SCIENTIFIC WORKING GROUP ON DNA ANALYSIS METHODS[1]

SWGDAM Interpretation Guidelines for Single Nucleotide Polymorphism (SNP) Analysis by Forensic DNA Testing Laboratories

Short Title: *SNP Interpretation Guidelines*

Approved and Effective: January 11, 2024

Scope

The SWGDAM Interpretation Guidelines for Single Nucleotide Polymorphism (SNP) Analysis by Forensic DNA Testing Laboratories provides guidance for the interpretation of single nucleotide polymorphism (SNP) typing results developed by Next Generation Sequencing (NGS) methods for human identification, ancestry/phenotype predictions and targeted kinship testing. This document provides guidance for the interpretation of commercially available NGS-based forensic SNP genotyping assays. Considerations specific to whole genome sequencing and/or microarrays for Investigative Genetic Genealogy (IGG) are not covered in this document. Additionally, this document does not provide guidance for interpretation of SNPs within mitochondrial DNA (mtDNA); see SWGDAM MtDNA Interpretation Guidelines (2019).

[1] The Scientific Working Group on DNA Analysis (SWGDAM; see SWGDAM.org) is comprised of forensic science practitioners and other experts who represent government laboratories within the U.S and Canada, as well as intra- and international professional groups and academia. SWGDAM recommends to the FBI Director revisions to the *Quality Assurance Standards for Forensic DNA Testing Laboratories* and the *Quality Assurance Standards for DNA Databasing Laboratories (QAS)*. SWGDAM provides a forum for its members and invited guests to discuss research, technologies, techniques, and training; and conduct or recommend studies to develop, test, and validate methods for use by forensic laboratories. SWGDAM's Guidelines and Recommendations represent best practices within the discipline. The term "should" is used herein to indicate good practices identified by SWGDAM. "Shall" distinguishes mandatory elements, which may be specified in the Quality Assurance Standards for Forensic DNA Testing Laboratories and/or Quality Assurance Standards for DNA Databasing Laboratories.

SWGDAM Interpretation Guidelines for Single Nucleotide Polymorphism (SNP) Analysis by Forensic DNA Testing Laboratories
Approval/Effective Date: January 11, 2024

Table of Contents

Key Concepts:

- Specific considerations are described for the interpretation of SNP data including analytical and stochastic thresholds, controls, locus and allele designations and mixtures.
- Guidance for statistical interpretation distinguishes between the various SNPs – Identity Informative, Ancestry Informative, Phenotype Informative and Ys – addressing requirements for exclusion and inclusion, population databases, statistical formulae, as appropriate.
- Additional statistical considerations for Identity SNP panels (including genetic linkage and linkage disequilibrium) are explained in Appendix A and ancestry and phenotype panels are addressed in Appendix B.

1. Introduction

A SNP is variation in a single nucleotide that occurs at a specific position in the genome. Autosomal SNPs are typically biallelic, with two alleles present in the population (e.g., C or T). Some SNPs influence phenotypes - the physical/observed traits determined or "expressed" by a given genotype - particularly in coding regions of the genome when they result in the incorporation of a different amino acid during translation. Other SNPs in the human genome are in non-coding regions and appear to have minimal or no effect on phenotype. SNP typing provides probative genetic information regardless of DNA fragment length but is often of particular benefit in the analysis of samples with degraded DNA that cannot be profiled using forensic Short Tandem Repeats (STRs), which require intact DNA between 100 and 400 base pairs.

SNPs used in forensic analyses are typically described using four main categories: Identity Informative (IISNP), Ancestry Informative (AISNP), Phenotype Informative (PISNP), and Lineage Informative (LISNP, e.g., SNPs on the Y chromosome, or Y-SNPs). These SNP categories are described in detail below.

IISNPs, similar to autosomal STRs, are used for human identification. They have high heterozygosity and a low fixation index across worldwide populations, which means they are randomly distributed and not unique to any one population group. SNPs have much lower mutation rates compared with STRs (Kondrashov 2003) and most SNPs are biallelic; therefore, each single SNP has low discrimination power. IISNP testing requires more loci to achieve levels of discrimination power comparable to STRs. IISNPs are also used for the interpretation of genetic relationships between two or more individuals, which means that many of the recommendations that apply to human identification applications also apply to kinship applications (in this context, referred to as KISNPs). However, significantly more SNPs are generally needed for the interpretation of genetic relationships, and there are some differences in the methods applied for kinship applications.

AISNPs are used to infer biogeographic ancestry and are often monomorphic for one allele (e.g., "C" allele) in one population and monomorphic for the other allele (e.g., "T" allele) in all other populations, excluding admixed populations where loci may be polymorphic. Ancestry inference from AISNPs requires reference population datasets with existing allele frequency data. The AISNP profile from the questioned sample is compared to the population reference datasets in order to make an ancestry prediction for investigative purposes.

PISNPs are used to predict externally visible characteristics in forensic investigations and typically refer to markers that are predictive of pigmentation levels in eyes, hair, and skin. PISNPs for pigmentation share some characteristics with AISNPs, and are often included in ancestry interpretation models. An example of this is rs12913832, in which the "G" allele is predictive of both blue eyes and European ancestry (Sturm et al. 2008).

LISNPs can be utilized to assess relatedness within a family pedigree and can also be informative for ancestry predictions. These include SNPs on the mitochondrial genome, X and Y chromosomes, and tightly linked autosomal loci (microhaplotypes), but this document will only address Y-SNPs.

The focus of this document is to provide guidance for the interpretation of commercially available NGS-based forensic SNP genotyping assays. The parts of a forensic SNP genotyping process include: (1) SNP panel; (2) Library preparation method; (3) Genotyping platform; and (4) Interpretation model. It is important for users to understand the research behind the panels, both for selecting a platform/assay and for validation/training purposes. The library preparation method and genotyping platform are independent of the SNP application (e.g., Identity, Ancestry, or Phenotype).

2. Interpretation of SNP Typing Results

2.1 Analytical Threshold. An analytical threshold defines the minimum level at and above which signal can be reliably distinguished from background noise. The analytical threshold (AT) shall be validated based on internally derived empirical data. Usage of an exceedingly high analytical threshold to minimize this sporadic noise signal increases the risk of allelic data loss.

2.1.1 Analytical thresholds should be based on fixed read count values, read count percentages (e.g., allele read count divided by total locus read count) or other validated methods. Analytical thresholds may vary by locus, as well.

2.1.1.1 Given the large number of SNP loci that may be multiplexed together in a single reaction, locus-specific analytical thresholds may be necessary. Examples of implementation include ATs for individual loci or AT values based on locus performance groups (i.e., loci that behave similarly).

2.1.1.2 If measures are used to enhance detection sensitivity by increasing signal magnitude, the laboratory shall perform studies incorporating those measures to establish criteria for analytical threshold(s). Such enhanced detection measures may include, but are not limited to, increased amplification cycle number, changes to sample normalization steps, and reduction in number of libraries pooled.

2.1.1.3 If, during the quality control verification of new lots of critical reagents, the laboratory observes signal magnitude variation that differs from expectations (i.e., by comparison to historical data), the analytical threshold(s) may need to be verified using additional samples, and approved by the DNA Technical Leader prior to use of each critical reagent lot.

2.1.2 Reproducible artifacts that may exceed the analytical threshold shall be characterized during validation, based on qualitative (sequence) and/or quantitative (read count) characteristics.

2.1.2.1 The laboratory shall establish expectations for the noise (e.g., amplification background, sequencing errors, or sequencing background) relative to signal, and data interpretation shall be based on the limits established by validation. With higher read counts, increased total reads are detected for the true allele as well as noise. As read count signal may vary across loci, and as noise may be proportional to signal, locus-specific analytical thresholds may be required.

2.2 Sequencing Run Evaluation. The laboratory shall develop criteria to evaluate the quality of the run and the run data. Run quality may be measured by results from positive controls and sequencing standards, for example, or by other defined run metrics/parameters. If run quality metrics are used for this purpose, they shall be defined during validation. Metrics used for this purpose may include such parameters as phasing, loading density, cluster density, total reads per sample, total reads per run, forward/reverse read balance, Q-scores, clusters passing filter, percent usable reads, etc. (see glossary for definitions).

2.3 Assessment of Controls. For data to be of requisite quality for interpretation, measures shall be established to demonstrate that the testing performed as expected. The use of controls is among the most important quality measures for DNA testing. Controls shall include, at a minimum, a positive amplification control, a positive sequencing control (which may be the same as the positive amplification control), a negative amplification control, a negative sequencing control (which may be the same as the negative amplification control), and a reagent blank control. Evaluation criteria shall be established for each control. Controls shall be assessed as outlined in the *FBI Quality Assurance Standards for Forensic DNA Testing Laboratories.*

2.3.1 Reagent blanks, negative amplification controls, and negative sequencing controls shall be used to monitor levels of contamination and also to assist in identifying at which step of the process contamination may have been introduced. Contamination is the unintentional introduction of exogenous DNA into a DNA sample or amplification reaction. Reagent blanks monitor contamination from

extraction to final analysis. Negative controls monitor contamination from amplification (or sequencing) to final analysis. Negative controls shall be processed concurrently with the corresponding samples using the same reagents and/or instruments. Additional negative controls may be included during sample preparation to assess contamination as needed. The laboratory shall have criteria for evaluating reagent blanks and negative controls that generate typing results to determine if associated sample data are reportable.

2.3.2 Positive controls shall be used to monitor the success of the laboratory process. A single positive control, of known genotype, may be processed starting at amplification and used throughout the process to monitor downstream steps. Alternatively, different positive controls may be used to monitor the success of different steps of the process. In either case, the positive control shall be processed concurrently with the samples being typed using the same reagents and/or instruments. If the positive amplification control is not used as a positive sequencing control, the laboratory shall have and follow procedures to evaluate the positive amplification control. The expected performance and/or genotype of the positive control(s) used for these purposes shall be well-characterized and detailed in validation or other documentation, to include tolerances for genotype agreement and allele drop-out within the assay.

2.3.3 To minimize the introduction of contamination during testing, a laboratory shall implement sample handling procedures and quality control practices designed for this purpose. Methods shall be in place to monitor contamination within the laboratory. A laboratory shall verify that all control results meet the laboratory's interpretation guidelines for all reported results. In addition, a laboratory shall have and follow policies and/or procedures that are supported by validation studies for interpreting data potentially affected by contamination.

2.4 Locus Designation. The laboratory shall have criteria to address locus and allele designations. A positive control may be used to verify correct locus designations.

2.4.1 Locus designations shall include an 'rs number' traceable to the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/) (Sherry et al. 2001), genome position(s), and associated human genome reference identification (e.g., hg 19, GRCh38). If no rs number is available for a given SNP, the genome position and associated human genome reference identification should be used.

2.4.1.1 The same SNP may have different rs numbers (for example rs200207348 is the same SNP as rs938283) or alternate names (rs312262906 has been described as "N29insA"), and it is important to ensure comparisons are being made with the same SNP. If the laboratory will perform cross-kit comparisons, there should be a procedure to address this.

2.4.1.2 Some SNP models have been developed that report SNP alleles based on the coding strand, which, in some cases, may not be the reported strand in databases. Commercially available software programs incorporate these strand reporting differences, but the laboratory should be aware of the locus/allele reporting for interpretation outside of commercial software applications.

2.5 Allele Designation. Given that no forensic database system presently prescribes reporting standards, the laboratory shall report SNP alleles in a manner that allows for intra- and inter-laboratory comparisons, in accordance with the following guidance.

2.5.1 DNA bases shall be designated by nucleotide identity (A, C, G, and T).

2.5.2 Insertions and deletions (INDELS) shall be reported in reference to a known database, such as dbSNP. If the SNP INDEL allele has not been previously documented within the known database, deletions shall be reported as "." or "-".

Nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
N	any base
. or -	gap

2.5.3 Homozygous SNP loci shall be reported in a manner that unambiguously indicates detection of a single allele (e.g., C/C, TT).

2.5.4 Heterozygous SNP loci shall report both alleles, (e.g., C/T or CT).

2.5.4.1 The laboratory shall use interpretation criteria for heterozygous alleles that are derived from internal validation and provide threshold values for read balance and/or coverage to correctly determine the presence of heterozygous SNPs, as described in section 2.6.

2.5.5 If drop out of a second allele is possible under the laboratory's interpretation criteria, results should be reported in a manner which distinguishes this scenario from a homozygous allele call (e.g., a single C or a C/- when homozygotes are reported as CC or C/C, respectively).

2.5.6 The presence of genotypes with more than two alleles for a SNP locus may occur, primarily from copy number variations and/or duplication of genome regions.

2.5.7 Unexpected/rare alleles may be observed. The laboratory may need a method to handle rare alleles during SNP analysis and should consider how their presence affects the interpretation of results.

2.5.8 Substitutions due to cytosine deamination are commonly observed in aged and degraded forensic DNA samples, such as skeletal remains, formalin-preserved remains, and rootless hairs, as shown in mtDNA sequencing studies (Gorden et al. 2018; Cuenca et al. 2020; Zavala et al. 2022). Cytosine deamination can lead to C→T and G→A substitutions in SNP data that may impact intralocus balance and/or cause erroneous genotypes (e.g., C/T instead of C/C; Loreille et al. 2022). Although cytosine deamination occurs at low frequency (<20%), the stochastic effects of PCR enrichment may inflate its detection. When cytosine deamination is believed to impact a genotype call, the locus can be excluded as long as a mixture is not suspected (based on validation data from a mixture study). Alternatively, replicate testing can be performed to confirm the allele(s), and a consensus genotype can be called (Marshall et al. 2020). Furthermore, the laboratory may choose to implement a DNA repair (uracil excision) step prior to target enrichment in order to minimize the impact of cytosine deamination on resulting sequence data (Gorden et al. 2018). The laboratory should be aware of this phenomenon and consider it in the interpretation of SNP genotypes from aged and degraded samples.

2.6 Stochastic Thresholds.

2.6.1 If using a binary interpretation method, the laboratory shall have a stochastic threshold (ST) which defines the signal magnitude value below which it is reasonable to assume that, at a given locus, dropout of a sister allele in a heterozygous pair may have occurred.

2.6.1.1 Stochastic thresholds shall be based on empirical data derived within the laboratory and specific to the genotyping system (e.g., kit) and detection platform used.

2.6.1.2 Stochastic thresholds may be based on fixed read count values and/or read count percentages (i.e., allele read count divided by total locus read count). Stochastic thresholds may vary by locus, as well.

2.6.1.2.1 Given the large number of SNP loci that may be multiplexed together in a single reaction, locus-specific stochastic thresholds may be necessary. Examples of implementation include STs for individual loci or ST values based on locus performance groups (i.e., loci that behave similarly).

2.6.1.3 If measures are used to enhance detection sensitivity (e.g., increasing signal magnitude), the laboratory shall perform studies incorporating those measures to establish criteria for stochastic threshold(s). Such enhanced detection measures may include, but are not limited to, increased amplification cycle number, changes to sample normalization steps, and reduction in number of libraries pooled.

2.6.1.4 If, during the quality control verification of new lots of critical reagents, the laboratory observes signal magnitude variation that differs from expectations (i.e., by comparison to historical data), the stochastic threshold(s) may need to be verified using additional samples, and approved by the DNA Technical Leader prior to use of each critical reagent lot.

2.6.1.5 For SNP assays that include loci exhibiting duplication (e.g., Y-SNPs) or multi-copy inheritance (e.g., X-SNPs in females), the laboratory shall establish specific stochastic thresholds for these markers.

2.7 Mixtures. Mixtures may be detected in SNP genotyping data. Most SNPs are biallelic; therefore, unlike STRs, SNP mixtures will not typically exhibit loci with more than two alleles. Indications of a mixture in SNP data include allele count ratios outside the expected ratio for heterozygous loci and/or an increase in the number of expected heterozygous loci versus homozygous loci. It is noted that allele count ratio imbalances may be seen in results from, for example, a primer binding site variant that results in attenuated amplification or signal of one allele of a heterozygous pair. Likewise, degraded, inhibited, and/or low-level single-source DNA samples may exhibit poor read count balance with heterozygous loci.

AISNPs are largely monomorphic within populations, and PISNPs may be as well; therefore, a mixture of individuals from the same population may present as a lower number of contributors, or single-source, when evaluating AISNP and PISNP data compared to IISNP data. For Y-SNPs, a sample is generally assumed to have originated from more than one male individual if two alleles are present at two or more single-copy loci. The assumed presence of a known contributor and/or collective data from more than one category of SNP (e.g., identity, lineage) may further support the presence of a mixture.

2.7.1 The laboratory shall establish criteria for assessing whether a result is consistent with a single-source sample or a mixture, including the minimum number of affected loci should be defined for determination of whether a sample is a mixture. Additional, existing data for the sample (e.g. STR genotyping and/or mitochondrial DNA sequencing results) should be considered.

2.7.1.1 The Allele Count Ratio (ACR) expectations based on empirical data may be useful for assessing whether a result is consistent with a single-source sample or a mixture.

2.7.1.2 The expected number of heterozygous versus homozygous loci in single-source samples may indicate the presence of a mixture.

2.7.1.3 If the evidence is assumed to consist of two or more close biological relatives, the true number of contributors to the evidence may be underestimated due to the high degree of SNP allele sharing.

2.7.2 If laboratories are interpreting SNP mixtures, they shall have a procedure, supported by validation studies, for mixture interpretation that addresses assessing the number of contributors, the separation of contributors, and the criteria for deducing potential contributors. Currently, guidance does not exist regarding interpretation of SNP mixtures; therefore, it is recommended that laboratories exercise caution if interpreting SNP data indicating the presence of more than one contributor.

3. Statistical Interpretation

3.1 Identity Informative SNPs. IISNPs are used for individualization and human identification in the same manner as autosomal STRs.

3.1.1 Exclusion/Inclusion criteria shall be established. Similar to STR interpretation, data that are not used in statistics may be used for exclusion purposes. It is important to note that SNP mutation rates are much lower compared with STR mutation rates, and it is therefore relevant to also account for other possible reasons for observed genetic inconsistencies (e.g., null alleles). Exclusion criteria should address the suitability of using a SNP for exclusion purposes (acceptable number of reads; distinguished from noise and typing errors; etc.).

3.1.2 Population databases. The laboratory shall document the source of the population database(s) used in any statistical analysis.

3.1.2.1 Due to the biallelic nature of most SNPs and the limited number of possible alleles and genotypes at a locus, the sample size needed for allele frequency estimates is smaller than for STRs; therefore, smaller population databases may be utilized.

3.1.2.2 The laboratory shall have a procedure to account for alleles not included in the population databases.

3.1.3 Statistical formulae. Statistical calculations for IISNPs generally utilize the same formulae and follow the same interpretation guidelines as those used with autosomal STRs, such as the use of likelihood ratios and random match probability. Probabilistic genotyping approaches may be considered for statistical interpretation. The choice of approach is influenced by the laboratory's current protocols, validation, and available resources.

3.1.3.1 The laboratory shall have a procedure to address population substructure (Fst).

3.1.3.2 The laboratory shall review the literature and determine an appropriate value for theta (θ). Currently, specific guidance does not exist regarding theta values

for SNP data; resources are included in: https://strbase.nist.gov/File_Share/LD-theta_SWGDAM_230417.xlsx

3.1.3.3 Supplemental to these guidelines, a summary of published population theta estimates from forensic IISNP panels is provided in the same file link above (3.1.3.2).

3.1.3.4 The following formulae are used in calculating the frequency of a single-source DNA profile.

3.1.3.4.1 Heterozygotes-NRC II 4.1b Probability (PQ) = 2pq

3.1.3.4.2 Homozygotes-NRC II 4.4a Probability (PP) = $p^2 + p(1-p)\,\theta$

3.1.3.5 Dependencies between genetic loci may exist when increasing the number of genetic markers included in an analysis, as for SNP panels. For more information see Appendix A.

3.1.3.6 The developmental validation shall establish if the SNP loci are linked and/or in linkage disequilibrium (LD) in reference population datasets. If the laboratory chooses to utilize a novel reference population, the laboratory shall establish whether the SNP loci are in LD.

3.1.3.6.1 If linkage and/or LD do not exist the product rule may be used.

3.1.3.6.2 If linkage and/or LD exist for a given set of SNP loci, the laboratory shall have a procedure to account for linkage and/or LD in order to minimize statistical bias.

3.1.3.7 Laboratories evaluating SNPs in microhaplotypes or haplotype blocks should apply a statistical framework based on phased haplotype frequency estimates rather than individual component SNP allele frequencies.

3.1.4 Kinship analyses

3.1.4.1 When performing kinship calculations with IISNPs, laboratories shall have guidelines for working with linked SNPs and SNPs in LD. This may include incorporation of the recombination frequencies into the likelihood ratio calculation, including only one of the two linked SNPs in the kinship analysis,

using the set of markers in LD as a haplotype block, and/or by using various software-based approaches listed in Appendix A.

3.1.4.2 Population substructure (Fst), null (silent) allele frequencies, and SNP-specific mutation rates may be relevant to consider in kinship calculations.

3.2 Ancestry Informative SNPs. AISNPs are used to estimate the ancestry of an individual for use as investigative leads.

3.2.1 The laboratory shall document the source of the population database(s) used in ancestry prediction. Example databases are given in Section 7.0 SNP Resources.

3.2.2 Statistical formulae

3.2.2.1 Different statistical models/methods have been developed, and can be used for the prediction of biogeographic ancestry.

3.2.2.2 Principal Component Analysis (PCA) is commonly used to cluster individuals with similar genotypes. Bayesian classifiers (e.g., Naive Bayes, STRUCTURE) are also used (Appendix B) to partition individual genotypes into ancestry proportions.

3.4 Phenotype Informative SNPs. PISNPs are used to estimate certain phenotypic characteristics that may include eye, hair and skin color. Such PISNPs can be used for investigative leads with an estimated probability to provide a weight to the estimate.

3.4.1 The laboratory shall document the source of the database(s) used in phenotype prediction.

3.4.2 Statistical formulae

3.4.2.1 Multinomial logistic regression (MLR) is commonly used to generate phenotype estimates, but likelihood ratios and Bayesian classifiers are also used (See Appendix B).

3.5 Y-SNPs

3.5.1 Currently, population databases and usage guidance are limited for Y-SNP statistical estimates; therefore, it is recommended that these markers are used only for exclusionary purposes or haplogroup predictions.

4. Reporting

4.1 The laboratory shall have and follow procedures for reporting SNP data, comparisons, and statistics.

4.2 If a laboratory is comparing SNP data across kits, rs number, genomic position, and strand should be considered.

4.3 The laboratory shall have and follow reporting procedures for ancestry and phenotypic predictions, which reflect the limitations of the prediction, based on the validation studies. The laboratory may consider the SWGDAM verbal scale equivalency for reporting LRs to qualify the confidence in the prediction.

4.4 The laboratory shall have and follow criteria for formulating inclusionary, exclusionary, and inconclusive results for human identification and kinship applications.

5. Glossary (for use with these Guidelines only)

Allele: a form of a gene that is located at a specific location on a specific chromosome. Alleles targeted in SNP analysis vary in composition (nucleotide) and size (indel), but are not repetitive like short tandem repeats.

Allele Count Ratio (ACR): the relative ratio, or intralocus balance, of two alleles at a given locus. This is commonly expressed as a percentage and is generally calculated for a given locus by dividing the count of the allele with the lower signal value by the count of the allele with the higher signal value. Allele Count Ratios may also be referred to as allele coverage ratios or read count ratios. In all cases, the ratios are analogous to CE-based Peak Height Ratios.

Analytical Threshold (AT): the minimum read count at and above which detected signal can be reliably distinguished from background noise.

Cluster Density: the density of clonal clusters on a sequencing flow cell. Optimal cluster density maximizes sequencing performance in terms of data quality and total sequence data output. The term applies to those NGS chemistries that employ glass flow cell sequencing technology.

Cytosine Deamination: the loss of an amine group in a cytosine base due to hydrolytic or oxidative DNA damage (Lindahl 1993). During the replication process, the deaminated cytosine is read by the polymerase as a uracil base that is paired with adenine. PCR then pairs the adenine molecule with thymine. Thus, cytosine deamination results in C→T substitutions (and G→A substitutions in complementary strands) (Jónsson et al. 2013).

Fixation Index (Fst): a measure of population differentiation due to genetic structure; commonly equated with the theta θ correction.

Forward/Reverse Read Balance (or Strand Balance): a measure of the distribution of forward and reverse reads aligned at each nucleotide position. A relatively even distribution of reads from both strands provides a measure of support for the nucleotide call. While strand imbalance or bias can, under certain circumstances, indicate reduced support for the affected nucleotide calls, in some assays, and in particular genomic regions, only one strand is routinely sequenced.

Genotype: results of SNP analysis of an individual at one or more genetic loci.

Haplotype: a set of DNA variations (polymorphisms such as SNPs and indels) adjacent to one another at the same locus that tend to be inherited together. This set of alleles is often referred

to as linked polymorphisms.

Haplotype Block (see Microhaplotype)

Library (or Sequencing Library): a work product consisting of genetic material prepared for analysis on a next generation sequencing instrument.

Linkage: the non-independent transmission of two genetic units, usually because of proximity on the same chromosome.

Linkage Disequilibrium (LD): the non-random association, in a population, of alleles at different loci.

Loading Density: the percentage of wells successfully loaded across the physical surface of a sequencing chip. Higher values reflect greater coverage or loading of the chip. The term applies to those NGS chemistries that employ sequencing chips with wells.

Locus: the specific physical location of a genetic marker on a chromosome, often denoted for SNPs as an rs number (e.g., rs12345678).

Microhaplotype (or Haplotype Block): two or more linked single nucleotide polymorphisms (SNPs) that occur within a short segment of DNA (able to be sequenced in a single read) and display multiple allelic combinations.

Mixture: DNA typing result originating from two or more individuals.

Next Generation Sequencing or NGS (or Massively Parallel Sequencing, Deep Sequencing, or High Throughput Sequencing): term used to describe modern sequencing technologies other than Sanger sequencing. This does not include array-based technologies.

Noise: background signal detected by a data collection instrument.

Null (Silent) Allele: an allele which cannot be detected due to lack of amplification product, often caused by a mutation in the primer binding site, or deletion of the primer binding site or locus.

Panel: a collection of markers that have been characterized and grouped for testing.

Phasing (sequencing chemistry): the rate at which single molecules within a sequencing cluster become out of sync with each other during the sequencing process. Individual strands may be a base (or more) ahead of the majority of the cluster (pre-phasing), or they may lag behind the majority of the cluster (phasing). Together, pre-phasing and phasing offer a measure

of the performance of the chemistry/sequencing run, with higher values indicating lower signal to noise ratios (i.e., more noise from phasing). The term applies to sequencing by synthesis chemistries.

Phasing (genetic): determining the association of alleles at different loci. For autosomal markers, phasing involves separating maternally and paternally inherited copies of each chromosome into haplotypes. Phasing may refer to the association of multiple variants at separate nucleotide positions on a single sequence read.

Prune: a process of selecting a subset of SNPs by removing SNPs that are in LD with each other.

Quality Score (or Q-Score): a metric that is used to indicate whether a base has been called correctly. Specifically, it is the probability that a given base has been miscalled. Mathematically, it is defined as: $-10\log_{10}(e)$, where e is the estimated probability of the base call being incorrect. Higher Q scores indicate a lower probability of base-calling error, while lower Q scores indicate a higher probability of error.

Reads: the raw sequence data produced by a sequencing instrument, generated as a result of detection of the sequence of nucleotides in a DNA strand and the translation of that information into digital sequence data.

Read Count: the absolute number of reads of a given sequence. Read count (X) is a measurement of signal and, as such, is analogous to relative fluorescent units (RFUs) in capillary electrophoresis based analysis.

Single-Source Profile: DNA typing results determined to originate from one individual based on Allele Count Ratio assessments and the number of alleles at given loci.

Stochastic Threshold: the read count below which it is reasonable to assume that, at a given locus, allelic dropout of a sister allele in a heterozygous pair may have occurred.

6. References

1000 Genomes Project Consortium. (2015) *A global reference for human genetic variation*. Nature 526 (7571): 68.

Abecasis, G. R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002) *Merlin--rapid analysis of dense genetic maps using sparse gene flow trees*. Nature Genetics 30 (1) (January 01): 97-101.

Amigo, J., Phillips, C., Lareu, M., and Carracedo, A. (2008) *The SNP for ID browser: An online tool for query and display of frequency data from the SNP for ID project*. International Journal of Legal Medicine 122 : 435-40.

Chaitanya, L., Breslin, K., Zuñiga, S., Wirken, L., Pośpiech, E. Kukla-Bartoszek, M., Sijen, T., de Knijff, P., Liu, F., and Branicki, W. (2018) *The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation*. FSI Genetics 35 : 123-35.

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., Donnelly, S., Gabriel, S., Gupta, N., Ferriera, S., Tolonen, C., Novod, S., Bergelson, L., Roazen, D., Ruano-Rubio, V., Covarrubias, M., Llanwarne, C., Petrillo, N., Wade, G., Jeandet, T., Munshi, R., Tibbetts, K., gnomAD Project Consortium, O'Donnell-Luria, A., Solomonson, M., Seed, C., Martin, A. R., Talkowski, M. E., Rehm, H. L., Daly, M. J., Tiao, G., Neale, B. M., MacArthur, D. G. and Karczewski, K. J. (2022) *A genome-wide mutational constraint map quantified from variation in 76,156 human genomes*. bioRxiv 2022.03.20.485034; available at https://doi.org/10.1101/2022.03.20.485034

Cuenca, D., Battaglia, J., Halsing, M., and Sheehan, S. (2020) *Mitochondrial sequencing of missing persons DNA casework by implementing thermo fisher's precision ID mtDNA whole genome assay*. Genes 11 (11): 1303.

Excoffier, L., and Lischer, H.E. (2010) *Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under linux and windows*. Molecular Ecology Resources 10 (3) (May 01): 564-7.

Gill, P., Phillips, C., McGovern, C., Bright, J.A., and Buckleton, J. (2012) *An evaluation of potential allelic association between the STRs vWA and D12S391: Implications in criminal casework and applications to short pedigrees*. FSI Genetics 6 (4): 477-86.

Gorden, E. M., Sturk-Andreaggi, K. and Marshall, C. 2(018) *Repair of DNA damage caused by cytosine deamination in mitochondrial DNA of forensic case samples*. FSI Genetics 34 : 257-64.

Jäger, A. C., Alvarez, M.L., Davis, C.P., Guzmán, E., Han, Y., Way, L., Walichiewicz, P., Silva, D., Pham, N., Caves, G., Bruand, J., Schlesinger, F., Pond, S.J.K., Varlaro, J., Stephens, K.M.,

and Holtet, C.L. (2017). *Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories*. FSI Genetics 28 : 52-70.

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L.F., and Orlando, L. (2013) *mapDamage2. 0: Fast approximate bayesian estimates of ancient DNA damage parameters*. Bioinformatics 29 (13): 1682-4.

Kidd, K. K., Speed, W.C., Pakstis, A.J., Furtado, M.R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F.R., Kidd, J.R. (2014) *Progress toward an efficient panel of SNPs for ancestry inference*. FSI Genetics 10 : 23-32.

Kling, D., Egeland, T., and Tillmar, A.O. (2012) *FamLink - A user friendly software for linkage calculations in family genetics*. FSI Genetics (March 03).

Kling, D., Dell'Amico, B., and Tillmar, A.O., (2015) *FamLinkX–implementation of a general model for likelihood computations for X-chromosomal marker data*. FSI Genetics 17 : 1-7.

Koenig, Z., Yohannes, M.T., Nkambule, L.L., Goodrich, Kim, H.A., Zhao, X., Wilson, M.W., Tiao, G., Hao, S.P., Sahakian, N., Chao, K.R., gnomAD Project Consortium, Talkowski, M.E., Daly, M.J., Brand, H., Karczewski, K.J., Atkinson, E.G., and Martin, A.R. (2023) *A harmonized public resource of deeply sequenced diverse human genomes*. bioRxiv 2023.01.23.525248; available at https://doi.org/10.1101/2023.01.23.525248.

Kondrashov, A.S. (2003) *Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases*. Human Mutation 21 (1): 12-27.

Kosoy, R., Nassir, R., Tian, C., White, P.A., Butler, L.M., Silva, G., Kittles, R., Alarcon-Riquelme, M.E., Gregersen, P.K., and Belmont, J.W. (2009) *Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America*. Human Mutation 30 (1): 69-78.

Lindahl, T. (1993) *Instability and decay of the primary structure of DNA*. Nature 362 (6422) (April 22): 709-15.

Loreille, O., Tillmar, A., Brandhagen, M.D., Otterstatter, L., and Irwin, J.A. (2022) *Improved DNA extraction and Illumina sequencing of DNA recovered from aged rootless hair shafts found in relics associated with the Romanov family*. Genes 13 (2): 202.

Marshall, C., Sturk-Andreaggi, K., Gorden, E.M., Daniels-Higginbotham, J., Sanchez, S.G., Bašić, Z., Kružić, Anđelinović, S., Bosnar, A., and Čoklo, M. (2020) *A forensic genomics approach for the identification of sister marija crucifiksa kozulić*. Genes 11 (8): 938.

Pakstis, A.J., Speed, W.C., Fang, R., Hyland, F.C.L., Furtado, M.R., Kidd, J.R., and Kidd, K.K. (2010) *SNPs for a universal individual identification panel*. Human Genetics 127 : 315-24.

Pakstis, A. J., Speed, W.C., Kidd, J.R., and Kidd, K.K. (2007) *Candidate SNPs for a universal individual identification panel*. Human Genetics 121 : 305-17.

Phillips, C. (2015) *Forensic genetic analysis of bio-geographical ancestry*. FSI Genetics 18 : 49-65.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000) *Inference of population structure using multilocus genotype data*. Genetics 155 (2) (June 01): 945-59.

Rajeevan, H., Soundararajan, U., Pakstis, A.J., and Kidd, K.K. (2012) *Introducing the forensic research/reference on genetics knowledge base, FROG-kb*. Investigative Genetics 3 (1) (September 01): 18.

Sanchez, J.J., Phillips, C., Borsting, C., Balogh, K., Bogus, M., Fondevila, M., Harrison, C.D., Musgrave-Brown, E., Salas, A., Syndercombe-Court, D., Schneider, P.M., Carracedo, A., and Morling, N. (2006) *A multiplex assay with 52 single nucleotide polymorphisms for human identification*. Electrophoresis 27 (9) (May 01): 1713-24.

Scientific Working Group on DNA Analysis Methods. (2019) *SWGDAM Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories*. Available at https://www.swgdam.org/publications.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001) *dbSNP: The NCBI database of genetic variation*. Nucleic Acids Research 29 (1): 308-11.

Sturm, R.A., Duffy, D.L., Zhao, Z., Leite, F.P.N., Stark, M.S., Hayward, N.K., Martin, N.G., and Montgomery, G.W. (2008) *A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color*. American Journal of Human Genetics 82 (2): 424-31.

Tillmar, A.O., and Phillips, C. (2017) *Evaluation of the impact of genetic linkage in forensic identity and relationship testing for expanded DNA marker sets*. FSI Genetics 26 : 58-65.

Tvedebrink, T., Eriksen, E.S., Mogensen, H.S., Morling, N. (2018) *Weight of the evidence of genetic investigations of ancestry informative markers*. Theoretical Population Biology 120 1–10.

Walsh, S., Chaitanya, L., Clarisse, L., Wirken, L., Draus-Barini, J., Kovatsi, L., Maeda, H., Ishikawa, T., Sijen, T., and de Knijff, P. (2014) *Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage*. FSI Genetics 9 : 150-61.

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W., and Kayser, M. (2013) *The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA*. FSI Genetics 7 (1): 98-115.

Weir, B.S. (1990) *Genetic data analysis. methods for discrete population genetic data*. Sinauer Associates, Inc. Publishers.

Zavala, E.I., Thomas, J.T., Sturk-Andreaggi, K., Daniels-Higginbotham, J., Meyers, K.K., Barrit-Ross, S., Aximu-Petri, A., Richter, J., Nickel, B., and Berg, G.E. (2022) *Ancient DNA methods improve forensic DNA profiling of Korean War and World War II unknowns*. Genes 13 (1): 129.

7. SNP Resources:

dbSNP

- National Center for Biotechnology Information
- https://www.ncbi.nlm.nih.gov/snp/
- Public domain archive for broad collection of simple genetic polymorphisms
- Accepts submissions for variations in any species and from any part of a genome

1000 Genomes

- https://www.internationalgenome.org/
- Catalog of common human variation
- Final phase included over 2,500 genomes

Catalogs of common human variation

- gnomAD - https://gnomad.broadinstitute.org/
  - Current version (v3.1) includes over 75,000 genomes
  - Chen et al. (2022)
- 1000 Genomes Project (1kGP) - https://www.internationalgenome.org/
  - Final phase included over 2,500 genomes
- Koenig et al. (2023) combines 1000 Genomes Project and gnomAD data

ALFRED

- The Allele Frequency Database- "a resource of gene frequency data on human populations supported by the Yale Center for Medical Informatics"
- https://alfred.med.yale.edu/alfred/index.asp
- Contains multiple published SNP datasets for IISNP, AISNP, PISNP (and corresponding online citations)

SNPforID

- Web-based tool for the query and visualization of the SNP allele frequency data generated by the SNPforID consortium
- http://spsmart.cesga.es/snpforid.php
- Contains data for 52-plex (individual identification markers) and 34-plex (ancestry informative markers)
- Combined population data from multiple sources
- Last updated 2012/2013
- Amigo et al. (2008)

GenoGeographer

- http://apps.math.aau.dk/aims/
- Classifies sample into a list of reference populations
- Built in datasets for 'Kidd loci', 'Seldin loci' and 'Precision ID'
- Tvedebrink et al. (2018)

FROG-kb

- Forensic Resource and Reference On Genetics- knowledge base
- https://frog.med.yale.edu/FrogKB/
- Formerly Ken Kidd lab- Yale
- Open access web application
- Underlying data is available in ALFRED
- Rajeevan et al. (2012)

Appendix A: Identity SNP Panels, Genetic Linkage & Linkage Disequilibrium

Commercially available forensic Identity Information Single Nucleotide Polymorphism (IISNP) assays contain combinations of SNPs found in publications from the SNPforID Consortium (Sanchez et al. 2006) and the Kidd laboratory (Pakstis et al. 2010). The manner in which statistical weight of probative inclusions using Identity SNP markers for single source profiles can be calculated similar to STR loci, using likelihood ratios or random match probabilities, after accounting for dependencies between loci as described below. The statistical weight in kinship testing is most often calculated using likelihood ratios, accounting for dependencies as described below.

Dependencies between genetic loci may exist when increasing the number of genetic markers included in an analysis, as is the case with SNP panels. Such dependencies could impact the biostatistical evaluations in identity and kinship testing. Two phenomena are important to consider:

> (1) Genetic linkage (also referred to as physical linkage or simply linkage) which causes closely located loci to be inherited as a unit to a higher degree than for unlinked loci within a family/pedigree. The degree of linkage is often measured by the recombination rate.
>
> (2) Linkage disequilibrium (LD, or allelic association), which exists at a population level when alleles at different loci appear together at rates that differ from those expected under independence (i.e., random association). It is important to note that even though genetic linkage and LD may coexist, they have different properties. Thus, different approaches and models are applied to account for their existence and impact on biostatistical calculations.

Genetic linkage exists when the recombination rate is less than 0.5. Linkage affects the transmission probabilities, and is normally only relevant to consider in kinship testing. It should however be considered in identity cases when the alternative hypothesis involves a close relative to the suspect. Linkage is accounted for by incorporating the recombination rate between loci in the likelihood ratio or random match probability calculation. Accounting for genetic linkage is, however, not always necessary, since the impact may depend on the relationship, individuals tested, as well as the zygosity of the loci included in the analysis (Gill et al. 2012).

LD exists when alleles at different loci are associated, causing haplotype frequencies to be significantly different from those expected under linkage equilibrium (LE). In both identity and

kinship testing, LD affects the genotype probabilities and can be addressed by using haplotype frequencies instead of allele frequencies in the calculation. An alternative approach could be to prune loci in LD in the analysis. If this latter approach is used, the degree of information will decrease, but excludes the need to estimate haplotype frequencies. If the pruning approach is used, it is also recommended that linked/associated markers are excluded from interpretation and that the report states which markers were used in the interpretation. It should be noted that all loci may still be incorporated in interpretations for the purpose of looking for exclusionary information.

Supplemental to these guidelines, a summary of published LD evaluations among markers in NGS-based forensic SNP assays is provided here:

https://strbase.nist.gov/File_Share/LD-theta_SWGDAM_230417.xlsx

Software and data references

- GDA (Weir 1990) (LD analysis)
- Arlequin (Excoffier and Lischer 2010) (LD analysis)
- ILIR (Tillmar and Phillips 2017) (Linkage analysis)
- FamLink (Kling et al. 2012) (Biostatistical calculations)
- FamLinkX (Kling et al. 2015) (Biostatistical calculations)
- Merlin (Abecasis et al. 2002) (Biostatistical calculations)

Appendix B: Ancestry and Phenotype Panels and Estimates

Currently, commercially available forensic assays for Ancestry Informative Single Nucleotide Polymorphisms (AISNPs) contain Ken Kidd's panel of 55 AISNPs (Kidd et al. 2014) and/or a panel of 128 SNPs from the Seldin Laboratory (Kosoy et al. 2009); whereas Phenotype Informative Single Nucleotide Polymorphism (PISNP) assays include the SNPs found in the HIrisPlex system (Walsh et al. 2013).

Different statistical models, tools and approaches exist for the prediction of phenotypes and ancestry. For the prediction of eye, hair and skin color, a multinomial logistic regression model has been validated (HIrisplex-S model, Chaitanya et al. 2018; Walsh et al. 2013; Walsh et al. 2014). The same model has also, for example, been implemented in commercially available software (Jäger et al. 2017).

In brief, based on a fixed set of output categories (e.g., "Blue", "Brown" and "Intermediates" for eye color) and a fixed set of DNA markers, the multinomial logistic regression model is trained with a large number of samples with known phenotypes and genotypes. Such model training results in estimates for a number of vital model parameters, which are then used for new predictions. The output from such a prediction tool is a prediction probability for each category, based on the observed genotypes and model parameter settings. It should be noted that these prediction probabilities do not necessarily correspond/represent the accuracy of the prediction. Accuracy and other performance statistics shall be studied using metrics such as sensitivity, specificity, positive predictive value (PPV), etc. (see Table 1).

Table 1. Example: Overall performance statistics for eye color predictions (https://hirisplex.erasmusmc.nl/).

		Category		
Metric	**Interpretation (example)**	**Blue**	**Brown**	**Intermediate**
Sensitivity	Proportion of samples with observed phenotype X, predicted as X (e.g., "how often is a true 'blue' predicted as 'blue'?")	0.928	0.935	0.001
PPV	Proportion of samples predicted as X, which are true X (e.g., "if the prediction is 'blue', how often is this true?")	0.903	0.772	0.085
Specificity	Proportion of samples predicted as not X, which are not X (e.g., "if the prediction is not 'blue', how often is it detected as 'not blue'"?)	0.866	0.859	0.999

Alternative statistical approaches exist. For example, the Snipper tool (http://mathgene.usc.es/snipper/) is based on a naïve Bayes model which applies the Hardy-Weinberg principle. Instead of training model parameters, this approach uses allele and genotype frequencies to calculate conditional probabilities for an observed set of genotypes (e.g., Pr[observed genotypes | Blue eye color]). Such likelihoods can then be converted into posterior probabilities, which would represent prediction probabilities.

For the prediction of ancestry, principal component analysis (PCA), naïve Bayes models and STRUCTURE are the most used statistical approaches (Phillips 2015). PCA is a cluster algorithm/tool that, based on observed genotype data, transforms existing variables (e.g., DNA markers and genotypes) into new variables that maximize the variation in the dataset. These new variables can then be used to create a scatter plot comprising all reference and unknown samples, from which interpretations can be performed. Additionally, it is possible to obtain numerical distances from the unknown sample to the center of each reference cluster to assign confidence to each PCA result.

In the basic naïve Bayes model, reference allele and genotype frequencies are used to calculate conditional probabilities for the observed genotype profile for unknown samples (e.g., Pr[observed genotypes | European ancestry]). These conditional probabilities can be used to calculate likelihood ratios for pairs of hypothesized ancestries, or to calculate posterior probabilities for all included categories. In more refined implementations, it is also possible to use these data to test whether there is at least one population in the reference data set that is sufficiently close to the unknown's ancestral population (Tvedebrink et al. 2018).

STRUCTURE (Pritchard et al. 2000) is another tool that, through an iterative analysis model, aims to find genetic clusters based on genotype similarities and dissimilarities among all samples included in a sample set. After the analysis, each sample gets a membership probability/coefficient for each cluster, which can be used to infer ancestry. This tool is one of the most widely used general population analysis programs.

Accuracies and other performance metrics can be analyzed for ancestry predictions in the same manner as for the phenotype prediction tools mentioned above.