|  |
| --- |
| Table of Contents  [CHAPTER 1: Background and Significance 4](#_Toc191632456)  [CHAPTER 2: Previous Work in Metabolomics & Machine Learning 6](#_Toc191632457)  [2.1 Memory-Efficient Searching of Gas-Chromatography Mass Spectra Accelerated by Prescreening [14] 6](#_Toc191632458)  [2.2 Metabolic Signature Discovery 9](#_Toc191632459)  [CHAPTER 3: Climate Adaptation and Genetic Differentiation in the Mosquito Species *Culex tarsalis* 13](#_Toc191632460)  [3.1 Introduction 13](#_Toc191632461)  [3.2 Materials and Methods 14](#_Toc191632462)  [3.2.1 Sample Collection 15](#_Toc191632463)  [3.2.2 Genome Sequencing, Assembly, and Annotation 15](#_Toc191632464)  [3.2.3 RAD-Seq Library Preparation, Sequencing, and SNP Calling 15](#_Toc191632465)  [3.2.4 Climate Data Extraction 16](#_Toc191632466)  [3.2.5 Population Structure Analyses 18](#_Toc191632467)  [3.2.6 Identification of Genotype-Environment Associations 21](#_Toc191632468)  [3.2.7 Detecting SNPs Under Selection 23](#_Toc191632469)  [3.3 Results 23](#_Toc191632470)  [3.3.1 Genomic Analyses 23](#_Toc191632471)  [3.3.2 Population Structure 24](#_Toc191632472)  [3.3.3 Isolation by Distance (IBD) verse Isolation by Environment (IBE) 28](#_Toc191632473)  [3.3.4 Partial Redundancy Analysis 29](#_Toc191632474)  [3.3.5 Genotype-Environment Associations 30](#_Toc191632475)  [3.3.6 SNPs Under Selection 34](#_Toc191632476)  [3.4 Discussion 40](#_Toc191632477)  [CHAPTER 4: West Nile Virus Prediction 43](#_Toc191632478)  [4.1 Data source 43](#_Toc191632479)  [4.1.1 Clinical Disease Surveillance Data 43](#_Toc191632480)  [4.1.1 Climate Data 46](#_Toc191632481)  [4.1.2 Land Cover Data 47](#_Toc191632482)  [4.1.3 Birds Data 48](#_Toc191632483)  [4.1.4 Mosquitoes’ abundance data 48](#_Toc191632484)  [4.1.5 Demographic Data 49](#_Toc191632485)  [4.2 Methodology 50](#_Toc191632486)  [4.2.1 Data Preprocessing 50](#_Toc191632487)  [4.2.2 Correlation Analysis using the Mantel Test 3](#_Toc191632488)  [4.2.3 Machine Learning / Statistical Models 3](#_Toc191632489)  [4.2.4 Hyperparameter Tuning 4](#_Toc191632490)  [4.2.5 Evaluation Metrics 5](#_Toc191632491)  [4.3 Results 6](#_Toc191632492)  [4.3.1 National CDC Dataset 6](#_Toc191632493)  [4.3.2 Inconsistency in Data Collection 11](#_Toc191632494)  [4.3.3 Significant Class Imbalance 12](#_Toc191632495)  [4.3.4 Data Granularity 15](#_Toc191632496) |

**Harnessing Computational Tools and Complex Biological Data for Environmental and Health Applications**

# Background and Significance

## Introduction to Mosquito-Borne Diseases and West Nile Virus

Mosquito-borne diseases remain one of the most pressing public health challenges worldwide, with West Nile Virus (WNV) emerging as the most prevalent arboviral disease in the United States. Since its initial detection in the United States in 1999, WNV has spread across the entire United States in a few years, with over 59,000 reported cases and 2,900 fatalities by the end of 2023. Approximately 20% of those infected develop West Nile fever, with a critical 1% suffering from severe neurological illnesses [[1]](https://paperpile.com/c/qo3oGi/UCe4U).

Beyond its human health impact, WNV imposes substantial economic and healthcare burdens. Neuroinvasive WNV cases often result in extended hospitalizations, long-term rehabilitation, and permanent disability in some survivors. The direct costs of hospitalization and treatment for severe WNV infections, combined with lost productivity, pose a significant financial burden. Additionally, vector control programs—which include insecticide spraying, habitat modification, and public awareness campaigns—require extensive resources, with varying levels of success depending on environmental conditions and community engagement.

WNV transmission dynamics are intricately influenced by a multitude of factors. Environmental conditions, including temperature, humidity, and precipitation, significantly impact mosquito populations and virus replication rates [[2]](https://paperpile.com/c/qo3oGi/PZO4z). Moreover, geographical features such as land cover patterns and water bodies play crucial roles in shaping vector habitats and breeding grounds [[3]](https://paperpile.com/c/qo3oGi/Q6Hrv). Population density and human behavior also contribute to WNV transmission, as urbanization and human activities can create favorable conditions for mosquito proliferation and contact with infected reservoir hosts [[4,5]](https://paperpile.com/c/qo3oGi/woVNO+6yNy8) . Additionally, In the continental U.S., here are about 12 distinct mosquito species which can transmit diseases to humans, but not all of them have comprehensive genetic resources. *Cx. tarsalis* is a major vector for WNV in the United States [[6]](https://paperpile.com/c/qo3oGi/Tloni), and is a predominant vector of the disease in the most severely impacted states in the West and Midwest [[7]](https://paperpile.com/c/qo3oGi/ZjnAa). Despite its significance, comprehensive genetic resources for *Cx. tarsalis* remains scarce, impeding our ability to elucidate its population dynamics and adaptability. Interestingly, our study of population genetics in *Cx. tarsalis* reveals a pattern of genetic differentiation that suggests a potential role for selection in addition to genetic drift. This pattern hints at environmental adaptations driving population divergence in *Cx. tarsalis*, suggesting that identifying the environmental factors and genetic determinants under selection is vital for predicting the spread of *Cx. tarsalis* and, by extension, WNV outbreaks.

Recent research endeavors have endeavored to predict WNV transmission dynamics through diverse methodological approaches across varied geographical contexts. Holcomb et al. [[8]](https://paperpile.com/c/qo3oGi/10xGH) highlighted the significance of historical disease incidences and population density over climate anomalies in the U.S., while José-María et al. [[9]](https://paperpile.com/c/qo3oGi/s8TOW) found climatic variables, human-related factors, and topo-hydrographic features to be key in Europe. Additionally, John M. Humphreys et al. [[10]](https://paperpile.com/c/qo3oGi/x43cb) pointed out the role of drought in amplifying virus transmission within the U.S. Despite these varied approaches, the overarching theme from these studies suggests that while predictive models can identify potential risk factors and outbreak patterns, the actual predictive power for specific outbreak events remains limited. These findings underscore the complex and multifaceted nature of WNV transmission dynamics, where single factors or models may not capture the full spectrum of variables influencing disease spread, indicating a need for more sophisticated and integrative predictive frameworks.

In summary, the collective research efforts across different regions and disciplines highlight the multifaceted nature of mosquito-borne disease transmission and the critical role of genetic, environmental, and ecological factors in shaping disease dynamics. Understanding the genetic diversity and population structure of vectors like *Cx.* *tarsalis*, alongside environmental and host factors, is paramount in developing comprehensive strategies to mitigate the impact of diseases like WNV and protect public health.

# Previous Work in Metabolomics & Machine Learning

The rapid expansion of metabolomics studies, particularly in databases like Metabolomics Workbench [[11]](https://paperpile.com/c/qo3oGi/8zjB8) and Metabolights [[12]](https://paperpile.com/c/qo3oGi/qyaic), underscores the need for sophisticated tools capable of efficiently managing and analyzing mass spectrometry data. My research focused on refining the ADAP-KDB algorithm [[13]](https://paperpile.com/c/qo3oGi/2CZyb), crucial for efficiently processing mass spectrometry data, including identifying and prioritizing spectra of both known and unknown compounds. This enhancement is key to effectively analyzing extensive metabolomics data, enabling the identification of distinct metabolic signatures. In parallel, I developed a pipeline to discern these signatures through untargeted metabolomics studies. This dual approach aims to reveal robust metabolic patterns, crucial for understanding disease mechanisms and enhancing diagnostic and treatment strategies in biomedical and public health research.

## Memory-Efficient Searching of Gas-Chromatography Mass Spectra Accelerated by Prescreenin*g* [*[14]*](https://paperpile.com/c/qo3oGi/UjBlf)

The original ADAP-KDB spectral search algorithm (Figure 1), which used a relational database for storing and querying spectral data, became increasingly slow with the growth of the spectral database. This slowdown was primarily due to the method used to calculate spectral similarity between a query spectrum and library spectra, which involved sqrt-cosine similarity calculations for each comparison. Each query spectrum needs to be compared with an increasingly large number of library spectra, a process that is computationally intensive. Additionally, using a general-purpose relational database like MySQL, while memory-efficient and cost-effective, may not be the fastest for spectral search, particularly when dealing with vast data. As the database grows, the time taken to match a query spectrum to all the library spectra in the database increases, leading to slower overall performance.

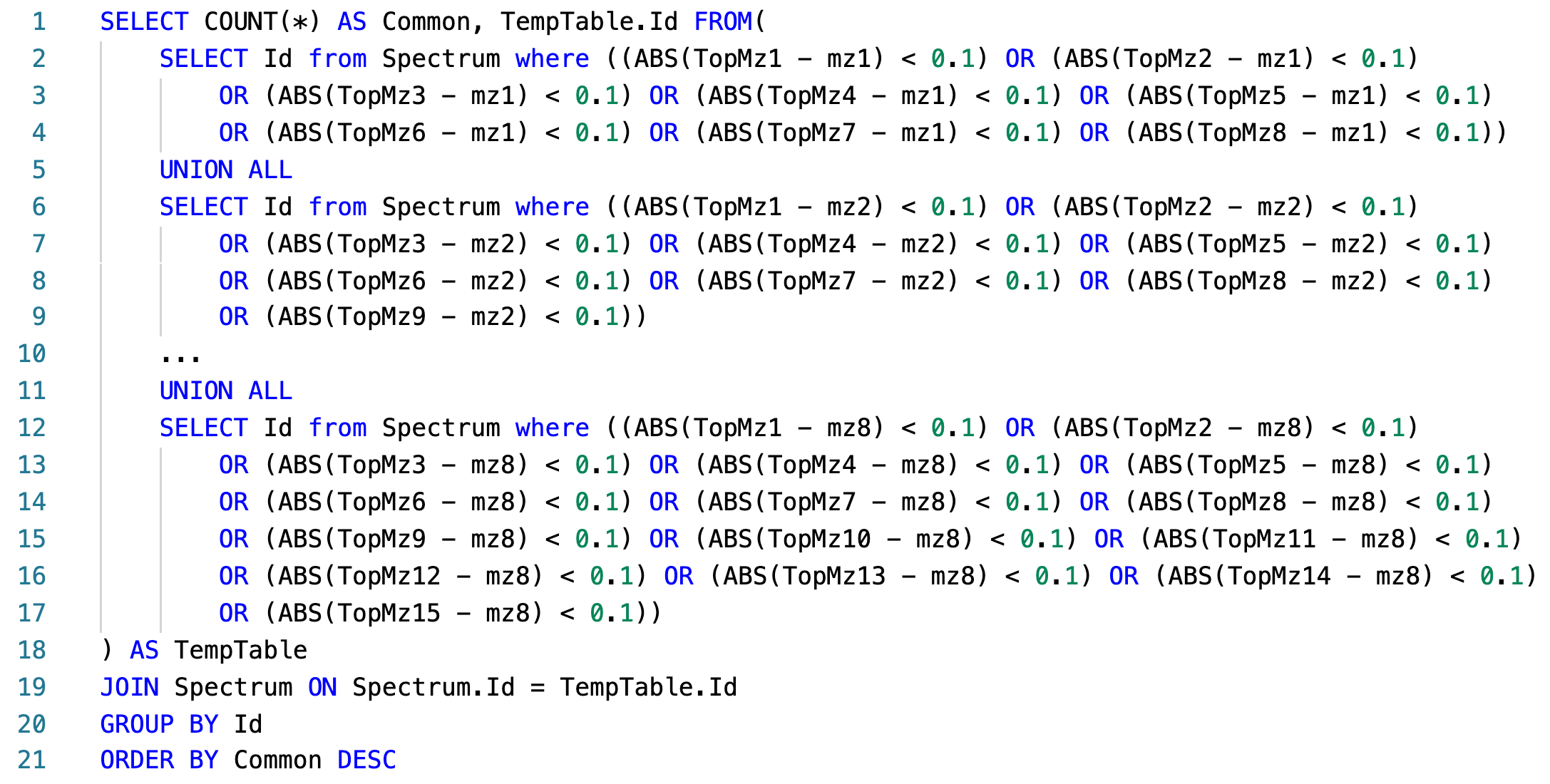


Figure 1: Pseudo SQL query for calculating similarity scores between a query spectrum and all library spectra.

The new algorithm in ADAP-KDB is faster due to the implementation of a prescreening search step (Figure 2). This step allows the algorithm to quickly identify candidate spectra, reducing the need to calculate similarity scores for every library spectrum. For this speed improvement to be effective, three conditions must be met: (1) the pre-screening algorithm is much faster than the main search algorithm, (2) it returns a relatively small number of candidate spectra, and (3) the returned candidate spectra include the correct match to the query spectrum.

In this new algorithm, the process involves pre-calculating 𝑚/𝑧 values of the largest peaks in all library and query spectra. Then, the 𝑚/𝑧 value of the largest peak in the query spectrum is matched to 𝑚/𝑧 values of n largest peaks in the library spectra, and this process continues until the 𝑚/𝑧 value of the n-th largest peak in the query spectrum is matched to 𝑚/𝑧 values of m largest peaks in the library spectra, as illustrated in Figure 2 where (n = 8, m = 15). The library spectra are ranked based on the number of matched peaks. Candidate spectra are determined based on certain criteria, such as the number of matched peaks and a threshold value R. Spectral similarities between the query spectrum and candidate spectra are then calculated, and the candidate spectra with the highest scores are returned to the user.

Calendar

Description automatically generated

Figure 2: New ADAP-KDB spectral search algorithm with the prescreening search: (A) m/z values of 8 largest peaks in the query spectrum are matched to m/z values of 15 largest peaks in every library spectrum; (B) all library spectra are ranked based on the number of matched m/z values, and top 50+ candidate spectra are returned by the prescreening search; (C) the similarity score is calculated for each candidate spectrum.

Figure 3 shows the comparison of execution time between the new library search algorithm and the original algorithm. The result shows that performance of the library search with prescreening is about four-times faster for the low-mass-resolution spectra and very similar for the high-mass-resolution spectra. Moreover, the execution time stays about the same for all pairs of parameters n and m, while pairs (n = 4, m = 7), (n = 6, m = 11), and (n = 8, m = 15) demonstrate slightly better inclusion rate than pairs (n = 4, m = 4), (n = 6, m = 6), and (n = 8, m = 8), respectively. Based on these results, comparing eight largest peaks in the query spectrum to 15 largest peaks in the library spectra seems to be the optimal approach for the preliminary search. Selecting threshold R is based on the tradeoff between the execution time and the inclusion rate. Based on the comparison results, value R = 50 seems to be optimal for keeping high inclusion rate and low execution time for both low-mass-resolution and high-mass-resolution spectra. Therefore, (n = 8, m = 15 and R = 50) were selected as optimal for the prescreening search of the new spectral search algorithm.

Chart, calendar

Description automatically generated

Figure 3: Execution times of the new search algorithm, estimated for prescreening parameters (n, m, R), where n is the number of query spectrum peaks participating in the prescreening search, m the number of library spectrum peaks participating in the prescreening search, and R determines the number of candidate spectra returned by the prescreening search: (A,B) search against low-resolution spectra; (C,D) search against high-resolution spectra.

## Metabolic Signature Discovery

Figure 4 illustrates the workflow for discovering metabolic signatures. Raw data from public metabolomics studies undergo preprocessing using ADAPBIG, a software for processing untargeted mass spectrometry data [[15]](https://paperpile.com/c/qo3oGi/r4uLR). This step includes formatting the data for compatibility with machine learning models. Subsequently, the data undergo cleaning, imputation, scaling, and target creation using a dummy matrix. The processed data is then fed into Partial Least Squares Discriminant Analysis (PLS-DA), a supervised learning algorithm used for identifying metabolite patterns that differentiate sample groups. VIP scores are computed in PLS-DA to quantify the contribution of each metabolite to group separation, with a threshold typically set at 1. Metabolites surpassing this threshold are considered metabolic signatures. Finally, PLS-DA performance is validated through cross-validation, and the results are visualized.

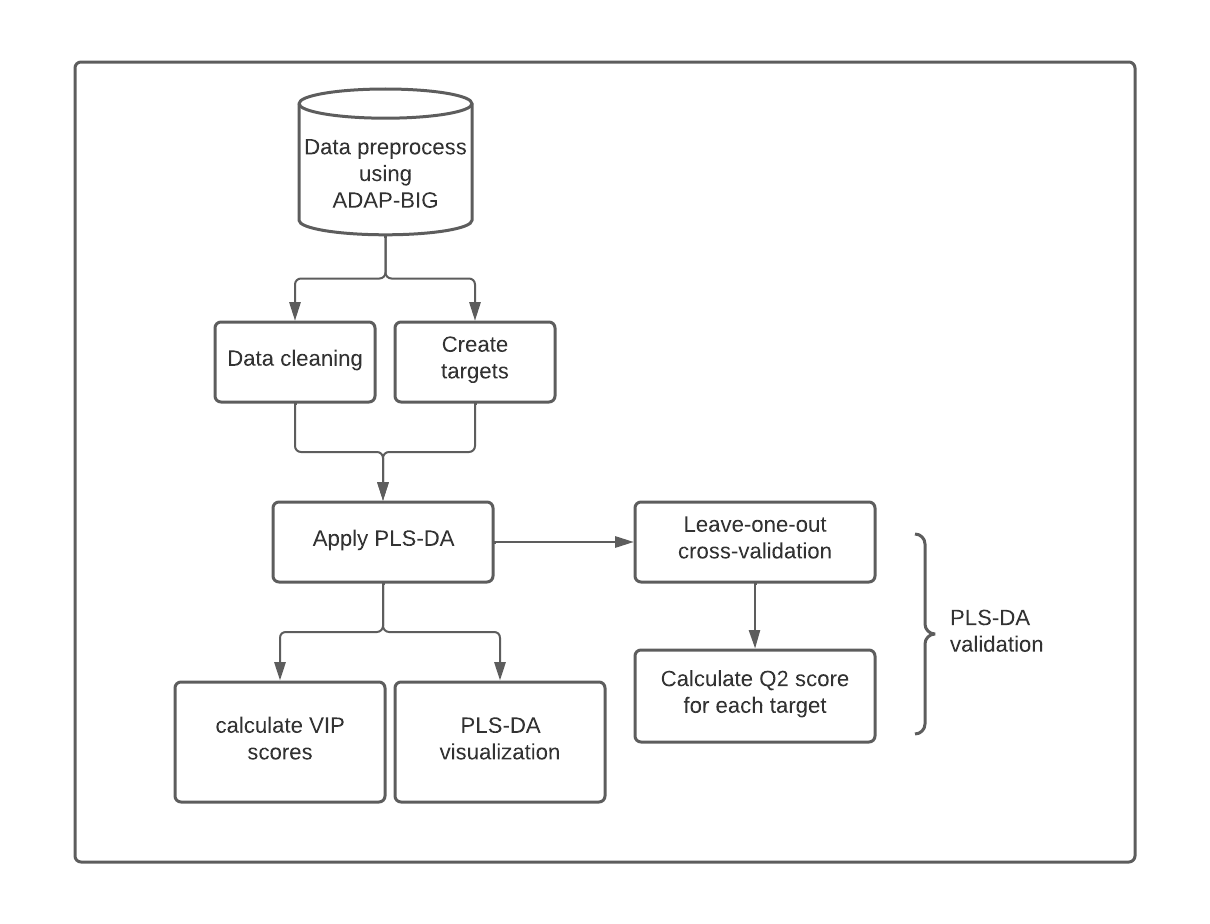


Figure 4: Metabolic Signatures Discover Pipeline

Study ST000058 [[16]](https://paperpile.com/c/qo3oGi/H2zM9) from the Metabolomics Workbench [[11]](https://paperpile.com/c/qo3oGi/8zjB8) investigates alterations in metabolite levels associated with methionine stress sensitivity in cancer using GC TOF MS analysis. The study comprises seven groups, each containing four samples. Group 1 serves as the control, receiving methionine treatment, while groups 2 to 7 undergo homocysteine treatment for varying durations from 2 hours to 48 hours. This design allows for the examination of metabolite responses under different stress conditions, offering valuable insights into cancer metabolism.

Figure 5 presents the PLS-DA visualization of results across six different pairs of groups, demonstrating clear separation between each group. The predictive capability of the model is assessed through Q2 scores after leave-one-out cross-validation, with the highest Q2 observed between the control group and the 48-hour treatment group (0.76), and the lowest between the control group and the 8-hour treatment group (0.36).

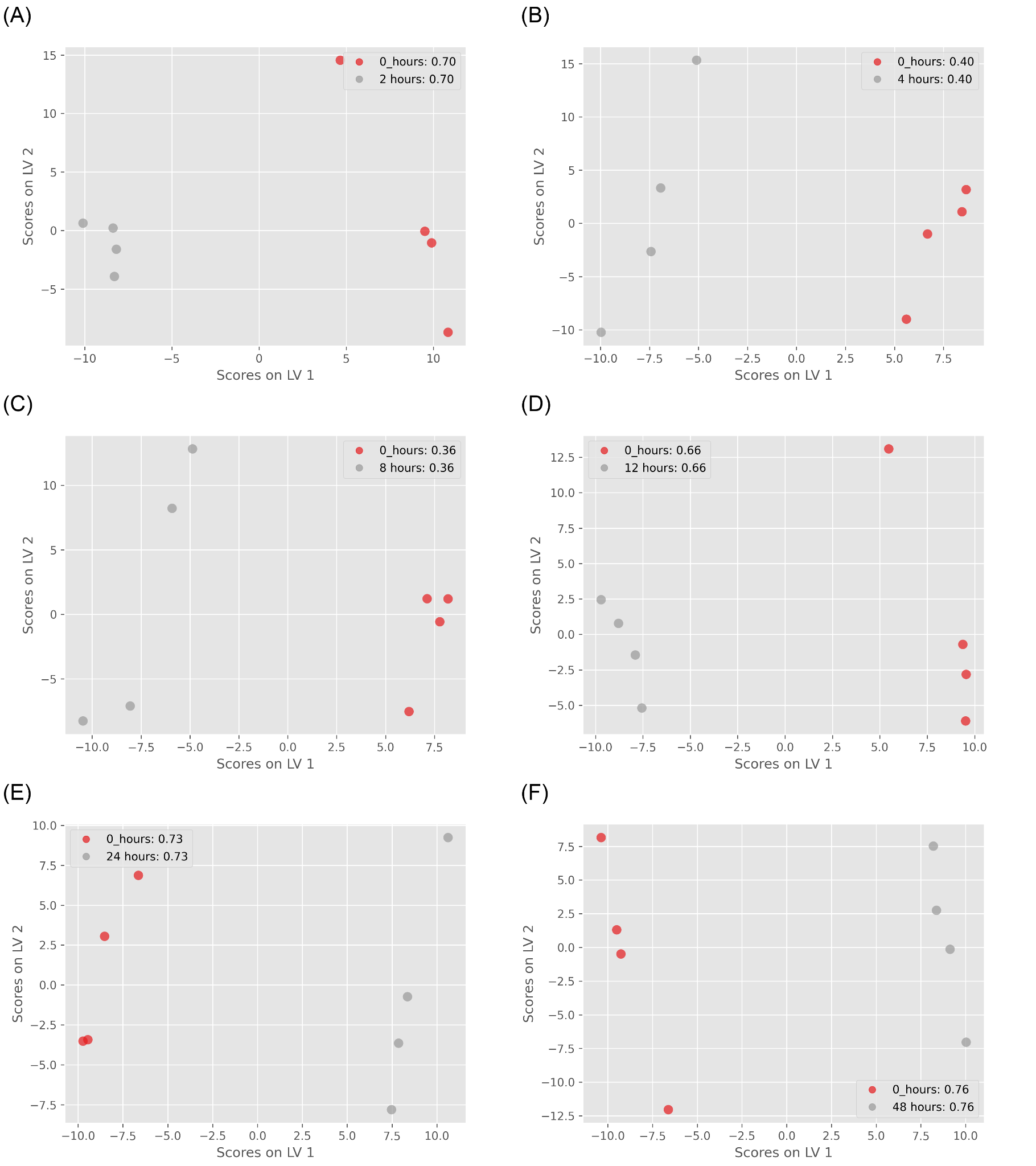


Figure 5: ST000085 PLS-DA results: (A) control group vs. 2 hours treatment group (Q2=0.7); (B) control group vs. 4 hours treatment group (Q2=0.4); (C) control group vs. 8 hours treatment group (Q2=0.36); (D) control group vs. 12 hours treatment group (Q2=0.66); (E) control group vs. 24 hours treatment group (Q2=0.73); (F) control group vs. 48 hours treatment group (Q2=0.76);

To identify metabolic signatures, I conducted PLS-DA analysis on control groups and 48-hour treatment groups, utilizing VIP cutoffs of 1.3 and 1.5 to select candidate metabolites. Additionally, I performed ANOVA tests on the same group pair to compare multivariate and univariate algorithms, correcting the findings with 5% and 10% FDR thresholds. Figure 6 presents the comparison between PLS-DA and ANOVA results. A total of 26 metabolites were identified as candidates by both methods, with 19 uniquely identified by PLS-DA (using a VIP score cutoff of 1.3), and all metabolites from ANOVA were found in PLS-DA results.

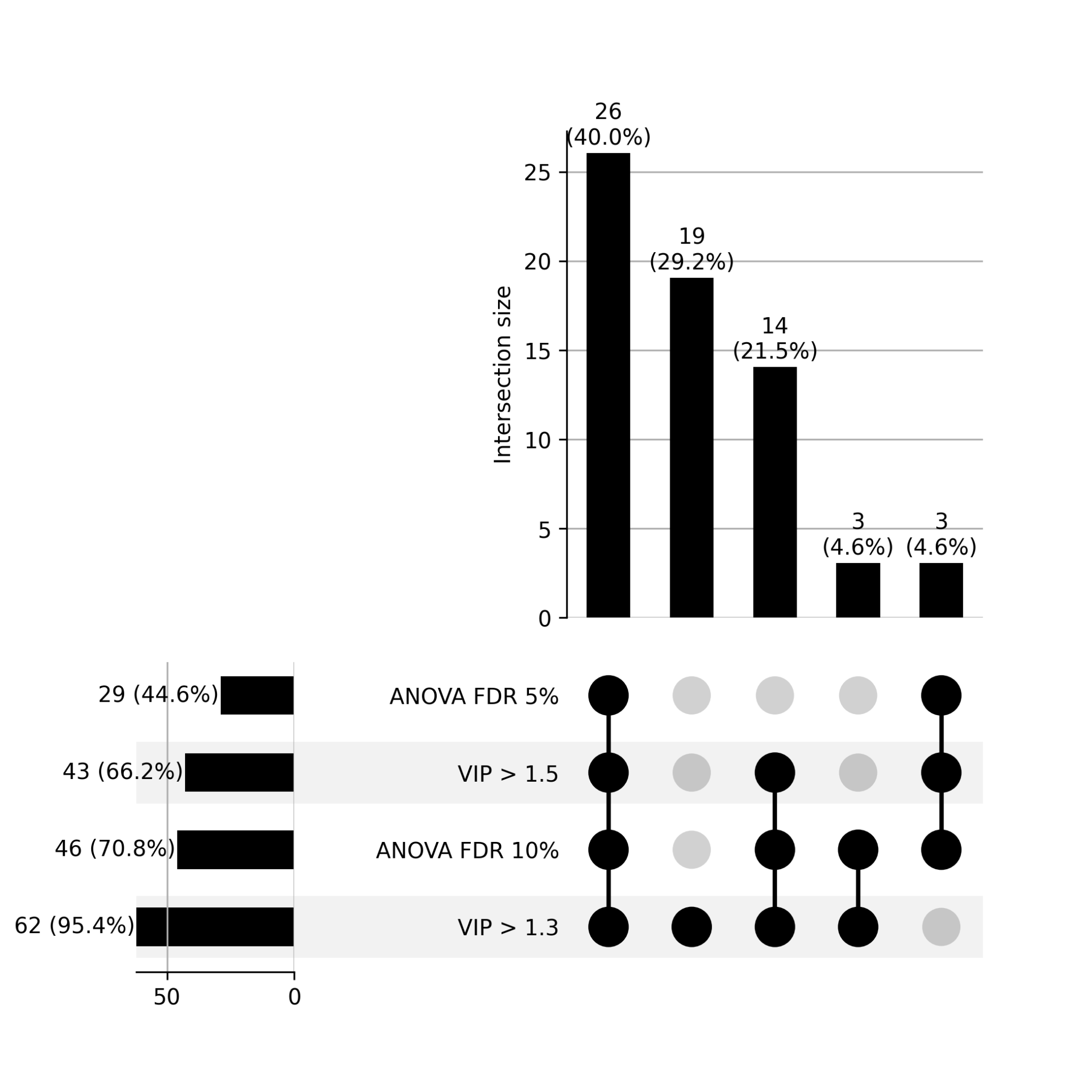


Figure 6: Upset plot between PLS-DA and ANOVA test. VIP cutoff is 1.3 and 1.5 in PLSDA; FDR threshold is 5% and 10% in ANOVA test.

In refining the ADAP-KDB algorithm for metabolomics data analysis, I significantly enhanced the efficiency of processing mass spectrometry datasets, enabling rapid identification of metabolic signatures. This experience honed my skills in algorithm optimization, data preprocessing, and machine learning, equipping me with the tools necessary for tackling complex datasets in genetic and disease prediction research.

Applying these skills, I am poised to contribute to *Cx. tarsalis* population genetics by efficiently analyzing genetic variations and to West Nile Virus prediction by developing models that integrate multifaceted data. The foundation laid in computational data analysis and pattern recognition prepares me for advancing these projects, ultimately aiming to inform public health strategies.

# Climate Adaptation and Genetic Differentiation in the Mosquito Species *Culex tarsalis*

## Introduction

Diseases transmitted by insects are a serious and growing health concern for both humans and livestock. Numerous mosquito species are viral vectors, and the global distribution plus high adaptability of these mosquitoes contributes to the rapid spread and evolution of dangerous diseases across multiple continents [(Gubler 1998)](https://paperpile.com/c/vD89YP/bm0UZ). Understanding how different species of mosquito have increased their ranges and adapted to new habitats in the past will be critical for predicting and managing their potential expansion in the future [(Githeko et al. 2000; Sutherst 2004)](https://paperpile.com/c/vD89YP/KWzYt+oC9xb).

The most abundant vector species found in the United States are derived from two genera: *Culex* and *Aedes*, each of which is comprised of both native and introduced species that occupy a variety of geographical ranges [(Darsie and Ward 2016)](https://paperpile.com/c/vD89YP/ZHIk9). *Culex tarsalis*, also known as the Western Encephalitis Mosquito, is a carrier of several forms of encephalitis that can infect humans as well as animals such as horses[(Reeves 1990; Darsie and Ward 2016)](https://paperpile.com/c/vD89YP/ZHIk9+NI0ml). It is a known vector of West Nile Virus (WNV), Japanese Encephalitis Virus, St. Louis Encephalitis Virus, and Rift Valley Fever Virus [(Evans et al. 2017; Main et al. 2018)](https://paperpile.com/c/vD89YP/ru3zp+b4xtT). *Cx. tarsalis* is abundant in the western continental United States, where it is responsible for the majority of WNV cases in the most severely impacted states [(Goddard et al. 2002; Evans et al. 2017)](https://paperpile.com/c/vD89YP/ru3zp+T5iHz).

Despite the economic and health risks posed by this mosquito species, there is little known about its genetics. Previous studies using microsatellite markers have consistently uncovered a distinct pattern of population structure that does not entirely correlate with current geographical features or indicate strong isolation-by-distance [(Venkatesan and Rasgon 2010; Pfeiler et al. 2013)](https://paperpile.com/c/vD89YP/OYvyJ+wqUhf), but genetic results do support the hypothesis that *Cx. tarsalis* originated on the southwest coast of North America and has since undergone a range expansion to spread eastward [(Venkatesan et al. 2007)](https://paperpile.com/c/vD89YP/a2D5J). While significant adaptive changes must have occurred to allow populations to overwinter in order to cross the Rocky Mountains [(Venkatesan et al. 2007; Diniz et al. 2017)](https://paperpile.com/c/vD89YP/a2D5J+MWr4O), the precise geographic and climatic variables driving divergence among populations are largely unknown.

Although the pattern of genetic differentiation observed in *Cx. tarsalis* could be the result of historical geographic divisions, it is also possible that local environmental adaptations may be driving some or all the population divergence in this species. Many mosquito species, including other *Culex* species, exhibit little to no population structure even after short time spans due to their large population sizes and short generation times [(Wilke et al. 2014; Kotsakiozi et al. 2017)](https://paperpile.com/c/vD89YP/wkZob+w9c2q), so the pattern observed in *Cx. tarsalis* is atypical and suggests a potential role for selection in addition to genetic drift. If the range of present-day populations of *Cx. tarsalis* is defined by adaptation to certain environmental factors, then identifying these factors as well as the genes and alleles under selection is essential to predicting whether or not *Cx. tarsalis* could continue to spread eastward and northward while being a more prevalent threat within the United States and possibly even other countries.

To advance our understanding of population structure and identify alleles linked to local adaptation in *Cx. tarsalis*, we first assembled and annotated a *de novo* reference genome and generated Restriction-Site Associated DNA sequencing (RAD-seq) data for over 300 individuals from 28 diverse geographic locations. We analyzed these RAD-seq markers through a comprehensive landscape genetics framework to explore how various environmental variables influence population differentiation and to identify alleles associated with adaptation to these conditions. By leveraging a broad spectrum of environmental variables, we assessed the adaptive responses of populations to their local environments, enabling the identification of critical genetic-environment associations. This approach reveals how specific climate variables and genetic variants underpin local adaptation strategies across 28 representative *Cx. tarsalis* collection sites. Our findings enrich our understanding of the complex interactions between genetics and environment, providing crucial insights into the ecological dynamics of this mosquito species.

## Materials and Methods

### Sample Collection

Individual mosquitoes were trapped and collected from 28 different locations across the United States and Canada as part of the North American Mosquito Project (NAMP) [(Cohnstaedt et al. 2016)](https://paperpile.com/c/vD89YP/L2cK6). All samples used in this study were collected in 2012 between the months of April and October.

### Genome Sequencing, Assembly, and Annotation

An F4 population was used to generate the reference genome assembly, and high molecular weight DNA was extracted and sequenced on a Pacific Biosciences (PacBio) RS II (University of Delaware). Thirty-five SMRTcells were generated. The resulting reads provided 76X coverage of the ~790Mb *Cx. tarsalis* genome, and were assembled with MECAT [(Xiao et al. 2017)](https://paperpile.com/c/vD89YP/lGV0P).

Gene annotation was completed by MAKER [(Cantarel et al. 2008)](https://paperpile.com/c/vD89YP/fkW6E) using EST and protein data from the *Culex quinquefasciatus* and *Aedes aegypti* mosquitoes. Sequences were downloaded from the NCBI Taxonomy database and both Trinotate and InterProScan were used for functional annotation of the MAKER predicted genes [(Jones et al. 2014; Bryant et al. 2017)](https://paperpile.com/c/vD89YP/N6Jnm+OTUop). The annotated assembly was assessed for completeness and quality using BUSCO [(Seppey et al. 2019)](https://paperpile.com/c/vD89YP/QwuOv) and QUAST [(Gurevich et al. 2013)](https://paperpile.com/c/vD89YP/tmypX).

### RAD-Seq Library Preparation, Sequencing, and SNP Calling

DNA was extracted from individual mosquitoes and libraries were constructed for Restriction-site Associated DNA Sequencing (RAD-Seq) according to previously established protocols [(Etter et al. 2011)](https://paperpile.com/c/vD89YP/lPHKP). The SbfI enzyme was used to digest purified DNA, and individual samples were barcoded prior to Illumina sequencing. Raw sequencing reads were subsequently filtered to remove any reads with an uncalled base, an error in the restriction enzyme cut site, or with an average Phred quality score less than 20 over 15 consecutive nucleotides. Filtered reads were then de-multiplexed using the Stacks software package [(Etter et al. 2011; Catchen et al. 2013)](https://paperpile.com/c/vD89YP/lPHKP+bQK3R).

After de-multiplexing, raw reads from each individual were aligned to the draft assembly of the *Cx. tarsalis* genome using BWA MEM [(Li and Durbin 2009)](https://paperpile.com/c/vD89YP/k0m3c), and individuals with poor mapping rates (less than 50%) were excluded from subsequent analyses. The mapped reads for the remaining 378 samples were then merged using the Samtools pipeline [(H. Li et al. 2009)](https://paperpile.com/c/vD89YP/D1hx9) and SNPs were called using the GATK HaplotypeCaller [(McKenna et al. 2010)](https://paperpile.com/c/vD89YP/Nb5AK). The SNPs were filtered using VCFtools v0.1.12a [(McKenna et al. 2010; Danecek et al. 2011)](https://paperpile.com/c/vD89YP/Nb5AK+NFX9b) to retain only sites with a minimum average individual read depth of 10X and a maximum of 20% missing data, resulting in a total of 457,387 sites. Individual samples were then filtered again to remove individuals with missing data at more than 20% of the remaining SNP sites, leaving 322 samples from 28 different locations for further analysis. The samples of the 28 different locations were group into 4 group based on the ADMIXTURE results. Coding and noncoding SNP effects were predicted using SIFT4G [(Vaser et al. 2016)](https://paperpile.com/c/vD89YP/aNT8Z).

### Climate Data Extraction

Climate data was extracted from the ERA5-Land monthly averaged dataset provided by the Copernicus Climate Change Service [(Copernicus Climate Change Service 2019)](https://paperpile.com/c/vD89YP/7hLfH). The original dataset was characterized by a temporal resolution of 1 hour and a native spatial resolution of 9 km on a reduced Gaussian grid (TCo1279). To facilitate broader accessibility and suitability for diverse analyses, the data underwent regridding to a regular lat-lon grid with a finer resolution of 0.1x0.1 degrees.

To analyze environmental adaptation in *Cx. tarsalis*, we selected environmental factors based on their potential influence on the mosquito’s life cycle and their role in West Nile virus transmission dynamics. [Humphreys JM, Pelzel-McCluskey AM, Cohnstaedt LW, McGregor BL, Hanley KA, Hudson AR, Young KI, Peck D, Rodriguez LL, Peters DPC. Integrating Spatiotemporal Epidemiology, Eco-Phylogenetics, and Distributional Ecology to Assess West Nile Disease Risk in Horses. Viruses. 2021 Sep 12;13(9):1811. doi: 10.3390/v13091811. PMID: 34578392; PMCID: PMC8473291.] In this study, we initially extracted a total of 13 environmental variables for comprehensive analysis. These variables encompass a diverse range of climate parameters: 10m eastward wind, 10m northward wind, 2m temperature, evaporation from bare soil, leaf area index for high vegetation, leaf area index for low vegetation, water retention capacity of land, snowfall, surface net solar radiation, surface runoff, total evaporation, total precipitation, and volumetric soil water layer 1. Table 3.1 provides detailed descriptions of each variable.

Table 3.1: Description of environmental variables

|  |  |  |
| --- | --- | --- |
| Variable Name | Short Name | Description |
| 10m eastward wind | eastward\_wind | The eastward component of wind speed at 10 meters above the ground. |
| 10m northward wind | northward\_wind | The northward component of wind speed at 10 meters above the ground. |
| 2m temperature | temperature | Air temperature measured at 2 meters above the ground. |
| evaporation from bare soil | evaporation | The amount of water evaporating directly from bare soil surfaces. |
| leaf area index for high vegetation | high\_vegetation | One-half of the total green leaf area per unit horizontal ground surface area for high vegetation type. |
| leaf area index for low vegetation | low\_vegetation | One-half of the total green leaf area per unit horizontal ground surface area for low vegetation type. |
| water retention capacity of land | water\_retention\_capacity | Amount of water in the vegetation canopy and/or in a thin layer on the soil. |
| snowfall | sf | The total amount of snow that falls over a specified period. |
| surface net solar radiation | ssr | The net amount of solar radiation reaching the earth's surface. |
| surface runoff | surface\_runoff | The amount of water that flows over the land surface after rainfall. |
| total evaporation | evaporation | The total amount of water evaporated from all sources, including soil and vegetation. |
| total precipitation | tp | Accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface. |
| volumetric soil water layer 1 | swvl1 | Volume of water in soil layer 1 (0 - 7 cm) of the ECMWF Integrated Forecasting System. |

Prior to processing this complex dataset, steps were taken to minimize variable redundancy. This involved evaluating the variance for each variable across all samples, examining pairwise correlation coefficients, and assessing the distribution of each variable. Variables with a total variance across samples of zero, a pairwise correlation coefficient exceeding 0.70, or showing extreme distribution patterns were identified and eliminated. This refinement process led to the removal of 5 variables: surface net solar radiation, total evaporation, total precipitation, volumetric soil water layer 1, and snowfall. Consequently, the variable set was reduced to 8 environmental variables for all subsequent analyses (Figure 3.1).

A screenshot of a graph

Description automatically generated

Figure 3.1: Correlations plot of each pair climate variables

The reduced dataset was processed using Python scripts utilizing the xarray and pandas libraries. These scripts calculated the averages for the 8 remaining climate variables for the year 2012. The data processing was specifically tailored to extract meaningful climate insights from NetCDF format files from the ERA5-Land monthly averaged dataset, with each variable's average computed based on its geographical location coordinates. This approach aimed to transform the extensive climate data into a format suitable for later Genome-Environment Association (GEA) [(Kamvar et al. 2017)](https://paperpile.com/c/vD89YP/9Ca9t) analyses.

### Population Structure Analyses

For population structure analyses, SNPs were further filtered using VCFtools v0.1.16 [(Danecek et al. 2011)](https://paperpile.com/c/vD89YP/NFX9b) to retain only bi-allelic sites with a maximum of 20% missing data per site and a minimum minor allele frequency (MAF) of 5% (Table 3.2).

Table 3.2: Filter on SNP file for each method in the paper

|  |  |
| --- | --- |
| Methods | Filter applied on original SNP file |
| Admixture | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |
| AMOVA | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |
| LFMM | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |
| RDA | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |
| Bayescan | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |
| PCAdapt | bi-allelic, Maximum of 20% missing data per site, Minimum minor allele frequency of 5 % |

The ADMIXTURE v1.3.0 program [(Alexander et al. 2009)](https://paperpile.com/c/vD89YP/OK9lU) was subsequently employed on the remaining 17,239 loci and 322 individuals. Multiple runs of ADMIXTURE were conducted for K values ranging from 1 to 13, with each K value analyzed across 10 independent runs using different random number seeds in order to ensure convergence. The best K value was determined through a cross-validation process ranging from K=2 to K=13 (Figure 3.2). Subsequently, ancestry proportion bar plots were generated using Python's pandas library (plot.bar()), and floating pie charts displaying ancestry proportions over the U.S. map were plotted using the ggplot2 package in R [(Villanueva and Chen 2019)](https://paperpile.com/c/vD89YP/O16MW) along with scatterpie() in the scatterpie package [(Yu 2024)](https://paperpile.com/c/vD89YP/Z3gyO).

A line graph with dotted lines

Description automatically generated

Figure 3.2: The cross-validation error plot of ADMIXTURE. Optimal number of Clusters (K) values ranging from 1 to 13.

Complementing this, a Principal Component Analysis (PCA) was performed in R to aid in determining the pattern of genetic differentiation and the optimal K value by visualizing the dataset in a reduced dimensional space. As PCA requires no missing data, the missing genotype values for a given SNP were imputed using the most common genotype at each SNP across all individuals.

An Analysis of MOlecular VAriance (AMOVA) [(Excoffier et al. 1992)](https://paperpile.com/c/vD89YP/918zU) was performed using the poppr package's poppr.amova() function in R [(Kamvar et al. 2014)](https://paperpile.com/c/vD89YP/DzFjg) [(Excoffier et al. 1992)](https://paperpile.com/c/vD89YP/918zU) to detect population differentiation among regions inferred from ADMIXTURE results. Pairwise FST values between all populations were computed using the hierfstat package's genet.dist() function in R [(Goudet 2005)](https://paperpile.com/c/vD89YP/5DUz2), employing the Weir-Cockerham estimator [(Weir and Cockerham 1984)](https://paperpile.com/c/vD89YP/wJ5jh). To assess whether there was a statistically significant level of population structure, a randomization test was conducted using the randtest() function from the ade4 package in R [(Thioulouse et al. 2018)](https://paperpile.com/c/vD89YP/iZiZg), employing 999 replicates.

To analyze isolation by distance (IBD) [(Slatkin 1993)](https://paperpile.com/c/vD89YP/v6dKx) and isolation by environment (IBE) [(Wang and Bradburd 2014; Jiang et al. 2019)](https://paperpile.com/c/vD89YP/PollG+MFQNO) (Wang IJ, Bradburd GS. Isolation by environment. Mol Ecol. 2014 Dec;23(23):5649-62. doi: 10.1111/mec.12938. Epub 2014 Oct 16. PMID: 25256562.) (Chang, CW., Fridman, E., Mascher, M. *et al.* Physical geography, isolation by distance and environmental variables shape genomic variation of wild barley (*Hordeum vulgare* L. ssp. *spontaneum*) in the Southern Levant. *Heredity* 128, 107–119 (2022).) patterns within the 28 mosquito populations, genetic distances, derived from Weir & Cockerham FST estimations [(Weir and Cockerham 1984)](https://paperpile.com/c/vD89YP/wJ5jh) based on SNP data, were compared with geographic and environmental distances computed based on latitude and longitude coordinates using the Haversine distance. Environmental distances were calculated in R using the dist() function, applying the Canberra method to emphasize relative differences in 8 non-negative environmental variables sourced from the Copernicus Climate Change Service for each geographic location. Scatterplots were created in R and linear regression models were fitted to each plot. Mantel tests [(Sokal and Rohlf 1995; Wagner and Fortin 2015)](https://paperpile.com/c/vD89YP/A3Glj+yDiCN) were used to assess the correlations between genetic distance and either geographic or environmental distances. The significance of these relationships was determined using 9999 permutations.

A mixed model was utilized to analyze the relationships between genetic distance, geographic distance, and environmental distance. To address potential collinearity issues that could create confounding results in the mixed model, variance inflation factor (VIF) assessments [(O’brien 2007)](https://paperpile.com/c/vD89YP/o5UWZ) were conducted to confirm low multicollinearity for each of the three distance matrices (Table 3.3).

Table 3.3: Variance Inflation Factor results for mixed model variables.

|  |  |
| --- | --- |
| Variables | VIF |
| genetic distance | 1.46 |
| geographic distance | 1.49 |
| Environmental distance | 1.04 |

To complement this analysis, a two-dimensional kernel density calculation was applied to visualize the concentration of data points within the scatterplots. To discern potential non-linear relationships between the distance matrices, a locally estimated scatterplot smoothing technique, implemented via the loess.smooth() function in R [(Gareth et al. 2013)](https://paperpile.com/c/vD89YP/l2KGE) , was employed. This nonlinear fit was compared against the linear model, with R² scores for both models calculated to assess and contrast their respective fits.

### Identification of Genotype-Environment Associations

To explore potential adaptive divergence in *Cx. tarsalis* populations, we employed two different approaches: a genome-environment association (GEA) [(Kamvar et al. 2017)](https://paperpile.com/c/vD89YP/9Ca9t) via latent factor mixed models (LFMM) [(Frichot et al. 2013; Caye et al. 2019)](https://paperpile.com/c/vD89YP/1Vr8V+Dp23A)[(Sokal and Rohlf 1995; Wagner and Fortin 2015)](https://paperpile.com/c/vD89YP/A3Glj+yDiCN) and a redundancy analysis (RDA) [(van den Wollenberg 1977)](https://paperpile.com/c/vD89YP/lbgKq) implemented in the vegan package in R [(Oksanen et al. 2019)](https://paperpile.com/c/vD89YP/gUt5c). For both analysis, SNPs were filtered to retain only bi-allelic sites with a maximum of 20% missing data per site and a minimum minor allele frequency (MAF) of 5%.

The LFMM approach (Figure 3.x) utilizes a univariate testing framework, modeling each SNP and environmental variable using the lfmm\_test() function from the LFMM package in R [(Caye et al. 2019)](https://paperpile.com/c/vD89YP/Dp23A). We initially conducted a Principal Component Analysis (PCA) on the environmental variables, focusing on the first principal component—a linear combination of the 8 climate variables—as the predictor in the LFMM. We also consider the inclusion of the second and third principal components as additional predictors in a subsequent LFMM, as detailed in the Supplemental Materials (Supplemental Figure 8 and Supplemental Table 15 - 17)

A diagram of a model

AI-generated content may be incorrect.

Figure 3.x LFMM Workflow to Extract the Environmental Adaptive Candidate SNPs

In assessing the association between SNPs and environmental variables using the LFMM, we evaluated the Genomic Inflation Factor (GIF) to determine the model's efficacy in handling potential confounding factors. In our analysis, a Genomic Inflation Factor (GIF) value of 1.23 was used to adjust the p-values for potential inflation due to confounding variables, a process that is automatically incorporated within the output of the lfmm\_test() function in R [(Caye et al. 2019)](https://paperpile.com/c/vD89YP/Dp23A). Subsequently, we converted these GIF-adjusted p-values to q-values using the qvalue() function in R. This conversion is to refine the significance thresholds for individual tests, particularly under the framework of multiple hypothesis testing. It enhances the precision of FDR control, crucial in large-scale testing scenarios like ours. We then employed FDR control measures using these q-values, identifying candidate results falling below our predefined FDR threshold of 0.1.

The RDA operates as a multifaceted ordination technique, evaluating multiple loci concurrently with environmental variables (Figure 3.xx). For this analysis, the significance of both the overall RDA model and its individual constrained axes was assessed. This assessment utilized the anova.cca() function from the vegan package in R [(Oksanen et al. 2019; Borcard et al.)](https://paperpile.com/c/vD89YP/gUt5c+B5cgb), facilitating a comprehensive examination of the null hypothesis (an absence of a linear relationship between SNP data and environmental variables). To select candidate SNPs for local adaptation, we identified SNPs that significantly deviated from the mean loadings, using a threshold of 3 standard deviations. Statistical significance was determined based on p-values below the threshold of 0.001.

A diagram of a system

AI-generated content may be incorrect.

Figure 3.xx RDA Workflow to Extract the Environmental Adaptive Candidate SNPs

To visualize results and identify relationships between RDA components and other factors, we utilized the vegan package in R [(Oksanen et al. 2019; Borcard et al.)](https://paperpile.com/c/vD89YP/gUt5c+B5cgb)to generate RDA tri-plots between each pair of the RDA components. To enhance the interpretability of our ordination plots, we employed symmetrical scaling [(Borcard et al.)](https://paperpile.com/c/vD89YP/B5cgb). This scaling method adjusts the SNP and individual scores by the square root of the eigenvalues, providing a clearer representation of the relationships between variables and samples.

### Detecting SNPs Under Selection

To find which SNPs that were significantly associated with environmental variables also showed signatures of natural selection, we performed a Bayesian selection inference as implemented in BayeScan [(Foll and Gaggiotti 2008; Foll et al. 2010; Fischer et al. 2011)](https://paperpile.com/c/vD89YP/Lnm2R+ZZizr+7B6Rz). We also independently identified outliers using a PCA-based method implemented in the pcadapt R package [(Privé et al. 2020)](https://paperpile.com/c/vD89YP/FVkK7), and then filtered for common candidate SNPs that were identified as significant across 4 different analyses: LFMM, RDA, BayeScan, and PCAdapt.

## Results

### Genomic Analyses

The final genome assembly contained 968,887,694 bases divided into 7,478 contigs. The N50 was 451,230 bp. The annotation included 43,905 predicted genes. The assembly quality, assessed using BUSCO version 5.1.3 with the Diptera\_odb10 lineage dataset (3285 BUSCOs), showed 88.2% complete BUSCOs (82.6% single-copy and 5.6% duplicated), 4.4% fragmented BUSCOs, and 7.4% missing BUSCOs. After aligning the RAD-seq reads and filtering for quality, there were 457,387 polymorphic sites identified across all populations.

### Population Structure

The ADMIXTURE analysis indicated a strong signature of population structure among the collected samples, with the optimal number of population assignments occurring at K=4 (Figure 3.2). The genetic clusters corresponded to four different broad geographic regions: (1) California/the West Coast, (2) the Southwest, (3) the Northwest, and (4) the Midwest (Figure 3.3).

The PCA results confirmed this pattern (Figure 3.3 and Figure 3.4), while also showing evidence of some sub-structure among the West Coast and Northwest populations (Figure 3.3C).

A map of the united states with different colored spots

Description automatically generated

Figure 3.3: Population Structure of Cx. tarsalis: (A) Floating pie charts of the admixture proportions in *Cx.* *tarsalis* populations sampled across the Western and Midwestern U.S and parts of Canada. Pie chart sizes are proportional to the sample size at each collection site. (B) ADMIXTURE results for K=4. Labels along the x-axis indicate sampling locations and colors correspond to the admixture proportion for each of the 4 clusters. (C) PCA results for the top 2 principal components, with points colored by the 4 geographic regions identified by ADMIXTURE.

A graph with a red line

Description automatically generated

Figure 3.4: Scree plot for PCA

The AMOVA results also indicated significant levels of population differentiation both between different populations within the same geographic region, and between different regions (Table 3.4 and Figure 3.5). The observed genetic variation within populations was significantly lower (p < 0.001) than expected (Figure 3.6 and Table 3.5), while variation between populations and between regions was significantly higher (p < 0.001) than would be expected by chance (Figure 3.6b and 3.6c, Table 3.5).

Table 3.4: Analysis of Molecular Variance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variation | Degree of Freedom (Df) | Sum of Squares | Mean Squares | % of Variation | Phi - statistics |
| Within samples | 290 | 410029.99 | 1413.9 | 78.86 | ɸ-populations-total = 0.21 |
| Between samples within regions | 28 | 51936.76 | 1854.88 | 2.54 | ɸ-populations-regions = 0.03 |
| Between regions | 3 | 85067.8 | 28355.93 | 18.6 | ɸ-regions-total = 0.19 |
| Total | 321 | 547034.55 | 1704.16 | 100 |  |

A group of graphs showing different types of data

AI-generated content may be incorrect.

Figure 3.6: Analysis of Molecular Variance (AMOVA). Gray bars indicate the expected distributions based on 999 random permutations, while black bars indicate the observed phi values. (A) Variation within populations, (B) Variation between populations within the same region, and (C) Variation between different regions.

Table 3.5: Permutation Tests of Observed and Expected Variation at different levels of population strata

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Monte-Carlo tests - permutation number: 999 | | | | |
| Test | Obs. | Std. Obs | Alter | P-value |
| Variations within populations | 1397.8 | -62.38 | less | 0.001 |
| Variations between populations within regions | 33.42 | 26.25 | greater | 0.001 |
| Variations between regions | 377.71 | 15.13 | greater | 0.001 |

### Isolation by Distance (IBD) verse Isolation by Environment (IBE)

Both geographic distance (IBD) and environmental distance (IBE) were tested for their relationships with genetic distance. The Mantel test revealed a strong and statistically significant correlation between genetic and geographic distances, with a Mantel statistic of 0.5661 (p < 0.001), indicating that geographic distance plays a notable role in genetic differentiation in this system. In contrast, the relationship between genetic and environmental distances was weak and non-significant, with a Mantel statistic of -0.05185 (p = 0.7741).

The mixed model analysis provided further insights. The full model, which incorporated both geographic and environmental factors, offered a more comprehensive explanation of genetic distance variations than models considering either factor alone (Table 3.6). Although the geographic distance model had a slightly lower Bayesian Information Criterion (BIC), the full model’s lower Akaike Information Criterion (AIC) suggests that combining geographic and environmental factors better captures the complexity of genetic differentiation. While geographic distance appears to have a stronger influence, these results suggest that environmental factors may contribute in more subtle or context-dependent ways to genetic variation. The kernel density and LOESS plots in Figure 3.7 highlight these relationships. Figure 3.7C shows a moderate to strong non-linear relationship between geographic and genetic distances, suggesting variability across different geographic ranges and the influence of complex factors beyond simple isolation by distance. The close alignment of the LOESS fit with the linear regression in Figure 3.7D suggests that most of the variation in the relationship between genetic and environmental distances can be captured by a simple linear model. However, the weak correlation (low R² of 0.0014) indicates that environmental factors may not exert a strong, direct influence on genetic variation at the global scale. This uniformity may mask subtle or context-specific interactions that are not evident in pairwise relationships.

A collage of different colored graphs

Description automatically generated

Figure 3.7: Isolation-by-Distance and Isolation-by-Environment. (A) Pairwise geographic distance versus genetic distance (FST) with best fit linear regression model (red line: ). (B) Pairwise environmental distance versus genetic distance with best-fit linear regression mode (red line: ). (C) Kernel density plot with best fit spline for geographic distance versus genetic distance. Areas of high, intermediate, and low density are represented by red, yellow, and blue colors, respectively. (D) Kernel density plot with best-fit spline for environmental distance versus genetic distance. Each point in panels A, B, C, and D represents one individual sample.

**Table 3.6: Mixed model results for IBD and IBE**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models | AIC | BIC | k | AICc | AICcmin | BICew |
| Full | 695.801 | 715.476 | 5 | 695.963 | 0.526 | 0.137 |
| Distance | 696.061 | 711.801 | 4 | 696.169 | 0.474 | 0.863 |
| Environment | 1009.195 | 1024.934 | 4 | 1009.302 | 0.000 | 0.000 |

### Partial Redundancy Analysis

To further investigate the potential role of environmental factors, we performed partial redundancy analysis (partial-RDA) to separate and evaluate the individual contributions of geographic (IBD) and environmental (IBE) factors to genetic differentiation. Partial RDA results revealed that environmental factors, when controlling for geographic effects, explained 3.1% of the variance in genetic differentiation *( = 0.0309, p = 0.001).* In contrast, geographic factors, when controlling for environmental effects, explained a smaller yet statistically significant proportion of variance *(1.34%,  = 0.0134, p = 0.001).* These findings suggest that environmental factors may play a more prominent role in shaping genetic differentiation in the studied system, even when accounting for geographic structure.

Notably, among the environmental variables, low vegetation and evaporation showed the strongest influence on the first canonical axis (RDA1), highlighting their significant roles in shaping genetic variation (Table 3.7). Eastward wind contributed heavily to the second axis (RDA2), while water retention capacity and high vegetation were key drivers on the third and fourth axes (RDA3 and RDA4). These results suggest that specific environmental factors, particularly those related to vegetation and water dynamics, play a notable role in genetic differentiation, even after accounting for geographic effects.

Table 3.7: Biplot Scores for Constraining Variables in Partial RDA (Environment Controlled for Geography)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | RDA1 | RDA2 | RDA3 | RDA4 | RDA5 | RDA6 |
| eastward\_wind | -0.04295 | 0.6138 | -0.38221 | -0.068225 | 0.2848 | 0.4225 |
| northward\_wind | -0.01733 | -0.1283 | -0.60759 | 0.469732 | -0.1585 | -0.2478 |
| temperature | 0.25899 | -0.2221 | -0.03984 | 0.01674 | 0.1733 | -0.1052 |
| high\_vegetation | 0.07189 | 0.4416 | 0.1104 | 0.411215 | 0.4095 | -0.3103 |
| low\_vegetation | -0.78515 | -0.3226 | 0.0345 | -0.004408 | 0.1817 | -0.2035 |
| water\_retention\_capacity | -0.50671 | -0.2401 | 0.33451 | 0.665514 | 0.0183 | 0.1856 |
| surface\_runoff | -0.37392 | -0.4824 | 0.22363 | 0.196856 | 0.353 | 0.256 |
| evaporation | 0.56079 | 0.3324 | -0.01991 | -0.094839 | -0.2216 | 0.1812 |

### Genotype-Environment Associations

In the PCA of the 8 environmental variables, described in the Section 3.2, the first component explained 34.75% of the variation in the predictors, with the second component accounting for 21.42% of the variance, and the third component accounting for 16.73%, for a total cumulative explained variance of 72.9% (Table 3.8 and Table 3.9). Using the first PC as the predictor in the latent factor mixed model (LFMM), 92 candidate SNPs were identified as being significantly associated with the environment (FDR < 0.1) (Supplemental Table 5).

Table 3.8: PCA results of environmental variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| Eigenvalue | 2.4326 | 1.4994 | 1.1709 | 0.69002 | 0.57216 | 0.457 | 0.17785 |
| Proportion Explained | 0.3475 | 0.2142 | 0.1673 | 0.09857 | 0.08174 | 0.06529 | 0.02541 |
| Cumulative Proportion | 0.3475 | 0.5617 | 0.729 | 0.82757 | 0.90931 | 0.97459 | 1 |

Table 3.9: Correlations between the PC axis and environmental variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| Eastward wind | 0.323 | -0.377 | 0.369 | -0.399 | -0.142 | 0.303 | 0.529 |
| Northward wind | -0.164 | 0.374 | -0.339 | -0.805 | 0.079 | -0.183 | 0.157 |
| Temperature | 0.124 | 0.727 | 0.04 | 0.224 | -0.101 | 0.19 | 0.168 |
| High vegetation | 0.213 | 0.346 | 0.602 | -0.233 | 0.138 | 0.282 | -0.415 |
| Low vegetation | -0.4 | 0.137 | 0.373 | 0.039 | -0.718 | -0.29 | 0.189 |
| Water retention capacity | -0.414 | -0.11 | 0.472 | -0.102 | 0.43 | -0.391 | -0.127 |
| Surface runoff | -0.496 | 0.103 | 0.063 | 0.199 | 0.407 | 0.406 | 0.517 |
| Evaporation | 0.481 | 0.165 | 0.14 | 0.189 | 0.28 | -0.598 | 0.42 |

The redundancy analysis (RDA) model revealed that environmental variables explain 10.42% of the genetic variance (constrained), while 89.58% of the variance remains unexplained. This unexplained variance is presumably influenced by geographical distance, as supported by the results of the IBD/IBE and partial RDA tests. Nevertheless, the environmental contribution to genetic variance offers valuable insights into adaptive processes. Even though the majority of the variance was not explained by environmental factors, both the overall RDA model and the first 4 axes still demonstrated statistical significance, indicating that there were substantive associations within the data, which suggests the presence of meaningful genotype-environment associations (Table 3.10). A total of 822 SNP candidates were detected as significantly associated with at least one of the first four RDA loadings (Supplemental Table 6). Among these, 658 were found within genes. Additionally, out of all the genes identified by RDA containing the 658 SNPs, 32 genes were also found to be significantly associated with the environment in the LFMM.

Table 3.10: Significant test results on each constrained axis of RDA

|  | Df | Variance | Proportion | F | Pr(>F) | Significance |
| --- | --- | --- | --- | --- | --- | --- |
| Full Model | 8 | 1796.3 | 10.42% | 4.5511 | 0.001 | \*\*\* |
| RDA1 | 1 | 894.1 | 5.19% | 18.1217 | 0.001 | \*\*\* |
| RDA2 | 1 | 469.8 | 2.73% | 9.5231 | 0.001 | \*\*\* |
| RDA3 | 1 | 158.7 | 0.92% | 3.2158 | 0.001 | \*\*\* |
| RDA4 | 1 | 69.8 | 0.40% | 1.4154 | 0.001 | \*\*\* |
| RDA5 | 1 | 56.8 | 0.33% | 1.1517 | 0.094 | . |
| RDA6 | 1 | 51.1 | 0.30% | 1.0361 | 0.591 |  |
| RDA7 | 1 | 48.6 | 0.28% | 0.9855 | 0.901 |  |
| RDA8 | 1 | 47.3 | 0.27% | 0.9592 | 0.901 |  |
| Constrained axis Residual | 313 | 15442.7 |  |  |  |  |

Significant Code: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01: ‘\*’ 0.05 “.” 0.1 ‘ ’ 1

Figure 3.8A captures the distribution of mosquito populations across the RDA 1 and RDA 2 axes, which together explained 76% of the environmentally influenced genetic variance (corresponding to 5.84% of the total genetic variance). The clear regional clustering depicted within this biplot aligns with the four geographic regions (Midwest, Northwest, Southwest, and West Coast), underscoring a significant regional influence on the portion of genetic variation shaped by environmental factors. Specifically, the alignment of populations with vectors for the temperature and evaporation signifies the role of temperature and humidity in this context. Figure 3.8B further explores the subtler environmental gradients within the RDA 3 and RDA 4 axes, which together explain approximately 13% of the environmentally responsive genetic variance (equating to roughly 1.56% of the total genetic variance). Here, the distributions suggest a more intricate interaction of genetic variance with environmental variables like the surface runoff and leaf area index for high vegetation, which could reflect micro-environmental adaptations.

A screenshot of a graph

Description automatically generated

Figure 3.8. Environmental Correlates of Genetic Variation in *Cx. tarsalis* Across Diverse North American Regions. Panels A and B display the relationship between environmental factors and the distribution of *Cx. tarsalis*, using RDA to illustrate how regional differences affect genetic variation. In these panels, the position of each circle (representing an individual mosquito) and color (indicating regional groupings from ADMIXTURE results) reflects their association with environmental variables, shown as purple vectors. The first plot (A) focuses on RDA1 and RDA2, the primary axes explaining the most variance, while the second (B) explores more subtle influences in RDA3 and RDA4.

As in Figure 3.9 and Supplemental Table 7, the majority of the candidate SNPs in the RDA1 group are negatively correlated to evaporation, temperature and high vegetation. As shown in Figure 3.8A, the first and fourth quadrants likely represent eastern populations, while the second and third quadrants represent western populations. This east-west gradient is evident among the RDA1 SNPs group. Similarly, the RDA2 group candidate SNPs are more negatively correlated with low vegetation index and water retention capacity but positively correlated with temperature.

This analysis reveals that most significantly associated SNPs are correlated with more than one environmental variable and often cluster together in distinct patterns (Figure 3.9). Although this complex interaction represents a smaller slice of the total genetic variance, it is critical for a comprehensive understanding of genetic-environment relationships.

A colorful lines on a screen

Description automatically generated with medium confidence

Figure 3.9. Correlation Heatmap between Environmental Variables and RDA Candidate SNPs. This heatmap illustrates the correlation between the environmental variables and candidate SNPs identified through the first four constrained axes of Redundancy Analysis (RDA). Each column on the x-axis represents a candidate SNP. For detailed SNP names and associated information on the 17 genes identified by SIFT4G, please refer to Supplemental Figure 3.

This suite of RDA plots collectively reveals the environmental portion of genetic variance within *Cx. tarsalis* populations. While most of the genetic variation correlates with geographic distribution, the environmental variance captured here affords a critical perspective on the selective forces at play, contributing to the broader evolutionary narrative of this species.

Overall, our findings underscore the intricate connections between geographic locations, environmental factors, and genetic variations, confirming that in addition to genetic drift and IBD, there is a profound influence of environmental variables on genetic variation in *Cx. tarsalis*.

### SNPs Under Selection

The BayeScan outlier analysis, performed with a Q-value threshold of 0.05 to identify loci under selection, revealed 1,836 loci potentially under diversifying selection, 10,166 potentially under balancing selection, and 5,237 neutral loci (Figure 3.10 and Supplemental Table 8). Of the 1,836 loci putatively undergoing selection for local adaptation, 1501 of these are found within 824 genes. Of these 824 genes, 24 also overlapped with genes that were identified by both the LFMM and RDA as significantly associated with environmental variables.

A graph showing a bullet

Description automatically generated

Figure 3.10: BayeScan results.

We also used PCAdapt independently to identify outlier loci with notable allele frequency differences across populations, potentially resulting from natural selection. The Q-Q plot in Figure 3.11(left plot) shows that most p-values follow the expected uniform distribution. However, the smallest p-values deviate significantly from expectation, indicating the presence of outliers. Similarly, the histogram of p-values (Figure 3.11 right plot) confirms this pattern, with most p-values aligning with a uniform distribution but an excess of small p-values highlighting the outliers. These outliers include 173 SNPs in the top 1% of extreme p-values. (Supplemental Table 9).

A graph with a line

Description automatically generated A graph of a number of columns

Description automatically generated with medium confidence

Figure 3.11 PCAdapt result. (Left) Q-Q plot to show the expected uniform distribution of the p-values. (Right) Histogram of p-values distribution

In total, 53 candidate local adaptation SNPs were identified (Table 3.13). Some candidates show a strong East-West gradient in terms of allele frequencies (Figure 3.12A and 3.12B), while others indicate that the alternate allele is present only in one or a few populations (Figure 3.12C and 3.12D).

A group of graphs with different colored bars

Description automatically generated

**Figure 3.12. Allele frequency distribution of selected candidate SNPs**

Looking across all four analyses examining significant environmental associations and evidence of natural selection, we identified 20 common genes containing candidate SNPs that were consistently significant in each instance (Figure 3.13 and Supplemental Table 11). Given BayeScan’s susceptibility to potential false positives, we included PCAdapt to provide additional validation and strengthen the robustness of our findings.A Chi-Square test (Table 3.11 and Table 3.12) indicated a significant association between the candidacy of SNPs and their location being within genes rather than intergenic (X-squared = 75.842, df = 1, p-value < 2.2e-16). This suggests that SNPs identified as candidates are more likely to be located within genes than would be expected by chance.

A graph of different sizes and numbers

Description automatically generated with medium confidence

**Figure 3.13: Upset Plot showing the overlap of genes annotated from candidate SNPs identified across four methods for detecting local adaptation and environmental associations in *Cx. tarsalis*: LFMM, RDA, PCAdapt, and BayeScan.** The bar heights indicate the number of genes annotated based on candidate SNPs in each unique or shared category, with percentages relative to the total genes annotated. Black circles below the bars represent intersections of methods, where connected lines indicate the methods contributing to the overlap. For example, the tallest bar corresponds to 438 genes uniquely identified by BayeScan (47.9%), with no overlap with other methods. Individual horizontal bars represent the total number of genes identified by each method: LFMM\_PC1 (44, 4.8%), PCA (110, 12.0%), RDA (459, 50.2%), and BayeScan (824, 90.2%).

Table 3.11: Data summary from LFMM, RDA, bayescan and PCAdapt

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | candidate SNPs | SNPs within genes | SNPs not within genes | matched genes | non-candidate SNPs | non-candidates SNPs within genes | non-candidates SNPs not within genes |
| LFMM | 92 | 73 | 19 | 44 | 17147 | 14863 | 2284 |
| RDA | 822 | 658 | 164 | 459 | 16417 | 14283 | 2134 |
| Bayescan | 1836 | 1501 | 335 | 824 | 15403 | 13440 | 1963 |
| PCAdapt | 173 | 144 | 29 | 110 | 17066 | 14797 | 2269 |
| Total | 2923 | 2376 | 547 | 1437 | 66033 | 57383 | 8650 |

Table 3.12: contingency table for Chi-square test

|  |  |  |
| --- | --- | --- |
|  | Candidates SNPs | Non-candidates SNPs |
| In Genes | 2376 | 57383 |
| Not in Genes | 547 | 8650 |

The SIFT4G predictions determined that 17 of the 20 overlapping candidate genes contained an environmentally-associated SNP located within the coding region, and 8 of these genes contained nonsynonymous mutations within our dataset (Table 3.13).

Table 3.13: SIFT4G results for candidate SNPs within Genes

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| gene\_ID | total\_SNPs | nonsynonymous | | | synonymous | | |
| total | tolerated | deleterious | total | tolerated | deleterious |
| Ct.00g025080 | 3 | 1 | 1 | 0 | 2 | 2 | 0 |
| Ct.00g026900 | 8 | 0 | 0 | 0 | 8 | 8 | 0 |
| Ct.00g030230 | 2 | 1 | 1 | 0 | 1 | 1 | 0 |
| Ct.00g032480 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| Ct.00g049290 | 3 | 2 | 1 | 1 | 1 | 1 | 0 |
| Ct.00g051300 | 3 | 2 | 2 | 0 | 1 | 1 | 0 |
| Ct.00g062900 | 3 | 0 | 0 | 0 | 3 | 3 | 0 |
| Ct.00g064410 | 2 | 0 | 0 | 0 | 2 | 2 | 0 |
| Ct.00g095350 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| Ct.00g154760 | 2 | 0 | 0 | 0 | 2 | 2 | 0 |
| Ct.00g176220 | 6 | 1 | 1 | 0 | 5 | 5 | 0 |
| Ct.00g179740 | 2 | 0 | 0 | 0 | 2 | 2 | 0 |
| Ct.00g237940 | 3 | 0 | 0 | 0 | 3 | 3 | 0 |
| Ct.00g238000 | 4 | 1 | 0 | 1 | 3 | 3 | 0 |
| Ct.00g280270 | 5 | 2 | 2 | 0 | 3 | 3 | 0 |
| Ct.00g280280 | 3 | 2 | 2 | 0 | 1 | 1 | 0 |
| Ct.00g290200 | 2 | 0 | 0 | 0 | 2 | 2 | 0 |
| Total | 53 | 12 | 10 | 2 | 41 | 41 | 0 |

Among these genes, several genes are noteworthy due to their intriguing functions. For instance, Ct.00g030230, identified as a seminal plasma protein, may play a role in reproductive success [(Boes et al. 2014; Amaro et al. 2021)](https://paperpile.com/c/vD89YP/8qRRN+r5eqP). Ct.00g095350, a defective proboscis extension response-related (DPR) protein has been shown to affect feeding behavior in fruit flies [(Nakamura et al. 2002)](https://paperpile.com/c/vD89YP/4hBgg). Ct.00g154760, which was annotated as a carnitine O-acyltransferase, is involved in fatty acid metabolism [(Jogl et al. 2004)](https://paperpile.com/c/vD89YP/pc9gu), a pathway crucial for energy provision under varying climatic conditions. Lastly, Ct.00g04290 encodes a PERIOD CIRCADIAN PROTEIN, which is essential for the regulation of circadian rhythms, aligning life cycle events with environmental cues [(Meuti et al. 2015; Chang and Meuti 2020; Shetty et al. 2022)](https://paperpile.com/c/vD89YP/Mabcd+vJyLU+8MHFo). Notably, within this gene, two nonsynonymous and one synonymous SNPs were discerned. SIFT predictions indicated that one of the nonsynonymous SNPs could be deleterious, potentially impacting the protein's functionality and, by extension, the organism's adaptability to environmental rhythmic changes. This was the one of the two candidate SNPs classified as deleterious by SIFT, so it represents a significant point of interest for its potential role in the ecological adaptation of *Cx. tarsalis*.

## Discussion and Conclusion

We successfully assembled a draft genome for *Cx. tarsalis* and identified hundreds of thousands of polymorphic markers across individuals sampled from diverse locations in North America, primarily in four regions: the West Coast, Southwest, Northwest, and Midwest. Each region is characterized by unique climatic conditions that likely influence genetic variation. The West Coast, known for its warm, seasonally dry climate, contrasts quite significantly with the hot, dry conditions and sparse vegetation of the Southwest. In comparison, the Northwest features a cool, wet climate with lush high vegetation, while the Midwest has a cool, dry climate characterized by a mix of both high and low vegetation. These data revealed a distinct pattern of population structure within this species, with clear differentiation among populations from these regions. Additionally, there was a strong correlation between geographic and genetic distances, highlighting the roles of genetic drift and selection in shaping genetic variation across these varied landscapes.

Our analysis also uncovered a significant link between environmental variables and genetic variation, particularly showing that evaporation rates, and vegetation density, temperature and wind are critical environmental factors with strong associations to genetic differentiation within *Cx. tarsalis* populations. Among these, the identification of 53 SNPs with strong evidence of both selection and environmental correlation across multiple tests suggests that these are most likely to be involved in local adaptation processes. These SNPs, linked to crucial biological functions such as circadian rhythms, reproductive success, feeding habits and fat metabolism, and, lay the groundwork for a detailed exploration of the genetic mechanisms driving adaptation in diverse environmental conditions.

Probably the most interesting SNP uncovered in our study was a deleterious mutation in a gene that we found to be the single copy ortholog of per, a well-studied circadian rhythm gene that encodes the regulatory period protein[(Konopka and Benzer 1971)](https://paperpile.com/c/vD89YP/BC1dN). Circadian clock genes are critical for synchronizing physiological and behavioral processes in insects as well as nearly all other living organisms, so a mutation in this gene could be key to the mechanisms that have allowed *Cx. tarsalis* to expand both northward and eastward across the North American continent. Research in other vector species have found that variations in circadian clock genes have a profound impact on mosquito behavior and fitness in both *Cx. pipiens* [(Meuti et al. 2015; Chang and Meuti 2020)](https://paperpile.com/c/vD89YP/Mabcd+vJyLU) and the more distantly related Aedes aegypti [(Shetty et al. 2022)](https://paperpile.com/c/vD89YP/8MHFo), and in *Cx. pipiens* it was found that circadian regulators (including per) were necessary for inducing diapause.

While it has been observed that at least some *Cx. tarsalis* populations must be capable of entering diapause in order to navigate the challenges of seasonal extremes and to traverse significant geographical barriers, such as the Rocky Mountains, the genetic mechanisms underlying this trait have not yet been identified in *Cx. tarsalis*. Our discovery of a candidate allele in the per ortholog represents the first identification of a potential genetic mechanism governing diapause in this species. The elucidation of circadian protein-involved feedback loops, as explored by Shetty et al., further emphasizes the potential universality of these mechanisms across mosquito species. This breakthrough underscores this gene's central role in broader adaptive strategies crucial for ecological adaptation in insect populations.

In addition to the circadian regulator gene, we also identified a handful of other highly significant nonsynonymous mutations in other genes related to traits that are known to be key for insect survival during seasonal changes. For example, Ct.00g030230 was identified as encoding a seminal plasma protein, a family of proteins which are known to influence fertility and post-mating behavior in mosquitoes [(Boes et al. 2014; Amaro et al. 2021)](https://paperpile.com/c/vD89YP/8qRRN+r5eqP). Ct.00g095350 was homologous to the DPR gene in Drosophila, where it has been shown to have a role in regulating feeding behavior [(Nakamura et al. 2002)](https://paperpile.com/c/vD89YP/4hBgg). In fruit flies, the DPR gene is involved in the gustatory (taste) response, particularly in the aversion to salt. In mosquitoes, feeding behavior is critical for both nutrient intake and disease transmission, so this gene may be crucial for not only understanding the ecology of *Cx. tarsalis*, but also its disease transmission dynamics. Finally, we also found a significant mutation in the gene Ct.00g154760, which encodes a carnitine O-acyltransferase. Carnitine O-acyltransferase facilitates the transport of fatty acids into mitochondria for β-oxidation, a process integral to energy production [(Jogl et al. 2004)](https://paperpile.com/c/vD89YP/pc9gu). Adaptations in Ct.00g154760 that enhance the enzyme's efficiency could provide *Cx. tarsalis* with survival advantages by optimizing energy utilization under fluctuating conditions, and could also be related to diapause behaviors, since insects must accumulate fat reserves prior to entering into the diapause state[(Denlinger 2002)](https://paperpile.com/c/vD89YP/aWUBU).

Overall, our investigation into the *Cx. tarsalis* genome has highlighted the critical importance of environmental adaptations for understanding this mosquito's distribution and adaptation to a wide range of habitats across North America. Identifying genetic markers linked to circadian rhythms, reproductive processes, and metabolic functions reveals how selection on a few key genetic variants may have sculpted the species' ability to adapt to an array of environmental challenges and expand its range both northward and eastward. Such genetic insights are pivotal as climate change continues to reshape habitats, potentially enabling vector species like *Cx. tarsalis* to spread into new areas and present new health challenges in the near future.

# West Nile Virus Prediction

The prediction of West Nile Virus (WNV) transmission is a critical component in public health strategies for disease prevention and control. Given its impact on human health and the complexity of its transmission dynamics, there is a pressing need for accurate predictive models. These models can inform timely interventions and guide resource allocation for vector control. My research aims to develop a comprehensive predictive model integrating environmental, demographic, and possibly genetic data, to enhance our understanding and forecasting of WNV outbreaks.

## Data source

### Clinical Disease Surveillance Data

This study utilized data from three primary sources to analyze West Nile Virus (WNV) cases and related metrics. The first source is ArboNET, a national arboviral surveillance system managed by the CDC in collaboration with state health departments. ArboNET provides comprehensive data on arboviral infections, including yearly counts of WNV cases in humans at the county level from 1999 to 2023. It also includes monthly counts of WNV-positive horses, dead birds (categorized as corvid and non-corvid), and mosquitoes for each county from 2000 to 2021.

The second data source is from the California Department of Public Health (CDPH), consisting of two datasets spanning the period from 2004 to 2023. These datasets provide weekly/monthly reports of human WNV cases across counties in California, offering high-resolution temporal and spatial coverage. The second data source were coming from two different ways. The first way is officially requested from Vector-Borne Disease Section of CDPH and the second way is scraped from CDPH’s weekly reports ([westnile.ca.gov/resources\_reports?report\_category\_id=6](https://westnile.ca.gov/resources_reports?report_category_id=6))

* Official CDPH Data:

This data was officially requested from the CDPH Vector-Borne Disease Section. It includes records specifically for 13 counties: Fresno, Kern, Los Angeles, Merced, Orange, Placer, Riverside, Sacramento, San Bernardino, San Joaquin, Solano, Stanislaus, and Tulare. Figure 2221 shows the annual geographic distribution of human WNV cases in California from 2004 to 2023.

A screenshot of a map

AI-generated content may be incorrect.

Figure 2221: Annual Geographic Distribution of Human West Nile Virus Cases in California (2004–2023)

* Scraped Data from CDPH Website:

This dataset was independently compiled by scraping more than 760 PDF files of weekly reports available on the CDPH website (<https://westnile.ca.gov/resources_reports?report_category_id=6>). It contains data for all counties in California and is entirely separate from the official CDPH data.

Figure 1112 shows the distribution of absolute differences in human WNV disease case counts between two different sources: the official CDPH dataset and the independently scraped dataset. The x-axis represents the magnitude of the case differences, while the y-axis indicates the frequency of these differences. This visualization helps identify the extent and frequency of discrepancies between the two data sources across counties, years, and months. The histogram shows that most case differences between the two datasets (CDPH and indecently scraped dataset) are close to zero, indicating that the datasets are largely consistent. However, a small number of large differences are evident, particularly in counties and months critical for WNV transmission. Table 1111 shows the top 20 largest case difference, highlighting the specific counties, years, and months where these differences occur. The result shows that Los Angeles County dominates the top differences, with multiple instances of significant differences, especially during late summer and fall months (August to October) in years like 2014, 2015, and 2017. Notably, the largest single difference is 97 cases in Los Angeles County in October 2017, followed by 94 cases in Orange County in August 2014. These outliers suggest potential differences in data collection or reporting methodologies, particularly during peak transmission periods. While the datasets are largely aligned, addressing these discrepancies through further investigation could enhance data reliability and ensure more accurate modeling and surveillance of WNV trends.

A graph with numbers and a bar

AI-generated content may be incorrect.

Figure 1112 Histogram plot of WNV cases different between CDPH and independently scraped data

Table 1111: Top 20 WNV case differences between CDPH and independently scraped data

|  |  |  |  |
| --- | --- | --- | --- |
| County | Year | Month | Case Absolute Difference |
| Los Angeles | 2017 | 10 | 97 |
| Orange | 2014 | 8 | 94 |
| Los Angeles | 2015 | 9 | 89 |
| Los Angeles | 2014 | 10 | 85 |
| Los Angeles | 2014 | 9 | 77 |
| Orange | 2014 | 10 | 76 |
| Los Angeles | 2014 | 8 | 63 |
| Los Angeles | 2017 | 8 | 60 |
| Los Angeles | 2004 | 8 | 58 |
| Riverside | 2015 | 10 | 58 |
| Los Angeles | 2015 | 11 | 57 |
| Riverside | 2015 | 8 | 48 |
| Los Angeles | 2017 | 9 | 47 |
| Los Angeles | 2023 | 8 | 46 |
| Los Angeles | 2015 | 12 | 45 |
| Los Angeles | 2016 | 8 | 45 |
| Los Angeles | 2014 | 11 | 44 |
| San Bernardino | 2004 | 7 | 42 |
| Los Angeles | 2013 | 10 | 40 |
| Stanislaus | 2023 | 8 | 37 |

The third source is from the Illinois Department of Public Health (IDPH), which includes weekly reports of WNV-positive dead birds, mosquitoes, and horses. This dataset focuses on non-human infection metrics in Illinois.

The granularity of the clinical disease surveillance data, including its geographic and temporal resolutions, is summarized in Table 2111.

Table 2111: Surveillance Data Granularity

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data source | Temporal level | Spatial level | Time period | Objects |
| CDC-ArboNET | Yearly | State level |  | Human, mosquito, bird, horse |
| CDPH | Weekly | County Level | 2004-2023 | Human |
| IDPH | Monthly | County Level | 2002-2022 | Mosquitoe, birds and horses |

### Climate Data

Same as in the *Cx. tarsalis* project environmental data section, utilizing ERA5-Land monthly averaged data (1950-present) from the Copernicus Climate Data Store [[27]](https://paperpile.com/c/qo3oGi/GyX6o) to extract annual average and monthly average climate data. Climate data representation for each county is based on county seat coordinates.

The influence of daylight length on West Nile Virus (WNV) dynamics was explored due to its potential role in regulating the biology of *Cx. tarsalis*, the primary vector of WNV. Insights from a population genetics study on *Cx. tarsalis* identified genes associated with circadian rhythm adaptation. Circadian rhythms are closely tied to photoperiods, such as daylight length, and play a critical role in regulating diapause—a physiological state that affects mosquito activity and seasonal disease transmission.

Average monthly daylight length (in hours) was calculated for each record in the dataset using geographic coordinates (latitude and longitude) and temporal data (year and month). These calculations provide an environmental variable to assess its relationship with WNV case dynamics and mosquito behavior.

### Land Cover Data

The land use information is derived from the global 1-km Consensus Land Cover dataset [28], an invaluable resource offering twelve distinct data layers. Each layer represents the prevalence of a specific land-cover type, with values ranging from 0 to 100, denoting the percentage of coverage within a given area. The dataset encompasses a diverse range of land-cover classes, including evergreen and deciduous needleleaf trees, evergreen broadleaf trees, deciduous broadleaf trees, mixed and other tree types, shrubs, herbaceous vegetation, cultivated and managed vegetation, regularly flooded vegetation, urban and built-up areas, snow and ice, barren landscapes, and open water. Spanning latitudes from 90ºN to 56ºS and longitudes from 180ºW to 180ºE, it provides comprehensive global coverage at a resolution of approximately 1 km per pixel at the equator [29], offering a remarkable perspective on land-cover patterns worldwide. Land cover data representation for each county is based on county seat coordinates.

### Birds Data

The eBird dataset, managed by the Cornell Lab of Ornithology, is one of the world’s largest biodiversity-related citizen science initiatives [30]. It aggregates bird observation data contributed by a global community of birders. For this study, eBird data includes observations of common North American bird species, aggregated by species, county, month, and year, along with avian phylodiversity metrics based on 2019 observations. The dataset provides valuable insights into bird distribution, abundance, and trends, with rigorous data quality checks overseen by regional experts. [john’s paper citing here]

### Mosquitoes’ abundance data

The mosquito abundance data, covering all species from 2004 to 2023, was obtained from the California Vectorborne Disease Surveillance System (CalSurv). It includes monthly data for 13 counties (Fresno, Kern, Los Angeles, Merced, Orange, Placer, Riverside, Sacramento, San Bernardino, San Joaquin, Solano, Stanislaus, and Tulare). The data was collected using two trap types: carbon dioxide-baited traps and gravid mosquito traps.

Access to this data is governed by the California Vectorborne Disease Surveillance Data Policy. Researchers or individuals interested in using this data must adhere to the policy outlined in the official document, which can be accessed at: <https://vectorsurv.org/assets/files/calsurv_data_policy.pdf>.

The data display clear gaps in mosquito collections during the early years in some counties, which could make using this data for prediction challenging (See in figure 4444). These gaps may introduce biases or reduce the reliability of predictive models, particularly when attempting to analyze temporal trends or make comparisons across counties. Careful handling of missing data, such as imputation or the exclusion of affected time periods, is necessary to ensure robust and accurate forecasting results.

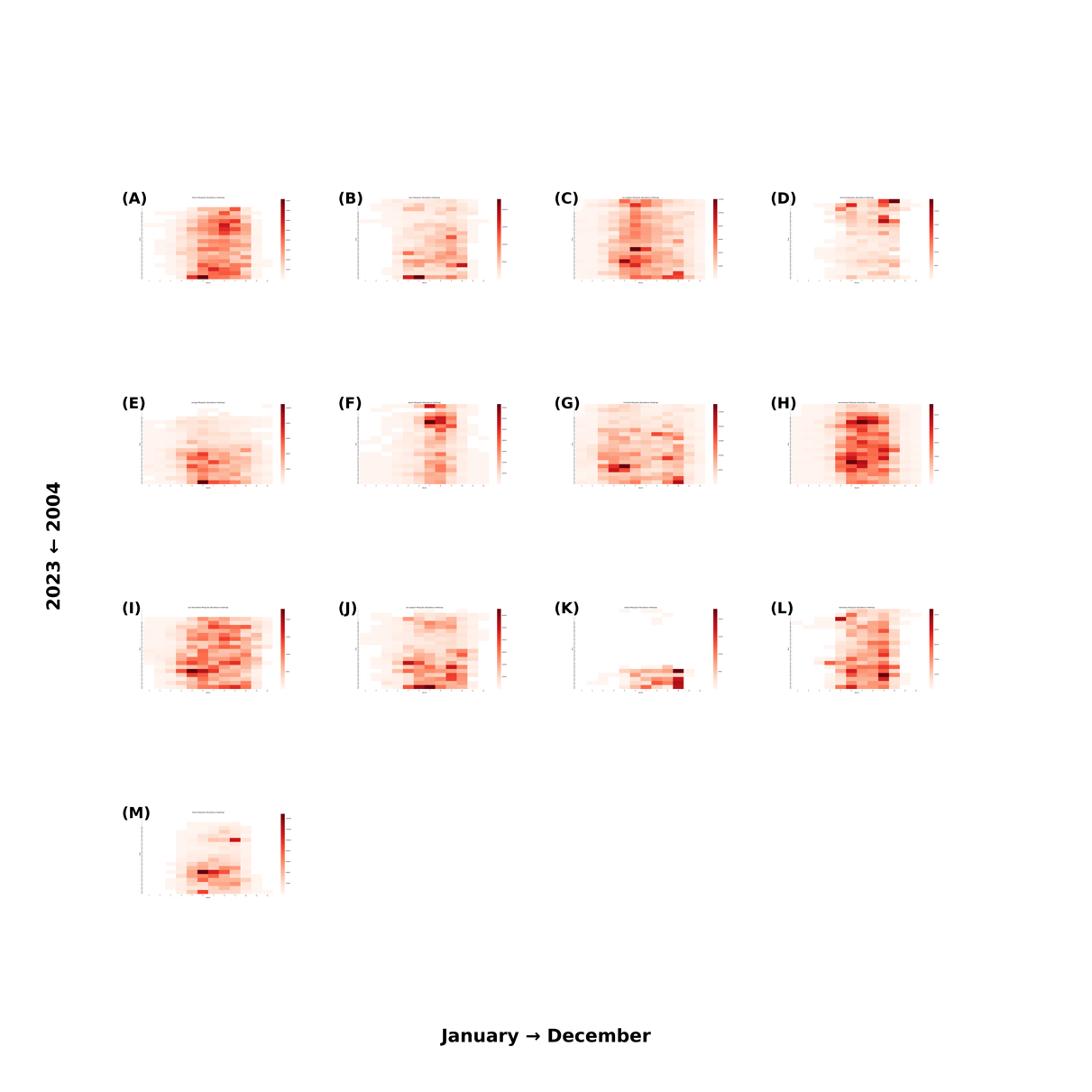


Figure 4444: Heatmap of Monthly Mosquito Abundance (2004–2023) Across 13 California Counties, by Month

### Demographic Data

Compiled from various sources including the U.S. Census Bureau (population, poverty estimates, and land area data) and Federal Communication Commission (county data). The focus is on data from 2000 to 2021, with manual adjustments for changes such as county name alterations.

## Methodology

### Data Preprocessing

#### Integration of multi-sources data

Harmonizing diverse datasets for machine learning is challenging, especially with the inconsistencies in county identifiers. I use FIPS codes as standard identifiers, yet not all sources include these, and county names often differ between datasets. This requires a thorough manual examination to ensure accurate FIPS code allocation, a task complicated by occasional county renames and FIPS updates. Meticulous effort is essential for successful data integration.

#### Climate data processing

For each data row representing a specific geographic location and year within our study, I employed the xarray library [[31]](https://paperpile.com/c/qo3oGi/h1QCg) in Python to process the ERA5-Land monthly averaged climatic data, which covers the period from 1950 to the present. From this dataset, I extracted the maximum, minimum, and mean values for each climate variable on an annual basis. This approach ensures that I accurately represent the climatic variability and trends for each location over the course of a year, providing a robust foundation for analyzing the influence of climate on West Nile Virus transmission dynamics.

For the monthly report dataset, I extracted the monthly average values for each climate variable within the monthly on the corresponding county. To better understand the past climate variables effect. The average monthly data of the previous one month, two months is also extracted for each location and climate variable.

#### Integration of land cover data

Additionally, I will incorporate land cover variables from the global 1-km Consensus Land Cover dataset [[29]](https://paperpile.com/c/qo3oGi/OvygF), employing the OpenCV library [[32]](https://paperpile.com/c/qo3oGi/arZbM) and xarray library [[31]](https://paperpile.com/c/qo3oGi/h1QCg) in Python to extract consensus land cover information pertinent to the county seat coordinates.

#### Standardization of data granularity

Annual Based Dataset

In the context of the data processing for my study, the human West Nile Virus (WNV) data is aggregated on an annual basis at the county level, whereas the non-human WNV data is reported with greater frequency, on a weekly basis, also by county. To harmonize the granularity for comparative and predictive analysis, I standardized all data to an annual county-level scale. This alignment allows for a consistent approach in analyzing the spread and impact of WNV across human and non-human populations, ensuring that the annual cycle of the disease's prevalence is accurately captured, and that the data is comparable across the different categories of subjects affected.

In preparing the dataset for predicting WNV cases, a series of data manipulation and imputation tasks are conducted to ensure data integrity and usability for predictive modeling. Initially, I filtered human infection records from 2004 to 2023 and extracted essential county information, removing duplicates and entries with missing FIPS or State data. Subsequently, I constructed a comprehensive data frame reflecting each unique FIPS code's yearly data, augmented with corresponding state, county, and geolocation details. To address the challenge of missing data, I employed a two-fold imputation strategy: setting case counts to zero in the absence of recorded activity and replacing missing human case reports with the average number of cases over the 20-year span for each county.

Monthly Based Dataset

For the monthly human case reports in California, the dataset was standardized in a manner similar to the annual dataset. A comprehensive data frame was constructed to reflect each unique Federal Information Processing Standards (FIPS) code and month in California from 2004 to 2023.

To address missing data, I contacted an epidemiologist at the California Department of Public Health (CDPH). Based on their confirmation, any missing records were treated as an absence of cases, and the missing entries were filled with zeroes. This approach ensures that the dataset accurately represents the absence of reported cases rather than incomplete data collection.

This standardization process provides a consistent temporal and spatial framework for analysis, minimizing bias from gaps in reporting and enhancing the reliability of subsequent modeling efforts.

### Correlation Analysis using the Mantel Test

Before delving into machine learning, understanding the relationships between environmental variables and the target variable is essential. For this purpose, I'll utilize Kendall's tau correlation, a robust non-parametric test developed by Maurice Kendall in 1938 [[33]](https://paperpile.com/c/qo3oGi/em6eD). Unlike parametric tests, Kendall's tau doesn't assume a specific data distribution, making it ideal for analyzing ecological and environmental datasets.

### Machine Learning / Statistical Models

The study employs various models to predict West Nile virus occurrences, considering the complex interplay between environmental, clinical, and land cover factors. Each model is selected based on its strengths in capturing non-linear relationships, temporal patterns and incorporating spatial dependencies. Criteria for model selection include their ability to handle high-dimensional data, robustness against missing information, and capacity to capture spatial and temporal dynamics intrinsic to disease spread. This selection process, informed by a literature review [[8–10,34–36]](https://paperpile.com/c/qo3oGi/x43cb+s8TOW+v5wwI+10xGH+vyyb4+BcWvD), underscores the value of utilizing a combination of traditional machine learning techniques and advanced models to deepen our understanding of disease determinants and improve predictive accuracy.

Table 2 summarizes key aspects of each model used in this study, including their descriptions, rationales for selection, as well as pros and cons. Understanding the strengths and limitations of each model is crucial for selecting the most appropriate approach based on the specific requirements and characteristics of the dataset and prediction task.

Table 2: Comparative Overview of Predictive Models for West Nile Virus Prediction

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Description | Rationale for Selection | Pros | Cons |
| Random Forest | Ensemble learning, multiple decision trees | Robustness, handles non-linear data and missing data | Robust, handles missing values | Complex, reduced interpretability |
| SVM | Finds hyperplane separating classes in feature space | High-dimensional effectiveness | Performs well in high-dimensional spaces | Complex parameter tuning |
| Neural Network | Computing system for pattern recognition | Learns complex relationships | Exceptional pattern recognition | Proneness to overfitting |
| Ensemble Model | Combines SVM/RF/HGBR with Autoregression | Leverages both strengths | Integrates spatial and temporal analysis | Constraints on temporal data structure |
| SGLMMs | Extend GLMs with spatially correlated random effects | Models’ spatial correlation | Accurate spatial dependencies modeling | Complex interpretation |
| HGBR |  |  |  |  |

### Hyperparameter Tuning

The hyperparameter tuning for the models was implemented using “hyperopt” python package with the Tree-structured Parzen Estimator (TPE) algorithm, optimizing for the negative Q² score on a validation set. The best hyperparameters were selected by iterating over 100 evaluations.

For the Support Vector Machine (SVM) model, the tuning focused on the regularization parameter (C) ranging from 0.1 to 10, the epsilon value (0.1 to 5), the kernel type (chosen among *“rbf”*, *“linear”*, *“poly”*, and *“sigmoid”*), and the gamma parameter (either *“scale”* or *“auto”*).

For the Random Forest (RF) model, the tuning involved a broader set of hyperparameters. The number of estimators was optimized between 100 and 1000 (in increments of 10), while the maximum tree depth varied from 1 to 20. The minimum samples required for a split and a leaf were tuned between 2 to 10 and 1 to 10, respectively. The maximum features parameter, which controls the number of features considered for splits, was chosen from both continuous values (0.1 to 1.0) and categorical options (“sqrt”, “log2”). Additionally, max leaf nodes (10 to 100), max samples (0.1 to 1.0), and min impurity decrease (0 to 0.1) were optimized.

For the Histogram-Based Gradient Boosting Regression (HGBR) model, hyperparameter tuning included the maximum depth (1 to 30), number of iterations (100 to 1000, in steps of 100), learning rate (0.01 to 0.5), and L2 regularization strength (0.0 to 1.0). The max leaf nodes (10 to 100, in steps of 10), minimum samples per leaf (1 to 10), and maximum number of bins (10 to 255, in steps of 5) were also optimized. Additionally, the scoring metric was selected from *“loss”*, *“neg\_mean\_squared\_error”*, and *“neg\_mean\_absolute\_error”*.

These tuning procedures ensured that each model leveraged the best possible parameter configurations for predicting West Nile Virus (WNV) cases, improving forecasting accuracy and robustness.

### Evaluation Metrics

To comprehensively assess our model's predictive capability for West Nile virus cases, I apply a few evaluation metrics on the testing dataset, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Squared Logarithmic Error (MSLE), predictive coefficient of determination (Q²) and bootstrapping technique to calculate the 95% confidence interval. Each metric offers a unique lens through which to view model performance.

Besides the evaluation metrics I mentioned above, I also add a baseline model to compare with each model which is a model that will always predict the result with the mean of the disease count.

In addition to evaluate the importance of variables and understand their contributions to the predictive performance of the machine learning models, we used SHAP (SHapley Additive exPlanations) values. SHAP is a widely adopted approach for interpreting complex models by quantifying the impact of each feature on model predictions. Derived from cooperative game theory, SHAP values provide an equitable and consistent measure of feature importance.

SHAP values allow for both local and global interpretability. Locally, they explain individual predictions by attributing the output to input features, making it clear which variables contribute positively or negatively. Globally, SHAP values are aggregated across the dataset to highlight the overall importance of each variable, offering insights into which features most strongly influence the model’s behavior.

The application of SHAP values in this study has several benefits:

* Enhanced Interpretability: By attributing predictions to specific features, SHAP values clarify the model’s decision-making process.
* Feature Selection: Variables with low SHAP importance can be identified and excluded, potentially improving model efficiency and reducing overfitting.
* Interaction Insights: SHAP values reveal interactions between features by showing how the presence of one variable affects the contribution of another.

In this study, SHAP values were computed for all models to ensure an in-depth understanding of variable importance. This not only supports model evaluation but also aligns the results with biological and ecological domain knowledge, ensuring that the models are both predictive and interpretable.

## Results

### National CDC Dataset

Figure 111 shows the total number of human WNV disease cases annually from the CDC national dataset. The plot shows significant year-to-year variation in WNV neuroinvasive disease cases from 2004 to 2023. The highest peak occurred in 2012, suggesting a major outbreak that year, while smaller peaks are visible in 2006 and 2021. Dips in cases are notable in 2008, 2010, and 2020, with the latter possibly linked to pandemic-related factors. Overall, the data reflects fluctuating case counts without a clear long-term trend.

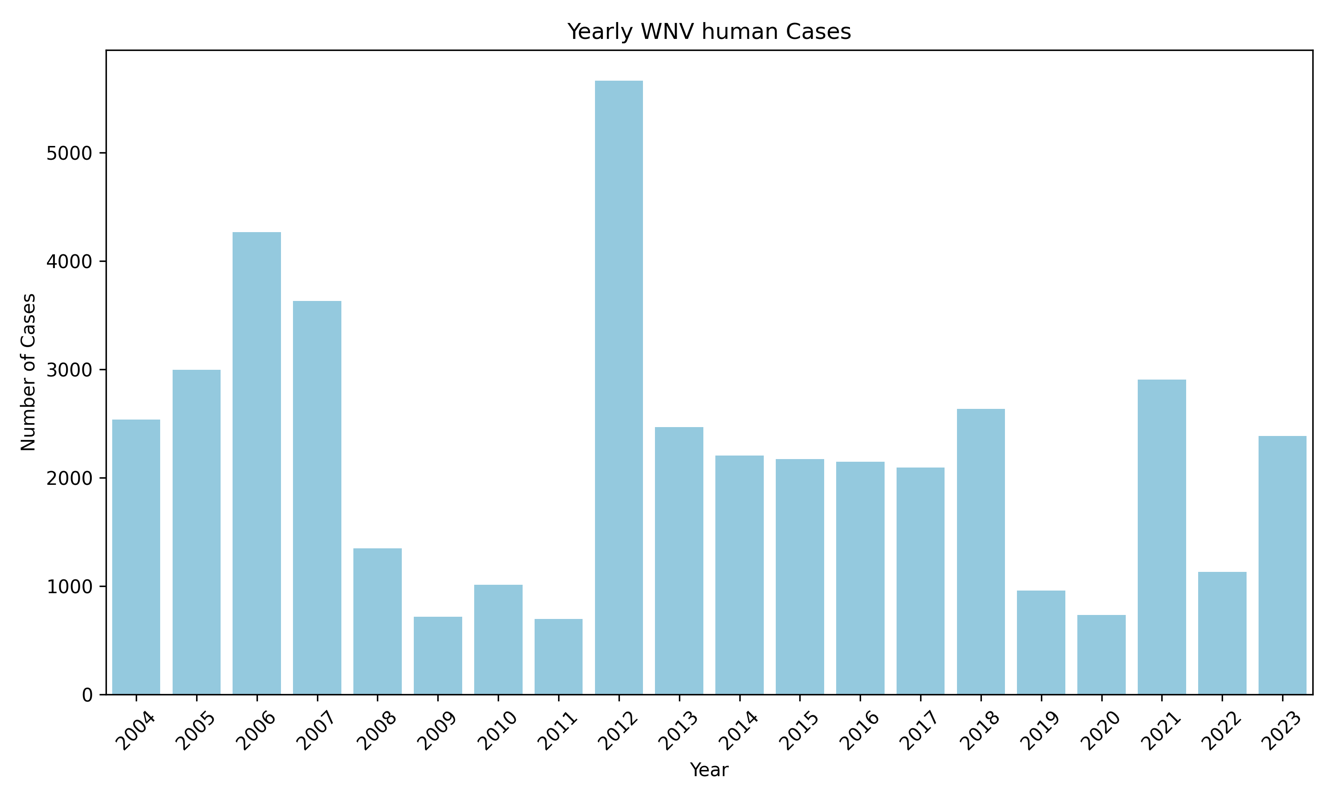


Figure 111: National Annual Total Number of West Nile Virus (WNV) Human Cases (2004–2023)

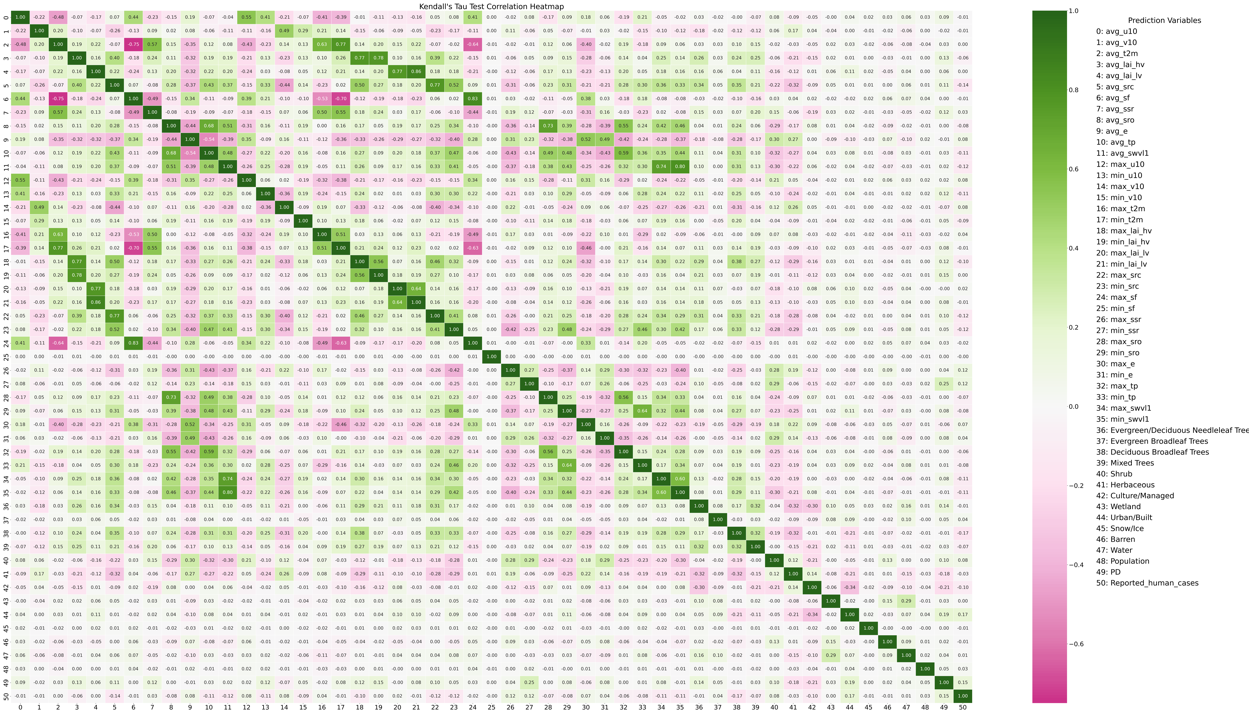
Rolling statistics were calculated for West Nile Virus (WNV) human disease counts using a 3-month window to capture short-term trends and variability in case numbers. This approach involved computing the rolling sum, mean, and standard deviation of cases over consecutive 3-month periods, providing insights into temporal patterns, such as seasonal fluctuations and periods of increased or decreased case activity. These rolling statistics help smooth out noise in the data while highlighting meaningful changes in case trends over time. The results visualized in the Figure 2222, reveal seasonal peaks and variations in WNV cases over the 2004–2023 period, providing a clearer understanding of the temporal dynamics of WNV outbreaks. This approach highlights both the magnitude of outbreaks and the consistency of their timing, helping to visually see the patterns in disease occurrence.

A graph of statistics with numbers

AI-generated content may be incorrect.

Figure 2222: Rolling statistics for number of WNV cases in California

Figure 9 features Kendall's tau correlation heatmaps, analyzing the national CDC dataset on annual human West Nile Virus (WNV) cases in relation to environmental and demographic factors. The heatmaps reveal significant correlations with variables such as minimum surface net solar radiation (min\_ssr), as well as land cover classifications—specifically, Cultivated/Managed and Urban/Built zones, and avian phylogenetic diversity (PD). However, most of the aggregated environmental variables exhibit minimal correlation with annual WNV case counts. This suggests that macro-level environmental metrics may fail to adequately reflect the intricate dynamics that drive disease transmission.



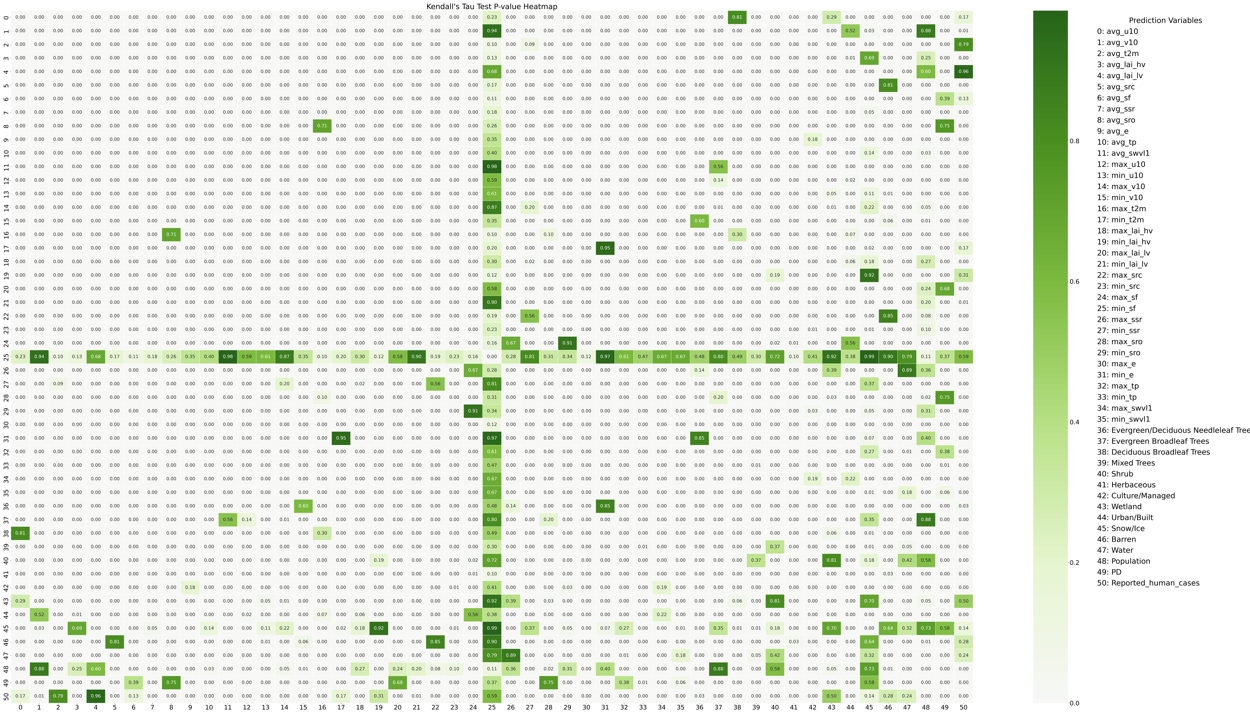


Figure 9: Kendall's Tau correlation (Upper) and p-values (Bottom) Heatmap for National CDC Dataset: Annual Human Cases vs. Environmental Variables

The initial approach aimed to predict nationwide human WNV cases utilizing several models. The outcomes, as illustrated in Table 3, suggest that most models perform comparably, implying that predictive capabilities are generally modest. Among these, the SVM model emerges as the most effective, with an MSE of 192 and a Q2 score of 0.09, indicating that while performance is limited, SVM may offer marginal predictive advantages.

Table 3: Performance of Different Predict Models on National Model

|  |  |  |
| --- | --- | --- |
| Model | MSE | Q2 |
| Linear Regression | 199 | 0.06 |
| Random Forest \* | 202.11 | 0.04 |
| SVM | 192 | 0.09 |
| Neural Network | 202.10 | 0.04 |
| SGLMM \*\* | 92.65 | 0.015 |
| HGBR | 198 | 0.06 |

\* The MSE and Q2 for RF is average results of 100 runs

\*\* Simplify the variables only use avg\_src + min\_ssr + Urban\_Built + PD

While tuning the prediction models for West Nile Virus (WNV) incidence, several significant data-related challenges have emerged:

**Inconsistency in Data Collection**: The collection of WNV case reports varies considerably across states, presenting substantial hurdles for modeling efforts at a national level. This inconsistency can lead to skewed data representations, which may compromise the accuracy and reliability of predictive analytics.

**Significant Class Imbalance**: The imputation of missing values exacerbates the existing imbalance between reports of zero cases and those documenting one or more instances of WNV. Such disproportionality poses a risk of biasing the model towards predicting the majority class (0 disease case) and necessitates the adoption of specialized techniques to achieve a balanced representation of case occurrences.

**Data Granularity**: An additional complexity encountered in predicting annual WNV cases is the granularity of environmental data. When forecasting yearly disease counts, environmental variables must be aggregated—typically as annual averages, maximums, or minimums. This approach, while necessary for aligning temporal scales, may obscure finer temporal correlations between environmental conditions and WNV transmission patterns. For instance, the use of yearly averages or extremes of temperature does not capture the subtleties of seasonal fluctuations that could be critical in understanding and predicting WNV outbreaks. The nuances of how and when these environmental factors influence mosquito populations and virus transmission may be lost, potentially diminishing the predictive power of our models.

Each of these issues requires tailored methodological adjustments to ensure the robustness of the predictive models. The following sections detail the specific approaches undertaken to mitigate these challenges and discuss the results yielded by these refined modeling strategies.

### Inconsistency in Data Collection

Given the variability in data collection and reporting of West Nile Virus (WNV) cases among states, we adopted a state-centric modeling approach. Figure 10 presents the outcomes of employing Support Vector Machine (SVM) algorithms to predict West Nile Neuroinvasive Disease (WNND) occurrences for each state. The displayed models, limited to those achieving a positive Q2 score, indicate a disparity in performance across states. For instance, California, Illinois, Michigan, and New York demonstrate superior model efficacy compared to those of Pennsylvania, North Dakota, and Washington, suggesting that the robustness of local WNV surveillance systems may significantly influence predictive accuracy.

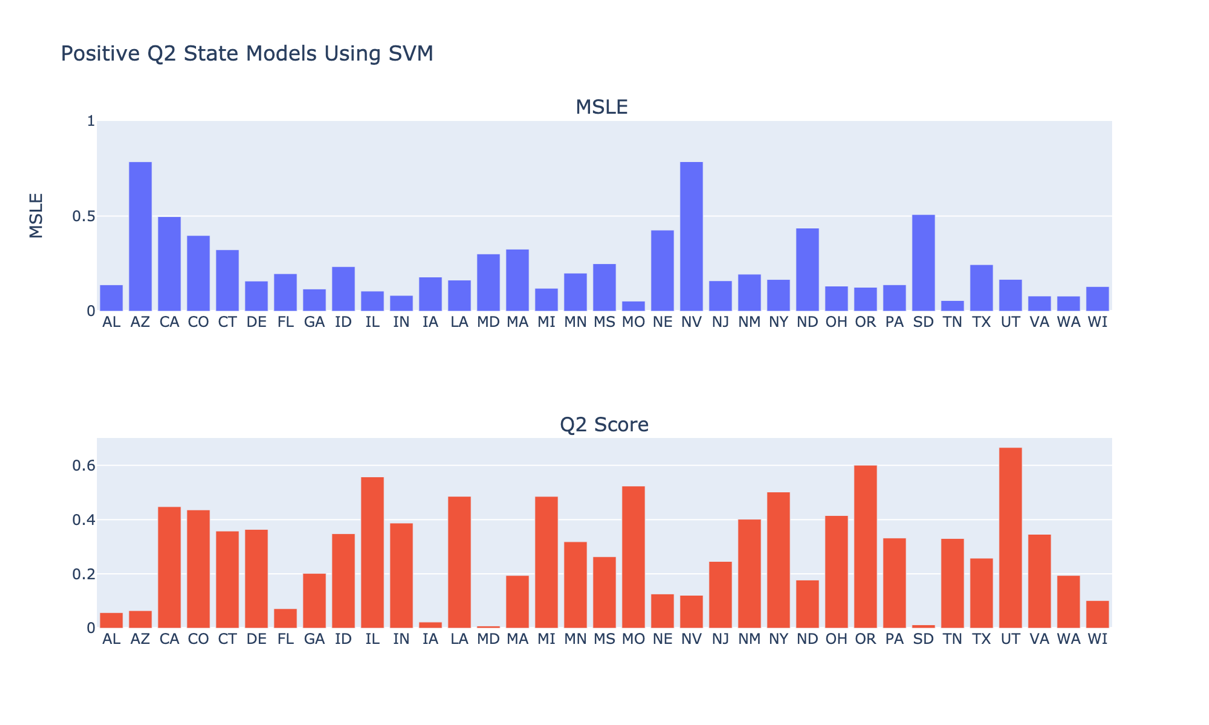


Figure 10: Comparative Analysis of SVM Models for WNV cases Across States

As part of an integrated strategy, I shifted my analytical focus from the quantitative assessment of West Nile Virus (WNV) incidences to a binary determination of WNV presence within individual counties. This pivot required reconceptualizing the issue from a regression framework to a classification paradigm. The newly adopted classification approach was applied to data from New York, California, and Illinois, which were specifically chosen due to their robust WNV case reporting systems that include both human and non-human instances. The efficacy of the SVM-based classification models is detailed in Table 4, where they are evaluated against a national model. These comparisons shed light on the models' capacity to accurately predict binary outcomes indicative of WNV presence.

To enhance model performance, I explored a range of probability thresholds from 0 to 1 to distinguish between zero cases (0) and non-zero cases (non-0). The optimal threshold was selected based on its ability to maximize the F1 score, a harmonic mean of precision and recall that balances the trade-off between false positives and false negatives. Among the state models, California's demonstrated the most robust performance. Conversely, the national model's high accuracy, coupled with a lower F1 score, highlights the persistent challenge posed by class imbalance within the dataset.

Table 4: SVM Predictive Outcomes for WNV Surveillance in State models,

New York, California, Illinois, Colorado and National Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| State | Threshold | F1 | Accuracy | AUC-ROC |
| New York | 0.58 | 0.694 | 0.846 | 0.82 |
| California | 0.85 | 0.860 | 0.826 | 0.87 |
| Illinois | 0 | 0.552 | 0.381 | 0.60 |
| Colorado | 0.33 | 0.743 | 0.842 | 0.85 |
| National | 0.33 | 0.499 | 0.764 | 0.74 |

### Significant Class Imbalance

Figure 11 displays the distribution of West Nile Virus Neuroinvasive Disease (WNND) counts nationally and by state. The data reveal a predominant number of zero-case instances compared to non-zero cases. Such an imbalance skews prediction models towards forecasting zero cases, resulting in overly conservative estimates. To counteract this, a subsampling technique was applied to down sample the majority class (zero cases) to align with the minority class (non-zero cases), aiming for a balanced dataset that would enable more accurate predictions.

A graph of a number of numbers

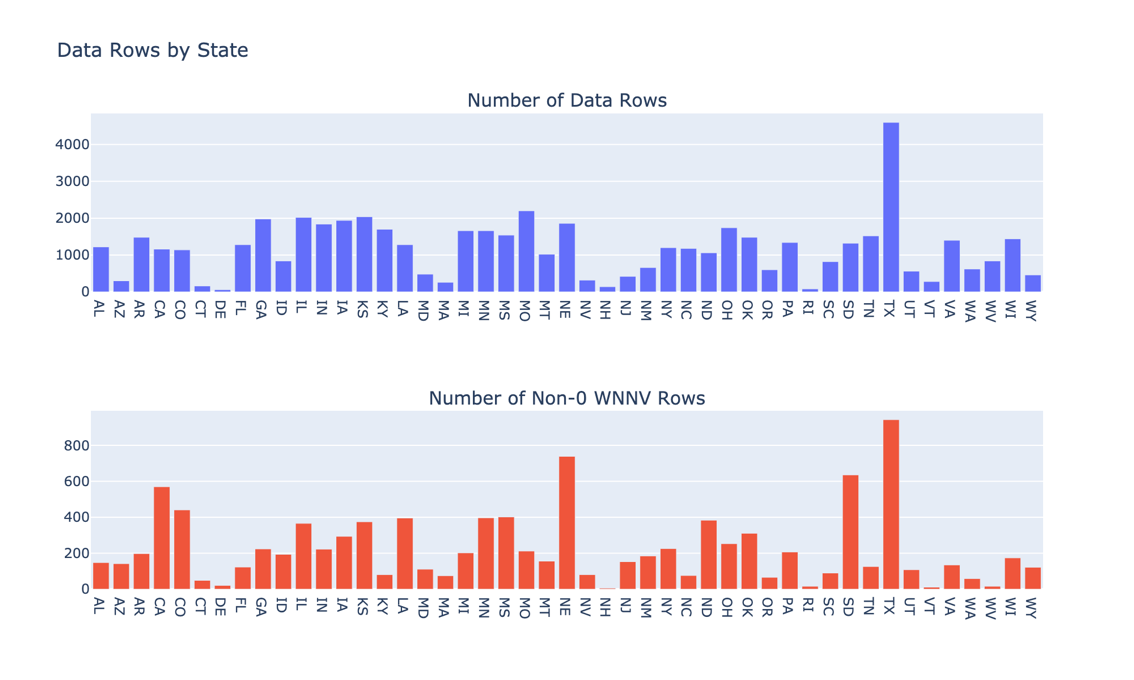
Description automatically generated

Figure 11 WNV case distribution across the nation (Top) and between each state (Bottom). In the plot, the blue represents the total number of rows in each state, where the red represents the total number of rows with non-0 WNV cases in each state.

Table 5 delineates the performance of three different models in predicting national annual WNV case counts. When compared with the results in Table 3, we observe that the Q2 scores for each model remain relatively consistent. However, there is a notable increase in Mean Squared Error (MSE), suggesting that the models are now less conservative and more likely to predict higher numbers of cases. Similarly, Figure 12 showcases the performance of state specific SVM models in predicting annual WNV case counts. The models exhibit Q2 scores comparable to those of the original models depicted in Figure 10, yet with increased Mean Squared Logarithmic Error (MSLE), further indicating a tendency towards predicting higher case counts.

Table 5: Performance of Different Predict Models on National Model Using Subsampling

|  |  |  |
| --- | --- | --- |
| Model | MSE | Q2 |
| Linear Regression | 200.46 | 0.05 |
| Random Forest \* | 202.52 | 0.05 |
| HGBR | 195.42 | 0.07 |
| SVM | 193 | 0.089 |

\* The MSE and Q2 for RF is average results of 100 runs

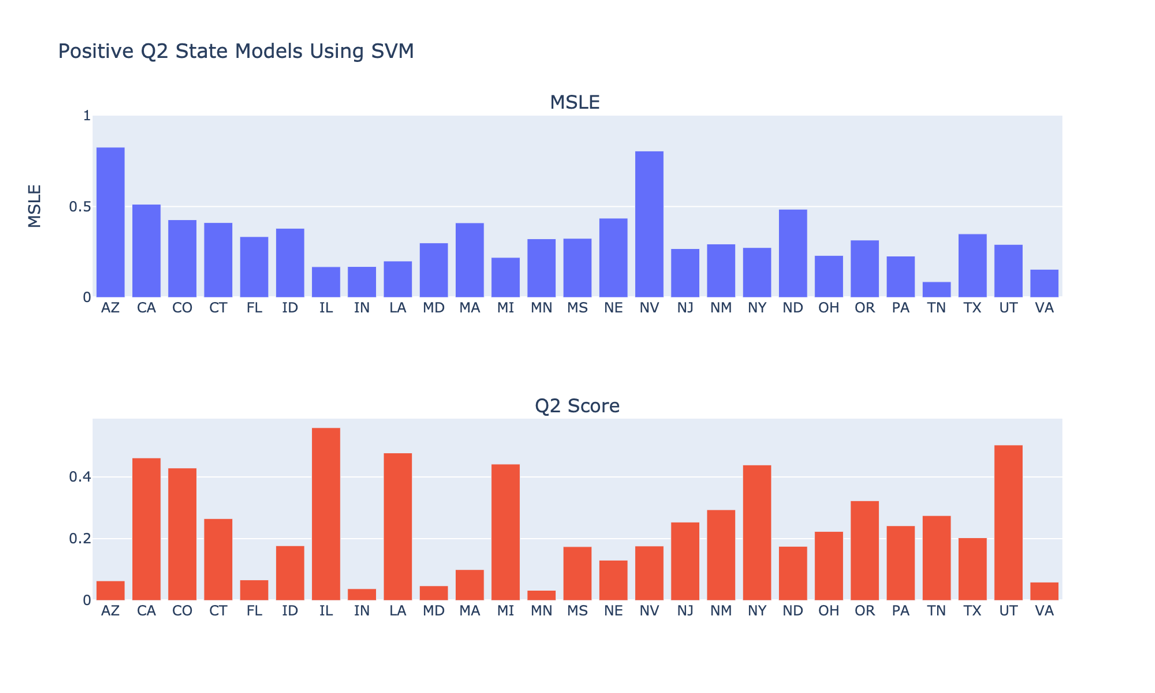


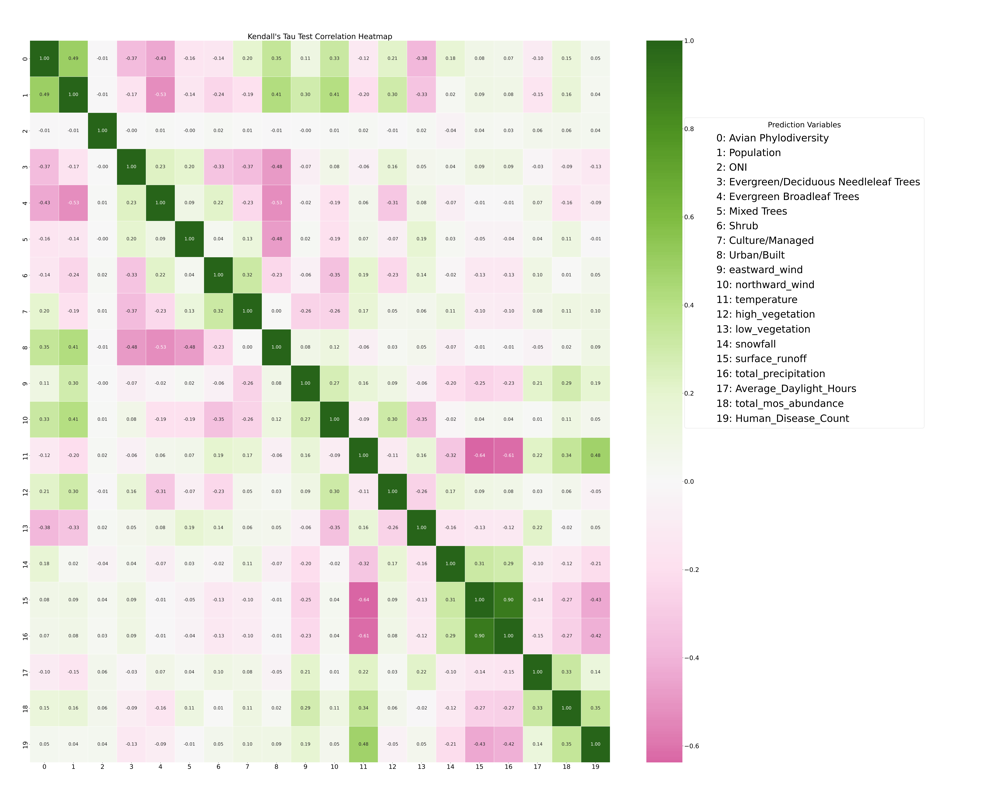
Figure 12: Comparative Analysis of SVM Models for WNV cases Across States Using Subsampling

### Data Granularity

To investigate how data granularity impacts model accuracy, we transitioned from using national CDC ArboNET data, which reports yearly WNV case counts at the state level, to more detailed state-specific datasets. The limitations of the CDC data, primarily its low spatial and temporal resolution, resulted in poor forecasting performance.

To address these challenges, we utilized two independent datasets with finer granularity, both focusing on California. The first dataset, referred to as the CDPH-Official dataset, was obtained through an official request to the CDPH Vector-Borne Disease Section and includes weekly WNV disease reports for 13 counties (see Clinical Disease Surveillance Data section) between 2004 and 2023. The second dataset, referred to as the CDPH-Scraped dataset, was independently compiled by scraping more than 760 PDF files of weekly reports available on the CDPH website. This dataset includes all counties in California and provides a comprehensive view of WNV case distribution across the state. These two datasets, treated independently, offered complementary insights into the temporal and spatial patterns of WNV cases at the weekly and county levels.

The enhanced granularity of these datasets revealed stronger and more noticeable connections between WNV case counts and predictors—including environmental variables and demographic factors—at the county level. Figure 13 illustrates the correlation improvements observed when using CDPH-Official data compared to national-level CDC data. Figure 13 reveals significant positive relationships between temperature, total mosquito abundance, eastward wind, and human disease count, as well as significant negative relationships between surface runoff, total precipitation, snowfall, and human disease count. These findings suggest that these variables are key drivers of WNV incidence. Additionally, average daylight hours and cultural and urban areas exhibit moderate positive correlations, while evergreen tree cover (variables 3 and 4) shows moderate negative correlations, further emphasizing their importance in modeling disease trends. The p-value heatmap confirms the statistical significance of these relationships. These insights highlight the importance of matching data granularity to the scale of environmental and epidemiological processes to enable more accurate forecasting.



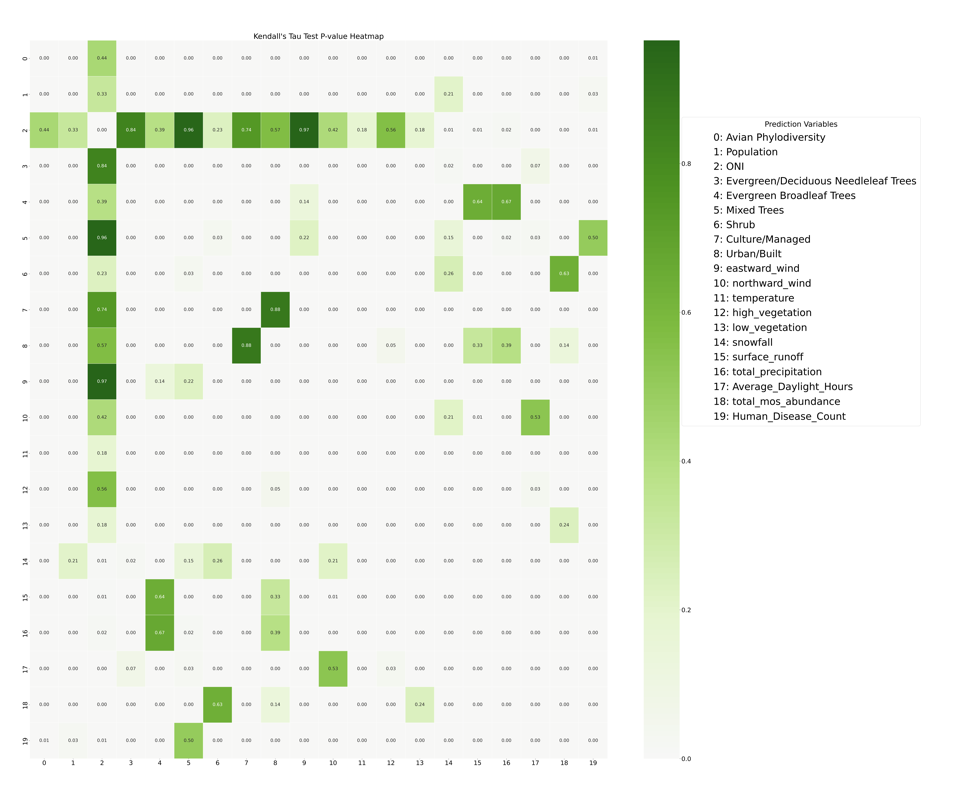


Figure 13 Kendall’s Correlation and P-value Heatmap between environmental variables and WNV disease count using CDPH-Official dataset

We implemented an iterative prediction framework to ensure that our models continuously adapted to new data while maintaining a realistic forecasting setup. The process began by training the model on data from the first available year (2004) and optimizing its hyperparameters using data from the following year (2005) through grid search. This optimization step involved systematically evaluating multiple hyperparameter combinations to identify the best-performing configuration based on a predefined evaluation metric (see Section 4.2.4 Hyperparameters tuning section). Once the optimal hyperparameters were determined, the tuned model was used to predict each subsequent year individually from 2006 to 2023. The process was then repeated iteratively: after training on data from 2004–2005, hyperparameter tuning was conducted on 2006 using grid search, and the model was applied to predict 2007–2023. This iterative procedure continued, incrementally incorporating all prior years for training, tuning hyperparameters on the most recent available year, and using the optimized model to forecast future cases.

By progressively refining the model with the latest data while leveraging historical trends, this iterative approach allowed us to adapt to evolving patterns in West Nile Virus (WNV) outbreaks, ultimately enhancing the robustness and accuracy of our predictions. Figure 14 presents the Q² results for Support Vector Machines (SVM), Random Forest (RF), and Histogram-Based Gradient Boosting Regression (HGBR). For improved visualization, Q² values lower than -1 were capped at -1. In Figure 14, the y-axis represents the year used for hyperparameter tuning, while the x-axis denotes the predicted year.

The results in Figure 14 indicate that predictions for the years 2009, 2010, 2018, 2019, and 2021 consistently exhibit poor performance. While the low predictive accuracy in earlier years (e.g., 2009 and 2010) may be attributed to the limited training data available, which hinders the model’s ability to generalize patterns effectively, this explanation does not account for the poor performance observed in 2018, 2019, and 2021. These findings suggest that additional factors, not captured by the model, may have specifically influenced these years, warranting further investigation.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 14: Iterative Forecasting Performance (Q2 Values) of SVM, RF, and HGBR Models

As an alternative approach, we partitioned the dataset into three subsets: 80% of the data from 2004 to 2018 for training, the remaining 20% as a validation set, and 2019 to 2023 as the test set. The models were trained on the training set, hyperparameter tuning was performed on the validation set, and final evaluations were conducted on the test set.

Figure 15 presents the Q² and Mean Squared Error (MSE) results for Support Vector Machines (SVM), Random Forest (RF), and Histogram-Based Gradient Boosting Regression (HGBR). Among the models, SVM demonstrated the best performance; however, its Q² value was only 0.2, highlighting the model’s limitations in capturing the underlying patterns effectively.

A diagram of a diagram

AI-generated content may be incorrect.

Figure 15: Performance of SVM, RF, and HGBR When Trained on 2004–2018 and Tested on 2019–2023

The poorer performance of the model trained on 2004–2018 and tested on 2019–2023 compared to the iterative approach suggests the presence of concept drift and temporal changes in the data. Factors such as climate variations, shifts in mosquito and bird populations, and changes in human intervention strategies may have altered WNV transmission patterns after 2018, making older data less predictive of recent years. Additionally, the iterative approach adapts yearly, leveraging the most recent data for hyperparameter tuning, whereas the fixed 2004–2018 model relies solely on historical patterns that may no longer be relevant. Overfitting to older trends, data distribution shifts, and the presence of anomalous years (e.g., 2018, 2019, and 2021) further contribute to the model’s reduced generalization ability. These findings indicate that recent data is more informative for forecasting than older historical data, highlighting the need for models that can dynamically adjust to evolving conditions.

To improve the accuracy of West Nile Virus (WNV) predictions, we incorporated time-series data into our modeling approach (Figure 16). We first calculated the month-to-month difference in WNV cases and created a new column to capture these changes. Next, we generated lagged features by shifting this difference by 1, 2, and 3 months, resulting in columns representing t-1, t-2, and t-3. These lagged variables were then integrated with the environmental factors used in our previous model, forming an enriched dataset for training and testing.

To leverage both historical case trends and external environmental factors, we developed a hybrid autoregressive model that integrates classical machine learning methods with time-series autoregression techniques. By incorporating temporal dependencies, this approach allows the model to better capture patterns in WNV outbreaks, ultimately enhancing prediction accuracy.

A screenshot of a table

AI-generated content may be incorrect.

Figure 16: Example of construction of Lagged Temporal Features for WNV Prediction

To evaluate the effectiveness of our predictive models, we applied Support Vector Machines (SVM), Random Forest (RF), and Histogram-Based Gradient Boosting Regression (HGBR) using a structured training, validation, and testing approach. Specifically, we trained the models on data from 2004 to 2015, allowing them to learn historical patterns and relationships. Next, we performed hyperparameter tuning using data from 2016 to 2018, optimizing model performance through grid search to identify the best parameter configurations. Finally, we assessed the models’ predictive accuracy by testing them on an unseen dataset from 2019 to 2023. The results of this evaluation are presented in Figure 17. Additionally, we plotted the global SHAP values for each predictive variable across all models (Figure 18) to interpret feature importance.

From Figure 17, we observe that incorporating autoregression into the model significantly enhances prediction accuracy compared to models that rely solely on environmental variables. Among the models tested, Random Forest demonstrates the strongest performance, achieving a Q² close to 0.5 and an MSE below 13, indicating a more reliable representation of observed case trends. This suggests that integrating temporal dependencies, such as past case trends, provides valuable predictive power beyond environmental factors alone.

A diagram of a diagram

AI-generated content may be incorrect.

Figure 17: Result of Autoregression model using SVM, RF and HGBR

Figure 18 further highlights the most influential predictors across the three models, revealing that t-1 (the difference in cases from the previous two months), temperature, average time length, and northward wind play the most significant roles in forecasting case numbers. Notably, temperature, average time length, and northward wind are not only key predictors in our models but also fundamental factors influencing the environmental adaptation of *Culex tarsalis*, as discussed in the previous chapter. Their strong predictive importance in this study reinforces their role in shaping mosquito population dynamics and, consequently, the transmission risk of West Nile Virus.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 18: Feature Importance Analysis Using Global SHAP Values for SVM, RF, and HGBR Models

To determine the optimal number of lagging months to incorporate into the model, we evaluated the Random Forest model using lag values ranging from 0 months (no lag) to 5 months. Figure 19 presents the model’s performance across different lag configurations. The results indicate that incorporating t-1 (1-month lag), t-2 (2-month lag), and t-3 (3-month lag) alongside environmental variables yields the highest Q² and the lowest MSE, suggesting that including short-term historical trends significantly enhances predictive accuracy. Beyond t-3, additional lagging months do not provide further improvements, likely due to diminishing temporal dependencies in WNV case trends.

A graph with lines and numbers

AI-generated content may be incorrect.

Figure 19: Impact of Temporal Lag on Random Forest Prediction Accuracy (Q² and MSE)

To quantify the uncertainty in our predictions, we conducted bootstrapping on the SVM, RF, and HGBR models. This involved resampling the test dataset (2019 - 2023) 1000 times with replacement and recalculating the model estimates, allowing us to derive a 95% confidence interval for the predictions. The results are showing in Figure 20. This statistical measure provides insight into the variability and reliability of our forecasts, ensuring robustness in our findings.

A group of blue and red graphs

AI-generated content may be incorrect.

Figure 20: Histogram of 95% Confidence Intervals of Q² from Bootstrapped Predictions for SVM, RF, and HGBR. SVM: (A) and (B); Random Forest: (C) and (D); Histogram-Based Gradient Boosting Regression: (E) and (F)

## Discussion

This study demonstrates the effectiveness of incorporating autoregressive components into machine learning models for West Nile Virus (WNV) prediction, significantly improving forecast accuracy over models relying solely on environmental variables. By integrating past case trends as predictive features, our hybrid autoregression approach enhances the temporal resolution of WNV forecasting, allowing for more robust early-warning systems.

Among the models tested, Random Forest consistently outperformed Support Vector Machines (SVM) and Histogram-Based Gradient Boosting Regression (HGBR), achieving a Q² close to 0.5 and a Mean Squared Error (MSE) below 13. This suggests that ensemble-based approaches, which capture complex interactions between features, are particularly well-suited for modeling WNV dynamics. The strong predictive influence of temperature, average daylight length, and northward wind aligns with prior ecological findings on *Culex tarsalis* adaptation, reinforcing the role of these environmental drivers in WNV transmission.

This study highlights several key advancements in West Nile Virus (WNV) prediction, demonstrating the advantages of integrating autoregressive modeling with machine learning techniques.

* Enhanced Predictive Accuracy with Autoregression

Incorporating time-lagged case data significantly improved model performance, underscoring the importance of historical trends in forecasting WNV outbreaks. The inclusion of past case trends enabled the models to capture temporal dependencies more effectively, leading to notable improvements in predictive accuracy. Compared to models relying solely on environmental variables, the autoregression-enhanced models demonstrated superior performance, as reflected in increased Q² values and reduced Mean Squared Error (MSE).

* Feature Importance and Ecological Relevance

Analysis of model feature importance revealed that temperature, average daylight length, and northward wind were among the most influential predictors of WNV incidence. These findings align with the known environmental adaptations of *Culex tarsalis*, further reinforcing the biological plausibility of the predictive models. The SHAP analysis (Figure 18) provided additional interpretability, offering insights into how both environmental and autoregressive variables contribute to model predictions. By capturing these critical ecological factors, the models offer a biologically meaningful framework for anticipating WNV transmission dynamics.

* Uncertainty Estimation and Model Robustness

Given the inherent variability in disease forecasting, quantifying uncertainty is essential for model reliability. The implementation of bootstrapping techniques (Figure 19) enabled the estimation of 95% confidence intervals, providing a measure of uncertainty around predictions. This approach reduces the risk of overconfident forecasts and ensures that the models remain statistically robust, even in cases of limited or noisy data.

Despite these promising advancements, several challenges remain that warrant further investigation:

* Temporal and Concept Drift

The models exhibited lower predictive performance in certain years (e.g., 2018, 2019, 2021), suggesting that unmodeled factors such as changes in vector control efforts, shifts in surveillance practices, or climate anomalies may have influenced disease transmission patterns. These findings highlight the need for adaptive learning frameworks that can dynamically update as new data becomes available, ensuring that the models remain responsive to evolving epidemiological trends.

* Data Quality and Missing Values

Discrepancies between the official CDPH dataset and the independently scraped dataset indicate potential inconsistencies in WNV surveillance data. While imputation techniques were used to address missing values, biases may persist, particularly in underreported cases. Improving data harmonization efforts across surveillance systems will be crucial for enhancing model reliability.

* Computational Complexity and Scalability

While ensemble-based models outperformed other approaches in our autoregression approach, their computational demands were substantially higher compared to simpler linear models. As the volume of surveillance data increases, optimizing model efficiency becomes a priority. Future efforts should explore techniques such as model distillation, feature selection, and distributed computing frameworks to balance predictive power with computational feasibility.

# Conclusion and Future Work