Supplementary material

Within-patient and global evolutionary dynamics of *Klebsiella pneumoniae* ST17

Marit A. K. Hetland [1,2*], Jane Hawkey [3], Eva Bernhoff [1], Ragna-Johanne Bakksjø [1], Håkon Kaspersen [4], Siren I. Rettedal [5,6], Arnfinn Sundsfjord [7,8], Kathryn E. Holt [3,9], Iren H. Löhr [1,10]

[1] Department of Medical Microbiology, Stavanger University Hospital, Stavanger, Norway
[2] Department of Biological Sciences, Faculty of Mathematics and Natural Sciences, University of Bergen, Bergen, Norway
[3] Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Australia
[4] Research Section Food Safety and Animal Health, Department of Animal Health and Food Safety, Norwegian Veterinary Institute, Oslo, Norway
[5] Department of Research, Stavanger University Hospital, Stavanger, Norway
[6] Faculty of Health Sciences, University of Stavanger, Stavanger, Norway
[7] Department of Medical Biology, Faculty of Health Sciences, UiT – The Arctic University of Norway, Tromsø, Norway
[8] Norwegian National Advisory Unit on Detection of Antimicrobial Resistance, Department of Microbiology and Infection Control, University Hospital of North Norway, Tromsø, Norway
[9] Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom
[10] Department of Clinical Science, Faculty of Medicine, University of Bergen, Bergen, Norway

* Corresponding author: Marit A. K. Hetland, marit.hetland@uib.no

Contents

Supplementary Methods:
See pages 2-5 of this document for methods supplementary to the main text.

Supplementary Figures:
Figure S1: Temporal signal amongst ST17 genomes.
Figure S2: Paired gene gain/loss between NICU and follow-up genomes.
Figure S3: Dated phylogeny of 145 *Klebsiella pneumoniae* ST17 genomes from around the world.
Figure S4: Global phylogeny of 300 *Klebsiella pneumoniae* ST17 genomes and plasmid content.
Figure S5: Zoom-in of the clade with the Stavanger NICU outbreak genomes and closely related genomes that all carried the *bla*$_{\text{CTX-M-15}}$ pKp2177_1 plasmid on the KL25/O5 sublineage.

Supplementary Tables:
Please see separate Excel file for the Supplementary Tables:
Table S1: Stavanger NICU outbreak genome information, BioSample accessions and genotypes.
Table S2: Global ST17 dataset with BioSample accessions, metadata and genotypes.
Table S3: Within-host colonisation time, number and type of SNPs, and number of gene gain-loss events during the Stavanger NICU outbreak.
Table S4: Stavanger NICU outbreak SNPs, SNP type and affected gene products.
Table S5: Summary of literature that mentions *Klebsiella pneumoniae* ST17 in PubMed abstracts or titles, as of November 2021.

Supplementary Methods

Collection and sequencing of the Stavanger NICU outbreak isolates

Kp2177, a *Klebsiella pneumoniae* ST17 isolated in November 2008 from the breast milk of a neonate's mother, was previously identified as the likely index isolate for the outbreak. To resolve the sequence of this genome, we whole-genome sequenced the isolate using the same DNA extraction on both Illumina MiSeq and Oxford Nanopore Technologies (ONT) MinION platforms. The ONT sequencing was performed on an R9.4.1 flow cell, basecalled with guppy v3.0.3+ (available for ONT customers at https://nanoporetech.com) and subsampled using Filtlong v0.2.0 (https://github.com/rrwick/Filtlong) by discarding reads shorter than 1 Kbp and removing the 10% worst read bases. The short- and long-read sequences were then hybrid-assembled using Unicycler v0.4.8 [1] and subsequently annotated with RAST v2.0 [2].

Between and within-host comparison of the Stavanger NICU outbreak genomes

Core genome single nucleotide polymorphism (SNP) alignments were produced with RedDog v1beta.11 (https://github.com/katholt/RedDog) to identify SNPs between and within the outbreak samples. Bowtie2 v2.3.5.1 [3] was used with option --sensitive-local for sequence alignment and SAMTools v1.9 [4] was used to identify SNPs. Filters were applied to exclude allele calls from variant sites with ambiguous base calls, Phred quality <30, read depth <5, or evidence of strand bias. For all variant sites, base calls were extracted from the genomes in the alignment, and core genome positions were identified as positions where ≥95% of the genomes had Phred quality ≥20. Genomes were excluded from the alignments if they had <50% sequence coverage and <5X sequence depth to the reference.

Further, to remove any artefacts caused by sequencing, mapping or assembly errors in the Kp2177 reference genome, we aligned the short-reads of Kp2177 against the hybrid-assembled closed genome of itself, and excluded any SNP calls in positions where the two disagreed (this comprised 12 positions in plasmid pKp2177_2). For the remaining 91 short-read sequenced isolates, we mapped the reads from each isolate against the *de novo* assembly of itself to identify any variable positions (there were none). There were 145 SNPs called against plasmid pKp2177_1. Of those, 142 were present in only two genomes and they were concentrated in two genes. Upon inspection these were most likely homologs and were therefore excluded from the final list of SNPs. The final alignments consisted of 113 (chromosome), 3 (pKp2177_1), 4 (pKp2177_2) and 0 (pKp2177_3) SNPs.

Gubbins v3.1.6 [5] was used to identify recombination events in the core chromosomal alignment of variable and invariable sites; there were none. RAxML v8.2.12 [6] was then used to infer a core genome maximum likelihood (ML) phylogeny of the SNPs, using a rapid bootstrap analysis searching for the best-scoring ML tree, a GTR substitution model and GAMMA distribution of rate heterogeneity. This was performed in five independent replicates, and the tree with the highest likelihood was retained (Figure 1).

Phylogenetic inference and molecular dating of the global ST17 collection

To assess the global dynamics of *K. pneumoniae* ST17, we included 254 genomes from around the world that were downloaded from GenBank or the European Nucleotide Archive. Fourty-one of these genomes were available only as *de novo* assemblies. To allow them to be included in the phylogenetic analysis, SAMtools wgsim was used to simulate 100 bp paired-end reads without error from the assemblies.

A core chromosomal SNP alignment of 46 Stavanger NICU outbreak genomes (one genome from each patient; 44 follow-up faecal, 1 blood, 1 breast milk) and the 254 global genomes against the Kp2177 chromosome was generated with RedDog. Five SNPs that were artefacts between the Kp2177 short-reads and the reference genome were excluded. Recombinant regions were filtered from the alignment using Gubbins v3.1.6 [5] with the weighted robinson foulds convergence method. The recombination-free alignment (n=10,313 SNPs) was then passed to RaxML v8.2.12 [6] to infer a core genome ML phylogeny, using a rapid bootstrap analysis searching for the best-scoring ML tree, a GTR substitution model and GAMMA distribution of rate heterogeneity. This was performed in five independent replicates. The final ML scores were highly similar and the phylogeny with the best score was imported to TempEST v1.5.3 [7] to investigate the relationship between the root-to-tip distances in the ML tree and the years of isolation. The slope of the root-to-tip regression under heuristic residual mean squared was negative (-0.0002), indicating low or no molecular clock levels [7]. As the dataset was quite large, compared to other dated clones such as ST307 (n=95) and CG147 (n=218) [8,9], we decided to subset the ST17 global dataset for the temporal analysis and repeated the steps above. We used patristic distances to group the 300 genomes into 125 groups and picked one isolate from each group to represent the diversity in the tree. Four of the resulting groups had >15 genomes. For these, we additionally picked ≥3 genomes to cover each country within the group. The root-to-tip regression slope was now positive, indicating that a molecular clock signal was present.

BEAST2 v2.6.5 [10] was used for the Bayesian phylogenetic analysis of the final recombination-free alignment of 11,345 SNPs in 145 genomes to estimate the phylogenetic tree and evolutionary rate. We specified exact date of sampling for the genomes where this information was available, and for the genomes that we only knew the year or the year and month of sampling (n=78), we used tip calibration in BEAUTi v2.6.5 [10] to sample the tip dates. We ran four different model combinations through BEAST2, using two clock models (strict and relaxed log normal) and two demographic models (constant population and exponential population). Each model was run in three replicates, to 350 million states until all parameters had >200 ESS and removing 10% of the states was sufficient to remove the burn-in. The model with relaxed log normal clock and constant population size was determined the best fit for the dataset: The rate coefficient of variance in the relaxed clock models did not touch 0, meaning the rate was not the same across the tree, so the relaxed model allowing for variation in rate was selected. The growth rate distribution in the constant exponential model went through 0, suggesting that the population was not growing, therefore the constant coalescent was selected. The trees from this model were subsampled with LogCombiner to 22,500 trees, which were passed to TreeAnnotator to summarise them into a single target maximum clade credibility tree (Figure S3). To confirm the temporal signal of the clone, ten independent date-randomisation tests were performed to show that the estimated rate of the true data did not overlap with those from the randomised dates. There was no overlap between

the 95% HPD evolutionary rate of the true and randomised data (Figure S1). To confirm that the prior was not driving the results, the analysis was also performed with sampling from priors only (i.e. without the sequence alignment); there was no overlap with the true data.

Comparing pKp2177_1 with non-Norwegian ST17 and non-ST17 Norwegian genomes

To determine the presence of the *bla*$_{\text{CTX-M-15}}$-encoding pKp2177_1 plasmid, which was persistently present in the Stavanger NICU outbreak, we used RedDog as described above to perform read mapping against two datasets: The collection of 254 *K. pneumoniae* ST17 genomes not from the outbreak, and a collection of 3,212 *K. pneumoniae* genomes of non-ST17 sequence types from Norway that were isolated between 2001 and 2020. They were from human clinical infection (n=2,073; 868 published in [11] and a further 1,205 not yet published), human gut colonisation of healthy adults (n=481) [12], marine bivalves and seawater (n=97) [13] and animals, including turkey and broilers (n=203) [14] and a further 206 isolates not yet published from turkey, broilers, pigs and dogs. Genomes were considered positive for a plasmid if they had <10 SNPs and ≥80% mapping coverage of the reference, as in [15].

References

1. **Wick RR, Judd LM, Gorrie CL, Holt KE.** Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13(6):e1005595. doi:10.1371/journal.pcbi.1005595

2. **Aziz RK, Bartels D, Best AA, *et al.*** The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008;9:75. doi:10.1186/1471-2164-9-75

3. **Langmead B, Salzberg SL.** Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9(4):357-359. doi:10.1038/nmeth.1923

4. **Li H, Handsaker B, Wysoker A, *et al.*** The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

5. **Croucher NJ, Page AJ, Connor TR, *et al.*** Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43(3):e15. doi:10.1093/nar/gku1196

6. **Stamatakis A.** RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30(9):1312-1313. doi:10.1093/bioinformatics/btu033

7. **Rambaut A, Lam TT, Max Carvalho L, Pybus OG.** Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* 2016;2(1):vew007. doi:10.1093/ve/vew007

8. **Wyres KL, Hawkey J, Hetland MAK, *et al.*** Emergence and rapid global dissemination of CTX-M-15-associated *Klebsiella pneumoniae* strain ST307. *J Antimicrob Chemother* 2019;74(3):577-581. doi:10.1093/jac/dky492

9. **Rodrigues C, Desai S, Passet V, Gajjar D, Brisse S.** Genomic evolution of the globally disseminated multidrug-resistant *Klebsiella pneumoniae* clonal group 147. *Microb Genom* 2022;8(1):000737. doi:10.1099/mgen.0.000737

10. **Bouckaert R, Vaughan TG, Barido-Sottani J, *et al.*** BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 2019;15(4):e1006650. doi:10.1371/journal.pcbi.1006650

11. **Fostervold A, Hetland MAK, Bakksjø R, *et al.*** A nationwide genomic study of clinical *Klebsiella pneumoniae* in Norway 2001-15: introduction and spread of ESBLs facilitated by clonal groups CG15 and CG307. *J Antimicrob Chemother* 2022;77(3):665-674. doi:10.1093/jac/dkab463

12. **Raffelsberger N, Hetland MAK, Svendsen K, *et al.*** Gastrointestinal carriage of *Klebsiella pneumoniae* in a general adult population: a cross-sectional study of risk factors and bacterial genomic diversity. *Gut Microbes* 2021;13(1):1939599. doi:10.1080/19490976.2021.1939599

13. **Håkonsholm F, Hetland MAK, Svanevik CS, Lunestad BT, Löhr IH, Marathe NP.** Insights into the genetic diversity, antibiotic resistance and pathogenic potential of *Klebsiella pneumoniae* from the Norwegian marine environment using whole-genome analysis. *Int J Hyg Environ Health* 2022;242:113967. doi:10.1016/j.ijheh.2022.113967

14. **Franklin-Alming FV, Kaspersen H, Hetland MAK, *et al.*** Exploring *Klebsiella pneumoniae* in Healthy Poultry Reveals High Genetic Diversity, Good Biofilm-Forming Abilities and Higher Prevalence in Turkeys Than Broilers. *Front Microbiol* 2021;12:725414. doi:10.3389/fmicb.2021.725414

15. **Hawkey J, Wyres KL, Judd LM, *et al.*** ESBL plasmids in *Klebsiella pneumoniae*: diversity, transmission and contribution to infection burden in the hospital setting. *Genome Med* 2022;14(1):97. doi:10.1186/s13073-022-01103-0

Supplementary Figures

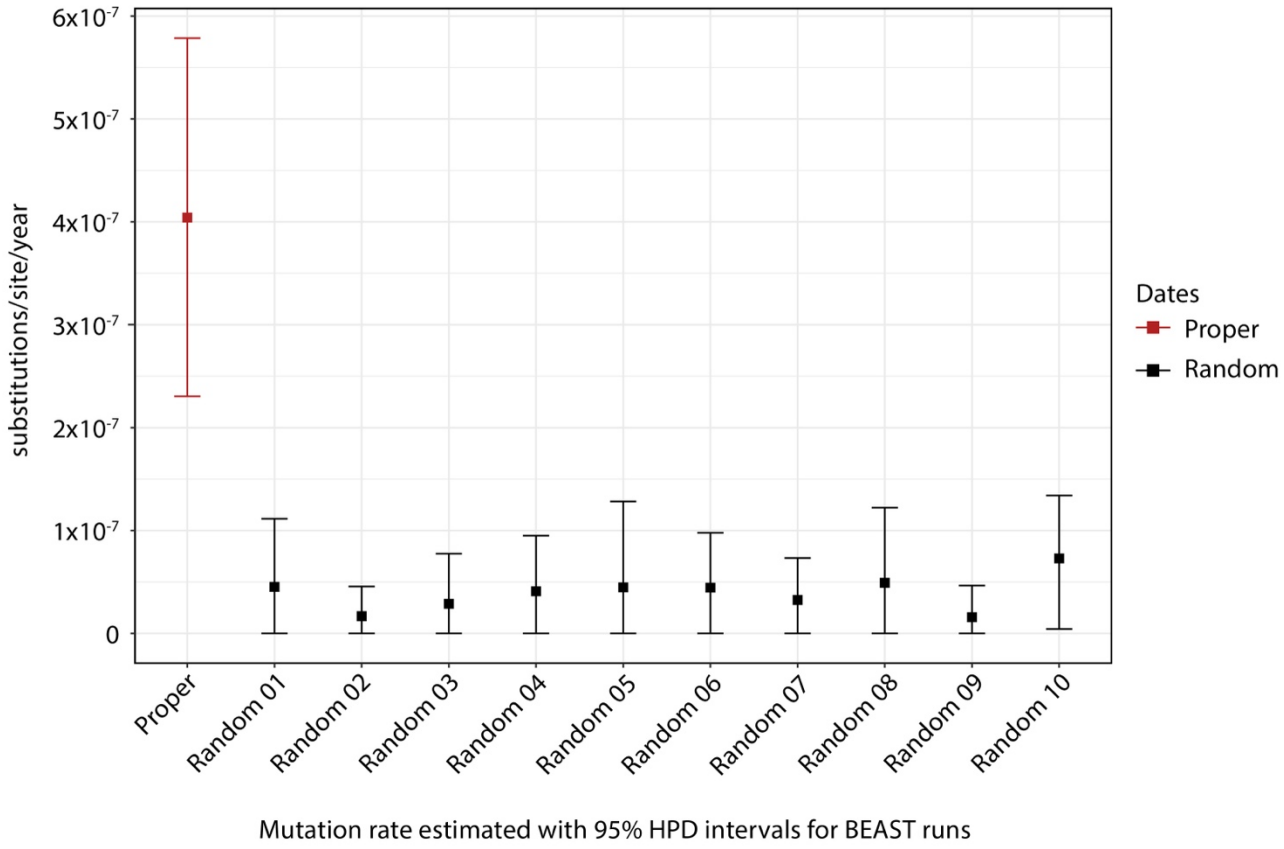

Figure S1. Temporal signal amongst ST17 genomes. Posterior distributions for evolutionary rate estimates generated in BEAST2 with the true isolate collection dates (red) and randomised dates (n=10, black).

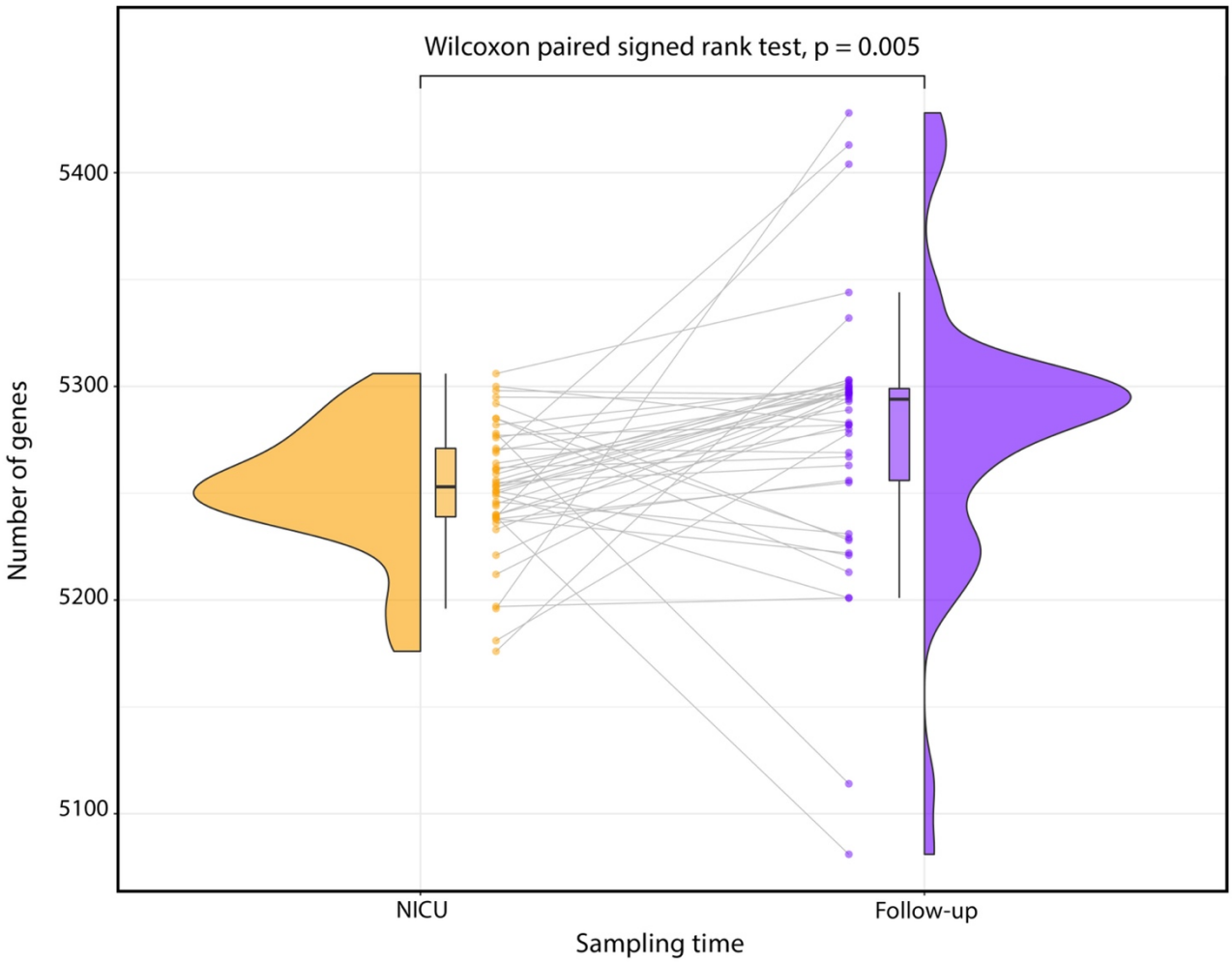

Figure S2. Paired gene gain/loss between NICU and follow-up genomes. Differences in the number of genes between the neonatal intensive care unit (NICU, on the left) and follow-up (taken 3-21 months after colonisation in the NICU, median 11, on the right) isolates. From outer to inner: Density plot, box plot with line indicating median, paired child data points. There was a significant within-host increase in the number of genes in the follow-up genomes (p=0.005).

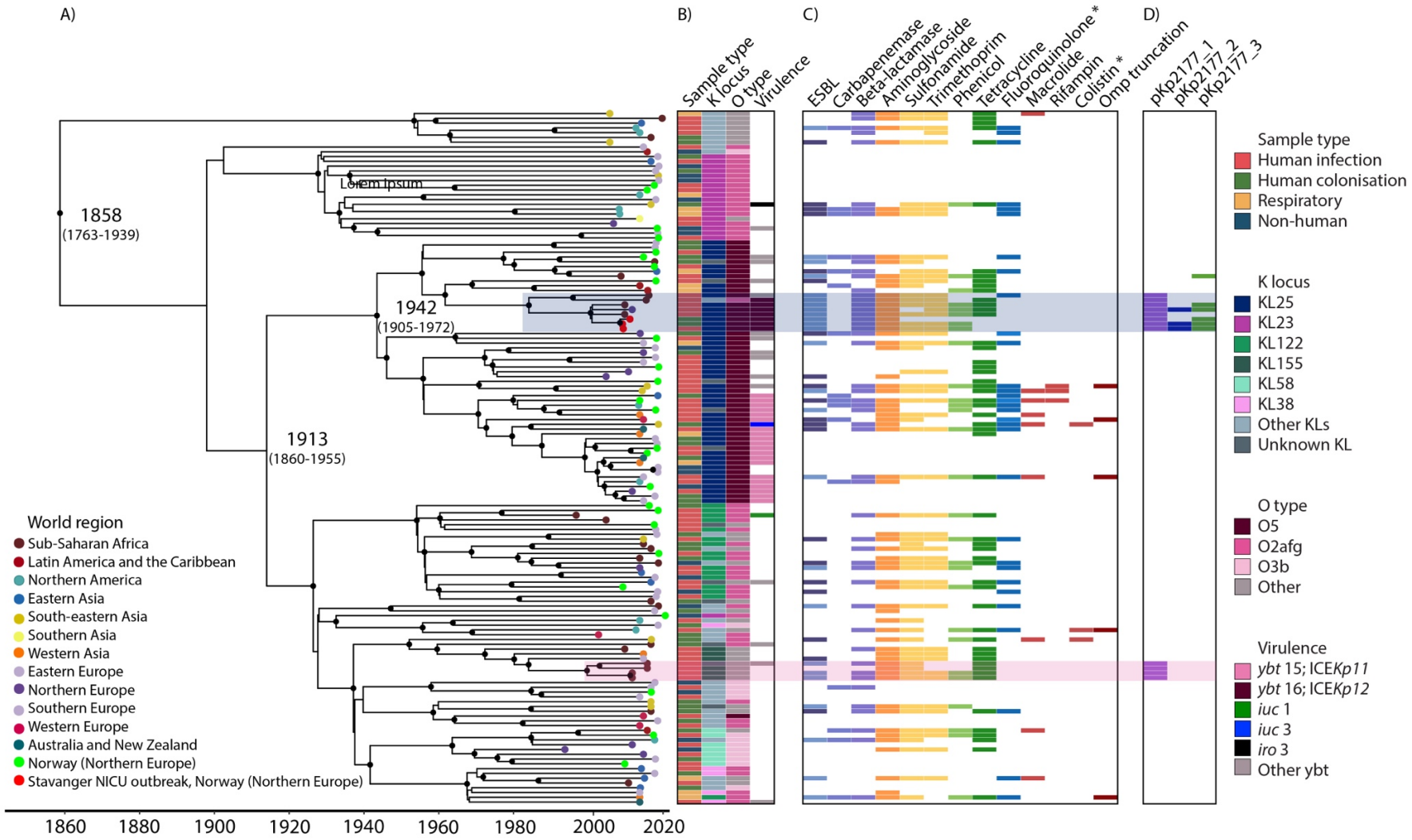

Figure S3. Dated phylogeny of 145 *Klebsiella pneumoniae* ST17 genomes from around the world. A) Dated tree with tips coloured by world region of collection. Additionally, the Stavanger NICU outbreak genomes are coloured red and other genomes from Norway green. Black dots on internal nodes indicate ≥95% posterior probability. **B)** Sample type and loci as indicated in the column names. The most prevalent loci are indicated in the inset legend. **C)** Presence (colour) or absence (white) of genes encoding resistance to the listed antimicrobial resistance (AMR) drug classes (blocks are coloured by drug class). Lighter colour in the ESBL column indicates $bla_{\text{CTX-M-15}}$. **D)** Presence (black) or absence of the Kp2177 plasmids. The highlighted clades include the representative sequences of the pKp2177_1-harbouring clusters: blue = KL25/O5 clade; pink = KL155/OL101. * Acquired genes and mutations.

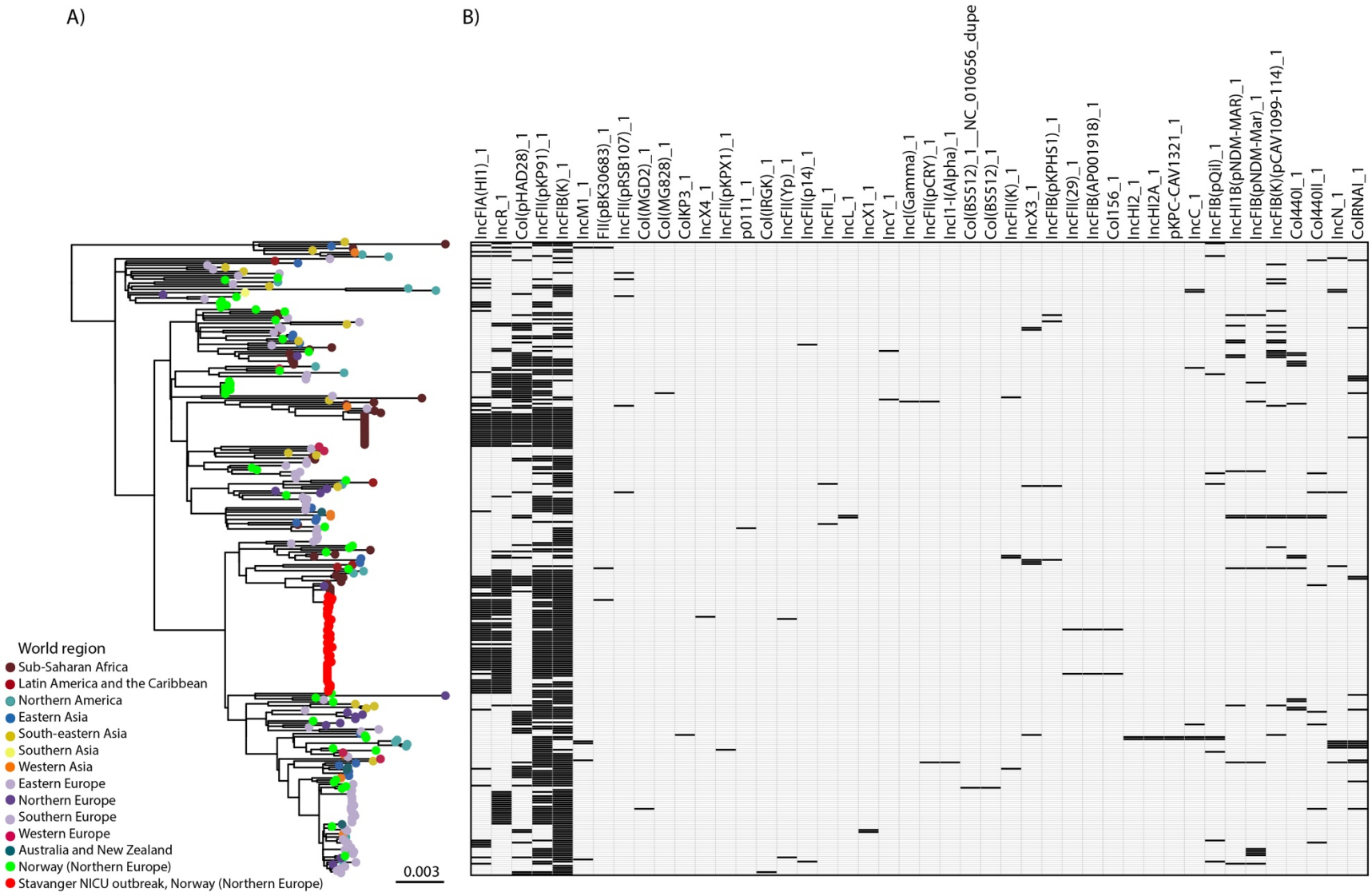

Figure S4. Global phylogeny of 300 *Klebsiella pneumoniae* ST17 genomes and their plasmid content. A) Maximum Likelihood tree reproduced from Figure 3. The tips are coloured by world region of collection. Additionally, the Stavanger NICU outbreak genomes are coloured red and other genomes from Norway green. **B)** Presence (black) or absence (white) of all detected plasmid replicon markers in the dataset.

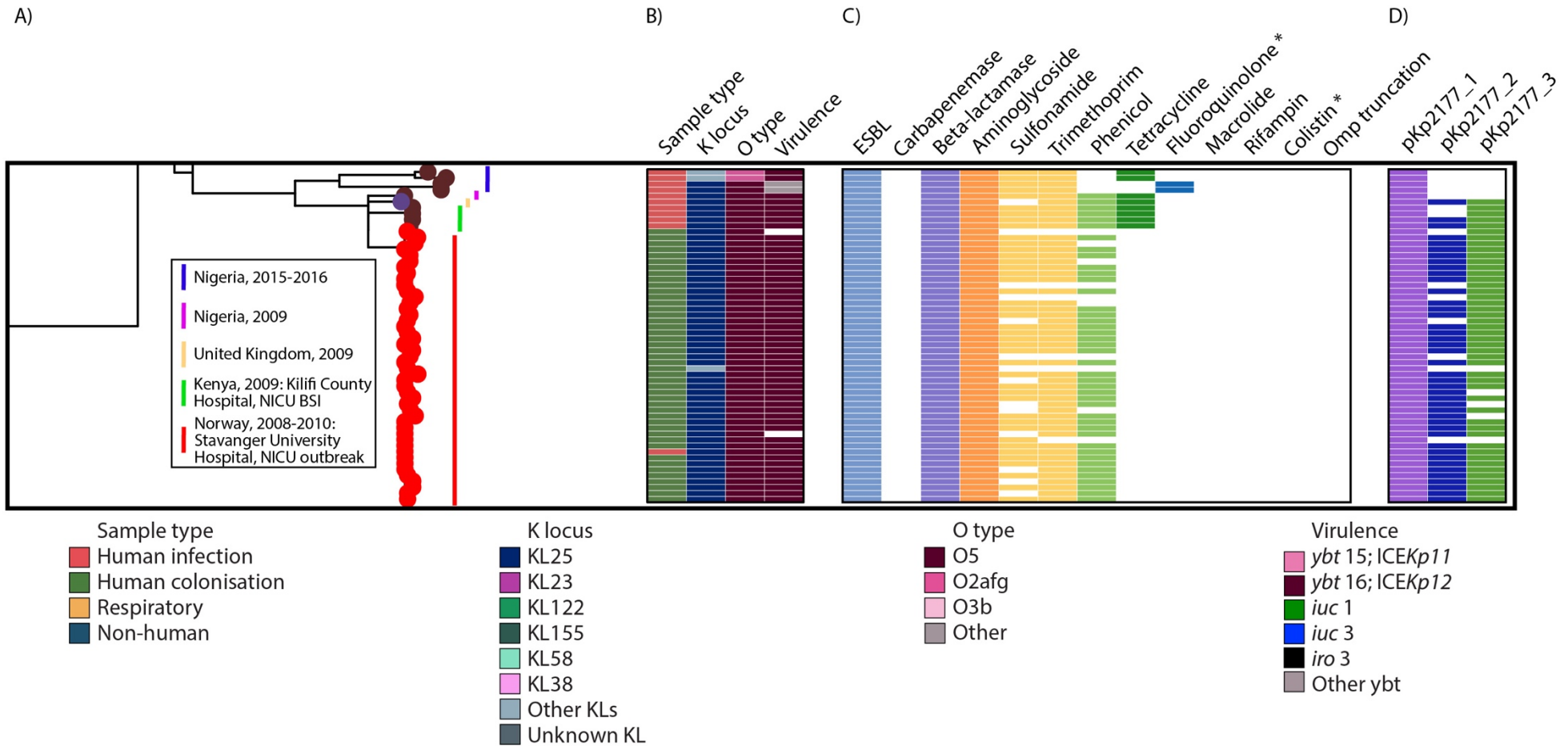

Figure S5. Zoom-in of the clade with the Stavanger NICU outbreak genomes and closely related genomes that all carried the *bla*$_{\text{CTX-M-15}}$ pKp2177_1 plasmid on the KL25/O5 sublineage. The figure is a zoom-in of the maximum likelihood tree in Figure 3. The lines next to the tree tips indicate which country and year the genomes were collected from in the inset legend.