

Whole genome sequencing detects minimal clustering among *Escherichia coli* Sequence Type 131 *H30* isolates collected from U.S. children's hospitals

Arianna Miles-Jay [1][2], Scott J. Weissmann [2][3], Amanda L. Adler [2], Janet G. Baseman [1],
Danielle M. Zerr [2][3]

Affiliations:

1. Department of Epidemiology, University of Washington, Seattle, Washington, USA
2. Seattle Children's Research Institute, Seattle, Washington, USA
3. Department of Pediatrics, University of Washington, Seattle, Washington, USA

Running header: ST131-*H30* clusters in U.S. children

Word counts:

Abstract: 99

Text: 1955

Footnotes

All authors report no conflicts of interest. This work was supported by Seattle Children’s Research Institute’s Center for Clinical and Translational Research Pediatric Pilot Fund program; and the National Institutes of Health via the National Institute of Allergy and Infectious Diseases (grant number R01AI083413), and the National Center for Advancing Translational Sciences (grant number TL1TR000422). A portion of this work was presented at the American Society for Microbiology Conference on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatics Pipelines in Tysons, Virginia, USA in September 2018.

Address correspondence to:

Dr. Arianna Miles-Jay

University of Michigan Medical School

1150 W. Medical Center Dr.

Medical Science Research Building I, Room 1511

Ann Arbor, MI 48109

amilesj@umich.edu

A. Miles-Jay’s affiliation has changed since the completion of this work. Her current affiliation is Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA.

1 **Abstract**

2 *Escherichia coli* sequence type 131 *H30* has garnered global attention as a dominant
3 antimicrobial-resistant lineage of extraintestinal pathogenic *E. coli*, but its transmission
4 dynamics remain undefined. We applied whole genome sequencing to identify putative
5 transmission clusters among clinical isolates of *H30* from children across the U.S. Of 126
6 isolates, 17 were involved in 8 putative transmission clusters; 4 clusters involved isolates with
7 some evidence of healthcare-associated epidemiologic linkages. Geographic clustering analyses
8 showed weak geographic clustering. These findings are consistent with a framework where
9 within-hospital transmission is not a main contributor to the propagation of *H30* in a pediatric
10 setting.

11

12 **Key words:** *E. coli* infections; ST131; antimicrobial resistance; pediatric infections

13 Background

14 Extraintestinal pathogenic *Escherichia coli* (ExPEC) cause a wide range of non-intestinal
15 illnesses, ranging from uncomplicated urinary tract infection to potentially fatal bacteremia [1].
16 Unlike intestinal pathogenic *E. coli*, which is commonly associated with outbreaks, ExPEC
17 infections are historically considered sporadic, and tracking ExPEC transmission has not been a
18 clinical or public health priority. However, the widespread dissemination of antimicrobial-
19 resistant lineages such as sequence type (ST) 131-*H30* (also known as ST131 Clade C), a
20 dominant multidrug-resistant (MDR) ExPEC lineage in both adults and children, has brought
21 new interest to understanding the transmission dynamics of these common pathogens [2,3]. In
22 particular, the relative importance of healthcare vs. community transmission to the
23 propagation of MDR ExPEC lineages remains poorly defined, though recent work indicates
24 community exposures may be more important in pediatric patients.[4]

25
26 The transmission dynamics of ST131-*H30* (hereafter, *H30*) are challenging to study. Like other
27 ExPEC lineages, *H30* is known to asymptomatically colonize the gut for extended periods of
28 time prior to—or potentially without ever—transitioning to extraintestinal infection [5]. These
29 instances of long-term intestinal colonization likely result in numerous “silent” transmission
30 events [6]. Whole genome sequencing (WGS) and phylogenetic methods can shed light on
31 pathogen transmission dynamics even in the presence of silent transmission events. Here, we
32 used WGS to identify putative transmission clusters among passively collected clinical *H30*
33 isolates from four children’s hospitals across the U.S. We also quantified genomic evidence of
34 geographic clustering to characterize the spatiotemporal dynamics of *H30* among children.

35 **Methods**

36 ***Strain collection and whole genome sequencing***

37 All isolates and clinical data came from a previously described multicenter case-control study.
38 Briefly, between September 1, 2009 and September 30, 2013, four freestanding children's
39 hospitals—referred to here as “West,” “Midwest 1,” “Midwest 2,” and “East”—collected *E. coli*
40 isolates during the course of standard clinical care from individuals <22 years old. All extended-
41 spectrum cephalosporin-resistant and a subset of extended-spectrum cephalosporin-sensitive
42 isolates were collected [7]. The Institutional Review Board at each hospital approved the study
43 protocol. *H30* isolates were identified using the *fumC/fimH* genotyping scheme [8]; only the
44 first *H30* isolate per individual was included.

45
46 All *H30* isolates underwent WGS on the Illumina NextSeq platform. Sequencing reads were
47 quality filtered, trimmed, mapped to a high-quality *H30* reference genome (EC958), and single
48 nucleotide variants (SNVs) were called and filtered.[9] Filtered SNVs were used to construct a
49 pairwise SNV distance matrix using snp-dists, and a maximum-likelihood phylogenetic tree,
50 using IQ-tree [10,11]. See Supplementary Methods for more details. Sequence data generated
51 are available from the NCBI Sequence Read Archive under BioProject PRJNA578285 (see
52 Supplementary Table 2 for study sample metadata).

54 ***Identification and characterization of putative clusters***

55 Pairwise SNV distances within and between all combinations of collection sites were visualized,
56 and the minimum SNV distance between two isolates from discordant collection sites was used

57 to define a threshold for identification of putative transmission clusters. This approach was
58 selected in an effort to capture direct and indirect transmission events that are
59 epidemiologically relevant, i.e. would warrant further investigation should the clusters have
60 been identified in real-time. Given the substantial geographic distance between the collection
61 sites in this study, we expect no epidemiologically relevant transmission events that span two
62 distinct collection sites. The transcluster package in R (version 3.5.1 ,R Core Team, 2018) was
63 used to estimate counts of uncaptured transmission events separating isolates in each putative
64 cluster while incorporating the sampling dates and estimated evolutionary rate and
65 transmission rate of *H30* [12]. See Supplementary Methods for more details.

66

67 ***Genomic evaluation of geographic clustering***

68 To examine the spatiotemporal dynamics of *H30*, we quantified genomic evidence of
69 geographic clustering using two approaches. The primary approach was a previously described
70 SNV-distance based approach where a ratio of the median SNV distance within collection sites
71 over the median SNV distance between collection sites ($SNV_{within} / SNV_{between}$) is calculated, with
72 a ratio closer to zero indicating more evidence of clustering by collection site [13]. Statistical
73 significance of the pairwise-SNV distance-based clustering was assessed using permutation-
74 based 95% interval estimates with 1000 permutations; a SNV distance ratio below the lower
75 bound of the 95% interval estimate indicated evidence of geographic clustering. Secondly,
76 we applied a previously described phylogenetically informed approach [14]. Briefly, the number
77 of isolates in well-supported phylogenetic clades of size 2 or greater that were homogeneous
78 for collection site (“pure” clusters) was tallied. Statistical significance of these counts was again

79 assessed using permutation-based 95% interval estimates with 1000 permutations; counts
80 higher than the upper bound of the permuted 95% interval estimate were considered evidence
81 of clustering. Both approaches were first executed on the full sample set and then on four
82 temporally segregated sample sets in approximate one-year increments to explore whether
83 geographic signal interacted with the temporal variability of sampling.

84
85 Clustering by “geographically close” sites compared to “geographically distant” sites was also
86 examined. In SNV-distance based analyses, pairs of isolates that spanned Midwest 1 and
87 Midwest 2 were classified as geographically close, while all other discordant site pairs were
88 classified as geographically distant. In phylogenetically informed analyses, the two sites in the
89 Midwest region of the U.S. were collapsed into one “Midwest” category. The same methods
90 described above were applied to the full data set and to the temporally segregated sample sets.

91

92 **Results**

93 One hundred thirty *H30* isolates were identified out of 1,347 *E. coli* screened over the four-year
94 study period. Three of the 130 *H30* isolates were determined to be non-*H30* after *in silico*
95 analysis and one isolate was identified as a subsequent isolate from an individual already in the
96 study, leaving 126 *H30* isolates in the remainder of the analyses. After quality filtering and
97 recombination masking, 3,433 variable sites were identified and included in the whole-genome-
98 based SNV alignment.

99

100 There were 7,875 different pairwise comparisons made, with the pairwise SNV distance ranging
101 from 0 to 165 SNVs. The minimum SNV distance between isolates from discordant collection
102 sites was 14 SNVs, and pairs of isolates separated by less than or equal to 14 SNVs were
103 considered to be members of a putative transmission cluster (Figure 1A). Using this threshold,
104 eight putative clusters were identified involving seventeen isolates, seven clusters containing
105 two isolates and one cluster containing three isolates (Figure 2A, Supplementary Figure 1). The
106 putative cluster with three isolates (Cluster 1) consisted of one pair separated by 15 SNVs,
107 which was just beyond the selected cutoff, but because the other two pairs within the cluster
108 were separated by <14 SNVs, all three isolates were included in further analyses. Clusters
109 contained a mix of community- and healthcare-associated isolates (Supplementary Figure 2).
110
111 Out of the eight identified putative clusters, documented epidemiologic data associated with
112 four clusters (Clusters 2,6,7,8) was consistent with possible nosocomial acquisition. Clusters 2
113 and 6 involved individuals with documented overlapping dates of hospitalization (Figure 2B).
114 The genomic evidence for direct transmission within these clusters is less clear: they were
115 separated by 12 and 10 SNVs, respectively, and the transcluster method estimated that there
116 were between 8 and 19 transmission events separating the isolates in these clusters (Figure
117 2A). Additionally, only one of the two isolates in Cluster 2 was phenotypically resistant to
118 trimethoprim-sulfamethoxazole (Supplementary Figure 2). However, the within-cluster
119 difference in isolation dates was 179 and 199 days, so long-term colonization and within-host
120 evolution may have inflated the estimated number of transmission events and resulted in a loss
121 of a resistance determinant. Clusters 7 and 8 consisted of isolates that differed by 0-1 SNVs

122 after quality filtering and were collected between 1 and 7 days of one another, with the
123 transcluster method estimating direct transmission (Figure 2A). While there was no
124 documentation of overlapping hospitalizations within Cluster 7 or 8, both individuals within
125 Cluster 7 had surgical site infections associated with neurological procedures, while both
126 individuals within Cluster 8 were paraplegic. These connections are consistent with a plausible
127 epidemiological link in inpatient or outpatient care, although conclusively establishing such a
128 link is outside the scope of these data.

129
130 Genomic clustering analyses demonstrated minimal evidence of geographic clustering by
131 collection site. Visual inspection of geographic sites on the phylogenetic tree did not show
132 remarkable evidence of geographic clusters (Supplementary Figure 1). The median SNV
133 distance within pairs across concordant sites was not significantly different from the median
134 SNV distance within pairs across discordant sites. ($SNV_{within} / SNV_{between} = 1.01$, 95% interval
135 estimate 0.99-1.01 Figure 1A). Similarly, the median SNV distance within pairs across the most
136 geographically proximate discordant sites (Midwest 1 and Midwest 2) was not significantly
137 lower than the median SNV distance within pairs across more geographically distant site pairs
138 ($SNV_{within} / SNV_{between} = 1.03$, 95% interval estimate 0.99-1.02, Figure 1B). The phylogenetic-
139 informed approach demonstrated weak clustering in the data by study site and no clustering by
140 Midwest sites vs. other sites (Supplementary Figure 3). Results for temporally segregated
141 sample sets were similar, with most measures not supporting evidence of clustering
142 (Supplementary Table 1 and Supplementary Figures 4 and 5).

143

144 **Discussion**

145 We applied WGS to clinical isolates collected from four freestanding U.S. children's hospitals
146 over four years to identify putative transmission clusters and investigate the spatiotemporal
147 dynamics of *E. coli* ST131-*H30*. We identified eight putative transmission clusters of *H30*
148 involving seventeen isolates, including two clusters with documented overlapping
149 hospitalizations and two clusters with other plausible healthcare-associated epidemiologic
150 links. Genomic spatiotemporal analyses demonstrated little evidence of geographic clustering
151 of *H30* more broadly.

152

153 To our knowledge, there are no available data about transmission of *H30* between children
154 within healthcare settings. Our observation of limited plausible within-hospital transmission is
155 not surprising, given the infrequent documentation of MDR-ExPEC transmission within
156 healthcare, generally [15]. These findings are consistent with a framework where within-
157 hospital transmission is not a dominant contributor to the propagation of *H30* in a pediatric
158 setting. However, the identification of some plausible nosocomial transmission highlights the
159 utility of WGS of isolates collected during the course of standard clinical care to uncover silent
160 transmission events.

161

162 There are also no data, to our knowledge, describing the spatiotemporal dynamics of *H30*
163 within the U.S. using geographically diverse isolates. Our observation of limited geographic
164 clustering was unexpected; we anticipated a stronger genomic signature associated with
165 sustained local circulation at the various geographic sites. These findings may reflect the rapid

166 and recent dissemination of *H30* at the time of this data collection. Whether these patterns
167 remain the same today, almost two decades after *H30* is believed to have disseminated rapidly
168 and globally, is worthy of further study [2].

169
170 The results of this study should be interpreted in the context of multiple limitations. First, the
171 available epidemiologic data were limited — including a lack of detail about the location of
172 specific wards during overlapping hospitalizations — and, as such, all observations of plausible
173 transmission should be interpreted cautiously. Second, as highlighted elsewhere [12], the
174 selection of a SNV threshold as a method of defining putative transmission clusters has
175 limitations. However, the multicenter design in this study provided the opportunity to apply a
176 conservative threshold where even indirect transmission was epidemiologically unlikely, which
177 we believe to be a reasonable approach given the limited transmission data available about
178 *H30*. Finally, the local epidemiology of *H30* may have changed since the collection of these
179 isolates. This study also had several strengths, including a multicenter design; a large collection
180 of *H30* isolates from an understudied pediatric population; and the use of WGS and
181 phylogenetically informed approaches to investigate both transmission-based and geographic
182 clustering.

183
184 As antimicrobial resistance rates among ExPEC rise, there is new urgency to improve our
185 understanding of the transmission dynamics of these common pathogens. Taken together, our
186 findings of minimal evidence of transmission clusters or broader geographic clustering are
187 consistent with the prevailing conceptualization of *H30* as a globally and recently disseminated

188 epidemic strain that is often community-associated [2]. Future studies should consider focusing
189 on community-based exposures when investigating the transmission dynamics of *H30*.

Funding

This work was supported by Seattle Children’s Research Institute’s Center for Clinical and Translational Research Pediatric Pilot Fund program; and the National Institutes of Health via the National Institute of Allergy and Infectious Diseases (grant number R01AI083413), and the National Center for Advancing Translational Sciences (grant number TL1TR000422).

Acknowledgements

The authors would like to thank Carey-Ann Burnham, Alexis Elward, Jason Newland, Rangaraj Selvarangan, Kaede Sullivan, Theoklis Zaoutis, and Xuan Qin for provision of bacterial isolates and associated clinical data; Brad Cookson for helpful advice; Jeff Myers and Huxley Smart for assistance with molecular typing of isolates; the Northwest Genomics Center for execution of whole genome sequencing; and members of the Snitkin Lab at the University of Michigan for their critical review of the manuscript.

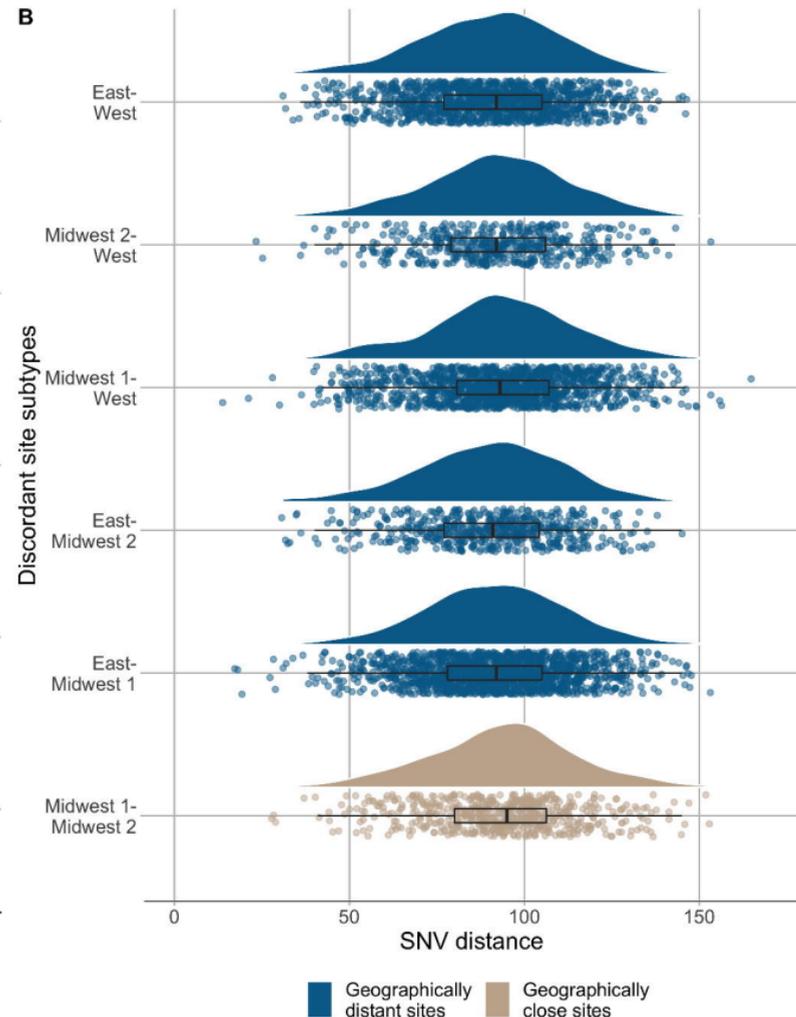
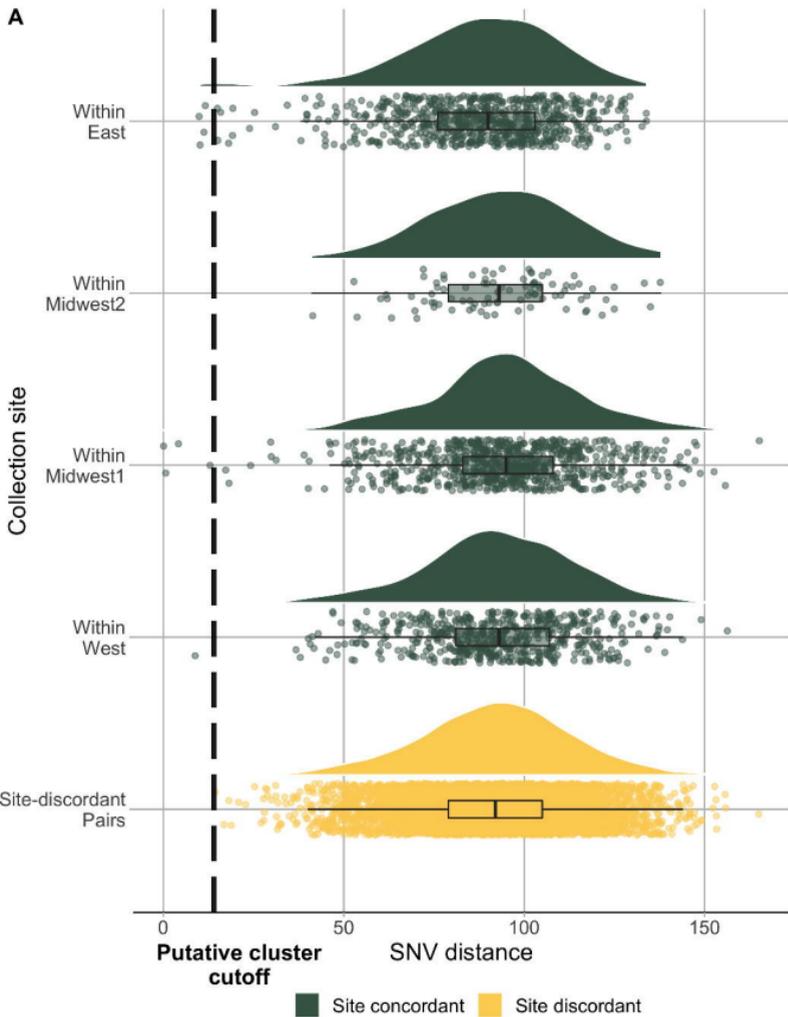
References

1. Russo TA., Johnson JR. Medical and economic impact of extraintestinal infections due to *Escherichia coli*: Focus on an increasingly important endemic problem. *Microbes Infect.* **2003**; 5(5):449–456.
2. Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL, Billig M, et al. Abrupt emergence of a single dominant multidrug-resistant strain of *Escherichia coli*. *J Infect Dis.* **2013**; 207(6):919–928.
3. Miles-Jay A, Weissman SJ, Adler AL, Tchesnokova V, Sokurenko E V., Baseman JG, et al. Epidemiology and Antimicrobial Resistance Characteristics of the Sequence Type 131-H30 Subclone among Extraintestinal *Escherichia coli* Collected from US Children. *Clin Infect Dis.* **2018**; 66(3):411–419.
4. Logan LK, Medernach RL, Rispens JR, Marshall SH, Hujer AM, Domitrovic TN, et al. Community Origins and Regional Differences Highlight Risk of Plasmid-mediated Fluoroquinolone Resistant Enterobacteriaceae Infections in Children. *Pediatr Infect Dis J.* **2019**; 38(6):595–599.
5. Torres E, López-Cerero L, Morales I, Navarro MD, Rodríguez-Baño J, Pascual A. Prevalence and transmission dynamics of *Escherichia coli* ST131 among contacts of infected community and hospitalized patients. *Clin Microbiol Infect.* **2018**; 24(6):618–623.
6. Logan LK, Hujer AM, Marshall SH, Domitrovic TN, Rudin SD, Zheng X, et al. Analysis of β -Lactamase Resistance Determinants in Enterobacteriaceae from Chicago Children: A Multicenter Survey. *Antimicrob Agents Chemother.* **2016**; 60(March):3462–3469.
7. Zerr DM, Miles-Jay A, Kronman MP, Zhou C, Adler AL, Haaland W, et al. Previous

- antibiotic exposure increases risk of infection with extended spectrum beta lactamase- and AmpC-producing *Escherichia coli* and *Klebsiella pneumoniae* in pediatric patients. *Antimicrob Agents Chemother.* **2016**; 60(7):4237–4243.
8. Weissman SJ, Johnson JR, Tchesnokova V, Billig M, Dykhuizen D, Riddell K, et al. High-resolution two-locus clonal typing of extraintestinal pathogenic *Escherichia coli*. *Appl Environ Microbiol.* **2012**; 78(5):1353–1360.
 9. Forde BM, Zakour NL Ben, Stanton-Cook M, Phan MD, Totsika M, Peters KM, et al. The complete genome sequence of *Escherichia coli* EC958: A high quality reference sequence for the globally disseminated multidrug resistant *E. coli* O25b:H4-ST131 clone. *PLoS One.* **2014**; 9(8).
 10. Seemann T. *snp-dists*: Pairwise SNP distance matrix from a FASTA sequence alignment [Internet]. [cited 2019 Jan 21]. Available from: <https://github.com/tseemann/snp-dists>
 11. Nguyen LT, Schmidt HA, Haeseler A Von, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* **2015**; 32(1):268–274.
 12. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. *Mol Biol Evol.* **2019**; 36(3):587–603.
 13. Eyre DW, Davies KA, Davis G, Fawley WN, Dingle KE, Maio N De, et al. Two Distinct Patterns of *Clostridium difficile* Diversity Across Europe Indicating Contrasting Routes of Spread. *Clin Infect Dis.* **2018**; 67(7):1035-1344.
 14. Popovich KJ, Snitkin ES, Hota B, Green SJ, Pirani A, Aroutcheva A, et al. Genomic and

Epidemiological Evidence for Community Origins of Hospital-Onset Methicillin-Resistant Staphylococcus aureus Bloodstream Infections. *J Infect Dis.* **2017**; 215:1640–1647.

15. Hilty M, Betsch BY, Bögli-Stuber K, Heiniger N, Stadler M, Küffer M, et al. Transmission dynamics of extended-spectrum beta-lactamase-producing enterobacteriaceae in the tertiary care hospital and the household setting. *Clin Infect Dis.* **2012**; 55:967–975.



A

Cluster	SNVs apart	Days apart	Estimated number of transmission events apart
1	10-15	129-478	7-17
2	12	179	9-19
3	4	104	2-6
4	13	242	10-22
5	9	116	6-15
6	10	199	8-16
7	0	7	1
8	1	1	1

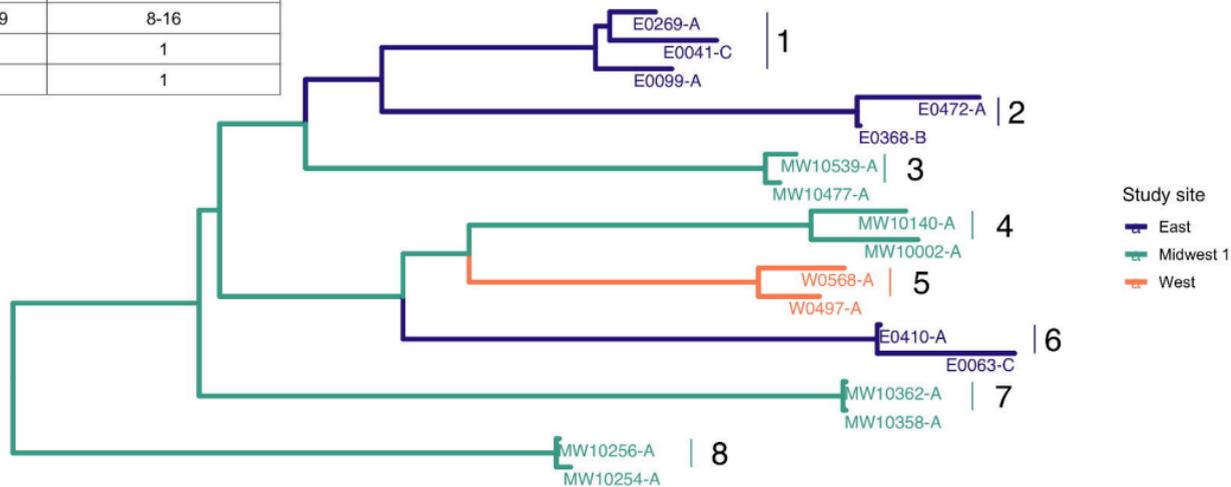
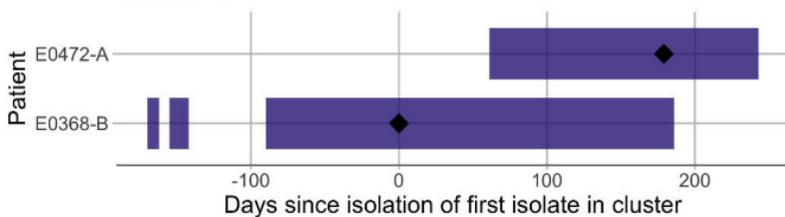
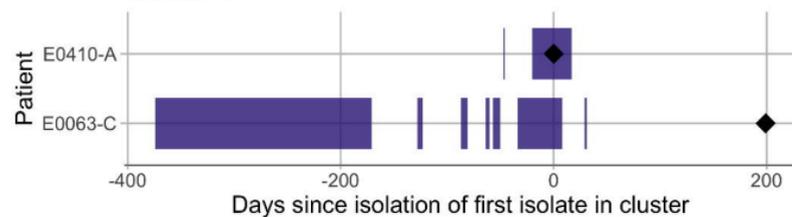
**B****Cluster 2****Cluster 6**

Figure 1: Distribution of pairwise single nucleotide variant (SNV) differences between *H30* clinical isolates from four children's hospitals in the U.S. A) Within single collection sites compared to between discordant collection sites. Dashed line indicates selected threshold for identifying putative transmission clusters; and B) Between geographically distant discordant collection sites vs. between geographically close discordant collection sites.

Figure 2: A) Phylogeny of eight identified putative transmission clusters of *H30* identified from four children's hospitals in the U.S. colored by study site, with the number of days separating their collection; the number of single nucleotide variants (SNVs) separating them after quality filtering; and a range of estimated number of transmission events separating them as calculated by the R package transcluster. The range of estimated transmission events reflect a range of reasonable values for substitution and transmission rates, but do not account for potential intra-host evolution or diversity. B) Temporal depiction of the overlapping hospitalizations of individuals in Cluster 2 and Cluster 6. The black diamond indicates the date of isolate collection and the purple bars represent time hospitalized. Time is measured in days since the isolation of the first isolate in each cluster.