West Nile Virus Genetic Diversity in California - DRAFT

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

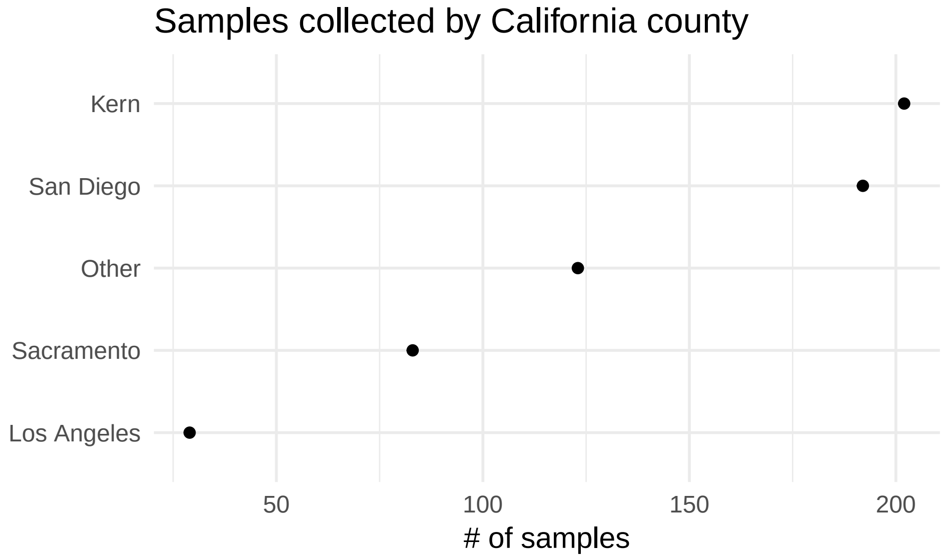
December 5, 2018

Introduction

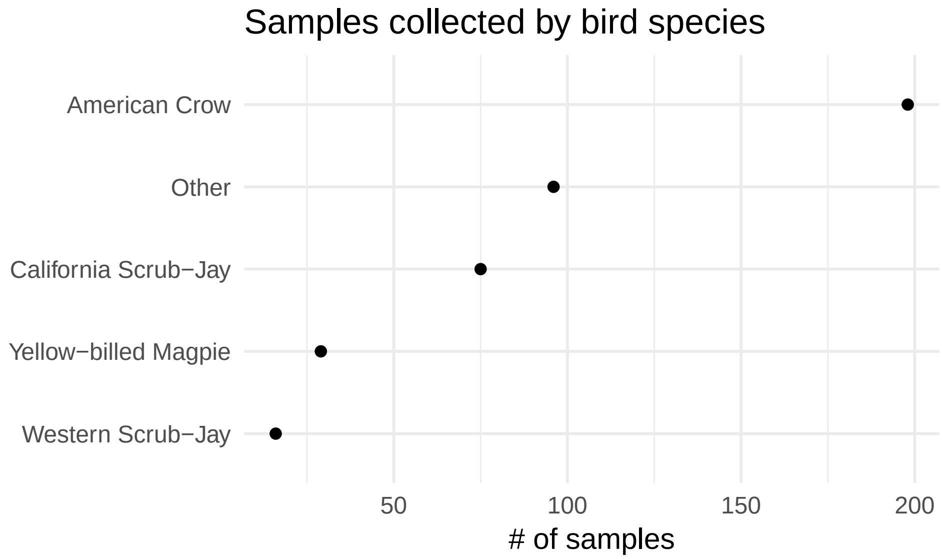
The following project aims to characterize the genomic data of West Nile Virus infections. It took us about a few days to figure out how we were going to approach such a project, and a few more days before we found a viable dataset to use to accomplish our project goals. The data set found and used in this genomic project characterizing West Nile Virus our group used a data set published through Dr. Andersen github: https://github.com/andersen-lab/west-nile. Once our group was assigned found usable data we developed an overarching question that we wanted our project to answer: How do infection rates and genetic structure vary within West Nile Virus in California? We now had a very large dataset with information on species, location, date, and sequence that could be used to answer a small set of questions we proposed to answer for this project.

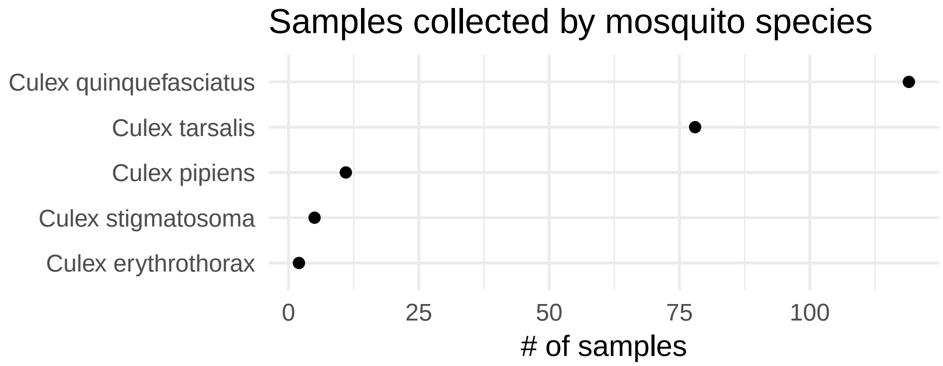
Sample Collection

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. Data was collected from 2004 to 2017.



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.

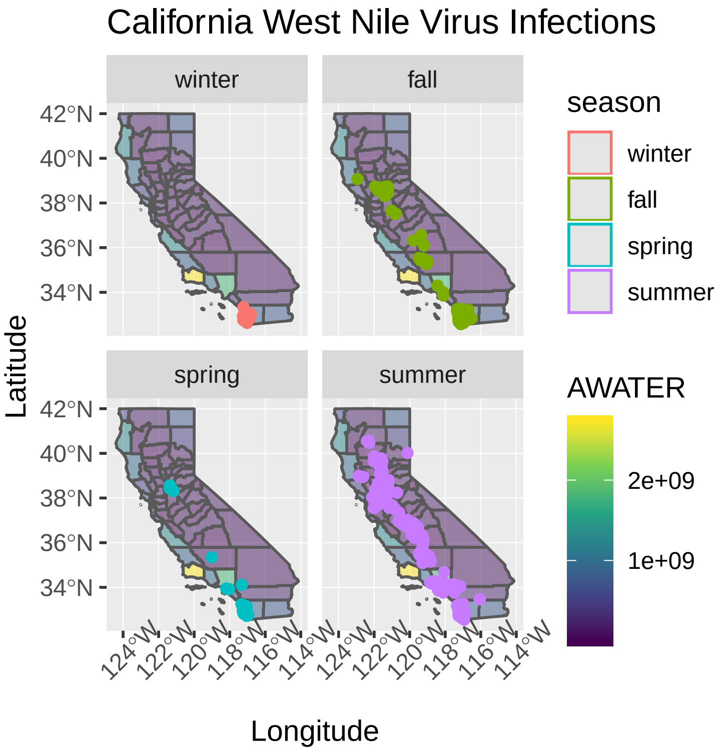


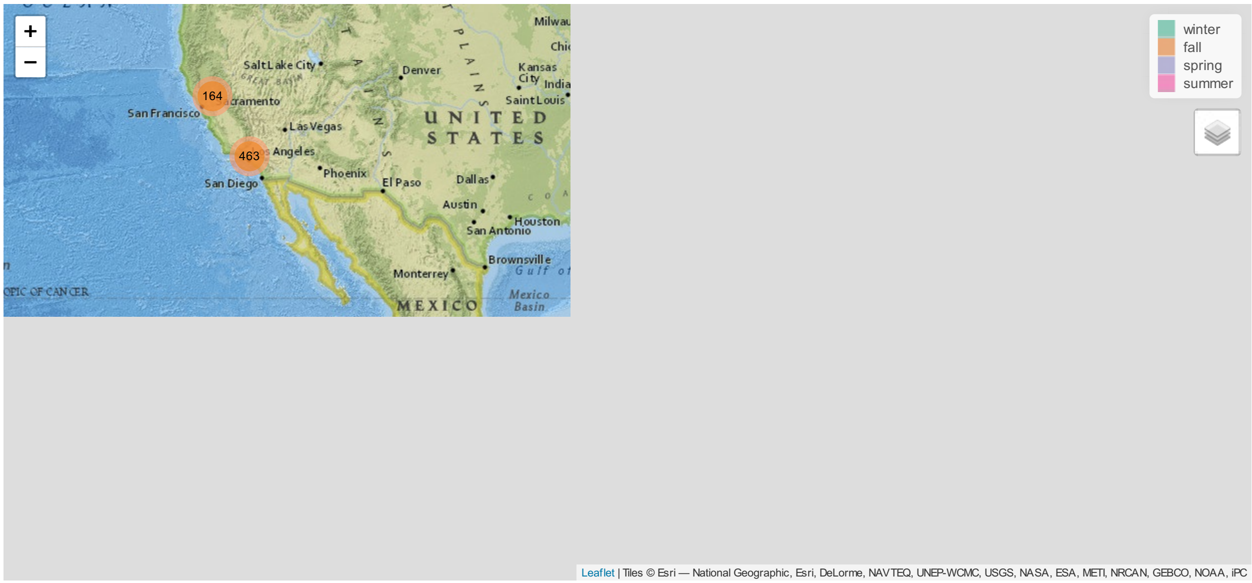


The above graphs demonstrate the variation in samples collected from either bird species or the pooled mosquito populations. It is very interesting to note that most samples were collected from the American Crow. Some literature suggests an overall high prevalence of West Nile Virus in corvids. The American 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.

Interactive map displaying seasonal infection by species

From the visual representations shown below it is clear that a vast majority of West Nile Virus infections occur in the summer. Surprisingly there are quite a few infections occurring in the fall, and almost none in the spring or winter. The data also shows that the american crow is by far the most highly infected species collected in California over this time period. The question I was trying to answer by this plot is how likely is an infection going to occur in the summer near water? There is a somewhat clustered amount of infections around areas of higher water concentration, but I would not say this is a strong correlating factor.





The interactive graph above has color coordinated points at locations of infections with popup information on species and year if wanted. This interactive map allows a much greater amount of detail on collection site to be determined as well as gives good spatial orientation on infection sites.

When I created this plot the questions I was trying to answer were: does WNV infection occur in one species preferentially? Does this occur at a specific time of year? Does the amount of water available in an area affect this occurance? From the visual representations shown above it is clearn that a vast majority of West Nile Virus infections occur in the summer (71%). There are quite a few infections occuring in the fall (20%), and almost none in the spring or winter. WNV infections are also clustered toward the southern portions of the state, while there are few indicences of infection in the northern portion of the state. The data also shows that the american crow (32%) is by far the most higly infected species collected in California over this time period. There is a somewhat clustered amount of infections around areas of higher water concentration, but I would not say this is a strong correlating factor. The interactive graph above has color coordinated points at locations of infections with popup information on species and year if wanted.

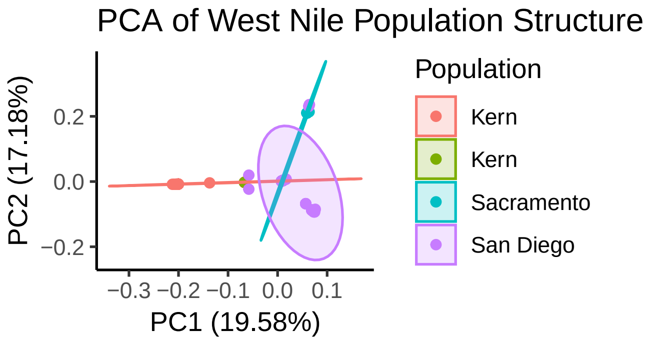
Genomics

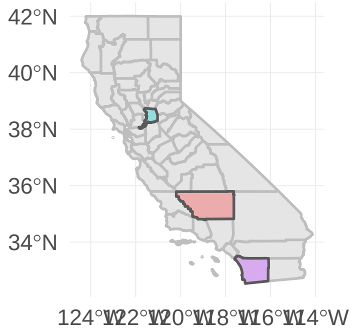
The genomic data from Dr. Andersen's git hub page included 649 consensus sequences of West Nile Viral DNA that has been isolated from multiple species of birds and mosquitoes through amplicon-based sequencing which sequences highly specific regions of the target genome. However, pre-processing of the genomic data for quality control and formatting was necessary before analysis could be conducted.

To start, consensus sequences were aligned using 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. A preliminary filter removed sequences with greater than 10% *N* content or shorter than 10317 bp long (568 sequences matched this criteria). A second, more conservative, filter criteria removed sequences with greater than 0% *Ns* and shorter than 10317 bp in length (157 sequences matched this criteria).

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 disctinct* 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. We then analyzed this polymorphic data set using a Principle Components Analysis (PCA) and plotted using ggplot and ggfortify. The purpose of the PCA is to represent the amount of genetic diversity represented by each principle component (axes of measure, which is the genetic site of variation in this case). Only the most conservative dataset could be used for this analysis as PCAs cannot handle missing or N values. If West Nile virus is panmictic across all of California due to high mobility of the virus preventing divergence, we would expect to see overlap in our three populations with all points grouping together on the axes.



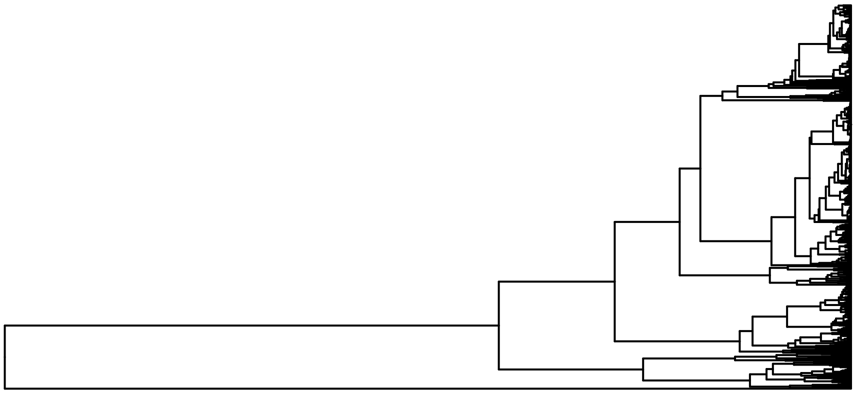


Above we can see our plotted PCA values, alongside a map of California highlighting Kern, Sacramento, and San Diego counties matched by color to the PCA groups. Though there is some amount of overlap between the three populations, there is a clear amount of structure between them as seen by the three elipses. The lines are very thin elipses showing tight grouping (genetic similiarity) of the points (each point represents an individual sequence. 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.



Colors:

- Blue = corvid
- Green = hawk
- 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.

Challenges

It is difficult to analyze genomic data in general; every year the cost of sequencing goes down, so the number of studies conducted is skyrocketing quickly. This rapid change in type of study caused a backlog on the data analysis and bioinformatics side of the research. There is simply not a single, well-established pipeline for analyzing these studies. So, for those of us new to analyzing genomics in R (which is our entire group!) it is difficult to determine which packages are important and valuable. This project required degrees in googling because most of our time was spent searching the internet for a way to do what we wanted. Most of our difficulties stemmed from this. Ideally, as the field progresses, the work will become more open and collaborative and the pipelines will become more standardized to prevent this sort of difficulty in the future.

The genomic data was in FASTA format which includes only sequence ID information and the genetic data. This forced us to use python to filter our data before it could be analyzed since most R Packages that filter sequences need some sort of quality information emedded in the file as FASTQ format or raw reads from the sequencers. In the future, we will bear in mind that FASTA files are only useful to strict R users if read quality information is included. Additionally, further processing of these consensus sequences was difficult because the sequences were different lengths and included varying numbers of missing base pairs. Given that an alignment is necessary to create consensus sequences, we originally believed that we could proceed as though they were aligned; however, the tools we found to do this were not able to process the samples as they were. It was lucky that Rebecca was able to track down a pipeline that allowed us to align the sequences to each other so we could proceed with our analyses.

There are many crs codes for California, which presented a challenge when it came to mapping the infection data to CA geometry. This took a lot of guessing before Dr. Anderson was able to show me a function that can return the crs code being used for the CA geometry, and therefore could be used for the infection locations.