Aalborg Universitet

Enhanced SNP genotyping with symmetric multinomial logistic regression

Nielsen, Malte Bødkergaard; Eriksen, Poul S.; Mogensen, Helle S.; Morling, Niels; Andersen, Mikkel M.

Published in:
Forensic Science International: Genetics

DOI (link to publication from Publisher):
10.1016/j.fsigen.2025.103291

Creative Commons License
CC BY 4.0

Publication date:
2025

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Nielsen, M. B., Eriksen, P. S., Mogensen, H. S., Morling, N., & Andersen, M. M. (2025). Enhanced SNP genotyping with symmetric multinomial logistic regression. *Forensic Science International: Genetics*, *78*, Article 103291. https://doi.org/10.1016/j.fsigen.2025.103291

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: January 25, 2026

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research Paper

Enhanced SNP genotyping with symmetric multinomial logistic regression

Malte B. Nielsen [a,*], Poul S. Eriksen [a], Helle S. Mogensen [b], Niels Morling [a,b], Mikkel M. Andersen [a,b]

[a] Department of Mathematical Sciences, Faculty of Engineering, Aalborg University, Aalborg, Denmark

[b] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

ARTICLE INFO

Keywords:
Symmetric multinomial logistic regression
Forensic genetics
Low DNA concentrations
Biallelic markers
SNP genotyping
Massively parallel sequencing

ABSTRACT

In genotyping, determining single nucleotide polymorphisms (SNPs) is standard practice, but it becomes difficult when analysing small quantities of input DNA, as is often required in forensic applications. Existing SNP genotyping methods, such as the HID SNP Genotyper Plugin (HSG) from Thermo Fisher Scientific, perform well with adequate DNA input levels but often produce erroneously called genotypes when DNA quantities are low. To mitigate these errors, genotype quality can be checked with the HSG. However, enforcing the HSG's quality checks decreases the call rate by introducing more no-calls, and it does not eliminate all wrong calls. This study presents and validates a symmetric multinomial logistic regression (SMLR) model designed to enhance genotyping accuracy and call rate with small amounts of DNA. Comprehensive bootstrap and cross-validation analyses across a wide range of DNA quantities demonstrate the robustness and efficiency of the SMLR model in maintaining high call rates without compromising accuracy compared to the HSG. For DNA amounts as low as 31.25 pg, the SMLR method reduced the rate of no-calls by 50.0% relative to the HSG while maintaining the same rate of wrong calls, resulting in a call rate of 96.0%. Similarly, SMLR reduced the rate of wrong calls by 55.6% while maintaining the same call rate, achieving an accuracy of 99.775%. The no-call and wrong-call rates were significantly reduced at 62.5–250 pg DNA. The results highlight the SMLR model's utility in optimising SNP genotyping at suboptimal DNA concentrations, making it a valuable tool for forensic applications where sample quantity and quality may be decreased. This work reinforces the feasibility of statistical approaches in forensic genotyping and provides a framework for implementing the SMLR method in practical forensic settings. The SMLR model applies to genotyping biallelic data with a signal (e.g. reads, counts, or intensity) for each allele. The model can also improve the allele balance quality check.

1. Introduction

In forensic genotyping, the accurate calling of single nucleotide polymorphisms (SNPs) is crucial but becomes a challenge when dealing with low amounts of DNA, which is typical for biological traces. While amplification-based genotyping tools, such as the HID SNP Genotyper Plugin (HSG) from Thermo Fisher Scientific (Waltham, MA, USA), excel with sufficient amounts of DNA, their performance declines with lower DNA quantities, resulting in increased erroneous genotype calls and reduced call rates [1–4].

To reduce the number of wrong calls (WCs), it seems natural to declare a no-call (NC) for genotypes not passing the quality checks (QCs) provided by the HSG. However, for low amounts of input DNA, this approach significantly decreases the call rate and still struggles to filter out WCs.

Other probabilistic approaches have been proposed to handle uncertainty in low-coverage sequencing [2], but they either integrate genotype likelihoods directly into downstream computations or rely on prior information about the genotypes, such as population allele frequencies, and some are limited to integer-valued data such as allele reads or counts.

We introduce a symmetric multinomial logistic regression (SMLR) model and a framework for refined NC declaration that improves genotyping accuracy and call rate, especially in challenging conditions. The SMLR model does not rely on prior genotype probabilities and can handle integer or continuous allele signals. Its symmetric formulation ensures it remains indifferent to the ordering or labelling of the two alleles, making the SMLR model robust and straightforward in its assumptions.

[*] Corresponding author.
E-mail addresses: maltebn@math.aau.dk (M.B. Nielsen), svante@math.aau.dk (P.S. Eriksen), helle.smidt@sund.ku.dk (H.S. Mogensen), niels.morling@sund.ku.dk (N. Morling), mikl@math.aau.dk (M.M. Andersen).

https://doi.org/10.1016/j.fsigen.2025.103291
Received 17 November 2024; Received in revised form 29 April 2025; Accepted 1 May 2025
Available online 17 May 2025
1872-4973/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

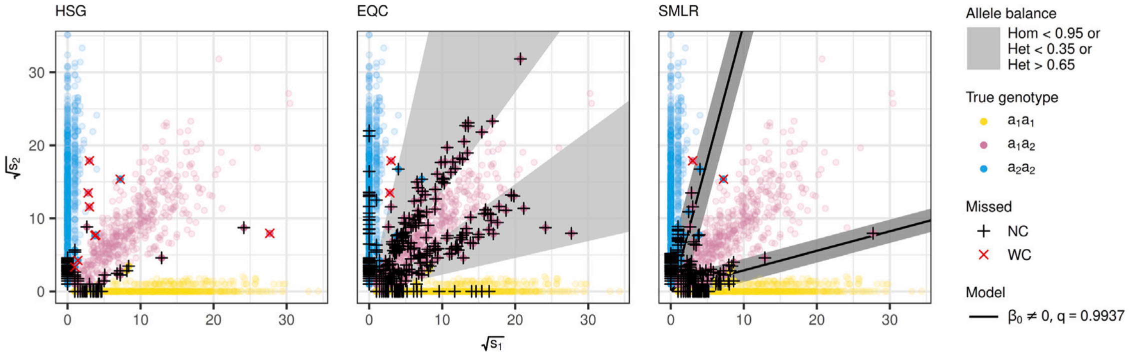

Fig. 1. Genotype predictions for the examinations of 31.25 pg DNA.
Each plot displays 1,931 SNP observations classified using the genotyping methods: HID SNP Genotyper Plugin (HSG), enforcing the quality checks (EQC), and symmetric multinomial logistic regression (SMLR). A dot represents a pair of SNP read counts (s_1, s_2). The true genotypes are coloured red for heterozygotes and blue or yellow for homozygotes. Failed genotype predictions are indicated by red crosses for wrong calls and black pluses for no-calls. In the EQC plot (middle), the grey areas show where the HSG is guaranteed to flag for allelic imbalance. In the SMLR plot (right), the solid lines show the decision boundaries of the SMLR model with an intercept fitted to square-root transformed allele signals. The grey area marks the no-call zone where genotype probabilities fall short of the threshold $q = 0.9937$ (a value chosen for illustrative purposes). Outside the grey area, the predicted genotype has $P(G \mid s_1, s_2) \geq q$.

We explore the SMLR model's efficiency and reliability, laying out guidelines for its application in forensic SNP genotyping. We also demonstrate the SMLR model's potential in quality control, particularly in identifying and managing allelic imbalance.

2. Materials and methods

2.1. The HID SNP Genotyper Plugin

The HSG software uses multiple metrics for SNP genotype determination [5, p. 35]. Along with its genotype calls, it outputs three quality checks:

- A locus-wise coverage check to indicate potential drop-outs (this QC flag was not observed in our data).
- A check of the strand balance where a percentage of positive coverage below 0.3 or above 0.7 results in a QC flag indicating imbalance.
- A check of the allele balance that flags homozygous calls if the ratio of the major allele's coverage to the total coverage of all four nucleotides falls below 0.95 and heterozygous calls if it falls outside the range of 0.35 to 0.65.

As seen in the leftmost and middle plots of Fig. 1, the HSG's genotype calls and determination of NCs are not based on these QCs alone, so genotypes are often called despite the presence of QC flags, and NCs are declared even when no QC flag is present [5, p. 35]. Therefore, enforcing the quality checks (EQC) by turning genotypes with QC flags into NCs will increase the number of NCs and thus reduce the call rate.

2.2. The symmetric multinomial logistic regression model

2.2.1. Model formulation

The SMLR model presents a statistical solution to biallelic genotyping challenges by using the allele signals to estimate conditional genotype probabilities. For a biallelic marker with alleles a_1 and a_2 having measured signals s_1 and s_2, we consider the unphased genotype G with outcomes $\{a_1a_1, a_1a_2, a_2a_2\}$ and the conditional genotype probabilities $p_{ij} = P(G = a_ia_j \mid s_1, s_2)$. Multinomial logistic regression is apt for modelling the conditional distribution of G given s_1 and s_2 with the heterozygous genotype as a baseline category for convenience and standardisation [6, p. 293]. However, it is desirable to model $P(G \mid s_1, s_2)$ in a way that is invariant to the labelling of the alleles by introducing a symmetry into the model equations, leading to the SMLR model:

$$\begin{aligned} \log\left(\frac{p_{11}}{p_{12}}\right) &= \beta_0 + \beta_1 f(s_1) + \beta_2 f(s_2), \\ \log\left(\frac{p_{22}}{p_{12}}\right) &= \beta_0 + \beta_2 f(s_1) + \beta_1 f(s_2). \end{aligned} \tag{1}$$

Here, the function f is a variance-stabilising transformation of the allele signals, e.g. $f(s_i) = \sqrt{s_i}$. In standard multinomial logistic regression, the second equation would have different β_i-parameters than the first, and the resulting parameter estimates would depend on which allele is labelled a_1 or a_2. The introduced symmetry eliminates this dependency and ensures that the model's behaviour is invariant to allele labelling. The conditional genotype probabilities become

$$\begin{aligned} p_{11} &= \frac{e^{\beta_0+\beta_1 f(s_1)+\beta_2 f(s_2)}}{1 + e^{\beta_0+\beta_1 f(s_1)+\beta_2 f(s_2)} + e^{\beta_0+\beta_2 f(s_1)+\beta_1 f(s_2)}}, \\ p_{22} &= \frac{e^{\beta_0+\beta_2 f(s_1)+\beta_1 f(s_2)}}{1 + e^{\beta_0+\beta_1 f(s_1)+\beta_2 f(s_2)} + e^{\beta_0+\beta_2 f(s_1)+\beta_1 f(s_2)}}, \\ p_{12} &= \frac{1}{1 + e^{\beta_0+\beta_1 f(s_1)+\beta_2 f(s_2)} + e^{\beta_0+\beta_2 f(s_1)+\beta_1 f(s_2)}}, \end{aligned} \tag{2}$$

where β_1 is expected to be positive, such that p_{11} increases with s_1 and p_{22} increases with s_2, and β_2 is expected to be negative, so that p_{11} and p_{22} decrease with increasing s_2 and s_1, respectively. It is expected that $\beta_1 < |\beta_2|$ such that p_{12} goes towards 1 as $s_1 = s_2$ grows large, aligning with the behaviour of heterozygous genotypes at high signal levels.

2.2.2. SNP genotype calling

With the SMLR model, the genotype calling for an observation is straightforward: the genotype to be called is the one associated with the highest conditional probability estimated from (2). The decision boundaries are the points (s_1, s_2) where at least one of the model equations from (1) equals zero. In other words, where the conditional probability of the heterozygous genotype equals the conditional probability of one of the homozygous genotypes. As shown in the SMLR plot in Fig. 1, these boundaries can be represented graphically in a plot of $f(s_2)$ versus $f(s_1)$ by the lines

$$\begin{aligned} f(s_2) &= -\frac{\beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} f(s_1), \\ f(s_2) &= -\frac{\beta_0}{\beta_1} - \frac{\beta_2}{\beta_1} f(s_1). \end{aligned} \tag{3}$$

The plot also illustrates how a measure of call confidence is implemented by introducing the user-defined probability threshold, q.

An observation is then declared an NC if its maximum conditional probability among the potential genotypes falls short of q. By setting $1/2 < q < 1$, observations on the decision boundaries are guaranteed to be declared NCs, yielding unambiguous call decisions even when multiple genotypes have equal conditional probability estimates. The call confidence increases with increasing q-values, corresponding to expanding a no-call zone around the decision boundaries within which all conditional genotype probabilities are below q.

2.2.3. Estimation

In multinomial logistic regression, maximum likelihood estimates (MLEs) of parameters are well-defined and unique when the data categories overlap and thus are not completely separable [7,8]. When fitting SMLR models to biallelic data, overlapping categories essentially mean that after applying the variance-stabilising transformation to the allele signals, at least one of the collections of homozygous points cannot be separated from the collection of heterozygous points by any straight line (see SMLR-plot in Fig. 1). For completely separable data, the true genotypes are perfectly partitioned by the decision boundaries. The ratios between the parameters, and thereby the decision boundaries in (3), are still well-defined, but the likelihood function will no longer have a unique maximum, and the MLEs will tend towards infinity unless constrained by a stopping criterion [6, p. 298].

The MLEs are determined by minimising the negative log-likelihood of the SMLR model, as derived in the supplementary material (S3). Using the R function ‘optim’ with the default Nelder–Mead method facilitates this process. It requires an initial guess for the parameter vector, which can be set to

$$\left(\beta_0^{\text{init}}, \beta_1^{\text{init}}, \beta_2^{\text{init}}\right) = (0, 1, -2) \quad \text{or} \quad \left(\beta_1^{\text{init}}, \beta_2^{\text{init}}\right) = (1, -2),$$

reflecting the expectations of $\beta_2 < 0 < \beta_1 < |\beta_2|$. These are merely expectations of the resulting parameters, not requirements or constraints on the model equations. Therefore, the initial guess can take values beyond these expectations, such as $(0, 0, -1)$.

2.3. Assessing the minimal sample size

Bootstrap analysis was applied to determine a reasonable sample size for fitting the SMLR model. By observing how the variance of parameter estimates decreases with increasing sample size, we estimated a point beyond which additional SNP observations or individuals would yield limited precision gains relative to the cost of further data collection [9].

The bootstrap analysis also demonstrates the stability of the SMLR model’s decision boundaries in scenarios of complete separation, which often occurs with smaller sample sizes, and it illustrates under which conditions complete separation is less likely.

2.4. Assessing the effectiveness of the SMLR model

The SMLR model is primarily designed as a predictive tool for classification, focusing on the practical application in forensic SNP genotyping rather than theoretical explanatory power [10]. As such, the evaluation metrics relevant for this study are call rate (CR) and accuracy (AC), where CR is defined as the percentage of calls that are not NCs, and AC as the percentage of correct calls when disregarding NCs:

$$CR = \frac{\text{Total calls - NCs}}{\text{Total calls}} \times 100,$$
$$AC = \frac{\text{Total calls - NCs - WCs}}{\text{Total calls - NCs}} \times 100.$$

To ensure fair comparisons between the SMLR model and the HSG, it is critical to analyse their performances under equivalent conditions. Therefore, cross-validation was used to assess the effectiveness of the SMLR model in increasing the call rate without compromising accuracy and vice versa. More precisely, the relative difference in call rates was assessed when the no-call zone of the SMLR model had a width providing the same level of accuracy as the HSG. Since it is not always possible to find a width that gives exactly the same accuracy for the SMLR model and the HSG, the accuracy of the SMLR model was set to the lowest value exceeding the accuracy of the HSG, i.e. its no-call zone was adjusted to the width where the model yielded the same number of WCs as the HSG or fewer. Conversely, the accuracies were compared when the SMLR model provided at least the same call rate as the HSG, i.e. the same number of NCs or fewer. This way, the comparisons are conservative by favouring the HSG over the SMLR model.

In general, the relative difference in call rates is proportional to the relative difference in NCs, and when the number of NCs for the SMLR model is equal to that of the HSG, the relative difference in accuracies is proportional to the relative difference in WCs:

$$\frac{CR_{\text{SMLR|HSG}} - CR_{\text{HSG}}}{CR_{\text{HSG}}} \propto \frac{NC_{\text{HSG}} - NC_{\text{SMLR|HSG}}}{NC_{\text{HSG}}}, \tag{4}$$
$$\frac{AC_{\text{SMLR|HSG}} - AC_{\text{HSG}}}{AC_{\text{HSG}}} \propto \frac{WC_{\text{HSG}} - WC_{\text{SMLR|HSG}}}{WC_{\text{HSG}}}. \tag{5}$$

Here, NC_{HSG} and WC_{HSG} are the no-calls and wrong calls of the HSG, while $NC_{\text{SMLR|HSG}}$ and $WC_{\text{SMLR|HSG}}$ are the corresponding counts for the SMLR model when its WCs and NCs are aligned to those of the HSG, respectively. The proportionalities (4) and (5) are derived in the supplementary material at (S1) and (S2).

For signals exhibiting variation consistent with a Poisson distribution, commonly observed for integer-valued data, the transformation $f(x) = \sqrt{x}$ is well established as an effective method for stabilising variance [11]. However, even a theoretically well-justified choice of f will not necessarily optimise the performance metrics in (4) and (5). Thus, to explore the effectiveness of various transformations and intercept configurations, six SMLR model formulations were tested: identity, square root, and logarithmic transformations

$$f\left(s_i\right) = s_i, \quad f\left(s_i\right) = \sqrt{s_i}, \quad \text{and} \quad f\left(s_i\right) = \log\left(s_i + 1\right),$$

each with and without an intercept (i.e. $\beta_0 \neq 0$ and $\beta_0 = 0$).

The models were evaluated through extensive cross-validation, a method that randomly divides the data into disjoint training and test subsets, the former used for model fitting and the latter exclusively to evaluate model performance [12,13]. Repeating this process with different data splits helps estimate how the model will perform on new datasets. To assess robustness and generalisability, several cross-validations were conducted, considering cases where training and test subsets were drawn from the same and different DNA dilutions.

This structured approach enables an objective comparison of the models and ultimately allows identification of the optimal model in forensic genetic contexts.

2.5. Software

For the analyses, we used R version 4.5.0 with the packages: ‘tidyverse’, ’future.apply’, and ‘xtable’ [14–20]. For creating figures, we used ImageMagick and the R packages: ‘ggplot2’, ‘ggnewscale’, ‘latex2exp’, and ‘patchwork’ [21–25]. The R scripts used for data analyses and figure generation are available on GitHub and Zenodo [26].

2.6. Data

This study analysed results of SNP typing with the Precision ID Ancestry Panel (Thermo Fisher Scientific), which includes 165 autosomal SNPs used to predict the biogeographic origin of humans. The laboratory methods were described by Pereira et al. [27].

The manufacturer recommends using 1ng DNA to increase the success rate with degraded DNA from compromised tissue samples. Others have demonstrated that good results can be obtained with smaller amounts of DNA if it is of good quality and modified experimental conditions are used [1].

Table 1
The SMLR model's genotyping improvements for all examined DNA quantities.

		First series:						Second series:			
		5 individuals, 4 examinations						18 individuals, 1 examination			
DNA quantity (pg)		1,000	500	250	125	62.5	31.25	50	25	12.5	6.25
Observed SNPs ($s_1+s_2>0$)		3,240	3,240	3,240	3,238	2,585	1,931	2,831	2,821	2,575	2,347
HSG[a]	No-calls	1	3	28	48	83	156	62	142	213	276
	Wrong calls	0	0	1	1	2	9	8	59	184	272
SMLR[d]	q aligning WCs[b]	0	0	0	0.5473	0.5473	0.7394	0.8115	0.8449	0.8302	0.8541
	Call rate (%)	100	100	100	99.8	99.7	96.0	98.4	95.7	95.1	92.2
	No-calls	0	0	0	5	7	78	46	121	125	184
	NC-reduction (%)	100	100	100	89.6	91.6	50.0	25.8	14.8	41.3	33.3
SMLR[d]	q aligning NCs[c]	0.8890	0.8808	0.8588	0.8592	0.8514	0.8414	0.8488	0.8592	0.8725	0.8841
	Accuracy (%)	100	100	100	100	100	99.775	99.747	97.878	92.760	88.048
	Wrong calls	0	0	0	0	0	4	7	57	171	248
	WC-reduction (%)	0	0	100	100	100	55.6	12.5	3.4	7.1	8.8

[a] Number of no-calls (NCs) and wrong calls (WCs) made by the HID SNP Genotyper Plugin (HSG).

[b] The smallest q that aligns the SMLR model's WCs according to (4).

[c] The smallest q that aligns the SMLR model's NCs according to (5).

[d] Results are conservative and based on the SMLR model with an intercept fitted to square-root transformed allele signals from examinations of 25 pg and 50 pg DNA.

Two DNA dilution series were analysed: The first series included six two-fold DNA dilutions from each of five individuals, with 1 ng, 500 pg, 250 pg, 125 pg, 62.5 pg, and 31.25 pg DNA, repeated in four examinations. Due to limitations in the initial DNA amounts, one of the examinations included only a single individual at 62.5 pg DNA, and another included only two individuals at 31.25 pg. This series included six complete SNP profile measurements for all individuals, with additional partial data. For each individual, these complete SNP profiles were identical and were adopted as the true genotypes for subsequent analyses.

The second series was a single examination of four two-fold DNA dilutions from each of 18 individuals with 50 pg, 25 pg, 12.5 pg, and 6.25 pg DNA. The individuals' SNP profiles were known from previous analyses. Details on the number of observed SNPs in each series are found in Table 1.

The project is registered in the University of Copenhagen's joint record of biobanks and record of research projects containing personal data (514-1056/24-3000). It complies with the rules of the General Data Protection Regulation (Regulation (EU) 2016/679).

2.6.1. Initial SNP calling

The primary sequencing analysis was performed with Torrent Suite Software v4.6 (Thermo Fisher Scientific). BAM files were generated using the HSG (v4.3.1). No noise filter was used.

Some SNPs are known to have rare variants. Therefore, it was investigated whether rare alleles, different from the two expected alleles, were present. The HSG does this through its QC of the allele balance [5, p. 35]. One marker, rs7722456, exhibited anomalous adenine reads and was excluded from the main analyses, along with the markers rs459920 and rs7251928, as recommended by Pereira et al. [27], leaving data from 162 SNPs for analysis. However, rs7722456 was retained as example data to illustrate how the SMLR model can enhance the detection rate of imbalance when used for QC of the allele balance.

SNP data with zero reads for both alleles were removed from the dataset, as they provided no useful information and consistently resulted in NCs for the HSG.

3. Results

For low amounts of DNA, the HSG-plot and EQC-plot in Fig. 1 reveal that using the QC-flags to improve the accuracy is inefficient: converting flagged observations to NCs has a significant cost to the call rate and does not eliminate all WCs. Hence, comparison with the EQC method is not meaningful. The ensuing results affirm the SMLR model's utility in forensic genotyping, demonstrating its capacity to improve call rate and accuracy across various testing scenarios.

3.1. Pre-analysis

Table 1 demonstrates how a fit of the SMLR model with an intercept and a square root transformation reduces the NCs for each DNA amount examined when compared to the HSG using (4) and similarly reduces the WCs when compared using (5). Thus, for each column in Table 1, this SMLR model gives fewer NCs and WCs when the q-threshold lies between the model's two alignment values displayed in that column, implying simultaneous improvements in call rate and accuracy. Since the metrics in (4) and (5) favour the HSG, and the cross-validations presented later show that the fit used in Table 1 is not necessarily the best performing one, the improvements displayed in the table are conservative.

3.1.1. Relevant dilutions

Fig. 2 shows how the distribution of $\left(\sqrt{s_1}, \sqrt{s_2}\right)$ becomes more spread out and how the HSG makes more NCs and WCs as the DNA amount decreases. It further illustrates that most WCs at low DNA amounts are heterozygous genotypes misidentified as homozygous due to a reduced signal for one of the alleles. For the SMLR model, this pattern necessitates higher q-threshold settings for accurate genotyping, resulting in lower call rates. Consequently, genotyping with the Precision ID Ancestry Panel requires more than 25 pg DNA to maintain acceptable call rates.

Table 1 highlights performance variations across different DNA input levels. At 500 pg DNA and above, the HSG performs adequately, rendering the SMLR model redundant. At 250 pg DNA, the HSG produces a few WCs and introduces significantly more NCs with each step down the dilution series. Fig. 2 and Table 1 show that the HSG makes significantly more WCs at 50 pg DNA and below.

Based on these observations, the following analyses focus on testing genotyping performance within the two ranges of DNA quantities: 31.25–50 pg and 62.5–250 pg. The former is where the call rate of the HSG becomes critically low and where it more frequently makes WCs. However, to assess robustness and generalisability, the SMLR models will also be fitted using data from exterior DNA quantities.

3.1.2. Parameter variance, sample size, and separation

Bootstrap analyses were performed for the six SMLR variants mentioned in the Materials and methods section. They all showed similar results to those depicted in Supplementary Fig. S1. Here, we see a less pronounced decline in the parameters' variances after a sample size of around nine individuals or roughly 1,460 SNPs. This sample size provides a good balance between the proportions of data used for fitting

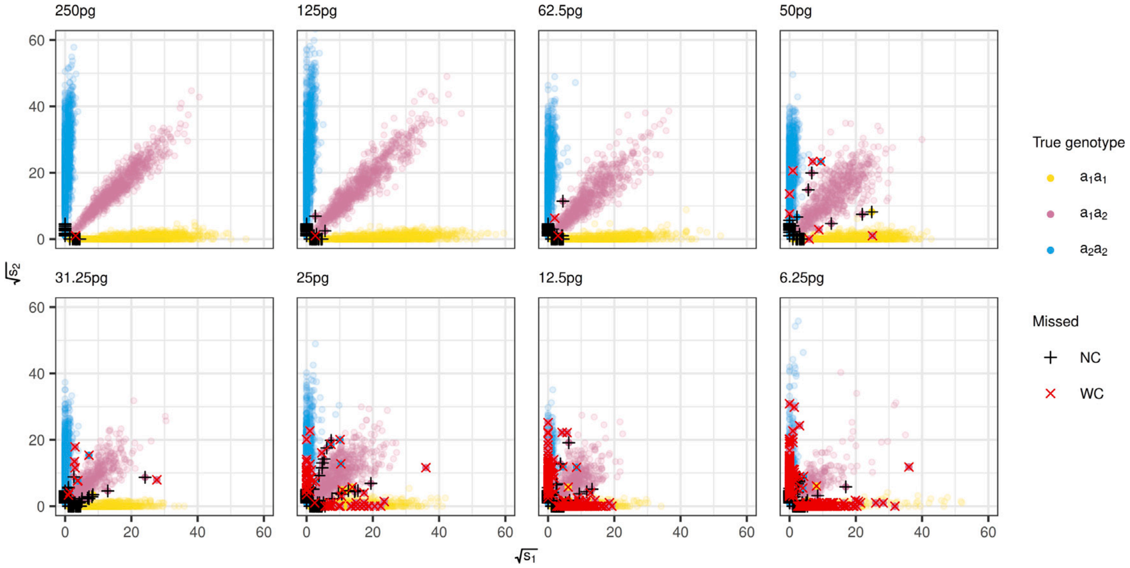

Fig. 2. Distribution of square-root transformed allele signals.
A dot represents a pair of SNP read counts (s_1, s_2). The true genotypes are coloured red for heterozygotes and blue or yellow for homozygotes. The displayed DNA quantities indicate where the accuracy of the HID SNP Genotyper Plugin falls below 100%, with its wrong calls marked by red crosses and no-calls by black pluses.

and testing in the cross-validation, especially in the least populated examination (31.25 pg DNA), where the use of nine individuals for fitting corresponds to 75% of the data.

Supplementary Fig. S1 indicates that the parameter estimates and decision boundaries depend on the DNA amount, with the effect becoming more noticeable at 12.5 pg and below. It also shows that complete separation can be avoided by including examinations of very low DNA amounts (e.g. 25 pg) in the fitting data, and that this does not substantially alter the decision boundaries.

3.2. Validation of the SMLR model

3.2.1. Cross-validation insights

The cross-validation experiments in Supplementary Fig. S2 demonstrate that the strong results from Table 1 were not coincidental. The figure shows that when either the square root or the logarithmic transformation was applied to the allele signals, the SMLR model generally outperformed the HSG with substantial reductions in NCs and WCs, leading to simultaneous improvements in call rate and accuracy.

The subplots in the figure's first and fourth columns show nearly horizontal performance lines, indicating that fitting to combined data from the examinations of 25 pg and 50 pg DNA gave particularly stable SMLR models for both the low (31.25–50 pg) and high (62.5–250 pg) DNA amounts. Therefore, the subsequent analyses are based on these fits to let the results inherit this stability.

Supplementary Fig. S2 intends to compare each SMLR variant to the HSG and not to make comparisons between the variants themselves, as a variant that excels when aligned to the HSG is not necessarily the best performing for other balances between the call rate and accuracy.

Fig. 3 compares the SMLR variants fitted to data subsets from the examinations of 25 pg and 50 pg DNA: the call rates from the 1,000 cross-validation iterations were rounded to one decimal place, and the median accuracy for each variant was determined at each rounded call rate. This approach allows for a more precise assessment of how the accuracy of each SMLR variant changes with the call rate.

The red dot in each plot of Fig. 3 represents the median accuracy and call rate of the HSG. The lines for the models with $f\left(s_i\right) = \sqrt{s_i}$ and $f\left(s_i\right) = \log\left(s_i + 1\right)$ are generally seen to move well to the left of and above these dots. Thus, these four SMLR models convincingly outperform the HSG in median accuracy and call rate. For the low DNA amounts, they increase the call rate to more than 95% while still achieving higher accuracies than the HSG. At the highest call rates, the square-root model without an intercept (dark blue line) achieves slightly higher median accuracies than the other variants.

To avoid overestimating the performance of the SMLR method framework, the remaining analyses were based on the square-root model with an intercept, as this is a more conservative variant given that better SMLR models exist.

3.2.2. Call rate and accuracy dependencies

The accuracy of the SMLR model depends on the call rate through the width of the no-call zone, which is controlled by the value of the probability threshold q. This dependency on q is depicted in Fig. 4, where the previously selected SMLR model is applied to the range of low DNA quantities. The left plot shows that the SMLR model surpasses the HSG's median accuracy of 99.567% at a probability threshold of $q = 0.74$. At this q-value, the right plot shows that the SMLR model has a median call rate of 96.6%, i.e. an increase of 3.0 percentage points compared to the HSG.

The model's median call rate remains higher than that of the HSG until $q = 0.84$, where it achieves a median accuracy of 99.727%. The left plot (Fig. 4) shows that this is above the maximum accuracy observed for the HSG (99.647%) during the 1,000 cross-validation iterations.

When applied to DNA quantities of 500 pg and above, Table 1 shows that the SMLR model achieves 100% accuracy and call rate. This demonstrates its capability to maintain high performance across varying DNA inputs.

3.2.3. SMLR for quality checks

The locus rs7722456 had many reads of an unexpected nucleotide (adenine) for all examinations in the first dilution series. The HSG's quality control flagged only 37 of the 108 unusual observations with its QC flag for allelic imbalance. The limited efficiency in identifying allele balance issues was consistent for all DNA quantities examined in this dilution series.

In Supplementary Fig. S3, the major allele signals for rs7722456 are plotted against the adenine signals, along with the decision boundaries of the previously selected SMLR model. Using the probability threshold $q = 0.99$, 106 points were in the no-call zone, indicating an allelic imbalance. The choice of $q = 0.99$ is a mere example illustrating how the SMLR model can also be used for QC and why this statistical approach is more effective than applying a static threshold to the coverage ratio [5, p. 35].

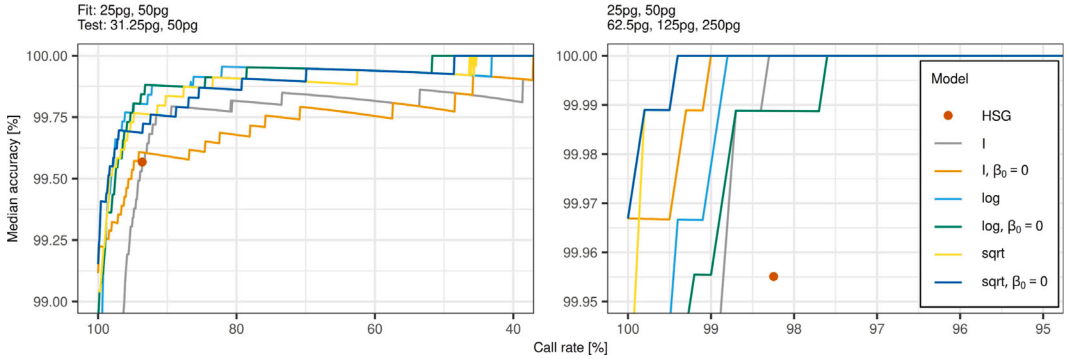

Fig. 3. Median accuracy versus call rate from aggregated results of cross-validations (legend applies to both plots).
Each line represents the median accuracies calculated from binned call rate data of 1,000 cross-validation iterations, where the six SMLR models were fitted to and tested on data from the examinations of the DNA quantities indicated above the plots. The red dot in each plot shows the median accuracy and call rate for the HID SNP Genotyper Plugin. Note that the two plots use different axes scales, and that the first axes have been reversed to align with Fig. 4, where increasing q leads to decreasing call rates.

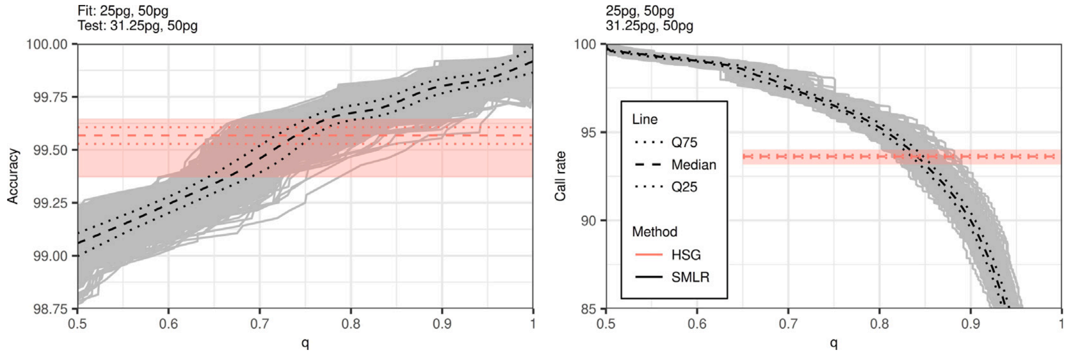

Fig. 4. Performance of the SMLR model across probability thresholds, q (legend applies to both plots).
The grey lines depict the SMLR model's performance in 1,000 cross-validation iterations. The model was fitted with an intercept to square-root transformed allele signals and tested on data from the examinations of the DNA quantities indicated above the plots. Dotted lines mark the 25th and 75th percentiles among the 1,000 cross-validations, while dashed lines indicate medians. The shaded horizontal bars represent the accuracy (left) and call rate (right) ranges for the HID SNP Genotyper Plugin.

4. Discussion

This study aimed to develop an effective method for SNP calling in forensic genetics. The developed SMLR model is simple, using only two or three parameters to compute three conditional genotype probabilities from two input signals. It is possible to use more complex variance-stabilising transformations than those investigated in this study. Still, the aim was to demonstrate the capabilities of the model while retaining as much of its inherent simplicity as possible.

Some immediate benefits of this simplicity are increased robustness and a straightforward implementation of a no-call zone. Conversely, for a non-symmetric model, changing the labels for a subset of the alleles will alter the parameter estimates, making the model less robust. Non-symmetric decision boundaries may also make some nucleotides more prone than others to be declared NCs, creating an NC bias. Furthermore, even with significant differences in the nucleotide-signal distributions (such that certain allele labellings make more sense than others), it is still likely that the SMLR model's equal treatment of alleles will yield higher accuracy compared to a non-symmetric model due to the bias–variance trade-off [10]. However, we do not see any alarming differences in the nucleotides' signals (see Supplementary Fig. S4).

An alternative approach to handling uncertainty in low-coverage data is the probabilistic model described by Mostad et al. (2023) [2], which derives genotype likelihoods $P(s \mid g)$ and combines them with prior genotype probabilities $P(g)$ to compute $P(g \mid s)$ via Bayes' theorem. This structure is particularly useful when allele frequency information is known and reliably specified, such as in population-based or familial relationship inference contexts. However, prior information is often unavailable in forensic genotyping scenarios, such as criminal investigations or cases involving unidentified remains. The SMLR model, by contrast, estimates $P(g \mid s)$ directly without relying on external priors, thereby avoiding potential bias from misspecified allele frequencies. This distinction makes the SMLR framework well-suited for general forensic use, especially in cases involving unknown individuals or populations.

Measuring the performance of the SMLR variants is not straightforward when NCs are introduced and should be judged neutrally, i.e. from the perspective that they are meant to remove WCs, whereby NCs should not be seen as incorrect genotype predictions. This is further complicated as performance changes with the amount of input DNA. However, the SMLR model using a square-root transformation emerged as the best all-in-one model, offering a good balance between NCs and WCs for DNA quantities as low as 31.25 pg, particularly when modelled without an intercept.

The SMLR method was compared with the commonly used HSG, which performs well with high DNA quantities but leaves room for improvement at lower DNA amounts. The SMLR models using a square-root transformation generally perform better than the HSG on the examined DNA quantities. Consistently outperforming a genotyping method with high accuracy and call rate, such as the HSG, is a significant achievement that demonstrates the applicability of the SMLR method for biallelic genotyping, particularly in forensic genetics settings.

While the improvements in call rate and accuracy were more modest for the high DNA amounts recommended by kit manufacturers, we demonstrated that the SMLR framework substantially increased the detection rate of allelic imbalance independently of the examined DNA quantity (Supplementary Fig. S3). As a result, any laboratory can benefit from the SMLR framework, regardless of the DNA quantities they typically work with.

Since data characteristics may vary when using SNP platforms and procedures different from those in this study, the SMLR model's parameters may need adjustment and validation before being implemented with new genotyping methods. It should be noted that the effect of changes in the probability threshold, q, on the width of the no-call zone depends on the parameter estimates, particularly on the choice of variance-stabilising transformation. A tailored dilution series can help define the optimal setup for each lab environment, especially in determining the ideal value for the probability threshold q.

The SMLR method can be adapted with minor modifications to analyse data from other biallelic genetic systems, e.g. insertion-deletions. The SMLR principle can also classify non-genetic data with similar structures.

The bootstrap analysis showed that using approximately 1,460 SNP observations markedly decreases the parameter estimates' variances when fitting the SMLR model (Supplementary Fig. S1). For the Precision ID Ancestry Panel, this corresponds to approximately nine complete SNP profiles, with the option of repeating examinations of the same individual, e.g. by measuring the SNP profiles of three individuals, each repeated in three examinations. However, if feasible, we recommend using larger sample sizes to enhance precision further.

For robust fitting, it is important to have a reasonable overlap between the genotype clusters, preferably by measuring sufficient data in the 31.25–50 pg DNA range. If this is not feasible, including a few examinations with as little as 25 pg DNA can effectively achieve overlap. However, the bootstrap analysis showed that adding data from examinations of extremely low DNA quantities, like 12.5 pg, may negatively distort the model's decision boundaries.

In the Estimation subsection, it was suggested that the initial value of the parameter vector be set to $(1, -2)$ or to $(0, 1, -2)$ if an intercept is included. Depending on the scale of the signals and the choice of variance-stabilising transformation, other initial values may be more suitable. For example, we encountered issues with the initialisation of 'optim' when using the suggested initial values for the models with an identity transformation. However, these issues were resolved by setting $\beta_1^{\text{init}} = 0$ and $\beta_2^{\text{init}} = -1$. Initialisation problems can occur due to numerical overflow, e.g. if the value of the log-likelihood function at the chosen initial value exceeds the limit the computer can represent. Such issues are of general concern for optimisation tasks and are not specific to fitting the SMLR model.

Maximum likelihood estimation (MLE) focuses on optimally placing the decision boundaries to classify genotypes based on the observed data. In this sense, MLE can be said to optimise accuracy when the probability threshold q is zero. However, the introduction of a no-call zone creates a second decision layer that alters the classification process in a non-trivial way. For example, before applying a probability threshold to convert WCs into NCs, the model with $\beta_0 \neq 0$ may achieve a higher likelihood and yield fewer WCs. Yet once a threshold is applied, the performance metrics (4) and (5) may favour the simpler model with $\beta_0 = 0$. This occurs because MLE does not account for how the parameter estimates influence the width and shape of the no-call zone.

Although the cross-validations demonstrated that MLE yields effective SMLR models, it may still be possible to improve accuracy and call rate further by considering estimation procedures that directly incorporate the no-call mechanism. Research into such methods – moving beyond pure classification to explicitly optimise performance in the presence of NCs – could be valuable for forensic genotyping.

Genetic software like GenoGeographer [28–30] typically assumes that genotype calls are correct once they pass quality filters, overlooking the inherent uncertainty in genotyping, particularly in forensic genetics, where samples of low quality and quantity are common. A statistical approach, such as the SMLR model, provides estimates of the conditional genotype probabilities, which can help mitigate this uncertainty. However, high accuracy of a model's predictions does not guarantee that its probability estimates align with the genotype frequencies in real data [10]. To integrate such probabilities successfully into existing software, empirical verification is crucial. We conducted separate preliminary empirical assessments for the homozygous and heterozygous genotypes within the combined data of 31.25 and 50 pg DNA. Through binomial testing on groups of observations, we evaluated whether the mean of the model's probability estimates for its predicted genotype could explain the observed proportion of that genotype. This method confirmed that the SMLR model's probability estimates align well with the empirical genotype distributions, reinforcing our confidence in its utility for forensic applications.

5. Conclusion

The SMLR model improved the SNP calling efficiency, particularly with suboptimal DNA amounts, as is often the case in forensic genetic examinations of stain material in criminal investigations. At 31 pg DNA, the NC rate was reduced by approximately 50% compared to the HSG, while maintaining the same WC rate as the HSG. The WC rate was reduced by over 50% while maintaining the same NC rate. At 62–250 pg DNA, the no-call rate was dramatically reduced. Using SMLR for quality checks of the allele balance substantially improved imbalance detection regardless of the DNA input level.

CRediT authorship contribution statement

Malte B. Nielsen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Poul S. Eriksen:** Writing – review & editing, Supervision. **Helle S. Mogensen:** Supervision, Data curation. **Niels Morling:** Writing – review & editing, Supervision. **Mikkel M. Andersen:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Martin Jensen, Sidsel Raaby, and Hai Trieu Nguyen for laboratory assistance and Martin Tuwel Jensen for technical assistance.

Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.fsigen.2025.103291.

References

[1] M. Al-Asfi, D. McNevin, B. Mehta, D. Power, M.E. Gahan, R. Daniel, Assessment of the precision ID ancestry panel, Int. J. Leg. Med. 132 (6) (2018) 1581–1594, http://dx.doi.org/10.1007/s00414-018-1785-9.

[2] P. Mostad, A. Tillmar, D. Kling, Improved computations for relationship inference using low-coverage sequencing data, BMC Bioinformatics 24 (1) (2023) http://dx.doi.org/10.1186/s12859-023-05217-z.

[3] J.M. Butler, Advanced Topics in Forensic DNA Typing: Methodology, third ed., Academic Press, 2012, http://dx.doi.org/10.1016/C2011-0-04189-3.

[4] Scientific Working Group on D.N.A. Analysis Methods (SWGDAM), SWGDAM Interpretation Guidelines for Single Nucleotide Polymorphism (SNP) Analysis by Forensic DNA Testing Laboratories, SWGDAM, 2024, URL https://www.swgdam.org/publications.

[5] Thermo Fisher Scientific, HID SNP Genotyper Plugin User Guide, v5.2.2, 2017, URL https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0010641_HIDSNP_Genotyper_Plugin.pdf, Publication Number: MAN0010641, Revision: D.0.

[6] A. Agresti, Categorical data analysis, third ed., in: Wiley Series in Probability and Statistics, vol. 792, Wiley, Somerset, 2013.

[7] A. Albert, J.A. Anderson, On the existence of maximum likelihood estimates in logistic regression models, Biometrika 71 (1) (1984) 1–10, http://dx.doi.org/10.1093/biomet/71.1.1, URL https://www.jstor.org/stable/2336390.

[8] E. Lesaffre, A. Albert, Partial separation in logistic discrimination, J. R. Stat. Soc.: Ser. B (Methodological) 51 (1) (1989) 109–116, http://dx.doi.org/10.1111/j.2517-6161.1989.tb01752.x, URL https://www.jstor.org/stable/2345845.

[9] B. Efron, Bootstrap methods: Another look at the jackknife, Ann. Stat. 7 (1) (1979) 1–26, http://dx.doi.org/10.1214/aos/1176344552, URL https://www.jstor.org/stable/2958830.

[10] G. Shmueli, To explain or to predict? Statist. Sci. 25 (3) (2010) 289–310, http://dx.doi.org/10.1214/10-STS330.

[11] M.S. Bartlett, The use of transformations, Biometrics 3 (1947) 39–52, http://dx.doi.org/10.2307/3001536, URL https://www.jstor.org/stable/3001536.

[12] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc.: Ser. B (Methodological) 36 (2) (1974) 111–147, http://dx.doi.org/10.1111/j.2517-6161.1974.tb00994.x, URL https://www.jstor.org/stable/2984809.

[13] R.R. Picard, R.D. Cook, Cross-validation of regression models, J. Am. Stat. Assoc. 79 (387) (1984) 575–583, http://dx.doi.org/10.1080/01621459.1984.10478083, URL https://www.jstor.org/stable/2288403.

[14] R. Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2025, URL https://www.R-project.org/, v4.5.0.

[15] H. Wickham, M. Averick, J. Bryan, W. Chang, L.D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T.L. Pedersen, E. Miller, S.M. Bache, K. Müller, J. Ooms, D. Robinson, D.P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the tidyverse, J. Open Source Softw. 4 (43) (2019) http://dx.doi.org/10.21105/joss.01686.

[16] H. Wickham, tidyverse: Easily install and load the 'Tidyverse', 2023, URL https://cran.r-project.org/package=tidyverse, R package v2.0.0.

[17] H. Bengtsson, A unifying framework for parallel and distributed processing in R using futures, R J. 13 (2) (2021) 208–227, http://dx.doi.org/10.32614/RJ-2021-048.

[18] H. Bengtsson, future: Unified parallel and distributed processing in R for everyone, 2023, URL https://cran.r-project.org/package=future, R package v1.33.1.

[19] H. Bengtsson, future.apply: Apply function to elements in parallel using futures, 2024, URL https://cran.r-project.org/package=future.apply, R package v1.11.3.

[20] D.B. Dahl, D. Scott, C. Roosen, A. Magnusson, J. Swinton, xtable: Export tables to LaTeX or HTML, 2019, URL https://CRAN.R-project.org/package=xtable, R package v1.8-4.

[21] ImageMagick Studio LLC, ImageMagick, URL https://imagemagick.org, v6.9.12.98.

[22] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, 2016, URL https://ggplot2.tidyverse.org.

[23] E. Campitelli, ggnewscale: Multiple fill and colour scales in 'ggplot2', 2025, URL https://CRAN.R-project.org/package=ggnewscale, R package v0.5.1.

[24] S. Meschiari, latex2exp: Use LaTeX expressions in plots, 2022, URL https://CRAN.R-project.org/package=latex2exp, R package v0.9.6.

[25] T.L. Pedersen, patchwork: The composer of plots, 2024, URL https://CRAN.R-project.org/package=patchwork, R package v1.3.0.

[26] M.B. Nielsen, SMLR-genotyping, 2025, http://dx.doi.org/10.5281/zenodo.15341884, Published on GitHub and preserved on Zenodo. URL https://github.com/maltebn/SMLR-Genotyping.git.

[27] V. Pereira, H.S. Mogensen, C. Børsting, N. Morling, Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers, Forensic Sci. Int.: Genet. 28 (2017) 138–145, http://dx.doi.org/10.1016/j.fsigen.2017.02.013.

[28] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Weight of the evidence of genetic investigations of ancestry informative markers, Theor. Popul. Biology 120 (2018) 1–10, http://dx.doi.org/10.1016/j.tpb.2017.12.004.

[29] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, GenoGeographer — A tool for genogeographic inference, Forensic Sci. Int.: Genet. Suppl. Ser. 6 (2017) 463–465, http://dx.doi.org/10.1016/j.fsigss.2017.09.196.

[30] T. Tvedebrink, genogeographer: Methods for analysing forensic ancestry informative markers, 2019, URL https://cran.r-project.org/package=genogeographer.