Final Report – WTD Phase 5

Screening SNP variation to model deer dispersal in Arkansas (2022)

Submitted to:

M. Cory Gray and Chris Middaugh

Arkansas Game and Fish Commission

771 Jordan Drive

Monticello, AR 72655

Prepared by:

Marlis R. Douglas

Zachery D. Zbinden

Tyler K. Chafin

Bradley T. Martin

Michael E. Douglas

Arkansas Conservation & Molecular Ecology Laboratory (aCaMEL)

University of Arkansas

Biological Sciences

850 W Dickson St

Fayetteville, AR 72701

Respectfully submitted:

30 June 2022

SUMMARY

This study was completed in support of AGFC's continued effort to predict potential threat of Chronic Wasting Disease (CWD) in Arkansas. The primary objective was to determine likely origin of seven CWD positive individuals collected in FY2022. Two of these were harvested in counties (Randolph and Union) outside the current Management Zone (MZ), whereas another three represented first documented cases of CWD positive WTD in counties within the 2022 MZ (Crawford, Franklin, and Van Buren).

A total of 144 white-tailed deer (WTD) were assayed for single nucleotide polymorphisms (SNPs). These data were incorporated into the existing state-wide SNP database derived from WTD collected from 2016-2020. The final data set contained 1,306 individuals and genotypes for 262,613 SNP loci. A probabilistic modeling approach (LOCATOR) was used to predict origin of the seven CWD positive individuals.

Analyses revealed that all seven individuals most likely originated from within or near by areas where the samples were collected. For the individuals from Randolph and Union counties, this documents CWD transmissions likely occur outside the current MZ in north central and south central Arkansas. For the individuals from Crawford, Franklin and Van Buren counties, the results document local presence of CWD in areas contained within the MZ, but previously without confirmed cases of CWD infections.

At first glance, the probabilistic location predictions for the samples from Crawford and Franklin counties seem to indicate a potential origin from counties in north central Arkansas, and hence would suggest long-distance dispersal of these individuals. However, the pattern is consistent with distribution of genetic subpopulations (ancestral gene pools) in Arkansas previously documented by the state-wide analysis and most likely reflects historic translocations from the Silamore district in north central to western Arkansas.

Results in this study underscores the importance of the state-wide genetic database of WTD the AGFC established and how it can be leveraged to understand WTD dispersal and inform CWD management efforts.

INTRODUCTION AND BACKGROUND

In Arkansas Chronic Wasting Disease (CWD) was first detected in wild cervids in 2016, and has since been documented in over over 1,200 white-tailed deer (WTD) from 19 counties (Figure 1). The Arkansas Game and Fish Commission (AGFC) has designated a CWD Management Zone (MZ) which currently includes 21 counties (Figure 2). As of March 2022, CWD positive cervids have been detected in 17 counties included in the MZ. In addition, in two CWD positive samples were identified in counties outside the MZ (i.e., Randolph and Union counties).

The AGFC continues surveillance of CWD positive WTD for risk assessment and to inform disease management efforts. As part of this approach, the AGFC has — in collaboration with the Arkansas Molecular Ecology and Conservation Laboratory (aCaMEL) at the University of Arkansas — conducted a state-wide SNP genetic assessment of WTD (Chafin et al., 2021). This genetic database can now be leveraged to determine likely origins of CWD positive WTD detected within and outside the previously known CWD range using a predictive model (LOCATOR analysis).

The current study (PHASE-5) continues this effort and assayed 144 WTD tissues collected during 2021 and 2022. The primary objective was to identify likely origin of seven CWD positive WTD (Table 1). In addition, samples from counties previously sparsly represented in the state-wide database were genotyped to help improve model predictions (Appendix 1).

METHODS

Library preparation

WTD tissue samples (*N*=146) were received in early March 2022 (Appendix 1) and genomic DNA extracted using standardized protocols (Chafin et al., 2020). A total of 144 DNA samples were processed for SNP generation (AR056662/83XX5N005 & AR056663/83XX5N005 were excluded due to lack of locality information). Note that sample numbers that are multiples of 48 work best with our pipeline, and samples must have sample coordinates to be useful as *training*

data for LOCATOR analysis (Battey et al., 2020). However, origins can be *predicted* for samples without locality information.

Library preparation followed previously established laboratory protocols for WTD (Chafin et al., 2021). Double digest restriction site-associated DNA (ddRAD) sequencing libraries were prepared by first digesting genomic DNA with *Nsi*I and *Msp*I restriction enzymes and using unique inline barcodes ligated to each sample prior to pooling (Peterson et al., 2012). Slight differences in sequencing are specific to PHASE-5: (i) the sequencing platform NovSeq 6000 instead of HiSeq 4000 (we expect full compatibility with previous Illumina libraries); (ii) pooled *N*=144 *vs.* 96 individuals. The higher throughput of the new sequencing platform allows for more samples to be pooled, reducing overall costs. Libraries were shipped to the University of Oregon GC3F on 28 March 2022, and sequencing data were received 13 April 2022.

Reference-guided alignment

Raw data for the new samples (PHASE-5; *N*=144) were demultiplexed (Eaton and Overcast, 2020), allowing no barcode mismatches (*N*=1 removed due to low quality; AR055599/83UN5N016). Out of 1,286 samples already demultiplexed from previous phases, *N*=2 were removed due to low quality: CWD-AR-18-177/83FU2N005, CWD-AR-16-0028/83NW1P006. Thus *N*=1,427 samples were assembled for DNA alignment.

Samples were clustered and aligned using ipyrad (Eaton and Overcast, 2020). Parameters were based on previous WTD pipelines, and the full parameters file is included in Appendix 2. The only exception was that this alignment used a WTD genome submitted by New England Biolabs as a reference, which should improve the overall assembly (https://www.ncbi.nlm.nih.gov/assembly/GCA_014726795.1).

Before analysis, some individuals from the 1,427 individuals in alignment were removed. Coordinates were unavailable for *N*=62 individuals from earlier phases, so they were removed. After mapping individuals, *N*=2 appeared outside of Arkansas and were removed (-/83DL3N005 & CWD-AR-18-675/83CT2N003; assumed Easting and Northing were erroneous). Variant call format (VCF) files were filtered by removing single nucleotide polymorphisms (SNPs) with

>25% missing data and individuals with >90% missing data (*N*=57 individuals removed). The remaining individuals used for analysis consisted of *N*=1,306 individuals.

Predicting geolocations

To determine the geolocation of 'origination' points for sampled deer, backward inference from genotypes was implemented via the deep-learning method LOCATOR (Battey et al., 2020; Chafin et al., 2021). There were *N*=23 individuals from Phase-5 that lacked geographic coordinates, which were not used in training or validation. Neither were the *N*=7 CWD positive target samples, but their origin locations were predicted.

The analysis first uses a portion of the data to train the neural network (e.g., 70%). The remaining proportion is used to validate the model (i.e., how close are the model predictions to the known sampling locations?). Like all deep-learning methods, LOCATOR is a parameter-rich method, meaning there are many *a priori* choices made by default. *We suggest exploring these choices more fully in the future using a grid search* (e.g., Martin et al., 2021). A manual heuristic search (non-exhaustive) was performed across several values (above and below defaults) for eight of the LOCATOR main parameters.

Overall, the LOCATOR analysis was repeated 68 times with different parameter values. Each of these separate runs was judged by its minimum validation loss (i.e., how well the locations of the remaining samples were predicted after training). Ultimately, the best-performing LOCATOR run presented here was chosen because it had the lowest validation loss (i.e., the best-trained model for predicting the data). This model included default parameters except for training split = 0.70, minor allele count = 20, and SNPs = 40,000. The parameters file is included in Appendix 3.

RESULTS AND DISCUSSION

Sequencing data for the seven target individuals was comparable in quality to the remaining samples (Table 2). Although missing data in the target samples were on average 17% higher than the mean of the remaining samples (43% vs. 26%), we note that Phase-5 samples overall had a higher mean percentage of missing data = 39% (Table 3). Missing data is typical for RAD-seq data panels and is less consequential the more total SNPs are cataloged.

Filtered DNA alignments with 1,306 individuals contained *N*=262,613 SNPs. The geographic location determined by AGFC for each of the seven target samples is provided in Table 1.

Individuals originating outside Management Zone

The predicted geolocation of origin for two CWD positive individuals appears to be from outside the current CWD Management Zone (MZ) (Figures 3 & 4).

Sample AR047669/83RA5P011 was collected in Randolph County, and its predicted origin based on genotype is likely Sharp County (neighboring county just to the east of Randolph County). Out of 100 independent replicates predicting the origination location, 86 were outside the MZ (Figure 3).

Sample AR05433/83UN5P008 was collected in Union County, and its genotype confirms this as the most likely county of origin. All 100 independent replicates placed this sample's origin outside the MZ (Figure 4).

Individuals originating within Management Zone

The remaining five individuals that tested positive for CWD likely originated within the MZ. Predicted origin for these samples generally associated closely with their sampling location (Figures 5–7), except for two individuals sampled in western Arkansas (Figures 8 & 9).

Predicted origin for two samples associated closely with the sampling location. Sample AR050916/83BO5P069 has a predicted origin concentrated around its collection location in Boone County (Figure 5). The same is true for AR047126/83VB5P036; its origin is likely near the sampling point in Van Buren County (Figure 6).

We did not have precise sample location information for AR048108/83NW5P324 other than it was collected in Newton County; its location of origin based on genotype is likely just to the south in either Johnson or Pope County (Figure 7).

Two samples collected on the western side of the MZ in Crawford (AR058733/83CW5P012) and Franklin (AR049847/83FR5P036) counties, respectively, display an unusual pattern of predicted origin: most predictions are concentrated around Independence county (eastern MZ), with some predictions scattered across the western MZ (Figures 8 & 9).

Upon first look, this might be inferred as long-distance dispersal out of Independence County. Alternatively, these samples could have originated from Oklahoma, where we lack genotypic training data, and therefore the model could be unable to accurately place these individuals. This alternative may be a more straightforward explanation, given that the location of the sampled deer was much closer to the Oklahoma border than to Independence County.

However, comparing the predicted locations for these two individuals from the western MZ to probabilistic distribution of 'genetic herds' (subpopulations) identified in the state-wide analysis of WTD offers another explanation (Figure 10). Assignment probabilities for subpopulation *k3* (shown in Figure 11 of Douglas et al., 2020) reflects a similar spatial distribution, with highest probabilities in west-central, north-central and south-western Arkansas. This pattern lacking spatial cohesion was interpreted as a signal of historic translocations, with a major source coming from the Sylamore District in the Ozark Mountains (Chafin et al., 2021). Given the spatial congruence between sampling and predicted locations for the Crawford and Franklin county samples and modeled distribution of subpopulation k3, the interpretation that these signals reflect historic translocations seems most likely.

REFERENCES

Battey, C. J., Ralph, P. L., & Kern, A. D. (2020). Predicting geographic location from genetic variation with deep neural networks. *ELife*, *9*, e54507.

Chafin, T. K., Douglas, M. R., Martin, B. T., Zbinden, Z. D., Middaugh, C. R., Ballard, J. R., Gray, M. C., White Jr, D. & Douglas, M. E. (2020). Age structuring and spatial heterogeneity in prion protein gene (*PRNP*) polymorphism in white-tailed deer. *Prion, 14*(1), 238-248.

Chafin, T. K., Zbinden, Z. D., Douglas, M. R., Martin, B. T., Middaugh, C. R., Gray, M. C., Ballard, J. R. & 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.

Douglas, M. R., Chafin, T. K., Zbinden, Z. D., Martin, B. T., & Douglas, M. E. (2020). White-tailed deer in Arkansas: Genetic connectivity and chronic wasting disease susceptibility. *Final Report* (10 February 2020), Arkansas Game & Fish Commission, Little Rock, AR. *pp*. 155.

Eaton, D. A., & Overcast, I. (2020). ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics*, *36*(8), 2592-2594.

Martin, B. T., Chafin, T. K., Douglas, M. R., Placyk Jr, J. S., Birkhead, R. D., Phillips, C. A., & Douglas, M. E. (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.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PloS one*, *7*(5), e37135.

TABLE 1: Sampling location information for seven CWD-positive WTD individuals. Sample ID = identifier given by AGFC; DNA ID = lab identifier for DNA sample; County = Arkansas county name; Easting = UTM X; Northing = UTM Y.

Sample ID	DNA ID	County	Easting	Northing	Latitude	Longitude
AR050916	83BO5P069	Boone	482566	4013530	36.266542	-93.194092
AR058733	83CW5P012	Crawford	374577	3938562	35.582834	-94.384318
AR049847	83FR5P036	Franklin	406500	3915569	35.379078	-94.029391
AR048108	83NW5P324	Newton	-	-	-	-
AR047669	83RA5P011	Randolph	660367	4033065	36.42941	-91.211005
AR054533	83UN5P008	Union	578326	3664774	33.118829	-92.160419
AR047126	83VB5P036	Van Buren	530649	3946963	35.666056	-92.66136

TABLE 2: Summary statistics of DNA sequence data for seven CWD-positive WTD individuals (i.e., target samples) compared to the remaining samples (Remainder; *N*=1420). Sample ID = identifier given by AGFC; DNA ID = lab identifier for DNA sample; Raw Reads = the number of fragments sequenced; Mean Depth = the average amount of redundancy for each fragment sequenced; Heterozygosity = the proportion of heterozygous alleles; Est Error = estimated error in genotype calls.

Sample ID	DNA ID	Raw Reads	Mean Depth	Heterozygosity	Est Error
AR050916	83BO5P069	7,173,398	109	0.0061	0.0018
AR058733	83CW5P012	3,307,285	67	0.0057	0.0018
AR049847	83FR5P036	2,522,040	51	0.0053	0.0019
AR048108	83NW5P324	3,069,793	58	0.0051	0.0018
AR047669	83RA5P011	3,713,502	68	0.0055	0.0018
AR054533	83UN5P008	4,159,566	74	0.0053	0.0017
AR047126	83VB5P036	3,309,462	62	0.0056	0.0018
REMAINDER (N=1,420):					
	Mean	3,597,032	57	0.0063	0.0017
	Stdev	1,939,756	20	0.0021	0.0016

TABLE 3: Proportion of Missing data and Mean Depth for individuals within SNP panels post-filtering. Seven target individuals are compared to the remaining individuals (Remainder; *N*=1,299). Sample ID = identifier given by AGFC; DNA ID = lab identifier for DNA sample; Missing = proportion of missing SNP data; Mean Depth = average amount of redundancy for each SNP call.

Sample ID	DNA ID	Missing	Mean Depth
AR050916	83BO5P069	0.40	76
AR058733	83CW5P012	0.47	39
AR049847	83FR5P036	0.45	29
AR048108	83NW5P324	0.43	36
AR047669	83RA5P011	0.41	45
AR054533	83UN5P008	0.41	50
AR047126	83VB5P036	0.43	38
REMAINDER (N=1,299):			
	Mean	0.26	47
	Stdev	0.18	27

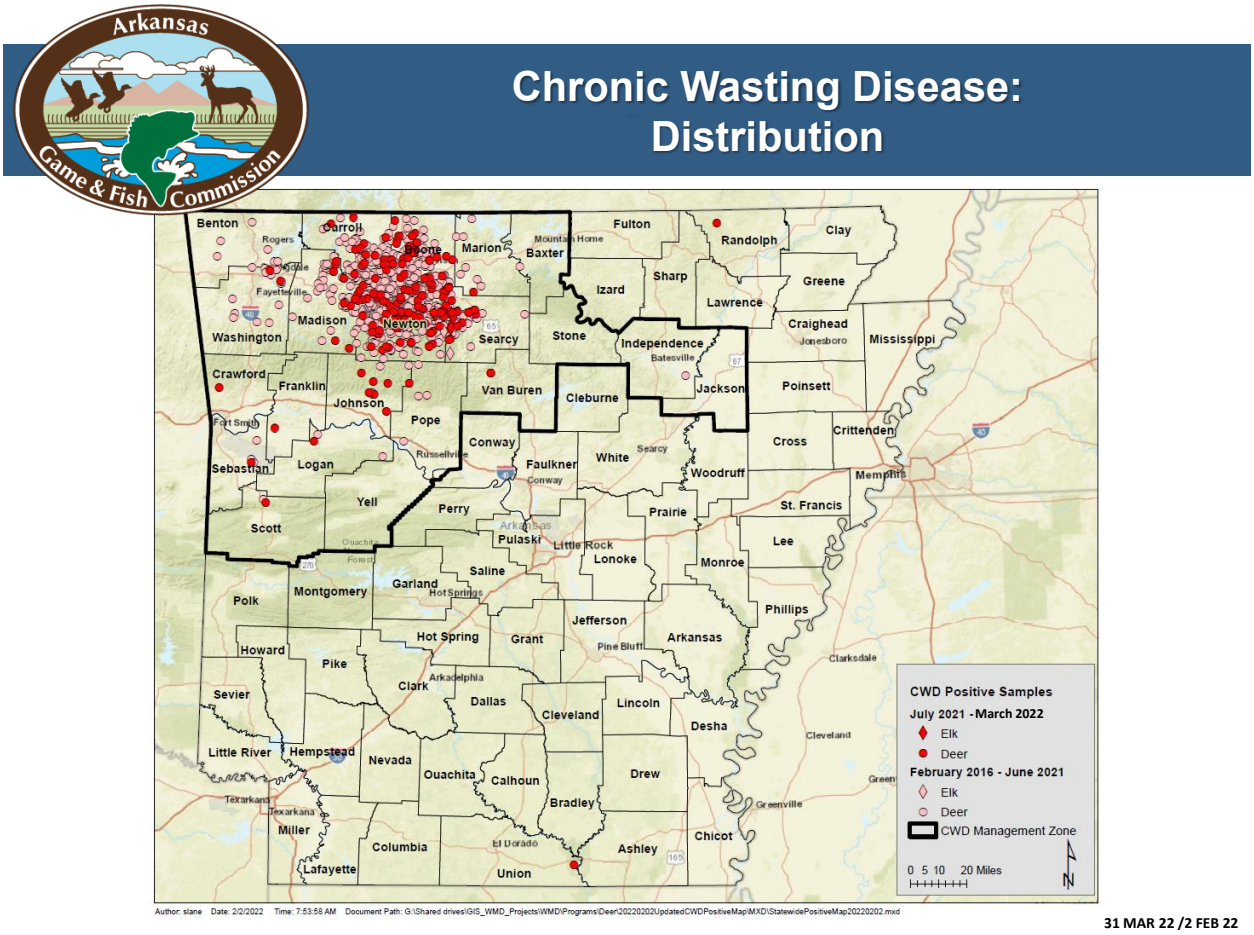

FIGURE 1: Status of Chronic Wasting Disease (CWD) detections in Arkansas as of 31 March 2022. The map shows geographic locations of white-tailed deer (WTD; dots) and elk (diamonds) that tested positive for CWD. Red: samples collected June 2021 through March 2022; pink = samples collected prior to June 2021. The CWD Management Zone (MZ) is outlined in black. From: https://www.agfc.com/en/hunting/big-game/deer/cwd/cwd-arkansas/ (accessed 27 June 2022).

CWD Detections by County by Fiscal Year* for White-tailed Deer and Elk

	FY2016		FY2017		FY2018		FY2019		FY2020		FY2021		FY2022		Totals		
	+WTD	+Elk	+WTD	+Elk	+WTD	+Elk	+WTD	+Elk	+WTD	+Elk	+WTD	+Elk	+WTD	+Elk	+WTD	+Elk	Total
Benton					2		1		2		2		2		9	0	9
Boone	5		7		24		34		54		45		38		207	0	207
Carroll	2		19		21		34		25		23		21		145	0	145
Crawford													1		1	0	1
Franklin													1		1	0	1
Independence									1						1	0	1
Johnson							1				5		7		13	0	13
Logan											2		1		3	0	3
Madison	1		6		12		28		6		18	1	21		92	1	93
Marion			2		2		2		1		1		1		9	0	9
Newton	87	5	78		79	4	123	5	121	4	140	6	77	4	705	28	733
Pope	1		1						1		1				4	0	4
Randolph													1		1	0	1
Scott							1						1		2	0	2
Searcy			1	2	3	3	10		5	1	22	1	19	4	60	11	71
Sebastian					1		1				1		1		4	0	4
Union													1		1	0	1
Van Buren													2		2	0	2
Washington					3		6		6		7		2		24	0	24
	96	5	114	2	147	7	241	5	222	5	267	8	197	8	1,284	40	1,324
	101		116		154		246		227		275		205				

AS OF 31 MAR 22
* = FY or Fiscal Year = July 1st to June 30th
FY2022 = Current Sampling Year In Progress

FIGURE 2: Overview of wild cervids in Arkansas that tested positive for Chronic Wasting Disease (CWD) as of 31 March 2022. Listed: Counties in Arkansas where CWD positive samples were detected; numbers of WTD and elk that tested positive by year; and totals. The 17 counties included in the FY2022 Management Zone are shown in Figure 1. First occurrence of CWD positive samples in 2022 were recorded in Crawford, Franklin, Randolph, Union and Van Buren counties. From: https://www.agfc.com/en/hunting/big-game/deer/cwd/cwd-arkansas/ (accessed 27 June 2022).

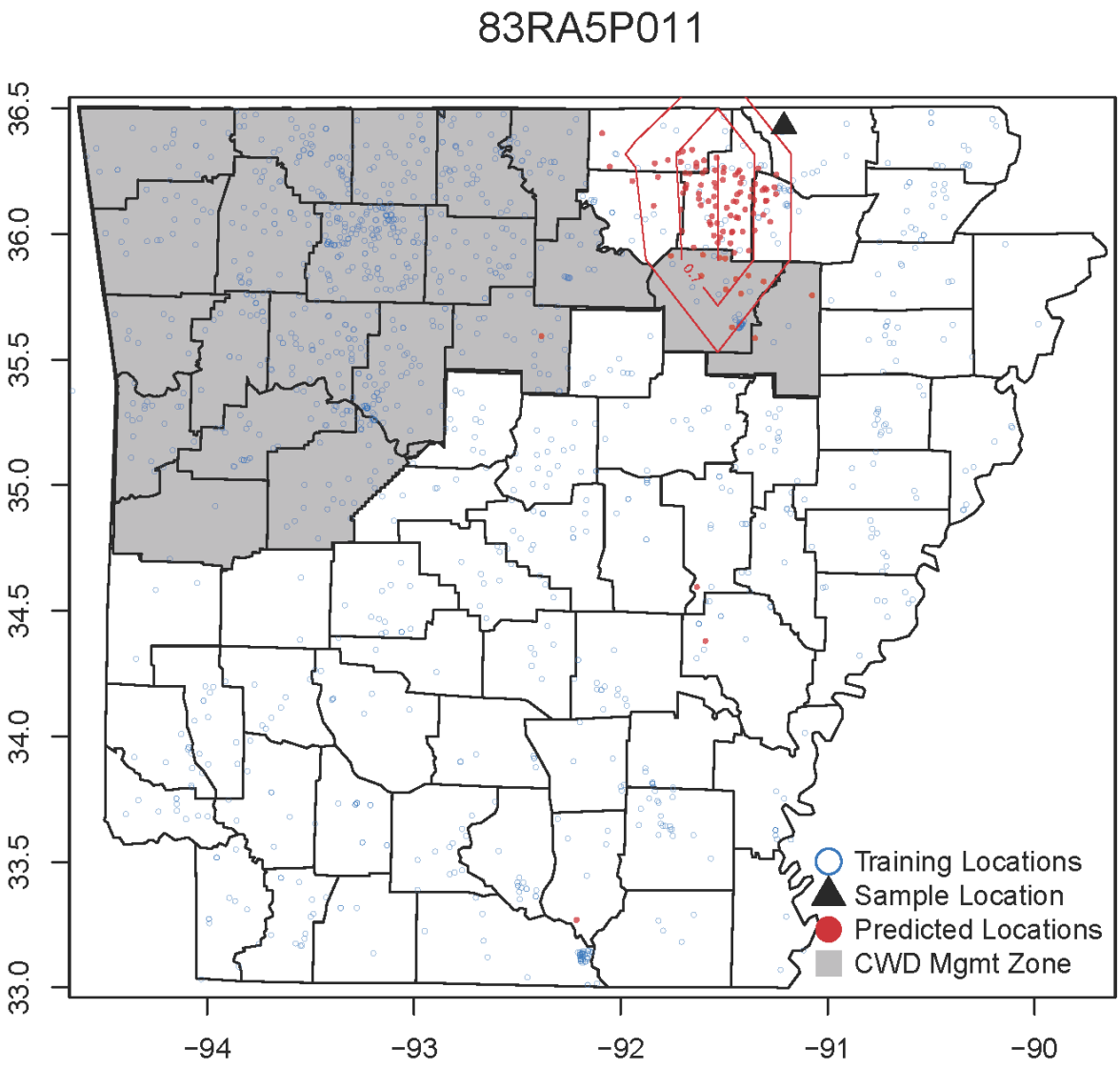

FIGURE 3: Locator predicted points of origin (red) for sample AR047669/83RA5P011. Locations are predicted based on the genotypes of training individuals and their locations. Lines drawn around predicted locations represent contours that contain 90%, 70%, and 50% of the predicted samples.

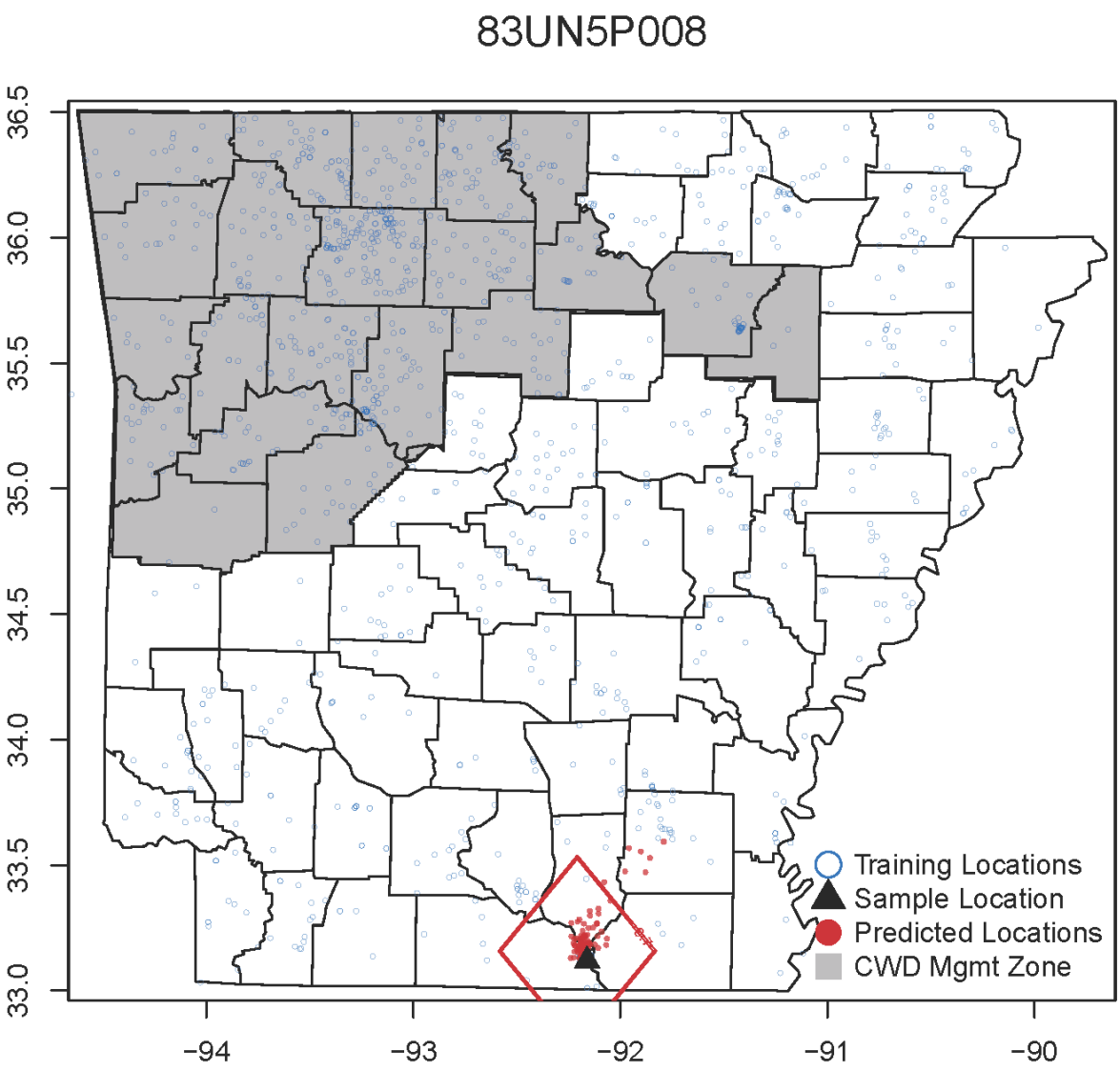

FIGURE 4: Locator predicted points of origin (red) for sample AR05433/83UN5P008. Locations are predicted based on the genotypes of training individuals and their locations. Lines drawn around predicted locations represent contours that contain 90%, 70%, and 50% of the predicted samples.

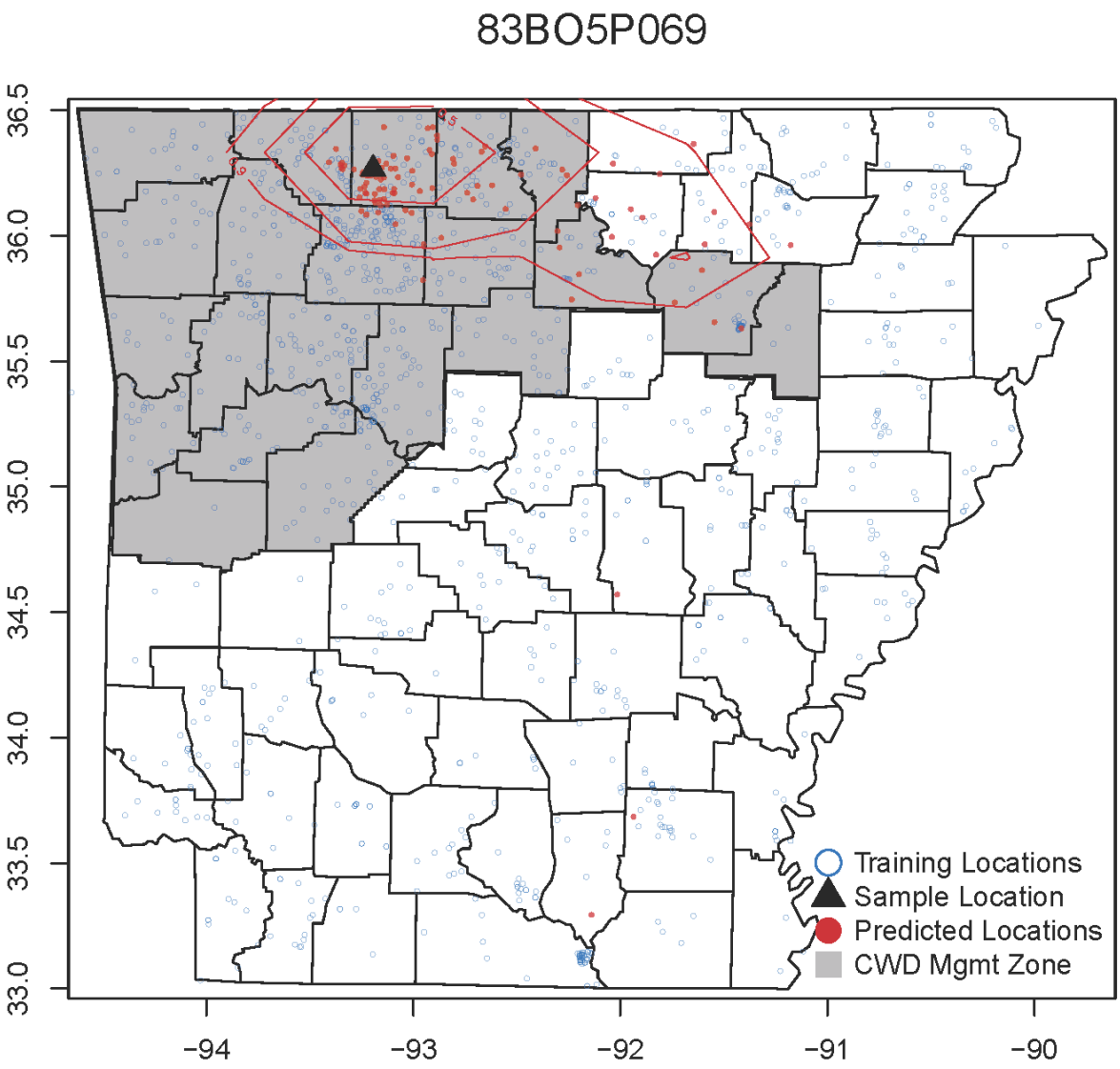

FIGURE 5: Locator predicted points of origin (red) for sample AR050916/83BO5P069. Locations are predicted based on the genotypes of training individuals and their locations. Lines drawn around predicted locations represent contours that contain 90%, 70%, and 50% of the predicted samples.

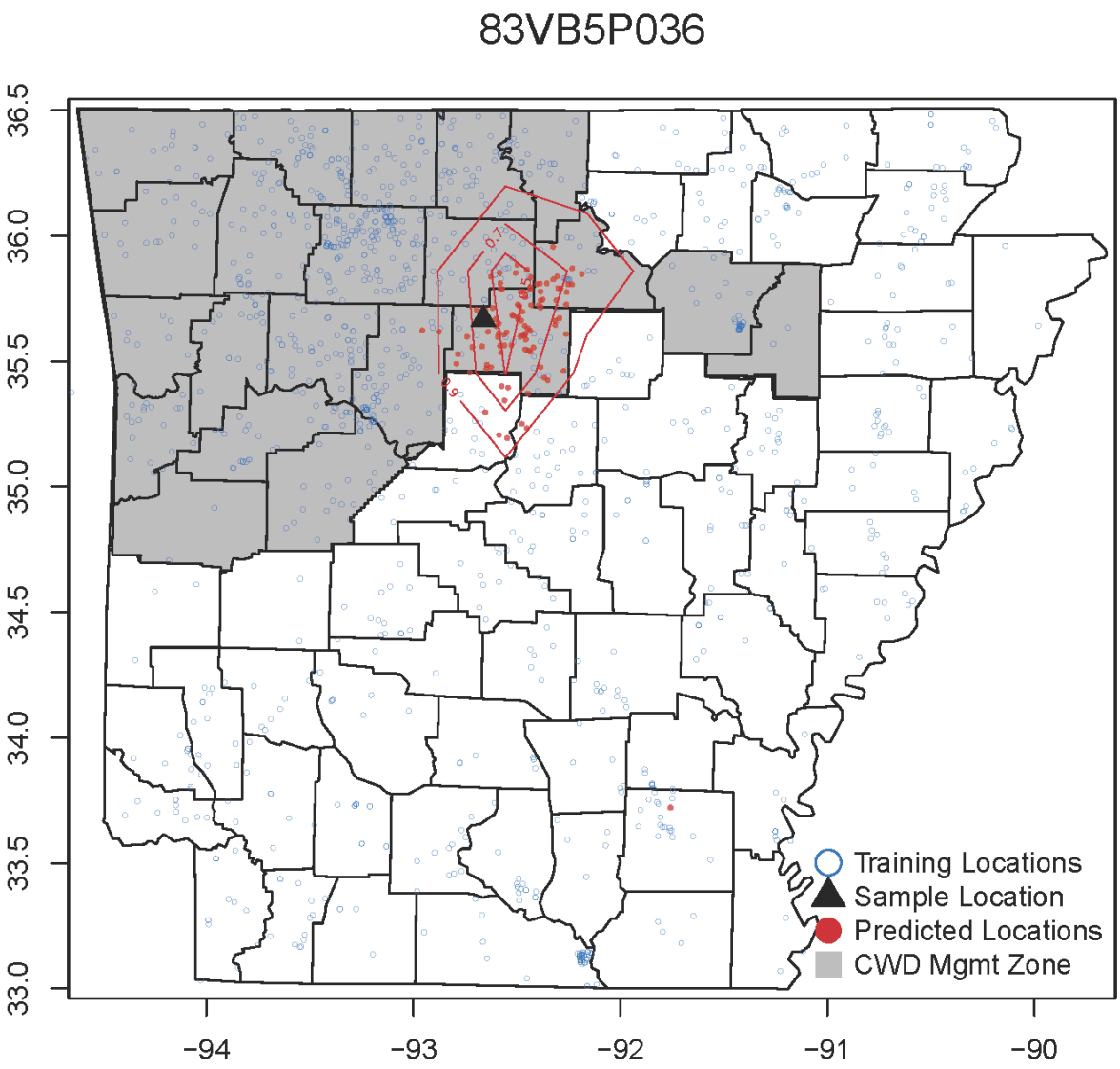

FIGURE 6: Locator predicted points of origin (red) for sample AR047126/83VB5P036. Locations are predicted based on the genotypes of training individuals and their locations. Lines drawn around predicted locations represent contours that contain 90%, 70%, and 50% of the predicted samples.

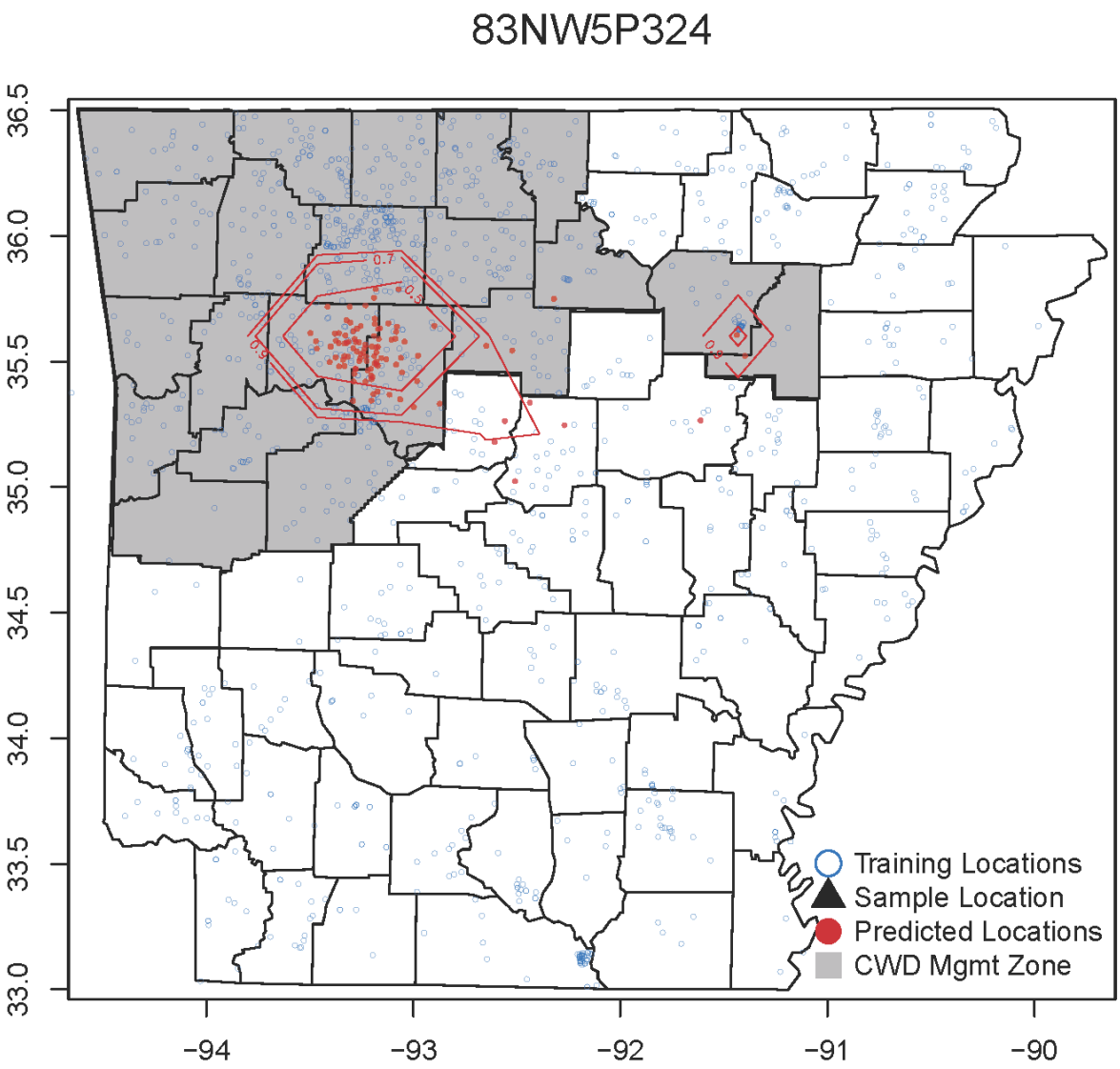

FIGURE 7: Locator predicted points of origin (red) for sample AR048108/83NW5P324. Locations are predicted based on the genotypes of training individuals and their locations. Lines drawn around predicted locations represent contours that contain 90%, 70%, and 50% of the predicted samples. The location of the sample was unknown for this individual.

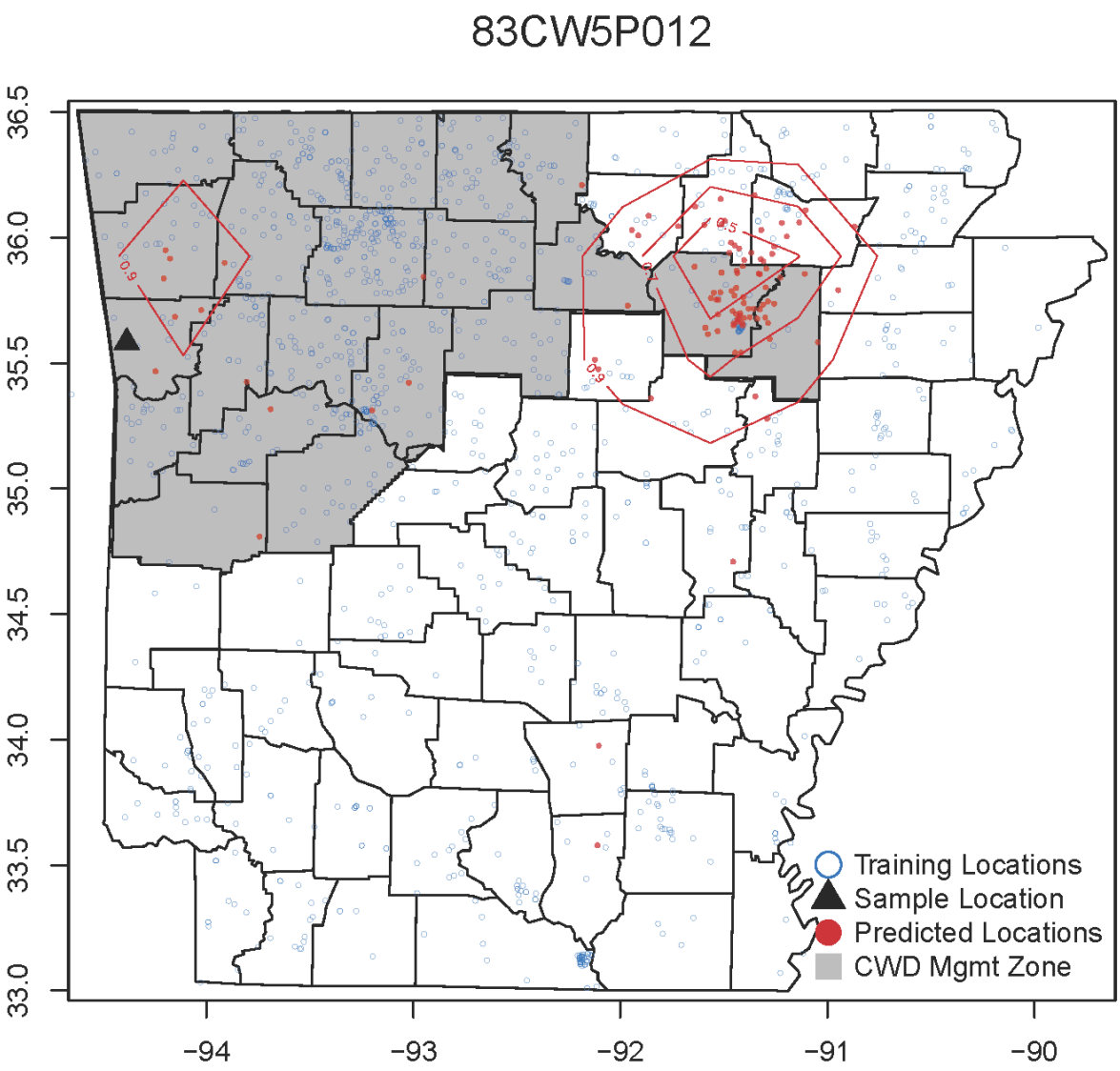

FIGURE 8: Locator predicted points of origin (red) for sample AR058733/83CW5P012. Locations are predicted based on the genotypes of training individuals and their locations. Lines drawn around predicted locations represent contours that contain 90%, 70%, and 50% of the predicted samples.

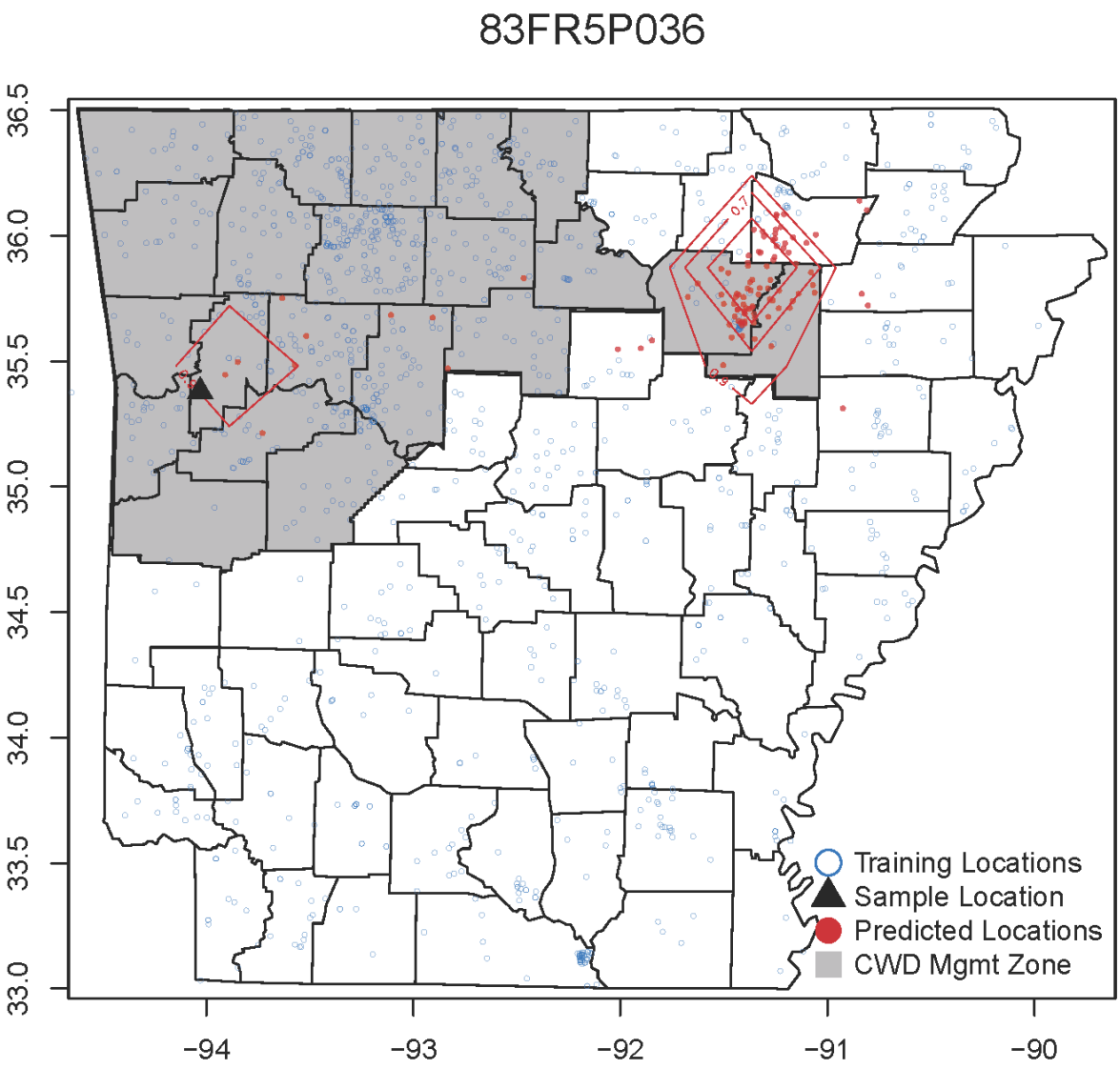

FIGURE 9: Locator predicted points of origin (red) for sample AR049847/83FR5P036. Locations are predicted based on the genotypes of training individuals and their locations. Lines drawn around predicted locations represent contours that contain 90%, 70%, and 50% of the predicted samples.

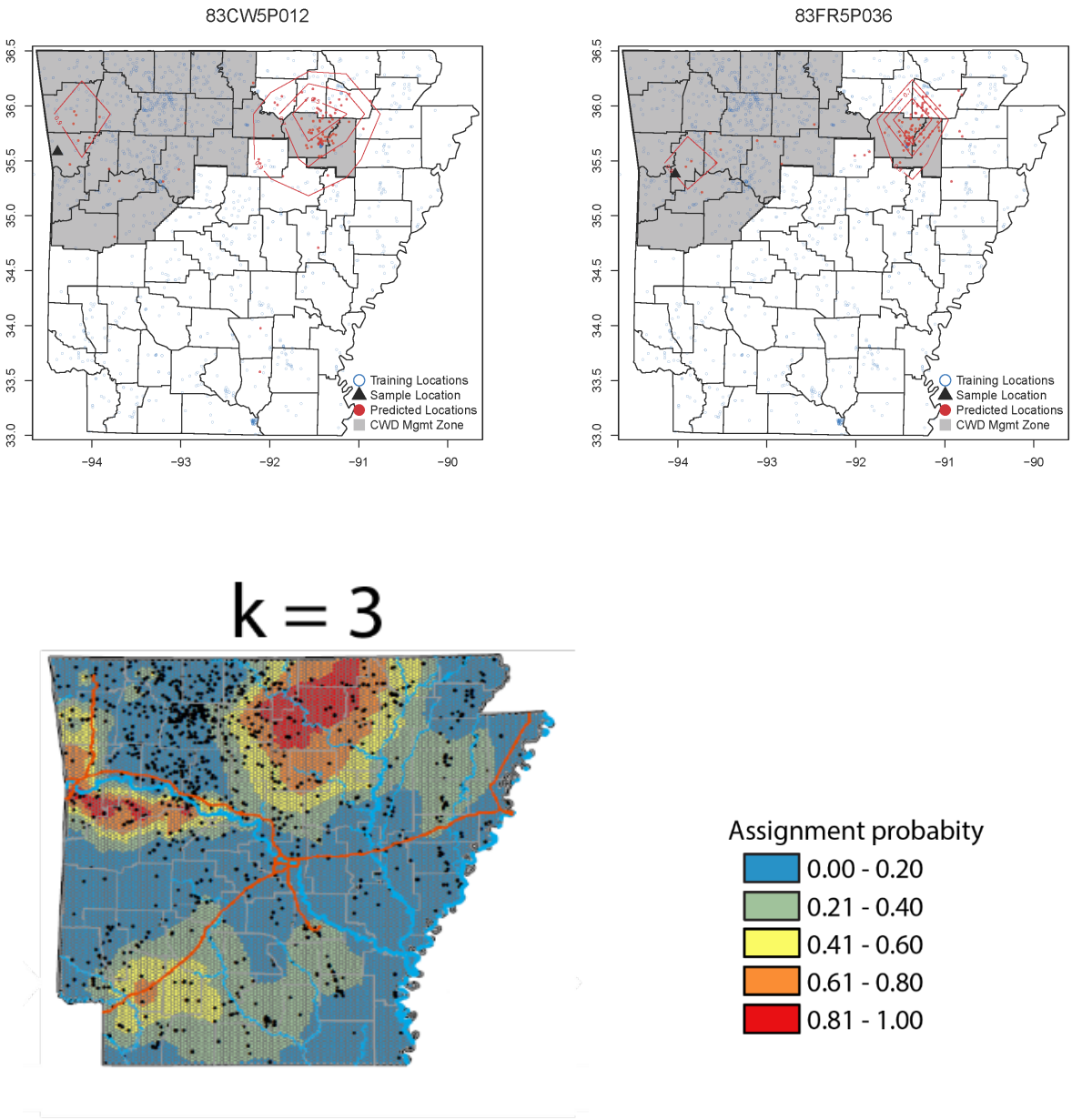

FIGURE 10: Comparison of spatial patterns of origin predictions *versus* genetic subpopulation of white-tailed deer in Arkansas. Top panels depict predict origins for two individuals sampled in FY2022 and analyzed in this study: (A) Crawford County (AR058733/83CW5P012 – Figure 8), and (B) Franklin County (AR049847/83FR5P036 – Figure 9). Bottom panel depicts probabilistic distribution of subpopulation *k3* identified in a state-wide analysis of 1,143 sampled from 2016-2019 (Figure 2 in Chafin et al., 2021).

APPENDIX 1

Overview of 146 white-tailed deer samples collected in 2021 and 2022 and genotyped in this study. Listed are: Sample ID = identifier given by AGFC; DNA ID = lab identifier for DNA sample; County = county where sample was collected; FY = fiscal year when sample was collected; Status = testing result for Chronic Wasting Disease (CWD); Comment = figure in this report depicting predicted location for sample.

SampleID	DNA ID	County	FY	Status	Comment
AR032436	83AS5N3	Ashley	2021	Negative	
AR052601	83AS5N4	Ashley	2022	Negative	
AR031204	83XX5N3	Baxter	2022	Negative	
AR043130	83BO5N68	Boone	2022	Negative	
AR050916	83BO5P69	Boone	**2022**	**Positive**	**Figure 5**
AR032409	83BR5N4	Bradley	2021	Negative	
AR052679	83BR5N5	Bradley	2021	Negative	
AR035750	83CL5N18	Clark	2021	Negative	
AR035732	83XX5N4	Clark	2022	Negative	
AR052696	83CV5N19	Cleveland	2022	Negative	
AR009091	83CN5N20	Conway	2021	Negative	
AR041372	83CG5N16	Craighead	2022	Negative	
AR058733	83CW5P12	**Crawford**	**2022**	**Positive**	**Figure 8**
AR000310	83CT5N9	Crittenden	2021	Negative	
AR052693	83DS5N13	Desha	2021	Negative	
AR000518	83DR5N18	Drew	2021	Negative	
AR000525	83DR5N19	Drew	2021	Negative	
AR035029	83DR5N20	Drew	2022	Negative	
AR052602	83DR5N21	Drew	2022	Negative	
AR052695	83DR5N22	Drew	2022	Negative	
AR007149	83FA5N23	Faulkner	2022	Negative	
AR009478	83FA5N24	Faulkner	2021	Negative	
AR052181	83FA5N25	Faulkner	2021	Negative	
AR033764	83FR5N35	Franklin	2021	Negative	
AR049847	83FR5P36	**Franklin**	**2022**	**Positive**	**Figure 9**
AR026106	83FU5N23	Fulton	2022	Negative	
AR034975	83FU5N24	Fulton	2022	Negative	
AR052694	83GR5N12	Grant	2022	Negative	
AR054301	83GE5N20	Greene	2022	Negative	
AR027653	83HO5N15	Howard	2022	Negative	
AR027657	83HO5N16	Howard	2022	Negative	
AR042315	83HO5N17	Howard	2022	Negative	
AR027650	83HO5N18	Howard	2022	Negative	

AR042311	83HO5N19	Howard	2022	Negative	
AR042312	83HO5N20	Howard	2022	Negative	
AR041761	83IN5N13	Independence	2021	Negative	
AR041783	83IN5N14	Independence	2022	Negative	
AR041784	83IN5N15	Independence	2022	Negative	
AR041785	83IN5N16	Independence	2022	Negative	
AR041786	83IN5N17	Independence	2022	Negative	
AR041787	83IN5N18	Independence	2022	Negative	
AR041788	83IN5N19	Independence	2022	Negative	
AR041789	83IN5N20	Independence	2022	Negative	
AR041790	83IN5N21	Independence	2022	Negative	
AR041791	83IN5N22	Independence	2022	Negative	
AR041792	83IN5N23	Independence	2022	Negative	
AR041793	83IN5N24	Independence	2022	Negative	
AR041863	83IN5N25	Independence	2021	Negative	
AR041864	83IN5N26	Independence	2021	Negative	
AR041865	83IN5N27	Independence	2021	Negative	
AR041866	83IN5N28	Independence	2021	Negative	
AR041867	83IN5N29	Independence	2021	Negative	
AR041868	83IN5N30	Independence	2021	Negative	
AR041869	83IN5N31	Independence	2021	Negative	
AR041870	83IN5N32	Independence	2021	Negative	
AR041871	83IN5N33	Independence	2021	Negative	
AR041872	83IN5N34	Independence	2021	Negative	
AR041873	83IN5N35	Independence	2021	Negative	
AR041874	83IN5N36	Independence	2021	Negative	
AR041875	83IN5N37	Independence	2021	Negative	
AR041876	83IN5N38	Independence	2021	Negative	
AR041877	83IN5N39	Independence	2021	Negative	
AR041878	83IN5N40	Independence	2021	Negative	
AR041880	83IN5N41	Independence	2021	Negative	
AR020681	83IZ5N12	Izard	2021	Negative	
AR041365	83IZ5N13	Izard	2021	Negative	
AR041366	83IZ5N14	Izard	2021	Negative	
AR041367	83IZ5N15	Izard	2021	Negative	
AR041368	83IZ5N16	Izard	2021	Negative	
AR041369	83IZ5N17	Izard	2021	Negative	
AR041763	83IZ5N18	Izard	2021	Negative	
AR032513	83JE5N15	Jefferson	2022	Negative	
AR016748	83LA5N17	Lafayette	2021	Negative	
AR056662	83XX5N5	Lafayette	2022	Negative	no ddRAD
AR056663	83XX5N6	Lafayette	2022	Negative	no ddRAD
AR031726	83LW5N21	Lawrence	2022	Negative	

AR057235	83LI5N10	Lincoln	2022	Negative	
AR042313	83LR5N11	Little River	2022	Negative	
AR042314	83LR5N12	Little River	2022	Negative	
AR019765	83MI5N13	Miller	2021	Negative	
AR041370	83MS5N4	Mississippi	2022	Negative	
AR008145	83NE5N15	Nevada	2022	Negative	
AR048108	83NW5P324	**Newton**	**2022**	**Positive**	**Figure 7** (no UTM)
AR027658	83PI5N9	Pike	2022	Negative	
AR042316	83PI5N10	Pike	2022	Negative	
AR035802	83PO5N15	Poinsett	2021	Negative	
AR035803	83PO5N16	Poinsett	2021	Negative	
AR041371	83PO5N17	Poinsett	2022	Negative	
AR001231	83PU5N19	Pulaski	2021	Negative	
AR008306	83PU5N20	Pulaski	2021	Negative	
AR009424	83PU5N21	Pulaski	2021	Negative	
AR012204	83PU5N22	Pulaski	2021	Negative	
AR012211	83PU5N23	Pulaski	2021	Negative	
AR012214	83PU5N24	Pulaski	2022	Negative	
AR012217	83XX5N2	Pulaski	2022	Negative	
AR047669	83RA5P11	**Randolph**	**2022**	**Positive**	**Figure 3**
AR020609	83ST5N18	Stone	2022	Negative	
AR054533	83UN5P8	**Union**	**2022**	**Positive**	**Figure 4**
AR055501	83UN5N9	Union	2022	Negative	
AR055502	83UN5N10	Union	2022	Negative	
AR055503	83UN5N11	Union	2022	Negative	
AR055595	83UN5N12	Union	2022	Negative	
AR055596	83UN5N13	Union	2022	Negative	
AR055597	83UN5N14	Union	2022	Negative	
AR055598	83UN5N15	Union	2022	Negative	
AR055599	83UN5N16	Union	2022	Negative	not analyzed
AR055600	83UN5N17	Union	2022	Negative	
AR060961	83UN5N18	Union	2022	Negative	
AR060962	83UN5N19	Union	2022	Negative	
AR060963	83UN5N20	Union	2022	Negative	
AR060964	83UN5N21	Union	2022	Negative	
AR060965	83UN5N22	Union	2022	Negative	
AR060966	83UN5N23	Union	2022	Negative	
AR060967	83UN5N24	Union	2022	Negative	
AR060968	83UN5N25	Union	2022	Negative	
AR060969	83UN5N26	Union	2022	Negative	
AR060970	83UN5N27	Union	2022	Negative	
AR060971	83UN5N28	Union	2022	Negative	
AR060972	83UN5N29	Union	2022	Negative	

AR060973	83UN5N30	Union	2022	Negative	
AR060974	83UN5N31	Union	2022	Negative	
AR060975	83UN5N32	Union	2022	Negative	
AR060976	83UN5N33	Union	2022	Negative	
AR060977	83UN5N34	Union	2022	Negative	
AR060978	83UN5N35	Union	2022	Negative	
AR060979	83UN5N36	Union	2022	Negative	
AR060980	83UN5N37	Union	2022	Negative	
AR060981	83UN5N38	Union	2022	Negative	
AR060982	83UN5N39	Union	2022	Negative	
AR060983	83UN5N40	Union	2022	Negative	
AR060984	83UN5N41	Union	2022	Negative	
AR060985	83UN5N42	Union	2022	Negative	
AR060986	83UN5N43	Union	2022	Negative	
AR060987	83UN5N44	Union	2022	Negative	
AR060988	83UN5N45	Union	2022	Negative	
AR060989	83UN5N46	Union	2022	Negative	
AR060990	83UN5N47	Union	2022	Negative	
AR060991	83UN5N48	Union	2022	Negative	
AR060992	83UN5N49	Union	2022	Negative	
AR060993	83UN5N50	Union	2022	Negative	
AR060994	83UN5N51	Union	2022	Negative	
AR009431	83VB5N35	Van Buren	2022	Negative	
AR047126	83VB5P36	**Van Buren**	**2022**	**Positive**	**Figure 6**
AR009432	83XX5N1	Van Buren	2022	Negative	
AR031168	83WD5N12	Woodruff	2021	Negative	
AR000138	83YE5N56	Yell	2021	Negative	

APPENDIX 2.

IPYRAD parameters file used for clustering and assembling of 1,427 WTD DNA samples collected across Arkansas.

```
------- ipyrad params file (v.0.9.62)-------------------------------------------
wtdphase5                      ## [0] [assembly_name]: Assembly name. Used to name output directories for assembly steps
./                             ## [1] [project_dir]: Project dir (made in curdir if not present)
                               ## [2] [raw_fastq_path]: Location of raw non-demultiplexed fastq files
                               ## [3] [barcodes_path]: Location of barcodes file
~/wtdPHASE5/fastq/*.gz         ## [4] [sorted_fastq_path]: Location of demultiplexed/sorted fastq files
reference                      ## [5] [assembly_method]: Assembly method (denovo, reference)
~/wtdPHASE5/genome/NEB_GCA_014726795.1_Odo_v1_genomic.fna ## [6] [reference_sequence]: Location of reference sequence file
ddrad                          ## [7] [datatype]: Datatype (see docs): rad, gbs, ddrad, etc.
TGCAG,CG                       ## [8] [restriction_overhang]: Restriction overhang (cut1,) or (cut1, cut2)
4                              ## [9] [max_low_qual_bases]: Max low quality base calls (Q<20) in a read
33                             ## [10] [phred_Qscore_offset]: phred Q score offset (33 is default and very standard)
20                             ## [11] [mindepth_statistical]: Min depth for statistical base calling
20                             ## [12] [mindepth_majrule]: Min depth for majority-rule base calling
10000                          ## [13] [maxdepth]: Max cluster depth within samples
0.85                           ## [14] [clust_threshold]: Clustering threshold for de novo assembly
0                              ## [15] [max_barcode_mismatch]: Max number of allowable mismatches in barcodes
2                              ## [16] [filter_adapters]: Filter for adapters/primers (1 or 2=stricter)
35                             ## [17] [filter_min_trim_len]: Min length of reads after adapter trim
2                              ## [18] [max_alleles_consens]: Max alleles per site in consensus sequences
0.05                           ## [19] [max_Ns_consens]: Max N's (uncalled bases) in consensus
0.05                           ## [20] [max_Hs_consens]: Max Hs (heterozygotes) in consensus
700                            ## [21] [min_samples_locus]: Min # samples per locus for output
0.2                            ## [22] [max_SNPs_locus]: Max # SNPs per locus
8                              ## [23] [max_Indels_locus]: Max # of indels per locus
0.5                            ## [24] [max_shared_Hs_locus]: Max # heterozygous sites per locus
0, 0, 0, 0                     ## [25] [trim_reads]: Trim raw read edges (R1>, <R1, R2>, <R2) (see docs)
0, 10, 0, 0                    ## [26] [trim_loci]: Trim locus edges (see docs) (R1>, <R1, R2>, <R2)
*                              ## [27] [output_formats]: Output formats (see docs)
                               ## [28] [pop_assign_file]: Path to population assignment file
                               ## [29] [reference_as_filter]: Reads mapped to this reference are removed in step 3
```

APPENDIX 3.

LOCATOR analysis parameters for WTD Phase 5 geolocation prediction of target samples.


```
"vcf": "/home/zdzbinde/wtdPHASE5/locate/vcfiles/3_filtered_N1306.vcf",
"zarr": null,
"matrix": null,
"sample_data": "/home/zdzbinde/wtdPHASE5/locate/indiv_coords/4_filtered_samplesANDcoords_N1306.txt",
"train_split": 0.7,
"windows": false,
"window_start": 0,
"window_stop": null,
"window_size": 500000.0,
"bootstrap": true,
"jacknife": false,
"jacknife_prop": 0.05,
"nboots": 100,
"batch_size": 32,
"max_epochs": 5000,
"patience": 100,
"min_mac": 20,
"max_SNPs": 40000,
"impute_missing": false,
"dropout_prop": 0.25,
"nlayers": 10,
"width": 256,
"out": "40koutfile",
"seed": null,
"gpu_number": null,
"plot_history": true,
"gnuplot": false,
"keep_weights": false,
"load_params": null,
"keras_verbose": 1
```