A networks-based approach to spatial-genomic associations of *Brucella spp.* in southern Kazakhstan

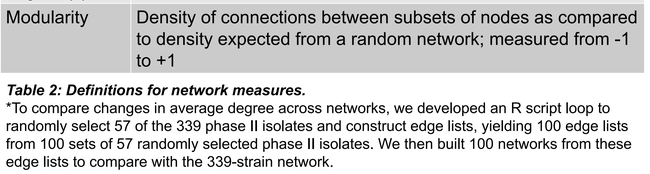
Lylybell Zhou[1,2], Sheldon G. Waugh[1,2], Igor Sytnik[3], Talgat Karibayev[3], Alim Aikimbayev[4], Mukhit Ornybayev[5], Nurgisa Rametov[5], Sue Hagius[6], Philip Elzer[6], Ted L. Hadfield[1,2], Gabriela Hamerlinck[7], José Miguel Ponciano[8], Mikeljon P. Nikolich[9], Jason K. Blackburn*[1,2]

[1]Spatial Epidemiology & Ecology Research Lab, Department of Geography, University of Florida, USA; [2]Emerging Pathogens Institute, University of Florida, USA; [3]National Reference Veterinary Center, Nur-Sultan 010000, Kazakhstan [4]Research Institute for Biological Special Problems, Otar, Zhambyl 080409, Kazakhstan; [5]Scientific Practical Center for Sanitary Epidemiological Expertise and Monitoring, Ministry of Health, Almaty 050008, Kazakhstan; [6]AgCenter, Louisiana State University, USA; [7]Department of Geography, University of Florida, USA; [8]Department of Biology, University of Florida, USA; [9]Bacterial Diseases Branch, Walter Reed Army Institute of Research, USA

Introduction

- Brucellosis is a bacterial disease caused by members of the *Brucella* genus, and it is one of the most common zoonoses found worldwide. Kazakhstan has persistently high brucellosis incidence rates, posing a serious public health and economic threat [1, 2]. In addition, the country faces several barriers for the control and prevention of brucellosis in animal and human populations [1, 3].
- To understand the molecular epidemiology of *Brucella,* previous efforts have employed phylogenetic trees and/or minimum spanning trees (MSTs). These tree building techniques search for or define relationships between individual strains characterized by genetic sequencing techniques [4]. However, phylogenetic trees are inherently aspatial, as they only serve to represent the genetic relationship between each node in the diagram.
- Previous work done by our group has demonstrated strong spatial-genomic associations in our *Brucella* MLVA data using the τ-statistic [5]. Here, we implement network analysis to describe the spatial distribution of *Brucella* genotypes and characterize the relationships between areas from which isolates were collected based on unique genotypes.

Data & Methods



- During a multi-part study, we used a database of 487 *B. melitensis* isolates collected from animals (domestic livestock) and humans during two survey phases. For all samples in the study, we used the MLVA-15 assay developed by Huynh et al. [6] using a modified protocol for the Beckman Coulter CEQ 8800 Genetic Analysis System.
- We used genotype designations from Waugh et al. [5]. A maximum likelihood phylogeny was constructed with the Lewis MK model using R 3.3.2 [7] and the R package *phangorn* [8]. Each isolate was subsequently categorized based on the position on the reconstructed tree to establish genotype groups. The analysis was analyzed with the Java-based program, *PhyloPart* [9].

	Phase I (2007 - 2008)			Phase II (2012 - 2013)		
	Total	Animal	Human	Total	Animal	Human
No. isolates	97	57	40	392	339	53
No. genotypes	26	10	16	30	18	12

Table 1: Overview of isolates and genotypes collected from each phase.

- A network can be defined as a set of nodes and a set of edges that connects nodes to each other.
- We defined nodes using geographic coordinates associated with each isolate. We created dummy villages for each coordinate point and assigned each village a unique ID number, resulting in a total of 114 villages (nodes).
- An edge was created between villages if the same genotype was found in each village; edges were then assigned weights based on the number of shared genotypes. For network analysis, we used Excel, Gephi [10] and R 4.0.2 with R packages *igraph* [11], *sf* [12] and *geokz* [13]. See Table 2 for network measures performed.

Measure	Definition
Degree	Number of connections a node has to other nodes; can be unweighted or weighted
Average degree (k)	Average number of edges per node and dependent on network size*
Modularity	Density of connections between subsets of nodes as compared to density expected from a random network; measured from -1 to +1

Table 2: Definitions for network measures.
*To compare changes in average degree across networks, we developed an R script loop to randomly select 57 of the 339 phase II isolates and construct edge lists, yielding 100 edge lists from 100 sets of 57 randomly selected phase II isolates. We then built 100 networks from these edge lists to compare with the 339-strain network.

Questions, comments, concerns? Leave a note below or scan the QR code:

SEER LAB UNIVERSITY OF FLORIDA

EMERGING PATHOGENS INSTITUTE
UNIVERSITY OF FLORIDA

Results

I. STUDY OVERVIEW

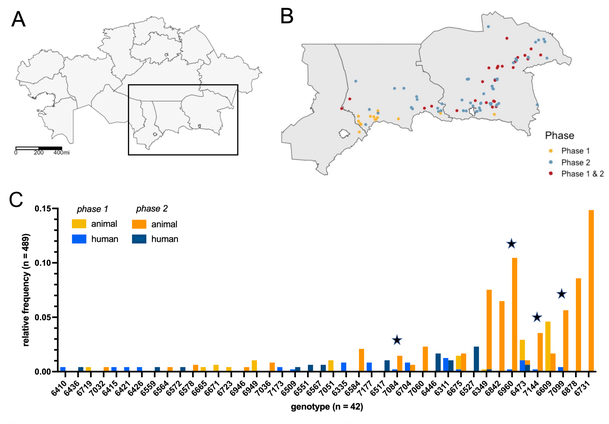


Figure 1: Study area and overview of isolates. (A) Isolates were collected from three districts in southern Kazakhstan, outlined in the black box. (B) Map of villages used as nodes in network construction, colored by the phases in which samples were collected from that village (phase I, phase II, or both phases). (C) Graph showing relative frequency of genotypes. Stars indicate genotypes in which isolates from phase I were found only in humans. All phase II genotypes were found only in animals.

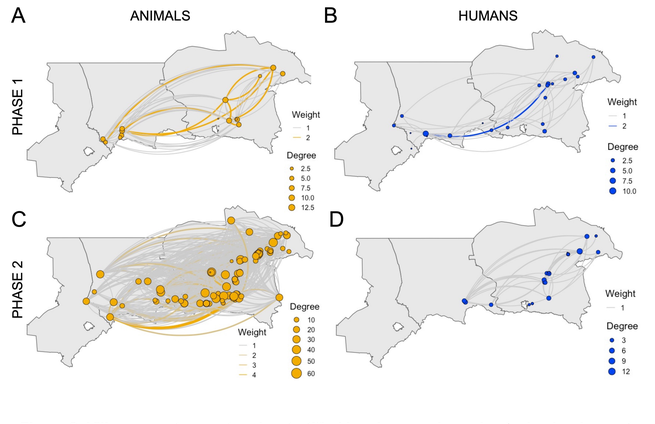


Figure 2: Village-genotype networks stratified by phase and species (animal or human). Animal-only (A) and human-only (B) isolates identified from phase I samples and animal-only (C) and human-only (D) isolates identified from phase II samples. Nodes (villages) represented sized by degree. Edges colored and sized based on edge weight, which was determined by the number of genotypes shared by the villages being connected by the edge.

II. NETWORK ANALYSIS

	Animals		Humans	
	Phase I	Phase II	Phase I	Phase II
Average degree	7.625	29.057, 5.438*	3.364	4.538
Assortativity (degree)	-0.038	0.0979, 0.447*	-0.347	0.642
Louvain communities	3	4	3	7
Modularity	0.131	0.300	0.528	0.572

Table 3: Network-level measures. *indicates the average value of metrics calculated from the 100 networks derived from random sampling.

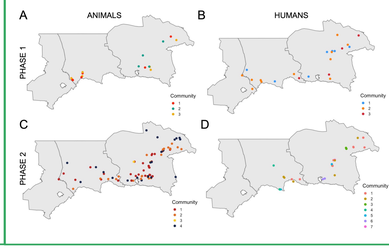


Figure 3: Isolate communities, stratified by phase and species. Animal-only (A) and human-only (B) isolates from phase I samples and animal-only (C) and human-only (D) isolates from phase II samples. Nodes colored by community membership as defined by the Louvain algorithm.

III. GENOTYPE NETWORKS

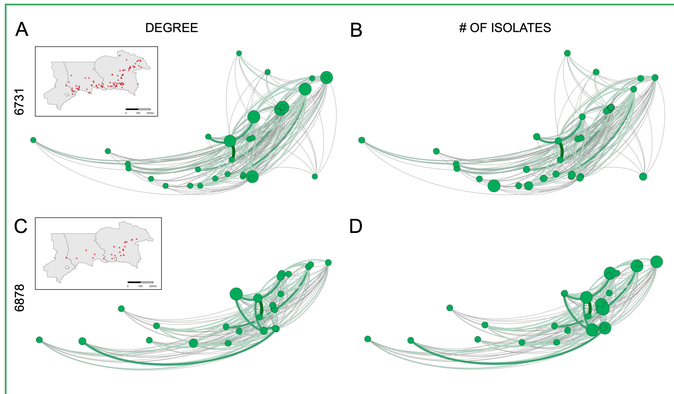


Figure 4: Genotype networks of 6731 with nodes sized by degree (A) and number of isolates (B); genotype networks of 6878 with nodes sized by degree (C) and number of isolates (D). Insets: (top) map of villages where each genotype was collected.

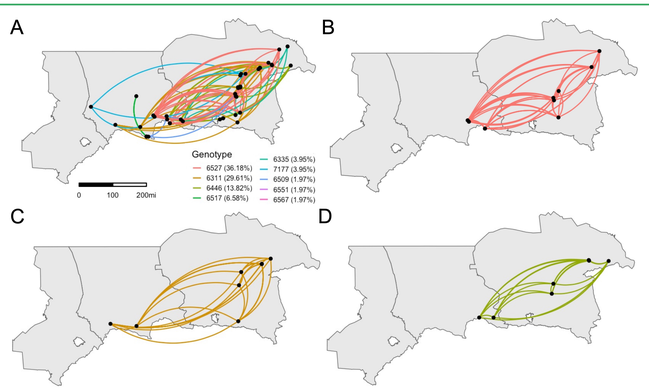


Figure 5: Individual networks of human-specific genotypes found across phases I and II. Edges colored by genotype (A). Individual networks for the three most prevalent human-only specific genotypes with edges colored as in panel A: 6527 (B), 6311 (B), and 6446 (C). Human-only genotypes found only in one or two villages were omitted.

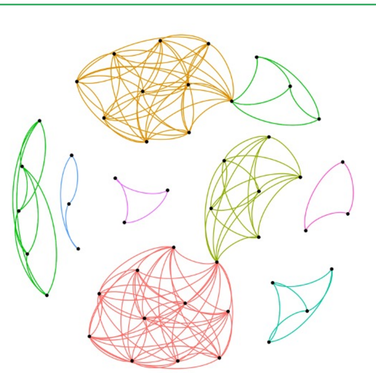


Figure 6: Network of human-only genotypes as visualized by the Fruchterman-Reingold layout. Nodes represent villages where human-only genotypes were isolated, and edges are colored by genotype. Genotypes found only in one or two villages are not shown.

Conclusion & Future Directions

- Calculating node degree and average degree for human and animal networks demonstrated an increase in the geographic spread and shared genotypes from phase I to phase II, most likely due to the expansion of sampling efforts.
- Lack of spatial structure and genotype-specificity to the Louvain communities could be due to a few genotype being widely spread across villages.
- Different spatial patterns of node sizes based on degree or number of isolates suggests that network analysis provides different insights from traditional spatial analyses, as typical spatial statistics would rely on quantity of isolates if applied in this context.
- We identified cases of human-only genotypes with highly localized transmission, further confirming previous findings and emphasizing the importance of a OneHealth approach to studying brucellosis.
- Future molecular surveillance efforts should include larger sample sizes. Ultimately, approaches that integrate spatial and molecular epidemiological data are needed to improve future brucellosis control efforts.

References

1. Shevtsova, E., Shevtsov, A., Mukanov, K., Filipenko, M., Kamalova, D., Sytnik, I., Syzdykov, M., Kuznetsov, A., Akhmetova, A., Zharova, M., Karibaev, T., Tarlykov, P., & Ramanculov, E. (2016). Epidemiology of Brucellosis and Genetic Diversity of Brucella abortus in Kazakhstan. PLOS ONE, 11(12), e0167496. https://doi.org/10.1371/journal.pone.0167496
2. Charypkhan, D., Sultanov, A. A., Ivanov, Nikolay. P., Baramova, S. A., Taitubayev, Mereke. K., & Torgerson, P. R. (2019). Economic and health burden of brucellosis in Kazakhstan. Zoonoses and Public Health, 66(5), 487–494. https://doi.org/10.1111/zph.12582
3. Beauvais, W., Coker, R., Nurtazina, G., & Guitian, J. (2017). Policies and Livestock Systems Driving Brucellosis Re-emergence in Kazakhstan. EcoHealth, 14(2), 399–407. https://doi.org/10.1007/s10393-015-1030-7
4. Volz, E. M., Koelle, K., & Bedford, T. (2013). Viral Phylodynamics. PLoS Computational Biology, 9(3), e1002947. https://doi.org/10.1371/journal.pcbi.1002947
5. Waugh, S. G., Sytnik, I., Talgat Karibayev, Aikim Alimbayev, Mukhit Ornybayev, Nurgisa Rametov, Mikeljon Nikolich, Sue Hagius, Philip Elzer, & Jason K11. Blackburn. (In review (2)). Genetic Evidence of Highly Localized Brucella Transmission in Southern Kazakhstan. Emerging Infectious Diseases.
6. Huynh, L. Y., Ert, M. N. V., Hadfield, T., Probert, W. S., Bellaire, B. H., Dobson, M., Burgess, R. J., Weyant, R. S., Popovic, T., Zanecki, S., Wagner, D. M., & Keim, P. (2008). Multiple Locus Variable Number Tandem Repeat (VNTR) Analysis (MLVA) of Brucella spp. Identifies Species-Specific Markers and Insights into Phylogenetic Relationships. In National Institute of Allergy and Infectious Diseases, NIH (pp. 47–54). Humana Press. https://doi.org/10.1007/978-1-59745-569-5_6
7. Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics, 5(3), 299–314. https://doi.org/10.1080/10618600.1996.10474713
8. Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. Bioinformatics, 27(4), 592–593. https://doi.org/10.1093/bioinformatics/btq706
9. Prosperi, M. C. F., Ciccozzi, M., Fanti, I., Saladini, F., Pecorari, M., Borghi, V., Di Giambenedetto, S., Bruzzone, B., Capetti, A., Vivarelli, A., Rusconi, S., Re, M. C., Gismondo, M. R., Sighinolfi, L., Gray, R. R., Salemi, M., Zazzi, M., & De Luca, A. (2011). A novel methodology for large-scale phylogeny partition. Nature Communications, 2, 321. https://doi.org/10.1038/ncomms1325
10. Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.
11. Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems, 1695.
12. Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal, 10(1), 439. https://doi.org/10.32614/RJ-2018-009
13. Rodionov, A. (2022). geokz: Offers Various Kazakhstani Maps as Data Frames and “sf” Objects.