1 **Supplementary Material for: Application of combined genomic and transfer**

2 **analyses to identify factors mediating regional spread of antibiotic resistant**

3 **bacterial lineages**

4 **Table of Contents**

Supplementary Methods

Genomic Analysis

We extracted DNA from single colonies grown overnight in brain-heart infusion broth at 37°C using the PowerSoil DNA extraction kit (Qiagen). Sample-specific barcoded libraries were prepared using the Nextera DNA Flex Library Preparation Kit and sequenced using Illumina MiSeq instruments at the University of Michigan Microbial Systems Molecular Biology Laboratory. Sequences were processed to generate a final alignment file with high-confidence single-nucleotide variants (SNVs).

The quality of raw sequencing reads quality was assessed using FastQC [1], and adapter sequences and low-quality bases were removed using Trimmomatic [2]. *De novo* genome assemblies were generated from trimmed sequencing reads using SPAdes [3], and one representative of each ARO species was selected as the reference genome for downstream alignment. Genome assemblies were evaluated using QUAST [4], and those with high quality further had their contigs ordered and oriented relative to finished reference genomes using ABACAS [5]. Single nucleotide variants (SNVs) were identified by mapping reads the reference genome using bwa [6], removing polymerase chain reaction duplicates with Picard [7], and calling variants with SAMtools and bcftools [8]. Prophage, mobile genetic elements and recombinant regions were identified using PHASTER, RAST annotation and Gubbins, [9–11], and excluded from the final SNV alignment.

Maximum likelihood phylogenetic reconstruction was performed by applying RAxML to sequences of variant positions within each ARO clade [12]. Phylogenetic analyses were performed by modeling SNVs with a Generalized Time Reversible model (GTR). Confidence of branches were quantified by running up to 1000 bootstrap replicates, with the precise number based upon convergence of bootstrap confidences. Final tree topology was determined by midpoint rooting as implemented in phytools [13].

To remove isolates that putatively resulted from within nursing facility (NF) transmission, we used a phylogenetic approach to identify and remove isolates collected during follow-up visits that were closely related to isolate(s) collected at other patients' time of admission, or on an early date. The effect of the removal of these potential transmission isolates was visualized by violin plots (**Figure S5**).

Multi-locus sequence typing (MLST) of Whole Genome Sequences

From whole genome sequences, we used ARIBA for *in silico* MLST analysis using the typing schemes from PubMLST [18]. Seven housekeeping loci were used to determine the sequence type of each MRSA, VREfm, and VREfc isolate; MRSA: *arcC*, *aroE*, *glpF*, *gmk*, *pta*, *tpi*, *yqiL*; VREfm: *atpA*, *ddl*, *gdh*, *purK*, *gyd*, *pstS*, *adk*; VREfc: *gdh*, *gyd*, *pstS*, *gki*, *aroE*, *xpt*, *yqiL*; eight housekeeping genes were used for CipREc in accordance with the Institute Pasteur scheme: *dinB*, *icdA*, *pabB*, *polB*, *putP*, *trpA, trpB, uidA.*

Population Structure Analysis

To measure the genetic differentiation between ARO isolates in different NFs, we used the Fsp method to calculate the degree of gene flow using whole-genome sequencing data [19]. Briefly, each SNV position in the core genome alignment was treated as a gene with four possible alleles. The difference in probabilities of alleles shared between NFs and within NFs was used to infer population homogeneity between NFs. The smaller the Fsp value, the more similar the two bacterial populations; the larger the Fsp value, the more isolated two populations are from each other.

Permutation Test

To evaluate whether isolates collected from patients recently exposed to the same acute-care hospitals (ACHs) or admitted to the same NFs were more likely to be phylogenetically clustered than expected by chance, we devised a permutation test with minor modifications from the procedure described in [14]. Briefly, the number of subtrees containing at least two isolates from the same facility was calculated for the original phylogeny, and the result was then compared to 1,000 permuted phylogenies where facility assignment of each isolate was shuffled randomly. An empirical P value was computed by determining the fraction of permuted phylogenetic clustering that was greater than the original statistic. ACHs with five or fewer isolates were collectively assigned as "999".

Nursing Facility Connectedness by Geographical Distance and Patient Sharing

The geographical distance between NFs was calculated using R package "geosphere" with the Haversine formula [15]. Patient transfer data was visualized with "igraph" [16].

The dissimilarity in patient transfer pattern between each pair of NFs was measured by Kullback-Leibler divergence using the "KL.plugin" function in R package "entropy" v1.2.1 [17]. The Kullback-Leibler method measures the divergence/dissimilarity between two probability distributions. As the proportion of patients discharged from a set of ACHs constitutes a set of frequencies, we adapted this method to calculate the dissimilarity of patient transfer patterns between each pair of NFs.

Supplementary Figures
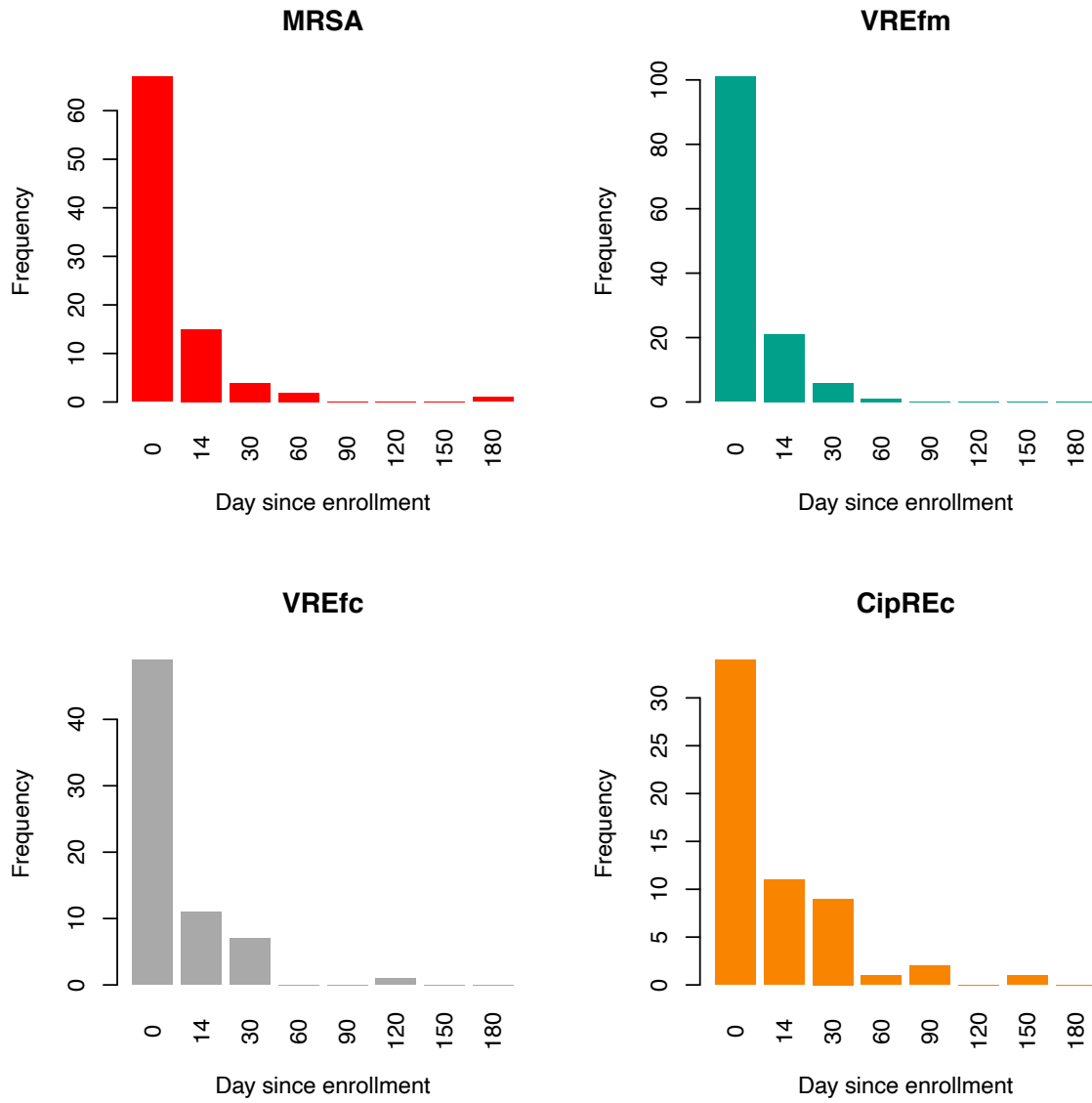
Figure S1. Time of isolation for each antibiotic-resistant organism.
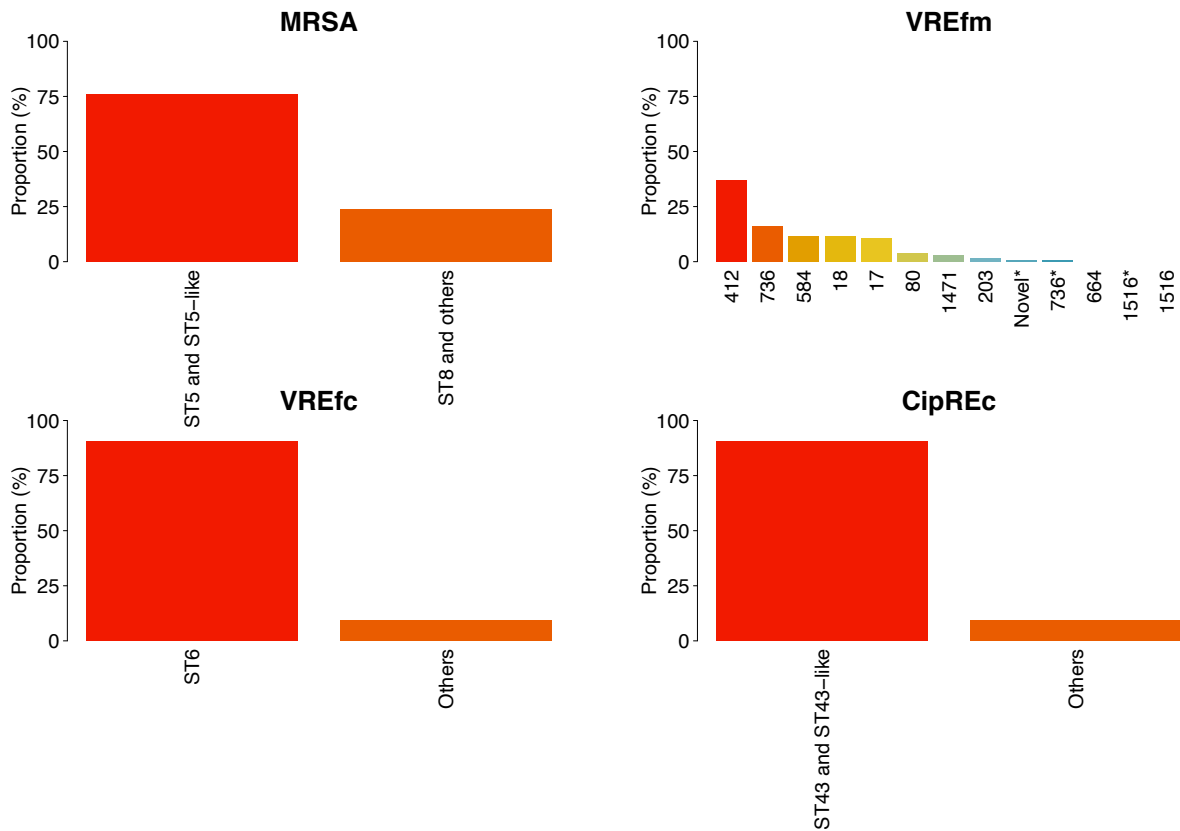
Figure S2. Proportion of sequence types (STs) of isolates subjected to whole-genome sequencing. For MRSA, VREfc and CipREc we focused our analyses to isolates belonging to or closely related to major lineages (MRSA: ST5, VREfc: ST6, CipREc: ST43, a subclade of ST131). Despite the presence of many different STs for VREfm, isolates are genetically closely related thus all VREfm isolates were analyzed regardless of ST. * denotes some uncertainty in ST designation.
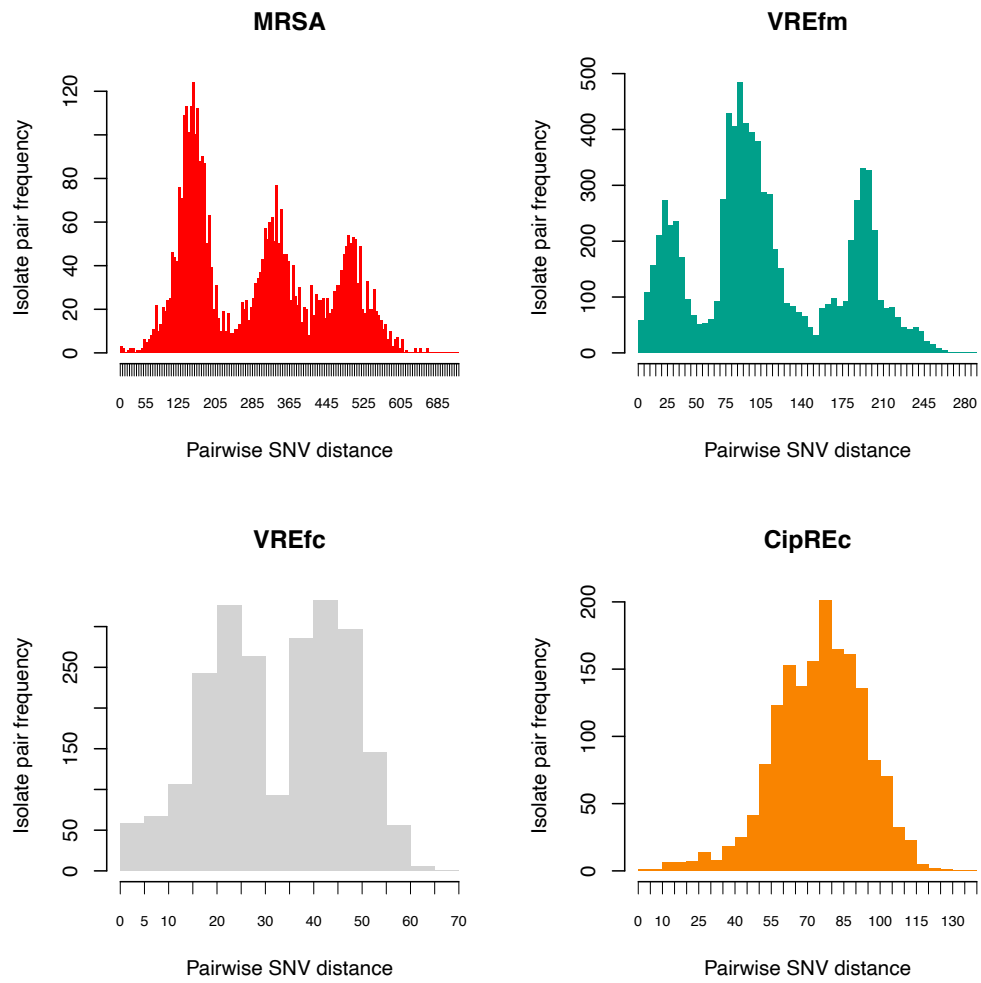
125
126
127 **Figure S3. Pairwise single nucleotide variant (SNV) distance between every pair**
128 **of analyzed isolates within the same ARO species.**
129

MRSA (ST5 and like)

VREfm (clade A)

Current

Previous

VREfc (ST6)

CipREc (ST43 and like)
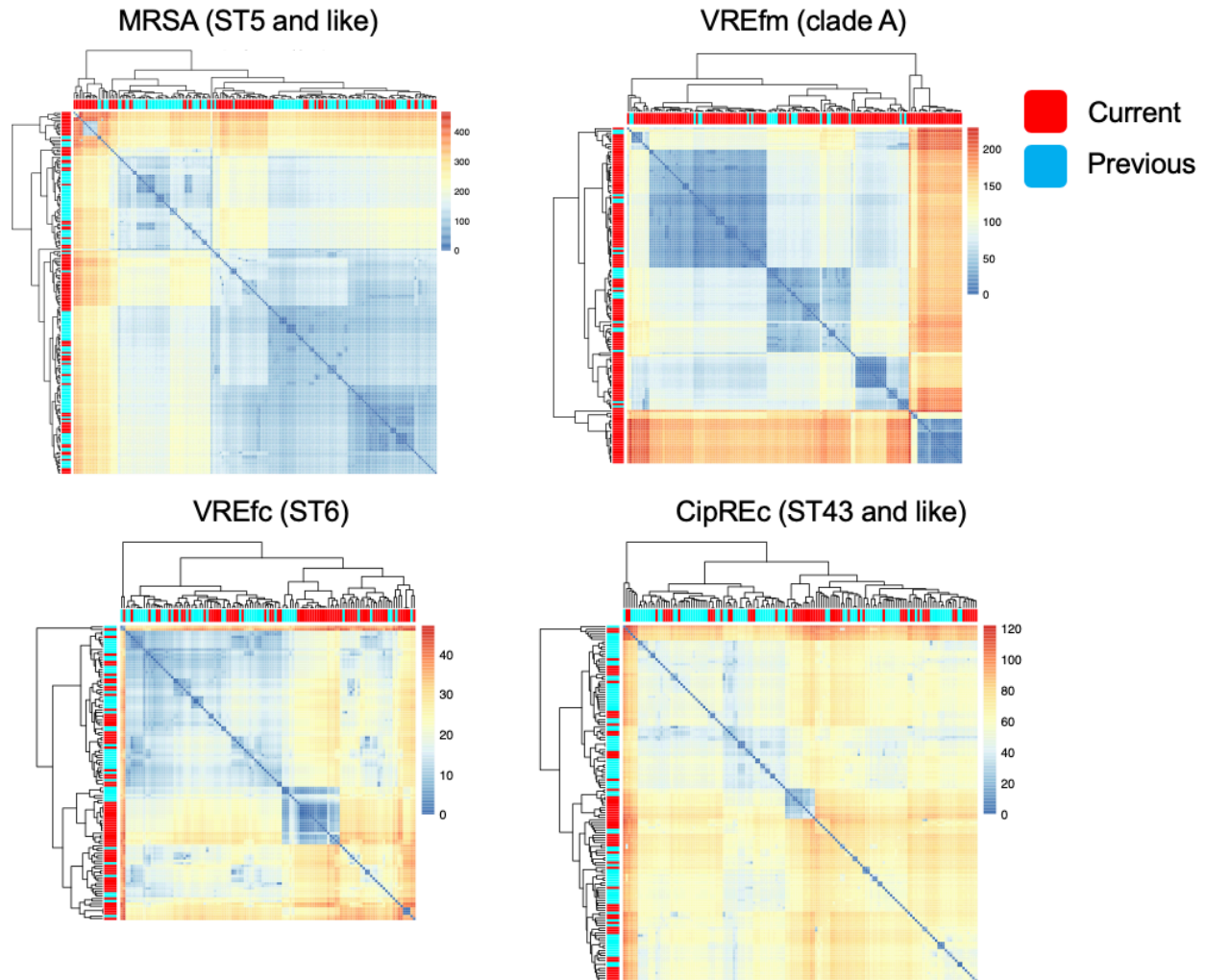
130
131
132 **Figure S4. A heatmap representing the pairwise genetic distance between**
133 **isolates** from current (annotated as red, 6 NFs, 2013 - 2015) and a previous study
134 (annotated as blue, 12 NFs, 2010 – 2013). Samples are clustered using a hierarchical
135 method implemented in R package (pheatmap). Colour key indicates pairwise genetic
136 distance.
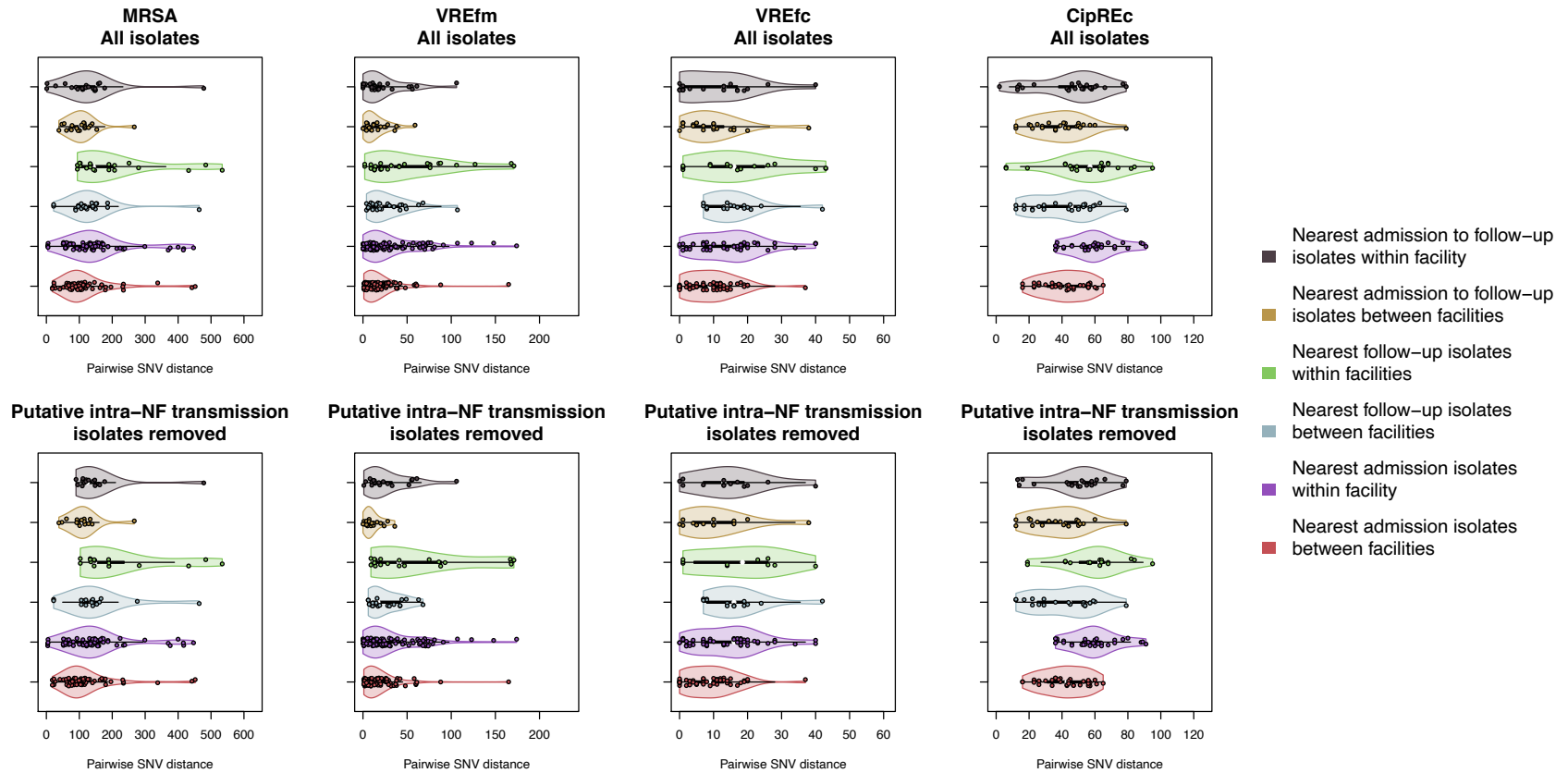
Figure S5. Violin plots showing the distribution of single nucleotide variant (SNV) distance among pairs of nearest neighbour isolates between and within nursing facility and collected at admission and during follow-up visits using all isolates (top panel) and only admission isolates and follow-up isolates that did not have a closely related neighbour within the same facility (bottom panel).

1

A

B
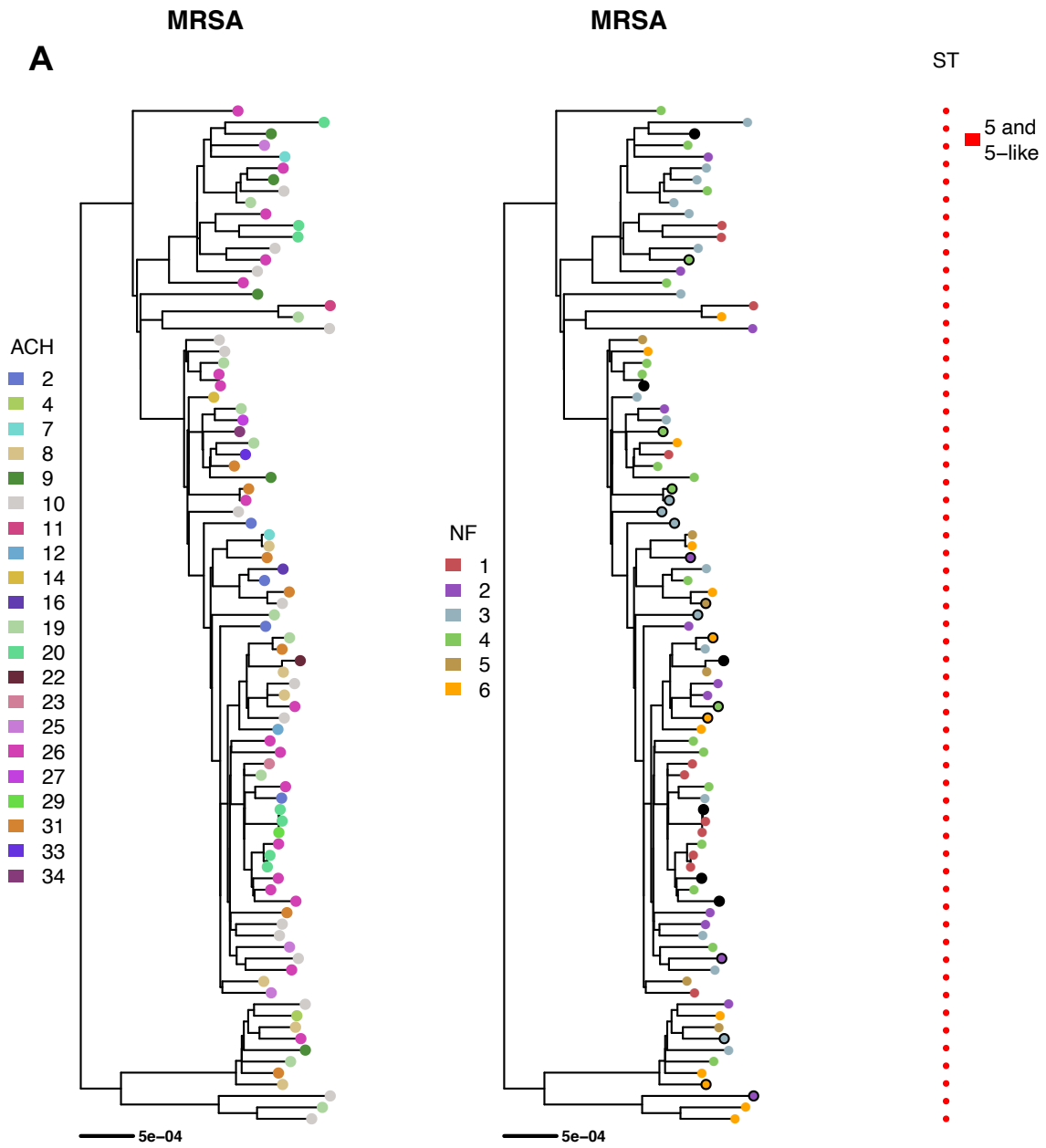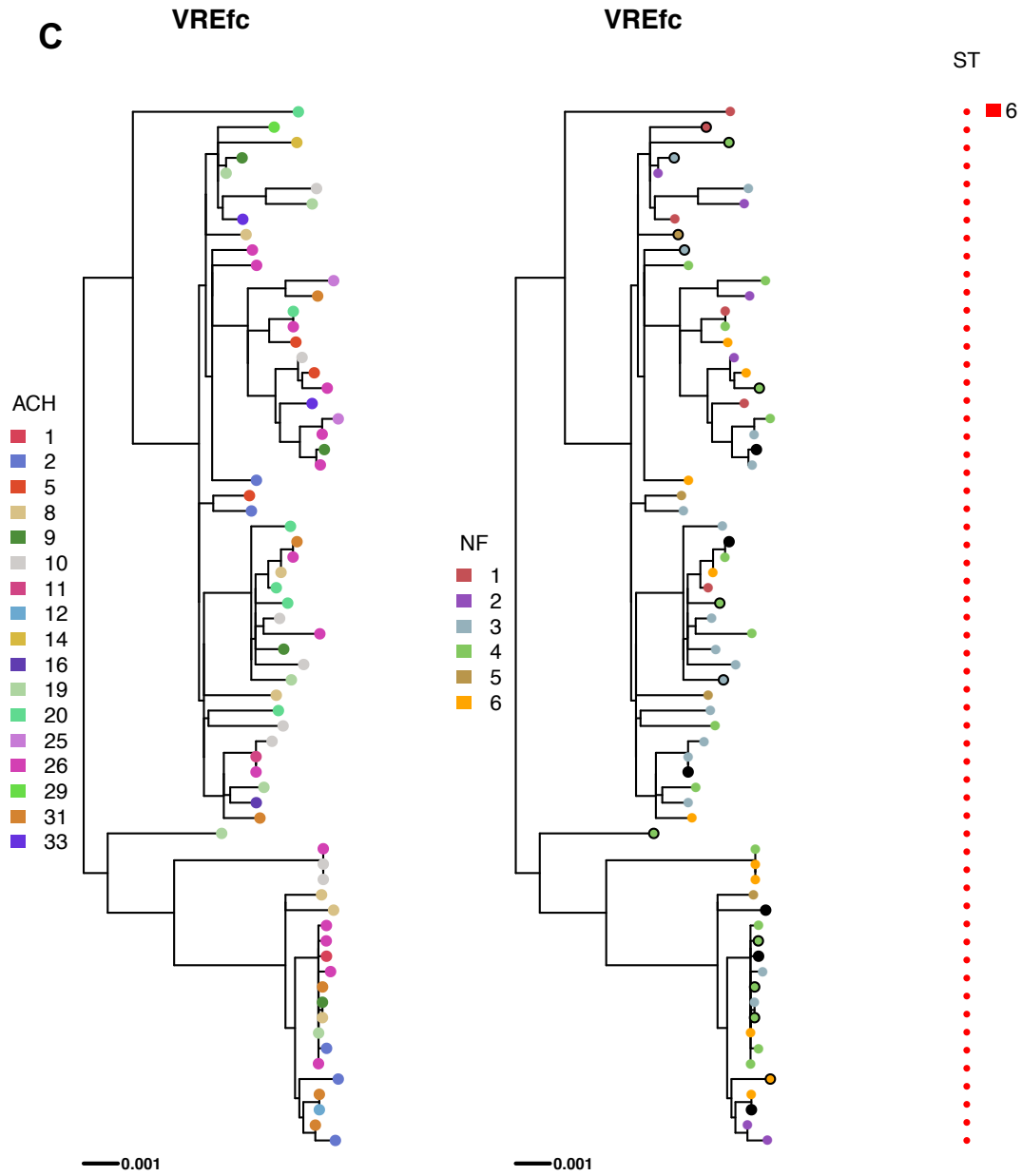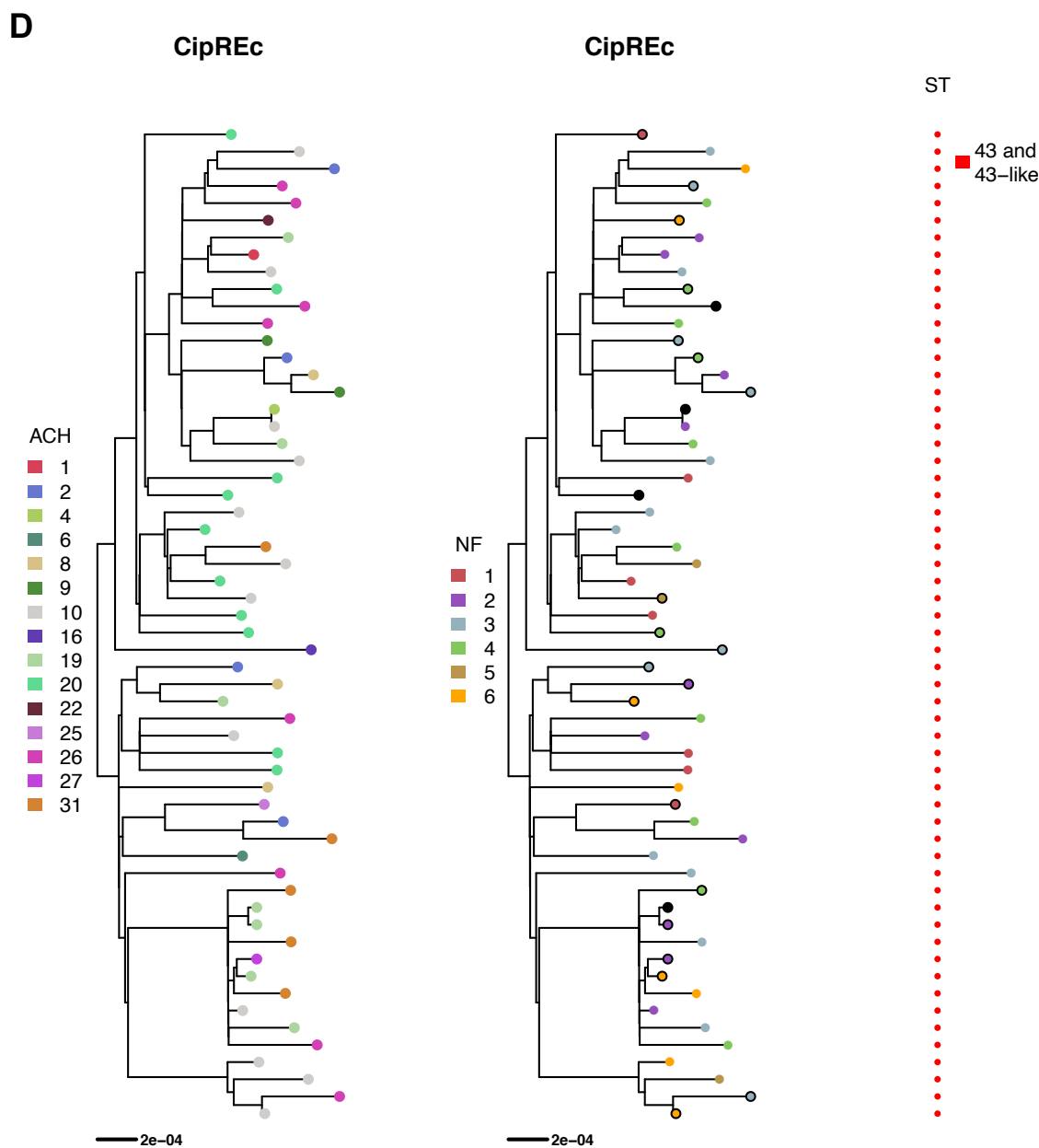
Figure S6. Phylogenetic tree of each ARO. A: MRSA; B: VREfm; C: VREfc; D: CipREc. Each isolate is labeled by patient's most recent acute-care hospital (ACH) exposure (left panel) or patient's nursing facility (NF) residence at the time of isolate detection (right panel). In the NF panel, isolates collected a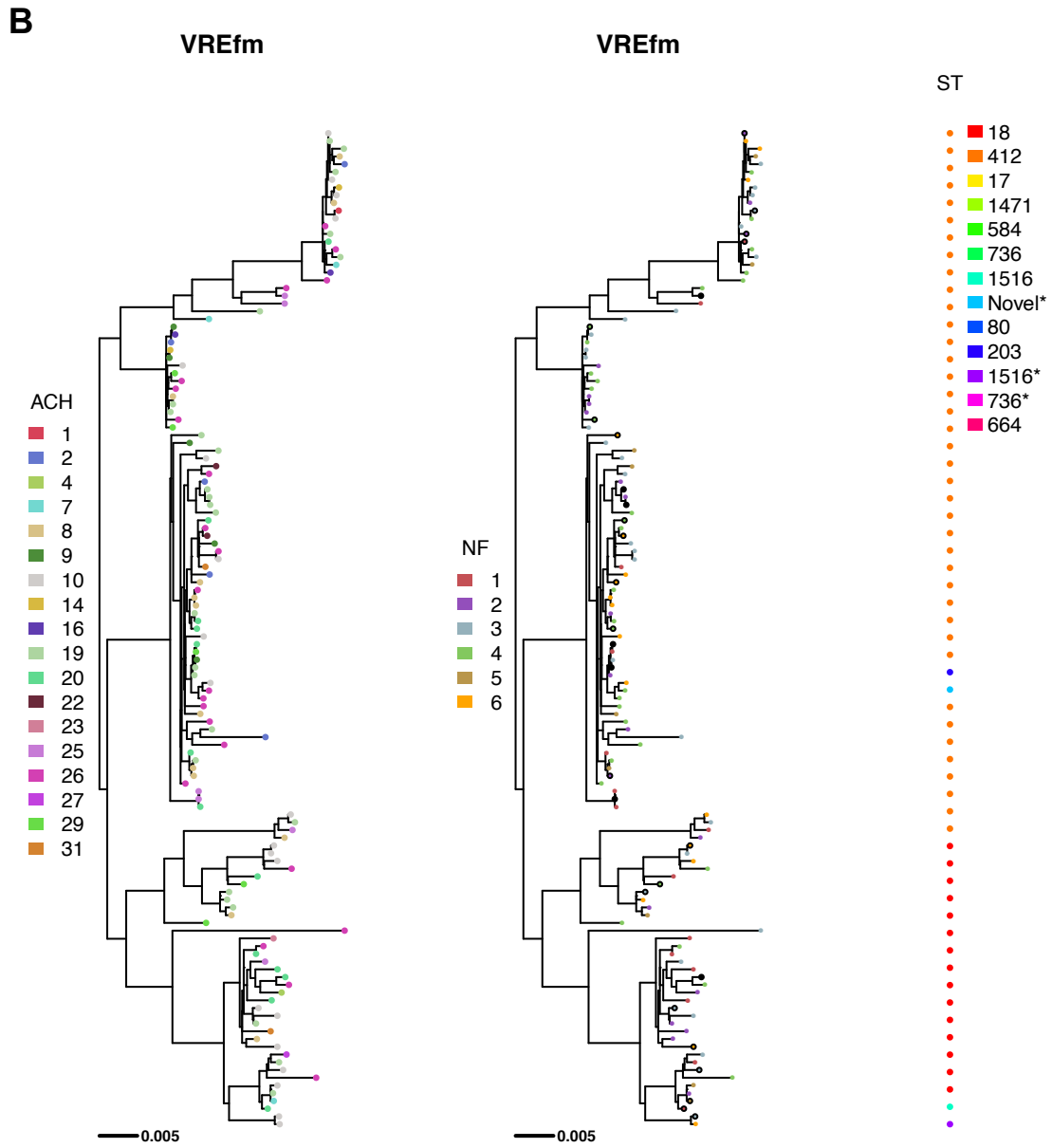t the time of NF admission are shown as solid circles. Follow-up isolates that are genetically distinct from other isolates within the same NF are shown as circles with a black border, otherwise indicated as a solid black circle and pruned from subsequent analysis. Sequence type (ST) of each isolate is shown on the right. The trees were inferred from maximum likelihood (RAxML) analysis with midpoint rooting. Scale bar represents substitutions per nucleotide site. * denotes some uncertainty in ST designation.
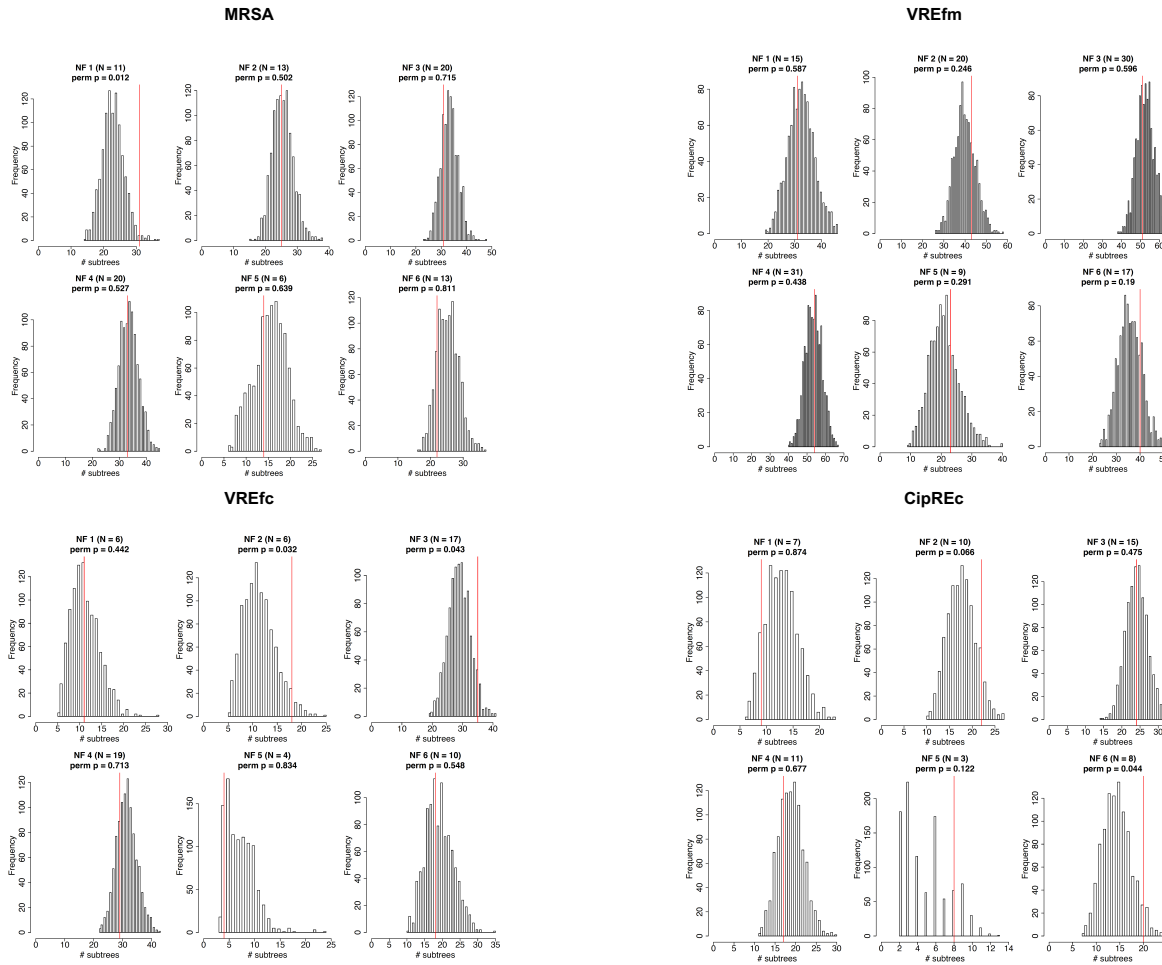
Figure S7. Histograms of each ARO indicating probability of phylogenetic clustering by nursing facility. The number of subtrees containing 2 or more isolates from the same nursing facility (NF) over the 1,000 permutations is plotted against the observed value which is indicated by the vertical red line. An empirical P value was computed by determining the fraction of permuted phylogenetic clustering that was greater than the observed statistic.
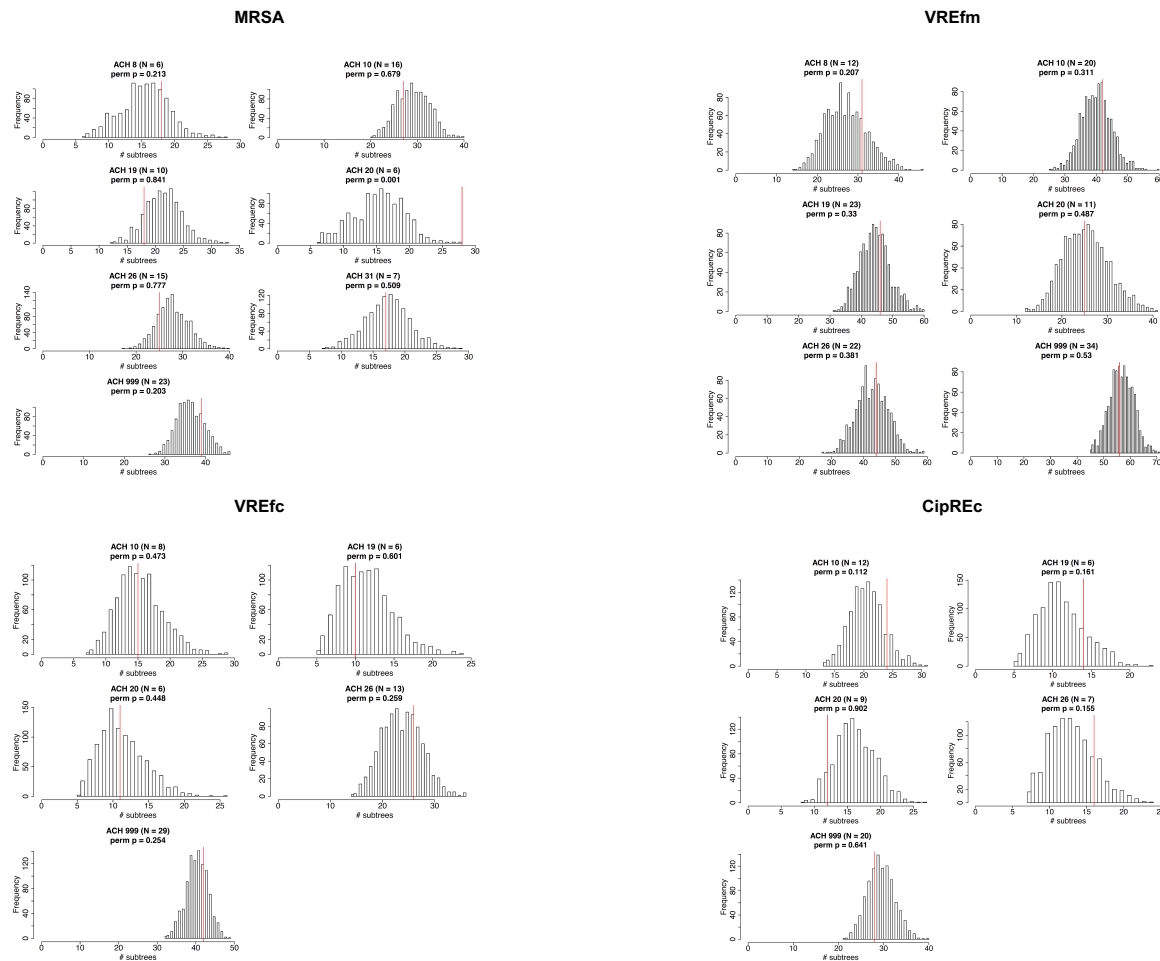
Figure S8. Histograms of each ARO indicating probability of phylogenetic clustering by acute-care hospital. The number of subtrees containing 2 or more isolates from the same acute-care hospital (ACH) over the 1,000 permutations is plotted against the observed value which is indicated by the vertical red line. An empirical P value was computed by determining the fraction of permuted phylogenetic clustering that was greater than the observed statistic. ACH999 is a collection of ACHs with five or fewer isolates.
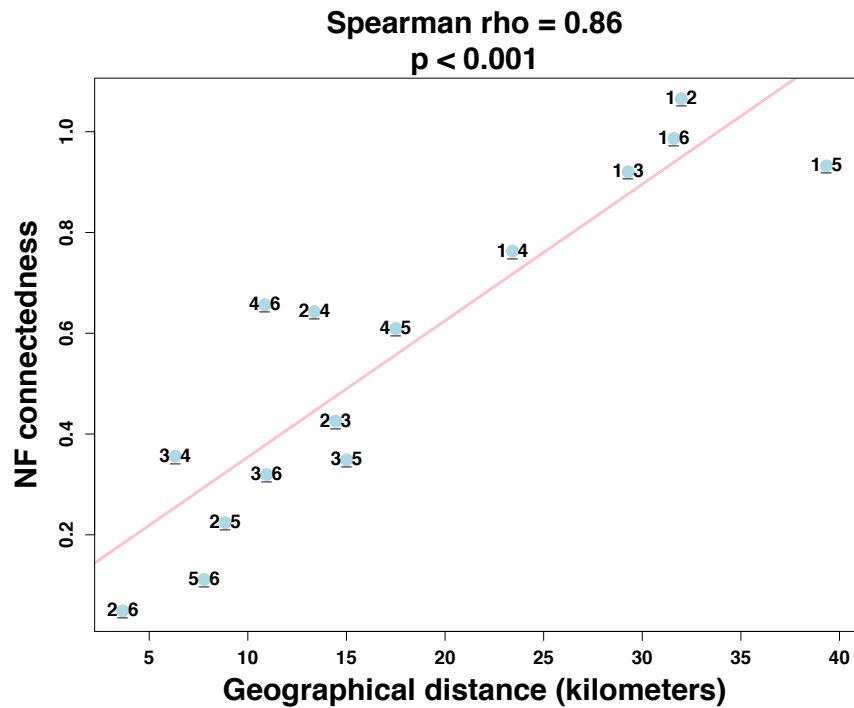
Figure S9. Spearman's rank correlation between geographical distance and connectedness between NFs. A low connectedness value between a pair of NFs indicates that these two NFs received patients from the same set of ACHs, and *vice versa*. Each dot represents a pair of NFs.
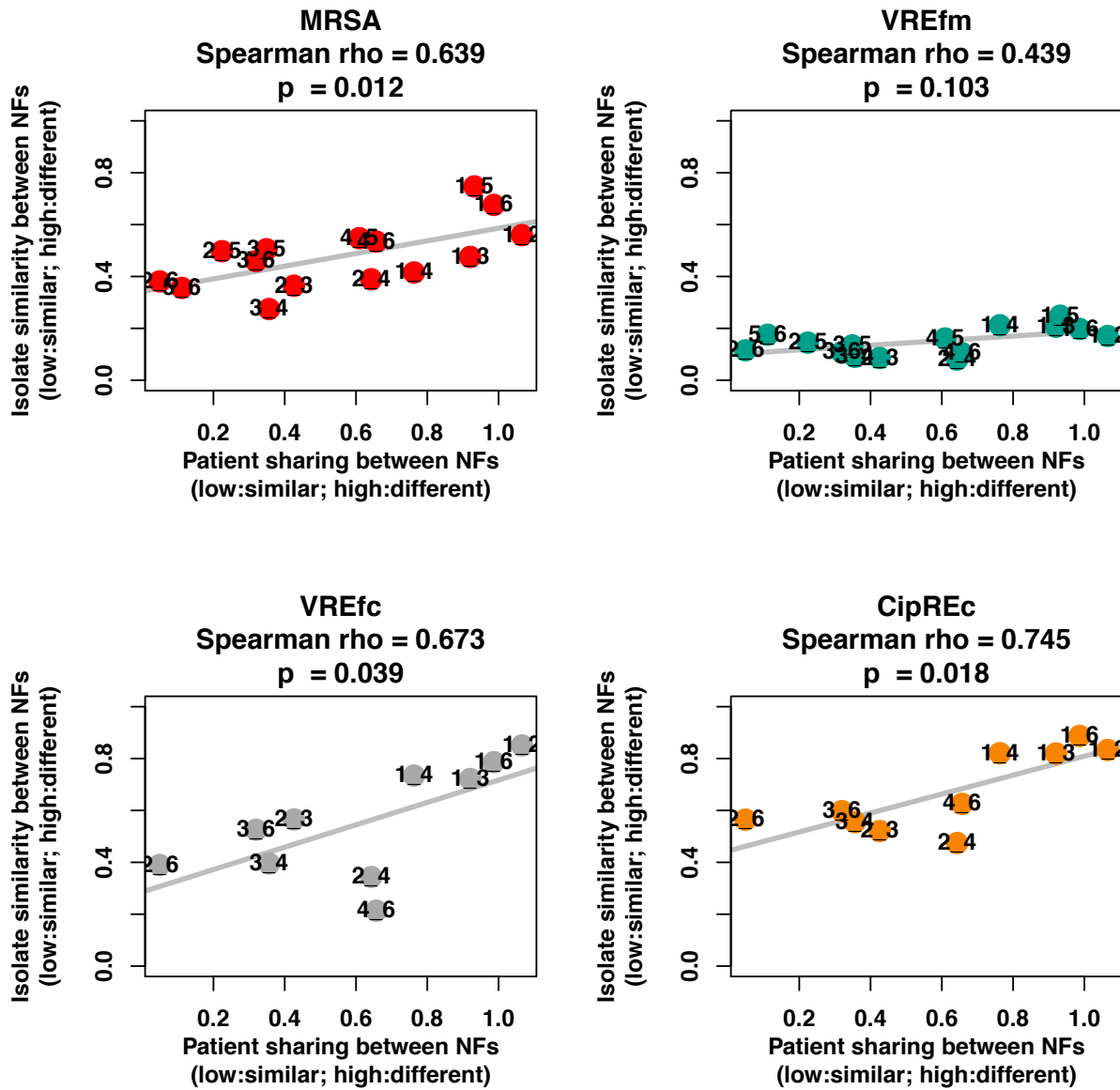
Figure S10. Spearman's rank correlation of patient transfer pattern and genomic similarity between nursing facility (NF) pairs. Each dot represents a pair of NFs. Only NF pairs both with at least five isolates are included in analysis.
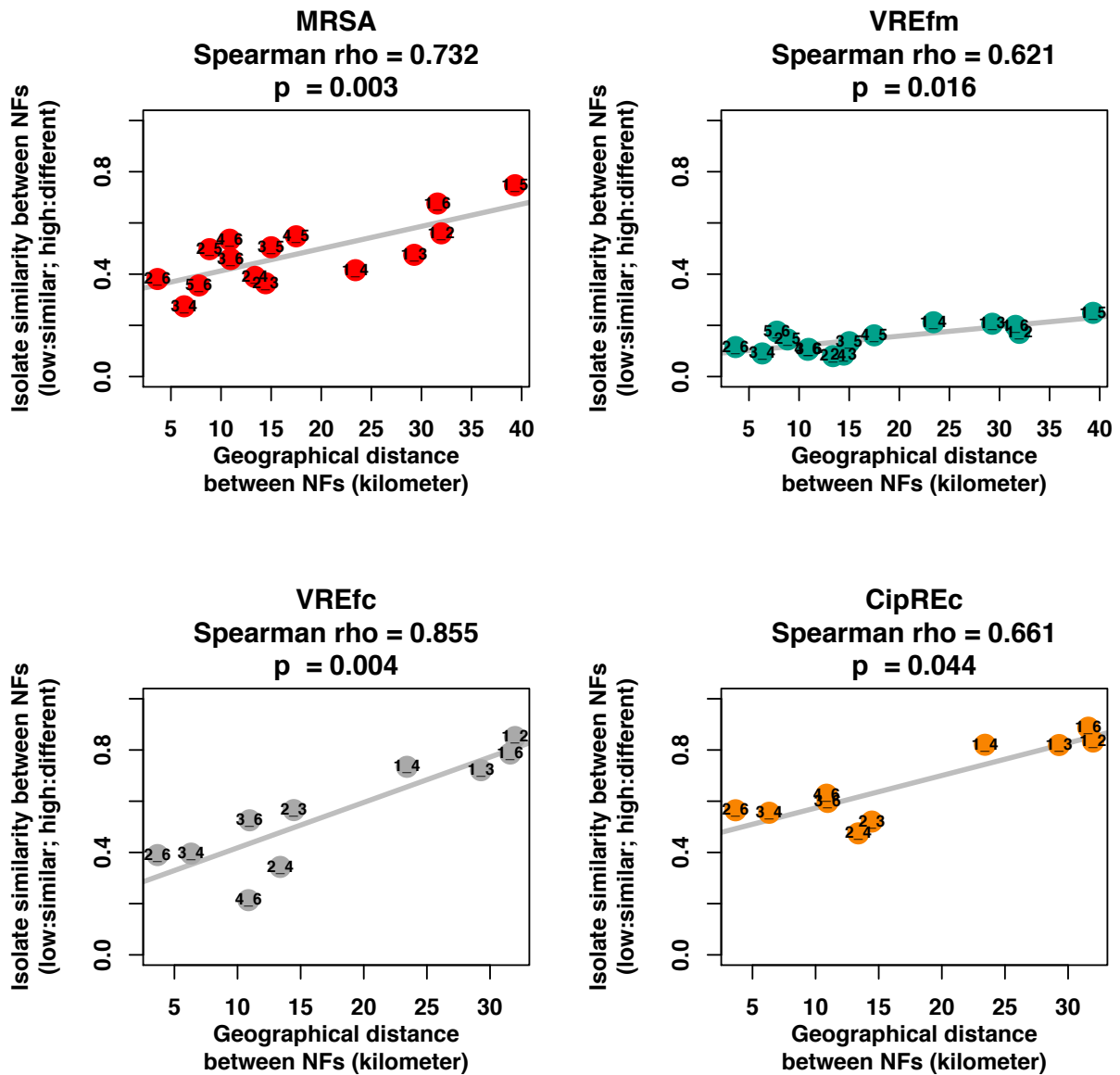
Figure S11. Spearman's rank correlation of geographical distance and genomic similarity between isolates of each ARO species between NF pairs. Each dot represents a pair of NFs. Only NF pairs both with at least five isolates are included in analysis.

Supplementary Tables

Risk Factor	MRSA (n = 67) OR (95% CI)	VREfm (n = 101) OR (95% CI)	VREfc (n = 49) OR (95% CI)	CipREc (n = 34) OR (95% CI)
Male sex	1.11 (0.64 - 1.91)	1.12 (0.7 - 1.79)	1.29 (0.69 - 2.38)	1.02 (0.48 - 2.1)
Urinary catheter use in past 30 days	1.9 (1.05 - 3.39)	2.33 (1.42 - 3.84)	1.47 (0.73 - 2.86)	1.35 (0.59 - 2.92)
Age	1.02 (0.99 - 1.04)	1 (0.98 - 1.02)	1.02 (0.99 - 1.04)	1.01 (0.98 - 1.04)
Functional disability	1.14 (1.08 - 1.22)	1.1 (1.04 - 1.15)	1.08 (1.01 - 1.15)	1.1 (1.02 - 1.18)
Length of hospital stay	1.04 (1 - 1.1)	1.1 (1.05 - 1.15)	1.04 (0.99 - 1.09)	1.02 (0.97 - 1.06)
Charlson comorbidity score	1.09 (0.96 - 1.24)	1.12 (1 - 1.25)	1.17 (1.02 - 1.33)	1.1 (0.94 - 1.29)
1st/2nd Gen Cephalosporins	0.51 (0.15 - 1.36)	1.08 (0.51 - 2.18)	0.91 (0.3 - 2.31)	0.24 (0.01 - 1.2)
3rd/4th Gen Cephalosporins	2.66 (1.23 - 5.59)	4.45 (2.4 - 8.37)	1.84 (0.69 - 4.42)	1.9 (0.6 - 5.1)
Glycopeptides	3.25 (1.51 - 6.88)	4.05 (2.12 - 7.85)	2.77 (1.13 - 6.42)	2.12 (0.66 - 5.76)

Table S 1. Univariate analysis of patient-level risk factors associated with antibiotic-resistant organism colonization at nursing facility admission. Abbreviations: MRSA, methicillin-resistant *Staphylococcus aureus*; VREfm, vancomycin-resistant *Enterococcus faecium*; VREfc, vancomycin-resistant *Enterococcus faecalis*; CipREc, ciprofloxacin-resistant *Escherichia coli*; OR, odds ratio; CI, confidence interval.

Supplementary Reference*s*

1. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 4 August 2018.

2. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **2014**; 30:2114–2120. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/. Accessed 4 August 2018.

3. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol **2012**; 19:455–477. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3342519/. Accessed 4 August 2018.

4. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics **2013**; 29:1072–1075. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624806/. Accessed 4 August 2018.

5. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics **2009**; 25:1968–1969. Available at: http://academic.oup.com/bioinformatics/article/25/15/1968/213094. Accessed 4 August 2018.

6. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics **2009**; 25:1754–1760. Available at: https://academic.oup.com/bioinformatics/article/25/14/1754/225615. Accessed 30 July 2019.

7. Picard Tools - By Broad Institute. Available at: http://broadinstitute.github.io/picard/. Accessed 30 July 2019.

8. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics **2009**; 25:2078–2079. Available at: https://academic.oup.com/bioinformatics/article/25/16/2078/204688. Accessed 30 July 2019.

9. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res **2016**; 44:W16–W21. Available at: http://academic.oup.com/nar/article/44/W1/W16/2499364. Accessed 4 August 2018.

10. 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:e15–e15. Available at: http://academic.oup.com/nar/article/43/3/e15/2410982. Accessed 4 August 2018.

11. Aziz RK, Bartels D, Best AA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics **2008**; 9:75. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2265698/. Accessed 23 July 2019.

12. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **2014**; 30:1312–1313. Available at: http://academic.oup.com/bioinformatics/article/30/9/1312/238053. Accessed 4 August 2018.

13. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution **2012**; 3:217–223. Available at: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2011.00169.x. Accessed 24 December 2019.

14. Han JH, Lapp Z, Bushman F, et al. Whole-Genome Sequencing To Identify Drivers of Carbapenem-Resistant Klebsiella pneumoniae Transmission within and between Regional Long-Term Acute-Care Hospitals. Antimicrobial Agents and Chemotherapy **2019**; 63:e01622-19. Available at: http://aac.asm.org/content/63/11/e01622-19. Accessed 25 October 2019.

15. Hijmans RJ, Williams E, Vennes C. geosphere: Spherical Trigonometry. 2017. Available at: https://CRAN.R-project.org/package=geosphere. Accessed 5 August 2018.

16. Csárdi G, Nepusz T. The igraph software package for complex network research. 2006.

17. entropy: Estimation of Entropy, Mutual Information and Related Quantities version 1.2.1 from CRAN. Available at: https://rdrr.io/cran/entropy/. Accessed 5 August 2018.

18. Hunt M, Mather AE, Sánchez-Busó L, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microbial Genomics **2017**; 3. Available at: https://mgen.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000131. Accessed 31 May 2019.

19. Donker T, Reuter S, Scriberras J, et al. Population genetic structuring of methicillin-resistant Staphylococcus aureus clone EMRSA-15 within UK reflects patient referral patterns. Microbial Genomics **2017**; 3. Available at: http://mgen.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000113. Accessed 9 May 2018.