List of Figures

Chapter 7

Whole genome sequencing as a high-resolution typing tool to track longitudinal ESBL-E colonisation

7.1 Chapter overview

In this chapter, I present an attempt to use short read whole genome sequence (WGS) data as a high resolution typing tool to track bacteria and mobile genetic elements (MGE) within the participants of the study. I use a hierarchical BAPS (Bayesian analysis of population structure) algorithm to classify the core gene alignment of 473 *E. coli* into 48 sequence clusters, which mapped well to the phylogeny. I used the *cd-hit* algorithm to cluster 488 ESBL-gene containing contigs into 99 clusters, which showed some lineage association on mapping them back to the phylogeny. Largely, hospital associated isolates were independent of sequence cluster assignment with the exception of sequence cluster 23 which was associated with hospital acquisition ($p = 6.3 \times 10^{-4}$), and corresponded to the putative recently arrived high-risk clone ST410 described inn Chapter 6.

The combination of hierBAPS sequence cluster and ESBL-contig cluster together were conserved within participants on longitudinal sampling compared to between-participants ($p = 1.1 \times 10^{-12}$) whereas either alone was not ($p = 0.4$ and $p = 1.0$). However beyond 35 days apart any two samples from a single a participant were only as likely to contain the same sequence cluster-ESBL cluster combination as any two samples randomly selected from the dataset. This suggests that, firstly, the unit of transmission in this system is likely to be the bacterium rather than the MGE. Secondly, the within-participant ESBL contig cluster

bacterial sequence cluster association suggests that a given bacterium-MGE combination is reasonably stable on the timescale of the study. Finally, these data suggest there is turnover of ESBL *E. coli* on a time scale of around 35 days suggestive of either frequent re-exposure or some other endogenous turnover in the microbiota.

7.2 Introduction and chapter aims

In Chapter 5, I described the epidemiology of ESBL-E carriage in study participants as they were exposed to antimicrobials and hospitalisation, showing a dramatic increase in carriage prevalence following hospitalisation and, particularly, hospitalisation and antibacterial exposure. In Chapter 6, I presented details of the whole genome sequencing of a sample of 473 *E coli* isolates recovered from the stool of these study participants, placed them in a global context, and described the antimicrobial resistance (AMR) determinants that the carried. In this chapter, I present an analysis whereby I combine the WGS data with the metadata from Chapter 5 in an attempt to use WGS as a high-resolution typing tool to track carriage of ESBL *E. coli* within participants in this study. Specifically, the aims of this chapter are:

1. To use clustering algorithms to classify biologically relevant groups of bacteria and mobile genetic elements (MGE) that can be used in further analyses to track bacteria and MGE within participants and associations between bacteria and MGE and metadata.
2. To explore whether apparent hospital acquisitions of ESBL *E. coli* can be distinguished on a genomic level from community associated *E. coli*.
3. To use bacteria and MGE clusters to determine, which, if any, element is conserved within participants over time and the time scale over which bacteria and MGE change over time within-host.

7.3 Methods

The collection, culture and whole genome sequencing of the isolates analysed in this chapter are described in Chapter 2, Methods, and Chapter 7. The methods of the further analyses covered in this chapter are described here.

The *rhierBAPS* v1.1.0 package in R[1] was used to cluster the core genome pseudosequence into sequence clusters (SCs). Two levels were used and these level 2 clusters used to test associations (see statistical analysis, below). To track putative mobile genetic elements ESBL-gene containing contigs were identified using BLASTn v2.7.0[2] of all contigs against

the SRST2 database[3] of AMR genes (the same database used with ARIBA in the previous chapter). Contigs containing any given ESBL gene were grouped by the ESBL gene they contained (for example, all $bla_{CTXM-15}$ gene-containing clusters were grouped together). Each was group clustered using cd-hit v4.6[4] to produce mutually exclusive ESBL-gene-containing contig clusters for each identified ESBL gene. Henceforth, these clusters will be referred to as ESBL-clusters, for brevity.

In order to attempt to determine the biological significance of the identified ESBL-clusters (i.e. what kind of MGE element they are likely to represent), basic statistics were plotted: number of samples contained within each cluster, length of longest contig in cluster in kbases, length distribution of all contigs is cluster relative to longest contig and distribution of sequence identity compared to the longest contig in the cluster). Presence of insertion sequences (i.e compound transposons), AMR determinants and plasmid replicons were identified for ESBL-cluster representative sequence (as determined by cd-hit i.e one, the longest, for each ESBL-cluster) using BLAST. BLAST default settings were used against the insertion sequence finder (ISfinder) database[5] and the SRST2 database, filtering such that sequence identify was greater or equal to 95%, taking the top hit (as determined by bitscore) for any given location if there were two overlapping hits, and visualising the results in *ggenes* v0.3.2 package in R. To assess lineage association, the ESBL-clusters were mapped back to the core gene phylogeny.

To explore hospital or community associations of any given *E. coli* clade, the location of isolation was first mapped onto the phylogenetic tree; location of isolation was classified as hospital, community, or recent hospital discharge (defined as a date of isolation within 2 weeks of hospital discharge). This latter category was used because it is possible that a patient could acquire an ESBL-E clone in hospital but only be sampled once leaving hospital; using only hospital isolated and community isolated categories could therefore introduce bias. Hospital or community association of each sequence cluster was assessed using a Fisher's exact test of proportion of hospital associated samples (defined as sum of hospital isolated and recent hospital discharge) for the given sequence cluster as compared to proportion of hospital associated samples in the remainder of the samples, with a Bonferroni correction for multiple comparisons. $p < 0.05$ was again considered statistically significant.

To compare within-patient to between-patient conservation of bacteria (as represented by sequence cluster) and ESBL-containing MGE (as represented by the ESBL-clusters) several approaches were taken. Firstly, I assessed whether either sequence cluster or ESBL-cluster were conserved within an individual at all. I hypothesised that any within-patient correlation is likely to be a function of time: samples closer together in time may be more likely to be similar. To assess if this was the case for bacteria, pairwise core genome pseudosequence SNP

distance was was calculated using snp-dists v0.4 (https://github.com/tseemann/snp-dists) for all samples and plotted against the time difference (in days) between samples, within and between patients, and with a smoothed curve fitted using a general additive model (GAM) with cubic splines. Because of significant overplotting, this was also plotted as a 2D density plot. Based on these plots, the within and between patient SNP distances were compared in two post-hoc defined groups binned by time distance between the samples (50 days or less vs. more than 50 days, these cutoffs determined from inspection of the pairwise SNP distance vs time plots), and distributions compared with Kruskal-Wallace tests.

I then compared the within patient temporal clustering of ESBL-clusters and sequence clusters, by estimating the proportion of within-patient samples that contain the same ESBL-cluster or sequence cluster, as a function of time; essentially a temporal auto-correlation function. To estimate this, I considered pairwise comparison of all within-patient samples. For any given time (t) between samples I defined a window of +/-5 days and estimated the probabilities as the number of all within-patient sample pairs in the window $[t-5, t+5]$ that contained the same sequence cluster or ESBL-cluster divided by the total number of all within-patient sample pairs within that time window. Exact binomial confidence intervals for these proportions were generated and probabilities plotted as a function of time. In order to estimate the probability of two samples containing the same sequence cluster of contig-cluster purely by chance, 1000 sample pairs were randomly drawn from all samples with replacement and the proportion of these samples that contained the same sequence cluster or ESBL-cluster calculated.

Finally, to inform the question as to what the likely unit of transmission in this system is, I assessed what was most conserved within patients, in pairwise sample comparison: bacteria (as represented by core gene sequence cluster), ESBL-containing MGE (as represented by ESBL-cluster), or both. Simple proportions in all-against-all pairwise comparison - stratified by whether between-patient or within-patient - were calculated: the proportion of samples that contain the same core gene sequence cluster only, the proportion of samples contain the same ESBL-cluster only, and the proportion that contain both sequence cluster and ESBL-cluster. Proportions were compared between within and between-patient strata in these three groups using Fisher's exact test, with $p < 0.05$ considered statistically significant.

7.4 Results

As described above, in order to test metadata associations of bacterial lineages or MGE, I used several techniques: considering core gene SNP distance between isolates to infer continuous carriage and/or transmission events, and clustering core gene pseudosequences and ESBL-containing contigs into mutually exclusive groups which can then be used to test associations.

Below, I first describe the outcomes of the clustering algorithms used, before describing tests of association of the results with metadata.

7.4.1 Hierarchical BAPS clustering of core gene pseudosequences

The hierarchical BAPS algorithm clustered the core gene alignments into 15 level one (top level) clusters, denoted sequence clusters A-O, and a total of 48 level two (lower level) clusters, denoted sequence clusters 1-48 that were almost exclusively monophyletic and often corresponded closely to the multilocus sequence types (STs, Figure 7.1A). Intracluster pairwise SNP distance varied (Figure 7.1B) but the clusters were often reasonably clonal: SC6, SC8 and SC23, for example (the three largest clusters) had median (IQR) intragroup pairwise SNP distance of 62 (34-97), 326 (18-378) and 18 (11-24) respectively.

7.4.2 ESBL-clusters

The 473 samples contained 486 ESBL genes (Figure 7.2A); 5 genes only occurred once in the collection and so no attempt was made to cluster them. Of the remaining 481 genes, BLAST failed to identify the ESBL-gene containing contig in 2 samples (one in which ARIBA had identified $bla-CTXM-15$ one $bla_{CTXM-27}$), but identified the remaining 479 ESBL genes on 478 contigs, with perfect agreement with ARIBA as to which AMR gene was present in which sample. Only one contig carried two ESBL genes: bla_{CTXM-3} and $bla_{CTXM-15}$; the remaining 477 contigs contained one. The *cd-hit* algorithm grouped the 477 unique contigs into 99 clusters (Figure 7.2B). In total, over 90% of the ESBL-genes (432/479 [90%]) were contained in the 52 largest contig clusters.

The *cd-hit* algorithm selects one member of a cluster (the longest) as the representative. The structure of these representative contigs was explored in an attempt to understand type of MGE they were likely to represent. The length of the representative clusters was very variable, ranging from 1.8kbp to 905.8kbp, with median (IQR) 46.1kbp (11.1-215.5kbp). The other cluster members were usually fragments of these representative contigs with varying sizes - a median (IQR) 60% (36-100%) of the representative contig length - but had high sequence identity, median (IQR) 100.0% (99.7-100.0%) (Figure 7.6 in the appendix to this chapter).

I then explored the insertion sequence (IS), AMR gene and plasmid replicon content of the representative contig for each cluster using BLAST against the SRST2, ISfinder and Plasmidfinder databases (Figures 7.7 and 7.8 in the appendix to this chapter). Every ESBL gene was closely associated with at least one IS, commonly ISEcp1, IS26 and IS903B. IS26 was frequently associated with an apparent 108bp fragment of a *catB4* chloramphenicol resistance

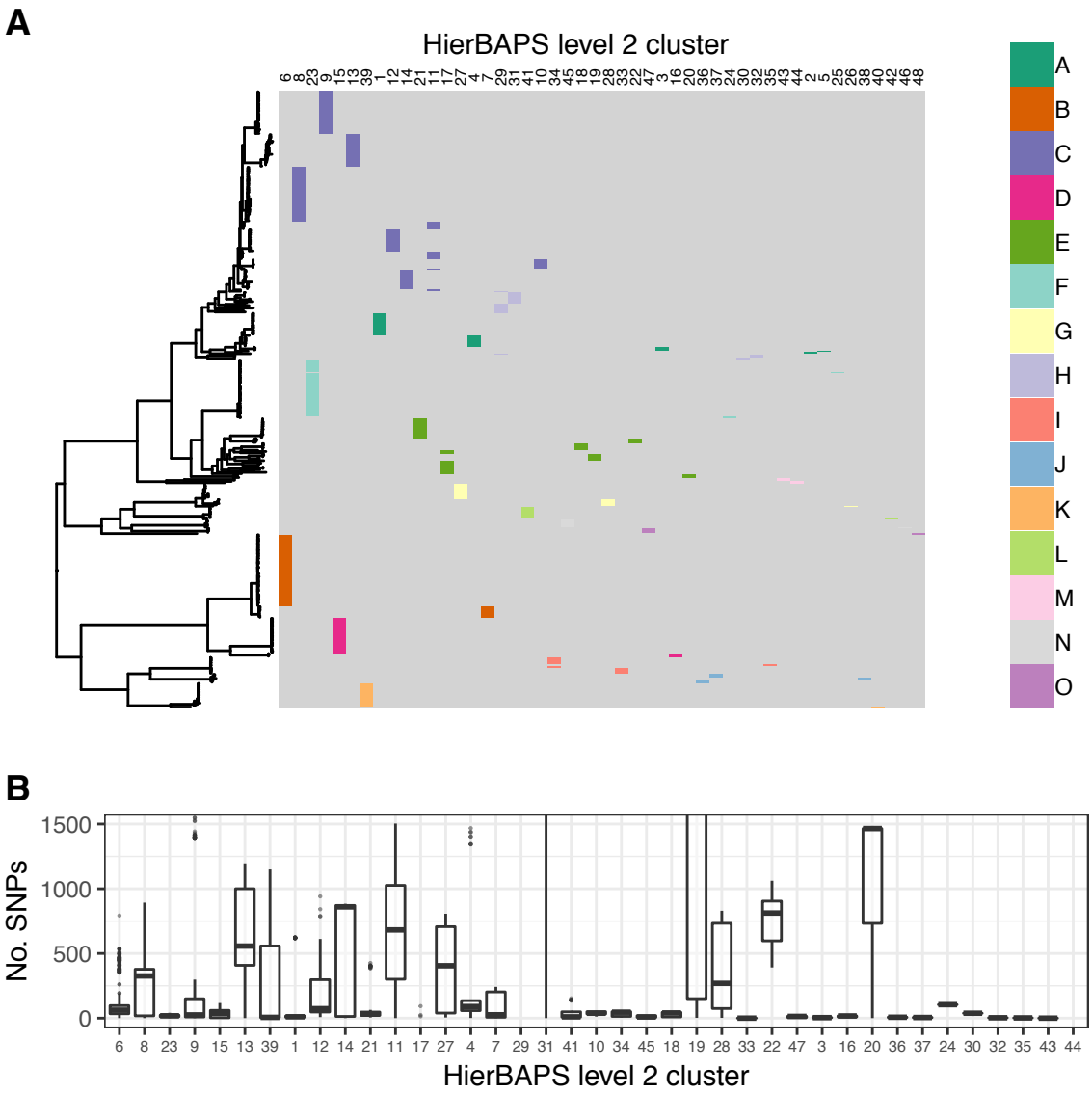

Figure 7.1: A: Core gene hierarchical BAPS clusters mapped back to phylogeny. Heatmap shows level 2 (lower level) with colour denoting level 1 (top level) cluster membership. B: Intracluster pairwise SNP distance for level 2 sequence clusters. Axis restricted to 0-1500 SNPs and as result SC17 (median 6881 SNPs), SC29 (median 2970 SNPs), SC31 (median 2970 SNPs) and SC44 (median 12322 SNPs) boxes are not shown. Boxplots show median and IQR, whiskers show 1.5 times IQR, and outliers are points falling beyond whiskers.

determinant. Some ESBL-genes were associated with particular IS; $bla_{CTXM-15}$, bla_{CTXM-9} and bla_{CTXM-1}, for example were very commonly associated with ISEcp1, whereas bla_{SHV-12} was associated with IS26. ESBL genes were not infrequently associated with other resistance determinants, including commonly $bla_{CTXM-15}$ with bla_{TEM-1}. Plasmid replicons were occasionally identified, including an IncFIB plasmid carrying $bla_{CTXM-15}$ and an IncQ1 plasmid carrying $bla_{CTXM-27}$. It is clear that the same configuration of AMR genes and IS are seen across different contigs, despite a varying backbone, implying historical transposition events. Finally, to assess lineage associations of the identified ESBL-clusters, I mapped the clusters back to the tree, and found that there was a strong lineage association (Figure 7.2C).

7.4.3 Assessing for healthcare-associated lineages

Having clustered bacteria and MGE using *rhierBAPS* and cd-hit respectively, I then mapped the location of sample collection back to the phylogeny and used the hierBAPS SCs to assess for healthcare associated lineages (Figure 7.3). In general, healthcare-associated isolates were distributed throughout the tree and across all SCs, rather than there being a clear hospital-associated lineage. The exception to this was SC23, corresponding to MLST 410, which was more likely to be healthcare associated. When comparing the proportion of healthcare associated samples within each SC to the remained of samples, only SC23 had a statistically significantly increased proportion of healthcare associated samples on Fisher's exact test ($p = 6.3 \times 10^{-4}$, threshold of significance following Bonferroni correction 1.0×10^{-3}), though it was by no means health-facility restricted: 50% (21/42) of SC23 samples were isolated in the community.

7.4.4 Assessing for within-patient conservation of lineage or MGE

To answer the question as to what elements (bacteria or MGE) are conserved within individuals across time I first compared all-against-all pairwise SNP distance between and within patients; first as a scatter plot, and then, because of significant overplotting, as a density plot (Figure 7.4). This suggested that there are a cluster of points close to the origin in the within-patient plot that are not seen in the between-patient plot: before approximately 50 days, there are more similar within-patient isolates than seen in the between-patent isolates. Dichotomising time at 50 days (based on inspection of the density plots) and performing a Kruskal-Wallace test found a statistically significant difference between the before 50 day and after 50 day pairwise SNP distance distribution in the within patient stratum ($p = 0.008$) but not in the between-patient stratum ($p = 0.07$). After 50 days, the distribution of between- and within-patent SNP distances are similar ($p = 0.45$). However it is clear from the plots that even at

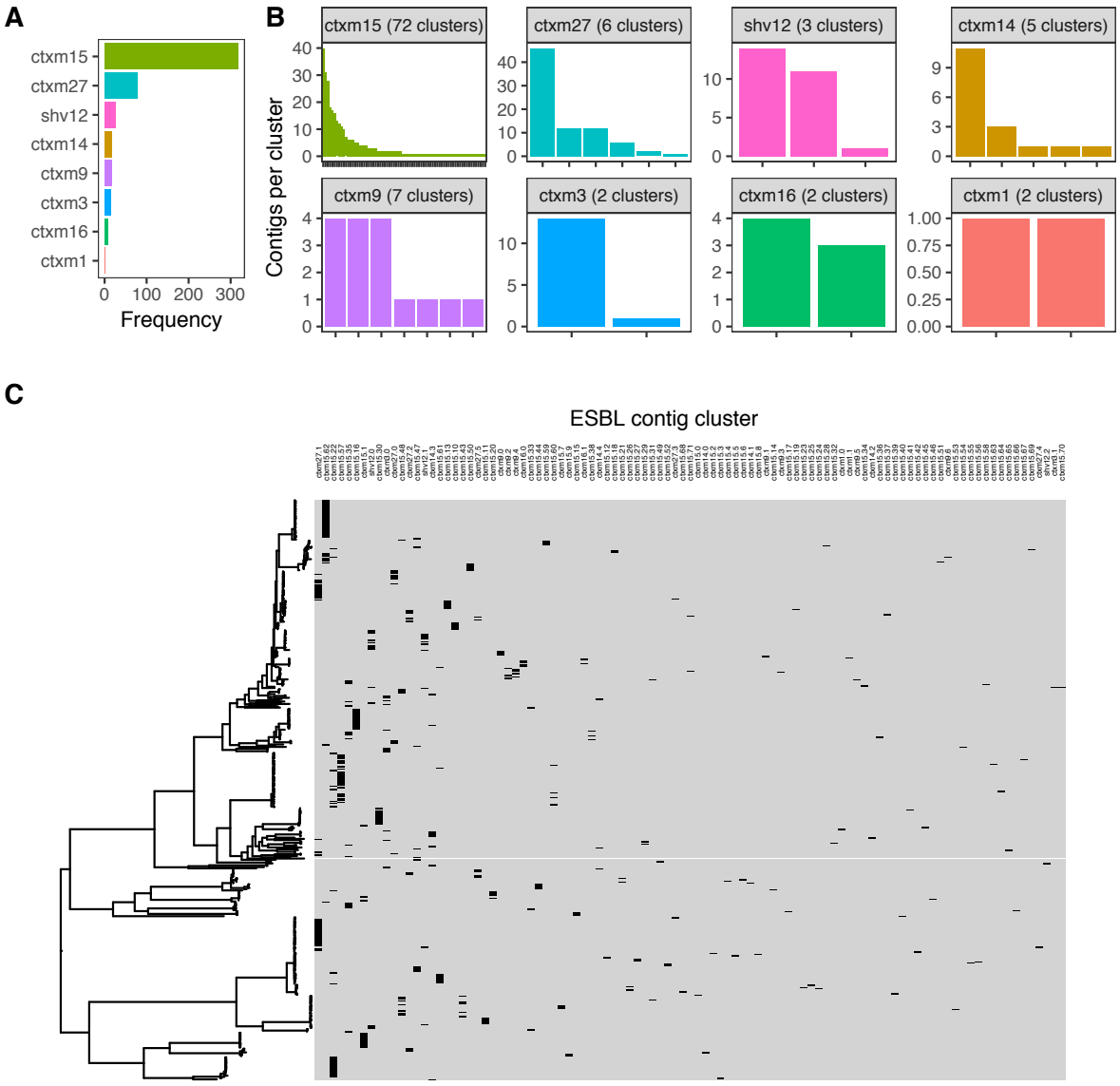

Figure 7.2: A: Frequency distribution of ESBL genes in included samples. B: Frequency distribution of samples per ESBL-cluster, stratified by gene. C: ESBL-cluster membership mapped back to phylogeny.

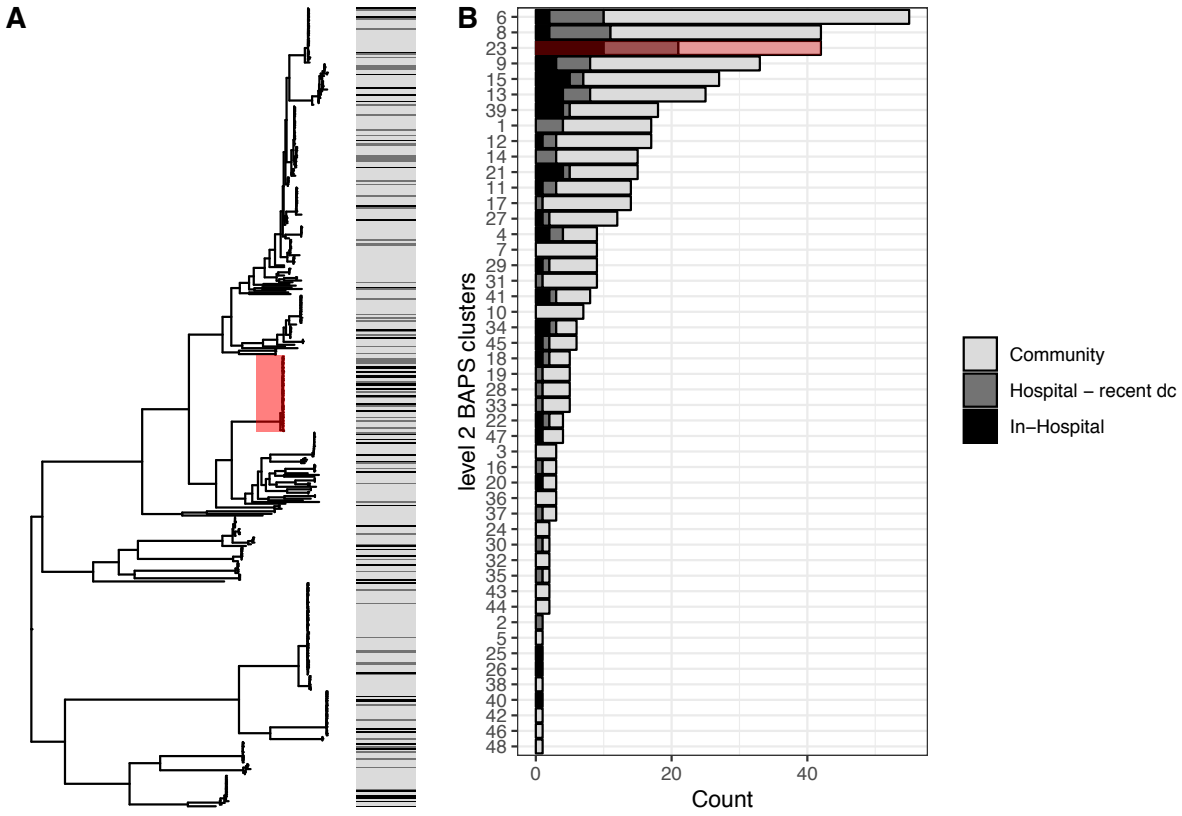

Figure 7.3: A: Location of sample isolation mapped back to phylogeny B: Distribution of location of sample isolation stratified by hierBAPS cluster. In each case, community isolates include those cultured from samples collected on the first day of hospital admissison, in-hospital isolates are from patients who have been hospitalised > 24 hrs and recent discharge isolates are from patients who have been discharged from hospital within the last 2 weeks. Sequence cluster 23, highlighted in red, showed a statistically sigificant association with hospitalisation (see text).

small t and within-participant, there is significant diversity is the SNP distances, and that some isolates close together in time, within the same participant, are only distantly related.

Having confirmed that there is a signal for within-participant temporal conservation of ESBL-E, I then sought to determine if the sequence clusters and ESBL-clusters were similarly conserved over time, and if so, which was the more conserved. The proportion of pairwise within-patient samples that contained the same ESBL-cluster and sequence cluster were significantly greater than would be expected by chance when the time between the samples is less than 35 days for sequence cluster and 32 days for ESBL-cluster (Figure 7.5A). After this time, the lower confidence interval of the sequence cluster and ESBL-cluster curve crossed the proportion of samples that would be expected to be the same by chance, suggesting that, after 35 or 32 days, the chance of any two within-patient samples having the same sequence cluster or ESBL-cluster (respectively) is the same as if the two samples were randomly picked from the data set without regard to patient. The two curves have a very similar appearance; to address the question of which element is most conserved within an individual - sequence cluster, ESBL-cluster, or both - I performed an all-against-all pairwise comparison of which elements were conserved (Figure (Figure 7.5C), and found that only ESBL-cluster and sequence cluster together are conserved within patients at a significantly greater proportion than between patients ($p = 1.1 \times 10^{-12}$).

7.5 Discussion

In this chapter, I have used clustering algorithms on WGS data to group bacteria and putative MGE. I have then assessed associations of these groups with metadata to attempt to better understand the determinants of ESBL-E carriage in Blantyre. I draw several conclusions.

First, I looked for healthcare associated *E. coli* sequence clusters. Healthcare associated bacteria were not associated with a particular sequence cluster, and were spread throughout the phylogeny rather than apparent hospital acquisitions being restricted to a single clade or clone. The exception to this was SC23, which corresponds to ST410, and was more likely to be healthcare associated. This could be consistent with the hypothesis that it is a recently arrived high-risk clone, which may be, at least initially, hospital-associated. Even so, it is clearly not hospital restricted, with half of the ST410 isolates being isolated from the community.

I showed in Chapter 5 that there was an increase in colonisation prevalence of ESBL-E in study participants following admission to hospital, particularly in the antibiotic-exposed. Despite this it seems as though the genomic diversity of ESBL *E. coli* apparently acquired in hospital is largely the same as *E. coli* isolated in the stool of community members. This result could be explained by two hypotheses: first, that these are true transmission events that are

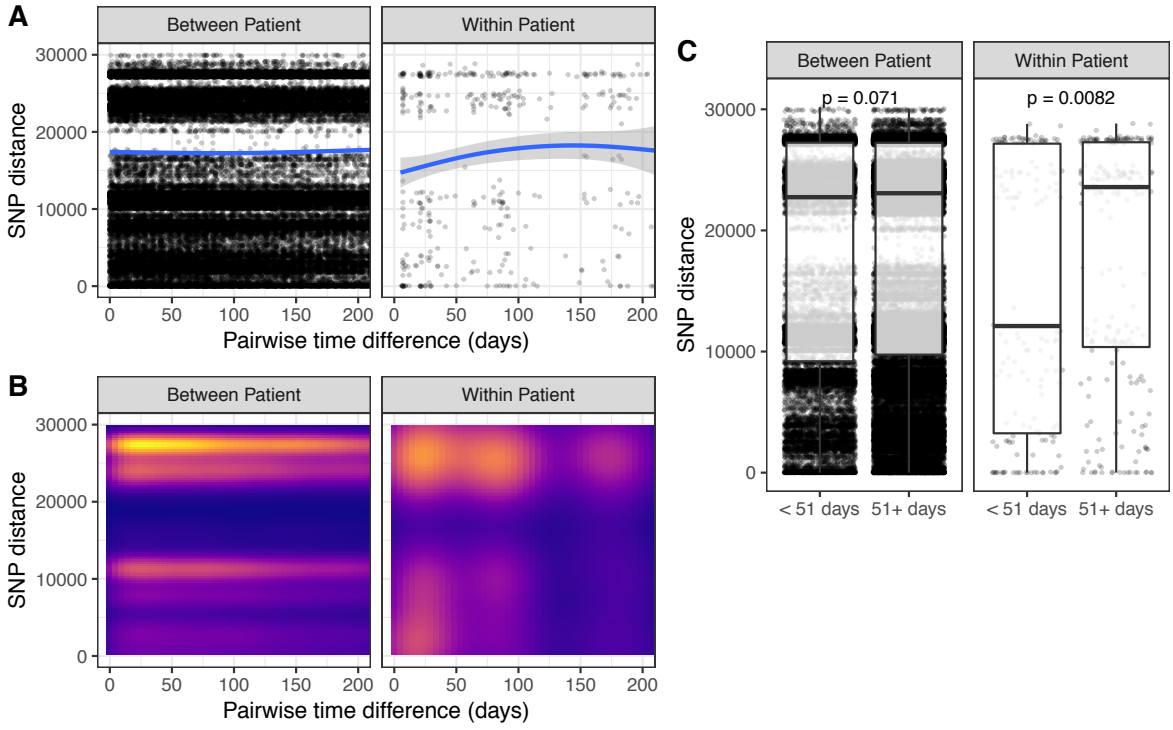

Figure 7.4: Within and between participant pairwise SNP distances. A: Scatterplot of pairwise SNP distances as a function of time with GAM model fitted curve. B: Pairwise SNP distance as function of time as a 2D density plot, showing cluster of isolates close to origin that are close together in time and SNP-distance. C: Pairwsise SNP distance distribution before and after 50 days, within and between patients, showing statistically significant descreas ein pairwise SNP distance within patients before 50 days. After 50 days, between- and within- patient distributuions are similar.

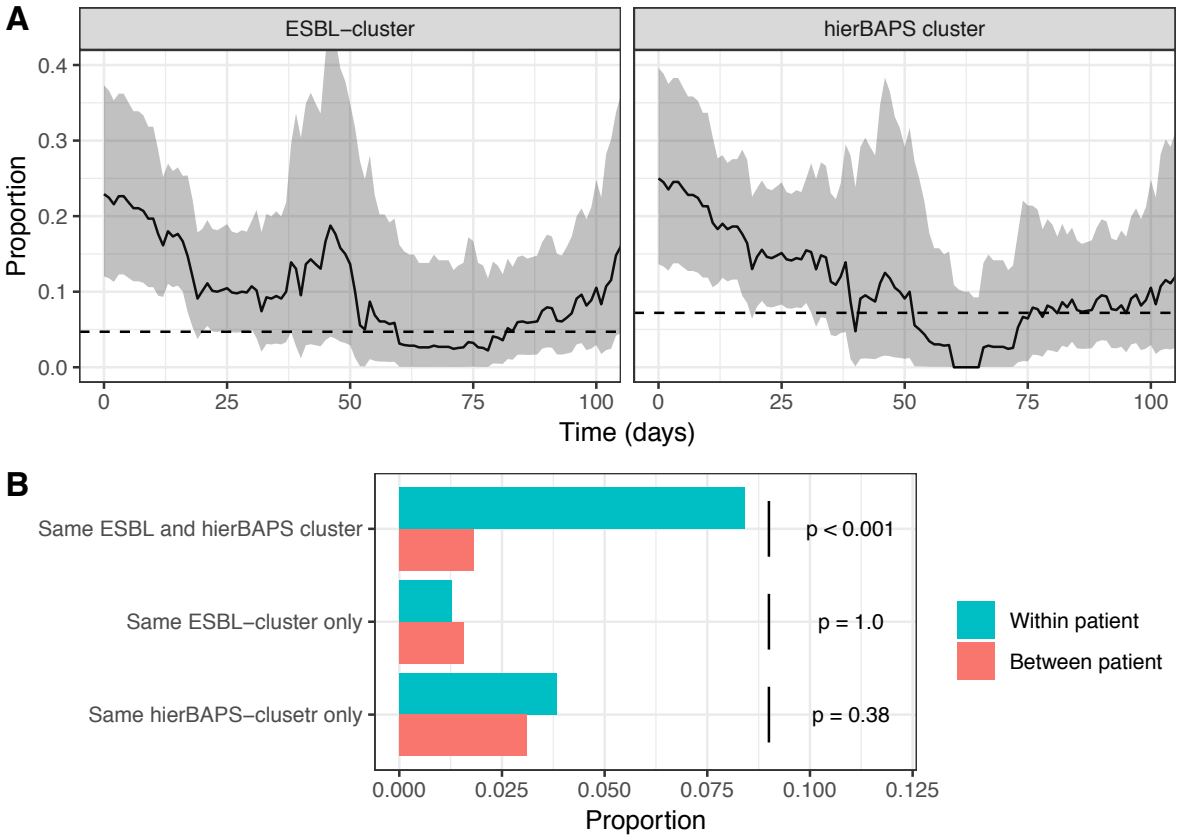

Figure 7.5: Probability of any two samples from within a given participant containing the same ESBL-cluster (A, left panel) or being a member of the same hierBAPS cluster (A, right panel). Time is windowed at +/- 5 days around the time indicated on the x axis. Dotted line is the probability that two samples would belong the the same group by chance, constructed by randomly sampling 1000 sample pairs. B: proportion of samples that contain the same herBAPS cluster alone, or ESBL-cluster alone, or both, demonstrating that the ESBL cluster-hierBAPS cluster pairing is the most conserved of the three.

occurring within the hospital, and that the diversity of ESBL *E. coli* within the hospital is the same as the community; or, second, that these "hospital acquisitions" are actually minority variant *E. coli* present in the microbiota (and therefore acquired in the community) at a low abundance and hence not detected by culture, and enriched for with antimicrobial exposure in hospital. Distinguishing between these two hypotheses is important as they would each require a different intervention: hospital infection control in the former case, or strategies to protect the microbiota from the deleterious effect of broad spectrum antimicrobials (such as pre- or probiotics, or oral β-lactamases) in the latter. It is of course possible that both hypotheses are true; they are not mutually exclusive. The genomic data are consistent with both, perhaps with a suggestion from the hospital association of ST410 that there is at least some true hospital acquisition. The longitudinal models I present in the next chapter can help to shed more light on the mechanism of increase of ESBL-E prevalence following admission, by quantifying the relative effects of hospital admission versus antimicrobial exposure.

By forming sequence clusters and ESBL-clusters, I was able to demonstrate that both bacteria and MGE are conserved together, within-patient, over time, whereas bacteria and MGE alone are not. Some previous longitudinal studies of ESBL-E found that *E. coli* STs tended to vary over time but that in many cases ESBL gene and plasmid replicons remained the same, which could be due to a conserved MGE transferring between bacteria[6]. Given my findings, this is unlikely to be the case. Though not directly addressed in this study it is possible to speculate therefore that the unit of transmission of ESBL between patients is likely to be the bacterium rather than, for example, horizontal gene transfer of ESBL genes on plasmids or transposons. The within-participant association of sequence cluster with ESBL-cluster suggests that MGE are reasonable conserved within bacteria, at least on the timescale of the study. Mapping the ESBL-clusters back to the *E. coli* phylogeny also shows some lineage association, which is consistent with this. The within-patient correlation of SC and ESBL-cluster lasts only for 32-35 days; two samples from a single patient more than 35 days apart are as likely to contain the same SC/ESBL cluster as two samples from two different patients. This implies either an exogenous re-exposure or some other endogenous mechanism whereby the dominant ESBL strain is replaced by a minority variant from within the microbiota. Again, the longitudinal models in the next chapter will investigate this further.

This analysis also is suggestive that there is significant within-participant diversity of ESBL *E. coli*. At a maximum (i.e. with samples that are days apart), only around 20% of within-patient samples contain the same SC/ESBL-cluster. This is many times more than would be expected by chance, but still implies that at any time point there is significant diversity of ESBL *E. coli* strains, that have been missed by only taking forward one colony pick for sequencing. Without multiple picks from one time point, however, it is not possible to fully define within-host diversity. Limited data are available to define this diversity; one study that examined this

question found that it varied between individuals, and that some individuals harboured widespread diversity of STs and ESBL genes whereas some did not[7].

7.5.1 Limitations

Only one colony pick from the ESBL selective media was taken forward for sequencing. In effect, we have randomly sampled one strain from all available strains at any give time point. This is likely to result in an underestimation of the extent to which strains persist within the individual over time, as strains that are present (but not sampled) are classed as absent in the above analysis.

Community members are under represented in this dataset; I have classified isolates as community or hospital associated, but there may have been healthcare contacts (especially in arm 1 and 2 of the study) which have not been recorded which would mean that isolates that were actually healthcare associated were classed as community associated. Healthcare contact is probably less likely in the true community arm of the study (arm 3).

I have attempted to overcome the difficulty of reconstructing MGE by defining ESBL-clusters as a proxy for MGE, but this approach has limitations. Some of the assembled contigs are short and likely represent transposons; the same transposons have likely inserted into multiple plasmids in the past and as such, these short contigs may cluster with other sequences that would be seen to be very different, were a full assembly available. In addition, the biological significance of these ESBL-clusters is not clear. It is not possible to say with certainty what they represent (e.g. plasmid) as they are only fragments. Nevertheless, the fact that I have seen within-patient associations of the ESBL-clusters lends some support to their use, as erroneous clustering would be expected to bias any associations towards to null.

7.6 Conclusions and further work

In this chapter, I have shown that apparent hospital acquired ESBL *E. coli* are largely as diverse as community isolates. This suggests either widespread mixing of strains between the hospital and community and/or an enrichment effect as study participants are admitted to hospital and exposed to antimicrobials and carried but undetected ESBL *E. coli* become detectable by culture. By clustering bacteria and putative MGE I find that the bacteria-MGE combination is the element that is most conserved within-participant over time, but that after 32-25 days this signal dissipaters, suggestive of a constant re-exposure or some other replacement mechanism. Many questions remain unanswered and further work is planned:

- Shotgun metagenomic sequencing of stool would allow testing of the competing acquisition and unmasking hypotheses of rapid increase in ESBL-E prevalence by defining the microbiota and total AMR gene content pre-, during and post- antimicrobial exposure. This would also provide an opportunity to explore the role of the microbiota to colonisation resistance to ESBL-E, and help to define the within-host ESBL diversity at any time point.
- Long read sequencing would allow a proper characterisation of the MGE that carry ESBL genes in the Malawian context, giving the resolution necessary to truly track MGE within and between patients and strains.
- Short-read sequencing of the *Klebsiella pneumoniae* isolates from this study (as discussed in the previous chapter) would allow a comparison between the mechanisms of AMR and MGE prevalent in this species as compared to *E. coli*, and assess the extent to which horizontal gene transfer between the two is driving ESBL spread in Blantyre.
- Sequencing one *E coli* from each sample in which *E coli* was identified but has not yet had a representative sequenced will give more power to detect metadata associations - in particular by expanding the number of isolates from true community members, arm 3 of the study, under represented in this dataset.
- Finally, incorporating the resolution afforded by sequencing into a longitudinal modelling approach may provide new insights in to the dynamics of ESBL-E carriage. This is taken up in the next chapter.

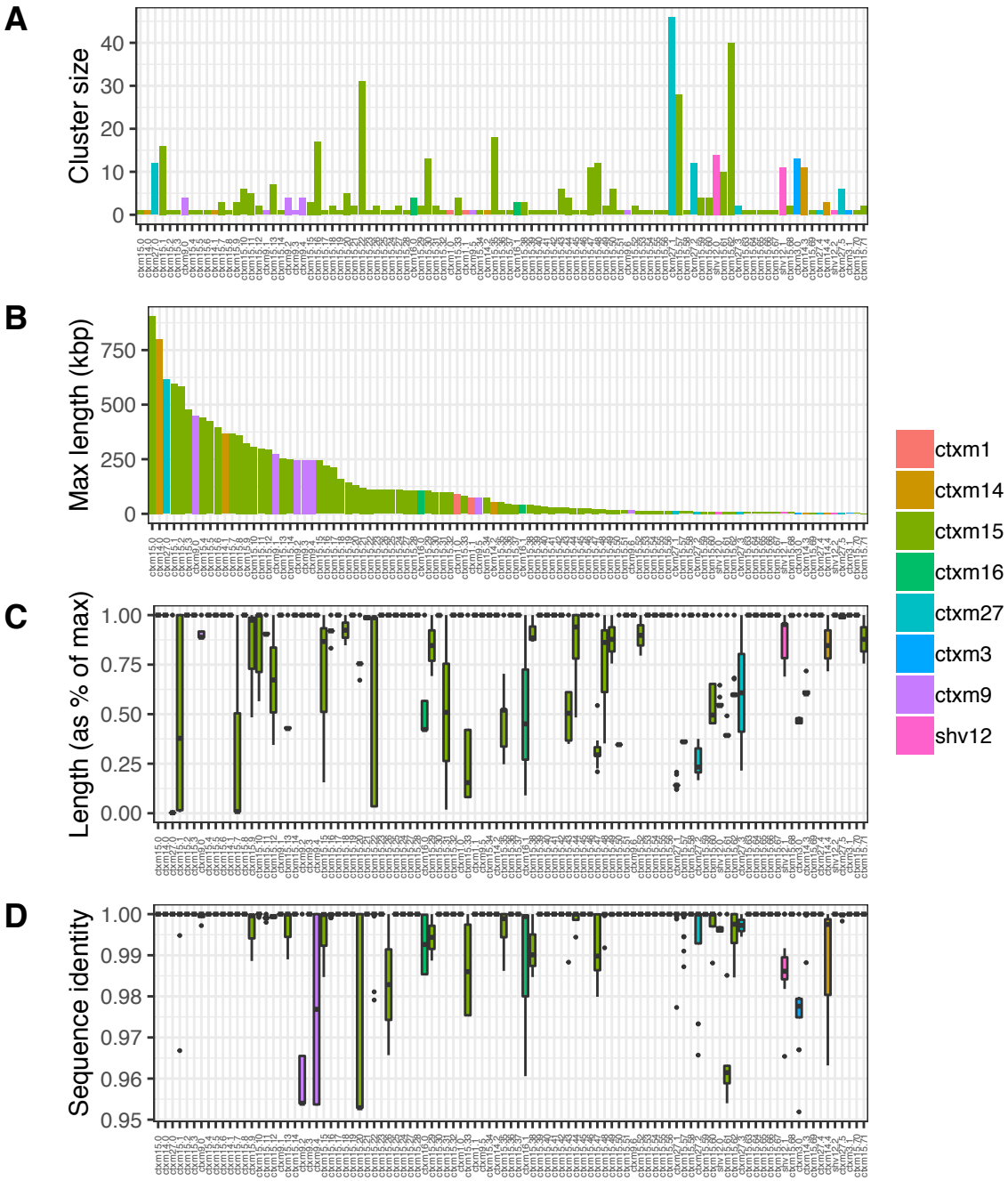

Figure 7.6: Summary statistics for 99 ESBL-containing contig clusters as deterined by *cd-hit*. A: Number of contigs per cluster. B: Length (kbp) of longest sample in each cluster. This is defined as the cluster representative sample by *cd-hit* to which all other samples are compared for the purposes of length and sequence identity. C: Distribution of contig lengths by cluster expressed as a proportion of longest contig length. D: Distribution of sequence identity of cluster members compared to representative member, by cluster.

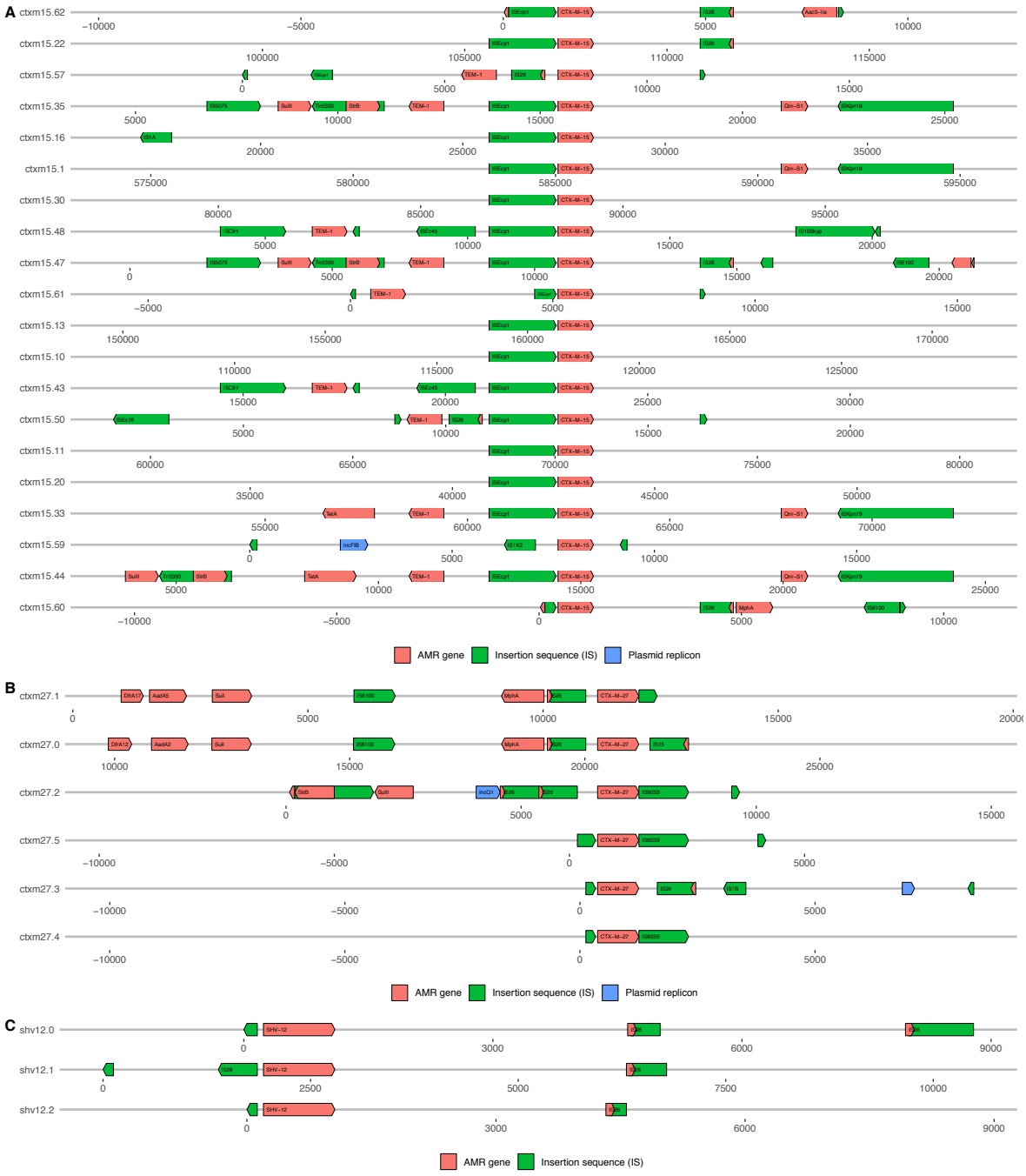

Figure 7.7: AMR genes, insertion sequences (IS) and plasmid replicons identified in the representative contig of each contig cluster, stratified by by ESBL gene and ordered by number of samples of cluster. IS26 is very frequently associated with a 108bp fragment of *catB4* chloramphenicol resistance gene, shown as a red fragnemt within the green IS26 element. A: *blaCTXM15*, B: *blaCTXM27* , C: *blaSHV12*. Plots show furthest IS/AMR gene or plasmid replicon up to +/- 10,000bp from the ESBL gene of interest.

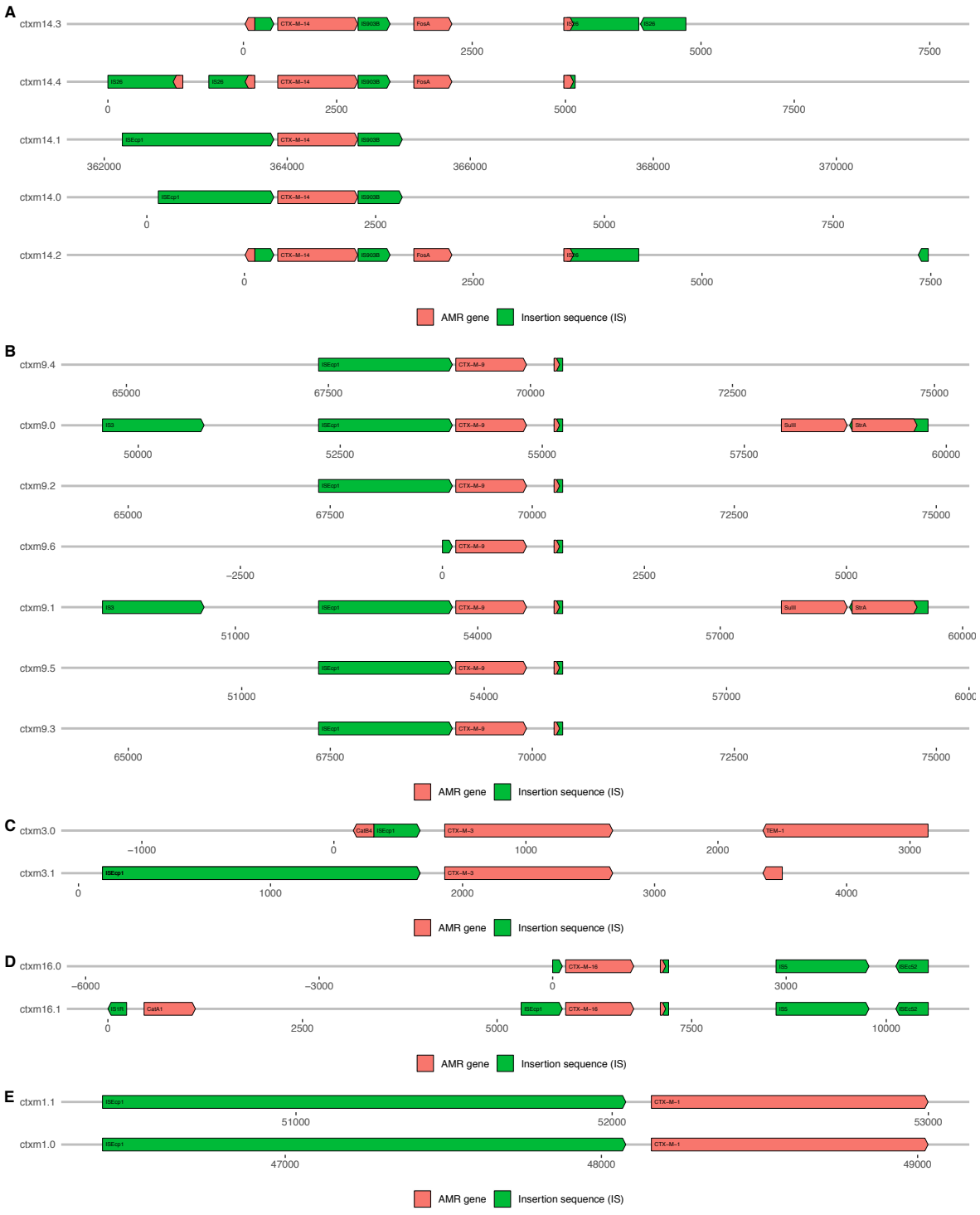

Figure 7.8: AMR genes, insertion sequences (IS) and plasmid replicons identified in the representative contig of each contig cluster, stratified by by ESBL gene and ordered by number of samples in cluster. IS26 is very frequently associated with a 108bp fragment of *catB4* chloramphenicol resistance gene, shown as a red fragnemt within the green IS26 element. A: *blaCTXM14*, B: *blaCTXM9* , C: *blaCTXM3*, D: *blaCTXM16*, E: *blaCTXM1*. Plots show furthest IS/AMR gene or plasmid replicon up to +/- 10,000bp from the ESBL gene of interest.

References

1 Cheng L, Connor TR, Siren J *et al.* Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. *Molecular Biology and Evolution* 2013;**30**:1224–8. doi:10.1093/molbev/mst028

2 Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *Journal of Molecular Biology* 1990;**215**:403–10. doi:10.1016/S0022-2836(05)80360-2

3 Inouye M, Dashnow H, Raven L-A *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine* 2014;**6**:90. doi:10.1186/s13073-014-0090-6

4 Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9. doi:10.1093/bioinformatics/btl158

5 Siguier P, Perochon J, Lestrade L *et al.* ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research* 2006;**34**:D32–6. doi:10.1093/nar/gkj014

6 Duijkeren E van, Wielders CCH, Dierikx CM *et al.* Long-term Carriage of Extended-Spectrum β-Lactamase–Producing Escherichia coli and Klebsiella pneumoniae in the General Population in The Netherlands. *Clinical Infectious Diseases* 2018;**66**:1368–76. doi:10.1093/cid/cix1015

7 Stoesser N, Sheppard AE, Moore CE *et al.* Extensive Within-Host Diversity in Fecally Carried Extended-Spectrum-Beta-Lactamase-Producing Escherichia coli Isolates: Implications for Transmission Analyses. *Journal of clinical microbiology* 2015;**53**:2122–31. doi:10.1128/JCM.00378-15