OSAC 2025-S-0021 Standard for Validation of Multilocus Databases

Wildlife Forensic Biology Subcommittee
Biology Scientific Area Committee (SAC)
Organization of Scientific Area Committees (OSAC) for Forensic Science

OSAC Proposed Standard

OSAC 2025-S-0021
Standard for Validation of Multilocus Databases

Prepared by
Wildlife Forensic Biology Subcommittee
Version: 1.0
June 2025

Disclaimer:

This OSAC Proposed Standard was written by the Wildlife Forensics Biology Subcommittee of the Organization of Scientific Area Committees (OSAC) for Forensic Science following a process that includes an open comment period. This Proposed Standard will be submitted to a standard developing organization and is subject to change.

There may be references in an OSAC Proposed Standard to other publications under development by OSAC. The information in the Proposed Standard, and underlying concepts and methodologies, may be used by the forensic-science community before the completion of such companion publications.

Any identification of commercial equipment, instruments, or materials in the Proposed Standard is not a recommendation or endorsement by the U.S. Government and does not imply that the equipment, instruments, or materials are necessarily the best available for the purpose.

To be placed on the OSAC Registry, certain types of standards receive a Scientific and Technical Review (STR). The STR process is vital to OSAC's mission of generating and recognizing scientifically sound standards for producing and interpreting forensic science results. The STR shall provide critical and knowledgeable reviews of draft standards to ensure that the published methods that practitioners employ are scientifically valid, and the resulting claims are trustworthy.

The STR consists of an independent and diverse panel, which may include subject matter experts, human factors scientists, quality assurance personnel, and legal experts as applicable. The selected group is tasked with evaluating the proposed standard based on a defined list of scientific, administrative, and quality assurance based criteria.

For more information about this important process, please visit our website at: https://www.nist.gov/organization-scientific-area-committees-forensic-science/scientific-technical-review-str-process

Foreword

This standard provides requirements for validating multilocus population genetic databases for wildlife forensics. The aim is to provide consistency in the wildlife forensics community. Forensic scientists using this standard are expected to have a working knowledge of sample acquisition, sample curation, DNA genotyping, population genetic theory and analyses, and the life histories of the species of interest. They are also expected to have a quality management system in place and documented procedures and protocols for all methods used.

Validated multilocus databases are intended for use in population genetic analyses. These databases are essential for accurate comparison among the individual subjects (e.g., individualization, relatedness) and genetic assignment (e.g., source population, geographic origin, taxonomic group).

Keywords: *wildlife forensics, population database, population genetics, multilocus, DNA, validation*

Table of Contents

Standard for Validation of Multilocus Databases

1 Scope

This standard sets forth the minimum requirements that shall be met when validating multilocus population genetic databases for wildlife forensics. This document covers validation of a multilocus population database for specific applications, such as individual and familial relationship evaluation, population assignment, or other scientific techniques performed in wildlife forensic casework. This document does not cover the construction of multilocus databases (e.g., criteria for the identification of samples or inclusion of associated biological information), reference collections obtained for the purpose of test development, or publishing databases. This document only applies to databases generated from reference samples and does not include samples derived from evidence items.

These minimum standards are not intended to replace standards in ISO 17025 or additional forensic laboratory standards but instead provide additional guidance for laboratories validating and modifying multilocus population genetic databases. Notes throughout this document offer clarifications and examples of how a laboratory may meet a specific standard.

2 Normative References

2.1 ANSI/ASB 19 Wildlife Forensics General Standards

2.2 ANSI/ASB 46 Wildlife Forensics Validation Standards—STR Analysis

2.3 ANSI/ASB 48 Wildlife Forensics DNA Standard Procedures

2.4 ANSI/ASB 216 Standard for Construction of Multilocus Databases

3 Terms and Definitions

3.1

false negative rate

A statistical measure that represents the proportion of actual positive cases that are incorrectly identified as negative by a test or classification model.

3.2

false positive rate

A statistical measure that quantifies the proportion of actual negative cases that are incorrectly identified as positive by a test or classification model.

3.3

haplotype

A set of linked DNA variations, or polymorphisms, that tend to be inherited together (e.g., commonly used for mitochondrial or Y-chromosome analysis). A haplotype can refer to a combination of alleles or to a set of single nucleotide polymorphisms (SNPs) found on the same chromosome.

3.4

kinship

The degree of genetic relatedness or shared ancestry between individuals.

3.5

probability of identity (PID)

The probability that two unrelated individuals have the same multilocus genotype.

3.6

private allele

A unique variant found in one population among a group of populations.

4 General Database Requirements

4.1 Protocols covering database validation shall adhere to standards in ANSI/ASB 19, ANSI/ASB 46, ANSI/ASB 48, and ANSI/ASB 216.

4.2 The validation of the constructed multilocus database shall address the criteria assessed during the construction of the database.

4.2.1 The validation shall identify key criteria (e.g., species, geographical region, mating system) that must be met for the database to be fit for its intended use.

4.2.2 These key criteria and records of the analysis showing how the database meets those criteria shall be documented.

5 Database Requirements for Different Applications

A species or population(s) differs based on demographic, ecological, and evolutionary factors, so quantitative values for the minimum number of individuals and genetic markers needed for a reference database are expected to vary according to the specific application, as well as the species or population(s) of interest. Because of the diversity of species, minimum numerical requirements are not feasible. During the process of validating a constructed database, samples or markers may be removed or added from the constructed database. The following standards identify requirements related to particular analysis method applications.

5.1 Individual identification and kinship determination

5.1.1 The PID and PID sibs for the genotypes in the database shall be determined.

5.1.2 The mean and standard deviation of the range of match statistics observed for the genotypes in the database shall be determined.

5.1.3 The false positive and false negative rates shall be estimated for kinship applications.

NOTE In relation to kinship applications, a false positive is concluding that a type of kinship exists when it actually does not; a false negative is excluding a type of kinship when that biological relationship actually exists.

5.1.4 At the conclusion of validation, the diversity, allelic richness, heterozygosity within and among populations, presence of null alleles, probability of identity, linkage disequilibrium, and Hardy-Weinberg equilibrium data for the finalized loci and markers shall be estimated and documented.

NOTE Some of this information may have already been captured during the developmental validation of the multilocus marker panel.

5.1.5 For individual identification application, the database shall include the calculated allelic frequencies of the population/subpopulations and shall use a minimum allele frequency when alleles have not been observed in the database profiles.

NOTE The National Research Council (in *The Evaluation of Forensic DNA Evidence* (1996)) recommended using 5/2N as the minimum allele frequency. This adjustment can be made during database validation or during case sample calculations.

5.1.6 The coefficient of co-ancestry shall be calculated and documented for the population/subpopulations.

NOTE In some subpopulations, calculating this type of statistic (i.e., F_{ST}, F_{IS}) is not possible. In those cases, the use of an estimated value is appropriate. For example, theta values of 0.01– 0.03 are frequently used in human match rarity calculations as a proxy for co-ancestry or population structure. Theta values are often far higher in wildlife species.

5.1.7 The degree of kinship amongst the samples in the database shall be documented.

NOTE Post-hoc analysis to confirm that relatedness in the database is consistent with what is known of the population is appropriate.

NOTE Closely related individuals (i.e., parent-offspring, full and half-siblings) should be minimized in the database, but that is not always possible, such as when working with herd populations with a bottleneck or populations with limited population size.

NOTE Overrepresentation of closely related groups of individuals in a population database may bias allele frequency estimates and cause deviations from Hardy-Weinberg Equilibrium and linkage equilibrium.

5.1.7.1 If the inclusion of related individuals is appropriate, the requirements of Section 5.1.4 still apply.

5.1.8 The database shall include the calculated frequency of each sex-linked STR marker and each sex-linked haplotype.

NOTE This type of database is used for assessing variability for non-autosomal markers (for example, Y-STRs).

5.2 The use of unique variants for taxonomic identification and phenotypic determination.

5.2.1 Multilocus databases used for taxonomic identification or phenotypic determination shall meet the standards listed in Section 5.1 Individual identification and kinship determination.

5.2.2 Unique (e.g., private alleles) and shared variants shall be defined and clearly identified in the database documentation.

5.2.2.1 If an allele previously identified as a private allele is seen in the non-target population, the database documentation shall be updated.

5.2.3 The number of unique (e.g., private alleles) and shared variants required, in order for a taxonomic identification to be made, shall be calculated.

NOTE This can be done with simulated genotypes, empirical data, or both (for example, leave-one-out type tests, jackknifing, or bootstrapping).

5.3 Population assignment analysis

NOTE Different types of assignments (e.g., geographic, temporal, or hybrid) use similar statistical methods, each requiring specialized reference databases. As such, the validation of the reference databases used for this application intrinsically includes validation of the statistical method(s) used in the analysis. While the data that makes up the database have already been validated as reliable during the multiplex panel validation, the statistical method is assessed for reliability during this validation.

5.3.1 Validation of a reference database to be used for population assignment shall include determination of the type and number of statistical methods that have discriminatory power for population assignment.

NOTE The best practice is to use more than one statistical program to evaluate the genetic database.

5.3.2 Source populations shall be genetically differentiated.

NOTE Whether populations are genetically differentiated is impacted by demographic, ecological, and evolutionary factors. Quantitative values are expected to vary according to the specific application, as well as the species/population(s) of interest.

5.3.2.1 Autosomal marker allele frequencies shall be calculated for each population for which individual membership is being estimated.

5.3.3 The method used for assigning an individual to a particular population shall include the following:

5.3.3.1 At minimum, an assignment test that includes at least one of the following methods:

5.3.3.1.1 Genetic distance-based analysis (e.g., evaluation of interpopulation distance, allele sharing distance, as in Cornuet et al. 1999).

5.3.3.1.2 Frequency-based analysis (e.g., evaluation of Hardy-Weinberg Equilibrium, allele frequency distribution, likelihood estimates, as in Paetkau et al. 1995)

5.3.3.1.3 Model-based analysis with Bayesian analysis (e.g. likelihood estimation as in Rannala and Mountain 1995, K clustering as in Pritchard et al. 2000, and Evanno et al. 2005).

5.3.3.2 The atypicality of the evidence sample shall be characterized (e.g., exclusion test) to account for the absence of the true population of origin in the multilocus population genetic database or to identify samples of mixed ancestry.

NOTE Correcting for the multiple comparisons should be assessed through simulated comparison studies, as it depends on the specific algorithms and inferences being made.

NOTE See Annex A for additional guidance.

5.3.4 The suitability and reliability of the statistical method(s) selected shall be characterized, including, but not limited to, the following:

5.3.4.1 The limits imposed by the geographic scope of the multilocus population genetic database.

5.3.4.2 The discriminating power of the test to resolve the genetic groupings of the multilocus population genetic database.

5.3.4.3 Variance of assignment power with heterogeneous sampling.

5.3.5 The uncertainty shall be described by running simulations for the method(s) selected with known and/or “mock” unknowns (i.e., simulated genotypes that are analogous to those encountered in casework).

5.3.5.1 Characterize the accuracy and precision of the statistical methods used for assignment.

5.3.5.2 Estimate the relative proportions of an individual’s membership in predefined groups, such as population units or species, and present the standard error of the relative proportion estimates.

5.3.6 When doing a hybrid assignment, natural and anthropogenic hybridization scenarios, including, but not limited to, intraspecific hybrids, interspecific hybrids, intergeneric hybrids, and interfamilial hybrids, shall be assessed.

NOTE In the case where interspecific hybridization events occur with species that only produce sterile F1 offspring, statistical analysis would be superfluous, provided there are fixed markers for each source population. In this case, standards relating to statistical methods would not apply.

Annex A

(Informative)

Statistical Method Supporting Information

The following statistical methods have been used in relation to population assignment. Each type of method is detailed below with references to various statistical packages that integrate that method.

Population Assignment Modeling

Programs may use a frequency-based or a model-based analysis with Bayesian methods. Users should understand the assumptions of each model used within each program and how violations of those assumptions may affect results. Applications of software may include genetic distance, frequency-based analysis, and model-based analysis with Bayesian methods.

Overview of Available Statistical Programs for Population Assignment Modeling

Note: Some programs can be sensitive to uneven sample sizes. The limitations of each program need to be considered during the validation process. Web addresses for these programs are subject to change. The information included in this Annex is current as of February 2025.

1. GenAlEx—A multipurpose Excel add-in that includes a function to determine the most likely population of origin using likelihood estimates.
    a. Using GenAlEx:
        i. https://biology-assets.anu.edu.au/GenAlEx/Welcome.html
    b. Additional reading:
        i. Peakall, Rod, and Peter E. Smouse. "GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update." *Bioinformatics*, vol. 28, no. 19, 2012, pp. 2537–2539, https://doi.org/10.1093/bioinformatics/bts460.
        ii. Smouse, Peter E., et al. "Converting quadratic entropy to diversity: Both animals and alleles are diverse, but some are more diverse than others." *PLOS One,* vol. 12, 2017, e0185499.
2. Rubias—An R package that implements Bayesian inference for genetic stock identification with modules to model mixtures and correct for bias introduced by uneven populations in a reporting group.
    a. Using Rubias:
        i. https://cran.r-project.org/web/packages/rubias/
        ii. Moran, Benjamin M., and Eric C. Anderson. "Bayesian inference from the conditional genetic stock identification model." *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 76, no. 4, 2018, 551-560.

   b. Additional reading:
      i. Anderson, Eric C., et al. “An improved method for predicting the accuracy of genetic stock identification.” *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 65, no. 7, 2008, pp. 1475–1486.
      ii. Kuismin, Markku, et al. “Genetic assignment of individuals to source populations using network estimation tools.” *Methods in Ecology and Evolution*, vol. 11, no. 2, 2020, pp. 333–344.

3. GeneClass2—A program that computes the probability of the multilocus genotype of each individual to be encountered in a given population using Monte Carlo sampling methods.
   a. Using GeneClass2:
      i. https://www1.montpellier.inrae.fr/CBGP/software/GeneClass/GeneClass2/Help/index.htm
      ii. Piry S., et al. “GENECLASS2: a software for genetic assignment and first-generation migrant detection.” *Journal of Heredity*, vol. 95, no. 6, 2004, pp. 536–539.

4. Structure—A Java run software that utilizes the systematic Bayesian clustering approach, applying Markov Chain Monte Carlo (MCMC) estimation to assess patterns of genetic structure in a set of samples.
   a. Using Structure:
      i. https://web.stanford.edu/group/pritchardlab/structure.html
      ii. Pritchard J.K., et al. “Inference of population structure using multilocus genotype data.” *Genetics*, vol. 155, 2000, pp. 945–959.
   b. Additional reading:
      i. Evanno, Guillaume, et al. "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study." *Molecular Ecology*, vol. 14, no. 8, 2005 pp. 2611–2620.
      ii. Wang, Jinliang. "The computer program structure for assigning individuals to populations: easy to use but easier to misuse." *Molecular Ecology Resources*, vol. 17, no. 5, 2017, pp. 981–990.
      iii. Porras-Hurtado, Liliana, et al. "An overview of STRUCTURE: applications, parameter settings, and supporting software." *Frontiers in Genetics,* vol. 4, 2013, 98.

5. WHICHRUN—Uses multilocus genotypic data to allocate individuals to their most likely source population. A C++ program that provides a variety of methods for evaluating population assignments, including maximum likelihood, jackknife, and critical population routines.
   a. Using WHICHRUN:
      i. https://marinescience.ucdavis.edu/research-programs/conservation/salmon-research/software

   b. Banks, M.A., W. Eichert. “WHICHRUN (version 3.2): a computer program for population assignment of individuals based on multilocus genotype data.” *Journal of Heredity*, vol. 91, no. 1, 2000, pp. 87–89. doi: 10.1093/jhered/91.1.87. PMID: 10739137.

Exclusion Testing (atypicality)

In the absence of the true population of origin in the baseline, a multilocus genotype may erroneously be assigned to a baseline population. The exclusion test identifies outliers in the database or calculates the probability that the genotype of an individual is not from any of the baseline populations.

Overview of Available Statistical Programs for Exclusion Testing

Note: The limitations of each program need to be considered during the validation process. Web addresses for these programs are subject to change. The information included in this Annex is current as of February 2025.

6. GeneClass2—*(see above for general program information)*
   a. GENECLASS2 calculates the probability that a new genotype of an individual in the baseline population of interest has a smaller likelihood of being observed than the actual individual of interest. It calculates this probability for each baseline population. It uses several Monte Carlo sampling algorithms that compute for each individual, its probability of belonging to each reference population, or being a resident (i.e., not first-generation migrant) in the population where it was sampled.
   b. Cornuet, J.M., et al. “New methods employing multilocus genotypes to select or exclude populations as origins of individuals.” *Genetics*, vol. 153, no. 4, 1999, pp. 1989–2000. doi: 10.1093/genetics/153.4.1989.
7. Rubias—*(see above for general program information)*
   a. Rubias compares simulated mixtures of varying sizes to the reference data set, with the likelihood being computed as well. After several simulations, the results can be used to predict the accuracy of the proportions that are estimated.
   b. The Overview of Rubias Usage section “Assessing whether individuals are not from any of the reference populations” provides information about the exclusion test module.
8. Additional reading on exclusion testing—
   a. *General reference*
      Ausdemore, M., et al. “Two-stage approach for the inference of the source of high-dimensional and complex chemical data in forensic science.” *Journal of Chemometrics*, 2021, 35:e3247. https://doi.org/10.1002/
   b. *Best practice*
      McLachlan, Geoffrey J. *Discriminant Analysis and Statistical Pattern Recognition*, Section 6.4. Wiley Series in Probability and Statistics, 1992. ISBN:9780471615316 |Online ISBN:9780471725299 |DOI:10.1002/0471725293

Annex B

(informative)

Bibliography

This is not meant to be an all-inclusive list, as the group recognizes that other publications on this subject may exist. At the time these standards were drafted, these were the publications available to the working group members for reference. Additionally, any mention of a particular software tool or vendor as part of this bibliography is purely incidental, and any inclusion does not imply endorsement by the authors of this document.

1] Aitkin, Colin G., et al., editors. *Statistics and the Evaluation of Evidence for Forensic Scientists*, 3rd ed., Wiley Series in Statistics and Practice, 2021.

2] Anderson, Eric C., et al. “An improved method for predicting the accuracy of genetic stock identification.” *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 65, no. 7, 2008, pp. 1475–1486.

3] Ausdemore, Madeline A., et al. “Two-stage approach for the inference of the source of high-dimensional and complex chemical data in forensic science.” *Journal of Chemometrics*, 2021, 35:e3247.

4] Banks, M.A., W. Eichert. “WHICHRUN (version 3.2): a computer program for population assignment of individuals based on multilocus genotype data.” Journal of Heredity, vol. 91, no. 1, 2000, pp. 87–89.

5] Beugin, Marie-Pauline, et al. “A fast likelihood solution to the genetic clustering problem.” *Methods in Ecology and Evolution*, vol. 9, no. 4, 2018, pp. 1006–1016.

6] Cornuet, J.M., et al. “New methods employing multilocus genotypes to select or exclude populations as origins of individuals.” *Genetics*, vol. 153, no. 4, 1999, pp. 1989–2000.

7] Evanno, Guillaume, et al. "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study." *Molecular Ecology*, vol. 14, no. 8, 2005, pp. 2611–2620.

8] Jombart Thibaut, et al. “Discriminant analysis of principal components: a new method for the analysis of genetically structured populations.” *BMC Genetics,* vol. 11, 2010, pp. 94–109.

9] Kuismin, Markku, et al. “Genetic assignment of individuals to source populations using network estimation tools.” *Methods in Ecology and Evolution*, vol. 11, no. 2, 2019, pp. 333–344.

10] McLachlan, Geoffrey J. *Discriminant Analysis and Statistical Pattern Recognition*, Section 6.4. Wiley Series in Probability and Statistics, 1992.

11] Moran, Benjamin M., and Eric C. Anderson. “Bayesian inference from the conditional genetic stock identification model.” *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 76, no. 4, 2019, pp. 551–560.

12] NRC II. National Research Council Committee on DNA Forensic Science. *The Evaluation of Forensic DNA Evidence.* National Academy Press, 1996.

13] Paetkau D., et al. “Microsatellite analysis of population structure in Canadian polar bears.” *Molecular Ecology*, vol. 4, no. 3, 1995, pp. 347–54.

14] Peakall, Rod, and Peter E. Smouse. “GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research.” *Molecular Ecology Notes*, vol. 6, no. 1, 2005, pp. 288–295.

15] Peakall, Rod, and Peter E. Smouse. “GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update.” *Bioinformatics*, vol. 28, no. 19, 2012, pp. 2537–2539.

16] Piry S., et al. “GENECLASS2: A Software for Genetic Assignment and First-Generation Migrant Detection.” *Journal of Heredity*, vol. 95, no. 6, 2004, pp. 536–539.

17] Porras-Hurtado, Liliana, et al. "An overview of *STRUCTURE*: applications, parameter settings, and supporting software." *Frontiers in Genetics,* vol. 4, 2013, 98.

18] Pritchard Jonathan K., et al. “Inference of Population Structure Using Multilocus Genotype Data.” *Genetics*, vol. 155, no. 2, 2000, pp. 945–959.

19] Rannala, Bruce, and Joanna L. Mountain. 1997. “Detecting immigration by using multilocus genotypes.” *The Proceedings of the National Academy of Sciences*, vol. 94, no. 17, 1997, pp. 9197–9201.

20] Smouse, Peter E., et al. “Converting quadratic entropy to diversity: Both animals and alleles are diverse, but some are more diverse than others.” *PLOS One,* vol. 12, 2017, e0185499.

21] Tonkin-Hill, Gerry, et al. “Fast hierarchical Bayesian analysis of population structure.” *Nucleic Acids Research*, vol. 47, 2019, pp. 5539–5549.

22] Valière, Nathaniel. “GIMLET: a computer program for analyzing genetic individual identification data.” *Molecular Ecology Notes*, vol. 2, 2002, pp. 377–379.

23] Wang, Jinliang. "The computer program structure for assigning individuals to populations: easy to use but easier to misuse." *Molecular Ecology Resources,* vol. 17, no. 5, 2016, pp. 981–990.