Analysis of SNPs in *Mycoplasma cynos* strains from canine hosts within the Pacific Coast animal shelter network

by
Anastasiya Prymolenna

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Biochemistry and Molecular Biology
(Honors Scholar)

Presented May 24, 2019
Commencement June 2019

AN ABSTRACT OF THE THESIS OF

Anastasiya Prymolenna for the degree of Honors Baccalaureate of Science in Biochemistry and Molecular Biology presented on May 24, 2019. Title: Analysis of SNPs in *Mycoplasma cynos* strains from canine hosts within the Pacific Coast animal shelter network.

Abstract approved:_____

Benjamin Dalziel

Single nucleotide polymorphisms (SNPs) can be used as markers to infer population genetic structure. For *Mycoplasma cynos*, a pathogen associated with infectious respiratory disease among canines and other animals, SNPs may help to elucidate ecological and evolutionary relationships among the strains that reside in the upper respiratory tract of canines and that circulate within a network of animal shelters. When hosts are transferred within a network and expose their microbiomes to different environments, the pathogens contained within the microbiome can accumulate single nucleotide mutations from environmental effects. The present study looks at SNPs at positions across the entire *M. cynos* genome using data from high throughput short-read sequences recovered from nasal swabs. The swabs came from five canines sampled from an animal shelter that regularly receives transfers from other locations. Results from the analysis of the SNPs indicates evidence of genetic structure among *M. cynos* populations resident within each sampled canine. This heterogeneity may be indicative of a variety of processes across different scales which can shape microbial biogeography in networks of hosts.

Key Words: SNP, genetic structure, Mycoplasma

Corresponding e-mail address: prymolea@oregonstate.edu

Analysis of SNPs in *Mycoplasma cynos* strains from canine hosts within the Pacific Coast animal shelter network

by
Anastasiya Prymolenna

A THESIS

submitted to

Oregon State University

Honors College

in partial fulfillment of
the requirements for the
degree of

Honors Baccalaureate of Science in Biochemistry and Molecular Biology
(Honors Scholar)

Presented May 24, 2019
Commencement June 2019

Honors Baccalaureate of Science in Biochemistry and Molecular Biology project of Anastasiya Prymolenna presented on May 24, 2019.

APPROVED:

_____

Benjamin Dalziel, Mentor, representing the Department of Integrative Biology, Oregon State University

_____

Felipe Barreto, Committee Member, representing the Department of Integrative Biology, Oregon State University

_____

David Hendrix, Committee Member, representing Department of Biochemistry and Biophysics, Oregon State University

_____

Toni Doolen, Dean, Oregon State University Honors College

I understand that my project will become part of the permanent collection of Oregon State University, Honors College. My signature below authorizes release of my project to any reader upon request.

_____

Anastasiya Prymolenna, Author

Acknowledgements

It is a pleasure to thank Dr. Benjamin Dalziel for his aid and guidance in the research and developmental process of this work, with special thanks and consideration to Elizabeth Knorr and Noah Silva de Leonardi. Also thank you to Dr. Felipe Barreto and Dr. David Hendrix for their intellectual insight, advice, and criticism.

Table of Contents:

INTRODUCTION

The spread and evolution of pathogens within host populations may depend on the structure of infectious transmission networks[1]. To elaborate, when pathogens are contained within a closed system, the way the pathogens move between locations and populations can promote environmental factors within the new location that influence pathogens at a genetic level. This can be quantified and modeled in numerous ways depending on the factors associated with a given system[2]. Evolutionary dynamics are always acting on any given population. In smaller populations, random speciation events and mutations have a greater impact on the allelic penetration in a population due to the fact that any new allele that presents itself in the population with make up a greater proportion of the whole gene pool that it would in a large population. When small subsets of a larger population become isolated, via the founder effect they become prone to influences of their novel environment as a selective pressure.

According to past studies, the contact structure of a network can suppress the selection of mutations within a population due to lowered fixation probabilities, a measure of how well a novel strain permeates the network and drives other strains to extinction[3]. For instance, in the case of an animal shelter network, heterogeneity in the dog population (i.e. spatial-temporal structure in contact among individuals) is maintained by the constant transfer of animals into and out of select points in the network.

The network transfers are based off of the principle that dogs in areas that have a high density of unowned or wild canines would be captured and distributed into areas where the unhoused canine population is less abundant. This is done to give those dogs a better chance of adoption. The samples in this study were pulled from one such network---animal shelters on the West Coast of the US, including OR, WA, and CA. Figure 1A depicts the West Coast network and

outlines the general size and relative location of shelters contained in the network. Additionally, Figure 1B lists all shelters contained within the network in order of their geographic distance from the Oregon Humane Society (OHS). Outlined in Figure 1B is the model of in-transfers and out-transfers of animals between shelters showing that OHS imports from larger California shelters and exports out to smaller shelter located within Oregon and Washington.

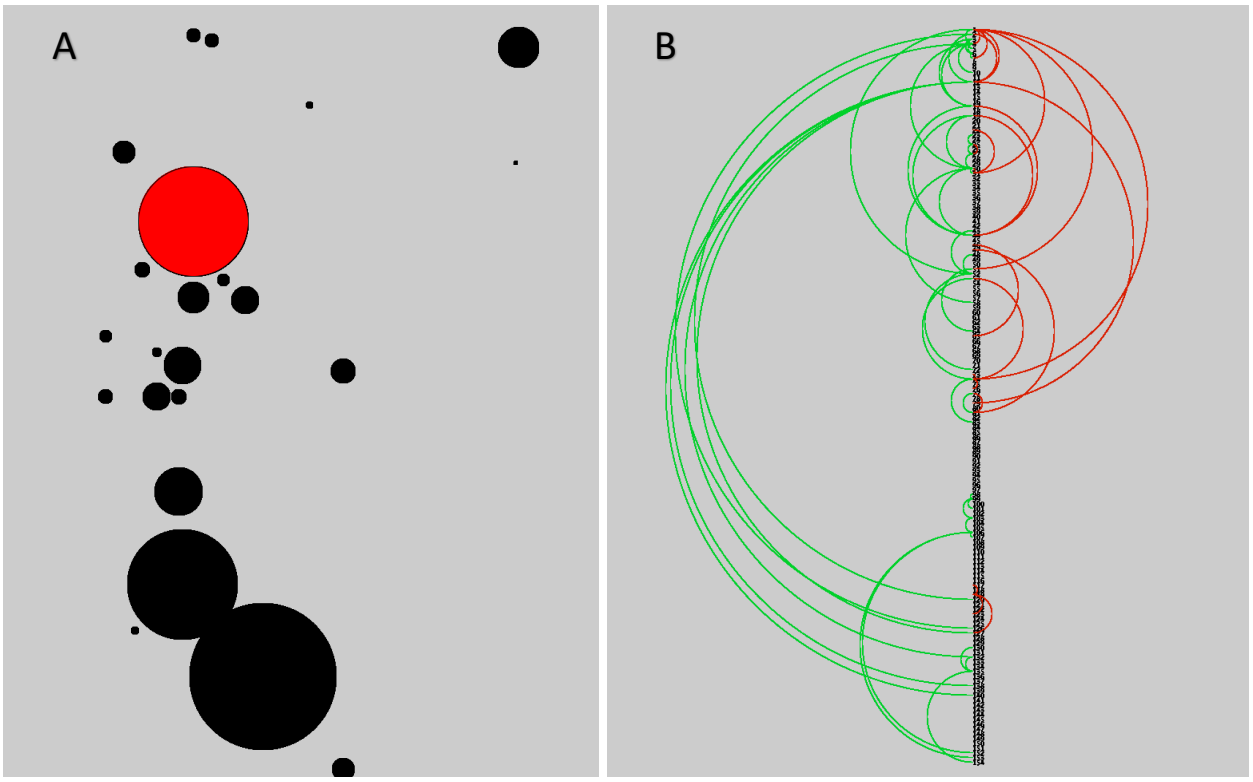

Figure 1: A) The size and relative location of the shelters comprising the West coast animal shelter network participating in canine transfers. Shown in red is Oregon Humane Society (OHS). The larger shelters toward the bottom represent Los Angeles and San Francisco hubs and smaller nodes represent the more rural shelters. **B) The shelters in the network are ordered by their relative distance from OHS, where OHS is the first in the list at the top of the figure.** The shelters located towards the top comprised Oregon shelters, while the ones towards the bottom include California. The pattern of transfers into a shelter (green) and out of a shelter (red) generally in a bottom to top direction.

A closed system, such as a regional animal shelter conglomerate, serves as a model to determine whether the pathogens that inhabit hosts also follow evolutionary patterns consistent

with the movement of the hosts within the system. Canine populations in animal shelters are at an increased risk of developing diseases from crowded conditions due to contact heterogeneity[4] and proximity of varied pathogen species[5]. One such bacterial pathogen is *Mycoplasma cynos. Mycoplasma,* a small free-living bacterium in the class Mollicutes, are common to the bacterial flora in the upper respiratory tract of dogs[6] and are increasingly being recognized for their association in canine upper respiratory infections[7,8].

As the animal shelter system regularly transfers animals from high extermination and densely populated shelters to ones where the animals are given a better chance to live and be adopted, the canine hosts carry and spread pathogens. In turn, *Mycoplasma* isolates transfer into new environments and may evolve by either acquiring single nucleotide mutations that alter fitness or through genetic drift. This evolution can be tracked on a genetic level in the form of single nucleotide polymorphism (SNP) presence to allow us to gain insight into the phylogenetic connections of *Mycoplasma* across different hosts and shelters. Combined with biological network modeling[9], an identified network can be analyzed with tools that will yield a greater understanding of the underlying biological processes[9]. In the case of this study, we used an *ad hoc* survey to gather information on shelter size and connectivity across the west coast of the US, and used bioinformatic approaches on *Mycoplasma* contained in five canine hosts in order to gather data regarding the genetic structure in resident *Mycoplasma* among dogs in the shelter network.

Canine *Mycoplasmas* are phylogenetically varied across their many various strains.[10] Their diversity is further reflected in the range of GC content of each species, which can vary from 23.8% to 36%.[11] The effect of environmental influence on the genome ofs a strain of mycoplasma can be measured in the prevalence and presence of SNPs. SNPs can be defined as nucleotide positions within an organism's genome that contains two or more varied bases, where each base has a 1%

or greater prevalence within a population[13]. Generally, SNPs are one of the most common markers of genetic variation in organisms.

The null hypothesis in this study is that the *Mycoplasma* bacteria contained within canine hosts will not show significant differences in SNPs across the samples. If this is the case, we will see no differences between the penetration of SNPs between distinct samples in individual nucleotide positions. Alternatively, polymorphisms in respective nucleotide positions could be unique to a single sample due to differing environmental influences acting on the *Mycoplasma*.

Microorganisms colonize animal bodies as a compound and highly diverse community. In many aspects, the greater roles of most organisms in the upper respiratory tract are unknown[14]. So far there have been no studies that have looked into microbial diversity in a survey of shelter dogs wherein the canines are randomly selected rather than only identified for study if the dog expressed symptoms of infection[7,8,15]. Therefore, the aim of this study was to characterize diversity within the microbiome of hosts within a closed transfer network system. We accomplished this by looking at SNPs of *Mycoplasma* strains acting as the respective model organism contained within the upper respiratory tract of several canine hosts that had migrated from one population to another within an animal shelter system.

The significance of this study lies in the fact that although we have access to sequencing technology and the genomes of many microbial species, we still lack accurate and comprehensive databases for the genetic diversity of many microorganism, especially *Mycoplasma*[16]. Insights into the significant systematic variation across a microbial population and its relatedness to the emergent genetic structure provides a novel look at how seemingly unrelated communities compare to one another across the scale of a whole genome. Studies such as this allow increased

access to a more thorough and in-depth accumulation of genetic resource databases that can be made available for all subsequent inquiries into microbial genomics.

METHODS and MATERIALS

Ethics Statement

The presented study did not involve interacting with animals with research intent. Nasal swab sampling was done as part of routine screening and within scope of veterinary practice at Oregon Human Society.

Network Structure Identification

Preliminary raw intake and outtake data was gathered from select shelters across the west coast. Shelters voluntarily submitted their recent transfers to the study and indicated how often and how many canines were brought in or taken out of their facility. The submitted data also stated the locations transfers came from and were sent to. Figure 1B was constructed based on this information.

Study Population

Table 1: The details of the host background and sampling information. Columns are ordered from left to right, respectively. 1) The index of the sample out of 104 total samples of 59 total dogs. 2) The index of the host sample applied to high-through put sequencing. 3) Traceable canine ID within the shelter database system. 4) Last known host residential history before coming to OHS. 5) Confirmation or lack of visible symptoms of infection at time of sampling. 6) Type of swab sample taken from the host. 7) How many times this animal has been sampled including the present sample 8) Confirmation or lack of signs of infection in the process of the hosts residence at OHS. 9) Comparison from evidence of infection at the time of swab to later after the sampling was completed

Sample Index	Shotgun Sample Index	Pet ID	Origin	Symptomatic	Sample Type	Sampling Sequence	Ever Symptomatic	Became Symptomatic
9	1	195349	Madera	TRUE	nasal	5	TRUE	TRUE
28	2	195352	Madera	FALSE	nasal	4	TRUE	TRUE
36	3	195637	Oakland	FALSE	nasal	1	TRUE	TRUE
43	4	195634	Oakland	TRUE	nasal	1	TRUE	FALSE
54	5	195635	Oakland	TRUE	nasal	1	TRUE	FALSE

Sample Collection

Nasal swabs were gathered from canines at the Oregon Humane Society in Portland, OR. Dry, sterile cotton swabs were used to collect a sample. Afterwards the samples were stored at -80°C in a phosphate-buffered saline solution.

DNA extraction and sequencing

To extract DNA from frozen swab samples protocol for Qiagen Allprep PowerViral DNA/RNA Kit was followed. The lysis step was from the Quick-Start Protocol for the kit. All remaining steps following the protocol were completed on the Qiacube and are adapted from the RNEasy PowerMicrobiome Protocol (Stool or Biosolid IRT). DNA was quantified using high sensitivity buffer and dye on Qubit. We obtained 2x300bp reads via Illumina MiSeq high throughput DNA sequencing, using a shotgun approach, implemented at the OSU Center for Genome Research and Biocomputing (CGRB) core facilities.

Sequence Analysis

Each of the five sets of paired end reads were profiled to the mpa_v20_m200 bowtie2 database using MetaPhlAn2 v 2.6.0, with the --bt2_ps sensitive flag; other settings were left as defaults[17,18].

Reads were aligned to the *Mycoplasma cynos* C142 (Assembly Accession GCF_000328725.1) genome using bwa mem (v. 0.7.12) with the -M flag enabled for Picard Tools compatibility[19]. GATK (v. 4.0.1.1) was used to convert SAM files to BAM files with the SamFormatConverter tool and the BAM files were coordinate sorted with the SortSam tool[20]. PCR duplicate reads were ignored in subsequent GATK analysis and marked with GATK's

MarkDuplicates tool. Each variant was called with HaplotypeCaller per sample, in GVCF mode, sample ploidy was set to one. Samples were combined into a single g.vcf file with CombineGVCFs and joint calling was performed with CombineGVCFs which resulted in a single vcf file. Variants were split, with SelectVariants. SNPs were filtered with parameters QD < 5.0 || MQ < 40.0 || FS > 60.0 || ReadPosRankSum < -8.0 || MQRankSum < -12.5 and INDELs were filtered by QD < 5.0 || FS > 200.0 || ReadPosRankSum < -20.0.

RESULTS

Analysis of the sequenced samples yielded approximately 99% canine DNA along with 1% *Mycoplasma* DNA. SNPs occurred across the Mycoplasma genome as shown in Fig. 2 and Fig. 3. A total of 1,754 SNPs were discovered across the *Mycoplasma* sequences from five samples. The hypothesis predicts the occurrence of systematic differences in the allelic depth of alternate nucleotides across different positions and samples (Fig. 3). Chi-Squared statistical significance tests showed a p-value less than 0.05 found in thirteen positions within the genome that contained SNPs at a read depth to be deemed significant based on calculations (Fig. 4).

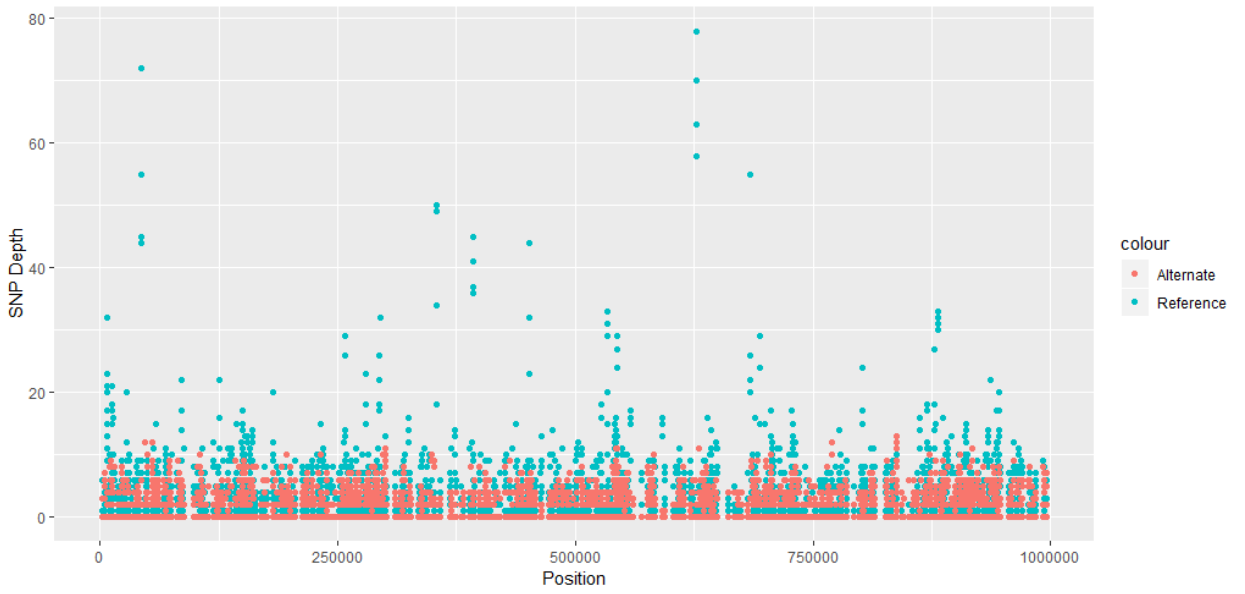

Figure 2: Dot plot of all positions containing possible significant SNPs. The frequency distribution for how many reads matched either the reference sequence or the alternative sequence was graphed against the position at which the SNP occurred. For all 1,754 SNP positions five samples of genomic data were compared against a reference sequence.

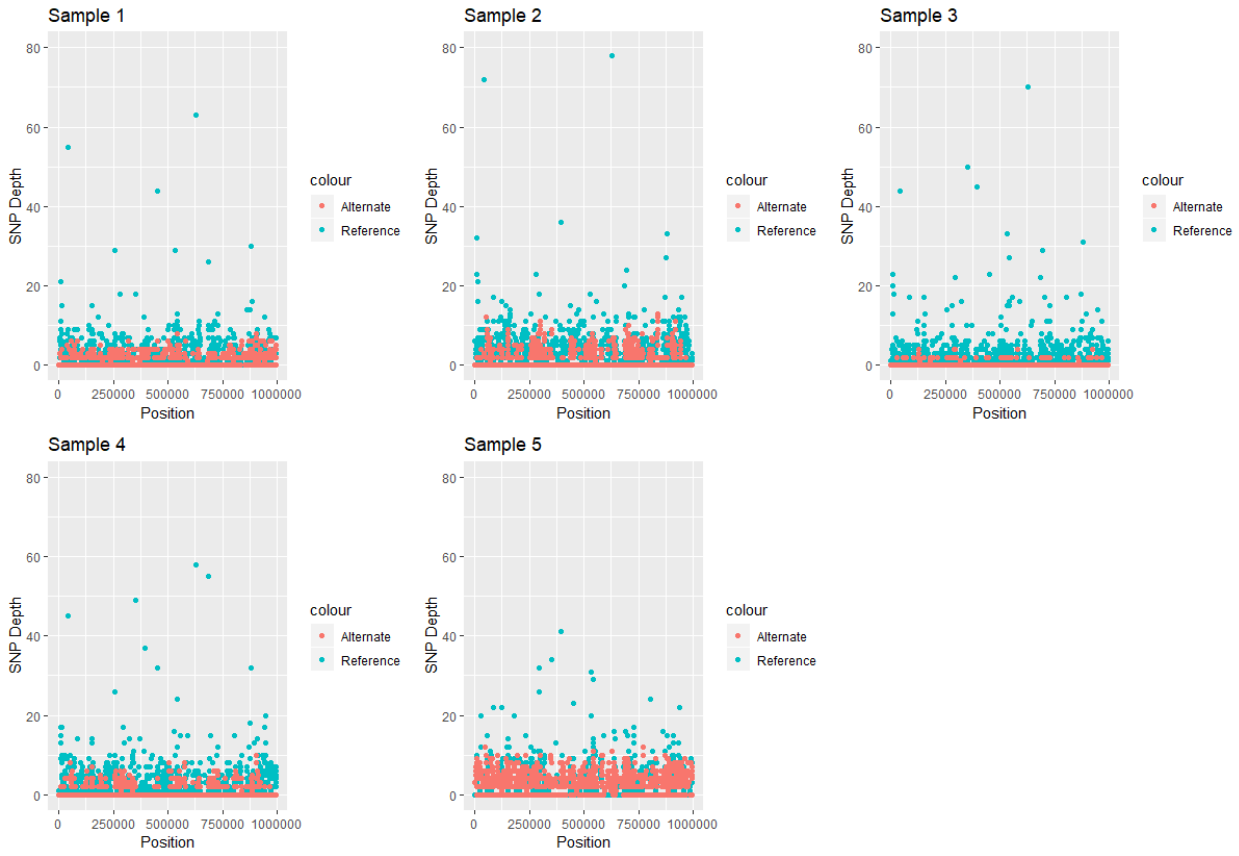

Figure 3: Dot plot of all positions containing possible SNPs in separate samples. In each of the five samples the frequency of a SNP read was graphed against the position at which that SNP occurred on for all significant positions across the *M. cynos* genome.

To find statistical significance, Figure 4 was constructed to differentiate between the probability that each SNP is equally likely to be present in any of the five samples – i.e. the probability that the null hypothesis of no genetic structure is correct at that locus – or if there are enough differences among individual samples to indicate genetic structure. In other words, the higher up in the graph a point is found the more likely there is genetic structure among the five samples at that locus, if the read depth is greater than five reads. This was assessed with a χ^2 test

at each locus. The p-values reported are unadjusted for multiple comparison. The evidence to support the alternate hypothesis is outlined in the red points in Fig 4 when (i) the p-value was less than 1/10000 based on the number of SNPs present and (ii) the total read depth was greater than 5. Under these criteria, 13 SNPs were found to be significant at positions shown in Figure 4.

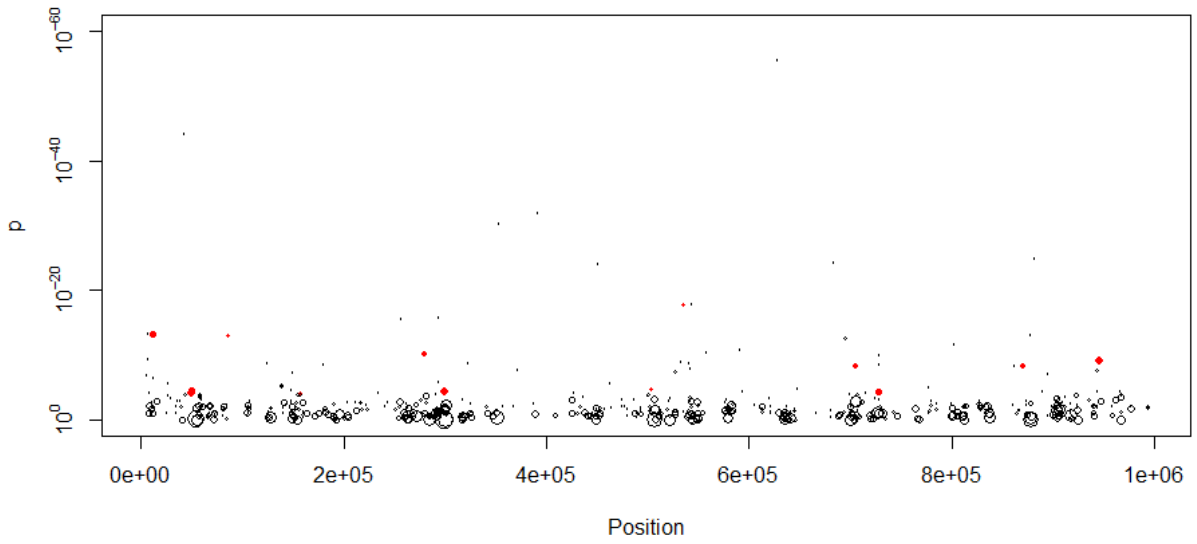

Figure 4: The p-value is plotted against each position at which a polymorphism occurred. The size of the plotted point corresponds to the frequency at which the SNP was observed and the red colored points indicate statistically significant values. The location of a point on the graph corresponds to a p-value in the y-direction and a position in the genome in the x-direction.

The expectation based on the null hypothesis was that there would be no difference in the SNP penetration at all positions in the sample at which an SNP was observed. However, the results indicate that this was not the case. Based on the data found in the study we reject the null hypothesis. The significance of the graphed SNPs is consistent with the presence of genetic structure among the samples from canines within the shelter network.

DISCUSSION

Each sample plotted previously in Figure 3 represents one population group, e.g. a relatively isolated island of genetic variation. The results shown previously in Figure 4 are plotted such that the graph can be used as a probabilistic model. In Figure 4 red points signify each significant SNP unique to a sample of mycoplasma. Points with low p-values, found in the upper regions of Fig. 4, indicated the lowest likelihood of observing that unique SNP in a sample other than in the one it was discovered. Regarding the null hypothesis of this experiment which predicted that there would not be any significant differences in SNP read depth across samples, the data shown in Figure 4 provides sufficient evidence necessary to reject the null hypothesis.

Of note is the relationship of the SNPs emergence as it may be related to the shelter network. Based on previous mathematical models and transfer simulations[3], the network could play a role in guiding the differentiation of SNPs at positions within the bacterial genome. The basis of some predictions includes differences in samples from various different parts of the network, including the hosts that reside and migrate to network regions that have dissimilar environmental conditions. The network effects on the outcome of mutations are even more greatly imposed by the transfer of canines from localized to dispersed shelters as shown in the preliminary data. At the same time, the polymorphisms in the bacterial genome could also be attributed to differences in host breed physiology. *Mycoplasma* communities in dog noses experience different conditions from host to host and these may lead to different selective pressures in different host environments.

Future directions and applications of the data presented in this study should look further into how the SNPs impact the functional genome and whether the mutations are occurring in functional or non-coding parts of the genome. The rate of mutation should then be adjusted and categorized depending on the significance of impact a mutation will have on the organism as a whole. Future applications can take the SNPs and their rate of occurrence to create functional annotations of not only *Mycoplasma*, but any organism that acts as an indicator of pathogenic infection. The use of annotated bacterial genomes will increase our ability to understand how variations at nucleotide level of genetic structure influence the resulting functional expression of the corresponding genes. Future work could also attempt to directly estimate rates of mutation within *Mycoplasma* since historically mutagenic inferencing has been done through the assumption of site-associated mutation rates[21,22].

We revealed noticeable patterns of genetic structure in *Mycoplasma* present in upper respiratory tract of canines. However, observed differences in the genetic structure could be imposed by a number of interacting processes. Selection could be a driving factor that generated differences in canine breed physiology or selection due to environmental differences at multiple scales. Differences or drift at multiple scales, within hosts, shelters, and regions could be affecting our results.

Although the basic evolutionary mechanics, such as accumulation of SNPs and their retention within populations of *Mycoplasma* are still not well understood in reference to structured host populations such as the shelter network, studies such as this allow us the opportunity to better look at organisms that define and correlate with canine upper respiratory infections[23]. The findings of distinct positions and unique polymorphisms that change the make-up of a bacterial genome in

a way that can be predicted and quantified empowers us to further consider, and perhaps one day predict, the evolutionary path of microorganisms critical for animal population health.

REFERENCES

(1) Lieberman, E.; Hauert, C.; Nowak, M. A. Evolutionary Dynamics on Graphs. **2005**, *433*, 5.
(2) Watts, D.; Strogatz, S. Collective Dynamics of Small-World Networks. **1998**, *393*, 3.
(3) Leventhal, G. E.; Hill, A. L.; Nowak, M. A.; Bonhoeffer, S. Evolution and Emergence of Infectious Diseases in Theoretical and Real-World Networks. *Nat. Commun.* **2015**, *6*. https://doi.org/10.1038/ncomms7101.
(4) Dalziel, B. D.; Huang, K.; Geoghegan, J. L.; Arinaminpathy, N.; Dubovi, E. J.; Grenfell, B. T.; Ellner, S. P.; Holmes, E. C.; Parrish, C. R. Contact Heterogeneity, Rather Than Transmission Efficiency, Limits the Emergence and Spread of Canine Influenza Virus. *PLoS Pathog.* **2014**, *10* (10), 1–13. https://doi.org/10.1371/journal.ppat.1004455.
(5) Estrada-Peña, A.; Ostfeld, R. S.; Peterson, A. T.; Poulin, R.; de la Fuente, J. Effects of Environmental Change on Zoonotic Disease Risk: An Ecological Primer. *Trends Parasitol.* **2014**, *30* (4), 205–214. https://doi.org/10.1016/j.pt.2014.02.003.
(6) Rosendal, S. Canine Mycoplasmas: Their Ecologic Niche and Role in Disease. *J. Am. Vet. Med. Assoc.* **1982**, *180* (10), 1212–1214.
(7) Mannering, S. A.; McAuliffe, L.; Lawes, J. R.; Erles, K.; Brownlie, J. Strain Typing of Mycoplasma Cynos Isolates from Dogs with Respiratory Disease. *Vet. Microbiol.* **2009**, *135* (3–4), 292–296. https://doi.org/10.1016/j.vetmic.2008.09.058.
(8) Decaro, N.; Mari, V.; Larocca, V.; Losurdo, M.; Lanave, G.; Lucente, M. S.; Corrente, M.; Catella, C.; Bo, S.; Elia, G.; et al. Molecular Surveillance of Traditional and Emerging Pathogens Associated with Canine Infectious Respiratory Disease. *Vet. Microbiol.* **2016**, *192*, 21–25. https://doi.org/10.1016/j.vetmic.2016.06.009.
(9) Xia, T. Network Modeling in Systems Biology. 85.
(10) Chalker, V. J.; Brownlie, J. Taxonomy of the Canine Mollicutes by 16S RRNA Gene and 16S/23S RRNA Intergenic Spacer Region Sequence Comparison. *Int. J. Syst. Evol. Microbiol.* **2004**, *54* (Pt 2), 537–542. https://doi.org/10.1099/ijs.0.02869-0.
(11) Chalker, V. J. Canine Mycoplasmas. *Res. Vet. Sci.* **2005**, *79* (1), 1–8. https://doi.org/10.1016/j.rvsc.2004.10.002.
(12) Nowak, M. A.; Sasaki, A.; Taylor, C.; Fudenberg, D. Emergence of Cooperation and Evolutionary Stability in Finite Populations. *Nature* **2004**, *428* (6983), 646–650. https://doi.org/10.1038/nature02414.
(13) Hartman, J. L. I.; Garvik, B.; Hartwell, L. Principles for the Buffering of Genetic Variation. *Science* **2001**, *291* (5506), 1001.
(14) Tress, B.; Dorn, E. S.; Suchodolski, J. S.; Nisar, T.; Ravindran, P.; Weber, K.; Hartmann, K.; Schulz, B. S. Bacterial Microbiome of the Nose of Healthy Dogs and Dogs with Nasal Disease. *PLOS ONE* **2017**, *12* (5), e0176736. https://doi.org/10.1371/journal.pone.0176736.

(15) Hong, S.; Kim, O. Molecular Identification of Mycoplasma Cynos from Laboratory Beagle Dogs with Respiratory Disease. *Lab. Anim. Res.* **2012**, *28* (1), 61–66. https://doi.org/10.5625/lar.2012.28.1.61.
(16) Stucki, D.; Gagneux, S. Single Nucleotide Polymorphisms in Mycobacterium Tuberculosis and the Need for a Curated Database. *Tuberc. Edinb. Scotl.* **2013**, *93* (1), 30–39. https://doi.org/10.1016/j.tube.2012.11.002.

(17) Truong, D. T.; Franzosa, E. A.; Tickle, T. L.; Scholz, M.; Weingart, G.; Pasolli, E.; Tett, A.; Huttenhower, C.; Segata, N. Erratum: MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling. *Nat. Methods N. Y.* **2016**, *13* (1), 101. http://dx.doi.org.ezproxy.proxy.library.oregonstate.edu/10.1038/nmeth0116-101b.
(18) Langmead, B.; Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **2012**, *9* (4), 357–359. https://doi.org/10.1038/nmeth.1923.
(19) Li, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. *ArXiv13033997 Q-Bio* **2013**.
(20) McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. *Genome Res.* **2010**, *20* (9), 1297–1303. https://doi.org/10.1101/gr.107524.110.
(21) Sung, W.; Ackerman, M. S.; Miller, S. F.; Doak, T. G.; Lynch, M. Drift-Barrier Hypothesis and Mutation-Rate Evolution. *Proc. Natl. Acad. Sci.* **2012**, *109* (45), 18488–18492. https://doi.org/10.1073/pnas.1216223109.
(22) Lynch, M. Evolution of the Mutation Rate. *Trends Genet.* **2010**, *26* (8), 345–352. https://doi.org/10.1016/j.tig.2010.05.003.
(23) Mitchell, J. A.; Cardwell, J. M.; Leach, H.; Walker, C. A.; Le Poder, S.; Decaro, N.; Rusvai, M.; Egberink, H.; Rottier, P.; Fernandez, M.; et al. European Surveillance of Emerging Pathogens Associated with Canine Infectious Respiratory Disease. *Vet. Microbiol.* **2017**, *212*, 31–38. https://doi.org/10.1016/j.vetmic.2017.10.019.
(24) Crozier, R. H. Preserving the Information Content of Species: Genetic Diversity, Phylogeny, and Conservation Worth. *Annu. Rev. Ecol. Syst.* **1997**, *28* (1), 243–268. https://doi.org/10.1146/annurev.ecolsys.28.1.243.