Final Report – Revision 1

Developing genetic approaches for sustainable long-term monitoring and modeling CWD spread in White-tailed Deer

Submitted to:
Chris Middaugh
Arkansas Game and Fish Commission
771 Jordan Drive
Monticello, AR 72655

Prepared by:
Marlis R. Douglas
Bradley T. Martin
Zachery D. Zbinden
Tyler K. Chafin
Nathan R. Campbell
Michael E. Douglas

Arkansas Conservation & Molecular Ecology Laboratory (aCaMEL)
University of Arkansas
Biological Sciences
850 W Dickson St
Fayetteville, AR 72701

Respectfully submitted:
25 October 2024

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Overview

This study represents an extension of ongoing collaborative research between the Arkansas Game and Fish Commission (AGFC) and the Arkansas Conservation and Molecular Ecology Laboratory (aCaMEL) at the University of Arkansas. The research program was initiated in 2017 to inform management and risk assessment of Chronic Wasting Disease (CWD) in White-tailed Deer in Arkansas by generating actionable insights derived from genetic data. The initial project established a state-wide genetic database of White-tailed Deer as a baseline for monitoring, risk assessment and modeling the potential spread of CWD. The goal of the current project was to enhance the genetic monitoring capacities and model predictions by developing standardized, accurate, and easy-to-reproduce workflows and protocols.

Approach

Our previous research generated actionable information that guided management decision and identified logistic and theoretical limitations of existing methods. Thus, the current project targeted several implementation issues: (i) time and expertise required to generate and analyze genotype data; (ii) lack of robust methods to detect and remove spurious statistical signals; and (iii) poor performance of existing analytical methods with biased sampling. The main objective was to untangle genetic population structure based on natural dispersal of wild deer from genetic signals reflecting historical management actions (i.e., translocations). A second objective was reducing analytical biases due to sampling variation (i.e., over- and under-sampled areas). We addressed these by (1) developing a standardized genotyping assay (i.e., SNP GT-seq panel), (2) creating novel computational approaches for bias detection and mitigation, and (3) generating user-friendly workflows and software tools to facilitate data analysis.

Key Outputs

Genotyping Assay: Tool SNP GT-seq panel

- To simplify genotyping of new samples, we developed and validated a standardized genotyping assay, **SNP GT-seq panel,** that screens 441 SNP loci using the GT-seq approach.
- SNPs (Single Nucleotide Polymorphisms) represent genetic variation throughout the genome of deer and are informative to determine population structure and genetic similarity of samples.
- Genotyping success rate is high (95.9%) for the SNP GT-seq assay, and genotypes are concordant (91.4% ± 3.02%) with data generated with the previous method (i.e., dRAD-seq).
- The new assay has the potential for ~75% cost-reduction/sample.

Genotyping Data Processing: **Tool GTSEQ2VCF**

- To facilitate processing of the SNP data generated with the new GT-seq assay and integrate these SNP genotypes into the existing database, we developed a bioinformatics **program, GTSEQ2VCF**.
- **GTSEQ2VCF** is a comprehensive pipeline and merges new SNP data produced via the GT-seq assay with the existing state-wide SNP dataset for analysis.
- It also integrates quality control steps to ensure reliability and accuracy of data.

Geolocation Analysis: **Tool GEOGENIE**

- To increase accuracy of geolocation predictions, we developed a new software program called **GEOGENIE** based on AI (Artificial Intelligence) approaches.
- **GEOGENIE** uses a novel deep learning approach to predict 'geographic origin' of samples.
- **GEOGENIE** is highly accurate (median predication error ~5km). It performs consistently across most of the state using less SNP data.
- **GEOGENIE** will be freely available via public repositories (i.e., GitHub) and actionable via a user-friendly manual.

Landscape Resistance Analysis: **Tool RESISTANCEGENIE**

- To facilitate fine-scale predications of deer dispersal across Arkansas, and hence inform risk assessment of potential CWD-spread, we developed a software program called **RESISTANCEGENIE**.
- **RESISTANCEGENIE** facilitates landscape resistance analyses using the state-wide SNP database.
- **RESISTANCEGENIE** establishes a 'best practices' automated bioinformatics workflow to visualize areas of high *versus* low population connectivity.
- **RESTISTANCEGENIE** will be freely available via public repositories (i.e., GitHub) and actionable via a user-friendly manual.

Conclusion

The tools and documentation generated for this project help overcome prior analytical limitations and make genetic information more 'actionable' to help guide White-tailed Deer management and CWD risk-assessment in Arkansas – and beyond. The genotyping assay and software tools developed in this project generate more accurate and consistent predictions at lower per-sample costs. Enhanced analytical accuracy and cost-efficiency will enable AGFC and partner agencies to make better informed, and thus more effective management decisions. This supports sustainable management practices, and facilitates long-term implementation and data integration across various agencies and regions.

1 | INTRODUCTION

1.1 | Background

Chronic Wasting Disease (CWD) was first confirmed in wild cervids in Arkansas in 2016. Since then, CWD has been detected in 20 counties and over 1,729 White-tailed Deer in the state (data AGFC June 30, 2024; https://www.agfc.com/hunting/deer/chronic-wasting-disease/cwd-in-arkansas/; Figure 1). To use science-based management and inform risk assessment of disease susceptibility and spread, the Arkansas Game and Fish Commission (AGFC) initiated a collaboration with the Arkansas Conservation and Molecular Ecology Laboratory (aCaMEL) at the University of Arkansas in 2017. This initial effort resulted in a state-wide database of genetic diversity and population structure of White-tailed Deer in Arkansas (Douglas et al. 2019). The project characterized variation and distribution in the White-tailed Deer prion protein gene *PRNP* (Chafin et al. 2020) and quantified spatial patterns in White-tailed Deer genetic diversity state-wide (Chafin et al. 2021). This information serves as a baseline and helps guide management decisions on mitigating the risk and spread of CWD in Arkansas (Ballard et al. 2021).

The AGFC continues to monitor CWD in Arkansas and the need to expand the CWD Management Zone as cases of CWD+ deer are confirmed in new areas. The state-wide genetic database can be leveraged to determine the likely geographic origin of any White-tailed Deer genetic sample using a modeling approach (i.e., probabilistic geolocation prediction). The model predictions can indicate whether the deer is from a local herd, has dispersed from a geographically distant population, or has potentially escaped from a captive population (Figure 2). Each of these scenarios has different management implications. Most importantly, this information is useful to evaluate the likelihood that a CWD+ White-tailed Deer dispersed from the Management Zone where the disease is already known to occur, or if a deer contracted the disease locally from another CWD+ deer, potentially indicating a new CWD hotspot (Douglas et al. 2022). **Accurate geolocation predictions enhance the AGFC's capacity to make informed decisions regarding CWD detections outside the current CWD Management Zone, where to focus management actions, such as targeted removal operations, and which strategies might be most effective at minimizing CWD spread into unaffected areas.**

1.2 | Current Project

In this study, we developed a more standardized genotyping approach, a **SNP GT-seq assay**, to simplify the screening of informative genetic variation in new White-tailed Deer collected in Arkansas. The method builds on our previous study (Douglas et al. 2019) and expands the existing state-wide database of SNP (Single Nucleotide Polymorphism) variation, but uses a more cost- and time-efficient screening method called GT-seq (Genotyping-in-Thousands-sequencing; Campbell et al. 2015). The approach simultaneously assays population genetic variation and *PRNP* gene variants and is more user-friendly than previously used approaches (Douglas et al. 2019; Chafin et al. 2020, 2021).

We also enhanced the geolocation predication analyses by developing several bioinformatics tools that integrate new SNP genotypes into the existing database (**GTSEQ2VCF)** and rely on an AI approach (i.e., Deep Learning; **GEOGENIE)** to predict 'geographic origin' of samples with high accuracy (median predication error ~5km). The new geolocation tool performs consistently across the entire state using less data and mitigates prediction errors caused by outlier data or sampling deficiencies. We also developed an automated bioinformatics tool (**RESTISTANCEGENIE)** that conducts a landscape genomics analysis using the state-wide SNP database. The analysis provides fine-scale predications of landscape resistance to deer dispersal and is a tool to visualize areas of high *versus* low population connectivity. These tools provide a simple **workflow** and facilitate sustainable genetic monitoring to inform management practices and enhance data integration across agencies and regions.

2 | RESEARCH GOAL AND OBJECTIVES

The goal of this project was to simplify genetic monitoring of White-tailed Deer in Arkansas to facilitate sustainable, long-term management and risk assessment of CWD spread. The objectives included designing and validating a standardized genotyping assay (**SNP GT-seq panel**), creating standardized workflows and bioinformatics pipelines for swift SNP data processing (**GTSEQ2VCF**), developing best practices for applying machine-learning in geolocation analysis (**GEOGENIE**), and pioneering statistical and computational methods to study landscape resistance and animal dispersal (**RESTISTANCEGENIE**).

Objectives

1. Design and validate a standardized genotyping assay to simplify screening of informative genetic variation in White-tailed Deer in Arkansas. → Tool: **SNP GT-seq panel**
2. Create a standardized workflow and custom bioinformatics pipelines to rapidly process SNP data from GT-seq. → Tool: **GTSEQ2VCF**
3. Generate best practices for using and applying machine-learning geolocation analysis with SNP data. → Tool: **GEOGENIE**.
4. Develop novel statistical and computational approaches to model landscape resistance and individual dispersal in species with translocation histories. → Tool: **RESTISTANCEGENIE**.
5. Disseminate research findings and information about user-friendly resources through open-source software, presentations, and peer-reviewed publications. → Tool: **GitHub**.

3 | METHODS

3.1 | Genotyping Assay (SNP GT-seq Panel Development)

To simplify screening of informative genetic variation in new White-tailed Deer samples, we developed a genotyping assay that standardizes genotyping of *N*=441 SNPs using the Genotyping-in-Thousands sequencing method (GT-seq; Campbell et al. 2015). The SNP data can be used for several purposes: (i) assignment of a White-tailed Deer sample to a genetic population and inferring its geographic origin; (ii) assessment of *PRNP* variants of individuals and populations; (iii) determining the sex of individual samples.

3.1.1 | SNP Discovery

First, we established a database for *N*=1,381 individual deer from Arkansas containing genotypes generated via ddRAD-seq (i.e., double-digest restriction-site-associated DNA sequencing) following protocols described in Chafin et al. (2021). This database comprised *N*=1,242 samples from our previous study collected from 2016-2019 (Chafin et al. 2020, 2021) and *N*=139 new samples collected by AGFC since 2019 (Douglas et al. 2022). DNA was extracted following Chafin et al. (2021).

Genetic variants called SNPs (=Single Nucleotide Polymorphisms) were identified across all individuals (*N*=1,381) by reference-guided assembly of sequences based on a chromosome-level genome of White-tailed Deer (London et al. 2022) using ipyrad (Eaton & Overcast 2020). This process identified *N*=1,046,465 SNPs.

3.1.2 | SNP Selection for GT-seq Panel Optimization

We first targeted the most informative SNP loci to optimize the GT-seq assay for White-tailed Deer management. This process required bioinformatic filtering of data to remove SNP loci and/or samples that did not meet specific criteria (i.e., quality control and filtering; details provided in Appendix 1). Our goal was to select ~500 most informative SNP loci out of the initial set of 2,156 SNP loci and involved reducing the dataset via quality control and filtering of individuals and loci using R statistical software (version 4.1.3; R Core Team 2022).

To confirm that the filtered set would retain enough genetic information to infer population structure, such as the eight genetic populations identified for White-tailed Deer in Arkansas by Chafin et al. (2021), we applied sparse non-negative matrix factorization (sNMF; Frichot et al. 2014). SNP loci were then ranked by their population divergence (Li et al. 2023), calculated from the weighted per-locus population divergence (F_{ST}) based on the matrix of ancestry proportions (see Appendix 1 for more details), a method suited for admixed and continuously structured populations (Martins et al. 2016).

The **SNP GT-seq assay** was developed by GTseek, Inc. (Twin Falls, ID) and involved the design of primers for PCR amplification and optimization of multiplex reactions. An initial set of SNP loci was selected for further testing: *N*=800 autosomal loci with the highest genetic divergence, *N*=2 loci associated with the *PRNP* gene (capturing all known variants in Arkansas; Chafin et al. 2020), and *N*=3 sex-linked loci on the Y-chromosome (Strickland et al. 2011). The initial GT-seq panel was tested by genotyping *N*=192 White-tailed Deer: *N*=96 samples previously genotyped with ddRAD-seq (Table S1) and evenly distributed across Arkansas counties (Figure S1); and *N*=96 collected during FY23 surveillance by AGFC as Phase 6 of the project (Table S2). Results were evaluated for amplification/genotyping success and consistency with data generated via ddRAD-seq. The final, filtered SNP GT-seq panel comprised *N*=436 autosomal loci, *N*=2 *PRNP* loci, and *N*=3 sex-linked loci, demonstrating desirable genotype capture rates (Figure 3).

3.1.3 | GT-seq Data Processing Pipeline

Bioinformatics pipelines for GT-seq data processing and quality control were developed. To validated GT-seq as a cost-effective, accurate method, genotypes for *N*=96 samples were generated with both ddRAD-seq and GT-seq approaches (Table S1). A first step requires conversion of the GT-seq genotype data into a standard VCF (Variant Call Format) file using **GTSEQ2VCF,** an open-source pipeline we develop for this purpose. The converted SNP GT-seq genotype data can then be merged with an existing SNP genotype database generated via ddRAD-seq, such as in our case (Chafin et al. 2021).

GTSEQ2VCF generates a full graphical report of missing data and genotype mismatch for redundant samples. It performs all the necessary steps to merge a GT-seq file with an existing VCF genotype database. Loci are only kept if they overlap between the two datasets, allowing a simple merger of SNP data matrices derived via the ddRAD-seq and GT-seq approaches. The output is a single standardized VCF file with a standard reference system, ready to use as input for downstream analytical pipelines.

3.1.4 | Testing Consistency between Genotyping Approaches Using Ancestry Analyses

We wanted to confirm our new genotyping assay (SNP GT-seq panel) that screens fewer, but more informative SNP loci was equally robust in inferrning patterns of genetic population structure as our initial dataset containing many more SNP loci generated via ddRAD-seq (reported in Chafin et al. 2021). To accomplish this, we conducted three independent ancestry re-analyses using ADMIXTURE (Alexander & Lange 2011, Mussmann et al. 2023). To avoid spurious results due to missing data or other artefacts, each analysis contained slightly different sets of individuals and number of SNP loci (Figure 4).

Analysis 1: ddRAD-seq data- original (Figure 4A)

The original ancestry analysis was based on *N*=1,143 individuals and genotypes represented *N*=35,099 SNP loci generated via ddRAD-seq (Douglas et al. 2019). This analyses revealed eight genetic populations

as the most likely genetic structure (published in Chafin et al. 2021). Geographic distribution of each genetic population was mapped via ancestery kriging – an oversimplified visualization of assignment probabilities by gene pool.

Analysis 2: ddRAD-seq data- referenced aligned (Figure 4B)

One anestery re-analysis was based on *N*=1,302 individuals and genotpes represented *N*=46,612 SNP loci generated via ddRAD-seq that were referenced-aligned. This data set resulted from filtering the initial ddRAD-seq dataset of *N*=1,046,465 SNPs, removing loci with >50% missing data, a minor allele count <3, and retaining one SNP per read. Individuals were removed if they had >90% missing data (*N*=146) or if they were deemed outliers (*N*=29) based on the geolocation analyses described below.

> *Note: This set did included samples from Phases 1-5, but not the FY23 samples (N=96; Phase 6) that were genoytped across 441 SNP loci with the new GT-seq approach.*

Analysis 3: ddRAD-seq data- reduced to 436 SNP loci (Figure 4C)

Another ancestery re-analysis was based on *N*=1,203 individuals and contained genotype data across the *N*=436 autosomal SNP loci (GT-seq panel). This data set was created by using the ddRAD-seq generated data (above), but reducing genotypes to just the subset of SNP loci included in the GT-seq panel. Individuals were removed if they had >75% missing data (*N*=53) or if they were deemed outliers (*N*=29) based on the geolocation analyses described below.

> *Note: This set did not include genotypes generated with the GT-seq assay (i.e., N=96 individuals from FY23/Phase 6; N=96 re-genotyped individuals from Phase 1-5). We intentionally did not include GT-seq generated data because this methods results in few missing loci, unlike the ddRAD-seq generated data that has a relative proportion of missing loci. The inclusing of 'low' and 'high' missing data genotypes can cause spurious analytical results (e.g., formation of a 'pseudo-population' - see below).*

Analysis 4: ddRAD-seq and GT-seq data- reduced to 436 SNP loci (Figure 4D)

A final ancestry re-analysis was based on *N*=1,286 individuals and contained genotype data across the *N*=436 autosomal SNP loci (GT-seq panel). This data set was a combination of genotypes generated via ddRAD-seq and the GT-seq approaches; it included all samples from analyses 3 (above) plus 183 samples genotyped with the GT-seq assay. Individuals were removed if they had >75% missing data (*N*=62; these included 9 of the 192 samples genotyped with GT-seq) or if they were deemed outliers (*N*=29) based on the geolocation analyses described below.

> *Note: This set included all samples received from Phase 1-6, and a mix of ddRAD-seq and GT-seq generated genotype data. It was thus comprised of genotypes with 'high' and 'low' missing data.*

For each analysis, models with one (*K*=1) and up to fifteen (*K*=15) populations were compared using 20 replicates each. To facilitate comparison with Chafin et al. (2021) visualizations were based on eight genetic populations. The individual sample ancestries and their geographic coordinates were used to create a 250,000-pixel raster (500 x 500) of spatially interpolated ancestry via kriging (Caye et al. 2016) to visualize patterns of dominant population ancestry across Arkansas, i.e., state-wide population structure. The highest resolution ddRAD data (Figure 4B) was concordant with the prior findings (i.e., minimized cross-validation error at *K*=8).

3.2 | Geolocation of Samples via Machine Learning

Determining 'point of origin' (also referred to as 'geographic provenance') is a common question in wildlife conservation, management, and forensics (e.g., Ogden & Linacre 2015). Recently developed analytical approaches (Battey et al. 2020) combined with population genomic data that encapsulate genetic variation at thousands of loci facilitate this approach at an unprecedented spatial resolution. However, the spatial and genomic density of sampling required for accurate predictions often precludes its application at scale for wildlife studies (Chafin et al. 2021; Douglas et al. 2018, 2019, 2022).

Accurately predicting the likely geographic origin of a CWD+ deer is essential to reliably inform wildlife management. We developed **GeoGenIE** (Geographic-Genetic Inference Engine), an AI model and software package based on the framework established in LOCATOR (Battey et al. 2020). GeoGenIE and LOCATOR feature deep learning AI models that predict geographic localities from genetic datasets. These models are particularly beneficial for species with high mobility, those affected by human activities like long-distance translocations, or cases where sample collection locality data are inaccurately recorded. They enable the inclusion of important samples with uncertain or missing locality information in further analyses, which could otherwise generate misleading results.

However, error estimates in our LOCATOR analyses for White-tailed Deer in Arkansas were often substantially higher than expected (Douglas et al. 2022). We attributed this in part to our uneven sampling distribution (i.e., an over-representation of samples from the Management Zones). This issue could be compounded if a limited number of loci is genotyped, as is the case with our SNP GT-seq assay; tests using LOCATOR suggested thousands of loci are needed for accurate locality predictions (Battey et al. 2020). Therefore, we incorporated several optimizations and additional features in GeoGenIE to counter the sampling imbalance and enhance prediction performance for GT-seq data.

GeoGenIE was developed with five primary objectives: (i) optimize a deep learning model based on LOCATOR for our GT-seq panel's loci; (ii) implement strategies to address imbalanced sampling efforts; (iii) generate informative visualizations and statistics, enabling users to identify regions with varying model

performance and quantitatively evaluate model effectiveness; (iv) introduce automated parameter searches to identify the best combination of settings for optimal model performance; and (v) proactively detect and exclude genetic samples that are inconsistent with other geographically neighboring samples or do not adhere to isolation-by-distance patterns (Chang et al. 2023), particularly considering the translocation history in Arkansas (Chafin et al. 2021). This fifth objective is crucial for removing potentially translocated individuals, or those that strongly reflect past translocations in their genotypes, and unusual long-distance dispersers because the genetic signal of such individuals contradicts model assumptions that are based on expected patterns under natural deer dispersal, i.e., isolation-by-distance signals of genetic similarity: individuals nearby are assumed to be genetically more similar than geographically distant individuals. By removing such 'outlier' data, the spurious signal is removed and thus the influence of 'noise' is reduced on the model's learning process. For a more comprehensive explanation of GeoGenIE's underlying model, refer to Appendix 2.

3.3 | Landscape Resistance to Deer Movement

We developed **ResistanceGenIE**, a user-friendly automated workflow to simplify resistance modeling analysis, including data pre-processing steps. As input, ResistanceGenIE requires a genotype database (standard VCF file as described above) with relevant sample metadata (e.g., coordinates and population assignments). It also requires GIS layers clipped to a shared spatial extent, sampled to a concordant spatial granularity, and output as modified files for any necessary downstream plotting. For our White-tailed Deer, we used layers representing land-cover (National Landcover Database 2021), major highways, major rivers (Stream Order >3), large waterbodies, and a digital elevation model. Shapefiles for linear or polygon features (i.e., highways, rivers, waterbodies) were converted to raster format. All layers were resampled for resolution scaling using bilinear interpolation (for continuous features), except NCLD-2021 (land-cover), which was resampled using a nearest neighbors algorithm due to its categorical features.

Outlier detection is integrated using the GGOutlieR algorithm (Chang & Schmid 2023). We first applied Principal Coordinates of Neighbor Matrices (PCNM) to model genetic resistance surfaces to construct spatial axes capturing the relationships among sample locations (Borcard et al. 2004). PCNM involves computing a geographic distance matrix, followed by a principal coordinates analysis (PCoA) to generate eigenvectors representing the spatial arrangement of samples, thus capturing spatial autocorrelation in the dataset. For our White-tailed Deer analysis, we retained half of the positive eigenvalues as variables for downstream analysis (Manel et al., 2010).

ResistanceGenIE next uses these spatial vectors, alongside candidate environmental predictors, in the gradient forest method [(VanHove & Launey 2023) (ResGF)]. The gradient forest approach employs regression trees to partition the data based on environmental gradients, with splits intended to capture

changes in allelic frequencies across the landscape. The predictive performance of each SNP locus is evaluated using the out-of-bag proportion (R^2), providing a cross-validated estimate of the generalization error. The output was a (Figure S2) surface representing the combined environmental resistance to individual movement, with the relative importance of each predictor ranked with R^2 (Figure S3)

As a final step, RESISTANCEGENIE infers 'population connectivity' on the estimated resistance surface output by RESGF as a 'predicted gene flow' graph (Dyer & Nason 2004). Genetic data are first used as input (PEGAS: Paradis 2010), to compute a pairwise genetic distance matrix - the proportion of shared alleles (GRAPH4LG: Savary et al. 2020). Mean pairwise geographic and genetic distances between samples among adjacent Arkansas counties are calculated using the GDISTANCE package (van Etten 2017). These distances are then used in an iterative process (van Strien et al. 2015) to find the maximum landscape distance where the correlation between geographic and genetic distances is strongest. This threshold is used to refine the population connectivity visualization by pruning 'predicted gene flow' and only leaving edges (=blue lines) that represent Arkansas counties with mean genetic similarities and geographic proximities below the threshold, i.e., indicating population connectivity).

4 | RESULTS

4.1 | Genotyping Assay: SNP GT-seq Panel

We developed a genotyping assay that facilities standardized screening of informative SNP variation via GT-seq (Campbell et al. 2015). The SNP GT-seq panel screens *N*=441 SNP loci: *N*=436 autosomal, *N*=2 PRNP, and *N*=3 sex-linked markers. The SNP loci are distributed throughout the genome of White-tailed Deer (Figure 3) and thus effectively encapsulate the genomic variation with only a small subset of markers. **The SNP genotypes contain sufficient genetic information to: (i) assign a sample to a genetic population and infer its likely geographic origin when compared to a state-wide genotype database; (ii) assess the *PRNP* gene variants of individuals and populations; (iii) identify the sex of an individual sample.**

This panel was tested and validated by genotyping *N*=192 individuals (Tables S1 and S2); addition of these samples increased the state-wide database to a total *N*=1,477 spatially referenced White-tailed Deer genetic samples collected in Arkansas (Figure 2). On average, each individual was successfully genotyped at 95.9% of the SNP loci with the GT-seq approach, but the genotyping success rate ranged from 0%–97.7% due to poor sequencing results for nine individuals. **The three sex-linked SNP loci consistently identified the correct genotype (XX or XY), indicating female or male for all individuals. In every case, the genotype was consistent with the known sex of the individual.**

The concordance of SNP genotypes produced via SNP GT-seq *versus* the ddRAD-seq approach was high and consistent (91.4% ± 3.02%). The primary reason for discordance was missing data in poor-quality samples (i.e., those with a genotyping failure rate). Ensuring high-quality DNA is obtained from tissue samples minimizes genotyping failures. **Overall, the SNP GT-seq genotyping assay is reliable and consistently produces high-quality data**.

Visualizations of spatial genetic structure in White-tailed Deer across Arkansas revealed similar patterns between in analyses based on the reduced set of SNP loci contained in the GT-seq assay (Figure 4C) and the full set of SNP loci based generated by ddRAD-seq (Figure 4B) and were consistent with patterns reported in our previous study (Figure 4A; Chafin et al. 2021). Analysis of the highest-resolution ddRAD dataset indicated eight (*K*=8) genetic populations of White-tailed Deer in Arkansas, with visualization of the ddRAD data subset to the 436 target loci showing similar spatial distribution when mapping the *K*=8 model (although note cross-validation was minimized at *K*=3; Figure 4C).

> *Note: When samples genotyped directly with the GT-seq panel were included in the ancestry analysis, they formed a 'pseudo-population' (K=4 in Figure 4D), which was likely due to artifacts caused by the very low proportions of missing data of those individuals compared to the others genotyped only with ddRAD-seq which is prone to higher (random) proportions of missing data (but yields ~100X more*

loci). This 'pseudo-population' is absent when the newly sequenced individuals are not included (Figure 4C). We expect this artifact would not exist if all samples were sequenced with the GT-seq panel.

This comparison demonstrates that the SNP GT-seq genotyping assay has the potential to generate very similar results with a much smaller amount of informative SNP data and is thus effective at capturing relevant genetic information to model gene flow as a proxy for predicting deer dispersal and hence potential spread of CWD across Arkansas.

4.2 | Geolocation of Samples via Machine Learning: GeoGenIE

Our novel approach for predicting 'geographic origin' (i.e., geolocation modeling) was implemented in the GeoGenIE software. It improved geolocation accuracy and was effective with the reduced number of SNP loci genotyped by our GT-seq assay (Table 1, Figure 5).

GeoGenIE's prediction error was significantly lower than that of the original LOCATOR software (Figure 5) when applied to our dataset of *N*=1,415 samples of White-tailed Deer from Arkansas (Figure 2) and genotyped across the *N*=436 autosomal SNP loci included in the GT-seq assay (Figure 3). Prediction error was reduced by approximately 2.62-fold for the mean, 5.77-fold for the median, and 1.52-fold for the standard deviation (Table 1) using the optimized model in GeoGenIE.

GeoGenIE also produces results with a substantial reduction in spatial bias, with accuracy distributed more evenly across the sampling area rather than confined to areas having dense sampling (Figure 6). This is also evidenced by a smaller negative correlation between sampling density and prediction error (GeoGenIE Pearson's R=-0.16, P=0.03; LOCATOR Pearson's R=-0.44, P<0.0001). **As a result, GeoGenIE can be deployed at scale, such as across the entire state of Arkansas, at reduced sampling density and consistently generate accurate predictions.**

This effect is particularly apparent in side-by-side comparisons of geolocator predictions derived using GeoGenIE *versus* LOCATOR (Figures 7 and 8). A visual indicator of this performance is the contours (circles) containing a % of the individual bootstrap replicate predictions: they are generally smaller for GeoGenIE than LOCATOR. **This increased accuracy leads to more reliable and informative geolocation predictions** (i.e., the 'predicted location' or centroids summarized across the bootstrap replicates)**. Consequently, this increases confidence in inferred geographic locations of a specific sample of interest and potential management decisions based on the modeled prediction.**

Model predictions have inherent uncertainties, and results are influenced by various factors, including input data and user-selected settings. Predictions for samples of high interest (e.g., CWD-positive samples from Randolph County collected during the FY2023 surveillance efforts) resulted in predictions with high uncertainty (Figure 8). **Unfortunately, prediction error is high for all Randolph County samples** (Figure 6), which could reflect high ancestry in an unsampled population outside of the study area (e.g., Missouri). **Obtaining samples from adjacent counties in other states to create a 'buffer zone' around Arkansas should be considered as a future project objective to further refine predictions.**

To provide flexibility for users and accommodate potential limitations in specific datasets, GEOGENIE enhancements focus on optimizing model predictions through (i) sample weighting, (ii) minimizing the effects of over- and under-sampling, and (iii) outlier detection and removal (e.g., *N*=29 'outlier' samples with aberrant genotypes in our data set). Each of these improvements contributed to reducing prediction error, though some had only a marginal impact. The most substantial error reduction was driven by synthetic oversampling and sample stratification, which highlights the importance of high-quality training data for model generalization. GEOGENIE is comparable to LOCATOR in computational efficiency for individual replicates, but real runtime is faster through parallel computation.

4.3 | Landscape Resistance to Deer Movement: RESISTANCEGENIE

Landscape resistance modeling with the RESISTANCEGENIE pipeline (implementing RESGF) identified broad-scale patterns of gene flow (i.e., deer dispersal) across Arkansas (Figure 9) that are concordant with previously characterized patterns of population structure (Chafin et al. 2021). High levels of gene flow—indicating high predicted population connectivity (visualized as 'blue lines'; Figure 9, bottom)—were detected within two areas in Arkansas: Northwest and East-Central parts of the state. However, denser sampling in these regions may have biased the model prediction because the procedure generating the population graph relies on an iterative process.

The relative importance of environmental variables and spatial features (Figure S3) in predicting landscape resistance was strongly impacted by outlier samples (Figure S4). The proportion of variance explained by environmental/spatial features was much higher in the absence of outliers. Without outliers, elevation (DEM) and spatial variables (PCNMs) were by far the strongest predictors, followed in order by land-cover (NLCD 2021), major highways, and major rivers (Figure S4).

Outlier detection had a pronounced impact on model performance to predict population connectivity; gene flow estimates based on an unfiltered dataset that included samples with genotypes reflecting extensive translocation histories, rather than natural dispersal (as per Chafin et al. 2021), resulted in spatially inconsistent predictions, rendering the visualization useless (i.e., 'blue blob' in Figure S4).

5 | DISCUSSION

Genetic approaches are essential tools in contemporary wildlife management and conservation. Because of the capacity of DNA to record population processes over time (i.e., 'DNA as an archive of population history,' Douglas & Douglas 2010), genetic data provide insights over various temporal and spatial scales (Douglas et al. 2006, 2016) and thus complement insights generated through ecological approaches. For example, genetic data provide a means to predict the likely geographic origin of a sample (i.e., geographic provenance; Ogden & Linacre 2015), detect hybridization (Bangs et al. 2020, Chafin et al. 2019, Martin et al. 2020, Zbinden et al. 2023), or quantify how environmental changes shape movement (Douglas et al. 2009, Epps & Keyghobadi 2015).

Our previous research on White-tailed Deer in Arkansas established a statewide genetic database to inform the management of the species and assist in risk assessment of CWD spread (Douglas et al. 2018, 2019; Chafin et al. 2020). But these data also reflected how past management actions, such as translocation, obscured 'natural' genetic patterns in the species, complicating interpretations of genetic data (Chafin et al. 2021). This project builds on this prior research but focuses on enhancing methodological tools to simplify genotyping and provide statistically robust predictive modeling tools to facilitate long-term genetic monitoring of White-tailed Deer and inform risk-assessment of CWD spread in Arkansas. These efforts addressed key objectives: (i) develop a standardized genotyping assay; (ii) mitigate intrinsic bias in the data by detecting and removing spurious signals to improve the accuracy of model predictions through novel computational approaches; and (iii) generate user-friendly workflows and establish best practices.

5.1 | Genotyping Assay Development: SNP GT-seq Panel

One of the key objectives was to design and validate a user-friendly genotyping assay by developing a SNP GT-seq panel (Campbell et al. 2015). This technology differs from our previous approach (e.g., Douglas et al. 2018, 2019, 2022) by identifying a subset of SNPs that are most informative and genotyping samples through targeted amplification rather than sequencing thousands of variable SNPs and subsequently extracting relevant information through bioinformatic processing of sequences (as done in ddRAD-seq). Because of the targeted, stream-lined approach, SNP GT-seq panels are ideal for standardizing recurring genotyping needs involving large numbers of individuals (e.g., stock assessment in fisheries; Meek & Larson 2019).

We designed a SNP GT-seq assay for White-tailed Deer in Arkansas by selecting *N*=441 validated SNPs from our statewide dataset of over 1 million SNPs, retaining SNP loci that maximize information content (Figure 3) while substantially decreasing sequencing and data processing efforts required to obtain the data. Our SNP GT-seq assay consistently recovers broad-scale patterns of population structure (Figure 4).

We observed a high genotyping success rate (95.9% on average) and 91.4% (CI ± 3.02%) concordance with ddRAD data in redundantly genotyped samples, affirming the panel's reliability and efficiency. This SNP GT-seq assay offers several advantages for long-term monitoring, including a lower sample failure rate (i.e., <7% *versus* >15%) and a potential for a 75% reduction in per-sample genotyping costs post-panel development and DNA extraction (note: per-sample costs can be drastically reduced when genotyping larger batches of samples, i.e., several hundred samples).

A benefit of a custom-built SNP GT-seq assay is the ability to incorporate additional functionalities by including objective-specific genetic markers, such as sex identification (e.g., Li et al. 2023, May et al. 2020). This flexibility of GT-seq makes it feasible to design custom panels incorporating SNP markers tailored to specific conservation and management questions. Our SNP GT-seq assay includes two SNP loci to genotype relevant variation in the *PRNP* gene (Chafin et al. 2020), as well as three sex-linked loci, to identify the sex of a sample. All genotype-derived sex predictions were consistent with the known sex of samples.

In addition, we designed our SNP GT-seq assay with the enhanced capacity to recover 'microhaplotypes' through sequencing longer, paired-end reads; microhaplotypes can yield greater statistical resolution, such as estimating relatedness amongst samples, though at a trade-off of higher per-sample data sequencing costs (Osborne et al. 2022). Because of their potential to estimate relatedness in wild populations, microhaplotypes are increasingly recognized to inform management of fish and wildlife populations (Delomas et al. 2023). While much of the commonly applied analytical infrastructure used in wildlife genetics lacks explicit compatibility with microhaplotype data, this design consideration provides a degree of 'future-proofing' for the developed assay.

5.2 | Software and Method Development: Enhancements in GeoGenIE + ResitanceGenIE

The statewide White-tailed Deer SNP genotype dataset presents several challenges stemming from uneven sampling densities among areas and species' management history in Arkansas. These issues necessitated the development of novel analytical approaches and their implementation in software, as outlined below.

5.2.1 | Detecting and Mitigating Bias

White-tailed Deer genomes retain the genetic signatures of historical processes, including human-mediated demographic collapse (e.g., Kessler & Shafer 2024) and admixture from restorative translocation efforts (Chafin et al. 2021). These patterns violate fundamental assumptions in landscape genetic analyses due to two artificial signals: (i) geographically distant individuals having high relatedness

due to common descent from translocated ancestors; and (ii) geographically proximate individuals exhibiting inflated genetic divergence due to bottleneck effects.

For White-tailed Deer in Arkansas, these can be disentangled (Chafin et al. 2021). We developed a diagnostic approach to remove spurious signals in our data based on GGOUTLIER (Chang & Schmid 2023). This method defines geo-genetic outliers using k-Nearest Neighbor (kNN) regression. Removal of outliers increases the accuracy of geographic provenance analysis (Figure 5) and produces more biologically meaningful results in landscape resistance modeling (Figure 9).

A second major type of bias in the statewide White-tailed Deer SNP genotype data set is due to opportunistic sampling, such as hunter-harvest (Hughes et al. 2021). Uneven spatial sampling results in the relative over- and under-representation of regions, while 'clumping' drives autocorrelation of model residuals (Kadmon et al. 2004, Reddy & Davalos 2003). This bias propagates downstream in analyses, skewing uncertainty and inflating confidence in under-performing models (e.g., Veloz 2009). Our results show that prediction error in the neural network-based regression approach correlates with local sampling density. Mitigative approaches developed here significantly reduce this effect in terms of per-sample accuracy (Figure 5) and error distribution across the landscape (Figure 6).

To establish best practices for generating meaningful empirical results, we explored various combinations of methods during development (Figure S2). By far, the most effective among these was using a stratified sampling approach, ensuring that uneven sampling density was accounted for using evenly distributed sample representation among subsets of data used for fitting, testing, and evaluating the model. Other actions, both individually and collectively, reduced model error, though to a lesser degree, including the 'weighting' of model accuracy and the use of 'synthetic oversampling' of areas with sparse sample representation, both of which aim to encourage model emphasis on under-represented areas, as well as the aforementioned outlier removal algorithm. The enhancements applied within our novel software, GEOGENIE, reduced prediction errors by approximately 2.90-fold for the mean, 5.77-fold for the median, and 1.76-fold for the standard deviation compared to LOCATOR (Table 1). Importantly, these improvements were consistently observed across most of the state rather than confined to regions with higher sampling density (Figure 6).

5.2.2 | 'Less is More': Increased Performance from Less Data

A successful outcome of the methodological tools developed in this project is that more accurate predictions for geolocation can be achieved with fewer samples and less data. GEOGENIE can generate a viable geographic representation for a given area with less. For example, applied to our GT-seq dataset of *N*=436 autosomal SNP loci, GEOGENIE outperforms LOCATOR applied to *N*=5,000 ddRAD-derived SNP loci reported in Chafin et al. (2021). Likewise, an analysis of optimal sampling density, as the threshold sampling density after which prediction error plateaus, found a ~5.8-fold decrease in the

number of samples required per county for a minimally viable state-wide reference database in Arkansas. **These enhancements make genotyping and data analyses simpler and more reliable across the entire state but also reduce the per-sample cost of continued genetic monitoring of White-tailed Deer in Arkansas (and elsewhere).**

5.2.3 | Applying Best Practices

Established guidelines were available for developing and testing the SNP GT-seq panel (e.g., Euclide et al. 2022). Other project objectives required implementing and experimentally validating best practices. For example, for GeoGenIE, combining automated hyperparameter optimization with stratified sampling and synthetic oversampling shows the most substantial overall effect.

Outlier removal through geo-genetic *k*-Nearest-Neighbor (*k*NN) regression identified 100% of artificially introduced translocation events in validation experiments, which involved manually assigning incorrect coordinates to a subset of samples, and 29 empirical samples identified as 'outliers.' While the number of outliers removed was modest, a profound impact could be seen in the distribution of contemporary gene flow events identified in the landscape resistance modeling (Figure S4) and on the relative assigned to environmental and spatial variables (Figure S3). **We recommend this as a standard practice for similar analyses in species with histories of translocations.**

Most theoretical models in landscape genetics are based on an 'isolation-by-distance' model of genetic population structure, modulated by environmental characteristics, and thus assume population connectivity to follow a form of 'stepping-stone' structure (e.g., Petkova et al. 2016). ResistanceGenIE leverages recent advances in the statistical frameworks available to perform resistance modeling (e.g., ResGF; VanHove and Launey 2023), and herein we predict the effect of environmental features on dispersal and connectivity for White-tailed Deer, yielding biologically meaningful outputs when combined with our outlier detection methodologies. **Despite these options, users are cautioned to carefully consider assumptions to guide their selection of data partition and parameter settings: The choices we make impact model predictions** (Martin et al. 2021).

In this context, it is essential to distinguish between statistically significant results—or model predictions—and what interpretations are biologically most meaningful. Slight variations in input data and software settings can produce different predictions (e.g., Figures 8 *versus* 9 and 10). **This requires effective communication between researchers developing methods (e.g., software developers) and biologists using the methods to generate 'actionable information' that guides management decisions (**Douglas et al. 2022a).

Our tools to perform geolocation and predict genetic connectivity among populations are all disseminated as open-source resources, including R-packages and bioinformatics pipelines, ensuring these advancements are accessible to the broader scientific community. By providing detailed documentation and user-friendly tools we facilitate the adoption of these methodologies for other regions and species.

5.3 | Implications for White-tailed Deer Management

5.3.1 | Contributions towards a Sustainable Genetic Monitoring Program

Genetic tools are widespread in biodiversity management and conservation (Hoban et al. 2022). However, the rapid pace of technological development and the high implementation costs make application at broad spatial and temporal scales difficult (Taylor et al. 2017). A pragmatic approach maximizes the continued usability of existing repositories (e.g., pre-existing datasets) while minimizing the cost of subsequent data acquisitions and expanding the database (Bertola et al. 2023)—particularly following the initial rounds of infrastructure investment.

With the SNP GT-seq assay developed in this study, we generated a resource that standardizes efficient genotyping and facilitates downstream analysis for consistent genetic monitoring in White-tailed Deer without re-assessment of samples previously assayed using other technologies (e.g., Douglas et al. 2018, 2019, 2022b). This reduces per-sample costs for data generation and integration of additional samples with the existing Arkansas database, particularly when scaled-up to process larger batches of samples (e.g., 200+ samples/year). Furthermore, the **SNP GT-seq assay enhances reproducibility and bolsters data integration opportunities across agencies, studies, and regions.**

Despite the widespread adoption of genetic tools in wildlife sciences, there remains, to some degree, a 'research-implementation gap' (Bourret et al. 2020). In addition to the importance of partnerships such as the one upon which the present project builds (e.g., Chafin et al. 2020, 2021, Douglas et al. 2018, 2019, 2022b), narrowing this gap requires that method development—often driven by theoretical motivations tailored to particular research questions—be aligned more directly with management needs (Merkle et al. 2019). Although tools for conservation and management are frequently published, their uptake rate in applied work is demonstrably low (Tkach & Watson 2023). **GeoGenIE is a successful example of development through collaborative partnerships—research priorities were developed through multiple iterations of management-centered use of geographic provenance analysis** (Douglas et al. 2018, 2019, 2022b), **with further inquiry identifying problems with current approaches, rooted in evolutionary theory** (Chafin et al. 2021).

5.3.2 | Refined Landscape Genetic Interpretations

Our prior research identified genetics signals in White-tailed Deer in Arkansas landscape that reflected historic population processes (Chafin et al. 2021). These distinct genetic patterns were shaped by natural population fluctuations and historic management actions. For example, re-expansion of populations following a bottleneck in the early 20th Century (Holder, 1951) resulted in rapid transitions in genetic ancestry on secondary contact (e.g., Excoffier & Ray 2008, Zellmer & Knowles 2009). While historic, such population processes generate distinct patterns in genetic structure still present today (considered artifacts) and need to be accounted for in statistical analyses. Notably, artifacts of the same process have been identified as a range-wide characteristic for White-tailed Deer (Kessler & Shafer 2024) and other cervid species (Burgess et al. 2023). Using appropriate statistical procedures, we were able to distinguish such signals from environmentally-mediated patterns, particularly for White-tailed Deer populations north of the Arkansas River in the Ozark Mountains.

A second type of artifact involved both inflated local genetic divergence and an apparent pattern of high similarity among geographically distant samples (Chafin et al. 2021). These were attributed to several historical translocation events from in-state and out-of-state sources (e.g., Holder 1951, Karlin et al. 1989, Wood 1944). Several 'genetic populations' lacked spatial cohesion or clear biogeographic/ environmental correlation (Figure 4). Thus, we concluded that signals of historic translocations reflected in our statewide genotyping data would violate assumptions of 'link-based' landscape resistance models and limit statistical power of standard landscape genetic approaches, as has been documented in other studies as well (Budd et al. 2018, Leberg et al. 1994, Leberg &Ellsworth, 1999). Despite these limitations, in our previous analyses, we were able to show the presence of major rivers as semipermeable dispersal barriers, as well as a possible effect linked to urbanization/land use (Chafin et al. 2021), conclusions concordant with other studies of differing spatial scales (e.g., Kelly et al. 2014, Locher et al. 2015, Miller et al. 2020, Robinson et al. 2012).

In the current study, we have combined alternative statistical frameworks for resistance modeling (VanHove & Launey 2023) with novel data processing algorithms (Chang & Schmid 2023) into an automated workflow, RESISTANCEGENIE, that facilitates landscape genetics in White-tailed Deer. **Removing aberrant signals increased the correlation of environmental predictors with genetic data** (Figure S4), **with several essential features in the landscape becoming apparent in the higher-resolution composite resistance model** (Figure 9). These include Crowley's Ridge, the elevational gradient of the Mississippi River valley transition from Ozark and Ouachita Mountains, and land-cover categorizations associated with both the Mississippi River valley and the Ouachita-Saline Rivers. The Arkansas and Red rivers were also evident as dispersal barriers, although to lesser degrees. These results are consistent with well-established physiographic boundaries and drivers and concordant with the known biogeography of Arkansas (e.g., Heidt et al. 1996).

The composite resistance surface (Figure 9) also corroborates the population genetic inferences by Chafin et al. (2021), underscoring how analytical artifacts are due to natural and human-induced population processes.

However, extreme uneven sampling in the southern part of Arkansas could not be completely mitigated by model settings, and caution is warranted when interpreting model predictions for that part of the state (Figure 9). Model predictions show very strong landscape resistance associated with river basins in southern Arkansas, which is inconsistent with empirical observations (pers. Comm. Chris Middaugh, AGFC). Extreme dense clustering of samples in Union County (stemming from a targeted removal effort) in an otherwise very sparsly region of the state (few samples from surrounding areas) likely induced artifacts, that, in turn, influenced estimates of relative contribution of variables (Figure S3). For example, NLCD 2021 map in Figure S3 reflects these basins as high resistance, but the NLCD variable did not contribute much to the model (Figure S4). **This underscores the importance of effective communication among modelers, geneticist and deer biologist to scruitinze model assumptions and predications.** In our case, while model predications are consistent with ecological knowledge of deer in Arkansas for areas with robust sampling, areas with uneven sampling might cause erroneous model performance and needs to be further explored.

The landscape genetic inferences provided in this study are in agreement with and further support the established Deer Management Units (DMUs) in use by AGFC (Meeker et al. 2019). **The software developed in this project provide a robust framework for landscape resistance inference in White-tailed Deer, albeit with caveats and a clear need for further development, they offer additional tools alongside disease surveillance data. Recently developed statistical methods** (Walter et al. 2024) **will allow epidemiological and environmental models to effectively leverage population genomic data for more accurate modeling CWD spread, risk assessment, and science-based management.**

6 | ACKNOWLEDGMENTS

Funding for this study was provided by USDA Award FAIN:AP22WSNWRC00C043 (Project ID APP-20248) by the Arkansas Game and Fish Commission (AGFC) to MRD, MED, TKC, and ZDZ. Supplemental funding was provided by generous endowments to the University of Arkansas: The Bruker Professorship in Life Sciences (MRD), the 21st Century Chair in Global Change Biology (MED). In addition, TKC was funded by the Scottish Government's Rural and Environmental Science and Analytical Services Division (RESAS). We thank the University of Arkansas and Arkansas High-Performance Computing Center for providing the computational resources that enhanced our research capabilities.

7 | DATA AND CODE AVAILABILITY STATEMENT

Data and code will be publicly released following procedures outlined in established agreements between the University of Arkansas and the Arkansas Game and Fish Commission, followed by an embargo period allowing for any publication review and release.

Code for reproducing all analyses presented herein will be made available via GitHub:

- GeoGenIE: https://github.com/btmartin721/GeoGenIE
- gtseq2vcf: https://github.com/btmartin721/gtseq_converter
- ResistanceGenIE: https://github.com/btmartin721/ResistanceGenIE

User-friendly manuals for software will be made available and disseminated via public repositories, too.

- GeoGenIE User Manual: https://github.com/btmartin721/GeoGenIE
- ResistanceGenIE User Manual: https://github.com/btmartin721/ResistanceGenIE

8 | REFERENCES CITED

Akiba T, Sano S, Yanase T, Ohta T, & Koyama M (2019, July) Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623–2631). https://doi.org/10.1145/3292500.3330701

Alexander DH & Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12:1–6. https://doi.org/10.1186/1471-2105-12-246

Ballard, J., J. Brown, B. Carner, S. Clark, A. Gramza, M. Gray, M. Hutchings, C. Middaugh, R. Meeker, A. Riggs, and W. Wright. 2021. Chronic Wasting Disease Management and Response Plan (2021-2025). Management Plan, Arkansas Game and Fish Commission, Little Rock, USA. DOI: https://doi.org/10.13140/RG.2.2.28573.44003

Bangs MR, Douglas MR, Brunner PC, & Douglas ME (2020) Reticulate evolution as a management challenge: Admixture in endemic fishes of Western North America. *Evolutionary Applications* 13:1400–1419. https://doi.org/10.1111/eva.13042

Battey C J, Ralph PL & Kern AD (2020) Predicting geographic location from genetic variation with deep neural networks. *Elife 9:*e54507. https://doi.org/10.7554/eLife.54507

Bertola LD, Bruniche-Olsen A, Kershaw F, Russo I-RM, MacDonals AJ, Sunnucks P, Bruford MW, Cadena CD, Ewart KM, de Bruyn M, Eldridge MDB, Frankham R, Guayasamin JM, Brueber, CE, Hoareau, TB, Hoban, S, Hohenlohe, PA, Hunter, ME, Kotze, A, Kuja, A, Kuja J, Lacy RC, Laikre L, Lo N, Meek MH, Mergeay J, Mittan-Moreau C, Neaves LE, O'Brien D, Ochieng JW, Ogden R, Orozco-terWengel P, Paez-VacasM, Pierson, J, Ralls K, Shaw RE, Sogbohossou EA, Stow A, Steeves T, Vernesi C, Watsa M & Segelbacker G(2023) A pragmatic approach for integrating molecular tools into biodiversity conservation. *Conservation Science and Practice* 6(1):e13053. https://doi.org/10.1111/csp2.13053

Borcard D, Legendre P, Avois-Jacquet C & Tuomisto H (2004) Dissecting the spatial structure of ecological data at multiple scales. *Ecology* 85(7):1826–1832. https://doi.org/10.1890/03-3111

Bourret V, Albert V, April J, Cote G & Morissette O (2020) Past, present and future contributions of evolutionary biology to wildlife forensics, management and conservation. *Evolutionary Applications* 13(6):1420–1434. https://doi.org/10.1111/eva.12977

Budd K, Berkman LK, Anderson M, Koppelman J & Eggert LS (2018) Genetic structure and recovery of White-tailed Deer in Missouri. *Journal of Wildlife Management* 82(8):1598–1607. https://doi.org/10.1002/jwmg.21546

Burgess BT, Irvine RL, Martin J-L & Russello MA (2023) Past population control biases interpretations of contemporary genetic data: implications for future invasive Sitka black-tailed deer management in Haida Gwaii. *Biological Invasions* 25:3871–3886. https://doi.org/10.1007/s10530-023-03145-w

Campbell NR, Harmon S A & Narum SR (2015) Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources* 15(4):855–867. https://doi.org/10.1111/1755-0998.12357

Caye K, Deist TM, Martins H, Michel O & François O (2016) TESS3: fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources* 16(2):540-548. https://doi.org/10.1111/1755-0998.12471

Chafin TK, Douglas MR, Martin BT & Douglas ME (2019) Hybridization drives genetic erosion in sympatric desert fishes of western North America. *Heredity* 123:759–773. https://doi.org/10.1038/s41437-019-0259-2

Chafin TK, Douglas MR, Martin BT, Zbinden ZD, Middaugh CR, Ballard JR, Gray CM, White Jr D & Douglas ME (2020) Age structuring and spatial heterogeneity in prion protein gene (PRNP) polymorphism in White-tailed Deer. *Prion* 14(1):238–248. https://doi.org/10.1080/19336896.2020.1832947

Chafin TK, Zbinden ZD, Douglas MR, Martin BT, Middaugh CR, Gray MC, Ballard JC, & Douglas M E (2021) Spatial population genetics in heavily managed species: Separating patterns of historical translocation from contemporary gene flow in white-tailed deer. *Evolutionary Applications* 14(6):1673–1689. https://doi.org/10.1111/eva.13233

Chang CW & Schmid K (2023) GGOUTLIER: an R package to identify and visualize unusual geo-genetic patterns of biological samples. *Journal of Open Source Software* 8(91):5687. https://doi.org/10.21105/joss.05687

Chawla NV, Bowyer KW, Hall LO, & Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16:321–357. https://doi.org/10.1613/jair.953

Delomas TA, Struthers J, Hebdon T, Campbell MR (2023) Development of a microhaplotype panel to inform management of gray wolves. *Conservation Genetics Resources 15*(1):49–57. https://doi.org/10.1007/s12686-023-01301-x

DeRaad DA (2022) SNPfiltR: an R package for interactive and reproducible SNP filtering. *Molecular Ecology Resources* 22(6):2443–2453. https://doi.org/10.1111/1755-0998.13618

Douglas ME, Douglas MR, Schuett GW & Porras LW (2006) Evolution of rattlesnakes (Viperidae: *Crotalus*) in the warm deserts of western North America shaped by Neogene vicariance and Quaternary climate change. *Molecular Ecology* 15:3353–3374. https://doi.org/10.1111/j.1365-294X.2006.03007.x

Douglas ME, Douglas MR, Schuett GW & Porras LW (2009) Climate change and evolution of the New World pitviper genus *Agkistrodon* (Viperidae). *Journal of Biogeography* 36:1164–1180. https://doi.org/10.1111/j.1365-2699.2008.02075.x

Douglas MR, Davis MA, Amarello M, Smith JJ, Schuett GW, Herrmann H-W, Porras LW, Holycross AT & Douglas ME (2016) Anthropogenic impacts drive conservation and ecosystem management of a niche conserved rattlesnake on the Colorado Plateau of Western North America. *Royal Society Open Science* 3:160047. https://doi.org/10.1098/rsos.160047

Douglas MR & Douglas ME (2010) Molecular approaches to stream fish ecology. pp. 157–195 – *In: Community Ecology of Stream Fishes*. K. Gido and D. Jackson (eds.). *American Fisheries Society Symposium 73.*

Douglas MR, Douglas ME, White D & Chafin TK. 2018. Family group assignment and genotyping of White-tailed Deer in Arkansas using Single Nucleotide Polymorphism (SNP) DNA markers extracted from ear tissue samples. *Final Report submitted to Arkansas Game and Fish Commission.*

Douglas MR, Chafin TK, Zbinden ZD, Martin BT & Douglas ME (2019) White-tailed Deer in Arkansas: genetic connectivity and chronic wasting diseases susceptibility. *Final Project Report to the Arkansas Game and Fish Commission*.

Douglas MR, Bahr K & Olson R (2022a). *The Narrative Gym for Science Graduate Students and Postdocs. Using the ABT Framework for Proposals, Papers, Presentations, and Life in General*. Prairie Starfish Press. October 2022.

Douglas MR, Zbinden ZD, Chafin TK, Martin BT & Douglas ME (2022b) Screening SNP variation to model deer dispersal in Arkansas. *Final Report submitted to the Arkansas Game and Fish Commission*.

Dyer RJ & Nason JD (2004) Population graphs: the graph theoretic shape of genetic structure. *Molecular Ecology* 13(7):1713–1727. https://doi.org/10.1111/j.1365-294x.2004.02177.x

Eaton DA & Overcast I (2020) IPYRAD: Interactive assembly and analysis of RADseq datasets. *Bioinformatics* 36(8):2592–2594. https://doi.org/10.1093/bioinformatics/btz966

Epps CW & Keyghobadi N (2015) Landscape genetics in a changing world: disentangling historical and contemporary influences and inferring change. *Molecular Ecology* 24(24):6021–6040. https://doi.org/10.1111/mec.13454

Euclide PT, Larson WA, Bootsma M, Miller LM, Scribner KT, Stott W, Wilson CC, & Latch EK (2022) A new GTSeq resource to facilitate multijurisdictional research and management of walleye *Sander vitreus*. *Ecology and Evolution* 12(12):e9591. https://doi.org/10.1002/ece3.9591

Excoffier L & Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology and Evolution* 23(7):347–351. https://doi.org/10.1016/j.tree.2008.04.004

Frichot E & François O (2015) LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution* 6(8):925–929. https://doi.org/10.1111/2041-210X.12382

Frichot E, Mathieu F, Trouillon T, Bouchard G & François O (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196(4):973–983. https://doi.org/10.1534/genetics.113.160572

Heidt GA, Elrod DA & McDaniel VR (1996) Biogeography of Arkansas mammals with notes on species of questionable status. *Journal of the Arkansas Academy of Science* 50:12. https://scholarworks.uark.edu/jaas/vol50/iss1/12

Hoban S, Archer FI, Bertola LD, Bragg JG, Breed MF, Bruford MW, Coleman MA, Ekblom R, Funk, WC, Grueber CE, Hand BK, Jaffe R, Jensen E, Johnson JS, Kershaw F, Jiggins L, MacDonals AJ, Mergeay J, Miller JM, Muller-Karger F, O'Brien D, Paz-Vinas I, Potter KM, Razgour O Vernesi C & Hunter ME (2022) *Biological Reviews* 97(4):1511–1538. https://doi.org/10.1111/eva.13233

Holder TH (1951) A survey of Arkansas game. *Arkansas Game and Fish Commission Federal Aid Publication Project II-R*, 57–79.

Hughes AC, Orr MC, Ma K, Costello MJ, Waller J, Provoost P, Yang Q, Zhu C & Qiao H (2021) Sampling biases shape our view of the natural world. *Ecography* 44(9):1259–1269. https://doi.org/10.1111/ecog.05926

Kadmon R, Farber O & Danin A (2009) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* 14:401–413. https://doi.org/10.1111/ecog.05926

Karlin AA, Heidt GA & Sugg DW (1989) Genetic variation and heterozygosity in White-tailed Deer in southern Arkansas. *The American Midland Naturalist* 121(2):273–284. https://doi.org/10.2307/2426031

Kelly AC, Mateus-Pinilla NE, Brown W, Ruiz MO, Douglas MR, Douglas ME, Shelton P, Beissel T & Novakofski J (2014) Genetic assessment of environmental features that influence deer dispersal: Implications for prion-infected populations. *Population Ecology* 56(2):327–340. https://doi.org/10.1007/s10144-013-0427-9

Kessler C & Shafer ABA (2024) Genomic analyses capture the human-induced demographic collapse and recovery in a wide-ranging cervid. *Molecular Biology and Evolution* 41(3):msae038. https://doi.org/10.1093/molbev/msae038

Knaus BJ & Grünwald NJ (2017) vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources* 17(1):44–53. https://doi.org/10.1111/1755-0998.12549

Leberg PL & Ellsworth DL (1999) Further evaluation of the genetic consequences of translocations on southeastern White-tailed Deer populations. *The Journal of Wildlife Management* 63(1):327–334. https://doi.org/10.2307/3802516

Leberg PL, Stangel PW, Hillestad HO, Marchinton RL & Smith MH (1994) Genetic structure of reintroduced wild turkey and White-tailed Deer populations. *The Journal of Wildlife Management* 58(4):698. https://doi.org/10.2307/3809684

Li P, de Groot PVC, Sun Z & Lougheed SC (2023) A new genomics tool for monitoring Arctic char (*Salvelinus alpinus*) populations in the Lower Northwest Passage, Nunavut. *Fisheries Research* 258:106523. https://doi.org/10.1016/j.fishres.2022.106523

Locher A, Scribner KT, Moore JA, Murphy B & Kanefsky J (2015) Influence of landscape features on spatial genetic structure of White-tailed Deer in human-altered landscapes. *Journal of Wildlife Management* 79(2):180–194. https://doi.org/10.1002/jwmg.826

London EW, Roca AL, Novakofski JE & Mateus-Pinilla NE (2022) A *de novo* chromosome-level genome assembly of the White-tailed Deer, *Odocoileus virginianus*. *Journal of Heredity* 113(4):479–489. https://doi.org/10.1093/jhered/esac022

Manel S, Poncet BN, Legendre P, Gugerli F & Holderegger R (2010) Common factors drive adaptive genetic variation at different spatial scales in *Arabis alpina*. *Molecular Ecology* 19:3824–3835. https://doi.org/10.1111/j.1365-294x.2010.04716.x

Martin BT, Douglas MR, Chafin TK, Placyk JS, Birkhead RD, Philips CA & Douglas ME (2020) Contrasting signatures of introgression in North American box turtle (*Terrapene* spp.) contact zones. *Molecular Ecology* 29:186–4202. https://doi.org/10.1111/mec.15622

Martin BT, Chafin TK, Douglas MR, Placyk JS, Birkhead RD, Philips CA & Douglas ME (2021) The choices we make and the impacts they have: Machine learning and species delimitation in North American box turtles *Terrapene* spp. *Molecular Ecology Resources* 21(8):2801–2817. https://doi.org/10.1111/1755-0998.13350

Martins H, Caye K, Luu K, Blum MG & François O (2016) Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Molecular Ecology* 25(20):5029–5042. https://doi.org/10.1111/mec.13822

May SA, McKinney GJ, Hilborn R, Hauser L & Naish KA (2020) Power of a dual-use SNP panel for pedigree reconstruction and population assignment. *Ecology and Evolution* 10(17):9522–9531. https://doi.org/10.1002/ece3.6645

Meek MH & Larson LA (2019) The future is now: Amplicon sequencing and sequence capture usher in the conservation genomics era. *Molular Ecology Resources* 19(4):975–803. https://doi.org/10.1111/1755-0998.12998

Meeker R, Brown J, Carner B, Stephens K, Dugger G, Groves B & White DJ (2019) *Arkansas Game and Fish Strategic Deer Management Plan*. Arkansas Game and Fish Commission.

Meirmans PG (2012) The trouble with isolation by distance. *Molecular Ecology* 21(12):2839–2846. https://doi.org/10.1111/j.1365-294X.2012.05578.x

Merkle JA, Anderson NJ, Baxley DL, Chopp M, Gigliotti LC, Gude JA, Harms TM, Johnson HE, Merrill EH, Mitchell MS, Mong TW, Nelson J, Norton AS, Sheriff MJ, Tomaski E & VanBeek KR (2019) A collaborative approach to bridging the gap between wildlife managers and researchers. *The Journal of Wildlife Management* 83(8):1644–1651. https://doi.org/10.1002/jwmg.21759

Miller WL, Miller-Butterworth CM, Diefenbach DR & Walter WD (2020) Assessment of spatial genetic structure to identify populations at risk for infection of an emerging epizootic disease. *Ecology and Evolution* 10(9):3977–3990. https://doi.org/10.1002/ece3.6161

Mijangos JL, Gruber B, Berry O, Pacioni C & Georges A (2022) DARTR v2: An accessible genetic analysis platform for conservation, ecology and agriculture. *Methods in Ecology and Evolution* 13(10):2150–2158. https://doi.org/10.1111/2041-210X.13918

Mussmann SM, Douglas MR, Chafin TK & Douglas ME (2023) ADMIXPIPE v3: facilitating population structure delimitation from SNP data. *Bioinformatics Advances* 3(1):vbad168. https://doi.org/10.1093/bioadv/vbad168

Ogden R & Linacre A (2015) Wildlife forensic science: A review of genetic geographic origin assignment. *Forensic Science International: Genetics* 18:152–159. https://doi.org/10.1016/j.fsigen.2015.02.008

Osborne MJ, Caeiro-Dias G & Turner TF (2023) Transitioning from microsatellites to SNP-based microhaplotypes in genetic monitoring programmes: Lessons from paired data spanning 20 years. *Molecular Ecology* 32(2):316–334. https://doi.org/10.1111/mec.16760

Paradis E (2010) Pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 26:419-420. https://doi.org/10.1093/bioinformatics/btp696

Petkova D, Novembre J & Stephens M (2016) Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics* 48(1):94–100. https://doi.org/10.1038/ng.3464

R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Reddy S & Davalos LM (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography* 30:1719–1727. https://doi.org/10.1046/j.1365-2699.2003.00946.x

Robinson SJ, Samuel MD, Lopez DL & Shelton P (2012) The walk is never random: Subtle landscape effects shape gene flow in a continuous White-tailed Deer population in the Midwestern United States. *Molecular Ecology* 21(17):4190–4205. https://doi.org/10.1111/j.1365-294x.2012.05681.x

Rousset F (1997) Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics* 145(4):1219–1228. https://doi.org/10.1093/genetics/145.4.1219

Savary P, Foltete,J-C, Moal H, Vuidel G & Garnier S (2020) Graph4lg: A package for constructing and analysing graphs for landscape genetics in R. *Methods in Ecology and Evolution* 12(3):539–547. https://doi.org/10.1111/2041-210X.13530

Strickland BK, Demarais S, Zamorano A, DeYoung RW & Dacus CM (2011) Accuracy for determining sex of white-tailed deer fetuses. *Wildlife Society Bulletin 35*(2):54–58. https://doi.org/10.1002/wsb.18

Taylor HR, Dussex N & van Heezik Y (2017) Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Global Ecology and Conservation* 10:231–242. https://doi.org/10.1016/j.gecco.2017.04.001

Tkatch K & Watson MJ (2023) Publication and use of genetic tools in conservation management applications – A systematic review. *Journal of Applied Ecology* 60(8):1522–1536. https://doi.org/10.1111/1365-2664.14433

Van Etten J (2017) R package gdistance: Distances and routes on geographical grids. *Foundation for Open Access Statistics* 76(13):1–21. https://doi.org/10.18637/jss.v076.i13

VanHove M & Launey S (2023) Estimating resistance surfaces using gradient forest and allelic frequencies. *Molecular Ecology Resources,* In press. https://doi.org/10.1111/1755-0998.13778

Van Strien MJ, Holderegger R & Van Heck HJ (2015) Isolation-by-distance in landscapes: considerations for landscape genetics. *Heredity* 114:27–37. https://doi.org/10.1038/hdy.2014.62

Veloz SD (2009) Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography* 29: 2290–2299. https://doi.org/10.1111/j.1365-2699.2009.02174.x

Walter WD, Hanley B, Them CE, Mitchell CI, Kelly J, Grove D, Hollingshead N, Abbott RC & Schuler KL (2024) Predicting the odds of chronic wasting disease with Habitat Risk software. *Spatial and Spatio-temporal Epidemiology* 49:100650. https://doi.org/10.1016/j.sste.2024.100650

Wood R (1944) Arkansas' deer transplanting program. *Transactions of the North American Wildlife Conference* 9:162–167.

Zbinden ZD, Douglas MR, Chafin TK & Douglas ME (2023) A community genomics approach to natural hybridization. *Proceedings of the Royal Society B* 290(1999):20230768. https://doi.org/10.1098/rspb.2023.0768

Zellmer AJ & Knowles LL (2009) Disentangling the effects of historic vs. contemporary landscape structure on population genetic divergence. *Molecular Ecology* 18(17):3593–3602. https://doi.org/10.1111/j.1365-294x.2009.04305.x

9 | GLOSSARY

Ancestry proportion: The estimated fraction of an individual's genome that originates from each of several predefined ancestral populations, typically derived from programs like ADMIXTURE that analyze genetic data to infer population structure and admixture.

Artificial intelligence: A branch of computer science focused on creating systems capable of performing tasks that typically require human intelligence, such as visual perception, speech recognition, decision-making, and language translation.

Bootstrapping: A statistical method that involves repeatedly sampling from a dataset with replacement to estimate the distribution of a statistic and assess the variability or uncertainty of the estimates.

Cross-validation: A statistical technique used in machine learning to assess the performance of a model by dividing the dataset into multiple subsets, training the model on some subsets while validating it on the remaining ones, and averaging the results to obtain a more reliable estimate of model performance.

Deep learning: A subset of artificial intelligence and machine learning that uses neural networks with many layers to model complex patterns in data.

ddRAD: Double-digest Restriction-site Associated DNA sequencing, a method used to discover and genotype large numbers of SNPs across the genome.

Gradient forest: A machine learning method to analyze and visualize complex relationships between genetic data and environmental variables.

GT-seq: Genotyping-in-Thousands by sequencing, a cost-effective and high-throughput method for genotyping large numbers of samples using targeted SNP panels.

Hyperparameter (in the context of an MLP model): A configuration parameter set before training a multilayer perceptron (MLP) model that governs the learning process, such as the learning rate, number of hidden layers, and number of neurons per layer, which are not learned from the data.

Isolation-by-distance: A population genetic pattern where the genetic similarity between individuals decreases as the geographic distance between them increases.

K-nearest neighbor (KNN): A non-parametric classification algorithm that assigns a data point to the class most common among its k nearest neighbors in the feature space.

K-means: A clustering algorithm that partitions a dataset into *k* distinct, non-overlapping subsets (clusters) by minimizing the variance within each cluster.

Machine learning: A subset of artificial intelligence that involves training algorithms to recognize patterns in data and make predictions or decisions without being explicitly programmed for each specific task.

Microhaplotype: A small, multi-SNP haplotype that can be used for fine-scale population genetic studies due to its high informativeness.

Multilayer perceptron: A type of artificial neural network composed of multiple layers of nodes (neurons), where each layer is fully connected to the next, used for modeling complex relationships in data.

Neural network: A computational model inspired by the human brain, consisting of interconnected nodes (neurons) that process data in layers to learn patterns and make predictions.

Next-generation sequencing: High-throughput DNA sequencing technologies that allow for the rapid sequencing of entire genomes or targeted regions of the genome.

PCNM: Principal Coordinates of Neighbor Matrices, a method used in spatial ecology to analyze spatial patterns by decomposing spatial relationships into orthogonal components.

Primer: A short single-stranded DNA sequence that provides a starting point for DNA synthesis during PCR amplification.

Sex-linked markers: Genetic markers on sex chromosomes are used to study sex-specific traits and inheritance patterns.

Single nucleotide polymorphism (SNP): A variation at a single position in a DNA sequence among individuals, which can be used as a genetic marker.

SMOTE: Synthetic Minority Over-sampling Technique, a method used in machine learning to address class imbalance by generating synthetic samples for the minority class based on the feature space similarities between existing minority class samples.

sNMF (sparse non-negative matrix factorization): A matrix factorization technique used in population genetics for inferring individual ancestry and admixture proportions, leveraging sparsity to enhance interpretability and computational efficiency.

Spatial autocorrelation: The measure of how much a given variable is correlated with itself in space, indicating how similar or dissimilar points are to each other over a geographic area.

Test/Train/Validation split: A technique in machine learning where the dataset is divided into three subsets: the training set for model training, the validation set for tuning hyperparameters and preventing overfitting, and the test set for evaluating the final model's performance.

TPE (tree-structured Parzen Estimator): A Bayesian optimization algorithm used for hyperparameter tuning in machine learning, which models the objective function using a probabilistic approach and evaluates hyperparameters by constructing a tree-structured density estimator.

VCF (Variant Call Format): A standardized text file format used in bioinformatics for storing gene sequence variations, including SNPs and indels.

10 | TABLES

Table 1. Performance comparisons for GeoGenIE *versus* Locator

Three performance indicators reflecting prediction error [km] of GeoGenIE *versus* the original Locator software in modeling 'origin location' in White-tailed Deer based on genotypes across 436 SNPs generated with the new SNP GT-seq assay. Summary statistics were calculated from 'test sets' of *N*=170 WTD samples randomly selected and not used during model training from a total sample size of *N*=1,415 WTD samples, and summarized across 100 bootstrap replicates. Prediction error represents the Haversine (i.e, Great Circle) distance, in kilometers, between the locality recorded upon tissue collection *versus* the 'predicted origin.' The statistics for GeoGenIE are based on the 'optimal' model as identified from the combination of enabled or disabled settings (e.g., weighted sampling, removing outliers, yielding N=1,386 non-outlier samples, and oversampling) that most effectively improved prediction error.

Summary Statistic	**GeoGenIE** [km]	**Locator** [km]
Mean Error	22.5	59.0
Median Error	5.7	32.9
StdDev Error	43.9	66.9

11 | FIGURES

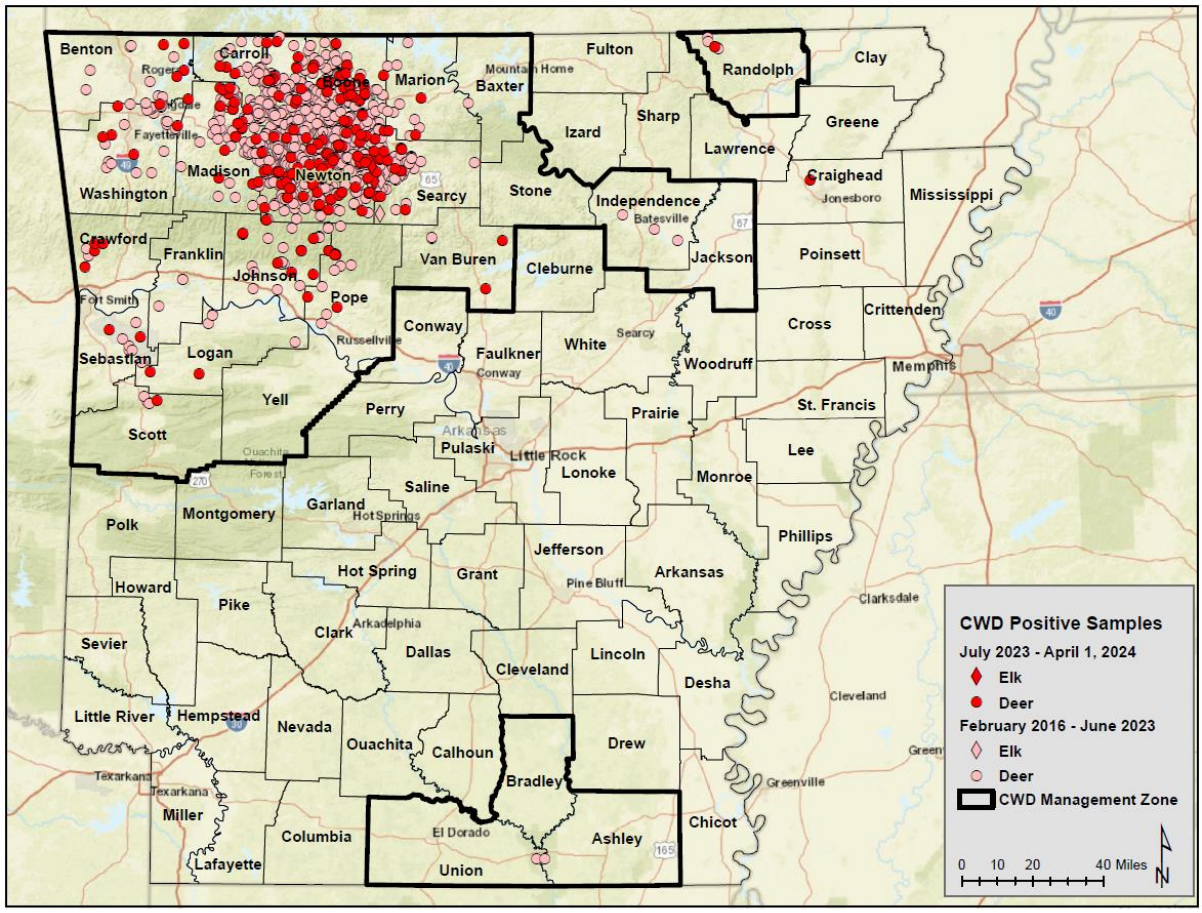

Figure 1. Chronic Wasting Disease (CWD) Management Zone in Arkansas

Status of Chronic Wasting Disease (CWD) detections in Arkansas as of 1-April-2024. The map shows geographic locations of White-tailed Deer (WTD; dots) and Elk (diamonds) that tested positive for CWD. Red: samples collected July 2023 through April 2024; pink = samples collected February 2016 through June 2023. Counties included in the CWD Management Zone (MZ) are outlined in black. From: https://www.agfc.com/hunting/deer/chronic-wasting-disease/cwd-in-arkansas/

(accessed 30 June 2024).

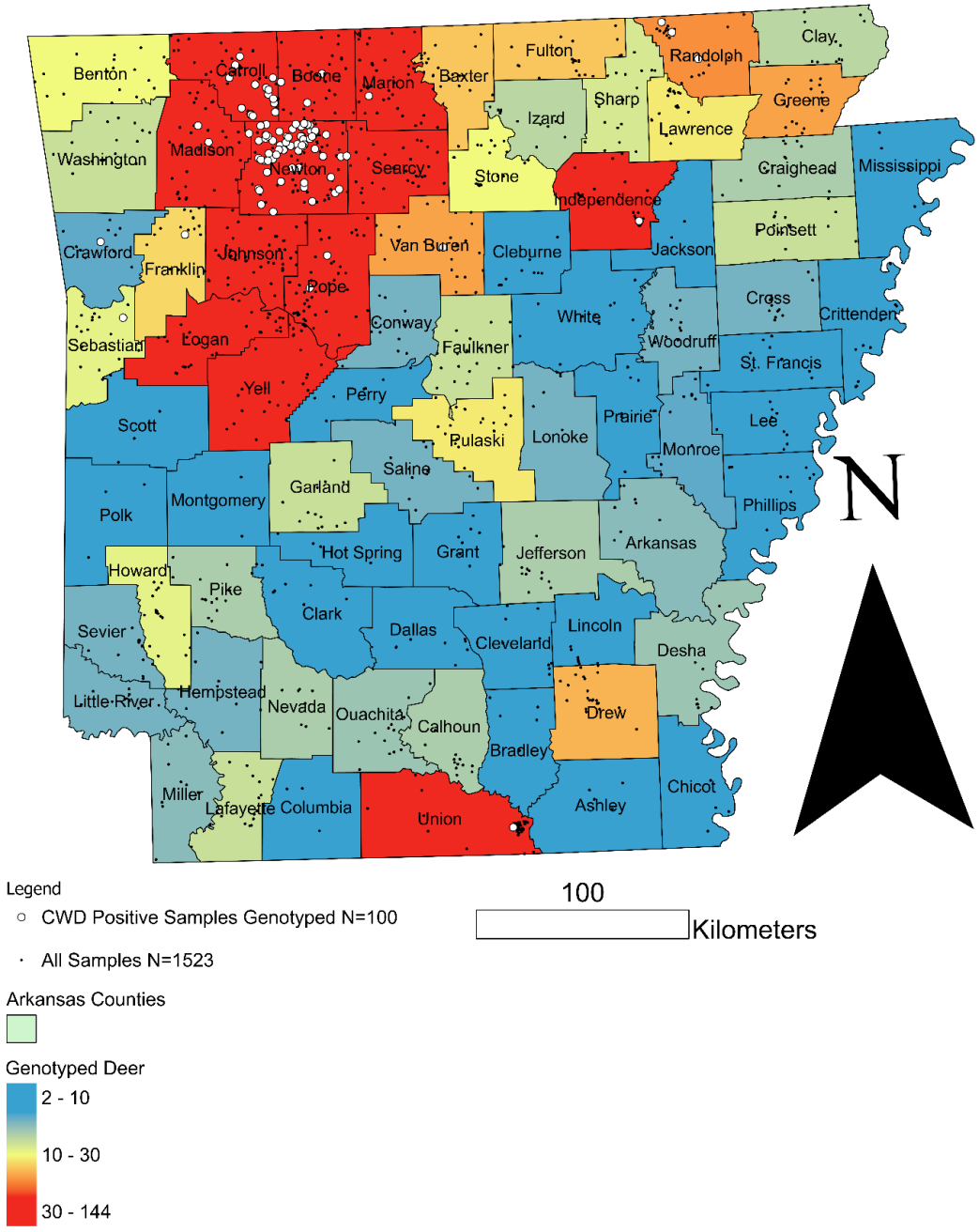

Figure 2. White-tailed Deer Samples by County

Geographic distribution of *N*=1,477 spatially referenced White-tailed Deer sampled across Arkansas and genotyped in this ongoing project. Closed circles: samples that tested negative for Chronic Wasting Disease (CWD-); Open circles: samples that tested positive for CWD. Colors reflect sampling density for each county based on the number of tissue samples representing that county in the analyses: warmer colors equal more genotyped deer).

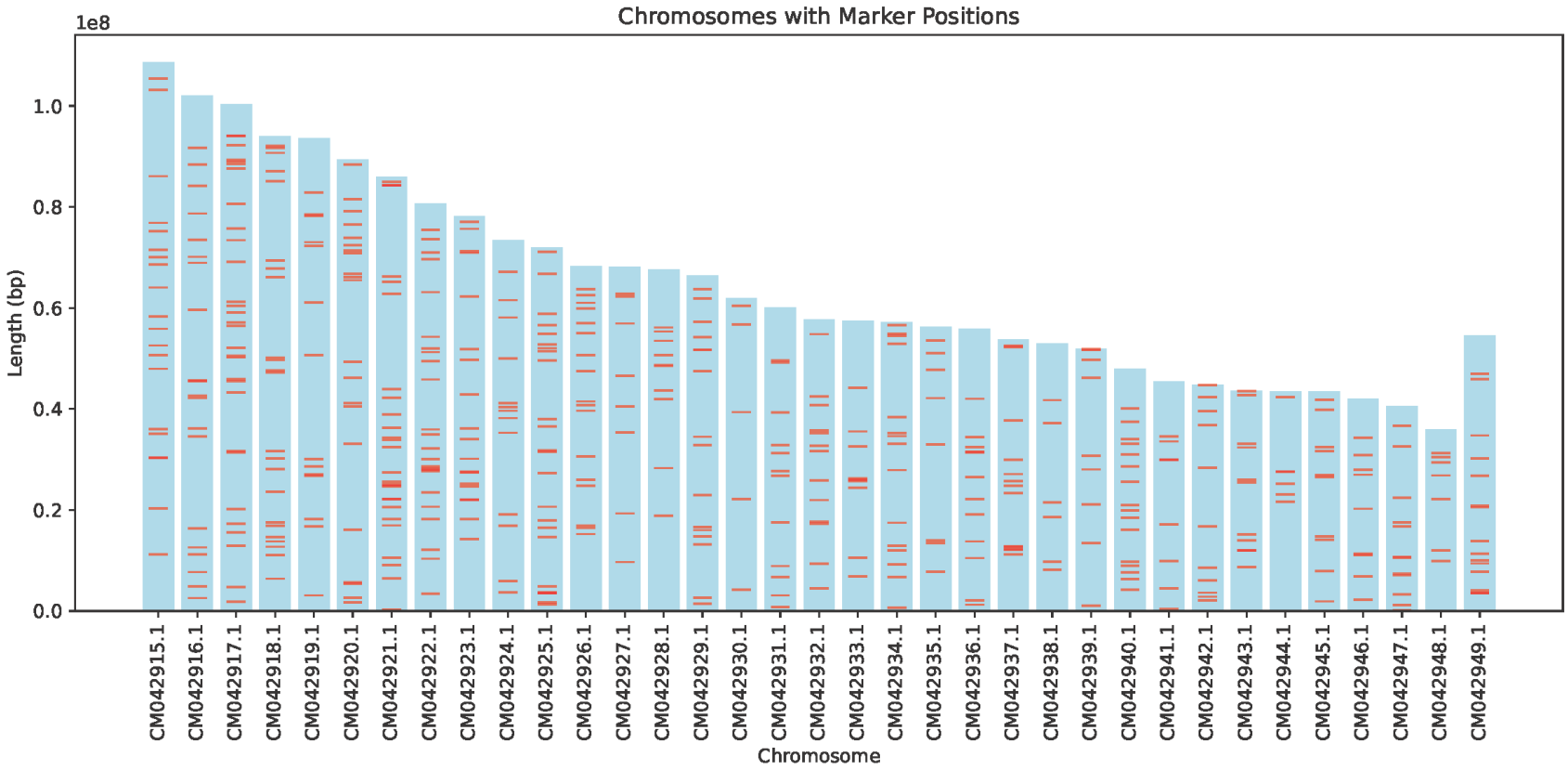

Figure 3. SNP GT-seq Panel Genome Map

Distribution of SNP loci included in the GT-seq genotyping assay across the genome of White-tailed Deer. Depicted are the autosomal chromosomes (blue bars) with positions of the 436 SNP loci denoted (red bars). Length of each chromosome is in base-pairs (y-axis).

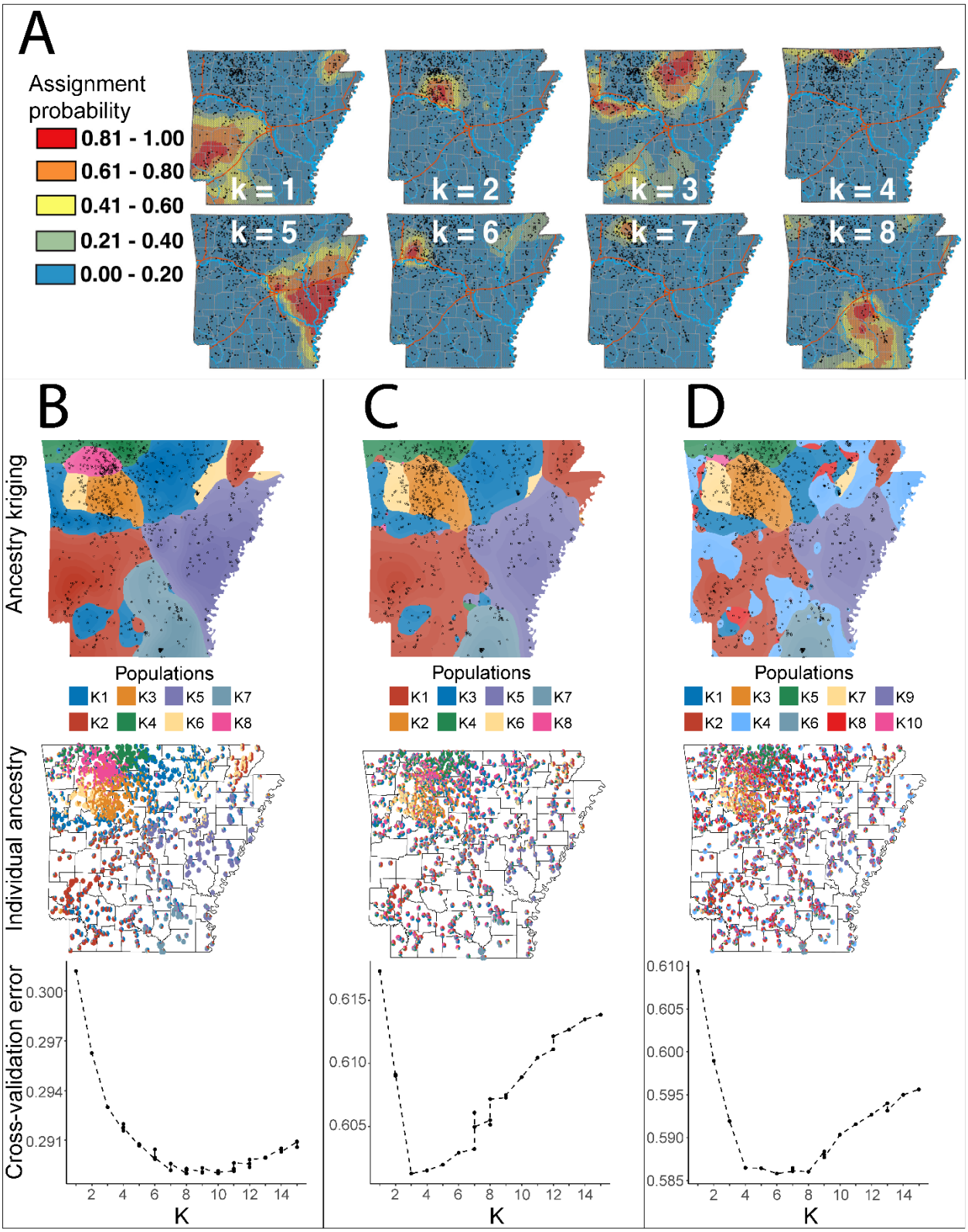
A
Assignment probability
0.81 - 1.00
0.61 - 0.80
0.41 - 0.60
0.21 - 0.40
0.00 - 0.20
k = 1
k = 2
k = 3
k = 4
k = 5
k = 6
k = 7
k = 8
B
C
D
Ancestry kriging
Populations
K1 K2 K3 K4 K5 K6 K7 K8
K1 K2 K3 K4 K5 K6 K7 K8 K9 K10
Individual ancestry
Cross-validation error
0.300
0.297
0.294
0.291
0.615
0.610
0.605
0.610
0.605
0.600
0.595
0.590
0.585
2 4 6 8 10 12 14
K

Figure 4. White-tailed Deer Population Structure Based on Different SNP Datasets

Spatial distribution of Arkansas White-tailed Deer genetic populations based on different sets of SNP genotypes (four sets: A, B, C, D) and determined via ancestry analyses using ADMIXTURE. Spatial distributions are based on the geo-referenced individual ancestry of each sample plotted via spatially interpolated ancestry kriging. Panels are based on different sets of individuals and loci: (A) **ddRAD-seq loci 2019 (original):** *N*=1,143 deer samples genotyped across 35,099 SNP loci via ddRAD-seq (Chafin et al. 2021); (B) **ddRAD-seq loci 2024 (reference aligned):** *N*=1,302 deer samples genotyped across 46,612 SNP loci via ddRAD-seq data and reference-aligned [29 outliers (GeoGenIE) and 146 samples with missing data >90% removed]; (C) **ddRAD-seq SNPs:** *N*=1,203 deer samples subset to just the 436 autosomal loci included in the GT-seq panel (29 outliers, 53 individuals with missing data >75%, and 192 samples with nearly complete GT-seq genotypes removed); and (D) **ddRAD-seq+GT-seq SNPs**: *N*=1,386 deer samples at 436 autosomal loci, including those *N*=183/192 samples genotyped with the new GT-seq panel (29 outliers, 62 samples with missing data >75% removed). For each panel, the number of population clusters, *K*, displayed was chosen based on that which minimized cross-validation error and intending to maximize the comparability among the different data sets (e.g., 'C' cross-validation error is minimized at *K*=3). Panel 'B' represents the highest-resolution population structure to date, and general patterns are consistent with previous results.

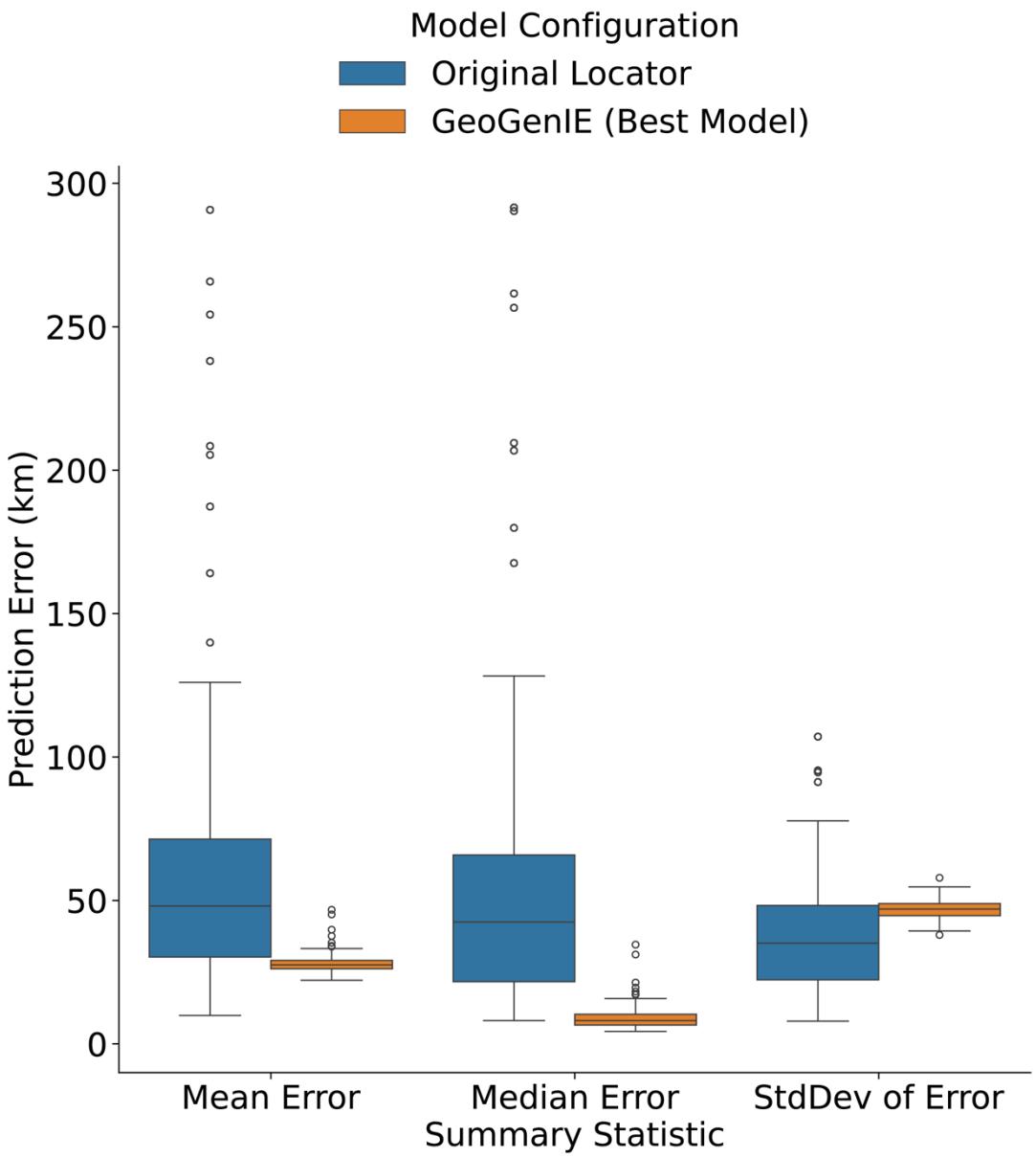

Figure 5. Performance Metrics for Geolocation Prediction: LOCATOR *versus* GEOGENIE

Comparison of performance metrics between LOCATOR and GEOGENIE. Statistics represent three metrics for 'prediction error' in kilometers [km]: Mean, Median and Standard Deviation. Estimates were based on *N*=1,415 deer samples from Arkansas genotyped across 436 SNP loci included in the GT-seq panel. Prediction error was measured as the Haversine distance (i.e., Great Circle) in kilometers between predicted and recorded localities for the 'test' subset of samples (i.e., unseen by the model during training). Summary statistics involve 100 bootstrap replicates. The GEOGENIE boxplots reflect the optimal model configuration, with sample weighting, outlier removal (yielding *N*=1,386 non-outlier samples), and oversampling enabled.

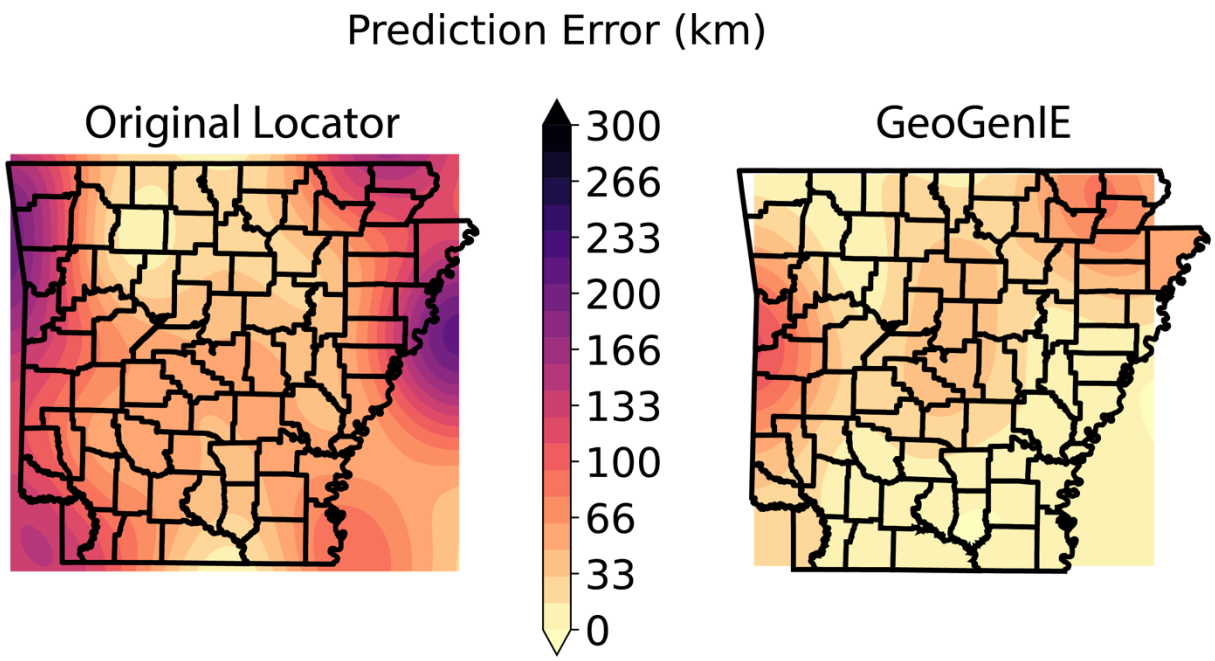

Figure 6. Spatial Interpolation of Geolocation Performance: LOCATOR *versus* GEOGENIE

Geolocation performance across the entire study area showing 'prediction error' in kilometers visualized as a map for LOCATOR (left) and GEOGENIE (right). Visualizations were generated via spatial interpolation of 'prediction error' across Arkansas. Geolocation predictions are based on *N*=1,415 deer samples genotyped across 436 SNP loci included in the GT-seq panel. Interpolated errors represent the Haversine distance (i.e., Great Circle) between recorded and predicted longitude and latitude coordinates, averaged over 100 bootstrap replicates on a held-out subset of samples unseen by the model during training (i.e., 'test' set). The GEOGENIE predictions reflect the optimal model configuration, with sample weighting, outlier removal (yielding *N*=1,386 non-outlier samples), and oversampling enabled.

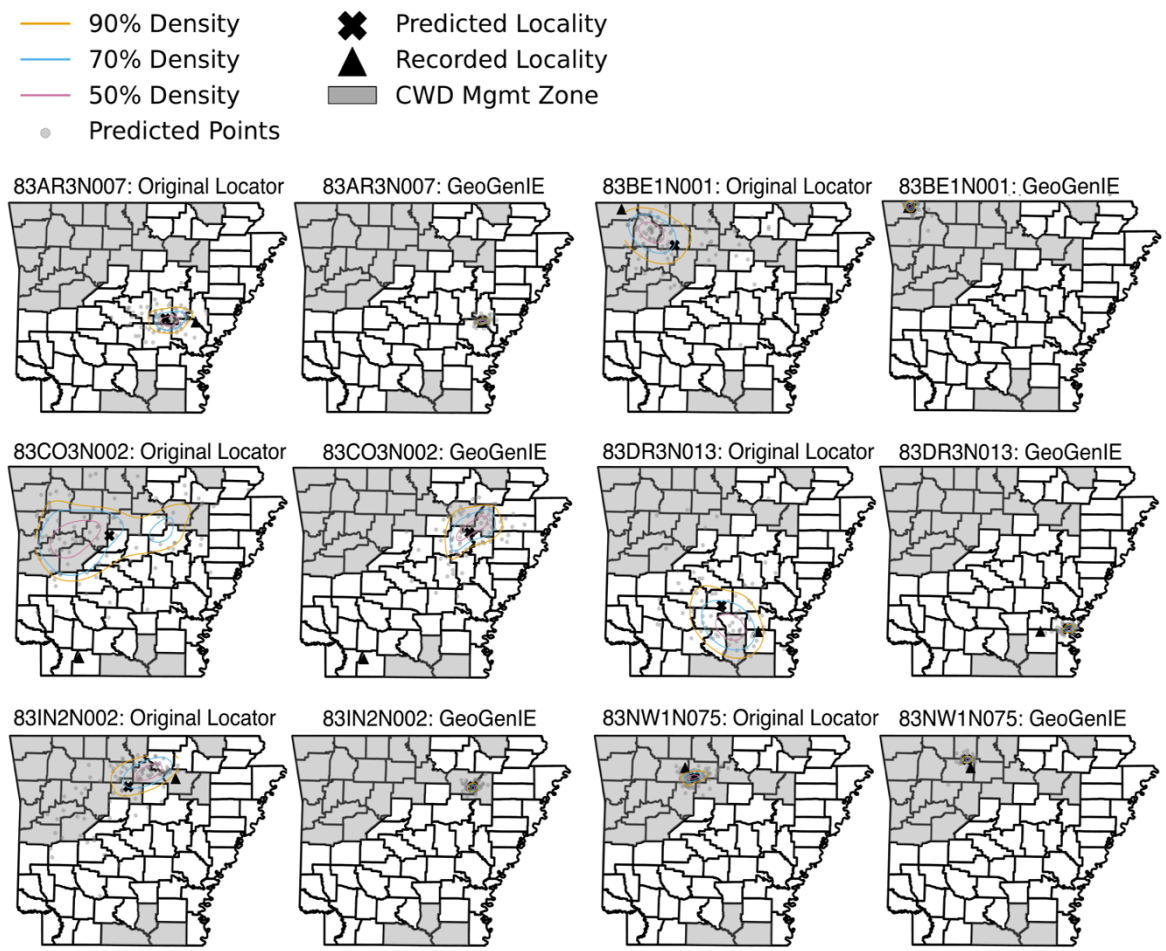

Figure 7. Geolocation Predictions of LOCATOR *versus* GEOGENIE for Select Samples

Comparison of geolocation prediction for *N*=6 select samples of White-tailed Deer derived via LOCATOR (left) and GEOGENIE (right). Predictions were modeled based on a dataset of *N*=1,415 deer sampled across Arkansas (Figure 2) and genotyped across 436 SNP loci. Depicted for each select sample are: Recorded locality (▲); Predicted locality (**X**) calculated as geographic centroid from predicted localities of 100 individual bootstrap replicates (gray circles); contour lines indicate the areas containing 90% (orange), 70% (blue), and 50% (pink) of the bootstrap predictions, respectively. Counties shaded in gray highlight the 2024 Chronic Wasting Disease (CWD) Management Zone (MZ). Map titles provide the sample identifier (DNA code; Table S1) and software package used.

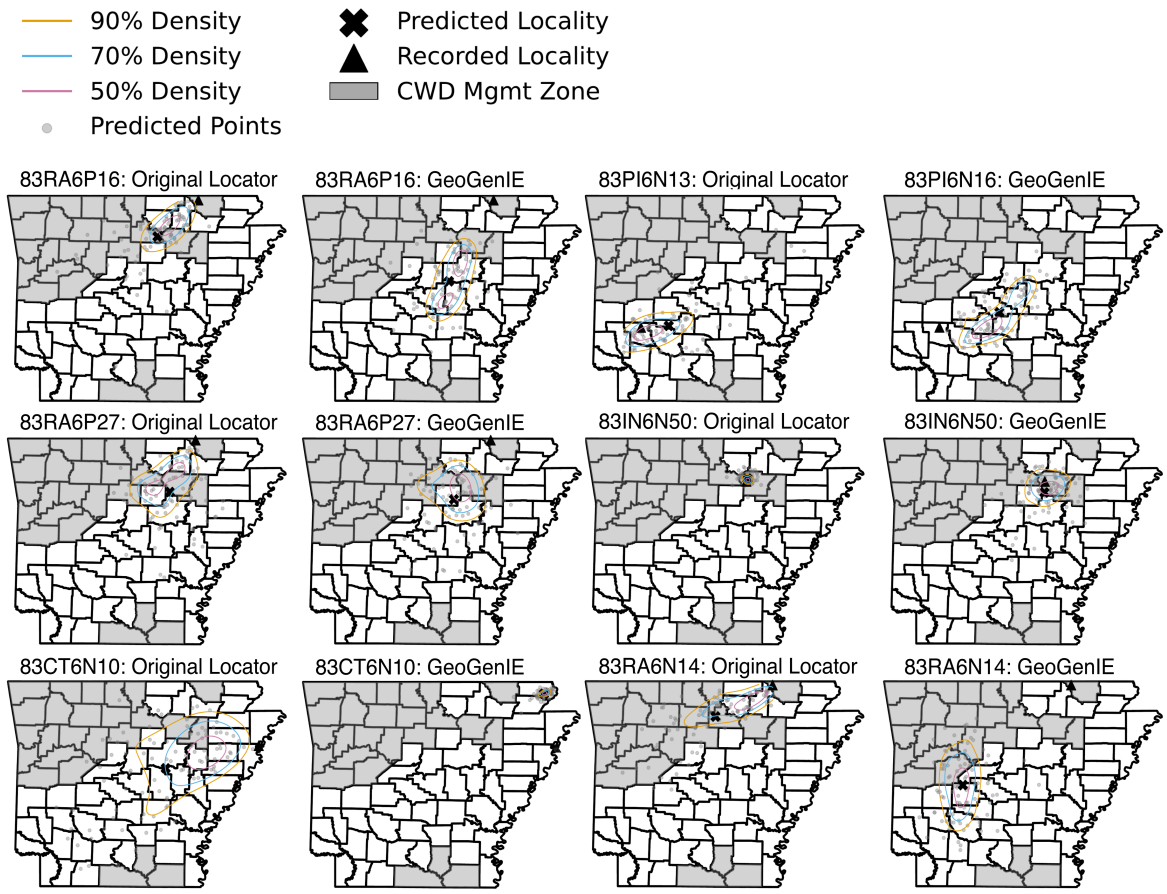

Figure 8. Geolocation Predictions of LOCATOR *versus* GEOGENIE for Select FY23 Samples

Comparison of geolocation prediction for *N*=6 select samples of White-tailed Deer collected for the 2023 Fiscal Year (FY) Surveillance efforts by AGFC (Phase 6). Predictions were modeled via LOCATOR (left) and GEOGENIE (right) and based on a dataset of *N*=1,415 deer sampled across Arkansas (Figure 2) and genotyped across 436 SNP loci. Depicted for each select sample are: Recorded locality (▲); Predicted locality (**X**) calculated as geographic centroid from predicted localities of 100 individual bootstrap replicates (gray circles); contour lines indicate the areas containing 90% (orange), 70% (blue), and 50% (pink) of the bootstrap predictions, respectively. Counties shaded in gray highlight the 2024 Chronic Wasting Disease (CWD) Management Zone (MZ). Map titles provide the sample identifier (DNA code; Table S2) and software package used.

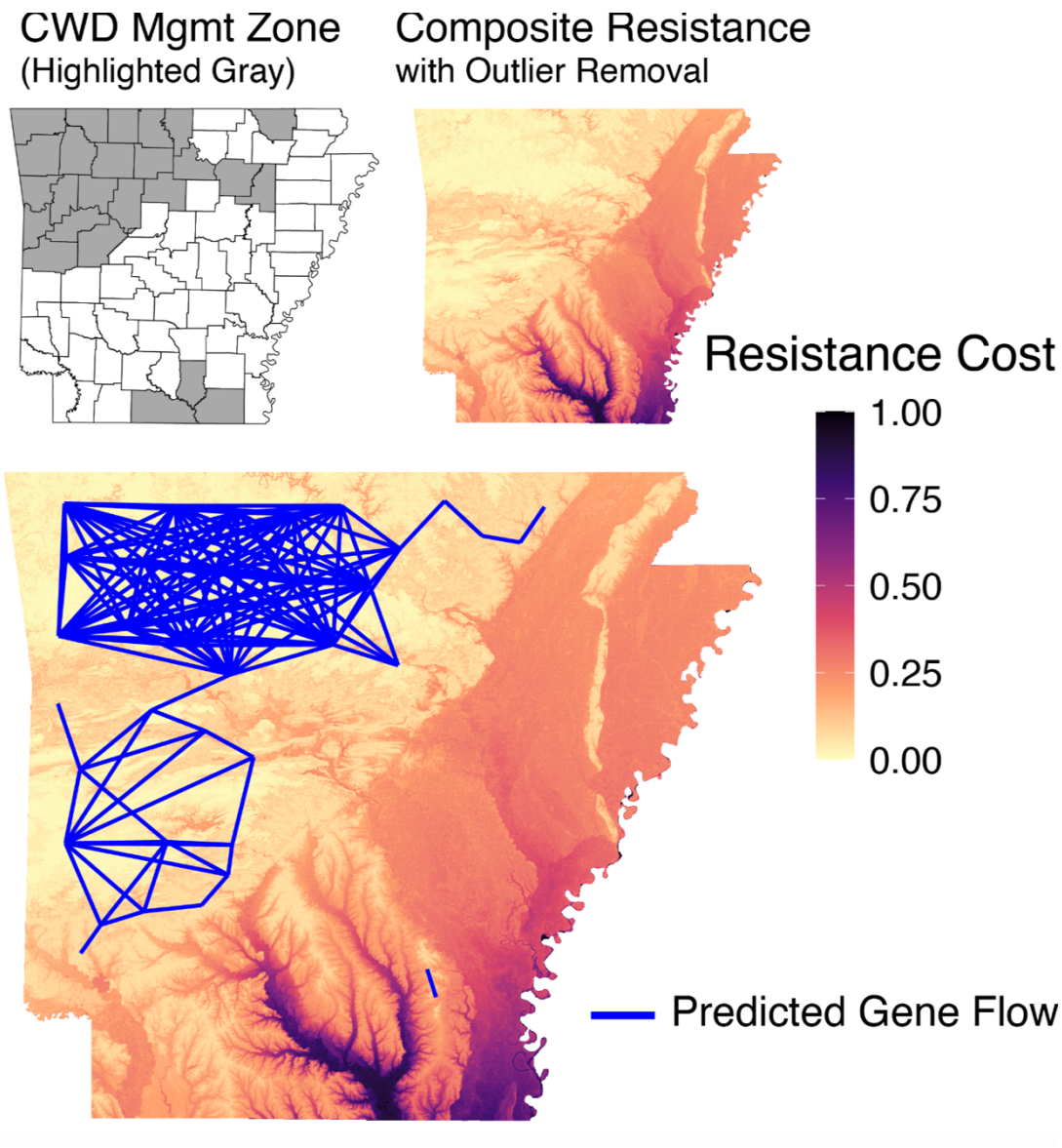

Figure 9. Landscape Resistance and Population Connectivity across Arkansas

Landscape resistance to dispersal of White-tailed Deer across Arkansas modeled using RESISTANCEGENIE to visualize broad patterns of population connectivity via predicted gene flow (blue lines). The maps show (Top Left) Map of Arkansas with county outlines and Management Zone highlighted in gray; (Top Right) Composite Resistance Surface (Top Right); and (Bottom) Population connectivity visualized as predicted gene flow (blue lines). The composite resistance surface was derived from a random subset of 1,000 SNP genotypes derived via ddRAD-seq and five environmental GIS (Geographic Information Systems) layers (for details see Appendix 2). Outliers removed.

12 | SUPPLEMENTAL MATERIAL

SUPPLEMENTAL TABLES

Table S1. SNP GT-seq Validation: 96 Samples Genotyped with ddRAD-seq and GT-seq

List of 96 samples of White-tailed Deer used to validate the SNP GT-seq assay. Samples that were previously genotyped via ddRAD-seq during Phase 1-5 (FY 2017-2021) were randomly selected from each county. Listed are: AGFC ID = identifier given by the Arkansas Game and FIsh Commission; DNA ID = corresponding aCaMEL identifier for DNA sample; County = county where sample was recorded; CWD = testing result for Chronic Wasting Disease; FY = Fiscal Year of surveillance effort; Fig = figure in this report depicting predicted location for sample. Recorded location of each sample is depicted in Figure S1.

AGFC ID	DNA ID	COUNTY	CWD	FY	Fig
CWD-AR-18-0241B	83AR2N005	Arkansas	Negative	2021	
AR032436	83AS5N003	Ashley	Negative	2021	
CWD-AR-00-9016	83BA3N010	Baxter	Negative	2019	
CWD-AR-16-5250	83BE1N002	Benton	Negative	2017	
CWD-AR-17-4120	83BE2N031	Benton	Negative	2018	
CWD-AR-17-15952	83BO2N040	Boone	Negative	2018	
CWD-AR-18-01133	83BR4N003	Bradley	Negative	2019	
CWD-AR-16-00261	83CA1N021	Carroll	Negative	2017	
CWD-AR-18-01469	83CA3U066	Calhoun	Negative	2019	
CWD-AR-17-18598	83CB2N001	Cleburn	Negative	2018	
CWD-AR-18-12	83CB2N004	Cleburn	Negative	2018	
CWD-AR-18-426	83CH2N001	Chicot	Negative	2018	
AR035750	83CL5N018	Clark	Negative	2021	
CWD-AR-17-19691	83CN2N004	Conway	Negative	2018	
n/a	83CO3U003	Columbia	Negative	2019	
CWD-AR-00-0846	83CO3N002	Columbia	Negative	2019	**7**
CWD-AR-00-0583	83CR3N004	Craighead	Negative	2019	
AR031106	83CS4N011	Cross	Negative	2020	
CWD-AR-18-699	83CT2N005	Crittenden	Negative	2018	
CWD-AR-18-829	83CW2N002	Crawford	Negative	2018	
CWD-AR-18-01375	83CW4N010	Crawford	Negative	2020	
n/a	83CY3U010	Clay	Negative	2019	
CWD-AR-18-01533	83DL3N009	Dallas	Negative	2019	
CWD-AR-00-0402	83DR3N013	Drew	Negative	2019	**7**
AR052695	83DR5N022	Drew	Negative	2021	
AR052693	83DS5N013	Desha	Negative	2021	
CWD-AR-18-0209HC	83FA2N012	Faulkner	Negative	2018	
CWD-AR-17-23913	83FR2N016	Franklin	Negative	2018	
AR034975	83FU5N024	Fulton	Negative	2021	
CWD-AR-18-01649	83GA3N016	Garland	Negative	2019	
n/a	83GE3N013	Greene	Negative	2019	
CWD-AR-00-0922	83GR3N011	Grant	Negative	2019	
AR035148	83HE3N007	Hempstead	Negative	2019	
CWD-AR-18-1059	83HM2N006	Hempstead	Negative	2018	

CWD-AR-00-1084	83HO3U011	Howard	Negative	2019
CWD-AR-18-563	83HS2N001	Hot Springs	Negative	2018
CWD-AR-18-152	83IN2N001	Independence	Negative	2018
AR041763	83IZ5N018	Izard	Negative	2021
CWD-AR-18-586	83JA2N001	Jackson	Negative	2018
AR032403	83JE4N014	Jefferson	Negative	2020
CWD-AR-16-04180	83JO1N018	Johnson	Negative	2017
AR032109	83LA4N016	Lafayette	Negative	2020
CWD-AR-17-3529	83LE2N003	Lee	Negative	2018
CWD-AR-17-3529	83LI3U004	Lincoln	Negative	2019
CWD-AR-18-0681HC	83LN2N005	Lonoke	Negative	2018
CWD-AR-18-681	83LN3N009	Lonoke	Negative	2019
CWD-AR-16-04458	83LO1N010	Logan	Negative	2017
CWD-AR-18-1051	83LR2N005	Little River	Negative	2018
CWD-AR-00-0625	83LW3N002	Lawrence	Negative	2019
CWD-AR-16-00796	83MA1N015	Madison	Negative	2017
CWD-AR-17-12542	83MA2N026	Madison	Negative	2018
CWD-AR-00-0963	83MI3U007	Miller	Negative	2019
CWD-AR-00-0966	83MI3U010	Miller	Negative	2019
CWD-AR-18-668	83MN2N012	Monroe	Negative	2018
n/a	83MO3N001	Montgomery	Negative	2019
n/a	83MO3N002	Montgomery	Negative	2019
CWD-AR-17-15261	83MR2N027	Marion	Negative	2018
CWD-AR-17-15261	83MR2N050	Marion	Negative	2018
CWD-AR-18-579	83MS2N001	Mississippi	Negative	2018
AR041370	83MS5N004	Mississippi	Negative	2020
n/a	83NE3U002	Nevada	Negative	2019
n/a	83NE3U004	Nevada	Negative	2019
CWD-AR-16-0017	83NW2N213	Newton	Negative	2018
CWD-AR-17-12460	83NW2P230	Newton	Positive	2018
CWD-AR-18-0502HC	83OU2N005	Ouachita	Negative	2018
CWD-AR-18-0129B	83PE3N23	Perry	Negative	2019
CWD-AR-17-3855	83PH2N005	Phillips	Negative	2018
CWD-AR-00-0993	83PI3N007	Pike	Negative	2019
CWD-AR-00-8769	83PL3N002	Polk	Negative	2019
CWD-AR-18-01359	83PO3N014	Poinsett	Negative	2019
CWD-AR-16-05798	83PP1N045	Pope	Negative	2017
CWD-AR-16-05799	83PP1N046	Pope	Negative	2017
CWD-AR-17-19923	83PP2N084	Pope	Negative	2018
CWD-AR-17-3704	83PR2N001	Prairie	Negative	2018
AR001231	83PU5N019	Pulaski	Negative	2021
n/a	83RA3N009	Randolph	Negative	2019
CWD-AR-17-19925	83SA2N001	Saline	Negative	2018
CWD-AR-17-1348	83SB2N006	Sebastian	Negative	2018
AR013441	83SC3N002	Scott	Negative	2019
AR041009	83SC4N003	Scott	Negative	2019

CWD-AR-17-13673	83SE2N019	Searcy	Negative	2018
CWD-AR-17-13675	83SE2N021	Searcy	Negative	2018
CWD-AR-17-13675	83SF3N008	St. Francis	Negative	2019
CWD-AR-00-1356	83SH3N013	Sharp	Negative	2019
CWD-AR-00-1356	83ST2N003	Stone	Negative	2018
CWD-AR-00-1356	83SV3N008	Sevier	Negative	2019
CWD-AR-00-0822	83UN3N001	Union	Negative	2019
AR054533	83UN5P008	Union	Positive	2021
AR055502	83UN5N010	Union	Negative	2021
AR055502	83VB2N007	Van Buren	Negative	2018
AR055502	83WA2N015	Washington	Negative	2018
CWD-AR-00-6659	83WA3N021	Washington	Negative	2019
CWD-AR-00-0340	83WD3N009	Woodruff	Negative	2019
CWD-AR-00-0340	83WH2N001	White	Negative	2018
CWD-AR-00-0340	83YE1N018	Yell	Negative	2017
CWD-AR-17-16447	83YE2N046	Yell	Negative	2018

Table S2. SNP GT-seq Validation: 96 Samples Genotyped with ddRAD-seq and GT-seq

List of 96 samples of White-tailed Deer used to validate the SNP GT-seq assay. Samples that were previously genotyped via ddRAD-seq during Phase 1-5 (FY 2017-2021) were randomly selected from each county. Listed are: AGFC ID = identifier given by the Arkansas Game and FIsh Commission; DNA ID = corresponding aCaMEL identifier for DNA sample; County = county where sample was recorded; CWD = testing result for Chronic Wasting Disease; FY = Fiscal Year of surveillance effort; Fig= figure in this report depicting predicted location for sample.

AGFC ID	DNA ID	COUNTY	CWD STATUS	FY	Fig
AR042193	83BA6N020	Baxter	Negative	2023	
AR082524	83BA6N021	Baxter	Negative	2023	
AR051018	83CA6N077	Calhoun	Negative	2023	
AR012156	83CV6N020	Cleveland	Negative	2023	
AR012157	83CV6N021	Cleveland	Negative	2023	
AR041258	83CG6N017	Craighead	Negative	2023	
AR041259	83CG6N018	Craighead	Negative	2023	
AR055820	83CT6N010	Crittenden	Negative	2023	**8**
AR055304	83DA6N010	Dallas	Negative	2023	
AR006971	83DR6N023	Drew	Negative	2023	
AR006973	83DR6N024	Drew	Negative	2023	
AR055296	83DR6N025	Drew	Negative	2023	
AR055297	83DR6N026	Drew	Negative	2023	
AR018531	83GE6N021	Greene	Negative	2023	
AR018532	83GE6N022	Greene	Negative	2023	
AR031730	83GE6N023	Greene	Negative	2023	
AR031731	83GE6N024	Greene	Negative	2023	
AR041256	83GE6N025	Greene	Negative	2023	
AR083359	83HO6N021	Howard	Negative	2023	
AR049508	83IN6N042	Independence	Negative	2023	
AR049509	83IN6N043	Independence	Negative	2023	
AR049510	83IN6N044	Independence	Negative	2023	
AR049511	83IN6N045	Independence	Negative	2023	
AR049512	83IN6N046	Independence	Negative	2023	
AR049513	83IN6N047	Independence	Negative	2023	
AR049514	83IN6N048	Independence	Negative	2023	
AR049515	83IN6N049	Independence	Negative	2023	
AR049516	83IN6N050	Independence	Negative	2023	**8**
AR049517	83IN6N051	Independence	Negative	2023	
AR049518	83IN6N052	Independence	Negative	2023	
AR049519	83IN6N053	Independence	Negative	2023	
AR049520	83IN6N054	Independence	Negative	2023	
AR049521	83IN6N055	Independence	Negative	2023	
AR082546	83IN6N056	Independence	Negative	2023	

AR082547	83IN6N057	Independence	Negative	2023	
AR082548	83IN6N058	Independence	Negative	2023	
AR082549	83IN6N059	Independence	Negative	2023	
AR082550	83IN6N060	Independence	Negative	2023	
AR082551	83IN6N061	Independence	Negative	2023	
AR082552	83IN6N062	Independence	Negative	2023	
AR082553	83IN6N063	Independence	Negative	2023	
AR082554	83IN6N064	Independence	Negative	2023	
AR082555	83IN6N065	Independence	Negative	2023	
AR082556	83IN6N066	Independence	Negative	2023	
AR082557	83IN6N067	Independence	Negative	2023	
AR082558	83IN6N068	Independence	Negative	2023	
AR082559	83IN6N069	Independence	Negative	2023	
AR082560	83IN6N070	Independence	Negative	2023	
AR026754	83LO6N044	Logan	Negative	2023	
AR026755	83LO6N045	Logan	Negative	2023	
AR081866	83LO6N046	Logan	Negative	2023	
AR012155	83LN6N014	Lonoke	Negative	2023	
AR041458	83MS6N005	Mississippi	Negative	2023	
AR041459	83MS6N006	Mississippi	Negative	2023	
AR041460	83MS6N007	Mississippi	Negative	2023	
AR041461	83MS6N008	Mississippi	Negative	2023	
AR073510	83MS6N009	Mississippi	Negative	2023	
AR041012	83PI6N011	Pike	Negative	2023	
AR041013	83PI6N012	Pike	Negative	2023	
AR056735	83PI6N013	Pike	Negative	2023	**8**
AR056736	83PI6N014	Pike	Negative	2023	
ARO41011	83PI6N015	Pike	Negative	2023	
AR055819	83PO6N018	Poinsett	Negative	2023	
AR012183	83PU6N025	Pulaski	Negative	2023	
AR006831	83RA6N012	Randolph	Negative	2023	
AR041225	83RA6N013	Randolph	Negative	2023	
AR055836	83RA6N014	Randolph	Negative	2023	**8**
AR055837	83RA6N015	Randolph	Negative	2023	
AR055845	**83RA6P016**	**Randolph**	**P**	**2023**	**8**
AR055846	83RA6N017	Randolph	Negative	2023	
AR055847	83RA6N018	Randolph	Negative	2023	
AR055850	83RA6N019	Randolph	Negative	2023	
AR078023	83RA6N020	Randolph	Negative	2023	
AR078026	83RA6N021	Randolph	Negative	2023	
AR078028	83RA6N022	Randolph	Negative	2023	
AR078032	83RA6N023	Randolph	Negative	2023	
AR078035	83RA6N024	Randolph	Negative	2023	
AR078225	83RA6N025	Randolph	Negative	2023	
AR078226	83RA6N026	Randolph	Negative	2023	
AR078257	**83RA6P027**	**Randolph**	**P**	**2023**	**8**

AR041009	83SC6N004	Scott	Negative	2023
AR041010	83SC6N005	Scott	Negative	2023
AR082528	83SH6N019	Sharp	Negative	2023
AR042192	83ST6N019	Stone	Negative	2023
AR082562	83ST6N020	Stone	Negative	2023
AR018093	83UN6N052	Union	Negative	2023
AR018096	83UN6N053	Union	Negative	2023
AR018097	83UN6N054	Union	Negative	2023
AR018099	83UN6N055	Union	Negative	2023
AR039804	83UN6N056	Union	Negative	2023
AR039805	83UN6N057	Union	Negative	2023
AR039809	83UN6N058	Union	Negative	2023
AR039810	83UN6N059	Union	Negative	2023
AR039811	83UN6N060	Union	Negative	2023
AR081177	83UN6N061	Union	Negative	2023
AR026727	83YE6N057	Yell	Negative	2023

SUPPLEMENTAL FIGURES

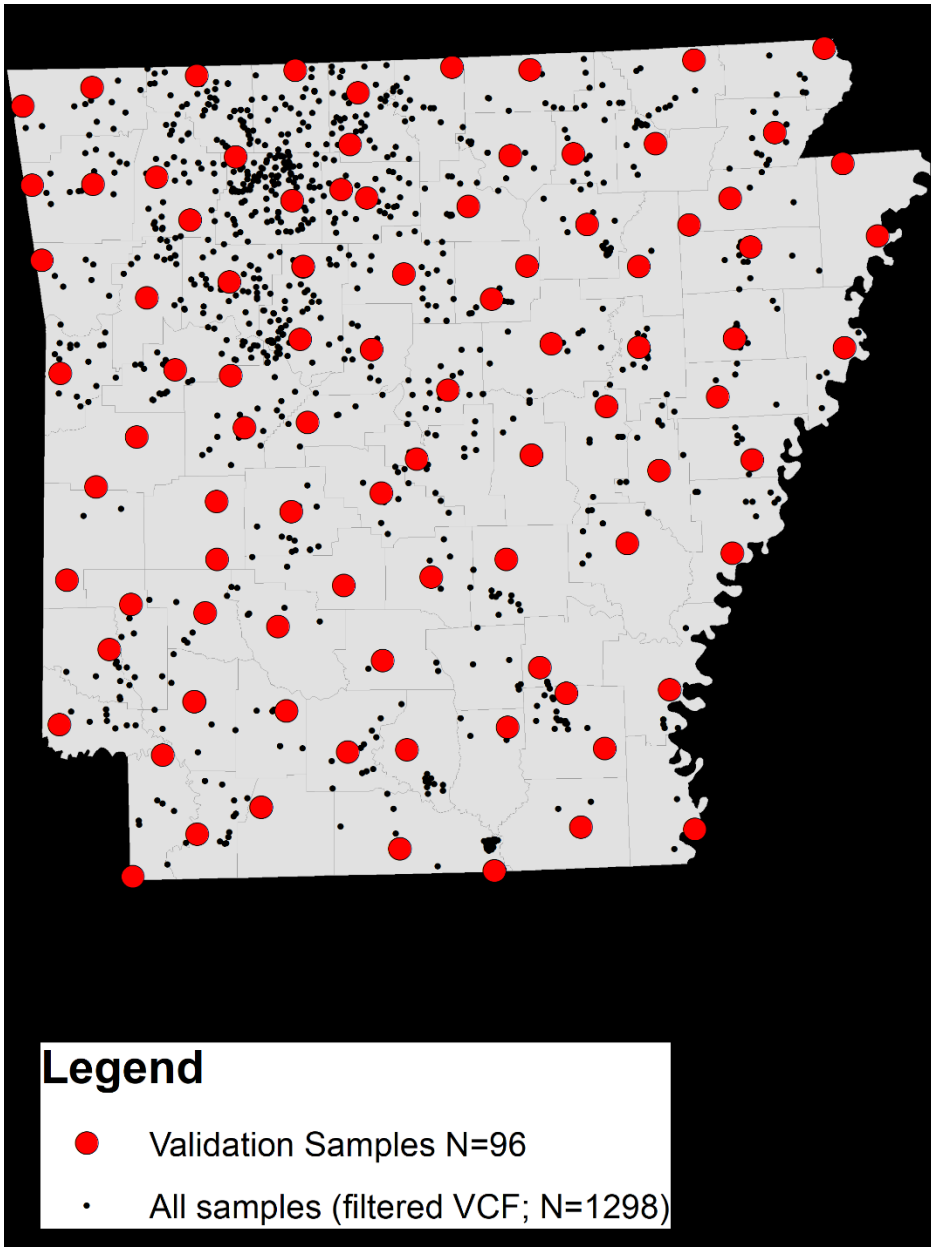

Figure S1. Sample Selection for SNP GT-seq Panel Validation

A map of White-tailed Deer samples (N=96) used to validate the GT-seq panel through genotyping. Samples were chosen that contained high proportion of genotyping success in previous ddRAD sequencing and were chosen from each Arkansas county to maximize spatial and genomic variation. Samples are listed in Table S1.

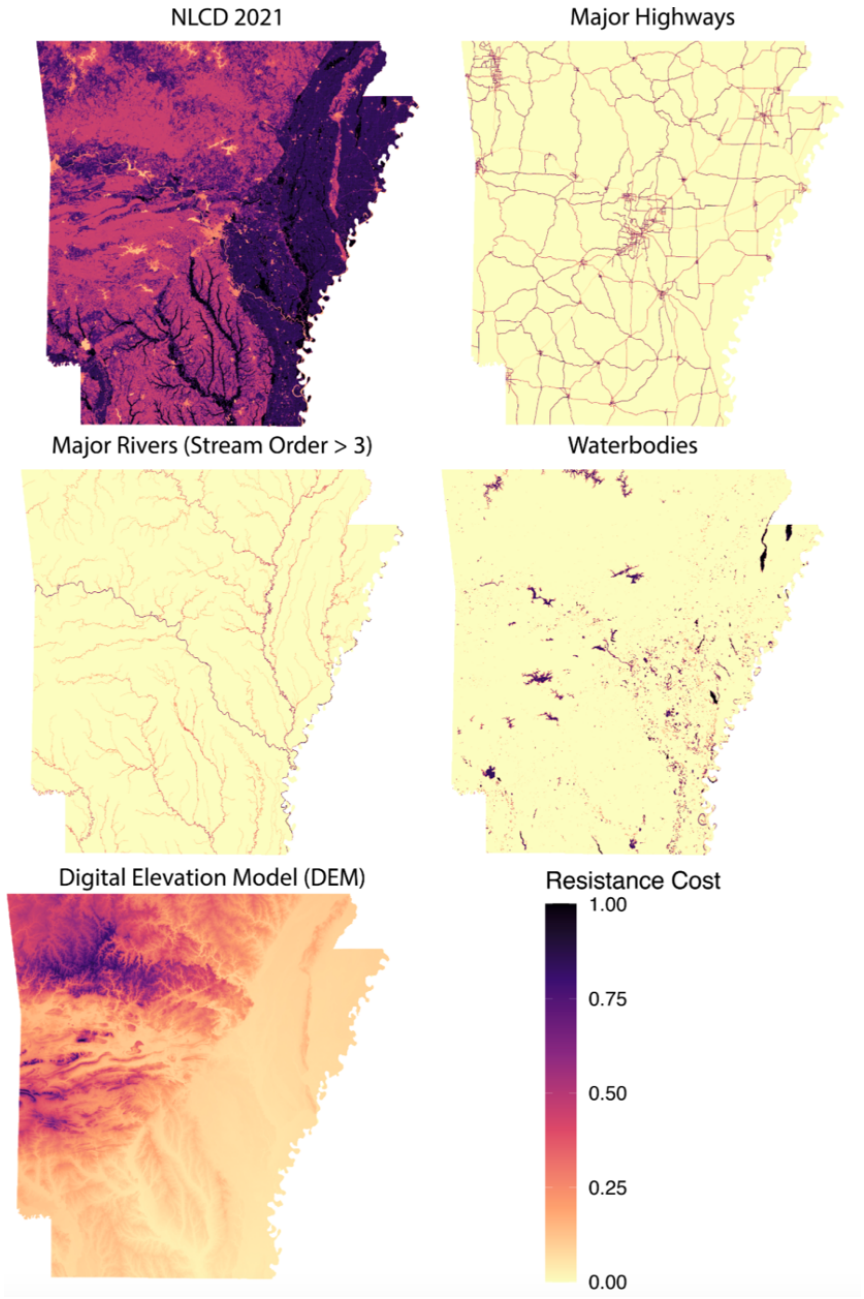

Figure S2. Resistance Surfaces of Five Environmental Layers

The individual resistance layers comprising the composite surface (Figure 9) were derived from a random subset of 1,000 SNP genotypes and five environmental GIS (Geographic Information Systems) layers: 1) 2021 National Land Cover Database (NLCD), with generalized land cover categories (i.e., similar categories were merged); 2) Major arterial highways; 3) Major Rivers (Stream Order > 3); 4) Waterbodies (e.g., lakes, ponds, wetlands); 5) a Digital Elevation Model (DEM). For additional detail, see Appendix 2.

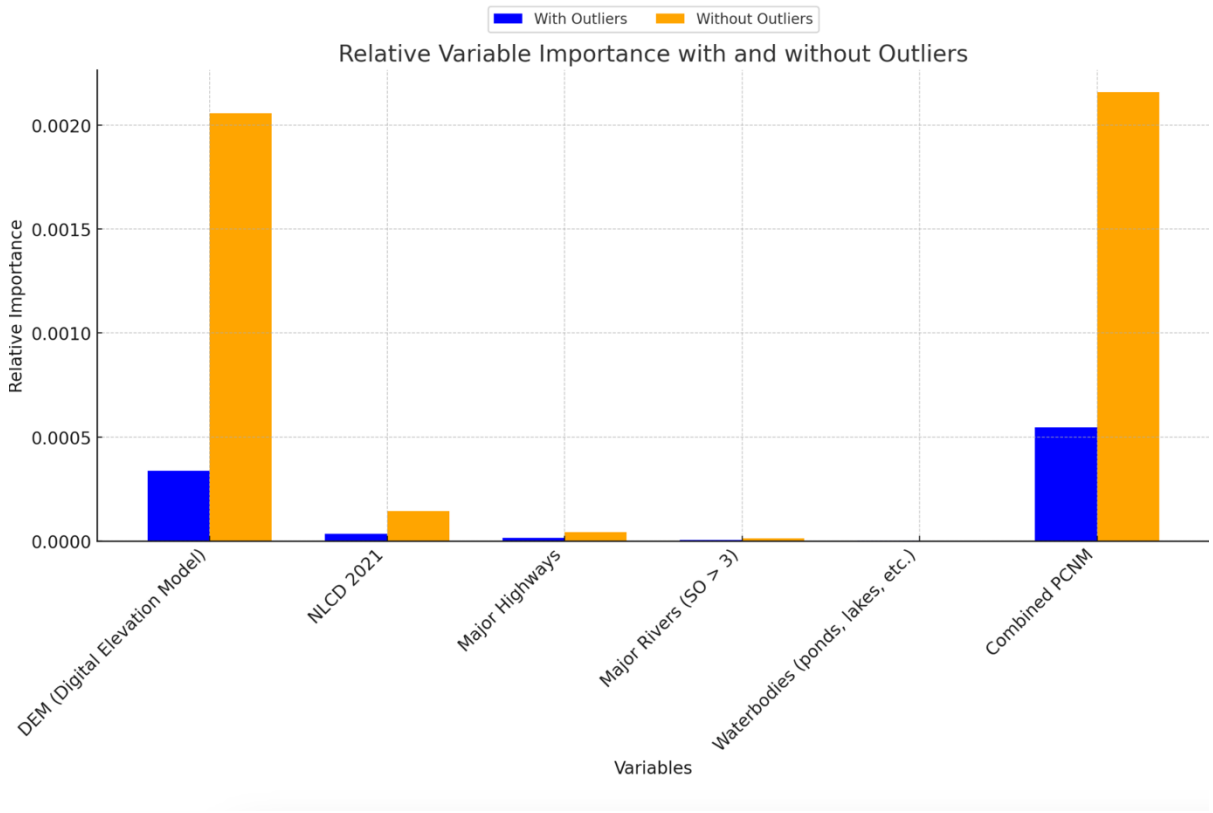

Figure S3. Relative Importance of Variables to Predict Landscape Resistance

Influence of genetic variance due to outlier data on the relative importance of environmental variables and spatial features to predict landscape resistance. Blue: outliers included; Orange: outliers removed. The relative importance metric quantifies the contribution of each environmental layer and combined PCNMs to the genetic variation observed in the SNP data. The individual resistance layers comprising the composite surface (Figure 9) were derived from a random subset of 1,000 SNP loci and five environmental GIS environmental layers: DEM = Digital Elevation Model; NLCD 2021 = National Land Cover Database, with generalized land cover categories (i.e., similar categories were merged); Major Highways = major arterial roads; Major Rivers = lentic systems with Stream Order >3; and Waterbodies = lotics systems (e.g., lakes, ponds, wetlands, etc.). Combined PCNMs (Principal Coordinates of Neighbor Matrices) = spatial variables capturing spatial autocorrelation in the genetic data. For additional detail, see Appendix 2.

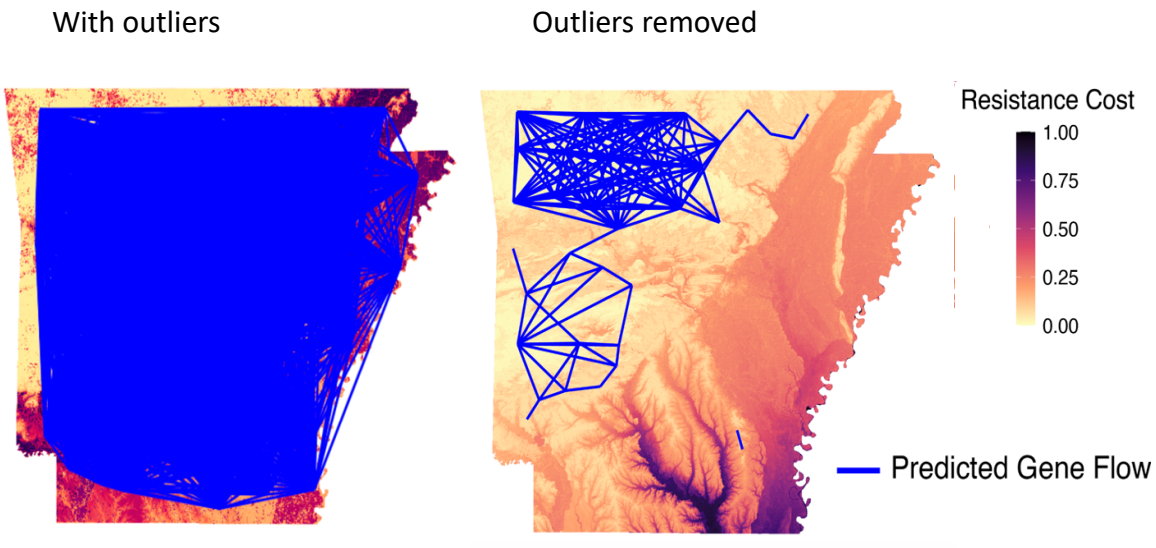

Figure S4. Predicted Gene Flow Pathways – Effects of Outliers

Predicted gene flow pathways (blue lines) across Arkansas counties visualize population connectivity with outliers included (left) and with outlier removal enabled (right). The presence of gene flow pathways is determined as being within a threshold distance, beyond which the correlation between genetic similarity and geographic proximity diminishes. Because historic translocations of White-tailed Deer across Arkansas remain as ancestry signals in the genotypes of samples, the correlation between spatial and genetic distances is spurious (i.e., deviates from model assumption of 'isolation-by-distance'). The uncorrected data (left) mask the nuanced signals reflecting natural deer dispersal over generations; the resulting visualization is meaningless. For details on the underlying composite resistance surface, see Appendix 2.

Appendix 1: GT-SEQ Panel Development

SNP Discovery

Our goal was to generate a Genotyping-in-Thousands (GT-seq) panel (Campbell et al. 2015) containing 400 to 500 SNP loci from which population assignment and approximate geographic origin for newly sampled individuals could be inferred. We generated ddRADseq data for a total of *N*=1,381 individual deer from Arkansas following Chafin et al. (2021) protocols, which initially produced *N*=1,242 of the individual sequences used herein, and augmented since by data generated via ddRAD-seq for additional samples (*N*=139) collected by AGFC during surveillance efforts from 2019-2022 (Douglas et al. 2022).

Genetic variants called containing 1,046,465 SNPs (= Single Nucleotide Polymorphisms) from all individuals (*N*=1,381) were aligned using reference-guided assembly based on a chromosome-level genome assembly of White-tailed Deer (London et al. 2022) using IPYRAD v.0.9.87 (Eaton and Overcast 2020). Quality control checks were performed on the raw sequence data using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Raw sequence files were then de-multiplexed, and reads were assigned to individuals based on their unique DNA barcodes, allowing no tolerance for mismatched barcodes.

Adapters and primers were removed and reads with >5 low-quality bases (Phred<20) were discarded. Clusters were removed via conditional criteria to ensure high-quality data: <10x and >1000x coverage per individual; >5% of consensus nucleotides ambiguous; >20% of nucleotides polymorphic; >8 indels present; or presence in <20 individuals. Putative paralogs were removed if clusters displayed >2 alleles per site in consensus sequence or excessive heterozygosity (>5% consensus bases or >50% heterozygosity/site).

SNP Filtering for GT-seq Panel

Quality control and filtering of the loci alignment were used to reduce the data to those SNPs most suitable for GT-seq using the R statistical software version 4.1.3 (R Core Team 2022). We first eliminated any individuals with more than 50% missing data. Subsequent SNP retention was based on stringent criteria: SNPs had to be located on one of the 36 autosomal chromosomes; positioned between bases 25-75 on a 100bp read; and found on a read with 12 or fewer SNPs. Additionally, the reads were required to have a minimum sequence depth of 10X and a maximum of 200X. Only SNPs that were present in at least 50% of the individuals and had a minor allele count of at least two were considered. When multiple SNPs per 100bp read met these criteria, we selected the SNP with the highest minor allele count to maintain a single SNP per read and minimize linkage disequilibrium. A combination of R packages was used for visualization, summary, and filtering, including VCFR v.1.13.0 (Knaus & Grünwald 2017), SNPFILTR v.1.0.0 (DeRaad 2022), and DARTR v.2.7.2 (Mijangos et al. 2022).

Population Structure and Ranking SNPs by Divergence

The resulting 2,156 SNP loci were used to infer population structure using an approach that clusters individuals (*N*=1,298) based on genotypes called sparse non-negative matrix factorization (sNMF; Frichot et al. 2014). *K*=1 to *K*=15 were analyzed for fit using cross-validation with 25 replicates and a regularization parameter of 100. Based on this analysis and the previously published population structure of Arkansas White-tailed Deer (Chafin et al. 2021), we grouped the individuals into *K*=8 genetic populations.

To determine which loci were most informative for identifying population assignment and we ranked them based on their population divergence (Li et al. 2023). We calculated the genetic divergence (F_{ST}) per locus based on the matrix of ancestry proportions estimated using the *K*=8 sNMF solution (Qmatrix = individuals X sub-populations). We used a slightly modified version of the function '*snmf.pvalues*' from the LEA package (Frichot & François 2015). The function was modified to output the F_{ST} values. This approach works well for admixed and more continuously structured populations by using the ancestry proportions to weight heterozygosity used to calculate F_{ST} (Martins et al. 2016).

GT-seq Primer Design and Panel Optimization

Our colleagues at GTseek (Twin Falls, ID) performed primer design and validation. Candidate loci chosen for primer design were *N*=800 of the autosomal loci with the highest population genetic divergence, *N*=2 loci associated with the PRNP gene (Chafin et al. 2020), and *N*=3 sex-linked loci on the Y chromosome (Strickland et al. 2011). A total of *N*=578 loci passed the primer design filtering steps, and *N*=500 were chosen for primer synthesis and testing. Testing the designed primer panel involved preparing an initial test primer mix with all 500 primer sets and then using all designed primers to prepare and sequence a GT-seq library. The sequencing data were analyzed using Python and Perl scripts (available at: https://github.com/Gtseq/Gtseek_utils). Primers with large amounts of associated “off-target” or primer artifact sequences are marked for omission from the primer mix. In this case, a set of 74 locus primers was chosen for omission because their primers either failed to amplify their intended target or produced a large number of off-target sequences. A final GT-seq panel of *N*=436 autosomal loci, *N*=2 PRNP loci, and *N*=3 sex-linked loci showed desirable genotype capture rates. It was used to genotype *N*=96 individuals previously genotyped with ddRAD and *N*=96 newly sampled individuals (*N*=192 total individuals).

Appendix 2: GeoGenIE Development

Data Preprocessing

GeoGenIE applies several preprocessing steps in order to mitigate biases associated with sparse and/or imbalanced datasets, as well as to detect and remove spurious signals (for example, associated with translocation). To address the unique translocation history of White-tailed Deer (Chafin et al. 2021), we integrated an algorithm inspired by GGOutlieR (Chang et al. 2023) which detects 'geo-genetic outliers', which may demonstrate genetic differentiation reflecting aspatial processes (Epps & Keyghobaldi 2015). This step considers outliers along both genetic and geographic axes, using k-nearest neighbor (KNN) regression, and an expectation that isolation-by-distance is ubiquitous at the local scale (Meirmans 2012; Rousset 1997).

The input data, either GT-Seq CSV (comma-delimited) or VCF formatted files, are then divided into subsets used for model training, testing, and validation. Here, the model is fitted to the 'training' subset, with its capacity to generalize evaluated using the validation sets – from this the performance of the model ('loss') is assessed across training iterations. True generalization is then tested with wholly unseen data (the "test" set), a design commonly applied to reduce overfitting (i.e, model unable to generalise) and data leakage (i.e, biased accuracy due to the model having 'seen' the validation set). We then further hold out samples as a 'prediction set', including samples with unknown localities for location prediction. Missing data imputation is applied using the most frequent allele per locus, and the data are encoded in a 0-1-2 format for reference, heterozygous, and alternate alleles, respectively.

Model Architecture and Optimization

GeoGenIE uses a deep learning model built with PyTorch >= 2.1.2 (Facebook AI Research 2019), building upon the multilayer perceptron (MLP) architecture of Locator (Battey et al. 2020). Rather than using a fixed architecture (e.g, number of hidden layers), we implemented automated parameter tuning using Bayesian optimization in Optuna (Akiba et al. 2019). This approach uses a tree-structured algorithm (TPE; tree-structured Parzen Estimator) which refining the search space for parameter distributions to heuristically identify the best combination which minimizes prediction error. Users can choose from three loss functions for prediction error assessment (i.e, Distance Root Mean Square, Huber, and Root Mean Squared Error), with the final optimized model used to generate various visualizations, performance metrics, and full results for downstream reporting. GeoGenIE provides several output metrics aiding in interpretation of accuracy and precision of predictions. We calculate prediction error of 'known' samples using the Haversine distance between the predicted and observed localities (recorded upon sample collection). GeoGenIE also implements a parallelized bootstrapping of loci within individuals, with final predictions visualized as a density kernel of pseudoreplicated predictions.

A novel feature implemented in GeoGenIE aims to tackle sampling imbalance. First, sample weights are calculated based on inverse sampling density [i.e, 1 / (samples / km^2)] relative to geographically proximal neighbors. Using K-means clustering (Lloyd 1982), we group locations so those within the same cluster

are closer to each other than to those in other clusters. The search for the optimal number of clusters is automated by identifying the minimal silhouette score. These weights are in turn optionally enabled to assigns higher penalties to less densely sampled regions, forcing a greater emphasis on these areas during model training. Further, GEOGENIE employs a stratified sampling design to evenly distribute samples from dense and sparse regions into training, validation, and testing datasets using a similar approach. As a final mitigative option, a custom synthetic over-sampling algorithm, adapted from SMOTE (Chawla et al. 2002), can be used to generate synthetic samples using a Mendelian inheritance-based interpolation, incorporating regression-based techniques to balance sample densities for relatively sparse areas of the sampled space.