

# 1 **regentrans: a framework and R package for using genomics to study regional** 2 **pathogen transmission**

3  
4 Sophie Hoffman, BS\*<sup>1</sup>; Zena Lapp, BA\*<sup>1</sup>; Joyce Wang, PhD<sup>2</sup>; Evan S Snitkin, PhD<sup>2,3</sup>

5  
6 <sup>1</sup>Department of Computational Medicine and Bioinformatics, University of Michigan, 1150 W.  
7 Medical Center Dr. Ann Arbor, MI, 48109-5680

8 <sup>2</sup>Department of Microbiology and Immunology, University of Michigan, 1150 W. Medical Center  
9 Dr. Ann Arbor, MI, 48109-5680

10 <sup>3</sup>Department of Medicine, Division of Infectious Diseases, University of Michigan, 1150 W.  
11 Medical Center Dr. Ann Arbor, MI, 48109-5680

12  
13 \*These authors contributed equally to this work.

14  
15 *Corresponding author:*

16 Evan Snitkin

17 [esnitkin@umich.edu](mailto:esnitkin@umich.edu)

## 18 19 **Keywords**

20  
21 R package, regional transmission, genomics, whole genome sequencing, software, pathogen  
22 spread

## 23 24 **ORCIDiDs**

- 25  
26
  - Sophie Hoffman: 0000-0003-2518-6422
  - Zena Lapp: 0000-0003-4674-2176
  - Joyce Wang: 0000-0002-8674-1015
  - Evan Snitkin: 0000-0001-8409-278X

## 27 28 29 30 31 **Repositories**

32  
33 <https://github.com/Snitkin-Lab-Umich/regentrans/>

## 34 35 **Abstract**

36  
37 Increasing evidence of regional pathogen transmission networks highlights the importance of  
38 investigating the dissemination of multidrug-resistant organisms (MDROs) across a region to  
39 identify where transmission is happening and how pathogens move across regions. We  
40 developed a framework for investigating MDRO regional transmission dynamics using whole-  
41 genome sequencing data and created regentrans, an easy-to-use, open source R package that  
42 implements these methods (<https://github.com/Snitkin-Lab-Umich/regentrans/>). Using a dataset  
43 of over 400 carbapenem-resistant *Klebsiella pneumoniae* isolates collected from patients in 21  
44 long-term acute care hospitals (LTACHs) over a one-year period, we demonstrate how to use

45 our framework to gain insights into differences in inter- and intra-facility transmission across  
46 different LTACHs and over time. These tools will allow investigators to better understand the  
47 origins and transmission patterns of MDROs, which is the first step in understanding how to stop  
48 transmission at the regional level.

49

## 50 **Impact statement**

51

52 Increasing evidence suggests that pathogen transmission occurs across healthcare facilities.  
53 Genomic epidemiologic investigations into regional transmission shed light on potential drivers  
54 of regional prevalence and can inform coordinated interventions across healthcare facilities to  
55 reduce transmission. Here we present a framework for studying regional pathogen transmission  
56 using whole-genome sequencing data, and a corresponding open-source R package,  
57 regentrans, that implements these methods. We also discuss how these methods can be  
58 extended to study transmission in other settings.

59

## 60 **Data summary**

61

62 The authors confirm all supporting data, code and protocols have been provided within the  
63 article or through supplementary data files.

64

- 65 • The regentrans R package can be downloaded from GitHub: [https://github.com/Snitkin-](https://github.com/Snitkin-Lab-Umich/regentrans/)  
66 [Lab-Umich/regentrans/](https://github.com/Snitkin-Lab-Umich/regentrans/)
- 67 • The manuscript figures are generated from regentrans example data and can also be  
68 found on GitHub: [https://github.com/Snitkin-Lab-](https://github.com/Snitkin-Lab-Umich/regentrans/tree/master/vignettes/manuscript_figures)  
69 [Umich/regentrans/tree/master/vignettes/manuscript\\_figures](https://github.com/Snitkin-Lab-Umich/regentrans/tree/master/vignettes/manuscript_figures)
- 70 • The example data used in the package and manuscript is from BioProject accession no.  
71 [PRJNA415194](https://ncbi.nlm.nih.gov/bioproject/PRJNA415194). The metadata corresponding to these sequences can be found on the  
72 SRA Run Selector (isolate column) and as example data in the regentrans package.

73

## 74 **Introduction**

75

76 Multidrug-resistant organisms (MDROs) are a global public health threat due to limited  
77 treatment options paired with widespread global transmission [1]. Healthcare facilities in  
78 particular, where critically ill patients reside in close proximity to one another, are hotspots of  
79 MDRO transmission [2]. Furthermore, increasing evidence suggests that substantial  
80 transmission occurs not only within facilities, but also between facilities in regional healthcare  
81 networks, and that intra- and inter-facility transmission does not occur evenly across these  
82 networks [3, 4]. This observation, paired with limited resources for state and regional public  
83 health efforts, necessitates the identification of optimal intervention locations to reduce overall  
84 regional prevalence. Investigating MDRO transmission from a regional perspective can shed  
85 light on the origin and spread of MDROs, providing critical information for precision public health  
86 interventions to allocate resources to maximally reduce transmission across a region [5, 6].

87

88 Understanding where and how recent transmission is occurring is an integral first step in  
89 developing interventions to curb transmission at the regional level. A powerful tool for studying  
90 regional pathogen transmission is whole-genome sequencing, which allows us to investigate  
91 pathogen movements at very high resolution [4, 6]. Several studies have used whole-genome  
92 sequencing, sometimes paired with additional epidemiological metadata, to gain insights into  
93 locations [6, 7] and drivers [3, 4] of elevated intra- or inter-facility transmission. These types of  
94 analyses have the potential to transform our public health response to MDROs if they are  
95 regularly performed at, or in collaboration with, regional public health centers.

96  
97 Here, we provide a framework for studying regional pathogen transmission using whole-genome  
98 sequencing data, and present the regentrans R package that implements these methods. We  
99 discuss methods to study transmission within and between healthcare facilities using whole-  
100 genome sequencing data from a single colony isolate from each patient, and discuss how these  
101 methods can be applied to study transmission within and between other locations such as zip  
102 codes. The methods presented here focus on studying recent transmission in a clonal set of  
103 isolates and can be applied to investigate overall transmission or transmission patterns over  
104 time, and to compare the transmission dynamics of different strains circulating in a region. We  
105 believe that these tools will help investigators better understand regional pathogen transmission,  
106 and thus potentially guide interventions to reduce transmission.

107

## 108 **Investigating regional transmission patterns**

109

110 Below we describe the questions, data, and methods for studying regional pathogen  
111 transmission. More details about using regentrans to implement these methods can be found in  
112 the vignette.

113

### 114 **Questions**

115

116 Our framework for studying regional pathogen transmission aims to help investigators  
117 interrogate the following questions (**Table 1**):

118

- 119 1. Is transmission occurring within and/or between facilities?
- 120 2. What facilities is transmission occurring within/between?
- 121 3. Have transmission dynamics changed over time?
- 122 4. Is transmission occurring along paths of higher patient/person flow?
- 123 5. Are there any observable geographic trends in prevalence/transmission?

124

### 125 **Data**

126

#### 127 Data required

128

129 Whole-genome sequences from studies such as prospective observational studies, point-  
130 prevalence surveys, and regional surveillance across different facilities in a region can be used  
131 to identify the genetic relatedness between isolates and subsequently investigate intra- and

132 inter-facility transmission. Depending on the method being used to study transmission, the  
133 genetic data required is either a recombination-filtered variant alignment or a phylogeny of all  
134 the isolates. We suggest using Gubbins [8] to mask recombinant sites and IQ-TREE [9] to  
135 generate a maximum-likelihood phylogeny. Researchers can also investigate the relationship  
136 between genetic distance and patient transfer between facilities, which requires a patient  
137 transfer network that minimally includes all facilities represented in the dataset. Finally, it is  
138 possible to visualize and quantify potential geographic trends in prevalence and transmission.  
139 We describe the specific inputs required for the regentrans package in the vignette and package  
140 documentation.

141

## 142 Data pre-processing

143

144 The suggested data preprocessing steps prior to performing a regional transmission analysis  
145 are shown in **Figure 1**. First, as the methods we present here are focused on identifying recent  
146 transmission events, we suggest that they be used only on closely related isolates, e.g. ones  
147 within the same sequence type (ST) or clonal complex. However, comparisons between the  
148 transmission dynamics of different groups can be performed. Furthermore, we suggest that the  
149 dataset be subset to include only one isolate from each unique colonization event per patient  
150 per facility, so that intra-facility transmission events are exclusively between different patients.  
151 One simple way of doing this is to use only unique combinations of patient, ST, and facility.

152

## 153 Datasets used in the package and for analyses

154

### 155 *Genomic data*

156

157 The genomic data used for this manuscript, and included in the regentrans package, were  
158 generated from whole-genome sequences of clinical isolates obtained from 21 long-term acute  
159 care hospitals across the U.S. [4]. The original study was reviewed and approved by the  
160 Institutional Review Board of the University of Pennsylvania with a waiver of informed consent.  
161 The data was processed as in [10]. Briefly, trimmed Illumina short reads were aligned to the  
162 KPNIH1 reference genome (BioProject accession no. PRJNA73191) using the Burrows-  
163 Wheeler short-read aligner (bwa v0.7.17) [11] and recombinant sites were masked using  
164 Gubbins v2.3.4 [8]. We used the Gubbins variant output fasta file to generate a pairwise single  
165 nucleotide variant (SNV) distance matrix using the `dist.dna()` function (`method = 'N'`,  
166 `pairwise.deletion = TRUE`, `as.matrix = TRUE`) in ape v5.5 [12]. IQ-TREE v1.6.12 [9] was used to  
167 generate a whole-genome phylogeny of all isolates. For all analyses, the data was subset to  
168 include only ST258 isolates, and only one isolate per patient. Sequence types were determined  
169 using Kleborate v0.4.0 [13].

170

### 171 *Patient transfer data*

172

173 Aggregate patient transfer data from all hospitals in the state of California was used to calculate  
174 paths of maximum patient flow. The data and methods are described in [4].

175

176 **Methods**

177

178 Q0: How do you choose pairwise SNV distance thresholds?

179

180 Several of the methods discussed below rely on interpreting, comparing, or thresholding  
181 pairwise SNV distances between isolates to make inferences. It is generally understood that  
182 small pairwise SNV distances between isolates implies recent transmission [14–16], but that this  
183 method is not entirely accurate due to within-host evolution and variable mutation rates [17, 18].  
184 To identify recent transmission pairs using pairwise SNV distances, investigators must choose a  
185 threshold to determine what pairs are considered closely related [14–16]. The threshold for  
186 “closely related” depends on the pathogen mutation rate and the setting; the mutation rate of  
187 pathogens in outbreak settings is often higher than endemic settings [19]. Thus, for a given  
188 pathogen, closely related isolate pairs in an outbreak setting will likely have a higher pairwise  
189 SNV distance than closely related pairs in an endemic setting. For this reason, knowledge of the  
190 epidemiologic context of the isolates, and the species or sequence type itself, is very important  
191 for interpreting pairwise SNV distance distributions.

192

193 One way to choose a pairwise SNV distance threshold is using the genome length and mutation  
194 rate. For instance, in the context of the dataset we use here, *K. pneumoniae* ST258 isolates  
195 from an endemic setting, we could calculate a pairwise SNV distance threshold based on the  
196 KPNIH1 reference genome length of 5,394,056 base pairs and a mutation rate of 1.03e-6 per  
197 base pair per year [20] (2 isolates \* 5,394,056 bases \* 1.03e-6 bases per year per isolate).  
198 However, it is often difficult or impossible to calculate the evolutionary rate of the pathogen in  
199 the particular instance being studied, and more general estimates of mutation rate may not be  
200 translatable.

201

202 Another way to identify potential pairwise SNV distance thresholds is by visualizing the fraction  
203 of isolate pairs from the same facility for various pairwise SNV distances and look for a  
204 decrease in the fraction of intra-facility isolate pairs, which suggests a potentially reasonable  
205 threshold under the assumption that intra-facility transmission is more likely than inter-facility  
206 transmission. Performing this analysis on our data indicated that using SNV distance thresholds  
207 of  $\leq 10$  and  $\leq 6$  are reasonable (**Figure 2**). We chose to use these two thresholds for our  
208 sensitivity analysis as they are more directly supported by our data compared to the more  
209 general mutation rate analysis.

210

211 When performing analyses where choosing a pairwise SNV distance threshold is necessary, we  
212 suggest that investigators evaluate the robustness of their results by comparing their findings for  
213 different pairwise SNV distance thresholds. Here, we compare the results using a threshold of 6  
214 and 10, but a wider range of values can be used in instances with more uncertainty about what  
215 the threshold should be.

216

217 Q1: Is transmission occurring within and/or between facilities?

218

219 One of the first questions an investigator might ask about isolates collected from a certain  
220 region is if transmission is occurring within particular facilities and/or between facilities.  
221 Phylogenetic and variant-based methods can be used to probe this question, and concordant  
222 findings between methods increase our confidence in the results.

223  
224 *Investigating intra-facility transmission using the phylogeny*

225  
226 One way to investigate the extent of intra-facility transmission is to identify maximum subclades  
227 that all originate from the same facility and quantify the size of these clusters. Larger clusters  
228 indicate more intra-facility transmission, as those isolates are all more closely related to one  
229 another than to the isolates from other facilities. In our dataset, we see that some facilities  
230 exhibit extensive intra-facility transmission as evidenced by large cluster sizes, while some  
231 facilities exhibit relatively little intra-facility transmission (**Figure 3**). However, it is important to  
232 note that isolates within a cluster may still be distantly related if, for instance, transmission is  
233 occurring at a facility that is more geographically isolated, or across longer timescales.

234  
235 *Investigating intra- and inter-facility transmission using pairwise SNV distances*

236  
237 Inspecting pairwise SNV distances between all isolates can provide information about the extent  
238 of recent transmission both within and between facilities, which will often manifest as an  
239 enrichment in closely related isolate pairs (i.e. isolate pairs with small pairwise SNV distances;  
240 see note above on what to consider closely related). In our example dataset, we observe an  
241 enrichment in both closely related intra-facility pairs and inter-facility pairs (pairwise SNV  
242 distance of  $\leq 10$  or  $\leq 6$ ), suggesting that recent transmission is occurring both within and  
243 between facilities (**Figure 4**).

244  
245 Q2: What facilities is transmission occurring within/between?

246  
247 Once we have investigated the extent of transmission occurring within and between facilities on  
248 a general scale, we can dig deeper into identifying certain facilities and facility pairs with closely  
249 related isolates. regentrans provides two methods to do this – one threshold-free approach, and  
250 one approach that requires the investigator to choose a pairwise SNV distance threshold to  
251 define closely related pairs.

252  
253 *Shared variants between facilities*

254  
255 Identifying variants that are shared among isolates at different facilities by calculating gene flow  
256 (Fsp) [21] provides a threshold-free population-level approach to investigating the extent of  
257 inter-facility transmission. This method is particularly useful in endemic scenarios when the  
258 relationship between individual isolates may be relatively diffuse due to frequent patient transfer  
259 over time. Using our dataset, we found that certain facilities have many more shared variants  
260 than others, indicating that there is likely more transmission between those facilities (**Figure 5**).

261  
262 *Pairwise SNV distance threshold*



263

264 Using a pairwise SNV distance threshold, the number of closely related pairs within and  
265 between facilities can be determined and used to identify facilities and with more or less putative  
266 spread. For instance, we observed a large number of closely related intra-facility pairs between  
267 some facilities, and few at other facilities (**Figure 6**). As there are limitations to choosing SNV  
268 cutoffs, we highly recommend doing a sensitivity analysis by choosing several different SNV  
269 thresholds and seeing how robust the results are to these changes.

270

271 Q3: Does transmission correlate with patient transfer?

272

273 In addition to only using genomic information to study transmission, inter-facility transmission  
274 can be studied in the context of patient flow between facilities. While sometimes investigators  
275 may have access to patient-level information regarding prior facility exposures, this information  
276 is often not available. In this case, aggregate patient transfer data can be used to study the  
277 relationship between patient flow and transmission. The simplest way to do this is to determine  
278 whether there is a relationship between direct flow between facilities and either the number of  
279 closely related pairs defined by pairwise SNV distances, the actual values of pairwise SNV  
280 distances, or Fsp. To take into account potential indirect transfers that may influence  
281 transmission, a more complex algorithm can be used to identify paths of maximum patient flow  
282 between facilities, and then this can be compared to metrics of genomic relatedness [4]. These  
283 analyses can provide insight into whether patient flow may be driving transmission between  
284 facilities. For instance, when subsetting our data to 11 Los Angeles area LTACHs, we observe a  
285 negative correlation between patient flow and Fsp, indicating that facilities connected by more  
286 patient flow often have more similar populations (**Figure 7A**). We also observed a positive  
287 correlation between patient flow and the number of closely related isolate pairs between  
288 facilities, suggesting that patient flow may, in part, drive inter-facility transmission (**Figure 7B**).

289

290 Q4: Have transmission dynamics changed over time?

291

292 All of the methods described above can be applied to discrete time chunks to gain insight into  
293 whether transmission dynamics have remained stable or changed over time. For instance, in an  
294 outbreak setting we observed an increase in intra-facility transmission followed by an increase  
295 in inter-facility transmission [7]. In an endemic setting, these trends may remain more stable  
296 over time. In our data, we observe an increase in the total number of pairs from 2014 to 2015,  
297 but no change in the distribution of closely related intra- or inter-facility isolate pairs (**Figure 8**).

298

299 Q5: Are there any observable geographic trends in prevalence/transmission?

300

301 Finally, it is often useful to visualize the geographic distribution of closely related isolates. This  
302 can provide insight into whether inter-facility transmission is concentrated in a certain  
303 geographic region, or is more diffuse. For instance, we can see in our data that facilities that are  
304 geographically more proximate tend to have more transmission between them, as indicated by a  
305 positive correlation between geographic distance and Fsp and a negative correlation between

306 geographic distance and number of closely related isolate pairs for a given facility pair (**Figure**  
307 **9**).

308

### 309 **Package implementation**

310

311 regentrans is implemented in R [22] and is available on GitHub ([https://github.com/Snitkin-Lab-](https://github.com/Snitkin-Lab-Umich/regentrans)  
312 [Umich/regentrans](https://github.com/Snitkin-Lab-Umich/regentrans)). Our package depends on several other packages including tidyverse [23]  
313 packages (dplyr and tidyr), ape [12], phytools [24], igraph [25], and future.apply [26]. The  
314 ggplot2 [23], ggtree [27], and pheatmap [28] packages are used in the vignette for plotting. The  
315 required and optional inputs to each function, as well as a reference to a manuscript that uses  
316 the method, can be found in **Table 1**. Each of the references describes in more detail the  
317 algorithm used in the underlying function [3, 4, 6, 7, 21]. Many functions require a phylogenetic  
318 tree read in by `ape::read.tree()` and/or a pairwise SNV distance matrix calculated using  
319 `ape::dist.dna()`, which requires a DNABin object input that can be read in using `ape::read.fasta()`.  
320 The example geographic data provided in the package was de-identified by adding random  
321 horizontal, vertical, and rotational shifts using the R package tangles v0.8.1 [29]. Our  
322 introductory vignette provides examples of how to read in data, use each function, and plot the  
323 corresponding output for interpretation.

324

### 325 **Additional possible uses**

326

327 While our expertise in studying regional transmission largely lies in investigating transmission  
328 within and between healthcare facilities, the methods implemented in regentrans can be used  
329 for many additional applications. Rather than investigating transmission between facilities, users  
330 could investigate transmission between different zip codes, different rooms or wards within a  
331 hospital, or even transmission between patient and environmental sources. As one example,  
332 Popovich and Thiede *et al.* [30] identified transmission signatures within a large urban jail by  
333 comparing pairwise SNV distances of community-onset MRSA to MRSA acquired within the jail.

334

### 335 **Cautionary notes on interpretation**

336

337 It is important to emphasize that there are several limitations to the methods we describe here.  
338 First, none of these methods include the use of epidemiological data to confirm or corroborate  
339 putative transmission links. Therefore, while we can gain useful insight into the likely extent of  
340 transmission within and between facilities, we cannot understand the nuances of actual  
341 transmission events. If epidemiological data is available, we highly recommend incorporating  
342 this information into the analysis to provide further insights into putative transmission pathways  
343 (examples: [6, 30]). Additionally, as mentioned above, for methods where choosing a threshold  
344 of genomic relatedness is required, care in choosing the threshold and investigating the  
345 sensitivity of the threshold on your interpretation of the results is warranted as the results may  
346 change drastically depending on what threshold is chosen [31]. Finally, the sampling schemes  
347 or time frames used in the study can influence the output of the methods presented here [32,  
348 33]. Therefore, as always, the strengths and limitations of the dataset being used must be  
349 considered carefully when interpreting the results.



350

## 351 **Conclusion**

352

353 Investigating regional pathogen transmission can provide insight into transmission dynamics  
354 and guide infection prevention and control. Here we provide a framework for studying regional  
355 pathogen transmission within and between healthcare facilities using whole-genome  
356 sequencing data, and implement these methods in the easy-to-use R package regentrans.  
357 regentrans allows users to interrogate the transmission dynamics of pathogens using various  
358 metrics of genomic relatedness, including SNV-threshold and threshold-free approaches. Using  
359 several complementary methods to investigate intra- and inter-facility transmission allows  
360 investigators to gain a better understanding of the robustness of their findings and provide  
361 different insights into transmission dynamics in the region of interest. Therefore, we believe that  
362 this tool will be a useful resource for researchers and public health practitioners interested in  
363 investigating regional pathogen transmission.

364

## 365 **Authors and contributors**

366

367 All authors developed methodology and reviewed and edited the manuscript. SH, ZL, and ESS  
368 authors conceptualized the project. SH and ZL curated the data, designed the software, and  
369 visualized the results. ZL, JW, and ESS acquired funding. ZL and ESS wrote the original  
370 manuscript draft. ESS supervised the project.

371

## 372 **Conflicts of interest**

373

374 The authors declare that there are no conflicts of interest.

375

## 376 **Funding information**

377

378 ESS received support from the National Institutes of Health under Grant No. 1R01AI148259-01.  
379 ZL received support from the National Science Foundation Graduate Research Fellowship  
380 Program under Grant No. DGE 1256260. JW received support from the Canadian Institutes of  
381 Health Research fellowship [grant number 201711MFE-396343-165736] and the Michigan  
382 Institute for Clinical and Health Research (MICHR) Postdoctoral Translational Scholars Program  
383 (PTSP). Any opinions, findings, and conclusions or recommendations expressed in this material  
384 are those of the authors and do not necessarily reflect the views of the National Science  
385 Foundation. The funding bodies had no role in the design of the study or collection, analysis,  
386 and interpretation of data, or in writing the manuscript.

387

## 388 **Ethical approval**

389

390 N/A

391

## 392 **Consent for publication**

393

394 N/A

395

## 396 Acknowledgements

397

398 We would like to thank Emily Benedict for testing out the alpha version of our package.

399

## 400 References

401

- 402 1. **Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, et al.** Discovery, research,  
403 and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and  
404 tuberculosis. *The Lancet Infectious Diseases*. Epub ahead of print 21 December 2017. DOI:  
405 10.1016/S1473-3099(17)30753-3.
- 406 2. **Blanco N, O'Hara LM, Harris AD.** Transmission pathways of multidrug-resistant organisms  
407 in the hospital setting: a scoping review. *Infection Control & Hospital Epidemiology*  
408 2019;40:447–456.
- 409 3. **Wang J, Foxman B, Pirani A, Lapp Z, Mody L, et al.** Application of combined genomic  
410 and transfer analyses to identify factors mediating regional spread of antibiotic resistant  
411 bacterial lineages. *Clin Infect Dis*. DOI: 10.1093/cid/ciaa364.
- 412 4. **Han JH, Lapp Z, Bushman F, Lautenbach E, Goldstein EJC, et al.** Whole-Genome  
413 Sequencing To Identify Drivers of Carbapenem-Resistant *Klebsiella pneumoniae*  
414 Transmission within and between Regional Long-Term Acute-Care Hospitals. *Antimicrobial*  
415 *Agents and Chemotherapy*;63. Epub ahead of print 1 November 2019. DOI:  
416 10.1128/AAC.01622-19.
- 417 5. **Paul P, Slayton RB, Kallen AJ, Walters MS, Jernigan JA.** Modeling Regional  
418 Transmission and Containment of a Healthcare-associated Multidrug-resistant Organism.  
419 *Clin Infect Dis*. DOI: 10.1093/cid/ciz248.
- 420 6. **Snitkin ES, Won S, Pirani A, Lapp Z, Weinstein RA, et al.** Integrated genomic and  
421 interfacility patient-transfer data reveal the transmission pathways of multidrug-resistant  
422 *Klebsiella pneumoniae* in a regional outbreak. *Science Translational Medicine*  
423 2017;9:eaan0093.
- 424 7. **Lapp Z, Crawford R, Miles-Jay A, Pirani A, Trick WE, et al.** Regional Spread of blaNDM-  
425 1-Containing *Klebsiella pneumoniae* ST147 in Post-Acute Care Facilities. *Clinical Infectious*  
426 *Diseases*. Epub ahead of print 17 May 2021. DOI: 10.1093/cid/ciab457.
- 427 8. **Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, et al.** Rapid phylogenetic  
428 analysis of large samples of recombinant bacterial whole genome sequences using  
429 Gubbins. *Nucleic Acids Res* 2015;43:e15–e15.
- 430 9. **Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ.** IQ-TREE: A Fast and Effective  
431 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol*  
432 2015;32:268–274.
- 433 10. **Lapp Z, Han JH, Choudhary D, Castaneda S, Pirani A, et al.** Fitness barriers to spread of  
434 colistin-resistant *Klebsiella pneumoniae* overcome by establishing niche in patient  
435 population with elevated colistin use. *medRxiv* 2021;2021.06.11.21258758.
- 436 11. **Li H, Durbin R.** Fast and accurate short read alignment with Burrows-Wheeler transform.  
437 *Bioinformatics* 2009;25:1754–1760.
- 438 12. **Paradis E, Schliep K.** ape 5.0: an environment for modern phylogenetics and evolutionary  
439 analyses in R. *Bioinformatics* 2019;35:526–528.
- 440 13. **Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, et al.** A genomic surveillance  
441 framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex.  
442 *Nat Commun* 2021;12:4188.

- 443 14. **Sherry NL, Lee RS, Gorrie CL, Kwong JC, Stuart RL, et al.** Pilot study of a combined  
444 genomic and epidemiologic surveillance program for hospital-acquired multidrug-resistant  
445 pathogens across multiple hospital networks in Australia. *Infection Control & Hospital*  
446 *Epidemiology* 2021;42:573–581.
- 447 15. **Gouliouris T, Coll F, Ludden C, Blane B, Raven KE, et al.** Quantifying acquisition and  
448 transmission of *Enterococcus faecium* using genomic surveillance. *Nature Microbiology*  
449 2020;1–9.
- 450 16. **Coll F, Raven KE, Knight GM, Blane B, Harrison EM, et al.** Definition of a genetic  
451 relatedness cutoff to exclude recent transmission of methicillin-resistant *Staphylococcus*  
452 *aureus*: a genomic epidemiology analysis. *The Lancet Microbe* 2020;1:e328–e335.
- 453 17. **Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, et al.** Beyond the SNP Threshold:  
454 Identifying Outbreak Clusters Using Inferred Transmissions. *Mol Biol Evol* 2019;36:587–  
455 603.
- 456 18. **Hawken SE, Yelin RD, Lolans K, Weinstein RA, Lin MY, et al.** Threshold-free genomic  
457 cluster detection to track transmission pathways in healthcare settings. *medRxiv*  
458 2021;2020.09.26.20200097.
- 459 19. **Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, et al.** Genome-scale rates of  
460 evolutionary change in bacteria. *Microb Genom*;2. Epub ahead of print 30 November 2016.  
461 DOI: 10.1099/mgen.0.000094.
- 462 20. **Bowers JR, Kitchel B, Driebe EM, MacCannell DR, Roe C, et al.** Genomic Analysis of the  
463 Emergence and Rapid Global Dissemination of the Clonal Group 258 *Klebsiella*  
464 *pneumoniae* Pandemic. *PLoS ONE* 2015;10:e0133727.
- 465 21. **Donker T, Reuter S, Scriberras J, Reynolds R, Brown NM, et al.** Population genetic  
466 structuring of methicillin-resistant *Staphylococcus aureus* clone EMRSA-15 within UK  
467 reflects patient referral patterns. *Microb Genom*;3. Epub ahead of print 4 July 2017. DOI:  
468 10.1099/mgen.0.000113.
- 469 22. **R Core Team.** R: A Language and Environment for Statistical Computing. [https://www.R-](https://www.R-project.org/)  
470 [project.org/](https://www.R-project.org/) (2020).
- 471 23. **Wickham H, Averick M, Bryan J, Chang W, McGowan L, et al.** Welcome to the  
472 Tidyverse. *Journal of Open Source Software* 2019;4:1686.
- 473 24. **Revell LJ.** phytools: an R package for phylogenetic comparative biology (and other things).  
474 *Methods in Ecology and Evolution* 2012;3:217–223.
- 475 25. **Csardi G, Nepusz T.** The Igraph Software Package for Complex Network Research.  
476 *InterJournal* 2005;Complex Systems:1695.
- 477 26. **Bengtsson H, Team RC.** future.apply: Apply Function to Elements in Parallel using  
478 Futures. <https://CRAN.R-project.org/package=future.apply> (2020, accessed 26 August  
479 2020).
- 480 27. **Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y.** ggtree: an r package for visualization and  
481 annotation of phylogenetic trees with their covariates and other associated data. *Methods in*  
482 *Ecology and Evolution* 2017;8:28–36.
- 483 28. **Kolde R.** *pheatmap: Pretty Heatmaps*. <https://CRAN.R-project.org/package=pheatmap>  
484 (2019, accessed 15 April 2020).
- 485 29. **Malone B.** *tangles: Anonymization of Spatial Point Patterns and Raster Objects*.  
486 <https://CRAN.R-project.org/package=tangles> (2019, accessed 24 July 2021).
- 487 30. **Popovich KJ, Thiede SN, Zawitz C, Aroutcheva A, Payne D, et al.** Genomic  
488 Epidemiology of MRSA During Incarceration at a Large Inner-City Jail. *Clin Infect Dis*  
489 2021;ciaa1937.
- 490 31. **Hall MD, Holden MT, Srisomang P, Mahavanakul W, Wuthiekanun V, et al.** Improved  
491 characterisation of MRSA transmission using within-host bacterial sequence diversity. *eLife*  
492 2019;8:e46402.
- 493 32. **Mooney SJ, Garber MD.** Sampling and Sampling Frames in Big Data Epidemiology. *Curr*

494 *Epidemiol Rep* 2019;6:14–22.  
 495 33. Murray M. Sampling Bias in the Molecular Epidemiology of Tuberculosis. *Emerg Infect Dis*  
 496 2002;8:363–369.  
 497

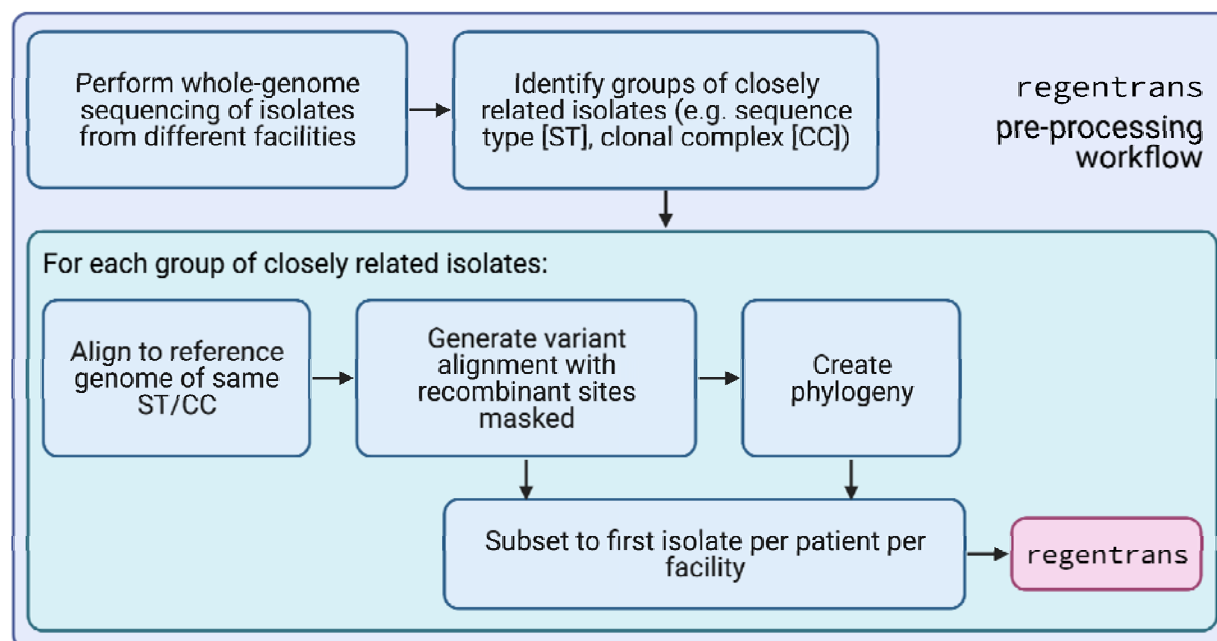
498 **Figures and tables**

499  
 500 **Table 1: Questions regentrans can help investigate, and corresponding regentrans**  
 501 **functions**  
 502

	Question	Method	regentrans function	Reference
Q0	How do you choose pairwise SNV distance thresholds?	Visualize the fraction of intra-facility pairs for various pairwise SNV distances	get_frac_intra	[4]
Q1	Is transmission occurring within and/or between locations?	Phylogenetic clustering of isolates from the same location	get_clusters	[6]
		Pairwise SNV distances within and between facilities	get_pair_types	[7]
Q2	What locations is transmission occurring within/between?	Population-level similarity between locations	get_facility_fsp	[21]
		Number of closely related pairs within and between facilities	-	-
Q3	Have transmission dynamics changed over time?	Methods above but split over time	-	[7]

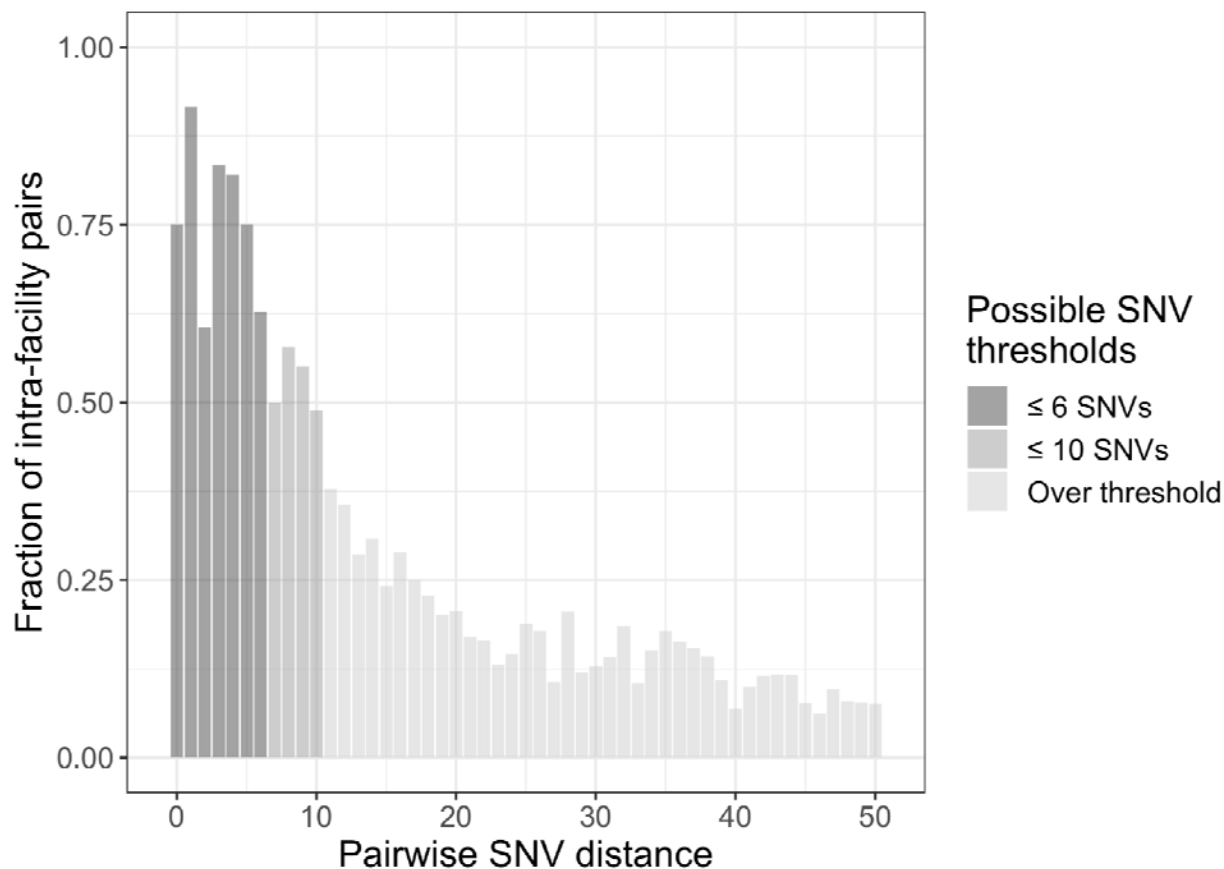
Q4	Is transmission occurring along paths of higher patient/person flow?	Compare patient/person flow between locations to inter-location pairwise SNV distances or Fsp	get_patient_flow w summarize_pairs	[3]
Q5	Are there any observable geographic trends in prevalence/transmission?	Visualize geographic distribution of prevalence and closely related pairs or Fsp	-	[4]

503  
504



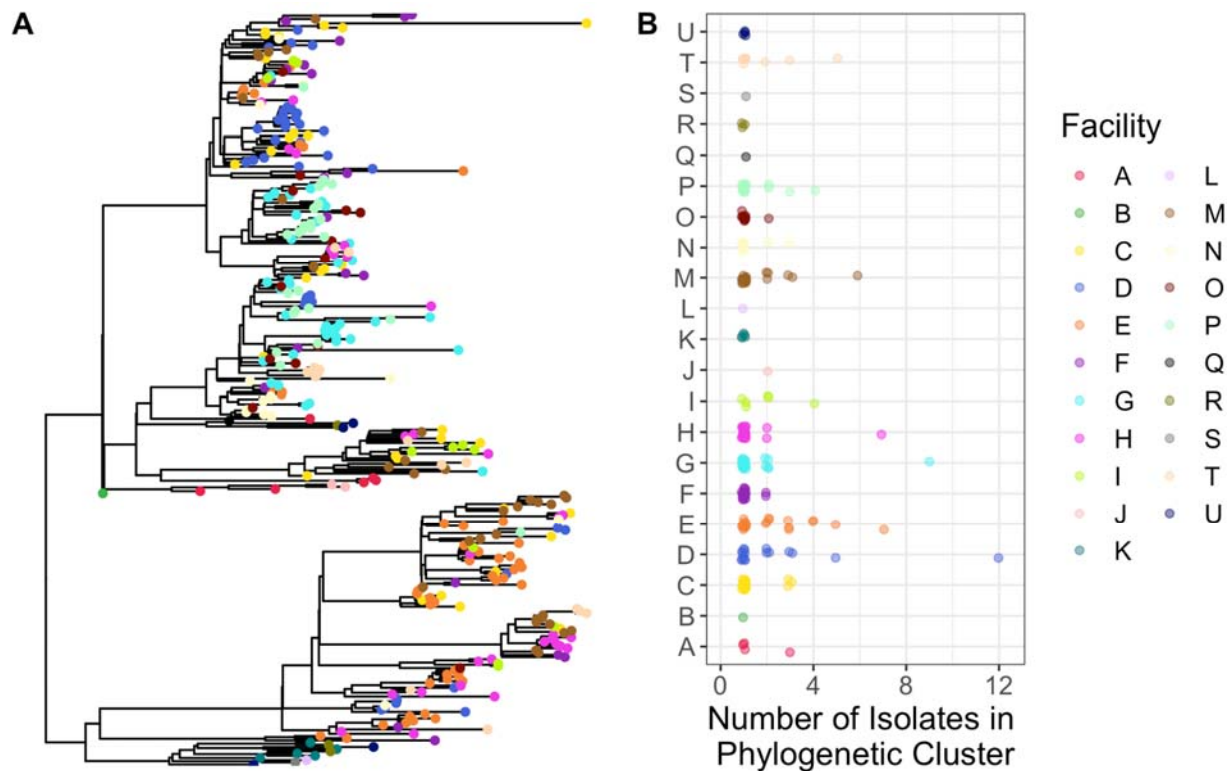
505  
506  
507  
508  
509

**Figure 1: regentrans data pre-processing workflow.** Whole-genome sequences of closely related isolates are aligned to a reference genome, non-recombinant variants are identified, a phylogeny is recreated, and the data is subset to the first isolate per patient per facility.

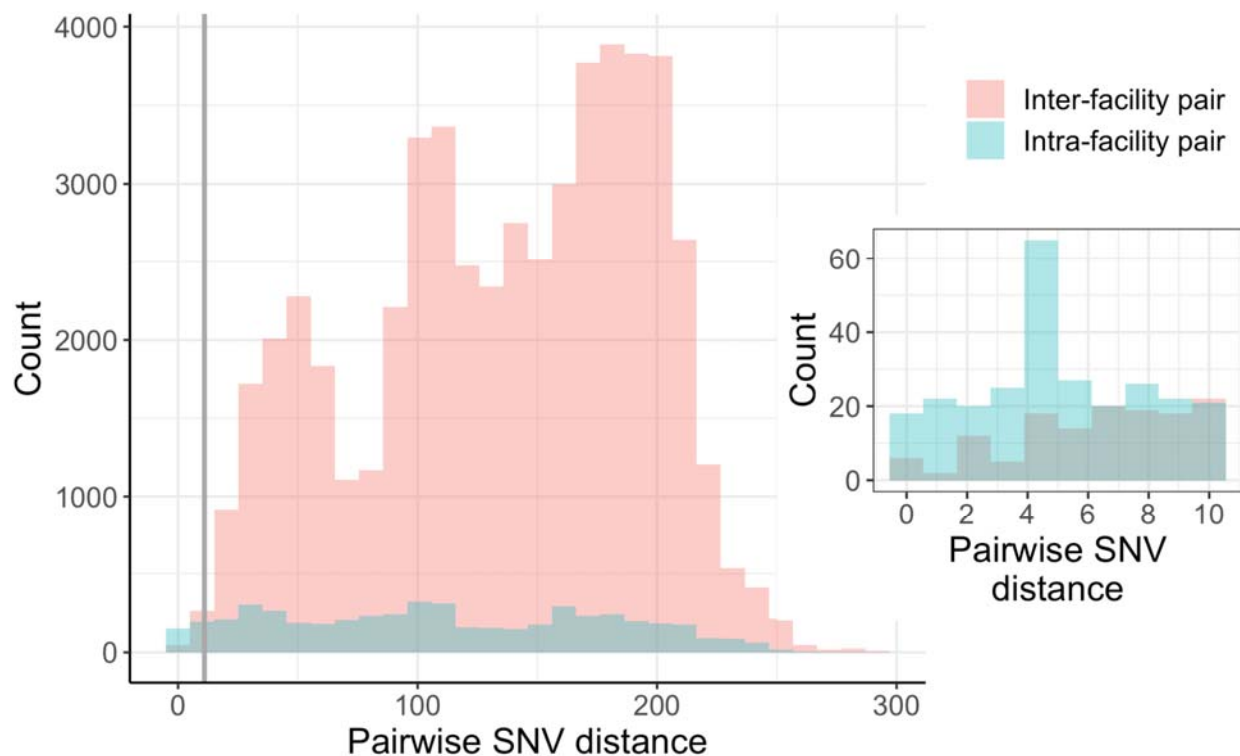


510  
511 **Figure 2: Choosing pairwise SNV distance thresholds.** Plotting the fraction of intra-facility  
512 pairs for various pairwise SNV distances can help identify drops in intra-facility pair fraction that  
513 may indicate a reasonable pairwise SNV distance threshold, assuming that intra-facility  
514 transmission is more common than inter-facility transmission. This data can be generated using  
515 the `get_frac_intra` function.  
516

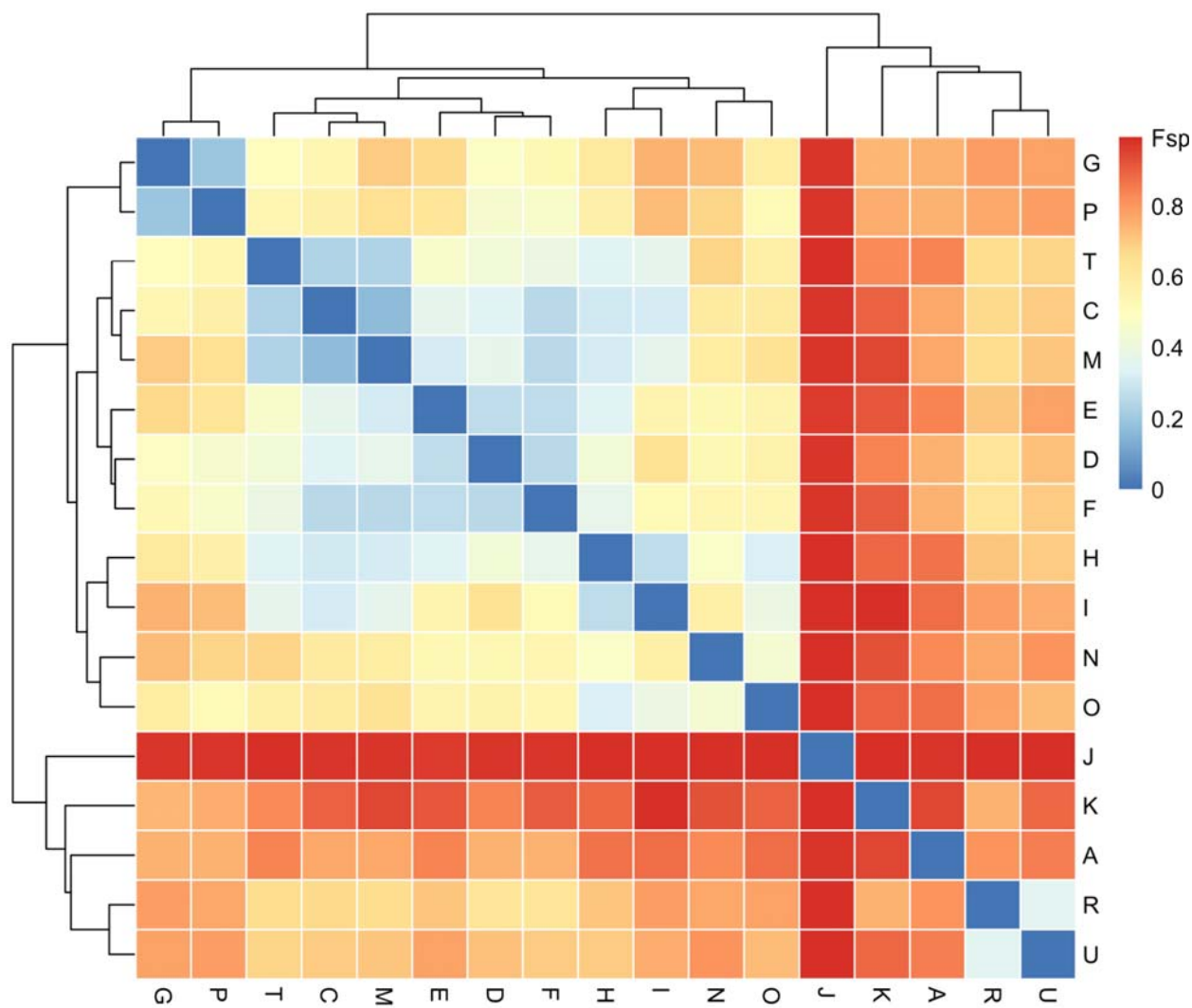




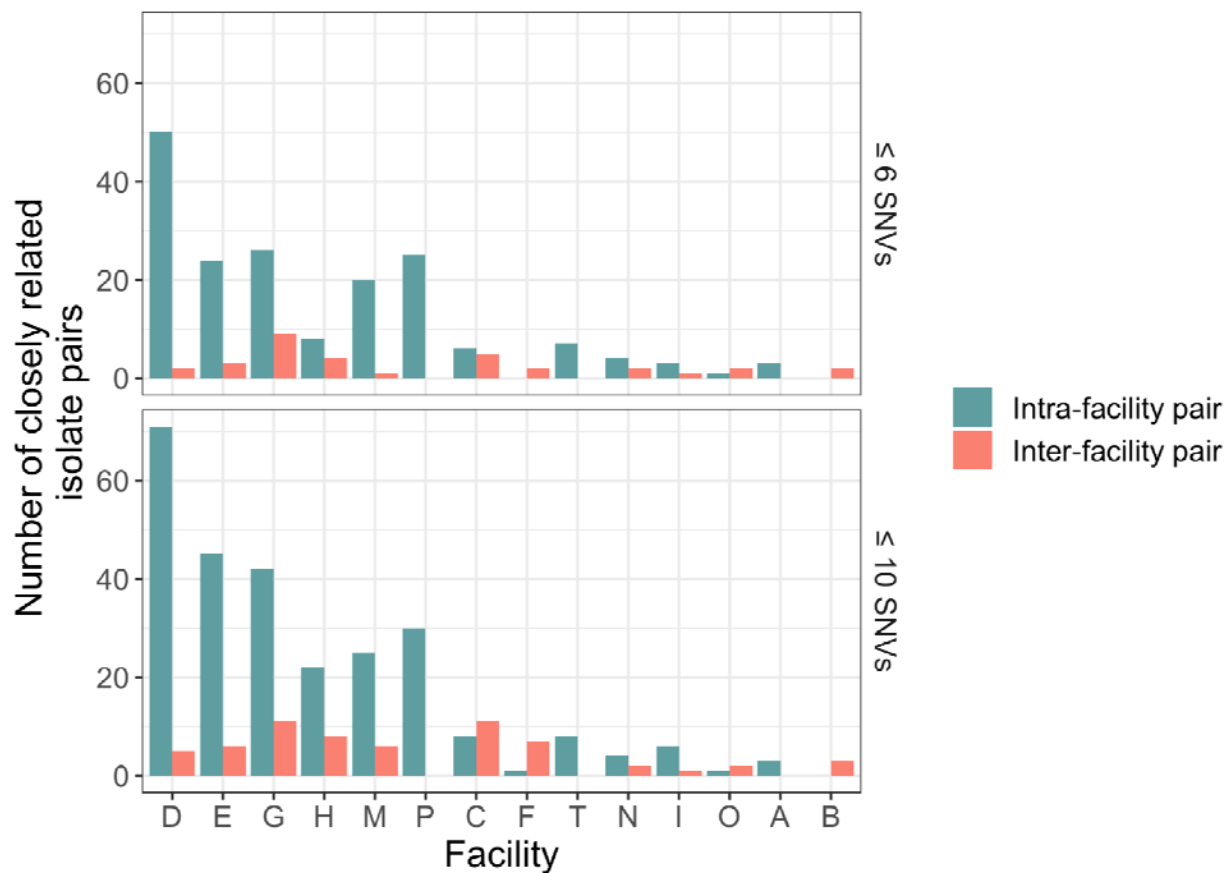
517  
518 **Figure 3: Clusters of isolates from the same facility suggest intra-facility transmission.**  
519 (A) Mapping isolate location on the phylogeny provides a visual for the extent of clustering by  
520 facility. Here we can see clustering of isolates from the same facility in several subclades of the  
521 phylogeny. (B) Quantification of the size of phylogenetic clusters from a single facility using  
522 `get_clusters`.  
523



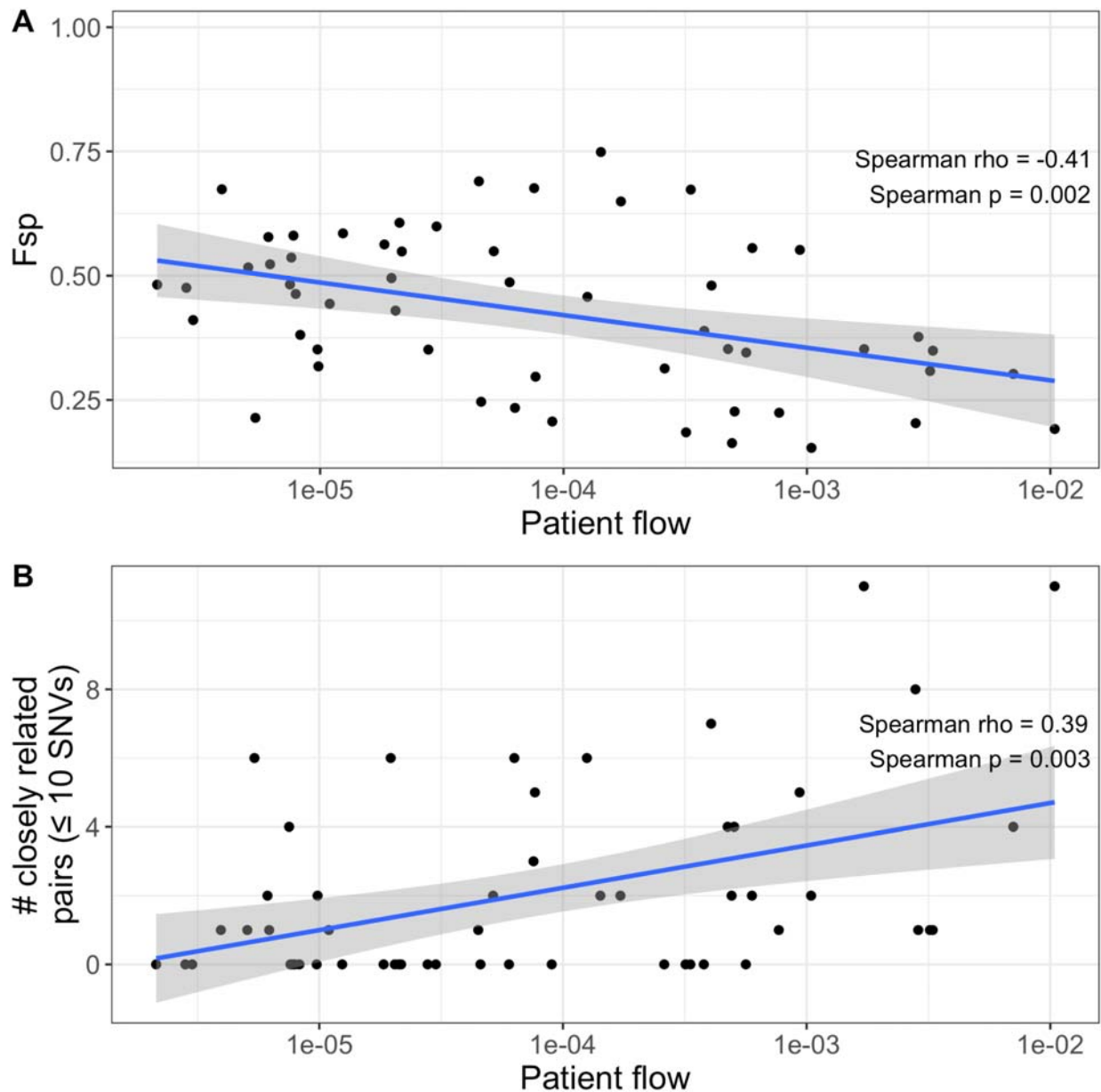
524  
525 **Figure 4: Pairwise single nucleotide variant (SNV) distances between facilities suggest**  
526 **recent intra- and inter-facility transmission.** Data generated using the `get_pair_types`  
527 function. Inset shows all pairs with a pairwise SNV distance  $\leq 10$ , which we consider indicative  
528 of recent transmission (see Q0 on SNV distance thresholds in main text). This plot also  
529 indicates that transmission is likely occurring both within and between facilities.



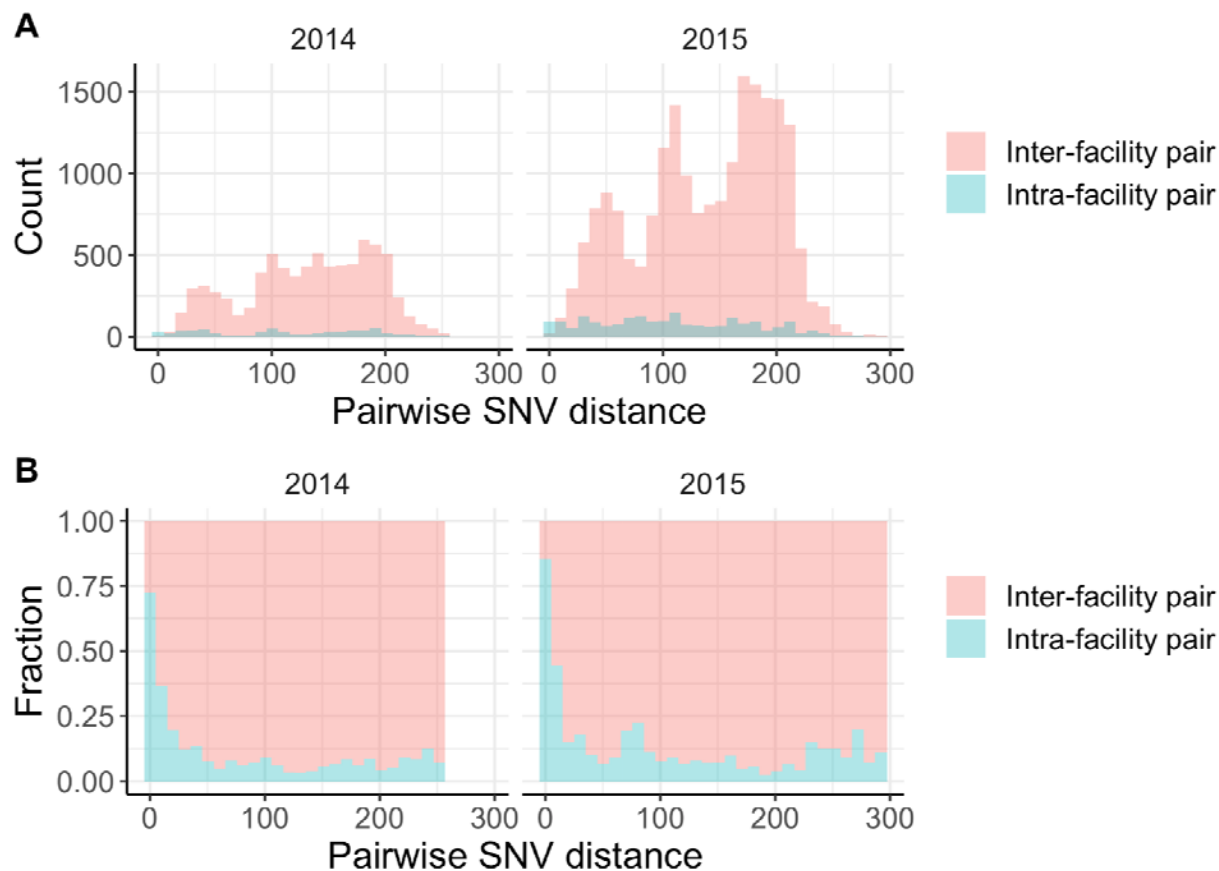
530  
531 **Figure 5: Some facility pairs have similar populations, indicating potential transmission**  
532 **between them.** Fsp was calculated using the get\_facility\_fsp function in regentrans. Rows and  
533 columns are facilities. Lower Fsp indicates more similar populations and thus more putative  
534 transmission.  
535



536  
537 **Figure 6: Some facilities have many closely related isolates, indicating potential intra-**  
538 **and inter-facility transmission.**



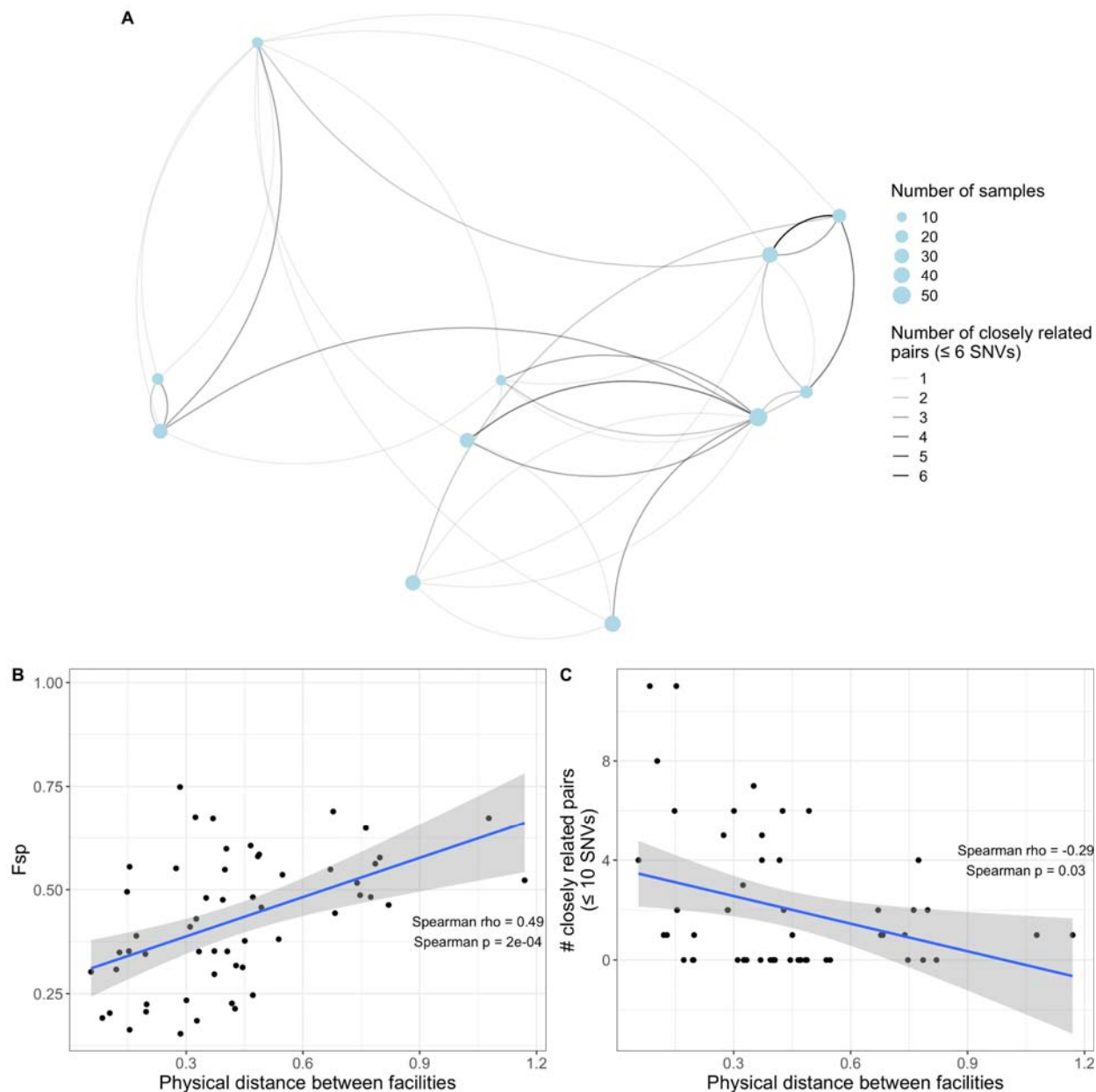
539  
540 **Figure 7: Facilities with more patient flow tend to have more similar CRKP populations.**  
541 (A) Patient flow and Fsp are negatively correlated. (B) Patient flow and number of closely  
542 related isolate pairs are positively correlated. Patient flow is the path of maximum patient flow.  
543 For this analysis we considered indirect transfers as LTACHs are often not connected by direct  
544 transfers, but rather are connected by transfers to an intermediate facility such as an acute care  
545 hospital. Lines were plotted using `ggplot::geom_smooth()` with the 'lm' method.  
546



547  
548 **Figure 8: Pairwise SNV distance distribution does not change over time.** (A) Count of  
549 pairwise SNV distances faceted by year. (B) Fraction of intra- vs. inter-facility pairwise SNV  
550 distances faceted by year. Trends are similar across both years.



551



552

553

554

555

556

557

558

559

560

561

**Figure 9: Geographically close facilities are often connected by closely related isolate pairs.** (A) Facilities are located as they are geographically in space but latitude and longitude are de-identified by horizontal, vertical, and rotational shifts. The smaller SNV threshold was chosen for visualization purposes. (B) Physical distance between facilities is positively correlated with F<sub>sp</sub>. (C) Physical distance between facilities is negatively correlated with number of closely related isolate pairs (≤10 SNVs). The larger SNV threshold was chosen to have a wider distribution of number of closely related isolate pairs. Physical distance was calculated as the shortest distance between the points of latitude and longitude for the facility pair. Lines in panels B and C were plotted using `ggplot::geom_smooth()` with the 'lm' method.