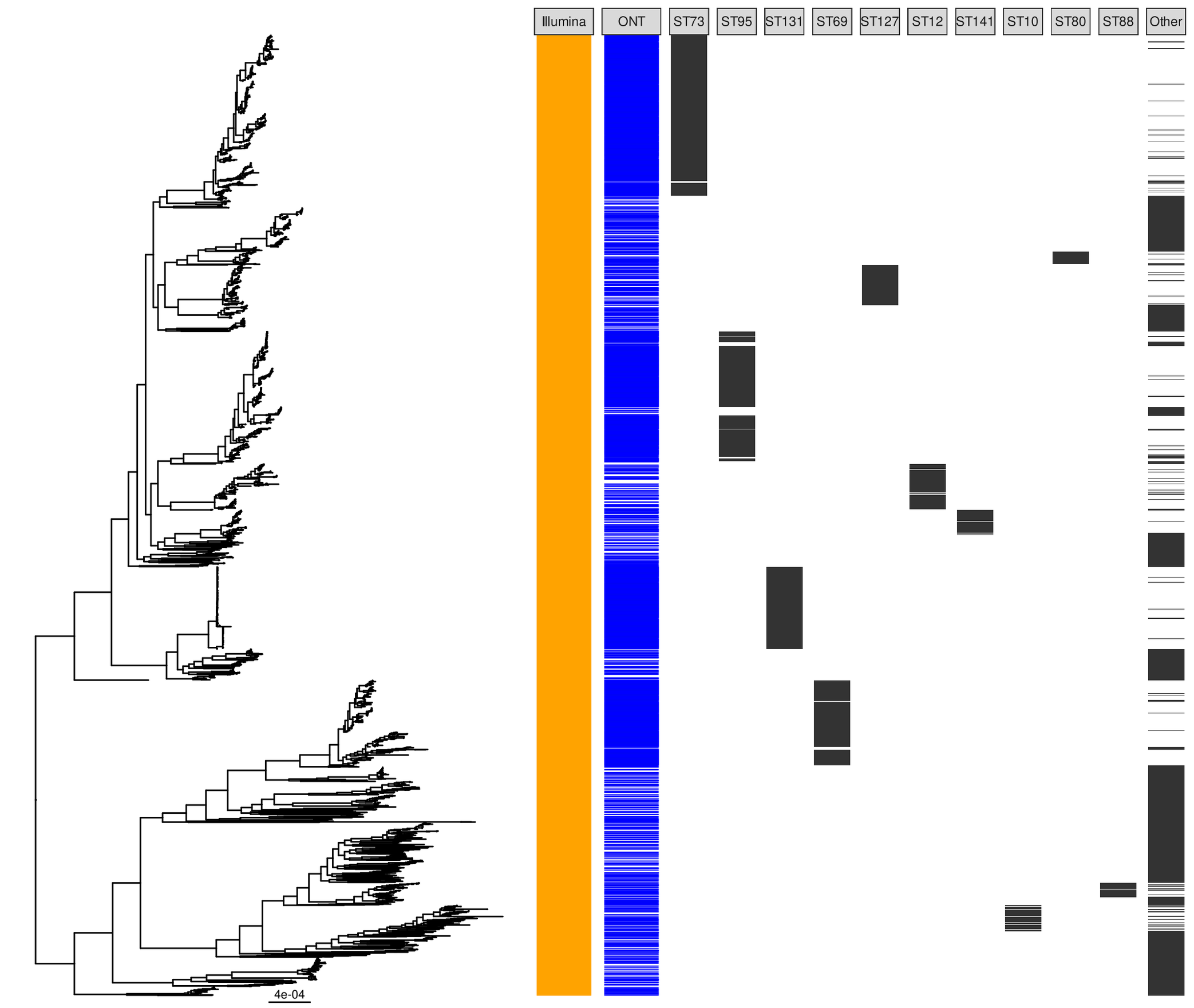
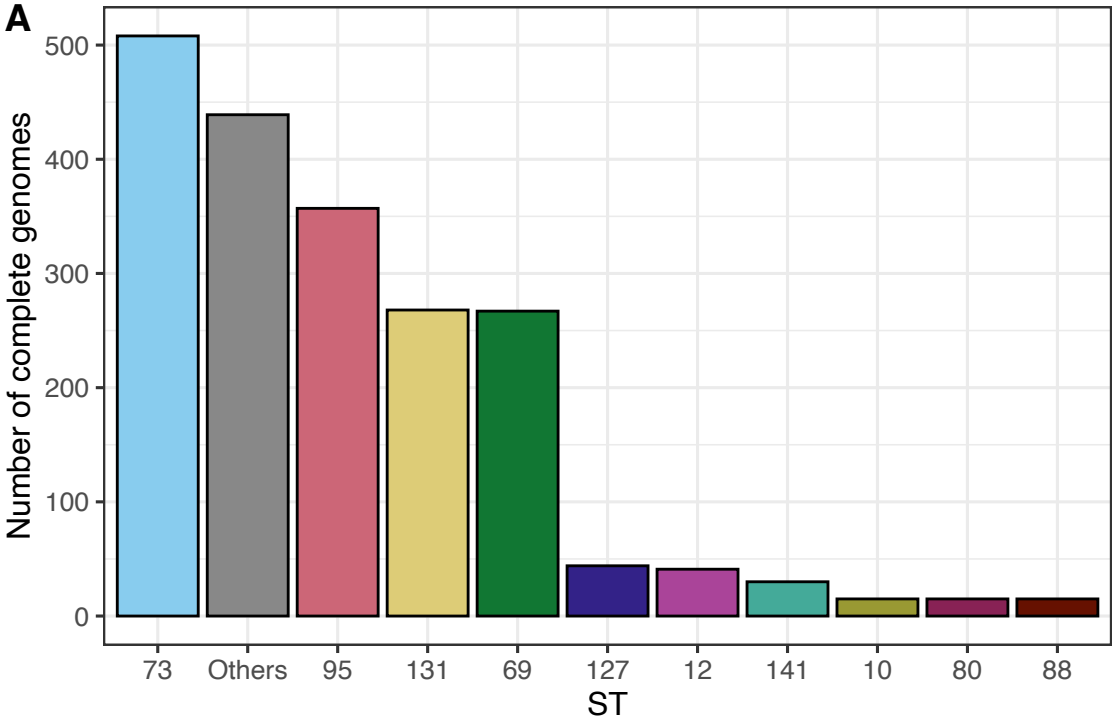
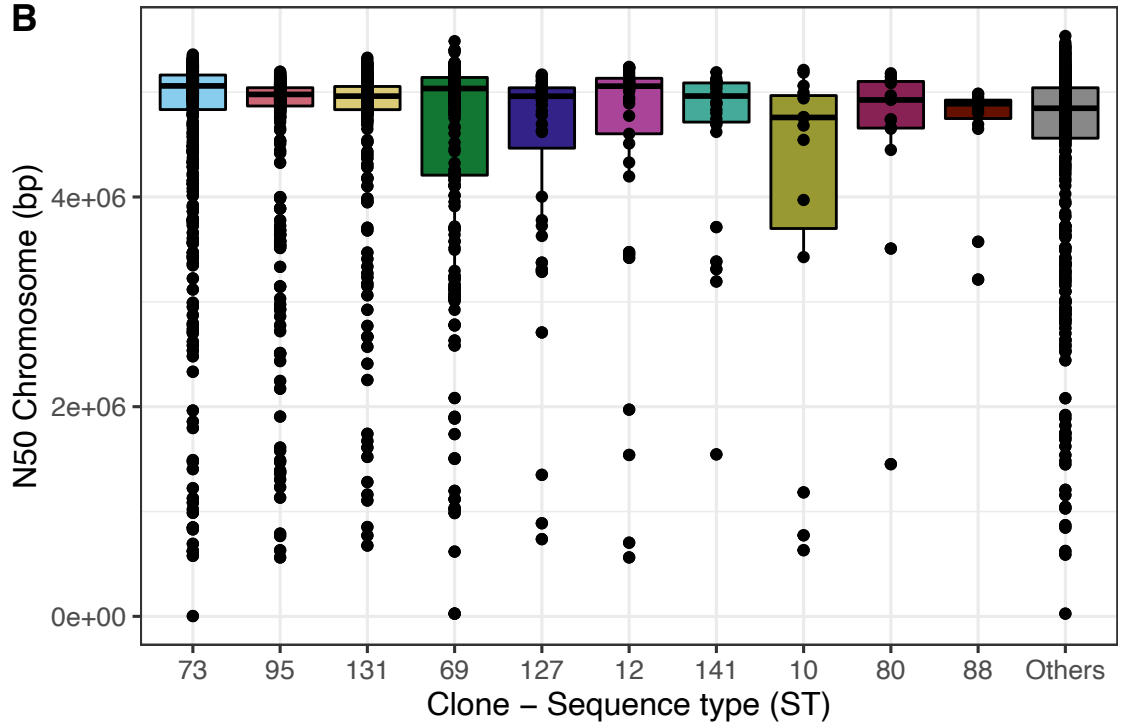

Illumina
ONT
ST73
ST95
ST131
ST69
ST127
ST12
ST141
ST10
ST80
ST88
Other
4e-04

Figure S2. Distribution of the isolates selected for long-read sequencing (2,045 isolates, ONT column) in the ExPEC phylogeny (3,245 isolates). This selection included 1,085 isolates chosen based on their accessory genome diversity and 960 isolates belonging to ST69, ST73, ST95 and ST131 not included in the first selection.

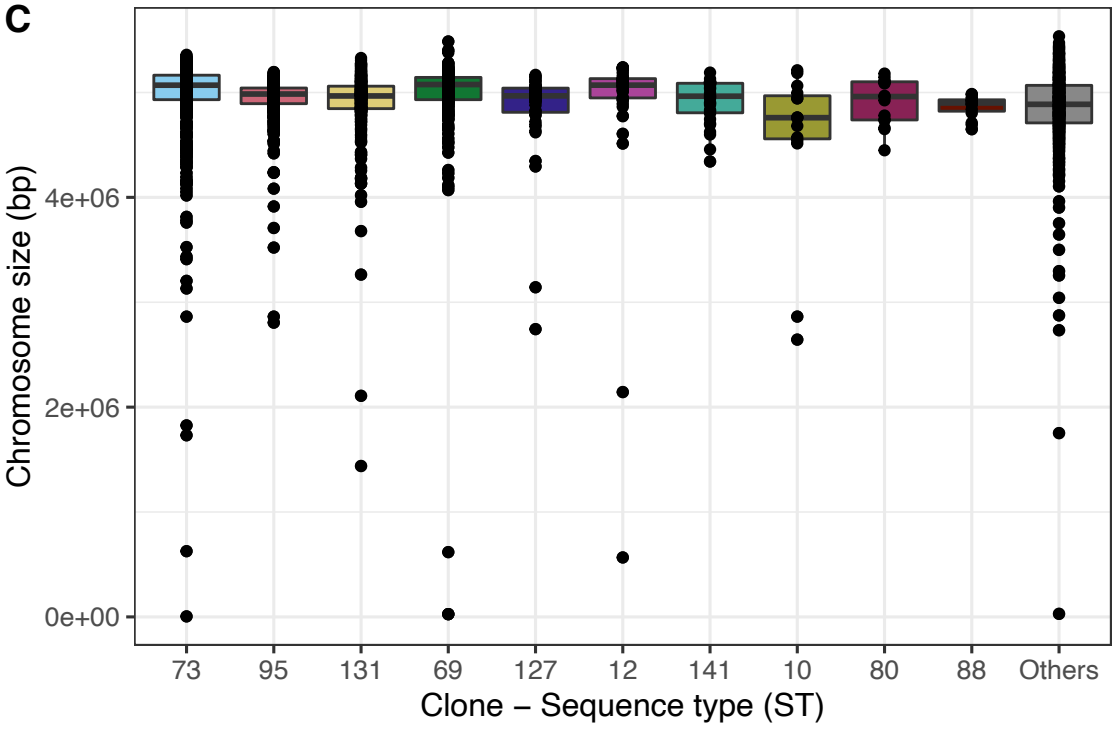
A


Number of complete genomes
500
400
300
200
100
0
73
Others
95
131
69
127
12
141
10
80
88
ST

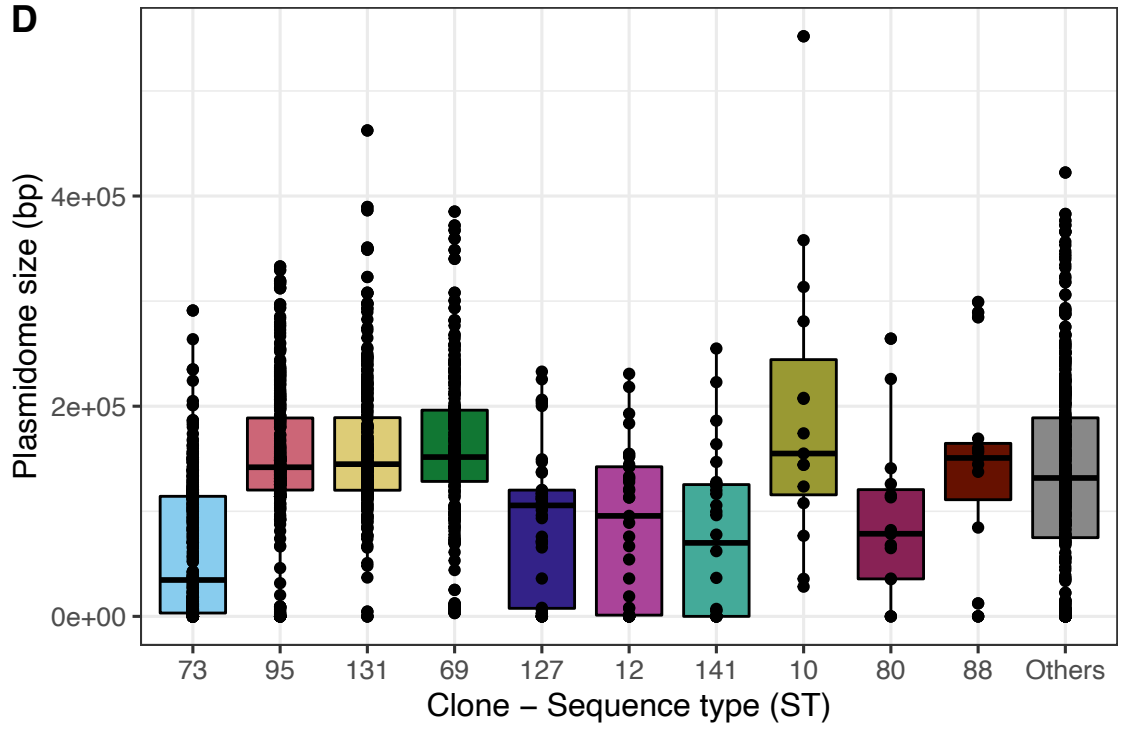
B


N50 Chromosome (bp)
4e+06
2e+06
0e+00
73
95
131
69
127
12
141
10
80
88
Others
Clone – Sequence type (ST)

C


Chromosome size (bp)
4e+06
2e+06
0e+00
73
95
131
69
127
12
141
10
80
88
Others
Clone – Sequence type (ST)

D


Plasmidome size (bp)
4e+05
2e+05
0e+00
73
95
131
69
127
12
141
10
80
88
Others
Clone – Sequence type (ST)

E

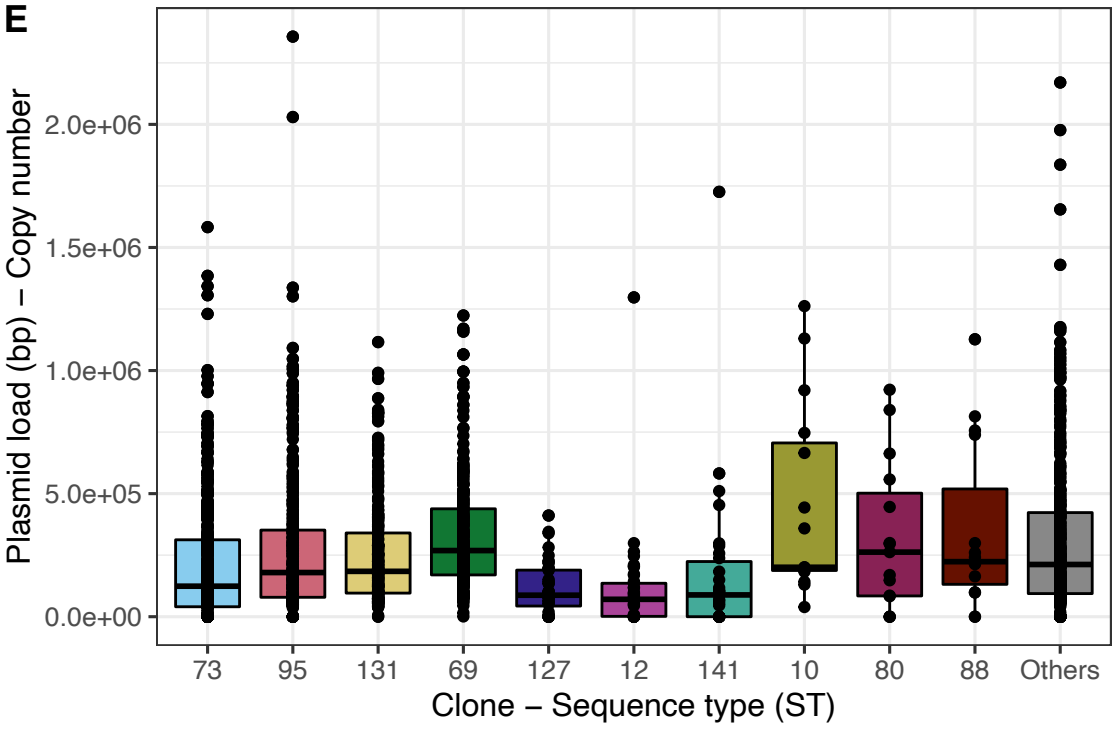

Plasmid load (bp) – Copy number
2.0e+06
1.5e+06
1.0e+06
5.0e+05
0.0e+00
73
95
131
69
127
12
141
10
80
88
Others
Clone – Sequence type (ST)

Figure S3. Genome statistics of the hybrid assemblies obtained for 1,999 ExPEC genomes. A) Barplot indicating the total number of assemblies available per ST. B) Boxplot of the N50 metric (in base-pairs) only considering contigs classified as chromosomal. C). Boxplot of the sum length (in base-pairs) considering contigs classified as chromosomal. D) Boxplot of the sum length (in base-pairs) of all contigs classified as plasmid (referred to as the plasmidome). E) Boxplot of the sum length (in base-pairs) of all contigs classified as plasmid but considering their copy number to estimate the total number of base-pairs present in each genome devoted to plasmid sequences. For this purpose, the length of each plasmid contig was multiplied by its normalized copy number as estimated by Unicycler.

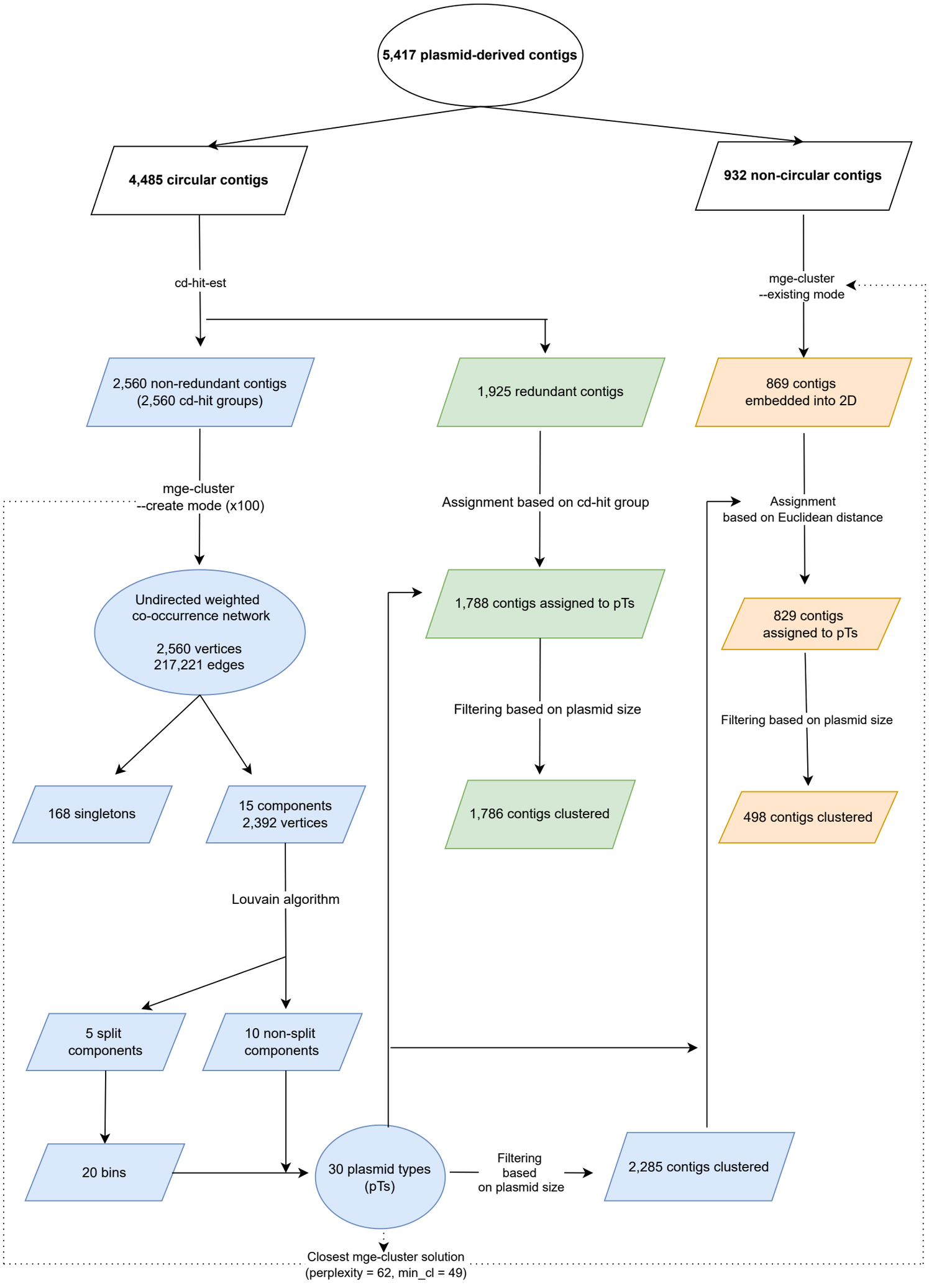

5,417 plasmid-derived contigs
4,485 circular contigs
932 non-circular contigs
cd-hit-est
mge-cluster
--existing mode
2,560 non-redundant contigs
(2,560 cd-hit groups)
1,925 redundant contigs
869 contigs
embedded into 2D
mge-cluster
--create mode (x100)
Assignment based on cd-hit group
Assignment
based on Euclidean distance
Undirected weighted
co-occurrence network
2,560 vertices
217,221 edges
1,788 contigs assigned to pTs
829 contigs
assigned to pTs
Filtering based on plasmid size
Filtering based on plasmid size
168 singletons
15 components
2,392 vertices
1,786 contigs clustered
498 contigs clustered
Louvain algorithm
5 split
components
10 non-split
components
20 bins
30 plasmid types
(pTs)
Filtering
based
on plasmid size
2,285 contigs clustered
Closest mge-cluster solution
(perplexity = 62, min_cl = 49)

Figure S4. Workflow followed for performing the plasmid type (pT) assignment on the set of 5,417 plasmid contigs derived from the hybrid assemblies.

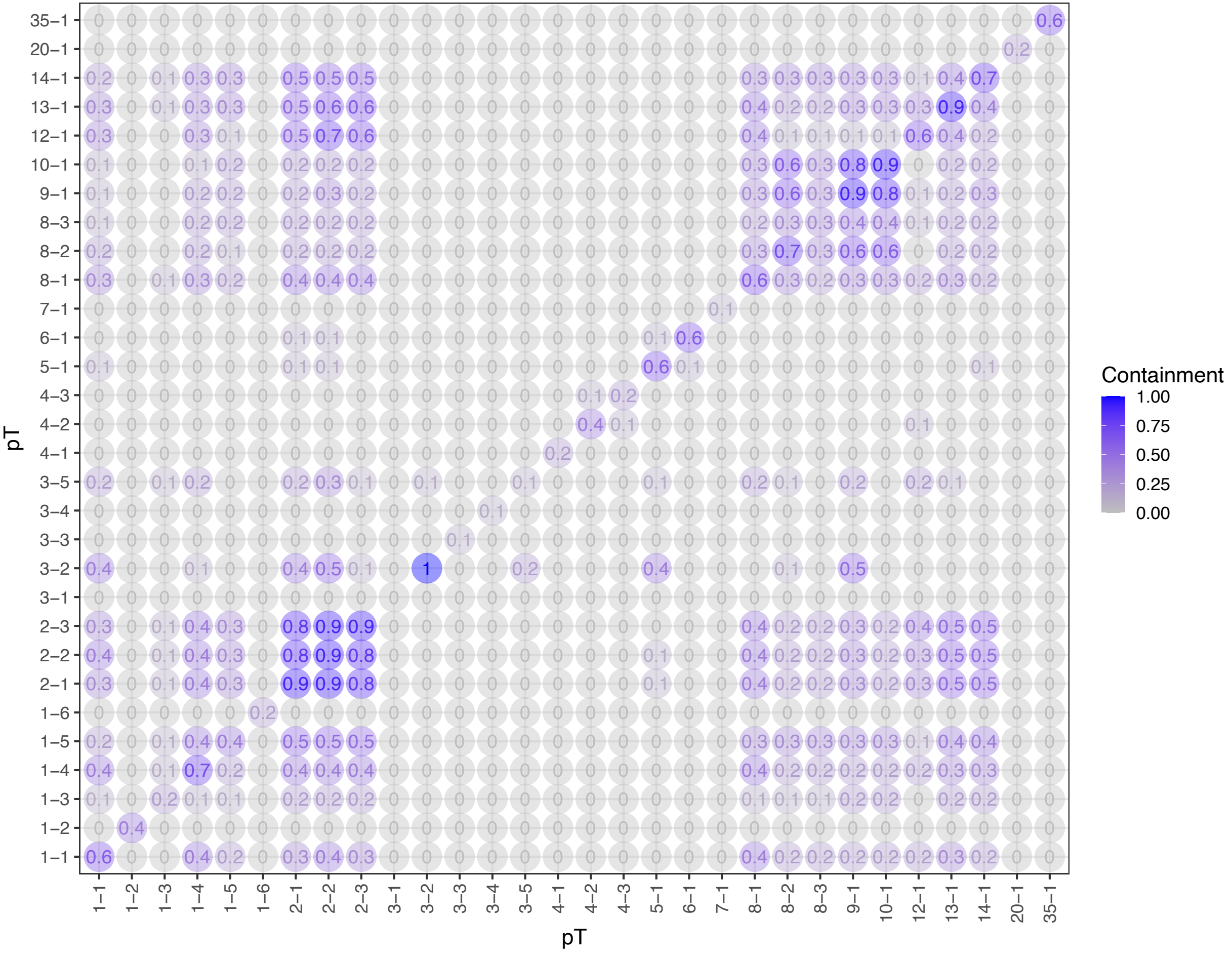

pT
35–1
20–1
14–1
13–1
12–1
10–1
9–1
8–3
8–2
8–1
7–1
6–1
5–1
4–3
4–2
4–1
3–5
3–4
3–3
3–2
3–1
2–3
2–2
2–1
1–6
1–5
1–4
1–3
1–2
1–1
1–1 1–2 1–3 1–4 1–5 1–6 2–1 2–2 2–3 3–1 3–2 3–3 3–4 3–5 4–1 4–2 4–3 5–1 6–1 7–1 8–1 8–2 8–3 9–1 10–1 12–1 13–1 14–1 20–1 35–1
pT
Containment
1.00
0.75
0.50
0.25
0.00

Figure S5. Sourmash containment analysis of the 30 plasmid types (pT) defined by mge-cluster. The containment value can range from 0 (pTs share no genomic region in common) to 1 (the entire genomic region of the pT indicated the y-axis is present in the pT indicated in the x-axis).

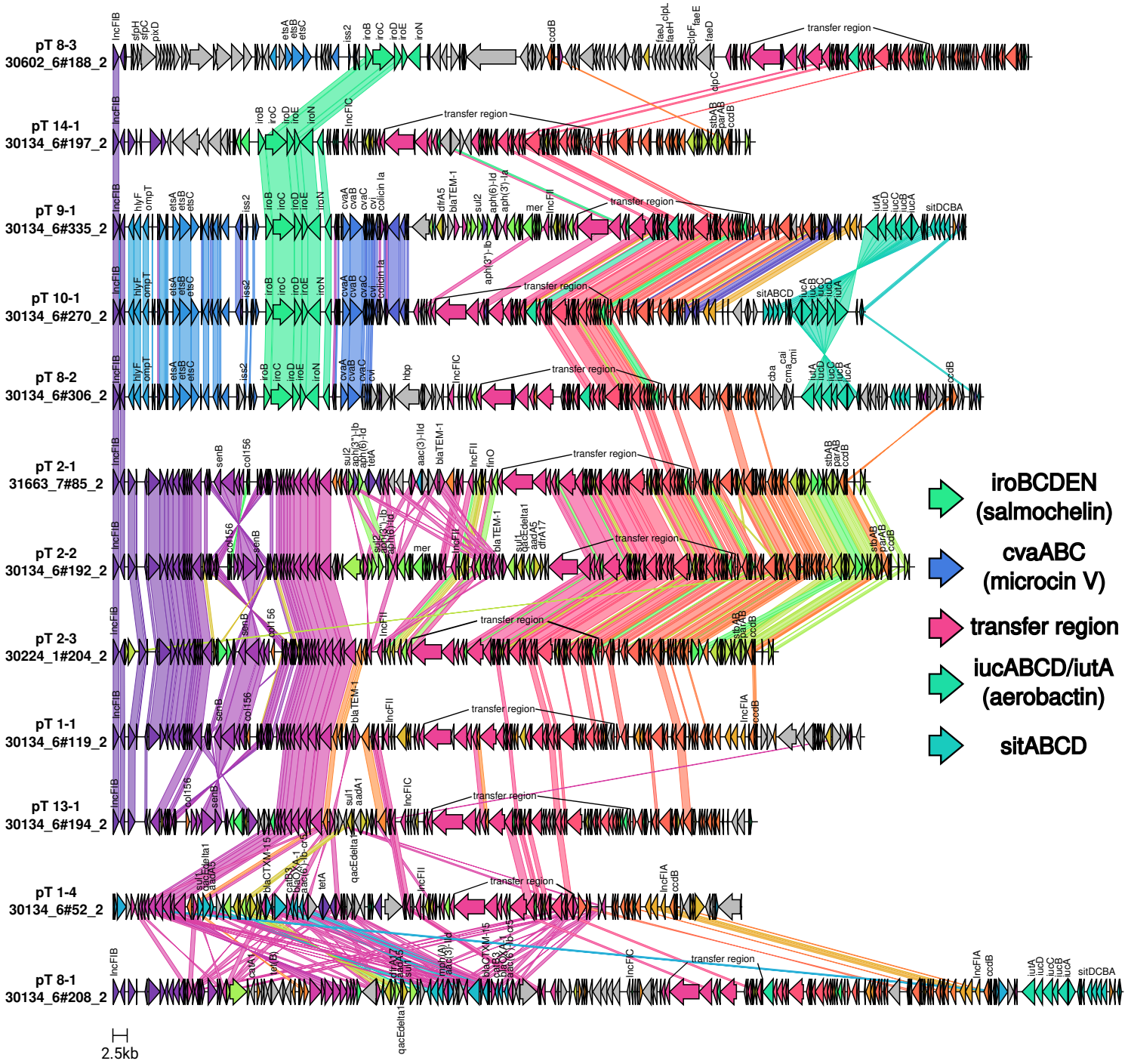

pT 8-3
30602_6#188_2
pT 14-1
30134_6#197_2
pT 9-1
30134_6#335_2
pT 10-1
30134_6#270_2
pT 8-2
30134_6#306_2
pT 2-1
31663_7#85_2
pT 2-2
30134_6#192_2
pT 2-3
30224_1#204_2
pT 1-1
30134_6#119_2
pT 13-1
30134_6#194_2
pT 1-4
30134_6#52_2
pT 8-1
30134_6#208_2
transfer region
iroBCDEN (salmochelin)
cvaABC (microcin V)
transfer region
iucABCD/iutA (aerobactin)
sitABCD
2.5kb

Figure S6. Clinker synteny visualization of the main large plasmid types (pT) considering a stringent minimum identity threshold of 0.99 to draw links between genes. Virulence and antibiotic resistance genes were identified using the EcOH database [82] (indexed in Abricate) and AMRFinderPlus [83] and their names curated according to well-studied ExPEC reference plasmids.

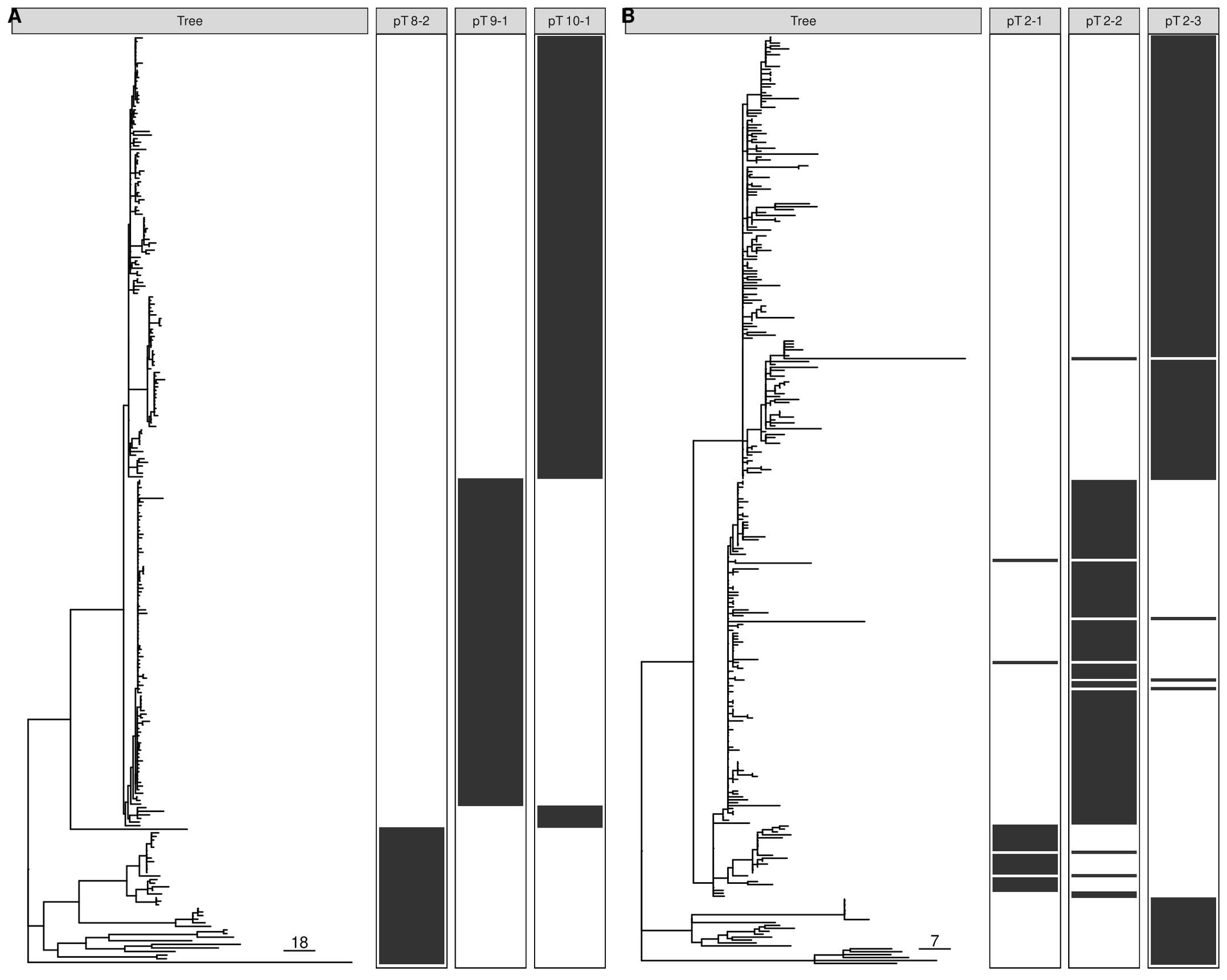

A
Tree
pT 8-2
pT 9-1
pT 10-1
18
B
Tree
pT 2-1
pT 2-2
pT 2-3
7

Figure S7. Recombination-free phylogeny based on plasmid sequences of pairs of plasmid types (pT) 8-2, 9-1, 10-1 (panel A) and 2-1, 2–2, 2-3 (panel B). The tree was computed using Gubbins based on a core-genome alignment estimated by Snippy. The tree branch length is measured as substitutions per site multiplied by the number of recombination-filtered SNPs detected by Gubbins.

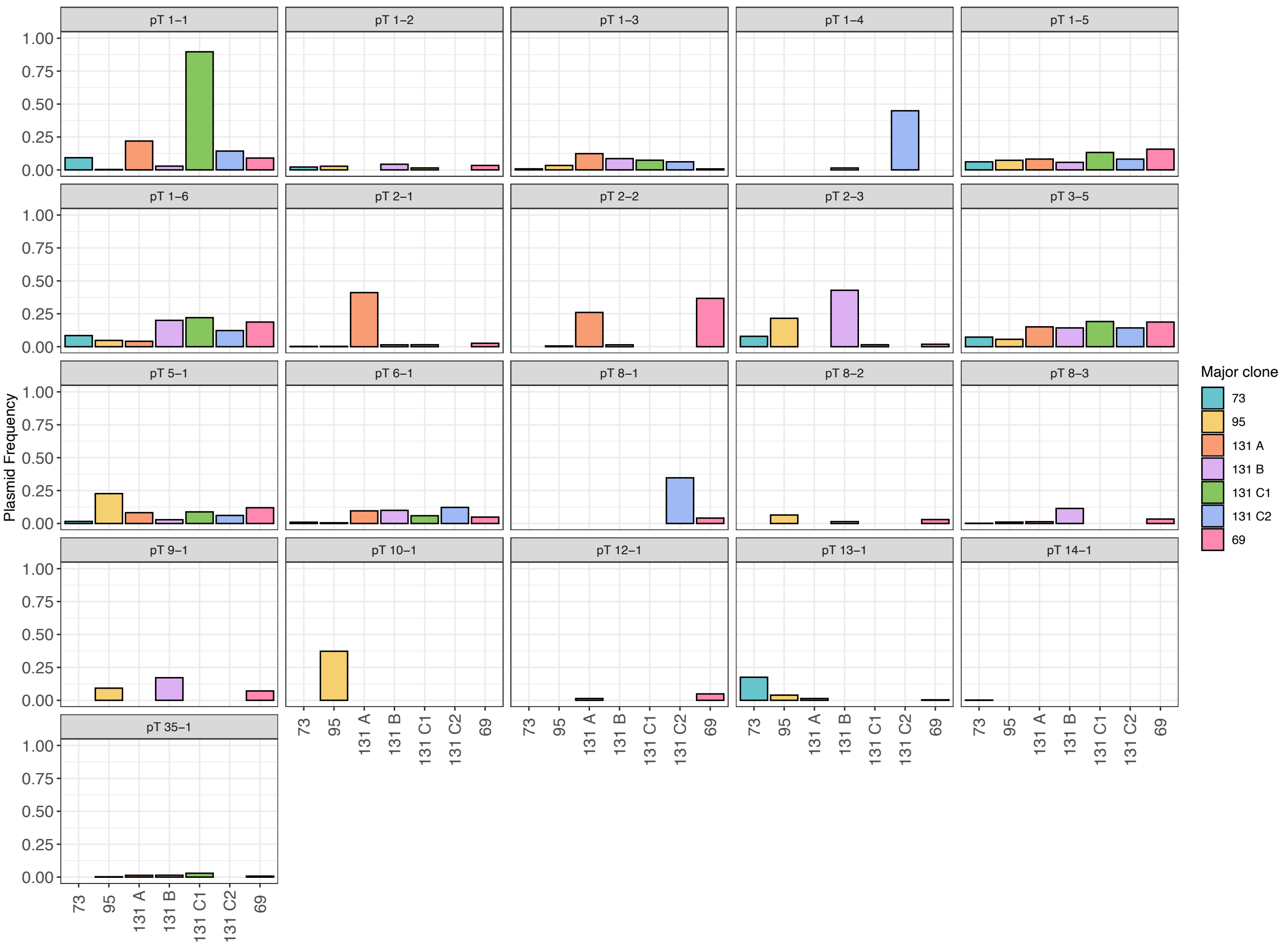
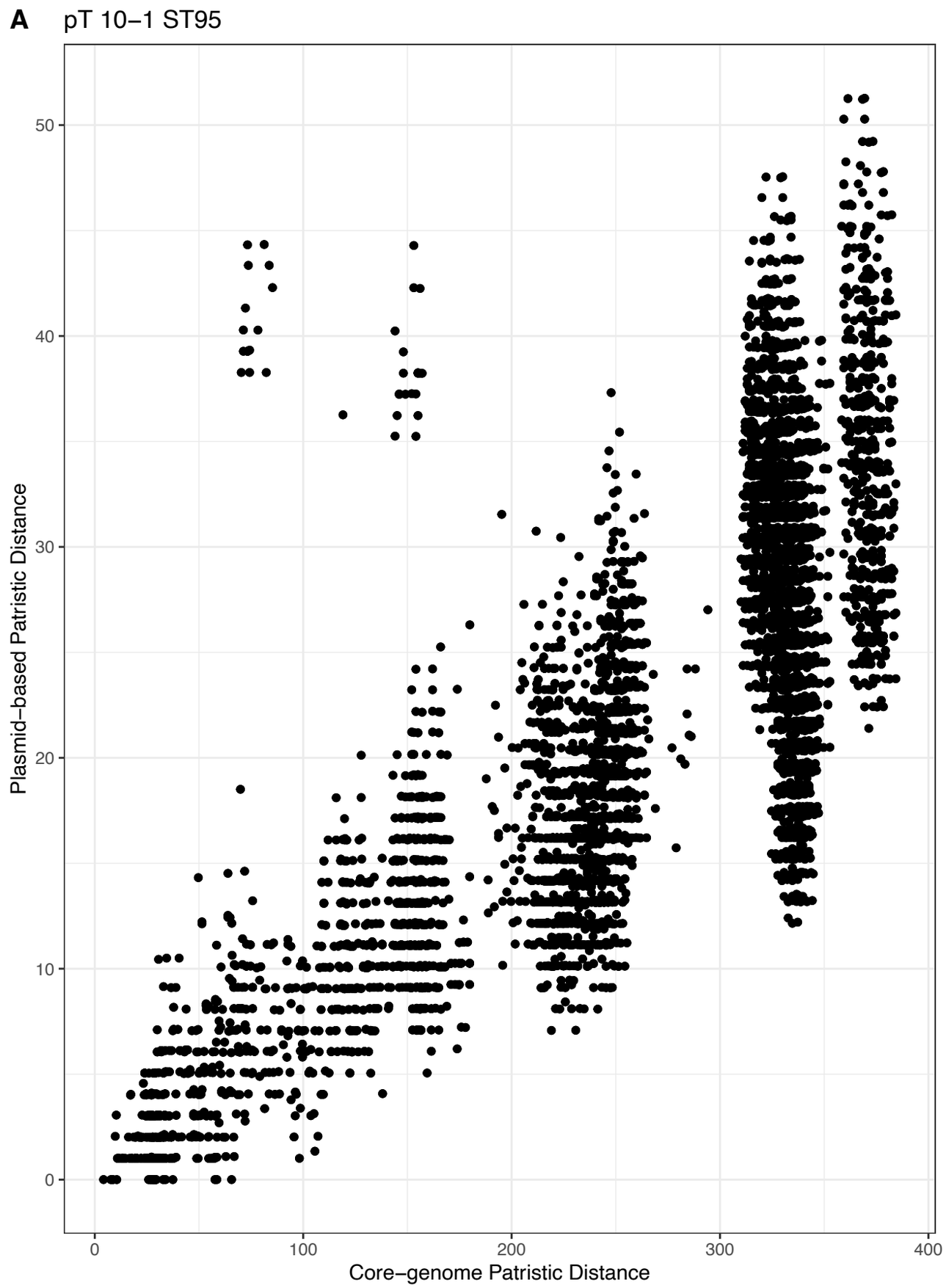

pT 1–1
pT 1–2
pT 1–3
pT 1–4
pT 1–5
pT 1–6
pT 2–1
pT 2–2
pT 2–3
pT 3–5
pT 5–1
pT 6–1
pT 8–1
pT 8–2
pT 8–3
pT 9–1
pT 10–1
pT 12–1
pT 13–1
pT 14–1
pT 35–1
Plasmid Frequency
1.00
0.75
0.50
0.25
0.00
73
95
131 A
131 B
131 C1
131 C2
69
Major clone

Figure S8. Plasmid type (pT) frequency among the four main ExPEC clones (ST69, ST73, ST95, ST131). In the case of ST131, we indicated the plasmid frequency according to the four clades present in the clone (A, B, C1, C2). Only pTs with an average size larger than 10 kb are represented.


A
pT 10–1 ST95
Plasmid–based Patristic Distance
Core–genome Patristic Distance
0
10
20
30
40
50
0
100
200
300
400

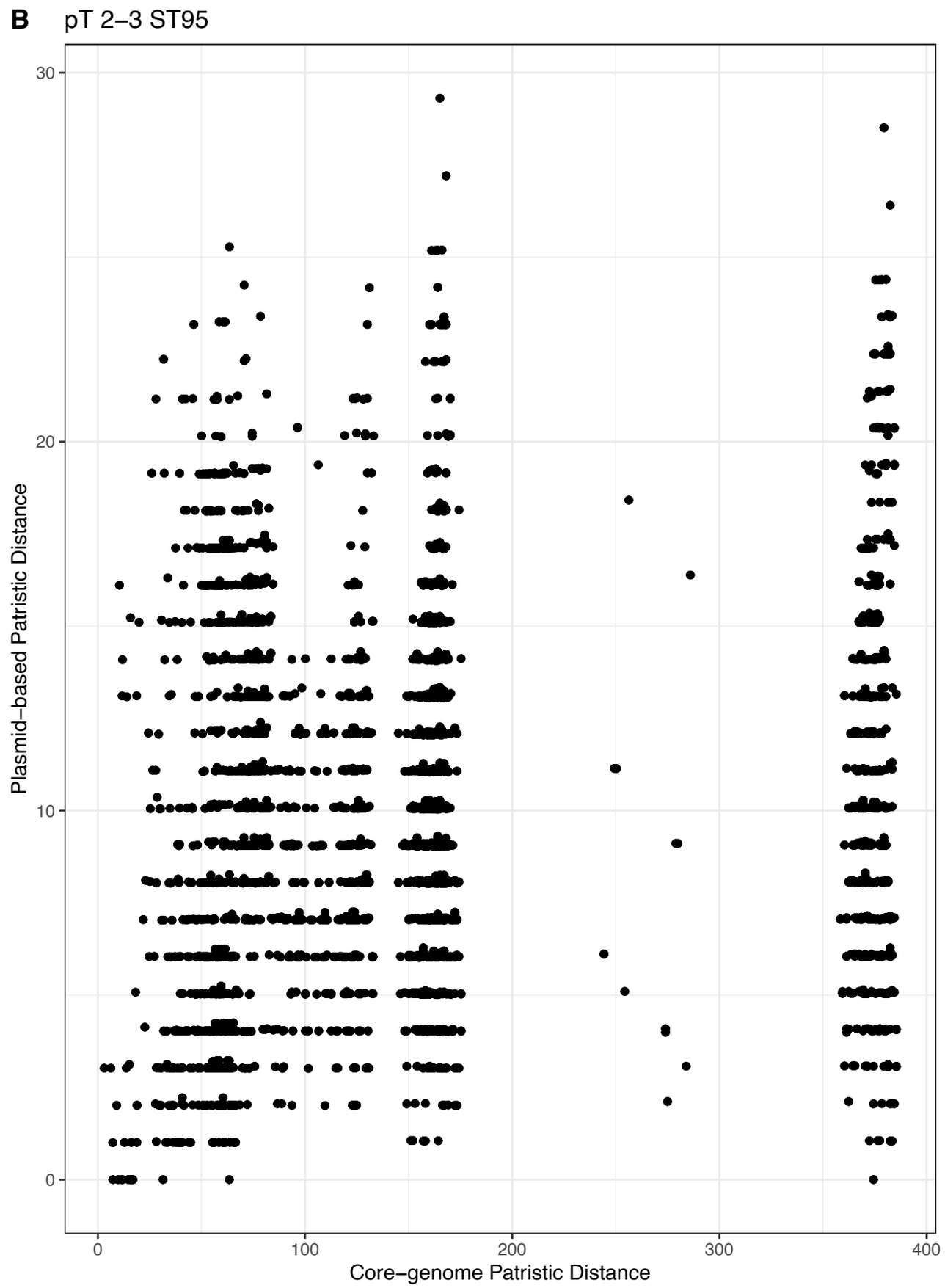

B
pT 2–3 ST95
Plasmid–based Patristic Distance
Core–genome Patristic Distance
0
10
20
30
0
100
200
300
400

Figure S9. Plasmid-based versus core-genome patristic distances in the main plasmid types (pT) present in ST95. A) A positive correlation between the two distances was suggestive of vertical transmission of the pT 10-1 in the ST95 clone. B) The lack of correlation among the two distances suggested that pT 2-3 was introduced multiple times in the ST95 clone.