International Journal of Legal Medicine (2026) 140:1313–1325
https://doi.org/10.1007/s00414-025-03716-7

ORIGINAL ARTICLE

Identification of biogeographically informative microssatelite markers for Brazilian *Cannabis sativa* samples: a machine learning approach for forensic origin prediction

Cássio Augusto Rodrigues Bettim[1] · Lucas de Oliveira Pereira Ribeiro[2] · Oscar Victor Cardenas Alegría[5] · Eduardo Filipe Avila Silva[2] · Franciele Maboni Siqueira[4] · Flávio Anastácio de Oliveira Camargo[7] · Clarice Sampaio Alho[2,6] · Márcio Dorn[1,2,3,4]

Received: 24 May 2025 / Accepted: 30 December 2025 / Published online: 6 February 2026

Abstract

In Brazil, *Cannabis sativa* is the most commonly trafficked illicit drug, with supply chains originating from both domestic and international sources, often involving sophisticated trafficking methods. In response to the Brazilian Federal Police (BFP) eradication efforts targeting large-scale plantations, alternative trafficking strategies have emerged, such as dispatching *C. sativa* seeds via postal services for indoor cultivation. In contemporary forensic genetics, microsatellites remain the gold standard for identification and have gained recognition as valuable tracking tools in investigative contexts. However, currently available microsatellite panels are not optimally informative for Brazilian samples. This study aimed to identify and evaluate highly informative microsatellites for origin tracking, using genomic data from 38 *C. sativa* samples sourced from Paraguay (PG), Colombia (CO), the Marijuana Polygon (MP) region, and a foreign group (FG) of commercial strains that constitute a substantial portion of national trafficking. Various combinations of Machine Learning (ML) Feature Selection (FS) techniques and classification models were used. To this end, diverse metrics for evaluating the performance of classification models were considered alongside the execution of principal coordinate analyzes (PCoA) and hierarchical clustering. Generally, Support Vector Classifiers (SVC) combined with different FS strategies demonstrated superior performance in predicting sample origin, achieving perfect classification of all samples using only 9 SSRs. These markers also effectively differentiated the four groups of seizures based on their biogeographic origin in PCoA and their hierarchical clustering. This study successfully identified informative SSR markers for distinguishing *C. sativa* samples associated with significant trafficking routes. Further research involving new and larger sample groups can be important to validate their application in practical investigative settings.

Keywords *Cannabis sativa* · Simple sequence repeats · Machine learning · Feature selection · Origin tracking

✉ Márcio Dorn
mdorn@inf.ufrgs.br

[1] Laboratory of Structural Bioinformatics and Computational Biology, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

[2] National Institute of Forensic Science and Technology, Porto Alegre, RS, Brazil

[3] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

[4] Center for Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

[5] Institute of Biological Sciences, Federal University of Pará, Belém, PA, Brazil

[6] Pathology Post Graduation Program, Health Sciences Federal University of Porto Alegre, Porto Alegre, RS, Brazil

[7] Faculty of Agronomy, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Introduction

Cannabis sativa L. is a dicotyledonous plant belonging to the Cannabaceae family. It is considered a socially and economically important crop, as it is increasingly cultivated and used in medicinal and recreational activities. This plant can be divided into the subspecies or varieties indica, sativa, and ruderalis, designations based on morphological characteristics (i.e., C. indica, shorter with a woody stem and *C. sativa*, taller with a fibrous stem) [19]. It can also be legally divided between hemp varieties and recreational drugs, popularly known as marijuana or weed, based on their chemical composition. It has a complex chemical profile with more than 500 constituents, including more than 100 identified phytocannabinoids, the most relevant being delta-9-tetrahydrocannabinol (THC) and cannabidiol (CBD) [32, 34].

In Brazil, the cannabis-type drug remains the most widely consumed illicit drug [28]. Approximately 30% of the country's illegal market is supplied by cultivation on the national territory originating from the marijuana polygon, a region made up of 13 cities in the states of Pernambuco and Bahia in the northeast and the middle and sub-middle hinterland regions of the São Francisco river, which supplies the North and Northeast regions of the country. Meanwhile, the remaining 70% is mainly supplied by international trafficking coming from cultivation areas spread across the Brazil-Paraguay border, especially Mato Grosso do Sul state, which supplies the South, South-West and Central-West regions [15, 36]. In addition to these two known routes, due to dedicated efforts of the Brazilian Federal Police (BFP) in eradicating large-scale plantations, alternative forms of trafficking, such as sending *C. sativa* seeds by post logistics, have emerged in recent years [33, 40]. The seeds are imported from neighboring countries such as Uruguay. However, they are mainly from the European continent, an important international supplier, and are intended for indoor cultivation on a smaller scale, which is more difficult for police forces to locate.

Different forensic methodologies for differentiating *C. sativa* samples about their biogeographical origin have been suggested in recent years and using different types of techniques, such as DNA barcoding, amplification of genetic markers in hotspot regions such as SNPs and microsatellites, and multi-element concentration in plants and soil [6, 11, 15, 30, 35]. Although some of these methodologies have been successful in individualizing and tracking the origin of the plant's cultivation, some of their limitations include: 1) the panels of genetic variants used in these studies were developed from foreign samples - in most cases, cultivated in Europe or North America - and do not present the most informative loci for identifying and classifying *C. sativa* that are present in the Brazilian illegal market, requiring a more significant number of markers to be needed for their differentiation, thus making the process more costly and demanding greater labor effort; 2) certain types of materials, such as seeds, young plants or more degraded samples are not suitable for chemical analysis [30, 35].

Microsatellites, also known as short tandem repeats (STR) or simple sequence repeats (SSR), serve as the cornerstone of contemporary forensic genetics, recognized as the gold standard for human identification [12, 22]. Their utility extends beyond forensics, as they are also widely employed in studies of genetic diversity and population structure in plants [42]. These markers are ubiquitously distributed across prokaryotic and eukaryotic genomes, with a particular prevalence within eukaryote euchromatin, encompassing both coding and non-coding regions. Building upon the established applications of microsatellites, recent research has demonstrated the efficacy of Machine Learning (ML) strategies in predicting various phenotypes and complex traits from multi-omics data across a diverse range of plant species. Specifically, techniques such as feature selection (FS) and the application of supervised classification models have been successfully utilized in soybeans, rice, coffee, grapevines, corn and Arabidopsis thaliana [1, 16, 20, 26, 41, 43, 45]. However, despite these advances, there is a notable gap in applying such methodologies to data derived from *C. sativa* for origin tracking. While recent studies have successfully applied ML and statistical methods to *Cannabis sativa* STR data [8, 9], these works primarily focused on establishing associations between genetic markers and chemical profiles (e.g., THC levels). To our knowledge, no study to date has used microsatellite data and ML specifically to infer the geographical provenance of *C. sativa* samples associated with trafficking routes in Latin America.

Microsatellite markers are powerful tools for elucidating population structure. Consequently, this study used these markers to determine the geographical origin of *Cannabis sativa* genomes associated with Brazilian narcotraffic routes. Specifically, we explore a range of supervised learning strategies, feature selection (FS) methods, and established classification models. We evaluated the efficacy of various combinations of these techniques to achieve a robust and accurate origin determination. This approach allowed for a comprehensive assessment of the potential of SSR markers and ML to trace the provenance of *C. sativa* samples in the context of the distribution of illicit drugs.

Materials and methods

Samples

This study investigated the chemical composition of 38 *C. sativa* samples provided by the Brazilian Federal Police to determine potential variations based on the culture origin. The samples were categorized into four distinct groups: Paraguay (PG), Colombia (CO), Marijuana Polygon (MP), and a Foreign Group (FG) of commercial strains with unspecified origins. PG (n=8) and CO (n=4) samples were derived from seeds extracted from compressed cannabis bricks - comprising dried leaves, flowers, and stems - seized during postal sorting operations between 2019 and 2022. Based on the analysis of imported materials, the investigation extended to domestically sourced samples. Specifically, MP samples (n=12) comprised leaf plant material confiscated during on-site BFP operations from 2015 to 2019. Complementing these domestic samples, the FG (n=14) comprised leaf plant material cultivated from seeds of various foreign commercial strains. These seeds were taken from postal services in 2017 and 2019 and subsequently cultivated by BFP for confirmation tests. This comprehensive sampling strategy allowed for a comparative assessment of *C. sativa* composition in diverse geographical and cultivation contexts. Furthermore, the collection of samples from distinct seizures over varying time periods (2015–2022) ensures temporal independence, minimizing potential batch effects associated with single seizure events.

Extraction and quantification

DNA isolation was performed using the *DNeasy mericon Food Kit* (Qiagen, Valencia, CA) for small-scale samples, and the *DNeasy Plant Mini Kit* (Qiagen, Valencia, CA) for leaf and seed materials. To optimize the protocols for specific sample types and sizes, modifications were implemented. Specifically, the *DNeasy mericon Food Kit* was adapted to use $\leq$20 mg of starting leaf material, and the *DNeasy Plant Mini Kit* protocol was modified to use a single seed per sample. Following isolation, the concentration of the extracted DNA was determined using the *Qubit dsDNA HS Assay Kit* (Invitrogen, Carlsbad, California, USA).

Sequencing

A total of 38 samples were sequenced using the Illumina NovaSeq 6000 platform, with 250 bp paired-end reads. Following sequencing, the demultiplexing of all libraries from each sequencing lane was executed using the Illumina bcl-2fastq v2.20 software. Consequently, 38 FASTQ files were generated, each containing sequencing adapter-clipped reads. These files were further complemented by FastQC v0.11.9 reports, which provided comprehensive read quality metrics.

Alignment analysis

Sequencing yielded up to 2.27×10^7 reads, with an average of 8.3×10^6 reads per sample. In particular, all samples exhibited Phred quality scores that exceeded 30.

For alignment, the Cannabis sativa reference sequence GCF_029168945.1 (v.2.0; GenBank accession number), available from the NCBI database and representing a genome size of 770.3 Mb, was employed. This process was executed using the Bowtie2 program [23] without restrictions imposed. Following alignment, a filtering criterion was applied. Only samples demonstrating an alignment percentage greater than 50% were retained for subsequent analysis. Consequently, initially in the ".sam" format, the resulting files underwent a series of transformations. First, they were sorted and converted into the ".bam" format. Subsequently, these ".bam" sequences were transformed into a consensus sequence in ".fasta" format, which approximates the size of the reference genome. This transformation was achieved using the Samtools program [24].

SSR extraction

The sequences, provided in ".fasta" format, together with the reference sequence GCF_029168945.1, were processed using the Micro and Mini SATellite IdentificatioN (SATIN) program [10] to identify SSRs. In particular, this tool can delineate distinct SSR sequences based on the number of repetitions of each marker at individual loci. Furthermore, a unique feature of SATIN is its ability to preselect a subset of SSRs by performing comparative analyzes between markers and each genome. This enables the identification of the most significantly divergent SSRs in the compared groups, even in the absence of a reference genome. Following SSR identification, all pertinent features were consolidated into a tabular data set within a single comma-separated value (CSV) raw file. This file also incorporated information from the cultivation origin site for each sample, which served as labels. Consequently, this comprehensive data set functioned as input data for subsequent analytical procedures.

Feature Selection

Based on the generated data set, feature selection (FS) was implemented to reduce the number of irrelevant or redundant features. The FS step was integrated into our analytical pipeline (Fig. 1) and executed using four distinct strategies in Python 3.10.12, using the packages numpy v1.26.4,

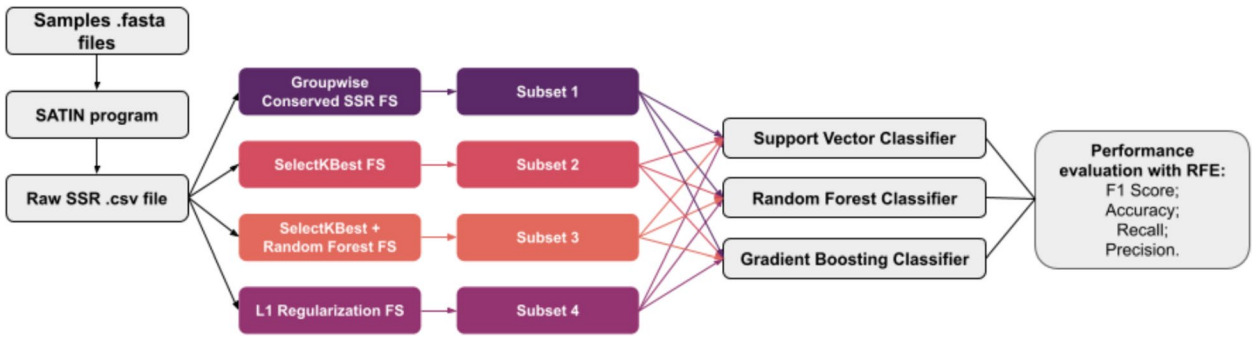

Fig. 1 Flowchart illustrating the pipeline used for SSR extraction, FS, ML analysis and evaluation of SSR markers

pandas v1.5.3, and scikit-learn v1.3.2. The objective was to determine which strategy would produce the optimal subset of SSRs from the raw CSV file for the *C. sativa* sample classification task.

Group-wise conserved SSR (GC-SSR) FS

The initial stage involved straightforward filtering of the total extracted SSRs. This process involved comparing samples, excluding the reference sequence, and eliminating SSR sequences containing "N" nucleotides. Only SSRs in all samples within at least one of the four origin groups (MP, FG, PG, or CO) were retained. This approach deviates from the classical feature selection (FS) methodologies found in the literature, instead representing an adaptation centered on conserving SSR markers across sample groups. Consequently, the Group-wise Conserved SSR Selection method served as a benchmark to determine the maximum number of features selected in subsequent feature selection techniques.

SelectKBest FS

The second method used was SelectKBest, a filter-based approach. This technique selects the top K features based on their statistical scores. Specifically, it uses the Analysis of Variance (ANOVA) F-value between labels and features for classification tasks, employing the f_classif score function. This ensures that the features with the highest discriminative power are retained, contributing to a more efficient and accurate classification model.

Hybrid FS

Subsequently, a two-stage hybrid FS process was implemented, combining the SelectKBest filtering method with Random Forest (RF) [5], an ensemble learning technique. This approach further incorporated hyperparameter tuning and Recursive Feature Elimination (RFE), a wrapper method. Initially, SelectKBest was applied to reduce the dimensionality of the raw SSR CSV file to a predetermined limit of K features. Following this initial reduction, RandomizedSearchCV was used to identify the optimal hyperparameters for an RF Classifier. A random search strategy, consisting of 150 iterations and four-fold cross-validation, was performed. The optimization criterion was the weighted F1 score, chosen to account for the inherent class imbalance within the dataset. Finally, RFE was performed using the optimized RF model, iteratively eliminating the least significant features to further refine the feature subset.

Least absolute shrinkage and selection operator FS

Finally, an embedded feature selection method was employed using a logistic regression model with L1 regularization (LASSO) [29]. A One-vs-Rest (OvR) strategy was implemented to facilitate multiclass classification, where a distinct binary classifier was trained for each class. Each classifier was designed to discriminate a single class against all others, with the overall loss function calculated as the sum of individual class losses. A *randomizedSearchCV* procedure was performed across a broad spectrum of inverse regularization values to determine the optimal regularization strength. This optimization process used Leave-One-Out Cross-Validation (LOOCV) to evaluate each parameter configuration. Subsequently, the logistic regression model was re-fitted using the identified optimal regularization strength. Following the training of the optimized model, the SelectFromModel function was applied to identify the most significant SSRs. This selection was based on the non-zero coefficients derived from the Lasso model, effectively highlighting the features with the most significant predictive power. Mathematically, the Lasso loss function, which combines cross-entropy loss with L1 regularization, is expressed as described by Eq. 1.

$$\text{Lasso loss} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}(y_i = k)\log(p_{i,k}) + \lambda\sum_{k=1}^{K}\sum_{j=1}^{m}|\omega_{k,j}| \quad (1)$$

Where:

- n is the number of samples.
- K is the number of classes.
- m is the number of features.
- y_i is the true class label of the i-th sample.
- $p_{i,k}$ is the predicted probability that the i-th sample belongs to class k.
- $\mathbb{1}(y_i = k)$ is an indicator function that is 1 if the true class label is $y_i = k$, and 0 otherwise.
- $|\omega_{k,j}|$ are the model coefficients (weights) for class k and feature j.
- λ is the regularization strength.
- $\sum_{k=1}^{K} \sum_{j=1}^{m} |\omega_{k,j}|$ represents the L1 regularization term, which sums the absolute values of all the model weights.

Geographic origin classification of *C. sativa* via ML and SSR analysis

To determine the geographical origin of *C. sativa samples*, we explored and evaluated three distinct classifier algorithms for each subset of SSRs obtained in the preceding steps: Support Vector Classifier (SVC), Random Forest (RF) Classifier, and Gradient Boosting (GB) Classifier. For the SVC algorithm, a linear kernel was pre-selected. Balanced class weights were configured for both the SVC and the RF models to mitigate the impact of class imbalance.

Before model evaluation, hyperparameter tuning was performed for each SSR subset and classifier combination. This process employed Bayesian optimization with 100 iterations, using 3-fold cross-validation and the weighted F1 score as the scoring function. Subsequently, the optimized algorithms were evaluated using Leave-One-Out Cross-Validation. To identify the most informative SSRs within each subset, recursive feature elimination was used with a step size of 1. Specifically, RFE was applied to select the optimal number of features in each combination of FS method / classifier algorithm, ranging from 1 to 30 SSRs (in ascending order), at each iteration. It is important to note that the hyperparameters optimized in the previous step were held constant during the RFE process to maintain computational feasibility.

For each selected feature count, i.e., SSR added to the model generated by the algorithm, the following performance metrics were calculated 30 times to ensure robustness and statistical significance: weighted F1 score, precision, weighted recall and weighted precision. The inclusion of these distinct metrics ensures a comprehensive evaluation, allowing the models to be assessed not only by their overall balance (F1) but also by their specific ability to minimize false negatives (Recall) or false positives (Precision), depending on the specific priorities of the forensic investigation. The mean weighted F1 score values for each added SSR were plotted on a learning curve. Furthermore, and also for each combination of FS method/classifier algorithm, the average confusion matrix is also presented for the entire learning curve construction process.

SSR-based clustering via PCoA and UPGMA

Clustering analyzes were performed using Python 3.10.12 to complement the classification modeling. These analyzes used the optimal number and combination of SSRs identified within the subset that yielded the maximum performance for each classification model. Principal Coordinate Analysis (PCoA) was performed, using the Jaccard distance metric, and implemented using the Scipy v1.11.2 and Scikit-bio v0.5.8 libraries. Subsequently, hierarchical clustering was executed. This used the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm, also based on the Jaccard distance, and was performed using the Scipy v1.11.2 library.

Experiments and results

The alignment of the 38 genomes to the reference genome yielded an average alignment rate of 90.465 ± 1.268%. The aligned sequences were subsequently processed, leading to the identification of 467,527 SSRs in a raw SSR dataset. This dataset served as input for the FS methods, generating their respective subsets, which were later used in classification tasks across different supervised learning models (Fig. 1). The implemented pipeline resulted in 12 distinct combinations of FS strategies coupled with ML classifiers. A more detailed description of the SSRs present in each subset generated from the raw SSR file can be found in Table S2 of the Supplemental Material.

Regarding the GC-SSR FS, following filtering steps - including intersample comparisons and removing SSRs containing nucleotide N -, 281 SSRs within coding sequences were selected. Of these, 97.49% were classified as perfect sequences, 2.49% as non-perfect sequences, and 7 as complex sequences. Among perfect sequences, dinucleotide repeats were the most prevalent, accounting for more than 48%, followed by mononucleotide repeats in 20% and trinucleotide repeats in 16.7%. Regarding the di-nucleotide composition, AG repeats were observed in the highest proportion, followed by TC, AT, and CT repeats. Retaining only those SSRs conserved across all samples within at least one origin group resulted in the remaining 30 SSRs (Subset 1). Except for the loci LOC115707685 and LOC115707221, each possessing two markers, all other SSRs were located at distinct loci. Of these 30 SSRs, 18 were identified in the genomes

of the CO origin group, with 9 exclusive to this region: (GAT)4, (T)12, (TC)21, (A)10, (T)11, (GAAT)3, (GAA)4, (TCTTCG)3, (A)16. Similarly, 16 SSRs were found in the MP origin group, with 6 exclusive: (ATC)5, (CT)5, (TAA)4, (A)11, (CACCAA)3, (CGAAGA)3. In the FG origin group, 8 SSRs were identified, with 4 exclusive: (GA)5, (AGC)6, (TCA)5, (GAA)4. In particular, SSR (AAG) 5 was present in all regions. In contrast, no conserved SSR was identified in all genomes within the PG origin group.

Given that the quantity of SSRs selected by subsequent feature selection (FS) strategies is arbitrary, the Group-wise Conserved SSR selection (which naturally yielded 30 markers) was used as a benchmark to establish an upper threshold of 30 SSRs for the other subsets. It is important to note that this threshold served only as an initial ceiling for comparative purposes; the final number of markers included in the predictive models was dynamically determined through Recursive Feature Elimination, which optimized the feature set based on model performance metrics.

For subsets 2 and 4, 30 SSRs exhibiting the highest statistical relevance were selected. This selection was achieved through the application of the SelectKBest and Lasso methods, prioritizing the ANOVA F-values and the nonzero coefficient values, respectively. With the exceptions of (A)21 and (T)21 in LOC115705433, and (CA)5 and (CG)5 in LOC115707672 for subset 2, as well as (CAC)4 and (TGTT)3 in LOC115705474 for subset 4, all other extracted SSR markers were located at different loci within each subset. Lastly, for subset 3, the reduction in dimensionality using the SelectKBest technique yielded the 5,000 most relevant SSRs from the raw file. Subsequently, these SSRs were then used as input for a Random Forest (RF) Classifier, from which 30 SSRs were selected based on feature importance derived from the mean decrease in impurity of decision trees, utilizing the Recursive Feature Elimination (RFE) technique. Again, with the exception of (GA)5 and (TC)5 in LOC115705181, all other SSRs were present at different loci within this subset. Comparative analysis revealed that subsets 2, 3, and 4 shared the following selected SSRs: (ATT)4 at the LOC115702144, (T)10 at the LOC133031214 and (TCA)4 at the LOC115704795 locus. Similarly, subsets 1, 2, and 4 shared markers (GAAT)3 at locus LOC115706559, (TCA)5 at locus LOC133034467, and (CACCAA)3 at locus LOC115707482. Furthermore, the following SSRs were present in at least two different subsets: LOC115704143 (GGAAA)3, LOC115704734 (TCC)4, LOC115705161 (TGG)5, LOC115705474 (TGTT)3,

LOC115706366 (T)11, LOC115707221 (GAT)4,
LOC115707563 (TCT)4, LOC115707882 (AT)5,
LOC115708333 (A)16 and LOC133035775 (GGAG)3.

The choice criterion for defining the best combination of the FS strategy and the classification algorithm was based on a trade-off between the maximum score achieved during the learning curve and the number of SSRs incorporated. This selection strategy effectively assigns equal weight to model accuracy and model parsimony, aligning with multi-objective optimization principles. By prioritizing the solution that offers the highest performance with the fewest features, we aim to minimize the risk of overfitting and facilitate the practical implementation of smaller, more cost-effective multiplex panels in forensic laboratories. Of all distinct combinations, optimal performance with the fewest markers was achieved using subset 1 within a Gradient Boosting model (Table 3 and Fig. S4, Supplementary material). Using a combination of 7 SSRs from the original Subset 1, the model successfully distinguished all samples based on their origins (Fig. S4B), demonstrating a mean F1 score, precision (Acc), precision (Prec), and recall (Rec) of 1.0. These SSRs include LOC133034467 (TCA)5, LOC115706559 (GAAT)3, LOC115705224 (TAA)4, LOC115706366 (T)11, LOC115707030 (GA)5, LOC115707832 (AAG)5, and

LOC115707482 (CACCAA)3. Conversely, the other Gradient Boosting models generated from Subsets 2, 3, and 4 exhibited the lowest scores among all tested combinations, ranked in descending order: Subset 3 (F1: 0.92, Acc: 0.92, Prec: 0.93, Rec: 0.92) using 23 SSRs, Subset 4 (F1: 0.83, Acc: 0.84, Prec: 0.84, Rec: 0.84) using 3 SSRs, and Subset 2 (F1: 0.81, Acc: 0.81, Prec: 0.82, Rec: 0.81) using 8 SSRs (Figs. S5, S6 and S7, Supplementary material).

The SVC generated models demonstrated superior overall performance across all four subsets, requiring only a few additional Simple Sequence Repeats (SSRs) for classification (Table 1). In particular, using only 9 SSRs from Subset 2, we achieved a maximum F1 score of 1.0 (Fig. 2 - B). These SSRs, in order of importance for the model

Table 1 Number of SSRs incorporated in the model, Mean Weighted F1-score (F1), Mean Accuracy (Acc), Mean Weighted Precision (Prec), and Mean Weighted Recall (Rec) metrics with their respective standard deviations (SD) calculated from 30 replicates for different subsets using Linear SVC. Subset with the best performance using the respective algorithm is marked with an asterisk (*)

	Linear SVC				
	SSRs incorp.	Mean F1±SD	Mean Acc±SD	Mean Prec±SD	Mean Rec±SD
Subset 1	10	1.0±0.0	1.0±0.0	1.0±0.0	1.0±0.0
Subset 2*	9	1.0±0.0	1.0±0.0	1.0±0.0	1.0±0.0
Subset 3	12	0.97±1.11e-16	0.97±2.22e-16	0.98±2.22e-16	0.97±2.22e-16
Subset 4	11	1.0±0.0	1.0±0.0	1.0±0.0	1.0±0.0

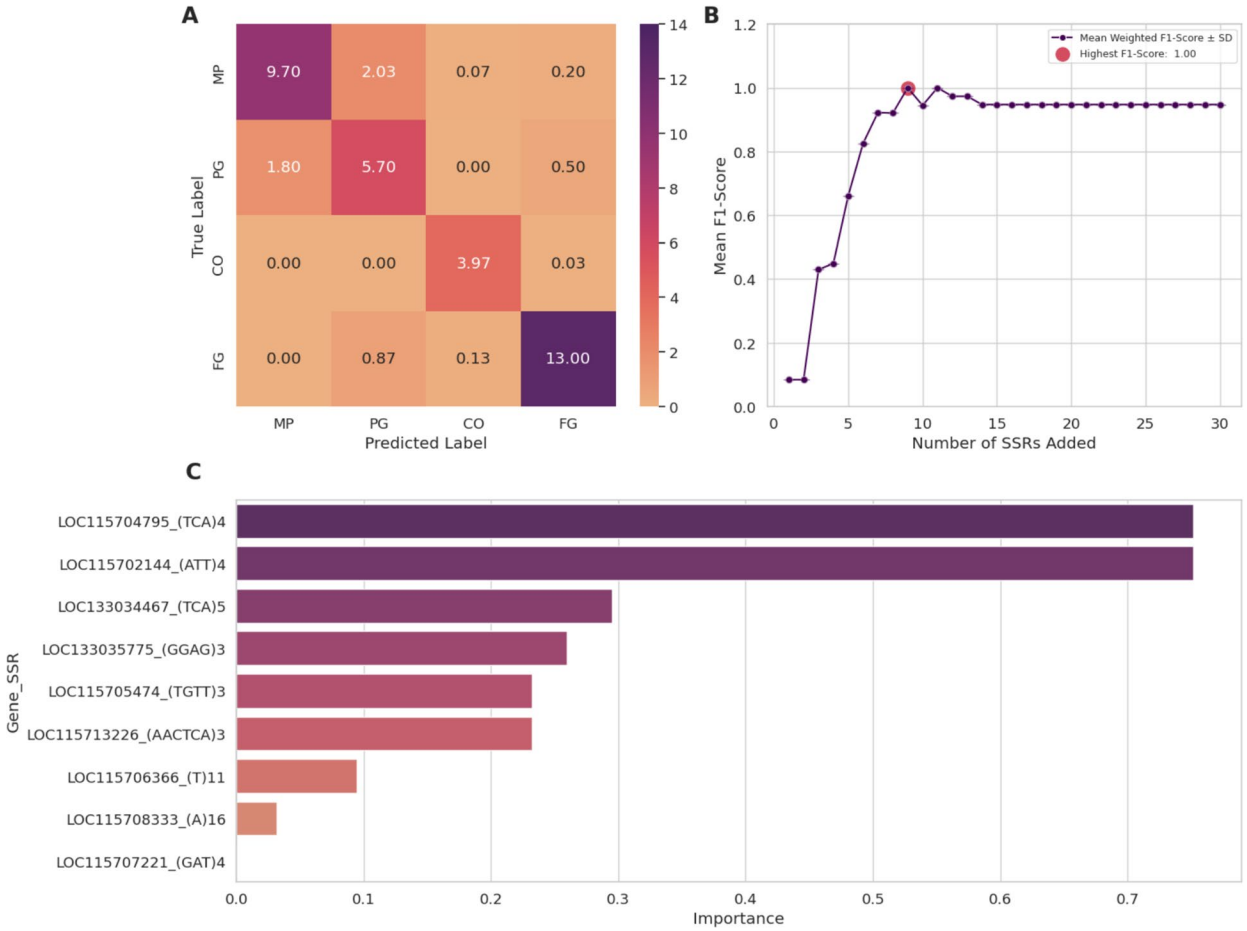

Fig. 2 Evaluation of the Linear SVC Model using Subset 2 SSRs Selected by SelectKBest Feature Selection. (A) The average confusion matrix across all replicates is presented, demonstrating the classification performance for the four sample origins: MP, PG, CO, and FG. (B) The weighted F1-score learning curve, including standard deviations, is depicted as a function of the number of added SSR markers. The red marker indicates the maximum observed F1-score. (C) Feature importances are shown for the 9 SSRs utilized in the Linear SVC model that achieved the highest score

(Fig. 2 - C), were as follows: LOC115704795 (TCA)4, LOC115702144 (ATT)4, LOC133034467 (TCA)5, LOC133035775 (GGAG)3, LOC115705474 (TGTT)3, LOC115713226 (AACTCA)3, LOC115706366 (T)11, LOC115708333 (A)16, and LOC115707221 (GAT)4. Subsets 1 and 4 exhibited identical performance with 10 and 11 SSRs, respectively. Subset 3 showed slightly lower classification performance using 12 SSRs (F1: 0.97; Accuracy: 0.97, Precision: 0.98, Recall: 0.97).

RF classifiers also showed high predictive ability (Table 2), although no subset reached the maximum possible score. The scores obtained for each subset in this classification model were in decreasing order of performance: Subset 1 (F1: 0.97, Acc: 0.97, Prec: 0.97, Rec: 0.97) using 14 SSRs, Subset 3 (F1: 0.95, Acc: 0.95, Prec: 0.95, Rec: 0.95) using 22 SSRs, Subset 2 (F1: 0.92, Acc: 0.92, Prec: 0.92, Rec: 0.92) using 14 SSRs and Subset 4 (F1: 0.85, Acc: 0.87, Prec: 0.85, Rec: 0.87) using 14 SSRs. RF classifiers learning curves, features importances and confusion matrices can also be found in supplementary material (Figs. S8-11).

Table 2 Number of SSRs incorporated in the model, Mean Weighted F1-score (F1), Mean Accuracy (Acc), Mean Weighted Precision (Prec), and Mean Weighted Recall (Rec) metrics with their respective standard deviations (SD), calculated from 30 replicates for different subsets using the Random Forest (RF) classifier. Subset with the best performance using the respective algorithm is marked with an asterisk (*)

	Random Forest				
	SSRs incorp.	Mean F1±SD	Mean Acc±SD	Mean Prec±SD	Mean Rec±SD
Subset 1*	14	0.97±0.02	0.97±0.02	0.97±0.02	0.97±0.02
Subset 2	14	0.92±0.02	0.92±0.02	0.92±0.02	0.92±0.02
Subset 3	22	0.94±0.03	0.94±0.03	0.94±0.03	0.94±0.03
Subset 4	14	0.85±0.03	0.87±0.02	0.85±0.04	0.87±0.02

Discussion

Feature selection

Advances in molecular biology technologies have led to an exponential increase in biological data available in recent decades, leading to the large p, small n issue, a very common scenario in life science studies where the number of observations or samples (n) is hundreds or thousands of times smaller than the number of markers or characteristics extracted (p). Within the field of ML, this can be translated into what we call the curse of dimensionality, that is, as dimensionality (p) increases, the number of data points (n) required for good performance of any ML algorithm increases exponentially [3]. In addition to the problem of high-dimensional data, another obstacle that can affect the performance of ML models is sparsity, which refers to the high number of noninformative features that have a small or insignificant effect on the characteristic of interest, leading to an unstable predictive model [2]. Therefore, the use of FS algorithms as a means of dimensionality reduction is a crucial step in bioinformatic pipelines applied to different omics datasets. In the present work, we explore four different SSRs filtering strategies: GC-SSR Selection, a personalized adaptation that served as a reference technique to determine the maximum number of selected markers, in addition to filter (SelectKBest), hybrid(RF associated with RFE) and embedded (LASSO) methods for FS, obtaining four partially distinct subsets of selected SSRs.

Surprisingly, subset 1 (GC-SSR FS) achieved the best performance in predicting the origin of the sample in two of three models tested compared to the other subsets (Tables 1, 2 and 3). Although it is a form of adapted selection based on biological intuition without the application of formal statistical analysis or the complexity of a model-based FS method, GC-SSR presented the most informative set of SSRs to separate samples regarding their origin of culture, likely indicating that these markers represent genomic regions that may be functionally important and evolutionarily conserved. While we hypothesize that this conservation implies functional constraints or selective pressure, we did not explicitly quantify evolutionary rates (e.g., dN/dS ratios) in this study. This decision was based on the primary focus on forensic discrimination and the current challenges regarding the precise functional annotation of these loci in diverse illicit strains. Since these SSRs were filtered to be present in all samples from at least one of the origin groups (MP, FG, PG or CO), they could reflect genetic stability or selection for traits specific to certain populations. Such features may be more biologically meaningful than other subsets. This subset might inherently carry less noise, which is beneficial for model performance, especially for classifiers like SVC with a linear kernel or RF, which might be sensitive to noisy or irrelevant features. On the other hand, methods like SelectKBest purely on statistical metrics and may retain features that, while statistically significant, are not as biologically informative.

Regarding subsets 3 (Hybrid FS) and 4 (LASSO FS), there are some combinations of factors that may justify their lower performance in certain models. Despite being a popular solution for dimensionality reduction, LASSO can be overly aggressive in feature elimination and tends to select only one feature from a group of highly correlated features. Although this can be beneficial in highly correlated or high-dimensional datasets, it can result in the loss of important features that are useful for classification [38]. As for the hybrid FS method, which combines filter (SelectKBest), ensemble (RF), and wrapper (RFE) approaches, the greater degree of complexity introduced in the attribute selection task may not always result in better performance, particularly in small and imbalanced datasets. Also, specifically in the RFE-associated RF application stage, if the hyperparameters were not optimized perfectly, the performance of the hybrid FS method might not reach its full potential since this model is more sensitive to hyperparameter tuning.

Furthermore, it is important to clarify that our approach treated each SSR motif as an independent feature during the selection process, even when multiple motifs originated from the same locus (e.g., LOC115705433 in Subset 2). We did not manually merge or filter these features; instead, we relied on the FS algorithms to determine their individual discriminative power. The selection of multiple motifs from a single locus by methods like SelectKBest suggests that these specific repeats may provide complementary information. It is possible that these markers exhibit linkage disequilibrium or distinct mutational patterns that, when combined, enhance the statistical separation of the sample groups.

Table 3 Number of SSRs incorporated in the model, Mean Weighted F1-score (F1), Mean Accuracy (Acc), Mean Weighted Precision (Prec), and Mean Weighted Recall (Rec) metrics with their respective standard deviations (SD) calculated from 30 replicates for different subsets using the Gradient Boosting (GB) classifier. Subset with the best performance using the respective algorithm is marked with an asterisk (*)

	Gradient Boosting				
	SSRs incorp.	Mean F1±SD	Mean Acc±SD	Mean Prec±SD	Mean Rec±SD
Subset 1*	7	1.0±0.0	1.0±0.0	1.0±0.0	1.0±0.0
Subset 2	8	0.81±0.03	0.81±0.03	0.82±0.03	0.81±0.03
Subset 3	23	0.92±0.02	0.92±0.02	0.93±0.02	0.92±0.02
Subset 4	3	0.83±0.06	0.84±0.06	0.83±0.07	0.83±0.06

Classification task and origin prediction

As for the task of predicting the origin and classifying samples of *C. Sativa*, the highlight goes to the linear kernel SVC, which presented excellent performance regardless of the subset of SSRs used (Table 1). These results indicate that, regardless of the FS method used, SSRs are extremely discriminative markers and provide practically perfect linear separability, using a simpler classification model, with low computational cost and easily interpretable when compared to RF and GB. The combination of subset 2 associated with SVC proved to be the most efficient, performing the task with only 9 SSRs.

The 9 SSRs are located at different protein-coding loci on chromosome 1. LOC115704795 (TCA)4, present in samples from MP, encodes for *CsNAC01*, which is part of the NAC transcription factor gene family, related to direct and indirect regulation of response to abiotic stresses, such as low temperatures, excessive heat, drought, etc. [25, 45]. LOC115702144 (ATT)4, also present in MP samples, and LOC115707221 (GAT)4, present in CO samples, encode the putative ubiquitin protein ligase E3 LIN-1 and ATL42, respectively, which are members of the E3 Ubiquitin Ligases family, a conserved group of enzymes among eukaryotes responsible for a range of regulatory functions that aid in covalent binding of ubiquitin to target proteins [21]. LOC115705474 (TGTT)3, present in MP and PG samples, encodes the F-box protein CPR1, which is responsible for negatively regulating plant immunity by promoting the degradation of resistance proteins (NLR), such as SNC1 and RPS2, through the ubiquitin proteasome system, essential to avoid excessive activation of immune responses, which can lead to impaired plant growth and autoimmunity [7, 17]. LOC115708333 (A)16, present in CO samples, encodes a probable intermediate-associated protein of complex I 30 (CIA30) or NADH dehydrogenase ubiquinone 1 alpha subcomplex assembly factor 1 (NDUFAF1), an important assembly factor of mitochondrial complex I, the largest protein complex operating in oxidative phosphorylation [37, 44]. LOC133034467 (TCA)5 and LOC133035775 (GGAG)3, present in samples from FG and some samples from the PG region, as well as LOC115713226 (AACTCA)3, present in MP and PG samples, and LOC115706366 (T)11 present in CO samples, are located in *C. sativa* uncharacterized genome regions. A plausible hypothesis is that these SSRs are a consequence of adaptive mutations important for the adaptation of *C. sativa* to different cultivation environments, such as the regulation of the response to environmental stresses, modulation of plant immunity, and metabolic and energetic adaptability of the plant, but additional studies would be necessary to prove such assumptions. In addition, because of the presence of several loci that have not yet been characterized, it is possible that there is some hidden correlation that has not yet been studied.

Combinations involving GB classification models obtained, on average, a lower performance than the others (Table 3). Subset 2 and subset 4 associated with GB obtained metrics just above 0.80 at their peak performance, despite using only 8 and 3 SSRs to classify the samples, respectively. Subset 3 associated with GB obtained a good peak performance above 0.90 for all metrics, but required 23 SSRs to achieve this score, a higher amount than the forensic microsatellite panels already validated and used for the identification and individualization of samples of Brazilian *C. sativa* in previous work [14, 30]. Subset 1 associated with the GB model, on the other hand, obtained a perfect score and correctly predicted the origin of all samples using only 7 SSRs (Fig. 4). Again, the 7 SSRs used for classification are also located at different loci of chromosome 1. LOC133034467 (TCA)5, present in MP samples, and LOC115706559 (GAAT)3/LOC115706366 (T)11, present in CO samples, are in uncharacterized coding regions. LOC115705224 (TAA)4, present in MP samples, encodes the late elongated hypocotyl protein (LHY), which regulates photoperiodic flowering through the circadian clock in *Arabidopsis thaliana* [31, 47]. LOC115707030 (GA)5, present in samples originating from the FG and MP groups, encodes Myosin XVII, which is part of a large family of motor proteins involved in various processes in eukaryotic cells [27, 39]. LOC115707832 (AAG)5 encodes the Rho proteins of plants (ROPs) guanine nucleotide exchange factor 12, involved in sending signals to downstream cellular targets, and is curiously present in practically all samples, with the exception of some from PG, which, added to its relatively low relevance within the model (Fig. S4C, Supplementary material), suggests that it was used to classify a specific sample within the group [4, 13, 48]. The LOC115707482 (CACCAA)3, present in MP and CO samples, encodes the early flowering protein MYB, a transcription factor responsible for regulating the flowering process in response to changes in light and temperature in *Arabidopsis thaliana* [18, 46]. Although the GB model also achieved a perfect score and used the lowest number of SSRs among all combinations tested, its inferior performance for the other subsets suggests a greater dependence on specific characteristics to achieve high performance. This indicates greater instability and possibly lower generalizability compared to SVC. Furthermore, GB models require higher computational power and greater hyperparameter tuning.

Interestingly, samples from the PG region did not present any SSRs that were present in all samples from the group, but rather were distributed among a few representatives. Because of this, the amount of SSRs required to predict the place of origin of samples from this group in

all classifiers was considerably higher when compared to those from MP, FG, and CO. This characteristic indicates a greater intragroup genetic diversity, possibly caused by factors such as variation in cultivation practices, different subpopulations within the same location, or the influence of varied environmental conditions, and may lead to a lower consistency between SSR markers among samples from the region, thus requiring more markers to represent the greater genetic heterogeneity present. Consequently, in the GC-SSR analysis (Subset 1), the PG samples were classified not by their own unique conserved markers, but by their distinct profile across the markers defined by the MP, CO, and FG groups. The classifiers effectively leveraged the specific absence or variable frequency of these intergroup conserved markers to distinguish the PG samples, demonstrating that a panel derived from specific groups can still be useful for identifying samples from diverse, highly heterozygous origins.

Clustering analysis

PCoA and hierarchical clustering were performed for each model using the combination and the number of SSRs from the subset with which they reached their peak performance, according to Tables 1, 2 and 3. Again, the 9 SSRs from subset 2 used by the SVC model were more prominent and could visually distinguish the four seizure groups according to their biogeographic origin (Figs. 3 and 4). PCoA and hierarchical clustering analysis related to the 7 and 14 most informative SSRs from Subset 1 applied to the models generated by GB and RF, respectively, can be found in the supplementary material (Figs. S12, S13, S14 and S15).

In PCoA based on SSRs used by the best model obtained with the SVC classifier (Fig. 3), the principal coordinates Coord.1 and Coord.2 explained approximately 41.13% and 16.54%, respectively, of the genetic variance explained, based on the Jaccard distance matrix. Interestingly, samples belonging to the foreign group (FG), coming from seeds of different commercial strains seized from postal services, formed a group spread across the X and Y axes. BFP reports suggest that the seeds came from Europe, although the exact location of the origin could not be determined. This dispersion in the PCoA likely reflects the high genetic diversity and admixture characteristic of commercial hybrid strains. Unlike the geographically distinct groups, the FG samples lack a single, defined biological origin, representing instead a mixture of diverse genotypes entering the country. We acknowledge that this heterogeneity limits the interpretation of clustering patterns for this specific group; however, it accurately represents the forensic reality of the 'foreign' class in the Brazilian illegal market, which is composed of a complex variety of commercial strains rather than a single lineage.

Meanwhile, hierarchical cluster analysis (Fig. 4) based on the Jaccard distance revealed that while the FG, MP and CO groups formed distinct, well-defined clusters, samples from the Paraguay group (PG) were scattered among these clusters. This pattern likely indicates higher genetic diversity or a more heterogeneous genetic structure within the PG group and reinforces the results found during the evaluation of the classification models, which show the absence of distinctive and informative SSR markers in the samples. This heterogeneity is consistent with the origin of these samples: they were derived from large-scale 'pressed brick' seizures intercepted along major entry routes (e.g., Mato Grosso do Sul and Paraná). Unlike single-site

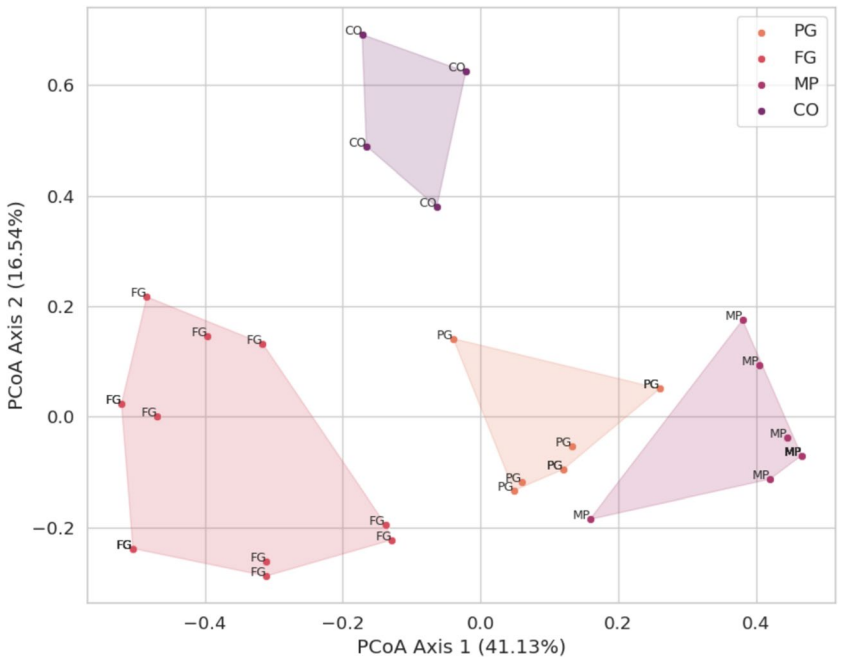

Fig. 3 Principal Coordinate Analysis (PCoA) plot based on Jaccard distance calculated from the 9 SSR markers used in SVC best model. The plot shows the ordination of samples across two dimensions (PC1 and PC2), which together explain 41.13% and 16.54% of the variation, respectively

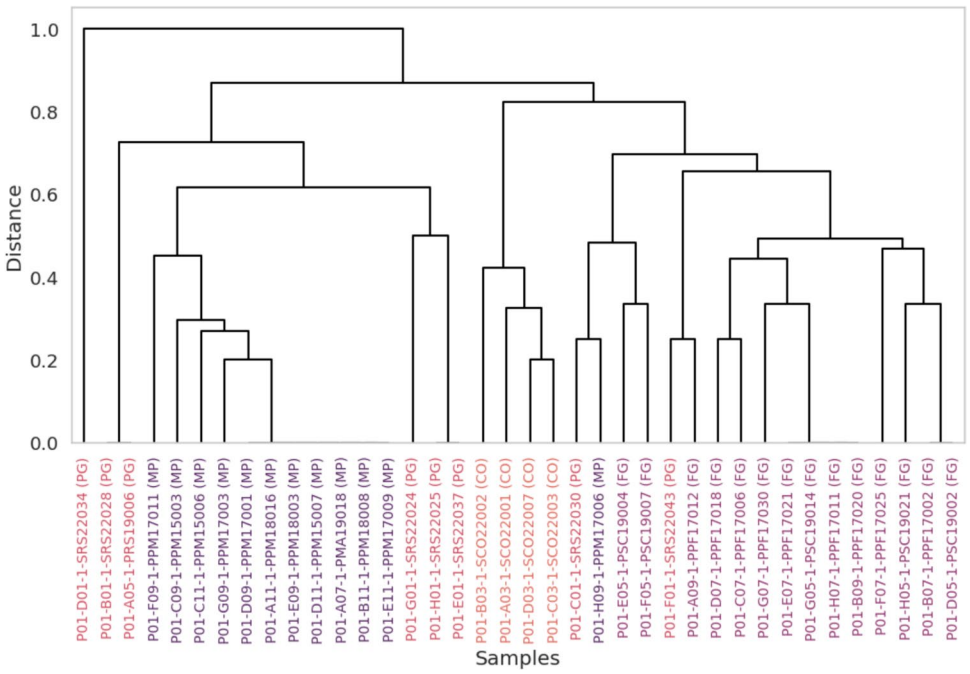

Fig. 4 Hierarchical clustering dendrogram based on UPGMA method and Jaccard distance. The dendrogram illustrates the genetic relationships among samples using the 9 SSR markers used in SVC best model. The samples are color-coded according to their origin, with MP in purple, PG in red, CO in orange-red, and FG in darker purple

cultivations, these shipments likely represent an aggregation of harvested material from multiple distinct farms or regions within Paraguay, resulting in a highly admixed genetic profile that lacks the fixed markers observed in more localized groups.

Conclusion

Our study successfully identified informative SSR markers capable of distinguishing *C. sativa* samples from defined geographic regions, as well as samples of uncertain provenance (FG), which are potentially related to trafficking routes supplying the illegal Brazilian market. Using simple filtering and feature selection techniques, associated with classification models widely used in the literature, with emphasis on linear kernel SVC, we obtained different panels of SSR markers with high predictive value and discriminative power. However, it is important to highlight the need for larger and balanced sample databases to evaluate the true discriminatory power of these markers and possibly discover and incorporate more informative markers for Brazilian *C. sativa* samples, also reducing the risk of overfitting. Specifically, regarding the limitation observed in the commercial strains (FG), future research must prioritize the use of plant samples with well-documented and certified origins. Establishing a 'ground truth' with certified samples will be crucial to obtaining more robust and meaningful results and to refining the identification of complex hybrid strains in practical investigative settings.

Although the results obtained are promising, highlighting the usefulness of these markers in the context of supervised ML, additional studies are necessary in new sample groups for experimental validation on the bench to validate their application in practical investigative settings. This would also allow the investigation of several forensic quality parameters commonly evaluated in these types of markers, such as the mean number of alleles, the effective number of alleles, the probability of random match, the power of exclusion and the power of discrimination, and population differentiation metrics such as the Fixation Index (F_{ST}), ideal for forensic and law enforcement purposes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00414-025-03716-7.

Acknowledgements This work was supported by grants from the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul - FAPERGS [24/2551-0001392-0 and 23/2551-0001894-2], Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq [314082/2021-2, 408154/2022-5, 440279/2022-4, and 404319/2024-6], and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES - Finance Code 001.

Author Contributions **Cássio Augusto Rodrigues Bettim:** Conceptualization, Methodology, Data Curation, Visualization, Investigation, Writing-Original draft. **Cássio Augusto Rodrigues Bettim**: Conceptualization, Methodology, Data Curation, Visualization, Investigation. **Oscar Victor Cardenas Alegría:** Conceptualization, Methodology, Data curation, Investigation, Software. **Eduardo Filipe Avila Silva:** Conceptualization, Investigation, Writing-review and editing, Methodology, Supervision, Formal analysis, Data curation. Franciele **Maboni Siqueira:** Conceptualization, Resources. **Flávio Anastácio de Oliveira Camargo:** Funding acquisition, Conceptualization. **Clarice Sampaio Alho:** Funding acquisition, Conceptualization, Resources, Formal analysis. **Márcio Dorn:** Conceptualization, Investigation, Funding acquisition, Writing-review and editing, Methodology, Supervision, Re-sources, Project administration, Formal analysis, Data curation.

Funding The Article Processing Charge (APC) for the publication of this research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) (ROR identifier: 00x0ma614).

References

1. Abbasi Holasou H, Panahi B, Shahi A, Nami Y (2024) Integration of machine learning models with microsatellite markers: New avenue in world grapevine germplasm characterization. Biochem Biophys Reports 38:101678
2. Al-Mamun HA, Danilevicz MF, Marsh JI, Gondro C, Edwards D (2024) Exploring genomic feature selection: A comparative analysis of gwas and machine learning algorithms in a large-scale soybean dataset. The Plant Genome
3. Bellman RE (1961) Adaptive Control Processes: A Guided Tour. Princeton University Press
4. Berken A, Thomas C, Wittinghofer A (2005) A new family of rhogefs activates the rop molecular switch in plants. Nature 436:1176–1180
5. Breiman L (2001) Random forests. Mach Learn 45:5–32
6. Cheng YC, Houston R (2021) Evaluation of the trnk-matk-trnk, ycf3, and accd-psal chloroplast regions to differentiate crop type and biogeographical origin of cannabis sativa. Int J Legal Med 135:1235–1244
7. Cheng YT, Li Y, Huang S, Huang Y, Dong X, Zhang Y, Li X (2011) Stability of plant immune-receptor resistance proteins is controlled by skp1-cullin1-f-box (scf)-mediated protein degradation. Proc Natl Acad Sci 108:14694–14699
8. Cisana S, Di Nunzio M, Brenzini V, Omedei M, Seganti F, Ververi C, Gerace E, Salomone A, Berti A, Barni F, Schiavone S, Coppi A, Di Nunzio C, Garofano P, Alladio E (2024) An initial exploration of machine learning for establishing associations between genetic markers and thc levels in cannabis sativa samples. Forensic Science International: Genetics 73:103123. https://doi.org/10.1016/j.fsigen.2024.103123
9. Cisana S, Omedei M, Di Nunzio M, Seganti F, Brenzini V, Coppi A, Berti A, Di Nunzio C, Garofano P, Alladio E (2022) Evaluation of genetic markers for the analysis of thc levels of cannabis sativa samples using principal component analysis – a preliminary study. Forensic Sci Int: Genetics Suppl Series 8:120–121. https://doi.org/10.1016/j.fsigss.2022.10.004
10. Dantas C, Neto S, Alves S, Pinheiro K, Franco E, Ramos R (2024) Satin: a micro and mini satellite mining tool of total genome and coding regions with analysis of perfect repeats polymorphism in coding regions. BMC Bioinf 25:217
11. Di Nunzio M, Barrot-Feixat C, Gangitano D (2024) Characterization and evaluation of nine cannabis sativa chloroplast snp markers for crop type determination and biogeographical origin on european samples. Forensic Sci Int Genet 68:102971
12. Dias Filho CR et al (2020) Introdução à genética forense. Ciência contra o crime, Millenium Editora Ltda., Campinas, SP
13. Engelhardt S, Trutzenberg A, Hückelhoven R (2020) Regulation and functions of rop gtpases in plant–microbe interactions. Cells 9:2016
14. Fett MS, Mariot RF, Avila E, Alho CS, Stefenon VM, de Oliveira Camargo FA (2018) 13-loci str multiplex system for brazilian seized samples of marijuana: individualization and origin differentiation. Int J Legal Med 133:373–384
15. Fett MS, Mariot RF, Ortiz RS, Avila E, de Oliveira Camargo FA (2020) Geographic origin determination of brazilian cannabis sativa l. (marihuana) by multi-element concentration. Foren Sci Int 315:110459
16. Gill M, Anderson R, Hu H, Bennamoun M, Petereit J, Valliyodan B, Nguyen HT, Batley J, Bayer PE, Edwards D (2022) Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. BMC Plant Bio 22
17. Gou M, Shi Z, Zhu Y, Bao Z, Wang G, Hua J (2011) The f-box protein cpr1/cpr30 negatively regulates r protein snc1 accumulation. Plant J 69:411–420
18. Hughes CL, Harmer SL (2023) Myb-like transcription factors have epistatic effects on circadian clock function but additive effects on plant growth. Plant Direct 7
19. Hussain T, Jeena G, Pitakbut T, Vasilev N, Kayser O (2021) Cannabis sativa research trends, challenges, and new-age perspectives. i Science 24:103391
20. John M, Haselbeck F, Dass R, Malisi C, Ricca P, Dreischer C, Schultheiss SJ, Grimm DG (2022) A comparison of classical and machine learning-based phenotype prediction methods on simulated data and three plant species. Frontier Plant Sci 13
21. Kelley DR (2018) E3 ubiquitin ligases: Key regulators of hormone signaling in plants. Molec Cellular Proteom 17:1047–1054
22. Kowalczyk M et al (2018) Molecular markers used in forensic genetics. Med Sci Law 58:201–209
23. Langmead B, Salzberg S (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9:357–9. https://doi.org/10.1038/nmeth.1923
24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G (2009) Genome project data processing s: The sequence alignment/map format and samtools. Bioinformatics 25:2078–9
25. Li Q, Zhang H, Yang Y, Tang K, Yang Y, Ouyang W, Du G (2024) Genome-wide identification of nac family genes and their expression analyses in response to osmotic stress in cannabis sativa l. Int J Mol Sci 25:9466
26. Liu Q, Zuo Sm, Peng S, Zhang H, Peng Y, Li W, Xiong Y, Lin R, Feng Z, Li H, Yang J, Wang GL, Kang H (2024) Development of machine learning methods for accurate prediction of plant disease resistance. Engineering 40:100–110
27. Madison SL, Nebenführ A (2013) Understanding myosin functions in plants: are we there yet? Curr Opin Plant Biol 16:710–717
28. Ministry of Justice and Public Security (2021) II Brazilian Drug Report: Executive Summary. https://www.gov.br. accessed: 08/19/2024
29. Muthukrishnan R, Rohini R (2016) Lasso: A feature selection technique in predictive modeling for machine learning, In: 2016 IEEE International Conference on Advances in Computer Applications (ICACA), pp 18–20
30. de Oliveira Pereira Ribeiro L, Avila E, Mariot RF, Fett MS, de Oliveira Camargo FA, Alho CS (2020) Evaluation of two 13-loci str multiplex system regarding identification and origin discrimination of brazilian cannabis sativa samples. Int J Legal Med 134:1603–1612
31. Park MJ, Kwon YJ, Gil KE, Park CM (2016) Late elongated hypocotyl regulates photoperiodic flowering via the circadian clock in arabidopsis. BMC Plant Bio 16

32. Pereira JFQ, Pimentel MF, Amigo JM, Honorato RS (2020) Detection and identification of cannabis sativa l. using near infrared hyperspectral imaging and machine learning methods. a feasibility study. Spectrochimica Acta Part A: Molec Biomolec Spectros 237:118385
33. Polícia Federal do Brasil (2022) Operação semilha combate tráfico de sementes entre uruguai e brasil. https://www.gov.br/pf/pt-br/assuntos/noticias/2022/02/operacao-semilha-combate-trafico-de-sementes-entre-uruguai-e-brasil. accessed: 2024-10-31
34. Rock EM, Parker LA (2020) Constituents of Cannabis Sativa. Springer International Publishing. p 1–13
35. Roman MG, Houston R (2020) Investigation of chloroplast regions rps16 and clpp for determination of cannabis sativa crop type and biogeographical origin. Leg Med 47:101759
36. Sander Fett M, Fogliatto Mariot R, Scorsatto Ortiz R, Tiecher T, Anastácio de Oliveira Camargo F (2021) The use of vegetal tissue multi-element content as an indicator of soil or substrate type employed to cultivate cannabis sativa l. (marijuana). Forensic Chem 23:100319
37. Subrahmanian N, Remacle C, Hamel PP (2016) Plant mitochondrial complex i composition and assembly: A review. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1857:1001–1014
38. Tolоși L, Lengauer T (2011) Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27:1986–1994
39. Trivedi DV, Nag S, Spudich A, Ruppel KM, Spudich JA (2020) The myosin family of mechanoenzymes: From mechanisms to therapeutic approaches. Annu Rev Biochem 89:667–693
40. United Nations Office on Drugs and Crime (UNODC) (2024) World drug report 2024. https://www.unodc.org/unodc/en/data-and-analysis/world-drug-report-2024.html. accessed: 2024-10-31
41. Upadhyaya SR, Danilevicz MF, Dolatabadian A, Neik TX, Zhang F, Al-Mamun HA, Bennamoun M, Batley J, Edwards D (2024) Genomics-based plant disease resistance prediction using machine learning. Plant Pathol
42. Vieira MLC, Santini L, Diniz AL, Munhoz, CdF (2016) Microsatellite markers: what they mean and why they are so useful. Genet Mol Biol 39:312–328
43. Vourlaki IT, Ramos-Onsins SE, Pérez-Enciso M, Castanera R (2024) Evaluation of deep learning for predicting rice traits using structural and single-nucleotide genomic variants. Plant Methods 20
44. Wang M, Ren LT, Wei XY, Ling YM, Gu HT, Wang SS, Ma XF, Kong GC (2022) Nac transcription factor twnac01 positively regulates drought stress responses in arabidopsis and triticale. Frontier Plant Sci 13
45. Wang P, Lehti-Shiu MD, Lotreck S, Segura Abá K, Krysan PJ, Shiu SH (2024) Prediction of plant complex traits via integration of multi-omics data. Nature Commun 15
46. Yan Y, Shen L, Chen Y, Bao S, Thong Z, Yu H (2014) A myb-domain protein efm mediates flowering responses to environmental cues in arabidopsis. Dev Cell 30:437–448
47. Yang M, Lin W, Xu Y, Xie B, Yu B, Chen L, Huang W (2024) Flowering-time regulation by the circadian clock: From arabidopsis to crops. The Crop J 12:17–27
48. Zhang Y, McCormick S (2007) A distinct mechanism regulating a pollen-specific guanine nucleotide exchange factor for the small gtpase rop in arabidopsis thaliana. Proc Natl Acad Sci 104:18830–18835