regentrans: a framework and R package for using genomics to study regional pathogen transmission

Sophie Hoffman, BS*[1]; Zena Lapp, BA*[1]; Joyce Wang, PhD[2]; Evan S Snitkin, PhD[2,3]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, 1150 W. Medical Center Dr. Ann Arbor, MI, 48109-5680
[2]Department of Microbiology and Immunology, University of Michigan, 1150 W. Medical Center Dr. Ann Arbor, MI, 48109-5680
[3]Department of Medicine, Division of Infectious Diseases, University of Michigan, 1150 W. Medical Center Dr. Ann Arbor, MI, 48109-5680

*These authors contributed equally to this work.

Corresponding author:
Evan Snitkin
esnitkin@umich.edu

Keywords

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

ORCIDs

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

Repositories

https://github.com/Snitkin-Lab-Umich/regentrans/

Abstract

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

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

Impact statement

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

Data summary

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

- The regentrans R package can be downloaded from GitHub: https://github.com/Snitkin-Lab-Umich/regentrans/
- The manuscript figures are generated from regentrans example data and can also be found on GitHub: https://github.com/Snitkin-Lab-Umich/regentrans/tree/master/vignettes/manuscript_figures
- The example data used in the package and manuscript is from BioProject accession no. PRJNA415194. The metadata corresponding to these sequences can be found on the SRA Run Selector (isolate column) and as example data in the regentrans package.

Introduction

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

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

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

Investigating regional transmission patterns

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

Questions

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

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

Data

<u>Data required</u>

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

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

Data pre-processing

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

Datasets used in the package and for analyses

Genomic data

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

Patient transfer data

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

Methods

Q0: How do you choose pairwise SNV distance thresholds?

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

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

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

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

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

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

Investigating intra-facility transmission using the phylogeny

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

Investigating intra- and inter-facility transmission using pairwise SNV distances

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

<u>Q2: What facilities is transmission occurring within/between?</u>

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

Shared variants between facilities

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

Pairwise SNV distance threshold

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

Q3: Does transmission correlate with patient transfer?

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

Q4: Have transmission dynamics changed over time?

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

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

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

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

Package implementation

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

Additional possible uses

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

Cautionary notes on interpretation

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

Conclusion

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

Authors and contributors

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

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding information

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

Ethical approval

N/A

Consent for publication

N/A

Acknowledgements

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

References

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

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

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

Figures and tables

Table 1: Questions regentrans can help investigate, and corresponding regentrans functions

	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 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]

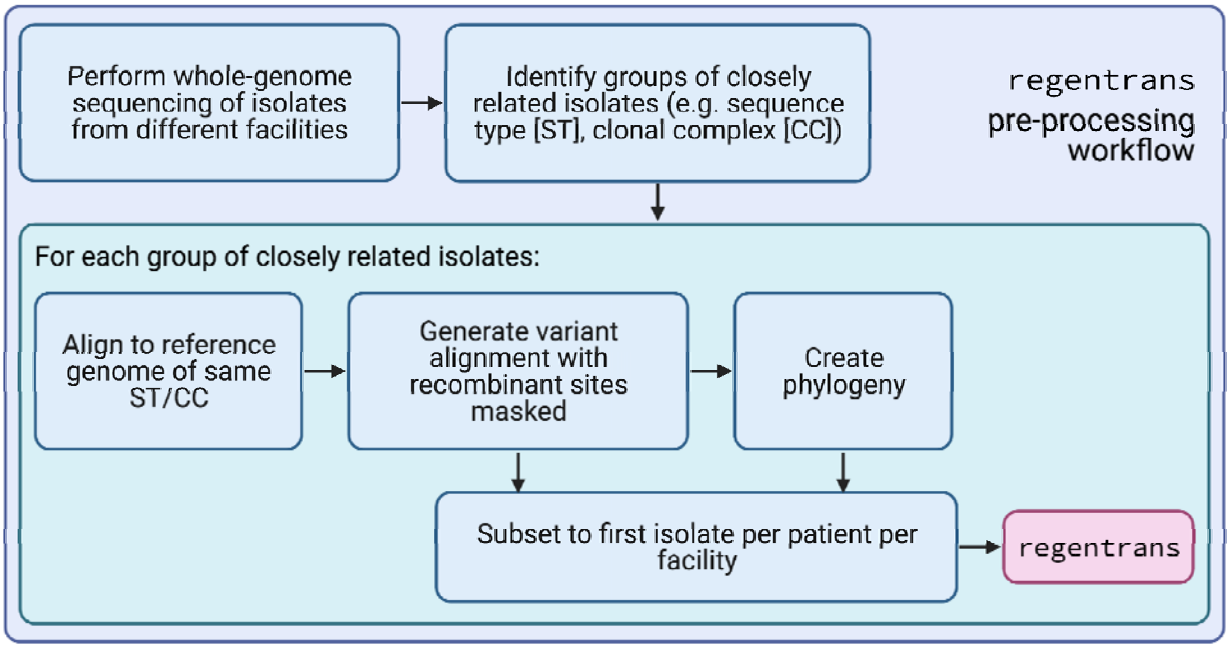

Figure 1: regentrans data pre-processing workflow. Whole-genome sequences of closely related isolates are aligned to a reference genome, non-recombinant variants are identified, a phylogeny is recreated, and the data is subset to the first isolate per patient per facility.

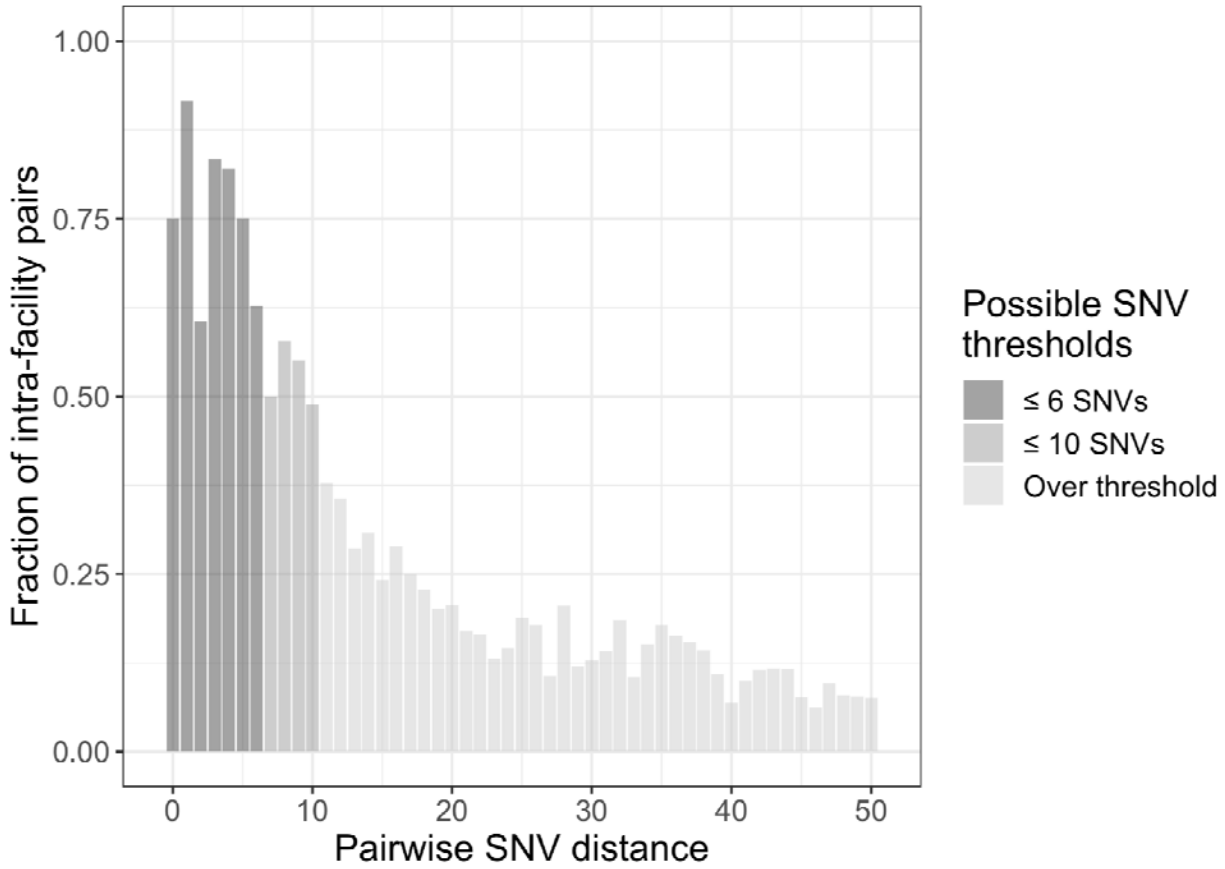

Figure 2: Choosing pairwise SNV distance thresholds. Plotting the fraction of intra-facility pairs for various pairwise SNV distances can help identify drops in intra-facility pair fraction that may indicate a reasonable pairwise SNV distance threshold, assuming that intra-facility transmission is more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

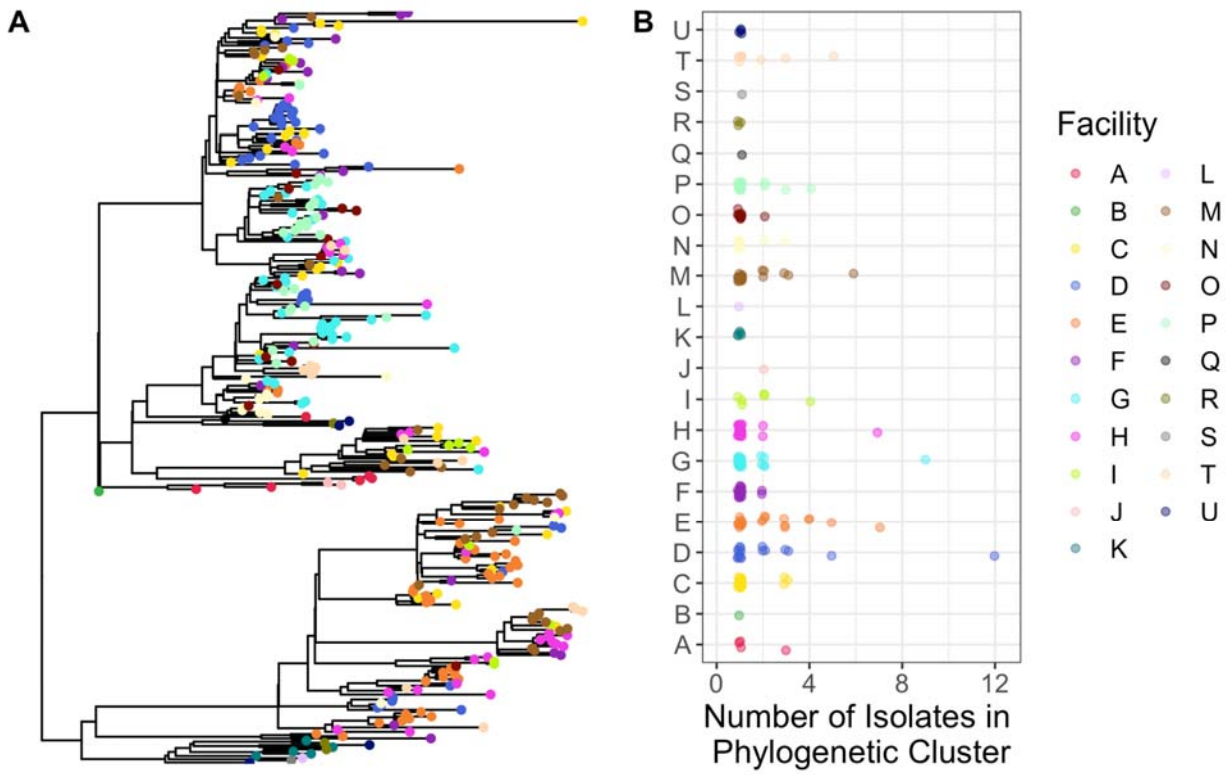

Figure 3: Clusters of isolates from the same facility suggest intra-facility transmission. (A) Mapping isolate location on the phylogeny provides a visual for the extent of clustering by facility. Here we can see clustering of isolates from the same facility in several subclades of the phylogeny. (B)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

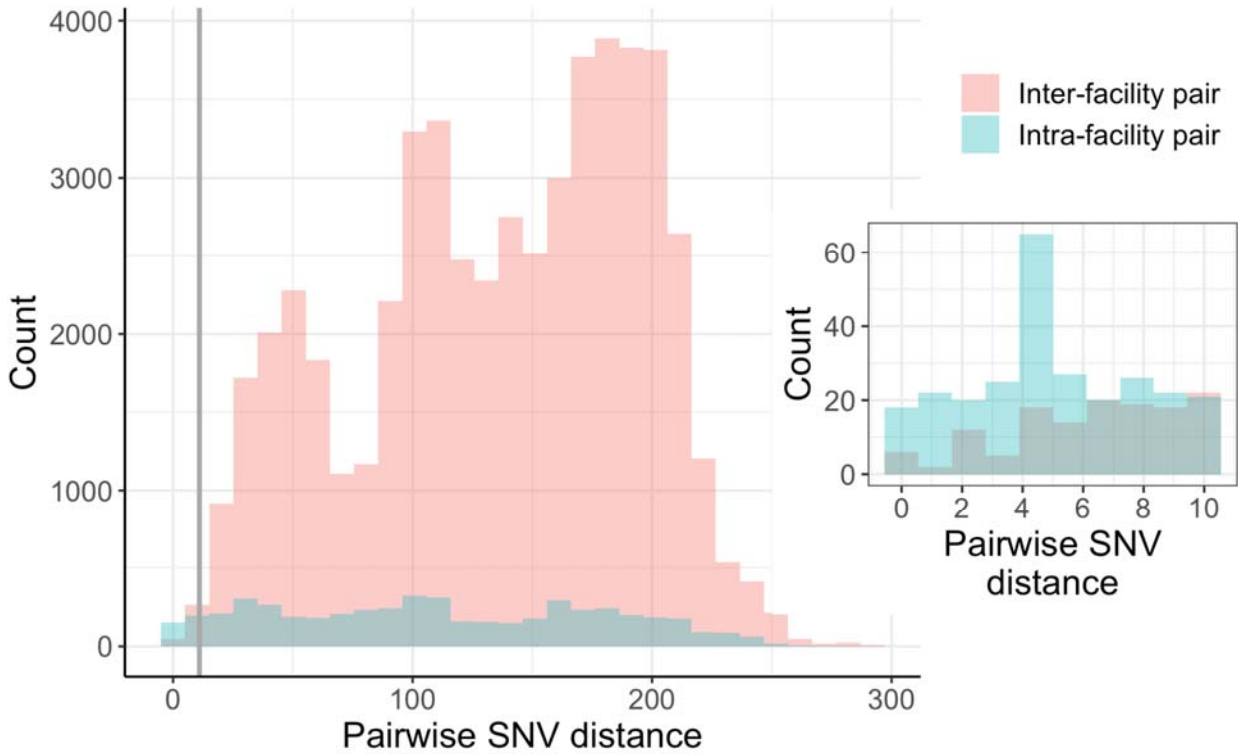

Figure 4: Pairwise single nucleotide variant (SNV) distances between facilities suggest recent intra- and inter-facility transmission. Data generated using the get_pair_types function. Inset shows all pairs with a pairwise SNV distance ≤ 10, which we consider indicative of recent transmission (see Q0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

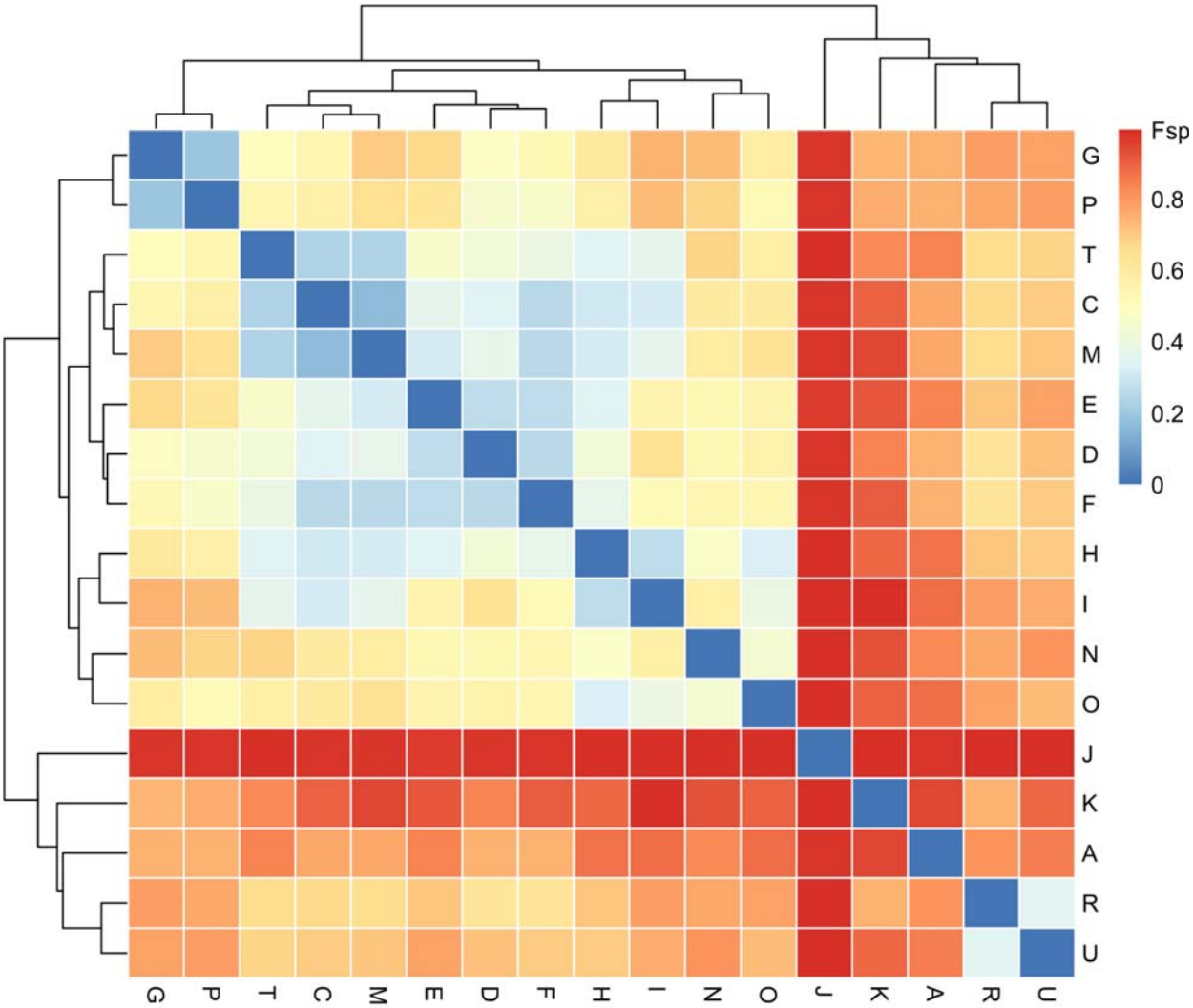

Figure 5: Some facility pairs have similar populations, indicating potential transmission between them. Fsp was calculated using the get_facility_fsp function in regentrans. Rows and columns are facilities. Lower Fsp indicates more similar populations and thus more putative transmission.

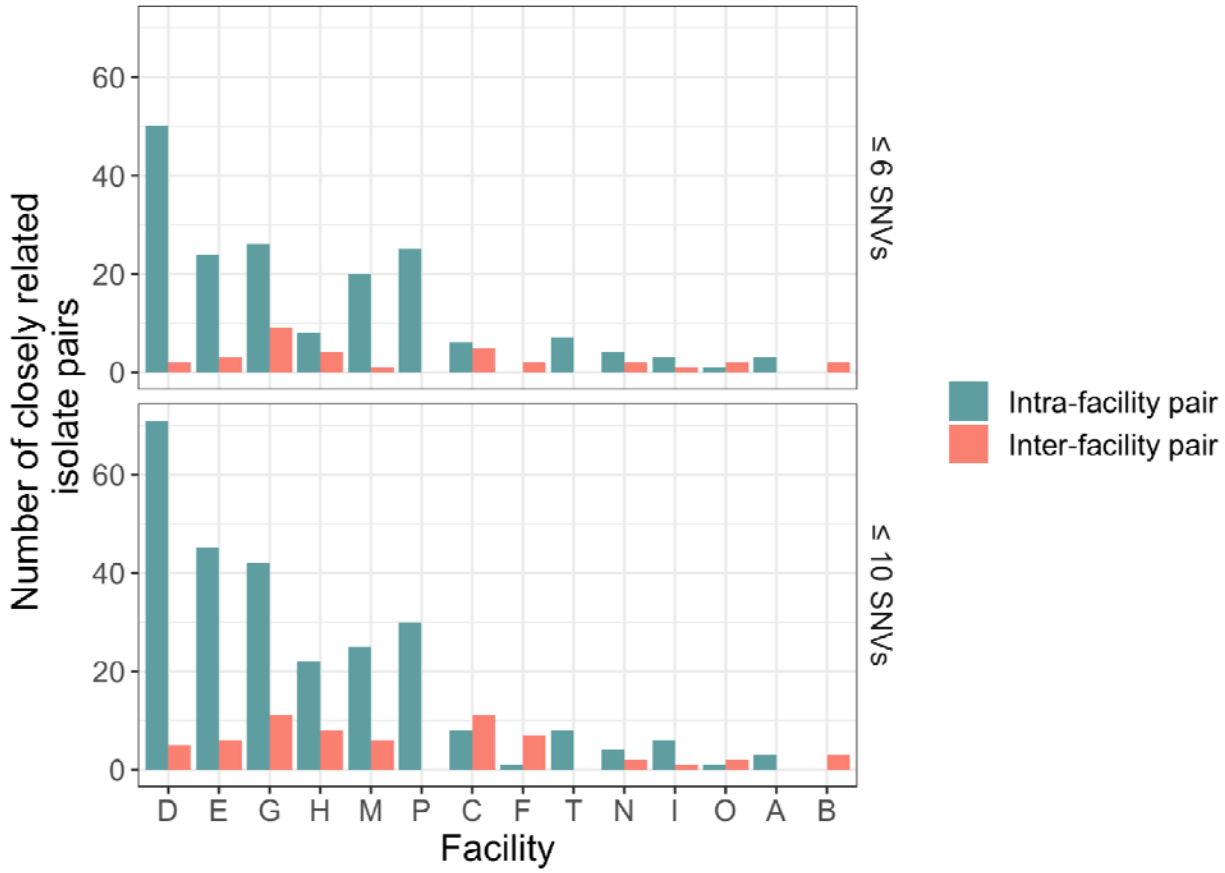

Figure 6: Some facilities have many closely related isolates, indicating potential intra- and inter-facility transmission.

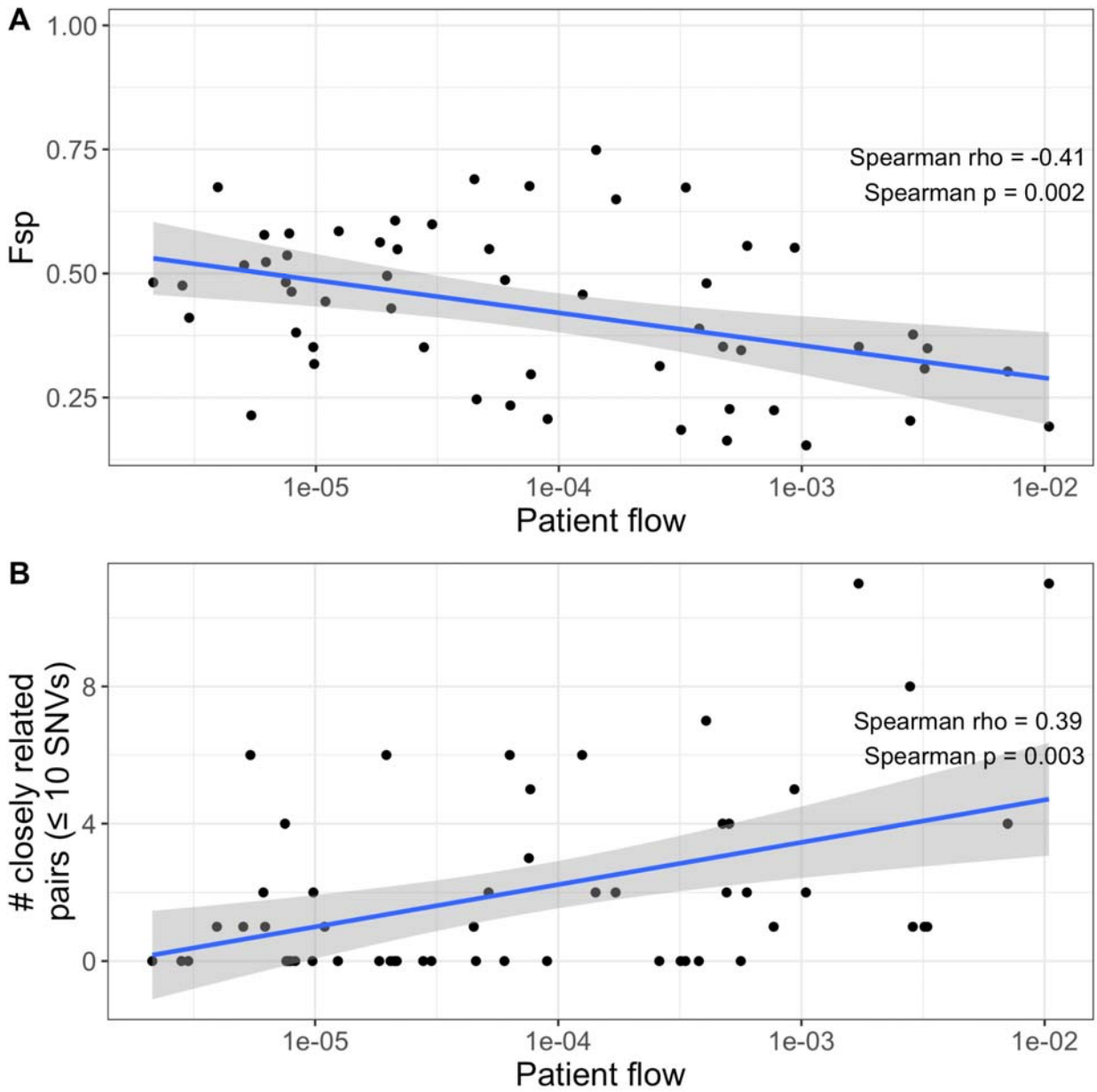

Figure 7: Facilities with more patient flow tend to have more similar CRKP populations. (A) Patient flow and Fsp are negatively correlated. (B) Patient flow and number of closely related isolate pairs are positively correlated. Patient flow is the path of maximum patient flow. For this analysis we considered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

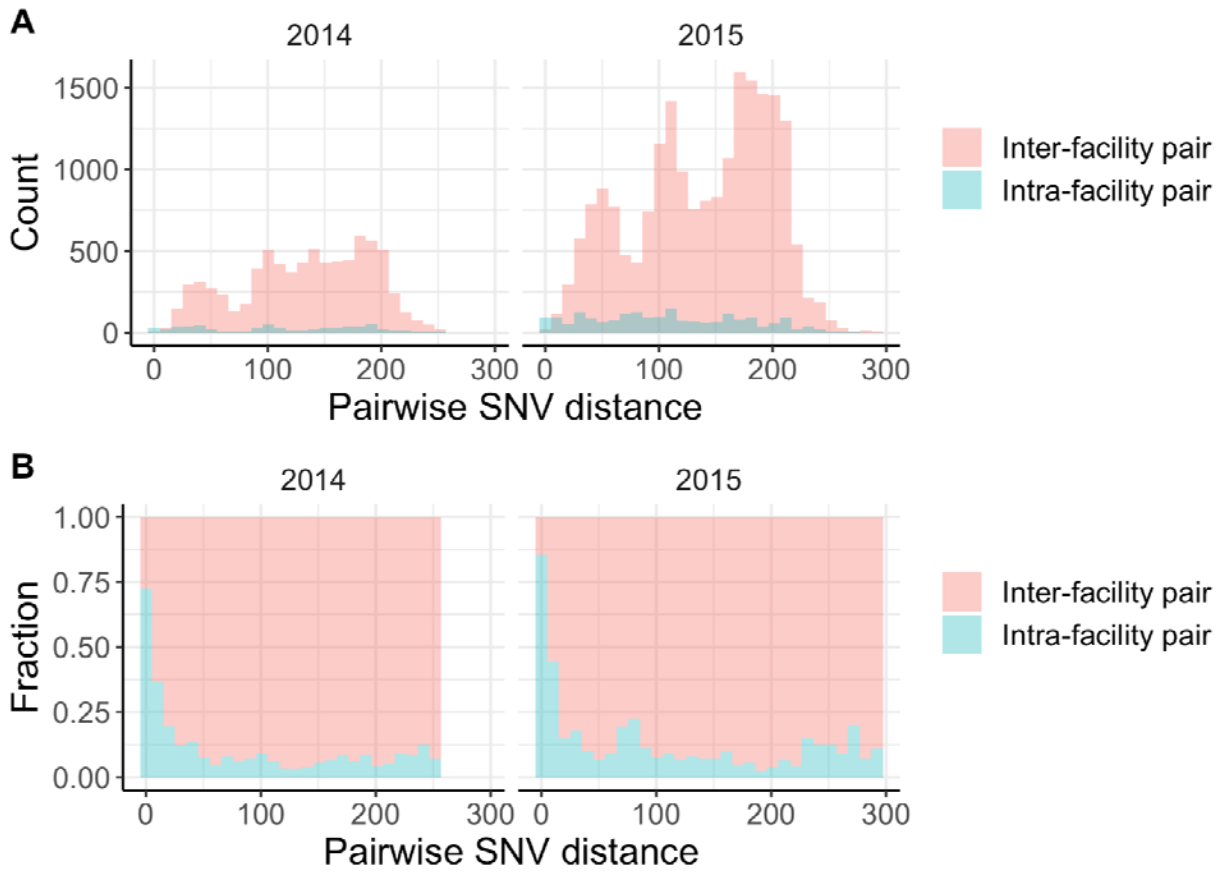

Figure 8: Pairwise SNV distance distribution does not change over time. (A) Count of pairwise SNV distances faceted by year. (B) Fraction of intra- vs. inter-facility pairwise SNV distances faceted by year. Trends are similar across both years.

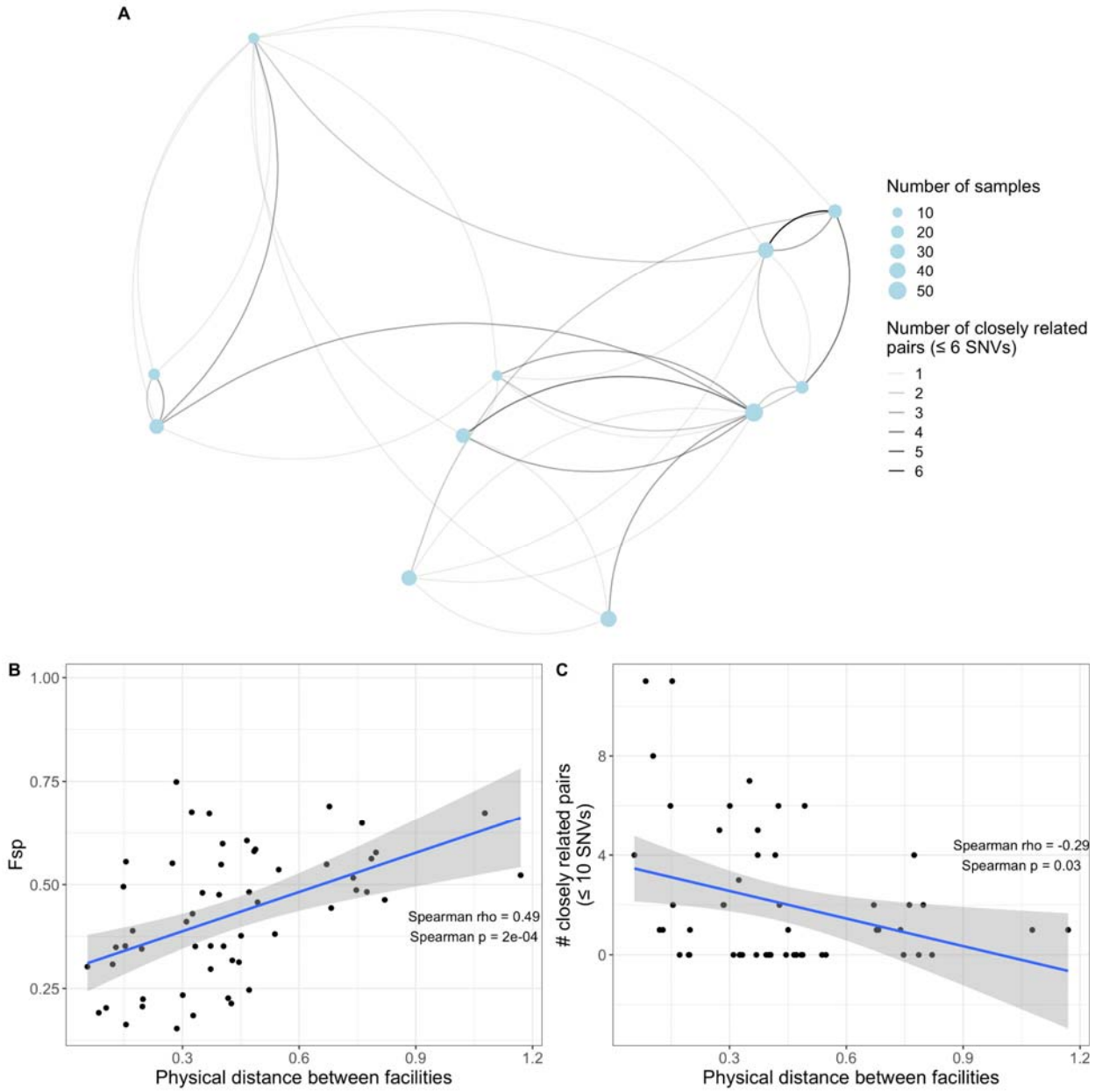

Figure 9: Geographically close facilities are often connected by closely related isolate pairs. (A) Facilities are located as they are geographically in space but latitude and longitude are de-identified by horizontal, vertical, and rotational shifts. The smaller SNV threshold was chosen for visualization purposes. (B) Physical distance between facilities is positively correlated with Fsp. (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.