West Nile Virus Genetic Diversity in California - DRAFT

← FYI: you may be able to figure out how to add a watermark with "Draft" throughout

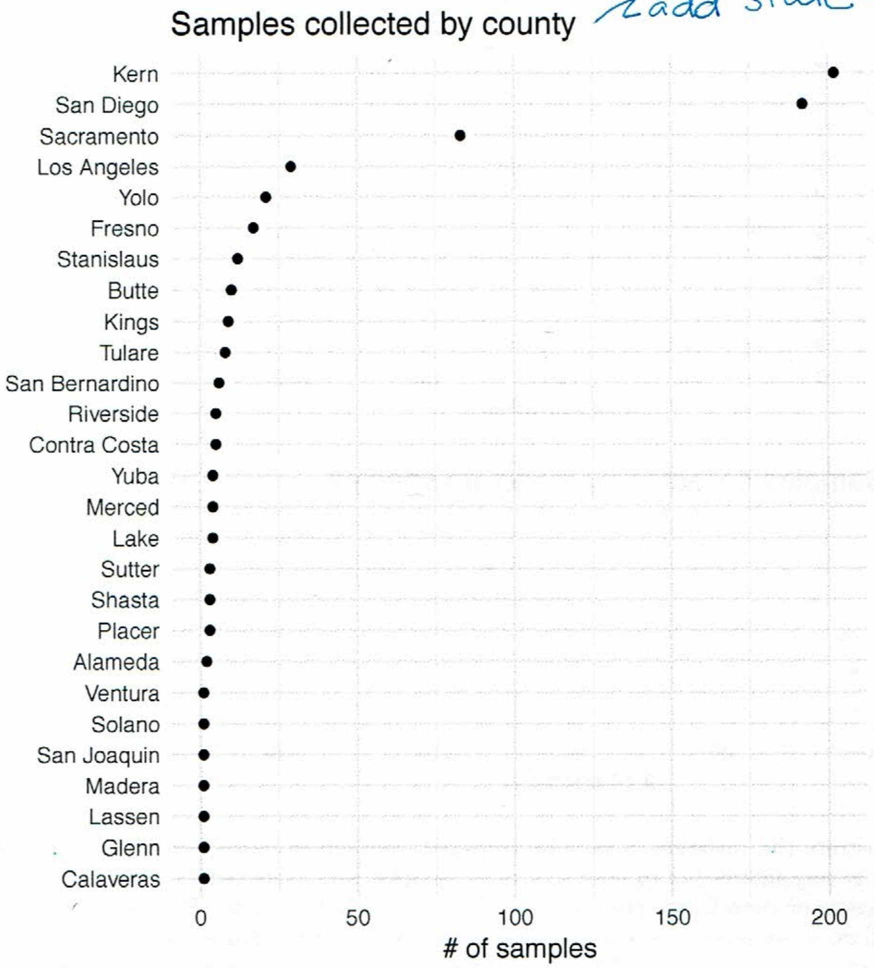
Aeriel Belk, Amanda Walz, Rebecca Cheek, & Kyle Root

December 5, 2018

Sample Collection

good

Samples were collected from both bird tissues (**n = 222**) and pooled mosquito populations (**n = 419**) in multiple locations with varying degrees of urbanization. Samples appear to have been collected by the researchers or submitted by other parties. Metadata reflecting the amount of samples collected from each county and each species, either avians or pooled mosquitoes, was examined to determine the potential for bias.

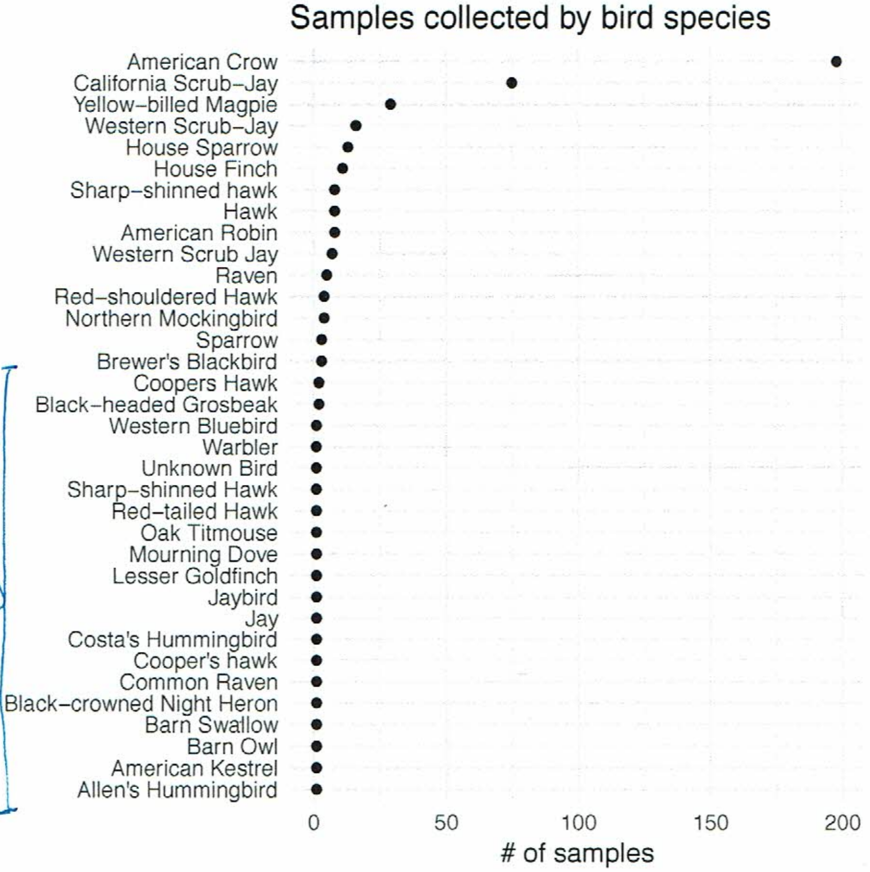


← add state → always in CA?

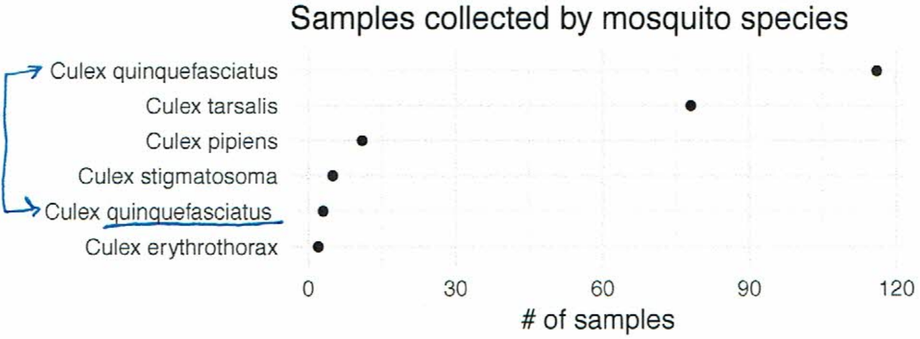
Nice plot!

Any information in how widely the date of collection varied across samples?

The majority of samples were collected from two counties, Kern and San Diego. What effect this may have on the results is unknown. However, if a particular strain of virus were more prevalent in these areas then the data would not be representative of West Nile Virus variation across all of California. Furthermore, other variables related to geographical location may be impacting the data.



Since some of these are just one or a few you might want to use 'fct_lump' to group those together in an "other" category



What's the difference between these two? Should they be the same?

The above graphs demonstrate the variation in samples collected from either bird species or the pooled mosquito populations. It is very interesting to note that most samples were collected from the American Crow. Some literature suggests an overall high prevalence of West Nile Virus in corvids. The American Crow is considered the third highest avian species for West Nile Virus prevalence behind Yellow-billed Magpies and Western Scrub-jays. However, this data suggests American Crows have a higher prevalence of the virus. This could be resulting from a disproportionate amount of American Crow samples submitted.

[illegible]

Genomics

GitHub

You can use BibTex as a reference manager for Rmarkdown documents

The genomic data from Dr. Andersen's git hub page included 649 consensus sequences of West Nile Viral DNA that has been isolated from multiple species of birds and mosquitoes through amplicon-based sequencing, which sequences highly specific regions of the target genome (see Wen et al. 2017 for a review). However, the nature of High Throughput Sequencing (HTS) technologies often results in significant variation between individual sequences due to poor quality DNA or sequencing error. So pre-processing of genomic data for quality control, and formatting is necessary before analysis may be conducted. To start, consensus sequences were aligned using an R package, DECIPHER, which scanned the sequences for k-mers (regions of genetic matching) that were 11 nucleotides long and aligned the reads to each other. ShortRead was then used to trim the poor-quality ends of the sequences from bp 31 to bp 10963. Many of the resulting sequences contained an obvious excess of *Ns* which are substitutes marking for either missing data or "any nucleotide". To rectify this variation of read quality, sequences were then filtered using the python script "Sequence_cleaner.py" available on Biopython. Once, to filter out any sequences with greater than 10% *N* content or shorter than 10317 bp long (568 sequences matched this criteria). And a second, more conservative, filter criteria that removed sequences with greater than 0% *Ns* and shorter than 10317 bp in length (157 sequences matched this criteria).

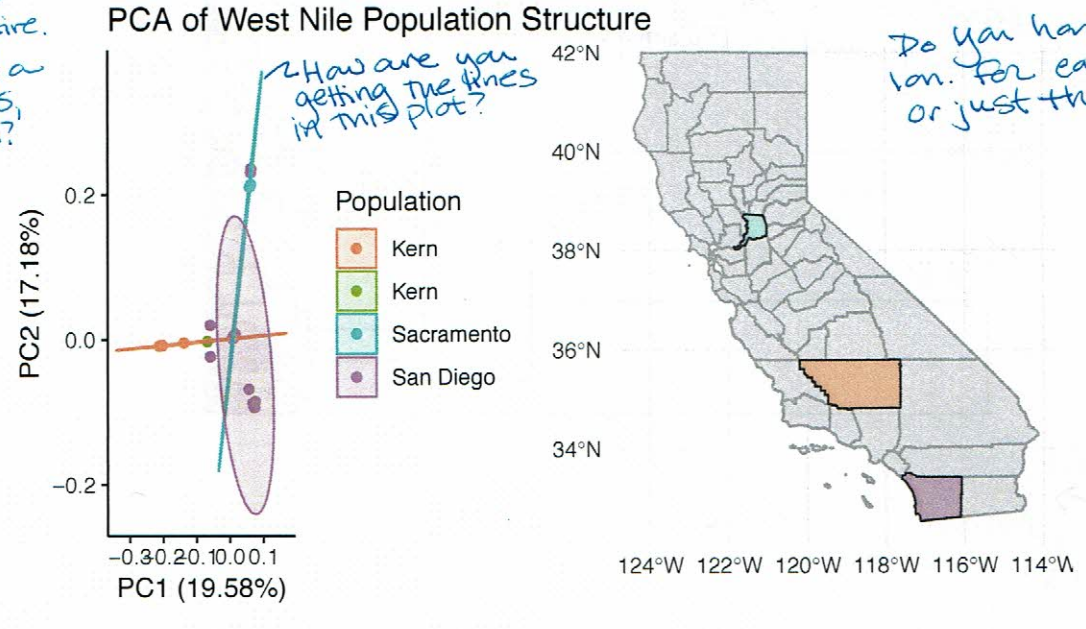
watch grammar here

great use of details like this!

Of the remaining 157 sequences that matched out filtering criteria, 133 were from samples collected in Sacramento, Kern, and San Diego counties. These counties conveniently reside along a large latitudinal gradient, which allows us to ask: "How closely related are these geographically distinct populations?" Considering that West Nile Virus is a vector borne disease that infects highly mobile taxa such as birds and mosquitoes, it would be interesting to determine if they may be described as one panmictic group or three *genetically distinct* populations.

To answer this question, a PCA analysis was performed on the Sacramento, Kern, and San Diego samples by first converting our filtered fasta files to a geid object using the adegenet and dartR packages, so that only the veritable regions of the sequences were conserved. This polymorphic data set was then run through a PCA and plotted using ggplot and ggfortify.

this sounds way too passive. You performed a PCA analysis, instead, maybe?



How are you getting the lines in this plot?

Do you have the lat + lon. for each sample? or just the county?

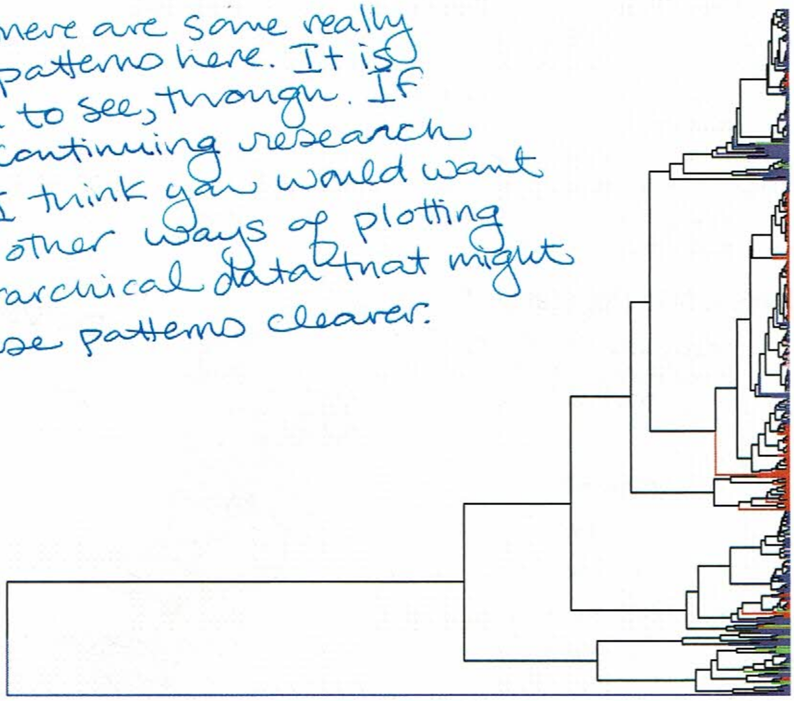
Though there is some amount of overlap between the three populations, there is a clear amount of structure between them. Additionally, there appears to be a single individual from Kern county that does not group with the others. While this could be due to a single mutant strain, it is also very likely that this individual sample was sequenced from a vagrant host, or the sample itself was mislabeled.

Comparative Dendrogram

Consensus sequences were obtained from the Andersen lab github. These sequences, collected from a variety of sources including several bird and mosquito species, were aligned against each other so we could determine point mutation differences in the sequences that could be attributed to distinct strains of West Nile Virus. The hypothesize investigated here is that sequences isolated from different organisms may be phylogenetically different. This is likely based on general knowledge of virus evolution. In general, evolutionary changes occur in response to pressures from the host immune system or other environmental effects. Thus, it is likely that strains that begin to differentiate in a bird may be different than those that begin to differentiate in mosquitoes. To investigate this, the individual species of birds were categorized into four larger groups: corvids, hawks, songbirds, and mosquitoes. Then, the aligned strains were hierarchically clustered based on Hamming distance using the R package biostrings. These clusters were visualized as a colored dendrogram to evaluate whether the sequences from individual species groups were more similar to each other than others.

interesting!

It seems there are some really interesting patterns here. It is a bit hard to see, though. If you were continuing research on this, I think you would want to explore other ways of plotting this hierarchical data that might make these patterns clearer.



floats

Colors:

- Blue = corvid
- Green = hawk

Ideally, this information would be included in a legend for the plot above

- Red = mosquito
- Pink = songbird

Based on this visualization, it is clear that the sequences did cluster by origin species. There were a greater number of corvid and mosquito samples, so these are predominant on the plot, making it difficult to interpret differences in other species. Interestingly, the first break separated a group of corvids from the other species, though not all of the corvid strains were different at this break. The lower levels of the plot further separate the strains. The mosquitoes all cluster together in the center of the plot, which was expected. The small group of hawk species also clustered together, separated at lower levels from the others. The songbird group is also clustered separately, but appear to be more closely related to corvids than mosquitoes, which is interesting. A deeper understanding of the relatedness of West Nile Virus between species might illuminate transmission patterns in the disease that could help researchers contain the illness and prevent further spread.

Maybe that weird Kern sample from your PCA?

Literature Cited Wen C, Wu L, Qin Y, Van Nostrand JD, Ning D, Sun B, Xue K, Liu F, Deng Y, Liang Y, Zhou J.2017. Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. PLoS ONE 12(4): e0176716. https://doi.org/10.1371/journal.pone.0176716

Very nice on describing the data, how you analyzed it, + what you found that was interesting. Need more content on challenges/what you would do differently if you were starting over.

- Did you get in touch with Andersen?
- Repeat of mosquito species
- GitHub remotely
- Map -> extra legends?
- What code does Andersen use?
- reproducibility with different languages
- Clustering legend -> not alphabetically
- Is clustering consistent scientifically?
- bioinf. club
- PCA in R?
- biologists coding?
- batch effects x geography
- Kern obs.?