**Table of Contents**

[CHAPTER 1: Background and Significance 4](#_Toc187677254)

[CHAPTER 2: Previous Work in Metabolomics & Machine Learning 6](#_Toc187677255)

[2.1 Memory-Efficient Searching of Gas-Chromatography Mass Spectra Accelerated by Prescreenin*g* 6](#_Toc187677256)

[2.2 Metabolic Signature Discovery 9](#_Toc187677257)

[CHAPTER 3: Climate Adaptation and Genetic Differentiation in the Mosquito Species *Culex tarsalis* 14](#_Toc187677258)

[3.1 Introduction 14](#_Toc187677259)

[3.2 Materials and Methods 15](#_Toc187677260)

[3.2.1 Sample Collection 16](#_Toc187677261)

[3.2.2 Genome Sequencing, Assembly, and Annotation 16](#_Toc187677262)

[3.2.3 RAD-Seq Library Preparation, Sequencing, and SNP Calling 16](#_Toc187677263)

[3.2.4 Climate Data Extraction 17](#_Toc187677264)

[3.2.5 Population Structure Analyses 20](#_Toc187677265)

[3.2.6 Identification of Genotype-Environment Associations 23](#_Toc187677266)

[3.2.7 Detecting SNPs Under Selection 24](#_Toc187677267)

[3.3 Results 24](#_Toc187677268)

[3.3.1 Genomic Analyses 24](#_Toc187677269)

[3.3.2 Population Structure 25](#_Toc187677270)

[3.3.3 Isolation by Distance (IBD) verse Isolation by Environment (IBE) 27](#_Toc187677271)

[3.3.4 Partial Redundancy Analysis 29](#_Toc187677272)

[CHAPTER 4: West Nile Virus Prediction 31](#_Toc187677273)

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

# Background and Significance

Mosquito-borne diseases, such as West Nile Virus (WNV), present formidable public health challenges worldwide. Since its initial detection in the United States in 1999, WNV has inflicted a significant toll, with over 56,000 documented cases and 2,700 fatalities. 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). Its pervasive spread has firmly established WNV as the foremost mosquito-borne disease in the country, warranting meticulous attention from public health authorities. This emergence underscores the intricate interplay of environmental conditions, geographical features, and vector biology in shaping disease transmission dynamics.

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 use 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.

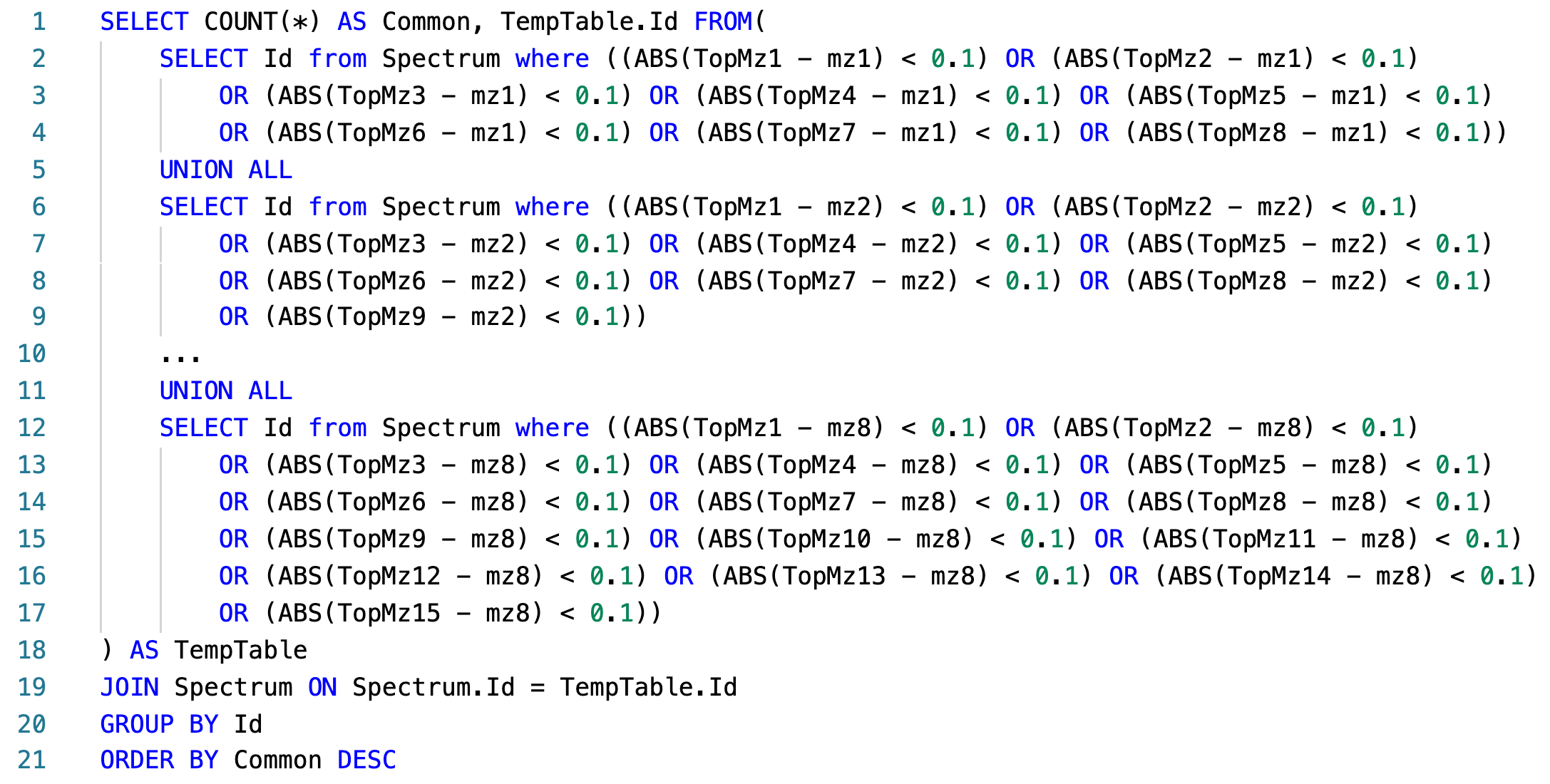


Figure 1: Pseudo SQL query for calculating similarity scores between a query spectrum and all library spectra.

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.

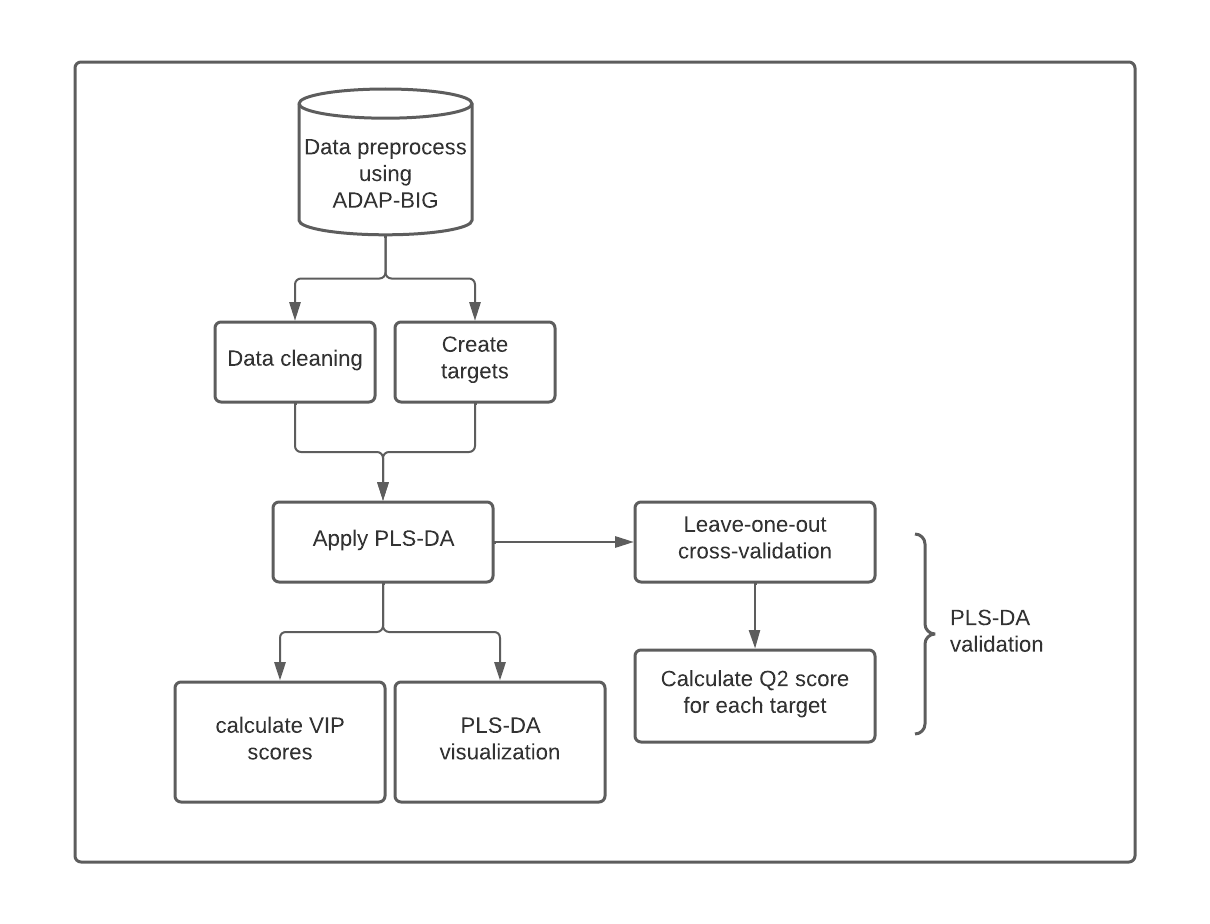


Figure 4: Metabolic Signatures Discover Pipeline

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).

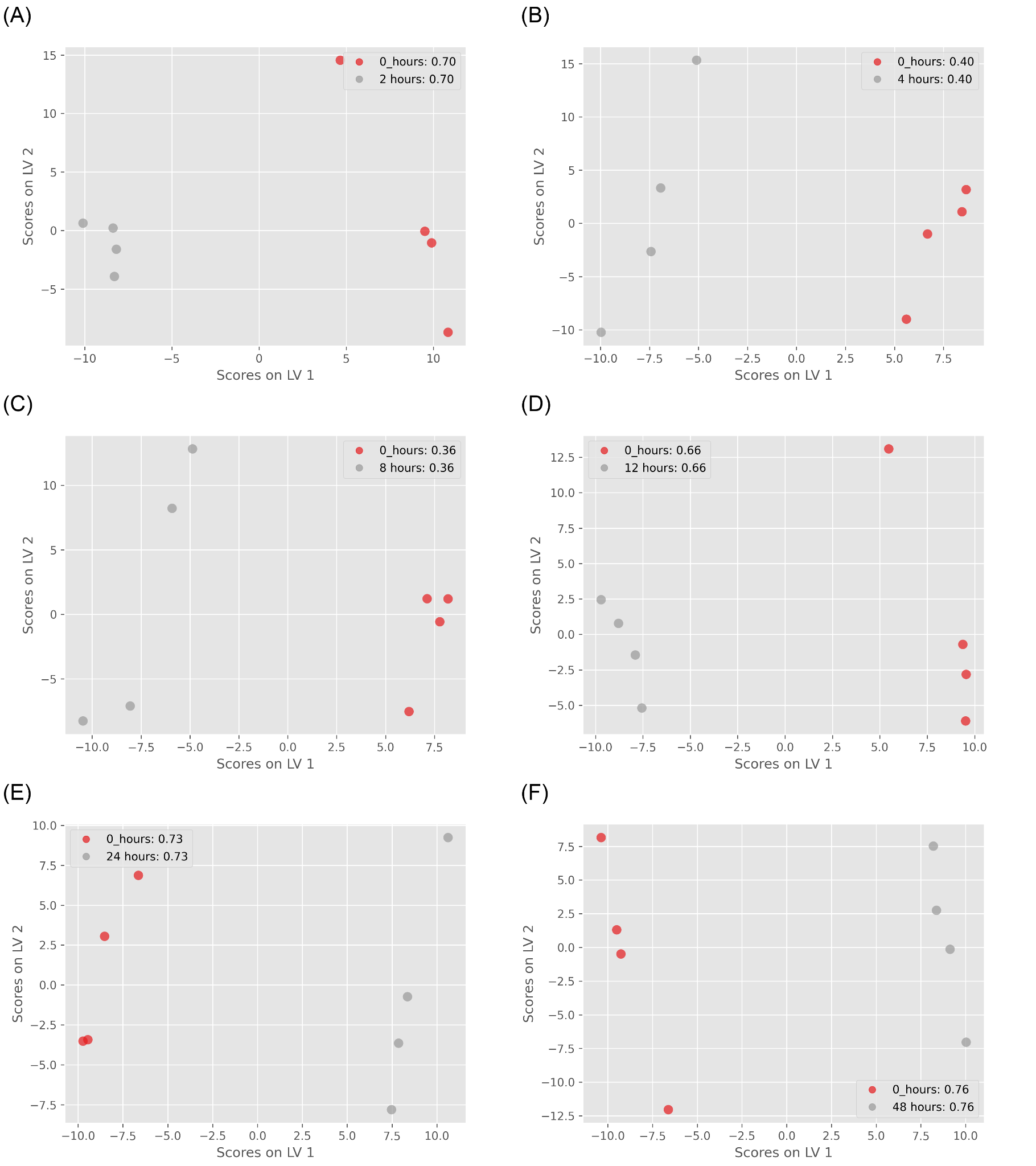


Figure 5: ST000085 PLS-DA results: (A) control group vs. 2 hours treatment group (Q2=0.7); (B) control group vs. 4 hours treatment group (Q2=0.4); (C) control group vs. 8 hours treatment group (Q2=0.36); (D) control group vs. 12 hours treatment group (Q2=0.66); (E) control group vs. 24 hours treatment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

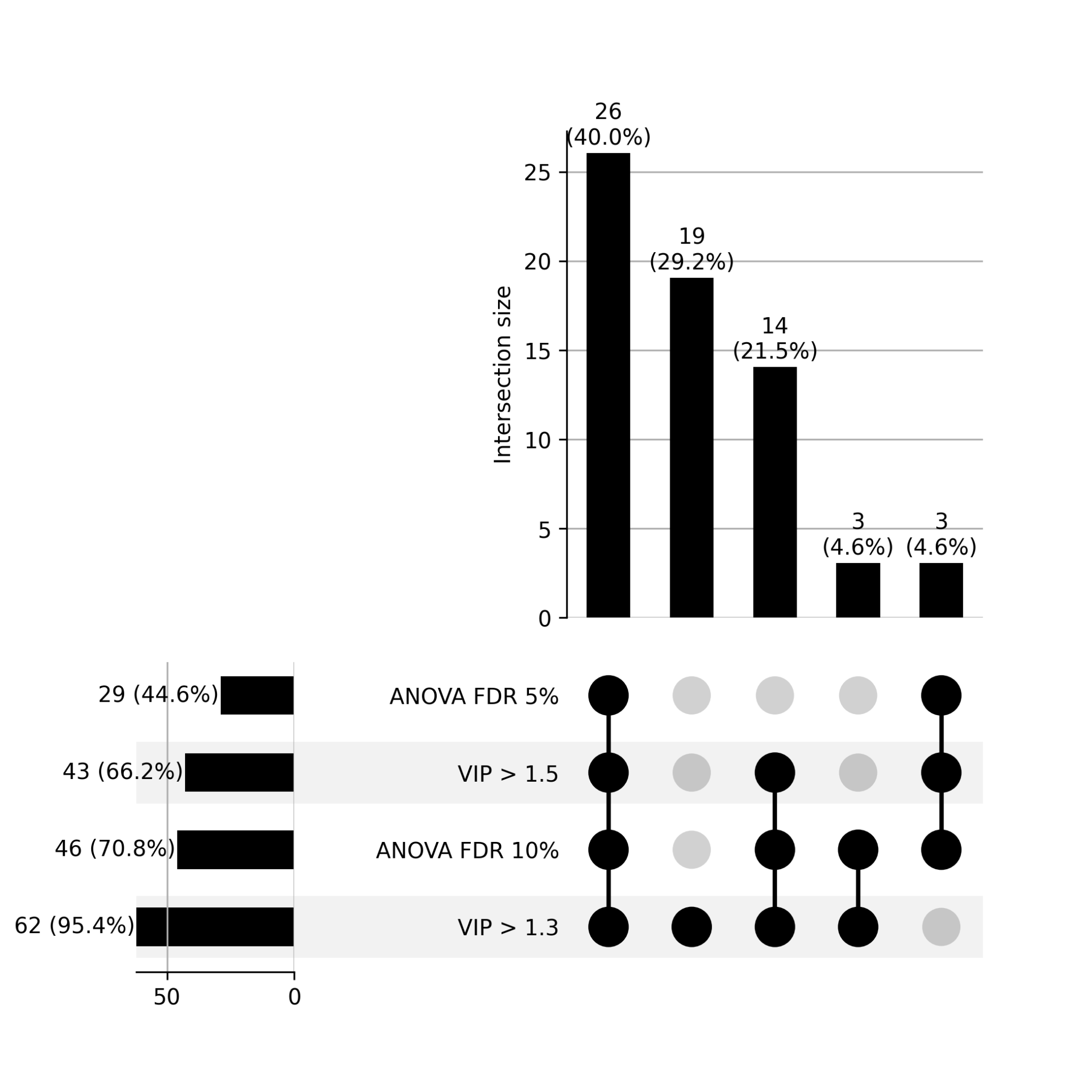


Figure 6: Upset plot between PLS-DA and ANOVA test. VIP cutoff is 1.3 and 1.5 in PLSDA; FDR threshold is 5% and 10% in ANOVA test.

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 qpplied 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 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)

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. 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.

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 population

Description automatically generated with medium confidence

**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 |

# West Nile Virus Prediction