



OPEN ACCESS

EDITED BY

Joao Carlos Setubal,
University of São Paulo, Brazil

REVIEWED BY

Soham Sengupta,
St. Jude Children's Research Hospital,
United States
Youtao Lu,
University of Pennsylvania, United States

*CORRESPONDENCE

Amel Haj Khelil,
✉ amel.hk@fsm.rnu.tn

RECEIVED 11 January 2025

ACCEPTED 27 May 2025

PUBLISHED 17 June 2025

CITATION

Attaoui A, Foddha H, Othman H, Ben
Abdennebi H and Haj Khelil A (2025) Utility of
regional STR marker variations in Tunisian and
sub-Saharan populations: insights into
forensic and population genetics.
Front. Bioinform. 5:1550730.
doi: 10.3389/fbinf.2025.1550730

COPYRIGHT

© 2025 Attaoui, Foddha, Othman, Ben
Abdennebi and Haj Khelil. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with
these terms.

Utility of regional STR marker variations in Tunisian and sub-Saharan populations: insights into forensic and population genetics

Asma Attaoui¹, Hajer Foddha¹, Houcemeddine Othman^{2,3,4},
Hassen Ben Abdennebi¹ and Amel Haj Khelil^{1,5*}

¹Laboratory of Human Genome and Multifactorial Diseases (LR12ES07), Faculty of Pharmacy, University of Monastir, Monastir, Tunisia, ²Department of Genetics, Farhat Hached University Hospital, Sousse, Tunisia, ³Laboratory of Cytogenetics, Molecular Genetics and Reproductive Biology (LR03SP02), Farhat Hached University Hospital, University of Sousse, Sousse, Tunisia, ⁴Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa, ⁵Department of Molecular and Cellular Biology, Higher Institute of Biotechnology of Monastir, University of Monastir, Monastir, Tunisia

Introduction: This study investigates the genetic variability and forensic applicability of Short Tandem Repeat (STR) loci including autosomal, X and Y-STR markers, across distinct Tunisian regions and among sub-Saharan African populations. Our objectives were to examine the regional allelic diversity of STR markers in Tunisia, and to assess the utility of these markers for forensic differentiation between Tunisian and sub-Saharan African.

Methods: Twenty two STRs were genotyped in 500 Tunisian individuals and 501 sub-Saharan corpses by capillary electrophoresis using commercial system kits. A Chi-square test for homogeneity was applied to assess allele distribution and Principal Component Analysis to assess geographical allele variations. Bioinformatic methods in R packages were used, such as Logistic Regression Model to predict geographic group membership and Random Forest models to evaluate the discriminative power of the analyzed STRs.

Results and Discussion: Statistical analyses revealed significant allelic variability between Northern, Central, and Southern Tunisia for markers such as D1S1656, D8S1179, and CSF1PO. PCA illustrated a clear genetic distinction between Tunisian and sub-Saharan populations, largely attributable to geographical and historical gene flow barriers. LRM achieved high accuracy (95.96%) in predicting geographic affiliation. RF analysis identified DYS391 as highly discriminative in population differentiation. Our findings align with prior research on Tunisian genetic diversity and extend this knowledge by illustrating allelic frequency variations in order to establish region-specific databases.

Conclusion: This study contributes valuable insights into the genetic structure of Tunisian and sub-Saharan populations, emphasizing tailored approaches in forensic practices.

KEYWORDS

allele frequency, genotype, power of discrimination, random forest, Y-STR

Introduction

STR (Short Tandem Repeat) markers have become indispensable in human genetics due to their high variability, making them essential tools for individual identification in criminal cases and differentiation of populations and human migration. These markers are particularly valuable because they have a high degree of polymorphism, allowing geneticists to generate unique DNA profiles.

Fundamental works such as DNA Typing (Butler, 2005) showed how DNA can be used to identify suspects or victims with a very high degree of certainty and highlighted cases where DNA has helped solve complex investigations or exonerate people convicted of crime. The work of John Butler, has contributed significantly to the professionalization and reliability of forensic practices. By introducing standard techniques and rigorous protocols, he helped make DNA analysis a fundamental pillar in the judicial system, recognized for its high level of accuracy (Butler, 2023). Genotyping of autosomal STR has been, for a long time performed by capillary electrophoresis (CE) (Kaushik and Sahajpal, 2020). In addition to autosomal STRs those on the Y-chromosome are of special utility in criminal cases (Ruitberg et al., 2001; Ding et al., 2021; Barni et al., 2024).

Actually, numerous kits are developed to sequence a combination of autosomal-/X-/Y-STRs (Guo et al., 2017), identity single nucleotide polymorphisms (SNP) (Yagasaki et al., 2022) and mitochondrial genome (Holt et al., 2019), and can be expanded to include phenotype-and ancestry-informative SNPs. The use of massively parallel sequencing for forensic applications has expanded rapidly in the last few years allowing the transition from forensic genetics to forensic genomics supported by high throughput sequencing (Kayser and Parson, 2017; McCord and Lee, 2018; Haddrill, 2021; Kayser et al., 2023).

In addition to the great usefulness of STR markers in forensic investigations, genetic diversity studies deepened forensic genetics by providing unique information on regional genetic composition. The populations of North Africa, and particularly Tunisia, are of particular interest because of their rich historical mix of diverse populations, including Berber, Arab, Phoenician and sub-Saharan African groups. This mixture has resulted in distinct genetic profiles that distinguish North African populations from neighboring regions (El Moncer et al., 2010; Mejri et al., 2022).

Genetic diversity studies highlight the importance of creating region-specific genetic databases in forensic applications, as allele frequencies in North Africa differ significantly from those in Europe or the Middle East. Such databases are essential to improve the accuracy and reliability of forensic identifications in these populations (Frigi et al., 2014; Kostiuchenko et al., 1997).

In this study, we aim to, firstly, test on a large scale with regional comparison, the most discriminative markers in the Tunisian population, secondly, to assess the degrees of genetic relationships between Tunisian and sub-Saharan populations and, finally, to exploit results for application in the identification of Tunisian criminals and sub-Saharan migrant corpses.

Materials and methods

Sampling

This study involved two distinct samples collected in the Tunisian Scientific and Forensic Laboratories. The first sample included 500 individuals with criminal records, representing three geographic regions of Tunisia: North (133 individuals), Center (217 individuals), and South (105 individuals). Oral swabs (Milne et al., 2006) containing DNA from these individuals were gathered as evidence. The second sample consisted of 501 DNA samples extracted from the remains of sub-Saharan migrants recovered from Tunisian coastal areas. Both sample sets were pre-analyzed in line with established recommendations (Freeman et al., 2003). Samples were collected following requisitions by the relevant authorities (judicial, police). In the present study, ethical standards were meticulously followed throughout the sampling process. First, formal approval was obtained from our institution's ethics committee prior to the commencement of any research activities. The confidentiality and anonymity of all data were vigilantly protected under the supervision of the Director of the Tunisian Forensic Science Department, who took proactive measures to mitigate any potential risks related to the unauthorized exploitation of participants' personal information and biological samples. All sensitive data were handled with the utmost care and in strict accordance with prevailing legal and ethical regulations. Biological samples were collected in full compliance with established ethical principles and local laws, with particular emphasis on securing informed consent and protecting the rights of participants, especially those involved in legal proceedings.

DNA extraction and genotyping

DNA extraction from oral swabs and genotyping were performed using the Investigator® 24plex QS Kit, Cat. No. 382415, QIAGEN, Hilden, Germany. This kit is designed for human identification, allowing directly multiplex amplification of 21 autosomal STR, one Y-STR and the Amelogenin marker for gender identification. It integrates an innovative quality sensor, which allows generating additional data that is very valuable for quality control and performance testing. In addition, it has high sensitivity and reliability with forensic samples (Hares, 2005).

DNA from human remains was extracted using the PrepFiler™ BTA Forensic DNA Extraction Kit, Applied Biosystems™, Catalog number 4463352, UK. This Kit has been developed for the extraction of DNA from calcified tissues (bone, tooth), as well as cigarette butts, tape lifts, and envelope flaps, thus increasing the potential to obtain probative information from downstream STR analysis. For genotyping, we used the PowerPlex® Fusion 6C System, PROMEGA, Madison, US, Catalog number DC2705. This system allows the amplification of 27 STRs including 25 autosomal loci, DYS391 and Amelogenin.

Allele values for each STR in the two samples were determined using the CE technology. Amplicons are separated by the 3,500 Genetic Analyzer, Applied Biosystems, UK, and analyzed using the Gene Mapper® ID version 3.2 software.

Statistical and bioinformatic analyses

For the Tunisian sample, statistical analysis was performed using the Superior Performing Software System (IBM SPSS version 30.0.0.0 172) for MS-Windows. A Chi-square test for homogeneity was applied to assess allele distribution across Tunisia's three regions (North, Center, and South). This test aimed to assess whether the allele distribution was homogeneous across the regions or whether there were differences between regions. Specifically, we compared the observed allele frequencies in each region, ensuring sample independence, with each individual being assigned to only one region. Before performing the test, we confirmed that the assumption of expected frequencies (greater than 5) was satisfied for each cell in the contingency table, thereby ensuring the validity of the results. The null hypothesis of this test posits that there is no significant difference in the allele frequency distributions across the three regions, while the alternative hypothesis is that there is a significant difference. A p-value of less than 0.05 was considered indicative of statistical significance.

For the expanded dataset (Tunisians and sub-Saharan individuals), we used the STRAF (STR Analysis for Forensics) R package (version 2.1.5) (Gouy and Zieger, 2017) with specific applications to autosomal STR data to calculate common forensic parameters. Forensic metrics were computed for the Tunisian and sub-Saharan African populations, as well as the combined cohort.

The data included a matrix of 22 STR markers. The Amelogenin was excluded as non-informative and missing values were addressed using the Simple-Imputer function from the scikit-learn Python library (version 1.5), with the imputation strategy set to the median number of repeats for each marker. To ensure that all features contributed equally to the subsequent analysis, the data were standardized to have a mean of 0 and a standard deviation of 1 using the 'StandardScaler' function. Principal Component Analysis (PCA) was then applied to reduce dimensionality while retaining maximum variance. Four principal components were computed, representing orthogonal directions of the highest variance in the data.

A logistic regression model was constructed to predict geographic group membership (Tunisian or sub-Saharan) based on the imputed matrix of the 22 STR markers. The model was trained using the 'caret' package (version 6.0.94) in R (version 4.3.3), specifying a binomial family with the generalized linear model method. This method models the probability that an individual belongs to one of the two geographical groups based on the allele frequencies of the STR markers. The dataset was split into training and test sets, with 80% of the data used for training and 20% for testing. Cross-validation was performed using 10-fold stratified cross-validation, ensuring that class probabilities were computed, and performance was evaluated based on the Area Under the Receiver Operating Characteristic curve (AUC). The 'glm.control' parameter was adjusted to allow for a maximum of 10,000 iterations to ensure model convergence. Model performance was assessed on the test set by generating predictions and class probabilities. A confusion matrix was used to evaluate classification accuracy, and the AUC was computed using the 'pROC' package (version 1.18.5) to quantify the model's discriminative ability. The final AUC score was reported along with other performance metrics derived from the confusion matrix. A random forest model was additionally

built to evaluate the discriminative power of the 22 STR markers, measured by the calculation of the decrease in Gini impurity, using the random forest package (version 4.7.1.1) in R. The number of trees was set to 500 (ntree = 500). For each split in the trees, 6 predictor variables were randomly selected and evaluated for the best split (mtry = 6).

Results

All STR genotypes are provided in the [Supplementary Sheet 2](#).

Regional analysis in the Tunisian population

The calculation of allelic frequencies, based on relative fluorescence units (RFU) and statistical analysis, demonstrated variable allelic distributions and frequencies across the three Tunisian regions (North, Center, and South) when comparing the 22 STR markers (21 autosomal STR + Y-STR). This regional analysis revealed considerable allelic diversity, with a total of 236 distinct alleles identified and frequencies ranging from 0.005 to 0.051. Among these, 31 rare alleles (frequency <0.005) were observed, with the SE33 locus exhibiting particularly high polymorphism, featuring 35 alleles, 10 of which were classified as rare. In contrast, the TH01 and CSF1PO loci showed the least polymorphism, with only six alleles each (Figure 1; Table 1). The diversity results for the Y-STR marker DYS391 were not generated by the GeneMapper® ID version 3.2 software, unlike the other 21 bi-allelic markers, as the software assumes that a monoallelic marker has limited power to assess genetic diversity. In addition, our study elucidates unique regional allelic patterns, particularly at the D1S1656 locus in Southern Tunisia.

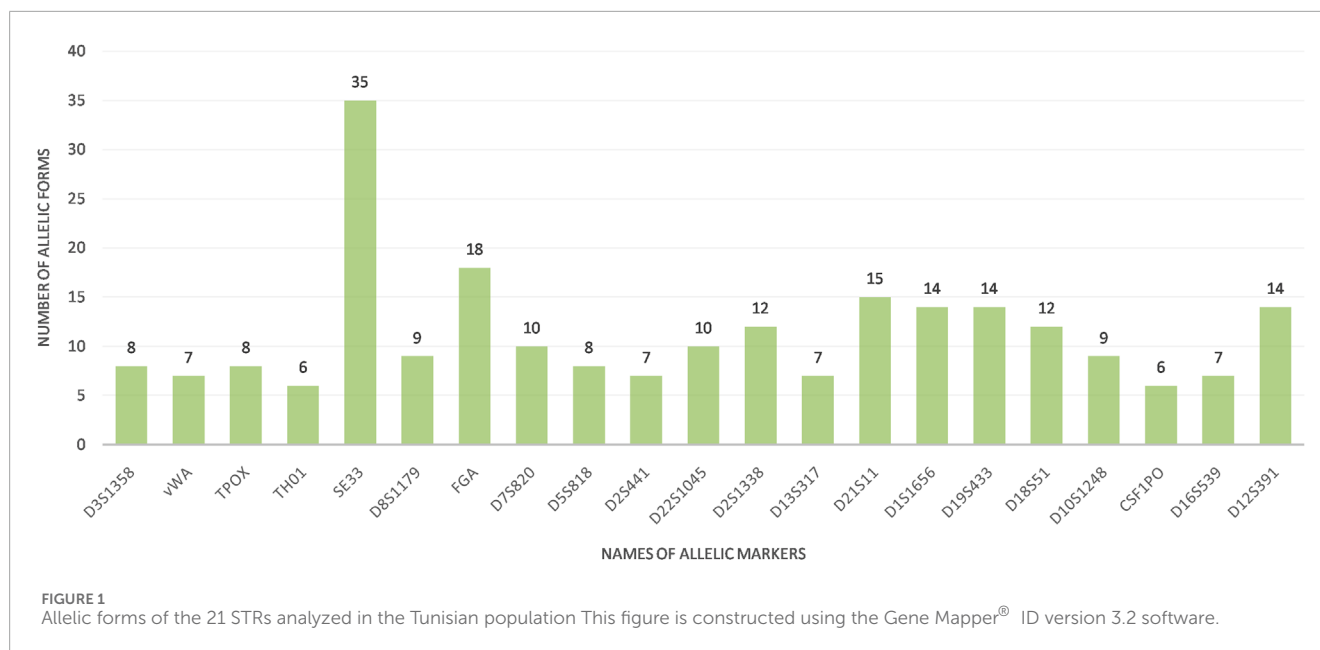
Further analysis of STR markers revealed differential allelic frequency distribution across Tunisia's regions for specific markers, such as DYS391 (Figure 2), underscoring the regional genetic variability.

The analysis highlighted high genetic diversity (GD), supported by elevated GD index values. Observed heterozygosity (Hobs) ranged from 0.625 to 0.938, while expected heterozygosity (Hexp) values ranged from 0.700 to 0.953 for TPOX and SE33 respectively (Table 2), indicating variations possibly due to inbreeding or migration within the population.

The polymorphism information content (PIC) values for each analyzed locus exceeded 0.6, with values ranging from 0.659 to 0.945 for TPOX and SE33 respectively, indicating the high polymorphism of these loci and their significant contribution to genetic variation in the Tunisian population.

The typical paternity index (TPI), which measures the STR marker's ability to establish paternity, ranged from 1.333 to 8.000 for TPOX and SE33 respectively, demonstrating the markers' strong utility in confirming biological relationships in paternity investigations within the Tunisian population.

The power of discrimination (PD), assessing a marker's ability to distinguish between individuals, showed values ranging from 0.859 to 0.986 for TPOX and SE33 respectively, confirming that all STRs studied possess robust discriminative capabilities within the Tunisian population.



Tunisian vs. sub-Saharan population analysis

The relationships between individuals on the PCA projection tend to reflect their genetic relatedness. The closer individuals are on the PCA projection, the more genetically related they tend to be. In this work, PCA showed clear separation between Tunisian and sub-Saharan African populations when examining the first two principal components (PC1 vs. PC2), with a fixation index (F_{ST}) value of 0.0246, indicating significant genetic differentiation (Figure 3). However, differentiation diminished in higher-dimensional subspaces beyond the third principal component (Supplementary Sheet 1), suggesting that most of the variance between these populations was captured in the initial components. Some data points displayed overlap in the PC1 vs. PC2 plot, potentially due to individual variation, shared genetic traits, or errors in population group assignment during data collection.

The logistic regression model showed excellent predictive performance in classifying geographic group membership between Tunisian and sub-Saharan African populations. The confusion matrix revealed that the model correctly predicted 94 out of 96 sub-Saharan African (SSA) individuals and 96 out of 102 Tunisian (TUN) individuals. This resulted in an overall accuracy of 95.96% (95% CI: 92.19%–98.24%), significantly better than the No Information Rate (NIR) of 50.51%, with a P-value of $<2 \times 10^{-16}$ indicating strong statistical significance. The Kappa statistic of 0.92 indicates a high level of agreement between the predicted and actual classifications. The McNemar's test yielded a P-value of 0.3, suggesting no significant difference in the misclassification rates between the two classes. The recall of the model, which measures the model's ability to correctly identify sub-Saharan African individuals, was 94%, while specificity, indicating the model's accuracy in identifying Tunisian individuals, was 97.96%. The positive predictive value (PPV) was 97.92%, and the negative predictive value (NPV) was 94.12%. The prevalence of sub-Saharan African individuals

in the dataset was 50.51%, with a detection rate of 47.47% and a detection prevalence of 48.48%. The balanced accuracy (F1 score), which accounts for both sensitivity and specificity, was calculated to be 95.98%. The model achieved an AUC score of 0.96, further validating its strong discriminative ability between the two geographic groups. These results indicate that the logistic regression model is highly effective in predicting geographic group membership based on the selected STR markers.

The random forest model which is used to measure the homogeneity of the groups formed during the data classification process, further identified the mean decrease in Gini values for different markers (Table 3). Notably, the DYS391 marker exhibited the highest values (140.85 and 149.00), underscoring its robustness in distinguishing between Tunisians and sub-Saharan Africans. This suggests that this marker facilitates the creation of more homogeneous genetic subgroups, thereby making distinctions between populations clearer and more reliable. This property is particularly advantageous in population genetics studies, where the ability to accurately distinguish subpopulations is crucial. Other markers, such as CSF1PO-A1 (8.94) and D18S51-A1 (12.86), also demonstrated relevance, although less than DYS391.

Discussion

In this study, we collected a sample size of 500 individuals. Although this sample size is sufficient for genetic studies, it represents only a small portion of the Tunisian population and does not capture its full genetic diversity, as it consists solely of individuals involved in legal cases. However, there are few, if any, studies indicating that STRs used in forensic science are influenced by socio-economic or criminal factors; rather, they are generally considered to be selectively neutral. As non-coding regions of DNA, they are not subject to the same selective pressures as coding genes. The variability of STRs is primarily due to random mutations, which

TABLE 1 Distribution of the 22 STR allelic forms in North, Center and South Tunisia. N: sample size.

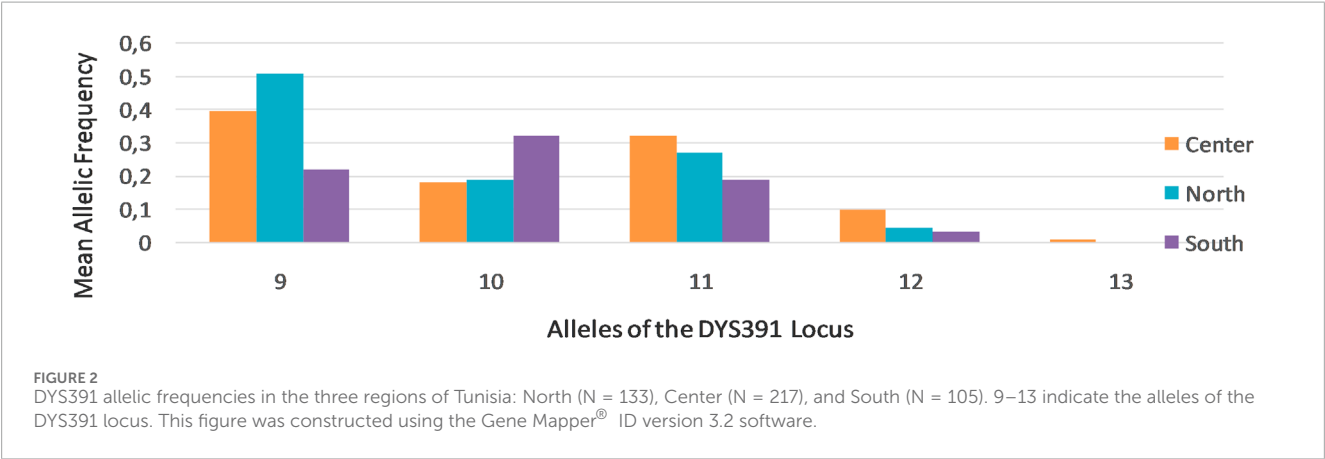
Marker	Total N = 500		North N = 133		South N = 150		Center N = 217		p value
	Allele	Frequency	Allele	Frequency	Allele	Frequency	Allele	Frequency	
TH01	9	29.3	9	32.3	9	26.3	9	29.5	0.13
	6	20.6	7	24.8	6	21	6	20.5	
	7	20.4	6	20.3	7	20	7	18.5	
D3S1358	16	27.5	15	29.7	17	31	16	30.2	0.16
	17	27	17	28.6	16	27.3	15	27	
	15	26.2	16	23.3	15	22	16	23.3	
vWA	17	26.5	17	31.2	17	28.7	16	28.8	0.06
	16	25	16	21.8	16	22.3	17	22.1	
	15	15	15	16.5	15	14.7	18	14.5	
D21S11	29	23.8	30	23.7	29	26	29	24.7	0.48
	30	23.8	29	19.9	30	23.3	30	24.2	
	32.2	12.5	32.2	13.9	32.2	13.7	28	14.7	
TPOX	8	47.9	8	48.1	8	45.7	8	49.3	0.42
	11	21.5	11	21.8	11	21.3	11	21.4	
	9	17.9	9	16.5	9	21	9	16.6	
DYS391	9	40.4	9	50	10	45.8	9	39.7	<0.001 ^a
	11	27.2	11	27.7	9	32.6	11	31.9	
	10	26.6	10	18.5	11	20.1	10	18.1	
D1S1656	15	17.1	12	15.4	16	22	15	20.7	<0.001 ^a
	16	16.2	16	14.3	12	15.3	13	15.4	
	12	14.3	15	13.5	15	15	16	13.4	
	13	14.3							
D12S391	18	18.2	17	16.9	18	17.7	18	19.4	0.71
	17	16.4	18	16.9	19	17.3	17	17.3	
	19	14.2	19	13.2	17	14.7	19	12.7	
SE33	18	9	17	10.2	18	10.7	16	8.5	0.27
	17	8.8	19	9	19	10.3	17	8.5	
	19	8.5	18	8.6	17	8	18	8.1	
D10S1248	14	33.3	14	28.2	14	38.7	14	32.7	0.20
	13	21.2	13	23.3	13	22	15	20.7	
	15	20	15	21.8	15	17.3	13	19.4	
D22S1045	15	39.8	15	39.8	15	43.3	15	37.1	0.58
	16	26.7	16	26.7	16	25	16	28.3	
	11	11.2	11	11.3	17	11	11	12	
D19S433	14	24.8	13	24.4	14	26.7	14	24.2	0.10
	13	23	14	23.7	13	26	13	20	
	15	13.7	15	14.7	15	14	12	13.1	
D8S1179	13	22.4	15	23.3	13	25	13	21.9	0.007 ^a
	14	18.9	13	20.3	15	19.3	15	18	
	15	17.1	11	15	12	13.3	14	17.3	
D2S1338	17	31	17	29.5	17	34	17	29.4	0.47
	20	17.2	20	16.3	20	20.4	20	15.3	
	19	12.3	19	13.1	19	8.7	19	13	
D2S441	14	33.3	14	31.6	11	34.3	14	34.8	0.15
	11	32.9	11	31.6	14	32.7	11	32.7	
	10	13.2	10	12	10	13.7	10	13.6	

(Continued on the following page)

TABLE 1 (Continued) Distribution of the 22 STR allelic forms in North, Center and South Tunisia. N: sample size.

Marker	Total N = 500		North N = 133		South N = 150		Center N = 217		pvalue
	Allele	Frequency	Allele	Frequency	Allele	Frequency	Allele	Frequency	
D18S51	14	14.3	17	15	14	18.7	13	16.6	0.25
	16	14.2	12	14.3	16	17.3	15	14.3	
	13	14.1	14	13.9	12	13.3	12	14.3	
FGA	23	15.1	23	15	21	12.4	23	16.1	0.10
	22	12.4	22	13.9	23	13.7	22	12.4	
	21	12.4	21	12.4	22	11	21	9.7	
D16S539	11	32.9	11	33.5	11	28.3	11	35.7	0.60
	12	27	12	25.9	12	25	12	29	
	13	16.5	13	15.8	13	17.7	13	16.1	
CSF1PO	10	31.7	10	36.1	12	35	12	31.8	0.002 ^a
	12	30.7	11	30.1	10	29.7	10	30.4	
	11	28.7	12	24.1	11	25	11	30.4	
D13S317	11	32.9	12	33.8	12	35.7	11	32.9	0.37
	12	32.4	11	32	11	33.7	12	29.3	
	13	10.6	13	9	13	10.7	13	11.5	
D5S818	12	36.2	12	30.5	12	30.5	12	38.7	0.11
	11	21.8	11	27.1	13	27.1	11	21.4	
	13	18	13	18.4	11	18.4	13	17.3	
D7S820	10	31	10	28.5	10	30.2	10	32.9	0.51
	11	24.8	11	28.1	11	25.2	11	22.3	
	12	16.4	12	15.9	12	21.3	12	15.9	

^aSignificant difference between North, Center and South allele frequencies.



makes their frequency independent of social or criminal behavior, and thus making them reliable for forensic analyses.

In the first part of this study, we expanded our analysis of autosomal loci from 21 (Bhambara et al., 2022) to 22 by incorporating the DYS391 as Y-STR. Due to its unique inheritance pattern, transmitted largely unchanged from father to son, mutations in DYS391 represent the primary source of its variation (Cockerton et al., 2012). This stability enhances its forensic applicability by enabling assessments of Y-STR evidence in criminal investigations using rapidly mutating Y-STR (Kayser, 2017), Y- single nucleotide variants (Zandstra et al., 2025) or combined with autosomal STRs using massively parallel sequencing applied in Forensic DNA Phenotyping (Kayser et al., 2023). Alternatively, X-STRs uses are developed to address complex kinship cases (Gusmão et al., 2025).

The primary aim of our study was to conduct a comprehensive statistical analysis of the 22 STR genotypes. We identified the most discriminative allelic forms for each marker across Tunisian regions. Specifically, significant allelic distinctions were noted for markers DYS391, D1S1656, D8S1179, and CSF1PO, with *p*-values <

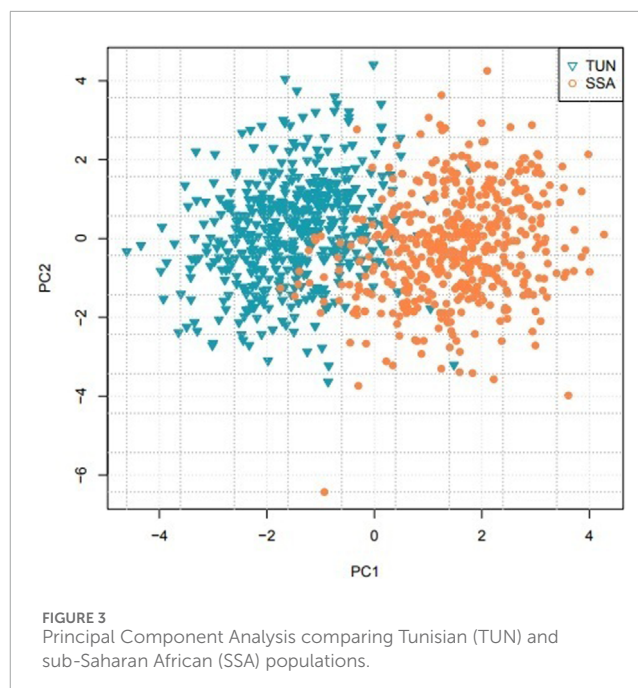
TABLE 2 Genetic diversity (GD) index of the 21 autosomal STR showing a gap between observed (Hobs) and expected (Hexp) heterozygosity.

Marker	GD index	
	Hobs	Hexp
CSFPO	0.677	0.720
D10S1248	0.792	0.794
D12S391	0.813	0.878
D13S317	0.646	0.780
D16S539	0.750	0.754
D18S51	0.865	0.877
D19S433	0.781	0.815
D1S1656	0.906	0.881
D21S11	0.750	0.831
D22S1045	0.656	0.752
D2S1338	0.813	0.851
D2S441	0.760	0.788
D3S1358	0.729	0.782
D5S818	0.667	0.760
D7S820	0.781	0.808
D8S1179	0.781	0.823
FGA	0.833	0.881
SE33	0.938^a	0.953^b
THO1	0.729	0.780
TPOX	0.625^c	0.700^d
VWA	0.729	0.818

^{a,b}The highest GD index for Hobs and Hexp respectively; ^{c,d}The lowest GD index for Hobs and Hexp respectively.

0.05. D1S1656 exhibited pronounced regional variation, especially between the North, South, and Center of Tunisia, with $p < 0.001$.

Our findings generally align with previous studies, though our population sample displays some distinct variations (Brinkmann, 1998; Ruitberg et al., 2001; Abdin et al., 2003; Butler, 2005; El Ossmani et al., 2007). According to Butler, the most polymorphic markers (FGA, D18S51, and D21S11) are distinguished by their complex repeat structures, which contribute to greater individual variability than simpler loci like TPOX and CSF1PO. In line with Butler's observations (Butler, 2005), we also observed high heterozygosity in SE33 and D18S51, while D2S1338 displayed moderate discrimination between Tunisian regions in our dataset.



Ruitberg et al. identified SE33, FGA, D18S51, D21S11, and D2S1338 as highly discriminative markers based on mutation rates, whereas THO1 and TPOX ranked lower (Ruitberg et al., 2001). EL Ossmani et al. confirmed the significant discriminating power of 15 STR markers, with D18S51 being particularly prominent (El Ossmani et al., 2007). Similarly, our findings corroborate the discriminating capacity of D18S51, which revealed distinctive allelic frequencies across Tunisia's North, South, and Center regions.

To enhance the generalizability of future studies and to gain a more nuanced understanding of the potential impact and limitations of selection bias, it is crucial to incorporate samples from diverse regions and a broader spectrum of social groups, particularly non-delinquent individuals. Nevertheless, when compared to other research on Tunisian samples from the general population (e.g., Cherni et al., 2005; Mahfoudh-Lahiani et al., 2006; Mejri et al., 2022), our findings reveal a striking similarity, thereby supporting the validity of our sample as a representative subset of the broader population.

Beyond forensic identification, we leveraged the 22 STR genotypes to examine the genetic divergence between Tunisian and sub-Saharan African populations. Logistic regression analysis demonstrated high efficacy in predicting geographic group affiliation based on these STR markers. PCA revealed a clear separation between Tunisian and sub-Saharan African populations, reflecting significant genetic differentiation. This outcome aligns with established knowledge on African population genetics, where geographic and historical factors, such as the Sahara Desert, have acted as barriers limiting gene flow between North and sub-Saharan Africa. This inter-population variance was predominantly captured in the first two principal components (PC1 and PC2). However, higher dimensions (beyond PC3) revealed diminished discriminatory information, indicating that additional components contribute largely to minor intra-population variations.

TABLE 3 Gini coefficient values of the 20 markers (19 autosomal and 1 Y-STR). D12S391 and D2S1338 are removed because of the lack of informativeness.

Marker	Mean decrease gini	
	A1 allele	A2 allele
CSF1PO	8.939	0.972
D10S1248	1.641	0.946
D13S317	3.070	1.905
D16S539	3.776	1.627
D18S51	12.859	4.455
D19S433	6.189	2.932
D1S1656	3.349	2.099
D21S11	1.327	2.377
D22S1045	1.209	2.233
D2S441	1.189	1.338
D3S1358	1.218	2.159
D5S818	1.150	1.080
D7S820	1.917	0.967
D8S1179	2.911	2.027
DYS391	140.852^a	149.002^b
FGA	2.290	3.773
SE33	2.103	8.359
TH01	2.575	4.172
TPOX	1.134	0.902
vWA	1.133	1.725

^{a,b}Highest values of Gini coefficient for A1 and A2 alleles respectively.

In addition, the observed overlap between some data points on the PC1 vs. PC2 graph can be interpreted in several ways. One possible hypothesis is that these overlaps result from historical or contemporary genetic mixing between populations north and south of the Sahara. Indeed, several genetic studies have shown traces of gene flow between these regions, probably related to ancient migrations (especially *via* trans-Saharan trade routes) or to more recent population movements. Another explanation could be normal individual variation within populations, where some individuals share common genetic traits with those from other groups due to genetic heterogeneity specific to each population. Finally, it is also possible that these overlaps are partly due to errors in the classification of individuals during data collection. Further investigation of the selection criteria and sampling methods may help clarify these overlaps.

The logistic regression model demonstrated strong predictive capability, achieving an overall accuracy of 95.96%, significantly surpassing the NIR. With high sensitivity and specificity (94% for sub-Saharan identification and 97.96% for Tunisian identification), the model highlighted the discriminative strength of the selected markers for these populations.

The random forest model identified the most informative STR markers for geographic classification. DYS391 emerged as particularly influential, contributing significantly to the predictive power of the model. Markers such as CSF1PO-A1 and D18S51-A1, while still relevant, had lower importance scores, indicating they may capture more intra-population rather than inter-population variation. These findings demonstrate that effective population differentiation is contingent upon marker selection, with certain loci being inherently more informative in distinguishing groups.

This study offers valuable insights into the genetic differentiation between Tunisian and sub-Saharan populations, underscoring the utility of multivariate methods (PCA) and advanced classification techniques (LRM, RF). However, the observed overlaps indicate possible limitations, suggesting the existence of genetic complexities unaccounted for by our models, potentially attributable to historical gene flow or sampling biases.

Our results align with prior studies on Tunisian genetic diversity, which also highlight the substantial variation within this population. Indeed, during successive historical periods, Tunisia, by its strategic position in the extreme North of Africa has been a crossroads of multiple civilizations and their corresponding key population movements. Throughout its history, many people arrived and settled in Tunisia among the Berbers (Brett and Fentress, 1997). Tunisia, by its standing mid-way between the Eastern and Western Mediterranean, played a major role as a route for historical migrations. This resulted in the present Tunisian population being a mixture of multiple origins. Many studies used autosomal and X-STR loci to describe the Tunisian genetic heterogeneity (Frigi et al., 2014; Messoussi et al., 2019; Mejri et al., 2022). These results, added to the knowledge of the migration and occupation routes that occurred in the past all around the Mediterranean, give us a great opportunity to reconstruct the migration patterns. Indeed, previous studies provided genetic information relating to the mixed origin of the Tunisian population using Alu/STR markers (El Moncer et al., 2010), which revealed a sub-Saharan component probably due to sub-Saharan historical migrations as shown by previous analysis on the beta globin gene mutations (Bloom, 1995; Bennani et al., 1994).

A migration pattern similar to that of Tunisia has been observed in Libya, the Tunisia's eastern neighboring, which was first inhabited, like Tunisia, by Berbers and then colonized by a variety of ethnic groups including Phoenicians, Greeks, Romans, Arabs and, more recently, Italians. A study, carried out on 175 Libyan males using haplotypes of 22 Y-chromosome-specific SNPs, revealed a predominant Northwest African component (signature of Berber speaking people, the autochthonous inhabitants) followed by one of a Middle Eastern origin (migration from Arabic populations). Overall, the comparative study with other populations (~5,400 individuals from North Africa, Middle East, Sub-Saharan Africa, and Europe) revealed a general genetic homogeneity among North African populations (Triki-Fendri et al., 2015).

Other studies based on comparisons with sub-Saharan populations underscore Tunisians' distinct genetic characteristics,

with significant allelic differences at loci such as DYS391 and D18S51 (Caglià et al., 2003; Frigi et al., 2014). On a large scale, using sequence haplotypes and cutting edge statistical machinery, previous data on the Eurasian populations had shown that North and West African ancestry had entered Southern Europe, suggesting a key role for the Mediterranean in supporting gene flow back into Europe, consistent with migrations associated with the Arabic Conquest of the Iberian peninsula and earlier movements in and around Italy (Busby et al., 2015).

Finally, alongside the development of bioinformatic tools (Wang et al., 2022), we are advancing Next-Generation Sequencing (NGS) techniques as documented in recent literature (Guo et al., 2017; Butler, 2023). These technologies should permit expanding the sample size and incorporating additional markers, especially SNPs and mitochondrial. Employing advanced genomic technologies would further improve genetic resolution. Additionally, we aim to establish a region-specific STR database to enhance investigations of challenging cases, an approach already adopted in various populations (Ge et al., 2013; Caputo et al., 2023).

Currently, no public or centralized STR database exists for the Tunisian population. While several studies have provided valuable insights into the genetic diversity of Tunisia and neighboring countries, offering data on allele frequencies for STR markers, these data remain fragmented and are not compiled into a unified, accessible database. Our research is specifically designed to address this critical gap by compiling and analyzing an extensive STR dataset from the Tunisian population. Through the publication of these findings and, if possible, the integration of the data into a publicly accessible database, we aim to provide a reliable and invaluable resource for forensic practitioners. This effort will not only enhance the precision of forensic genetic profiling but will also significantly improve the interpretation of genetic data in forensic investigations, particularly within the Tunisian and wider Maghreb context.

Furthermore, integrating genetic, historical, demographic and anthropological data would provide a more complete picture of the genetic landscape, enhancing forensic capabilities and contributing valuable insights into human migration patterns across the Sahara and beyond.

Conclusion

Overall, the combination of allelic frequency distribution, PCA, logistic regression, and random forest analyses provides a comprehensive view of the genetic diversity within the Tunisian population and its distinction from sub-Saharan African groups. The significant inter-population F_{ST} value, high discriminatory power of individual markers notably DYS391, and substantial TPI values indicate that the chosen STR markers effectively capture both intra- and inter-population genetic variation. This is particularly valuable for applications in forensic analysis and paternity testing within Tunisia and possibly across North Africa, where genetic markers can serve as reliable tools for identifying individual ancestry and familial relationships and predicting geographical origins.

The use of multiple statistical methods in this study not only reinforces the robustness of findings but also highlights the influence of specific loci on genetic diversity and population structure.

The study also highlights the need for comprehensive, population-specific databases to ensure accurate forensic identifications. We think that our study has direct forensic applications, particularly in the context of unidentified remains of clandestine migrants recovered along Tunisia's coast. Many of these bodies, found in an advanced state of decomposition, make traditional identification methods impossible, as visual features are no longer distinguishable.

Our study offers a valuable tool in these situations. By establishing a genetic profile of the Tunisian population with 22 STR markers, we provide a means to identify the origin of remains, even when they are severely decomposed. This reference database could significantly aid forensic investigations, allowing for the differentiation between Tunisian and sub-Saharan populations, among others, in cases where visual identification is not possible. In light of the challenges posed by clandestine immigration, the development of such a database would enhance the accuracy of forensic identification, offering a more reliable method to identify victims and provide closure to families, while also assisting in legal and administrative processes.

Overall, a multidisciplinary approach integrating demographic, historical, anthropological, geographical and environmental data would provide deeper contextualization, enriching our understanding of human migration and genetic diversity.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Material](#), further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by Ethics committee of the Higher Institute of Biotechnology of Monastir. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because the sample consists of either corpses of deceased individuals or a large number of individuals with a criminal record, the majority of whom are no longer contactable. The study was approved by the responsible authorities at the Center of technical and scientific police in Tunisia.

Author contributions

AA: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Writing – original draft. HF: Conceptualization, Formal Analysis, Software, Writing – original draft. HO: Formal Analysis, Software, Validation, Writing – original draft. HB: Project administration, Supervision, Writing – review and editing. AH: Conceptualization, Data curation, Formal Analysis, Project administration, Software, Supervision, Validation, Visualization, Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

The authors would like to acknowledge ABS society, the representative of QIAGEN Company in Tunisia for technical assistance and Professor Abdelwaheb OMRI from the School of Natural Sciences, Laurentian University, Canada for reviewing and editing the text.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

References

- Abdin, L., Shimada, I., Brinkmann, B., and Hohoff, C. (2003). Analysis of 15 short tandem repeats reveals significant differences between the Arabian populations from Morocco and Syria. *Leg. Med. Tokyo, Jpn. 5* (Suppl. 1), S150–S155. doi:10.1016/s1344-6223(02)00129-3
- Barni, F., Ralf, A., Della Rocca, C., Cannistrà, F., Gigliucci, M., Trombetta, B., et al. (2024). Analysis of rapidly mutating Y-STRs enables almost complete discrimination of unrelated and related males from the African continent. *Forensic Sci. Int. Genet.* 73, 103127. doi:10.1016/j.fsigen.2024.103127
- Bennani, C., Bouhass, R., Perrin-Pecontal, P., Tamouza, R., Malou, M., Elion, J., et al. (1994). Anthropological approach to the heterogeneity of beta-thalassemia mutations in northern Africa. *Hum. Biol.* 66 (3), 369–382.
- Bhambara, A. K., Singh, A., Sahajpal, V., Thakur, M., Bhandari, D., Sharma, S., et al. (2022). Evaluation of genetic polymorphisms at 21 autosomal STR loci in Ramgharia Sikh population of Punjab, India. *Ann. Hum. Biol.* 49 (5–6), 263–268. doi:10.1080/03014460.2022.2126888
- Bloom, M. (1995). *Understanding sickle cell disease (understanding health and sickness series)*. Jackson: University Press of Mississippi.
- Brett, M., and Fentress, E. (1997). *The Berbers (repr)*. Oxford: Blackwell.
- Brinkmann, B. (1998). Overview of PCR-based systems in identity testing. *Methods Mol. Biol. Clift. N.J.* 98, 105–119. doi:10.1385/0-89603-443-7:105
- Busby, G. B. J., Hellenthal, G., Montinaro, F., Tofanelli, S., Bulayeva, K., Rudan, I., et al. (2015). The role of recent admixture in forming the contemporary west Eurasian genomic landscape. *Curr. Biol. CB* 25 (19), 2518–2526. doi:10.1016/j.cub.2015.08.007
- Butler, J. M. (2005). Constructing STR multiplex assays. *Methods Mol. Biol. Clift. N.J.* 297, 053–066. doi:10.1385/1-59259-867-6:053
- Butler, J. M. (2023). Recent advances in forensic biology and forensic DNA typing: INTERPOL review 2019–2022. *Forensic Sci. Int.* 6 (Synergy), 100311. doi:10.1016/j.fsism.2022.100311
- Caglià, A., Tofanelli, S., Coia, V., Boschi, I., Pescarmona, M., Spedini, G., et al. (2003). A study of Y-chromosome microsatellite variation in sub-saharan Africa: a comparison between F st and R st genetic distances. *Hum. Biol.* 75 (3), 313–330. doi:10.1353/hub.2003.0041
- Caputo, M., Sala, A., and Corach, D. (2023). Reference population database for 24 STR loci of selected provinces of Argentina. *Mol. Biol. Rep.* 50 (7), 6287–6291. doi:10.1007/s11033-023-08497-z
- Cherni, L., Loueslati Yaacoubi, B., Pereira, L., Alves, C., Khodjet-El-Khil, H., Ben Ammar El Gaaied, A., et al. (2005). Data for 15 autosomal STR markers (powerplex 16 system) from two Tunisian populations: kesra (Berber) and zriba (Arab). *Forensic Sci. Int.* 147 (1), 101–106. doi:10.1016/j.forsci.2004.04.009
- Cockerton, S., McManus, K., and Buckleton, J. (2012). Interpreting lineage markers in view of subpopulation effects. *Forensic Sci. Int. Genet.* 6 (3), 393–397. doi:10.1016/j.fsigen.2011.04.020
- Ding, Q., Hu, Y., Koren, A., and Clark, A. G. (2021). Mutation rate variability across human Y-chromosome haplogroups. *Mol. Biol. Evol.* 38 (3), 1000–1005. doi:10.1093/molbev/msaa268
- El Moncer, W., Esteban, E., Bahri, R., Gayà-Vidal, M., Carreras-Torres, R., Athanasiadis, G., et al. (2010). Mixed origin of the current Tunisian population from the analysis of Alu and Alu/STR compound systems. *J. Hum. Genet.* 55 (12), 827–833. doi:10.1038/jhg.2010.120
- El Ossmani, H., Bouchrif, B., Talbi, J., El Amri, H., and Chafik, A. (2007). La diversité génétique de 15 STR chez la population arabophone de Rabat-Salé-Zemmour-Zaer. *Antropo* 15, 55–62. Available online at: <https://www.didac.edu.es/antropo>.
- Freeman, B., Smith, N., Curtis, C., Hockett, L., Mill, J., and Craig, I. W. (2003). DNA from buccal swabs recruited by mail: evaluation of storage effects on long-term stability and suitability for multiplex polymerase chain reaction genotyping. *Behav. Genet.* 33 (1), 67–72. doi:10.1023/a:1021055617738
- Frigi, S., El Gaaied, A. B. A., and Cherni, L. (2014). An investigation of the genetic diversity of the Kerkennah islands and Mahdia (Tunisia) using biparental markers. *Ann. Hum. Biol.* 41 (1), 53–60. doi:10.3109/03014460.2013.824025
- Ge, J.-Y., Yan, J.-W., Xie, Q., Sun, H.-Y., Zhou, H.-G., and Li, B. (2013). Development of Chinese forensic Y-STR DNA database. *Fa Yi Xue Za Zhi* 29 (3), 212–215.
- Gouy, A., and Zieger, M. (2017). STRAF-A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci. Int. Genet.* 30, 148–151. doi:10.1016/j.fsigen.2017.07.007
- Guo, F., Yu, J., Zhang, L., and Li, J. (2017). Massively parallel sequencing of forensic STRs and SNPs using the Illumina® ForenSeq™ DNA signature prep kit on the MiSeq FGx™ forensic genomics system. *Forensic Sci. Int. Genet.* 31, 135–148. doi:10.1016/j.fsigen.2017.09.003

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fbinf.2025.1550730/full#supplementary-material>

SUPPLEMENTARY FIGURE 1 |

Principal Component Analysis comparing Tunisian (TUN) and sub-Saharan African (SSA) populations beyond the third principal component: PC3 vs PC4. TUN: Tunisians; SSA: sub-Saharan Africans.

SUPPLEMENTARY TABLE 1 |

STR Genotypes of all samples imputed. TUN: Tunisians, NE: North-East; NO: North-West; CE: Center-East; CO: Center-West; SO: South-West; SSA: Sub-Saharan Africans, N: Number; Amel: Amelogenine marker.

- Gusmão, L., Antão-Sousa, S., Faustino, M., Abovich, M. A., Aguirre, D., Alghafri, R., et al. (2025). X-chromosomal STRs: metapopulations and mutation rates. *Forensic Sci. Int. Genet.* 76, 103232. doi:10.1016/j.fsigen.2025.103232
- Haddrill, P. R. (2021). Developments in forensic DNA analysis. *Emerg. Top. Life Sci.* 5 (3), 381–393. doi:10.1042/ETLS20200304
- Hares, D. R. (2015). Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 17, 33–34. doi:10.1016/j.fsigen.2015.03.006
- Holt, C., Walichiewicz, P., Eagles, J., Daou, A., Didier, M., Edwards, C., et al. (2019). Mitochondrial DNA data analysis strategies that inform MPS-based forensic casework implementation. *Forensic Sci. Int. Genet. Suppl. Ser.* 7 (1), 389–391. doi:10.1016/j.fsigss.2019.10.024
- Kaushik, S., and Sahajpal, V. (2020). Capillary electrophoresis issues in forensic DNA typing. In *Forensic DNA typing: principles, Applications and advancements*. Editor P. Shrivastava, H. R. Dash, J. A. Lorente, and J. Imam Springer, 223–238. doi:10.1007/978-981-15-6655-4_11
- Kayser, M. (2017). Forensic use of Y-chromosome DNA: a general overview. *Hum. Genet.* 136 (5), 621–635. doi:10.1007/s00439-017-1776-9
- Kayser, M., Branicki, W., Parson, W., and Phillips, C. (2023). Recent advances in Forensic DNA Phenotyping of appearance, ancestry and age. *Forensic Sci. Int. Genet.* 65, 102870. doi:10.1016/j.fsigen.2023.102870
- Kayser, M., and Parson, W. (2017). Transitioning from forensic genetics to forensic genomics. *Genes* 9 (1), 3. doi:10.3390/genes9010003
- Kostiuchenko, O., Anna, V., Yurii, S., Ivo, S., and Lilia, P. (1997). The impact of DNA databases on the investigation of crimes. Available online at: <https://www.pjcriminology.com/publications/the-impact-of-dna-databases-on-the-investigation-of-crimes/>.
- Mahfoudh-Lahiani, N., Rebaï, A., and Makni, H. (2006). Data for 10 autosomal STR markers in South Tunisian population. *Forensic Sci. Int.* 164 (2-3), 254–256. doi:10.1016/j.forsciint.2005.10.018
- McCord, B., and Lee, S. B. (2018). Novel applications of massively parallel sequencing (MPS) in forensic analysis. *Electrophoresis* 39 (21), 2639–2641. doi:10.1002/elps.201870175
- Mejri, A., Baeta, M., Al-Haj-Taib, R., Herrera, R. J., Benammar-Elgaaied, A., de Pancorbo, M. M., et al. (2022). Genetic polymorphisms of 17 X-STR loci in two Tunisian populations from Sousse and Makthar. *Ann. Hum. Biol.* 49 (7-8), 342–347. doi:10.1080/03014460.2022.2136755
- Messoussi, M., Prieto-Fernández, E., Baeta, M., Núñez, C., Gaaied, A. B. A.-E., de Pancorbo, M. M., et al. (2019). Genetic variation of 17 X-chromosome STR loci in Tunisian population of Nabeul. *Int. J. Leg. Med.* 133 (1), 85–88. doi:10.1007/s00414-018-1827-3
- Milne, E., van Bockxmeer, F. M., Robertson, L., Brisbane, J. M., Ashton, L. J., Scott, R. J., et al. (2006). Buccal DNA collection: comparison of buccal swabs with FTA cards. *Cancer Epidemiol.* 15 (4), 816–819. doi:10.1158/1055-9965.EPI-05-0753
- Ruitberg, C. M., Reeder, D. J., and Butler, J. M. (2001). STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res.* 29 (1), 320–322. doi:10.1093/nar/29.1.320
- Triki-Fendri, S., Sánchez-Diz, P., Rey-González, D., Ayadi, I., Carracedo, Á., and Rebaï, A. (2015). Paternal lineages in Libya inferred from Y-chromosome haplogroups. *Am. J. Phys. Anthropol.* 157 (2), 242–251. doi:10.1002/ajpa.22705
- Wang, X., Budowle, B., and Ge, J. (2022). USAT: a bioinformatic toolkit to facilitate interpretation and comparative visualization of tandem repeat sequences. *BMC Bioinforma.* 23 (1), 497. doi:10.1186/s12859-022-05021-1
- Yagasaki, K., Mabuchi, A., Higashino, T., Hao Wong, J., Nishida, N., Fujimoto, A., et al. (2022). Practical forensic use of kinship determination using high-density SNP profiling based on a microarray platform, focusing on low-quantity DNA. *Forensic Sci. Int. Genet.* 61, 102752. doi:10.1016/j.fsigen.2022.102752
- Zandstra, D., Ralf, A., Ozgur, Z., van Ijcken, W. F. J., Ghanbari, M., and Kayser, M. (2025). Unprecedented male relative differentiation with Y-SNVs from whole genome sequencing. *Forensic Sci. Int. Genet.* 78, 103265. doi:10.1016/j.fsigen.2025.103265