DEVELOPMENT OF A COMPREHENSIVE GENETIC TOOL FOR IDENTIFICATION OF *CANNABIS SATIVA* SAMPLES FOR FORENSIC AND INTELLIGENCE PURPOSES

_____

Dissertation

Presented to

The Faculty of the Department of Forensic Science

Sam Houston State University

_____

In Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

_____

by

Rachel Michelle Houston

May, 2018

DEVELOPMENT OF A COMPREHENSIVE GENETIC TOOL FOR IDENTIFICATION OF *CANNABIS SATIVA* SAMPLES FOR FORENSIC AND INTELLIGENCE PURPOSES

by

Rachel Michelle Houston

_____

APPROVED:

David Gangitano, PhD
Committee Director

Sheree Hughes-Stamm, PhD
Committee Co-Director

Sibyl Bucheli, PhD
Committee Member

Bobby LaRue, PhD
Committee Member

Phillip Lyons, PhD
Dean, College of Criminal Justice

DEDICATION

"I can do all things through Him who strengthens me." – Philippians 4:13

I would first like to thank my advisor, Dr. David Gangitano, for always believing in me and steering me in the right the direction. Many thanks to my co-advisor, Dr. Sheree Hughes-Stamm, for lending an ear in a time of need and pushing me to accomplish so much more than just my PhD dissertation.

I would also like to thank my PhD cohort for the intellectually stimulating conversions, the collaborations, and for all the laughs we have had in the past four years.

Finally, I must express my very profound gratitude to my family and to my boyfriend for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this dissertation. This accomplishment would not have been possible without them. Thank you.

ABSTRACT

Houston, Rachel Michelle, *Development of a comprehensive genetic tool for the identification of Cannabis sativa samples for forensic and intelligence purposes.* Doctor of Philosophy (Forensic Science), May, 2018, Sam Houston State University, Huntsville, Texas.

Cannabis sativa L. (marijuana) is the most commonly used illicit drug in the United States. Due to partial legalization, law enforcement faces a unique challenge in tracking and preventing flow of the legal marijuana to states where it is still prohibited. Moreover, significant illegal *C. sativa* traffic from Mexico exists at the US border. To date, no DNA method for *Cannabis* using short tandem repeat (STR) markers following International Society of Forensic Genetics (ISFG) or Scientific Working Group on DNA Analysis Methods (SWGDAM) recommendations (e.g., use of sequenced allelic ladder, use of tetra-nucleotide STR markers, etc.) has been reported. In addition, there is no existing *Cannabis* STR reference population database that can be used for forensic purposes (e.g., population in Hardy-Weinberg and linkage equilibrium, parameters of forensic interest). There have been very limited chloroplast (cpDNA) and mitochondrial DNA (mtDNA) studies investigating *C. sativa* haplotypes in the Americas. Lastly, massively parallel sequencing (MPS) technology has not yet been applied to targeted sequencing of *C. sativa* for forensic purposes. This project explores the use of genetic tools to identify and determine the origin of *C. sativa* for forensic purposes. Results provide the forensic DNA community a comprehensive genetic tool (STR, cpDNA, mtDNA, and MPS) that allows for the individualization of *Cannabis* samples, the association of different cases as well as origin determination of samples for forensic and intelligence purposes.

First, a previously reported 15-loci STR multiplex was evaluated. Results of the evaluation indicated that this STR system is not suitable for forensic identification due to

several issues; namely high heterozygote peak imbalance in some markers, overlapping alleles between two closely located STR markers, high stutter peaks in dinucleotide markers, inter-loci peak imbalance and presence of null alleles in four of the markers.

Therefore, a novel 13-loci STR multiplex was developed and optimized for *C. sativa* identification (3500 Genetic analyzer), according to ISFG and SWGDAM recommendations, using primer and multiplex STR design software, and a gradient PCR approach for optimal annealing temperature determination. This STR multiplex was validated according SWGDAM guidelines. Case-to-case comparisons were performed by phylogenetic analysis using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method and parsimony analysis with statistically significant differences detected using pair-wise genetic-distance comparisons. Homogeneous subpopulations (low F_{ST}) were determined by phylogenetic analysis and confirmed by bootstrap analysis (95% confidence interval). Results revealed a homogeneous subpopulation that could be used as a *Cannabis* reference STR population database (N=101) with parameters of population genetics (observed heterozygosity, expected heterozygosity, Hardy-Weinberg equilibrium, and linkage disequilibrium) and of forensic interest (allele frequencies and power of discrimination, etc.).

Another previously reported multi-loci system was modified and optimized to genotype five chloroplast and two mitochondrial markers. For this purpose, two assays were designed: a homopolymeric STR pentaplex and a SNP triplex with one chloroplast (Cscp001) marker shared by both methods for quality control. For successful mitochondrial and chloroplast typing, a novel real-time PCR quantitation method was developed and validated to accurately estimate the quantity of the chloroplast DNA (cpDNA) using a

synthetic DNA standard. Moreover, a sequenced allelic ladder was also designed for accurate genotyping of the homopolymeric STR pentaplex.

And finally, as a proof of concept, a custom panel for MPS was designed to interrogate 12 *Cannabis*-specific STR loci by sequence. A simple workflow was designed to integrate the custom PCR multiplex into a workflow compatible with the Ion Plus Fragment Library Kit, Ion Chef, and Ion S5 system. For data sorting and sequence analysis, a custom configuration file was designed for STRait Razor v3 to parse and extract STR sequence data. The study resulted in a preliminary investigation of sequence variation for 12 autosomal STR loci in 16 *Cannabis* samples. Results revealed intra-repeat variation in eight loci where the nominal or size-based allele was identical, but variances were discovered by sequence. In addition, full concordance was observed between the MPS and capillary electrophoresis (CE) data. Although the panel was not fully optimized and only a small number of samples were evaluated, this study demonstrated that more informative STR typing of *Cannabis* samples can successfully be performed on a MPS platform.

KEY WORDS: *Cannabis sativa*, Forensic DNA, Forensic plant science, Massively parallel sequencing, Organelle DNA, Short tandem repeats

ACKNOWLEDGEMENTS

This dissertation was partially funded by a Graduate Research Fellowship Award #2015-R2-CX-0030 (National Institute of Justice, Office of Justice Programs, U.S. Department of Justice). The opinions, findings, conclusions, or recommendations expressed in this presentation are those of the authors and do not necessarily reflect those of the National Institute of Justice.

The author would like to thank all staff and personnel at the U.S. Customs and Border Protection LSSD Southwest Regional Science Center for their great assistance and help with this project. The author would also like to thank Roberta Marriot and Alejandra Figueroa for their kind donation of marijuana DNA extracts. Lastly, the authors greatly appreciate Haleigh Agot for her assistance with the chloroplast quantitation method.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ABBREVIATIONS

AFLP	Amplified Fragment Length Polymorphism
ANOVA	Analysis of variance
bp	Base pairs
CBD	Cannabidiol
CE	Capillary electrophoresis
CI	Confidence interval
cpDNA	Chloroplast DNA
%CV	Percent coefficient of variation
DNA	Deoxyribonucleic acid
F_{ST}	Fixation index
He	Expected heterozygosity
Ho	Observed heterozygosity
HID	Human identification
HWE	Hardy-Weinberg equilibrium
InDel	Insertion/deletion polymorphism
IRMS	Isotope ratio mass spectrometry
ISFG	International Society of Forensic Genetics
ISSRs	Inter Simple Sequence Repeats
LD	Linkage disequilibrium
mtDNA	Mitochondrial DNA
MPS	Massively parallel sequencing
NGS	Next generation sequencing
NIST	National Institute of Standards and Technology
NJ	Neighbor Joining
OSAC	Organization of Scientific Area of Committees
PCA	Principal component analysis
PCR	Polymerase chain reaction
PD	Power of discrimination
qPCR	Quantitative PCR

RAPD	Random Amplified Polymorphic DNA
RMP	Random match probability
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SNP	Single nucleotide polymorphism
STR	Short tandem repeat
SWGDAM	Scientific Working Group on DNA Analysis Methods
THC	Tetrahydrocannabinol
tRNA	Transfer ribonucleic acid
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

GLOSSARY

Allele	Versions of a gene or other locus.
Fixation	Genetic drift can result in the fixation (~100% frequency) of one allele due to loss of other alleles.
Genetic drift	A change in the allele frequencies over time due to chance (sampling error). Genetic drift affects small, isolated populations at a higher rate.
Hardy-Weinberg equilibrium	A theorem stating that genetic variation (allele and genotype frequencies) in a population will remain constant from generation to generation in the absence of disturbing forces. Assumptions for Hardy-Weinberg equilibrium to hold true include: random mating, a closed infinitely large population size, no mutation, no natural selection, and no genetic drift.
Linkage	Linkage refers loci that are linked and inherited together. Loci that are physically close on a chromosome tend to be inherited together and are not independent. The greater the physical separation of the loci, the less likely they are linked.
Linkage equilibrium	The random association of alleles from different loci in a population.
Linkage disequilibrium	The non-random association of alleles from different loci. Disequilibrium is observed when the association frequency between two alleles is higher or lower than expected if the loci were independent from one another and associated randomly.

Polymerase chain reaction	A technique used to amplify or make copies of a specific DNA region. A thermostable *Taq* polymerase is used for the replication process along with primers designed to amplify a specific target. The PCR cycle consists of a series of temperature changes that allows for many copies of the target region to be produced. This cycling is repeated several times to generate millions of copies of the target region(s).
Unweighted Pair Group Method with Arithmetic Mean	A simple, agglomerative algorithm for phylogenetic tree construction based on a distance-based matrix. This algorithm assumes populations or taxa evolve or mutate at a constant rate.

CHAPTER I

Introduction

Botany

Cannabis sativa Linnaeus (*Cannabis sativa* L.) is an annual herb that is classified as an angiosperm or a flowering plant. It is dioecious, meaning there are distinct male and female flowers. This is rare for flowering plants as more than 90% of angiosperms are known to be hermaphrodites or monoecious [1]. *Cannabis* plants vary in height, with most between one and five m tall. The female plant has an organ containing eggs known as the pistil while the male plant includes a pollen-producing organ, the stamen. Male plants are taller and less robust than female plants. The phloem (bast) from the stalks of the plant is targeted for fiber while the flowering and leaf parts are preferred for drug use. There are three primary forms of drug-type *Cannabis*: marijuana, which is dried flowers and leaves, hashish, which consists of dried resin and compressed trichomes, and hash oil, which is a distilled form of hashish. Additionally, the seed and oilseed can be used as a source of food or nutritional supplement.

Genetics of *Cannabis*

Genome

Cannabis sativa has a diploid genome (2n=20) with nine pairs of autosomes and a pair of sex chromosomes [2]. The estimated haploid genome size of *C. sativa* for female plants is 818 Mb and 843 Mb for males [2, 3]. Completion of a draft genome (GCA_000230575.1) of a purple kush variety in 2011 revealed a transcriptome of approximately 30,000 genes [3]. Comparison of the purple kush transcriptome to the transcriptome of a hemp cultivar, finola, showed that many genes associated with the

cannabinoid synthesis pathway were more highly expressed in the drug variety, purple kush [3]. The chloroplast contains a circular, double-stranded genome that has been fully sequenced and mapped [4, 5]. Four annotated varieties are available on NCBI: carmagnola (KP274871), dagestani (KR779995), cheongsam (KR184827), and yoruba nigeria (KR363961). The chloroplast genome is AT-rich (63%) and 153,871 bp in length [4]. There are 127 genes including 83 protein-coding genes, four unique ribosomal RNAs (rRNAs), and 37 transfer RNAs (tRNAs) [4, 5]. The mitochondrial genome of two hemp varieties, carmagnola (KR059940.1) and Chinese hemp (KU310670), has been fully mapped and annotated [6]. The mitochondrial genome is 415,499 bp in length and contains 54 genes (38 protein coding, 15 tRNA, and three rRNA) [6].

Sex determination

Sex determination of *Cannabis* is an important trait to determine for agricultural and drug production purposes. Female plants are more desired due to their higher content of cannabinoids [7]. Additionally, the female plant is more robust and stable for fiber production.

Unlike mammals, the Y-chromosome is larger than the X-chromosome in *Cannabis* [2]. The Y-chromosome is reported to be essential for pollen development [8]. In most dioecious plants, sex determination seems to be related to the ratio of X- and autosomal chromosomes [8, 9]. Cytological studies have revealed that the long arm of the Y-chromosome contains several copies of retrotransposon elements believed to contribute to the evolutionary differentiation of sex in *Cannabis* [2]. Several, male-associated DNA sequences in *C. sativa* (MADC) have been identified and studied [10-16]. Sakamoto et al. described 729 bp fragment, MADC1, obtained from Random Amplified Polymorphic

DNA (RAPD) analysis [10]. Additional MADCs have been described from RAPD analysis: MADC2 [12], MADC3 [14], and MADC4 [14]. RAPD analysis has also been used to identify female specific markers [17, 18]. Furthermore, Flachowsky et al. developed an Amplified Fragment Length Polymorphism (AFLP) marker that could differentiate between dioecious male and female hemp [15]. More classical studies with progenies have led to the discovery and classification of male-associated markers that are present only in the Y-chromosome without the possibility of recombination [16]. Other markers were identified as being located on both the Y- and X- chromosome. These regions are considered to be pseudoautosomal markers due to the recombination that occurs between the X- and Y- chromosome at this location [16].

Chemotype

Determination of a plant's chemotype is also an essential factor for breeding purposes. Tetrahydrocannabinolic acid synthase (THCAS) is responsible for the production of the psychoactive compound, delta-9-tetrahydrocannabinol (THC) [19]. Kojoma et al. sequenced the THCA synthase genes and observed 63 nucleotide substitutions differentiating drug-type and fiber-type *Cannabis* [20]. In 2003, De Meijer et al. proposed a genetic model to explain the inheritance of *Cannabis* chemotypes [21]. The model postulated that there is a single co-dominant locus B that determines the ratio of THC to cannabidiol (CBD) production. Other studies mirrored this single-gene model of chemotype inheritance and sought to identify polymorphisms within genes coding for cannabinoid production [22-24]. Recent sequenced-based research has highlighted that drug and fiber type *Cannabis* differ across the whole genome and not just in cannabinoid

production genes [25, 26]. Soorni et al. identified loci outside the cannabinoid pathway to be targeted in future studies evaluating the genetics associated with chemotype [26].

Taxonomy

Taxonomy refers to the classification and nomenclature of a species. Classification is the identification and categorization of an organism while nomenclature describes the name of an organism. There has long been a debate over the taxonomy of marijuana, and still, there is a lack of agreement on a practical and workable nomenclature for *Cannabis* [27, 28]. The central point of contention is whether the genus *Cannabis* is polytypic or monotypic. In 1753, Linnaeus first named and described a single species of hemp, *Cannabis sativa* L., in his text *Species Plantarum* [29]. Later in 1785, Lamarck coined the term *C. indica* for cannabis plants he found in India, Southeast Asia, and South Africa. Lamarck noted that *C. indica* was distinctly different from the European hemp species, *C. sativa,* in eight different morphological characteristics namely the different plant heights and leaf shapes [30]. Lamarck concluded that marijuana strains were polymorphic and could be differentiated into species based on chemotype, ecotype, and leaf morphology [30]. In 1976, the formal taxonomy of *Cannabis* was assigned by Small and Cronquist [31]. They recognized that *Cannabis* was a monotypic species with two subspecies: *C. sativa* subsp. *sativa* and *C. sativa* subsp. *indica.* After studies and surveys, Small and Cronquist determined that the variations observed within the *Cannabis* genus were primarily due to man's cultivation and selection practices. Other studies have evaluated polymorphisms within alloenzymes and proposed that *Cannabis* is composed of three species: *Cannabis sativa, Cannabis indica,* and *Cannabis ruderalis* [32, 33]. Though there is a lack of agreement on a practical and workable nomenclature for cannabis, most botanists still

consider the genus *Cannabis* to be monotypic. This dissertation will work on the principle that cannabis is a single species with polymorphic characteristics. Additionally, *C. sativa* is referred to as "hemp" when used as a fiber and "marijuana" when utilized for its intoxicant properties.

C. sativa belongs to the Cannabaceae family which until recently only contained two genera: *Cannabis* and *Humulus* [34]. The Cannabaceae family now comprises ten genera and roughly 100 species [35-37]. However, *Cannabis* and *Humulus* are still the closest genera, forming a phylad. The complete taxonomic classification of *Cannabis* is displayed in Table 1.1.

Table 1.1. Current taxonomic classification of *Cannabis sativa* L.

Domain:	Eukayota (Eurkayotes)
Kingdom:	Plantae (plants)
Subkingdom:	Tracheobionta (vascular plants)
Superdivision:	Spermatophyta (seed plants)
Division:	Magnoliophyta (flowering plants)
Class:	Magnoliopsida (diotyledons)
Subclass:	Hamamelididae
Order:	Urticales
Family:	Cannabaceae
Genus:	*Cannabis*
Species:	*Cannabis sativa* L.

History

Cannabis sativa is a plant that is cultivated worldwide for its use as a fiber, medicine, or intoxicant. Although no precise origin has been identified, it has been widely speculated that *Cannabis sativa* originated in western or central Asia [27, 38]. Origin determination is difficult because *Cannabis* has been heavily transported for the last 6000 years and has established itself in several areas outside its indigenous location. It is known

that *Cannabis* has been intentionally grown and cultivated for the past 6000 years [39], but the earliest human use of *Cannabis* may have occurred as early as 10,000 BCE. However, this evidence embodies weak archeological evidence in the form of hemp strands in clay pots from tombs estimated to be as old as 10,000 BCE [40, 41]. Additionally, *Cannabis* may have been harvested 8500 years ago by the Chinese, most likely from the wild-plant and not a domesticated form [42]. Hemp was later introduced to western Asia, Egypt, and finally Europe between the years 1000 and 2000 BCE [43]. By 500 CE, cultivation in Europe was widespread [43]. With the era of exploration, hemp was first transported to South America in 1545 and to North America in 1606 [43]. For most of its recorded history, *Cannabis* (hemp) has been primarily used for its distinctive fiber properties including strength, durability, and water resistance [44, 45]. Additionally, *Cannabis* seeds have been used a source of food for humans and livestock for 3000 years in China [46].

Cannabis has been used for its medicinal properties in traditional Chinese, Indian, and Tibetan medicine [47-49]. The Chinese have exploited *Cannabis* for its analgesic effects dating back to 2700 BC [48]. Indeed, Jiang et al. documented a 2500-year-old gravesite in Xinjiang, China that contained high-THC *Cannabis* [50]. DNA typing of ribosomal and chloroplast *Cannabis* specific regions revealed an uncertain relationship to modern strains [51]. Evidence suggests *Cannabis* has been used for rituals and religious ceremonies in southern Asia, especially Afghanistan and India, even before written history [52]. For these ceremonies, high THC *Cannabis* was commonly prepared as hashish. Hashish is still a common form of *Cannabis* in Europe and Asia.

Chemistry

Cannabinoids

Cannabis contains more than 100 cannabinoids that belong to a class of terpenophenolic secondary metabolites, of which only a few are psychoactive [53-55]. In the living plant, the cannabinoids are in a carboxylic acid form which is decarboxylated into its neutral constituent when heated (e.g., smoked or cooked). Delta-9-Tetrahydrocannabinolic acid (Δ^9-THCA) is the precursor of the primary psychoactive agent, delta-9-tetrahydrocannabinol (Δ^9-THC or THC) (Fig. 1.1.) A THC concentration of 0.9% in the plant has been proposed as a minimum level for intoxication [28]. Another cannabinoid, Cannabidiol (CBD), is the primary cannabinoid in fiber type *Cannabis* and can serve as a potentiator or antagonist to THC [28]. Due to the antagonistic relationship between THC and CBD, the differentiation of drug-type versus fiber-type *Cannabis* is dependant upon the concentrations of both THC and CBD.

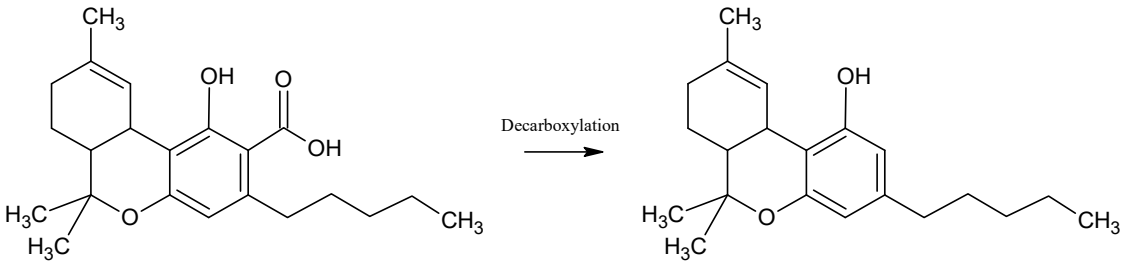


Fig. 1.1. Chemical diagram of the carboxylated form (THCA) and decarboxylated form (THC)

Chemotypes

Cannabis can be divided into three main chemotypes on the basis of chemical profiling: chemotype I (drug-type) which contains THC in concentrations greater than 0.5% and cannabidiol (CBD) in concentrations less than 0.5%, chemotype II (intermediate type), with CBD as the major cannabinoid but with THC also present at various concentrations, and chemotype III (fiber-type or hemp), with CBD as the major cannabinoid combined with exceptionally low THC content [56].

Cultivation

Growing conditions

Cannabis sativa is an annual plant that can be cultivated both indoors and outdoors. Under outdoor conditions, the plant's life cycle takes approximately five to seven months. Successful outdoor cultivation is affected by many factors such as wind, rain, humidity, and sunlight. Indoor cultivation allows for more control of the life cycle of the plant, but the environment must be strictly controlled to ensure optimum growth. *Cannabis* requires an optimum quantity and quality of light for photosynthesis. Studies have shown that *Cannabis sativa* benefits from high Photosynthetic Photon Flux Density (PPFD) [57]. Also, photosynthesis is dependent upon temperature (25-30 °C), humidity (75%), and levels of carbon dioxide (1600 ppm) [57, 58].

Propagation

Cannabis is commonly propagated through seeds or vegetative cuttings. Seeds are planted in moist, aerated soil and germination occurs in two to seven days. Although seed propagation is a conventional technique, it is difficult to maintain quality and THC/CBD levels. When growing from seeds, a significant portion of the plants will be male plants

which will result in lower levels of the desired cannabinoid (THC or CBD). As such, the number of male plants is strictly controlled in the production of marijuana for intoxicant purposes. Identical genotypes occur due to cultivation via vegetative propagation or clonal propagation instead of sexual reproduction. Most growers and dispensaries prefer clonal propagation to maintain consistent quality and potency of their products. For clonal propagation, clippings from the desired female plants, which contain higher THC levels, are directly rooted in the soil or a liquid medium (hydroponics). Clonal propagation results in genetically identical plants, while seed propagation results in plants with a unique genetic makeup [59]. In the case of clonal propagation, DNA typing will allow direct linkage of cases to a common grower/distributor.

Polyploidy

Polyploidy refers to an organism that contains more than two sets of chromosomes. A plant that contains two sets of chromosomes is known as a diploid (2x), whereas one with three sets would be a triploid (3x), four sets a tetraploid (4x), etc. [60]. Polyploidy has not been shown to occur in *Cannabis* naturally; however, it may be artificially induced with colchicine treatments [61]. Colchicine is a poisonous compound derived from the roots of certain colchicum species that prevents segregation of chromosomes and cell wall formation. This inhibition will lead to larger daughter cells with multiple chromosome sets. Induction of polyploidy may serve as a powerful tool for improving the characteristics desired within each plant. Two studies have reported the occurrence of polyploidy observed during short tandem repeat (STR) analysis of drug type *Cannabis* [62, 63].

Legal Status

History

The legal status of *Cannabis* varies worldwide; however, possession is still illegal in most countries. During the 1800s and early 1900s, *Cannabis* was dispensed by physicians for various medicinal purposes. In the 1930s, there was a widespread prohibition of *Cannabis* worldwide. The Marijuana Act of 1937 prohibited possession of marijuana except for medicinal or industrial uses [64]. Though legal for medicinal purposes, several reporting requirements were implemented by the Act that effectively discouraged physicians from prescribing *Cannabis*. In the Netherlands, the Opium Law of 1976 allows consumers to purchase *Cannabis* in legal coffee shops [65]. In 2013, Uruguay was the first country to legalize *Cannabis* [66].

The Controlled Substances Act

The use or possession of *Cannabis* (marijuana) is illegal under federal law in the United States as per the Controlled Substances Act (CSA) of 1970 [67]. Under this act, marijuana is recognized as a Schedule I substance meaning that it has a high potential for abuse, no accepted safety for use, and no accepted medical use [68]. Other drugs in Schedule I include heroin, psilocybin, peyote, and D-Lysergic acid diethylamide (LSD) [68]. Cannabidiol is a Schedule I substance as a derivative of marijuana (21 USC 802). This is true for all other cannabinoids with THC specifically listed separately. Currently, three cannabinoid drugs (Marinol®, Syndros®, and Cesamet®) can be legally prescribed to patients by federal law.

State laws

There are conflicting laws at the state level with various degrees of *Cannabis* use allowed (Fig. 1.2.). In 1996, California became the first state to legalize *Cannabis* for medical use [69]. Currently, 29 states and the District of Columbia have laws allowing for various levels of medicinal marijuana [69, 70]. In addition, recreational use of *Cannabis* for persons over 21 is currently allowed in eight of the 29 states: Alaska, California, Colorado, Maine, Massachusetts, Nevada, Oregon, and Washington, as well as the District of Columbia. Though the federal law is supreme in the land, the Cole Memorandum in 2013 provided some protection against the enforcement of the federal law. In January 2018, Attorney General Jeff Sessions rescinded this memorandum making the future of federal *Cannabis* prosecutions unknown.

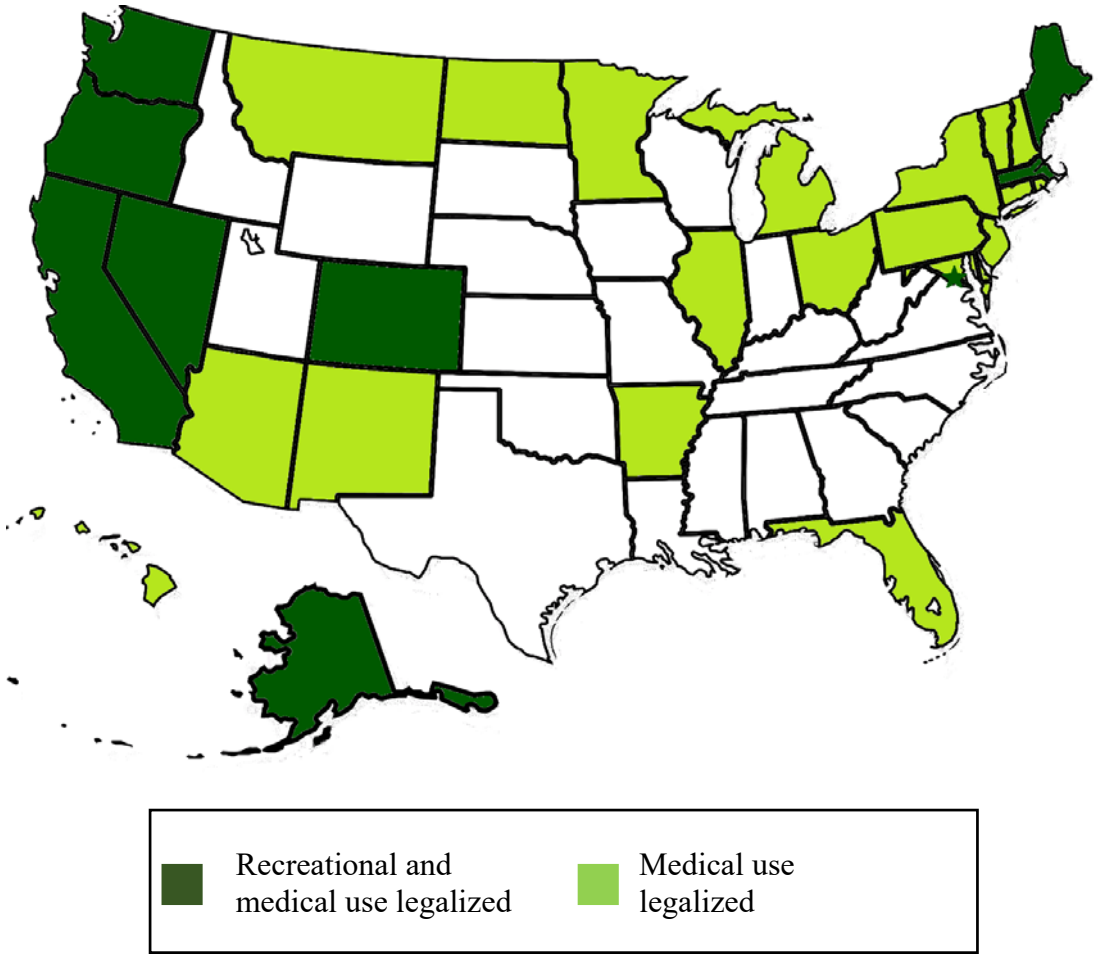


Fig. 1.2. Map of the vary levels of cannabis legalization across the United States

Research

As a Schedule I substance, research on *Cannabis* plant material can only be performed through the National Institute on Drug Abuse (NIDA). Historically, only the University of Mississippi has been given a contract to cultivate *Cannabis* for research. The DEA has announced that it will add additional *Cannabis* cultivation sources for research and development of FDA-approved products [71].

Forensic identification of *Cannabis* material

When identifying *Cannabis* (marijuana) for prosecutorial purposes, regulations require the confirmation of THC via gas chromatography-mass spectroscopy (GCMS), the confirmation of the presence of cystolithic hairs, and a positive Duquenois-Levine color test [72, 73]. However, some evidence such as wash up samples may be compromised for identification via morphology. Additionally, more than 80 different plant species are reported to contain cystolithic hairs nearly identical to those of *Cannabis sativa* [73]. While the chemical identification of *Cannabis* may be sufficient at prosecuting an individual for possession of marijuana results, do not provide meaningful intelligence about the origin or provide individualization of the plant.

Individualization/Origin determination

Many methods have been proposed to individualize and determine the origin of a marijuana sample. These methods include but are not limited to palynology [74], chemical profiling [75], isotope ratio mass spectrometry (IRMS) [76, 77], and DNA analysis [78, 79]. Palynology or the study of pollen is one field that is used to predict the origin of *Cannabis* based on the type of pollen found in the sample [74]. Depending on the region the *Cannabis* is grown in, the native plants seen may vary. These native plants will contain different pollen types that can be differentiated using a scanning electron microscope (SEM). Although useful, palynology is a field that is time consuming, expert-based, and not easily integrated into a forensic laboratory. Chemical profiling is a more common technique that evaluates different ratios of both major and minor cannabinoids in a *Cannabis* plant [75]. The ratios of these compounds may vary depending on the environment in which a plant was cultivated. However, storage and time since removal can

also affect the ratios making results inconsistent and unreliable [75]. IRMS is another technique that has shown promise in the association of *Cannabis* to a source [76]. IRMS relies on the stable isotope ratios of carbon and nitrogen that are intrinsic to a region. While growing, both carbon and nitrogen are incorporated into the marijuana plants. The isotope ratios vary region to region and can be observed in the evaluation of plants. Studies in Brazil have shown that IRMS has a poor power of discrimination in regions with overlapping isotope patterns [77]. DNA has been shown to provide higher resolution to the individualization of *Cannabis* plants compared to these alternate chemical and morphological techniques [79, 80].

Nuclear DNA Identification

Ribosomal DNA (rDNA)

Early genomic-based studies focused on the botanical identification of *Cannabis*. Though several techniques had been employed to identify *Cannabis*, they were susceptible to false positives. Several plant species may contain cystolithic hairs and not all *Cannabis* contains THC. Though Random Amplified Polymorphic DNA (RAPD) identification could be used to compare cultivars, it was not suitable for the identification of the species in question.

The nuclear ribosome is composed of three subunits (18S, 5.8S, and 25S) [81, 82]. (Fig. 1.3.) Hundreds to thousands of copies of rDNA are found within the nucleus [81]. Each copy of rDNA codes for the three subunits, and each subunit is separated by an Internal Transcribed Spacer (ITS1 or ITS2) while each ribosomal unit is separated by an Inter Genic Spacer (IGS) [81, 82]. (Fig. 1.3.)

Fig. 1.3. Diagram of the structure of eukaryotic rDNA

The genes coding for the subunits are largely conserved, but the non-coding regions (ITS1, ITS2, and IGS) show high variability between species [82-84] Through two studies, Gigliano demonstrated that ITS2 could distinguish *Cannabis* from *Humulus* [83, 84]. These studies identified variants using a sequence-based assay [83] and Restriction Fragment Length Polymorphism (RFLP) fragments [84]. While both could distinguish *C. sativa* from other members of the Cannabaceae family, sequencing yielded a higher power of discrimination. Gigliano also evaluated the utility of sequencing the ITS1 region for identifying *Cannabis* [85]. Results revealed that ITS1 was also suitable for correctly identifying a *Cannabis* sample. Through the use of restriction site mapping, the IGS region has also been shown to be highly variable between *C. sativa, H. lupulus,* and *H. japonicus.* [82].

Random Amplified Polymorphic DNA (RAPD)

Law enforcement became interested in being able to compare seizures to make associations and origin determinations. Traditional methods like gas chromatography and High-Performance Liquid Chromatography (HPLC) did not provide enough individualizing characteristics for distinctions to be made. DNA, being a stable marker with variability across samples, could serve as a tool to distinguish different seizures. The technique of RAPD had previously been used in other plants to study phylogenetic relationships [86, 87]. RAPD allows for random sampling across the whole genome with no prior sequence knowledge necessary. Primers for amplification can be universally

designed for all eukaryotes with polymorphisms detected based on the presence or absence of bands (Fig. 1.4.).

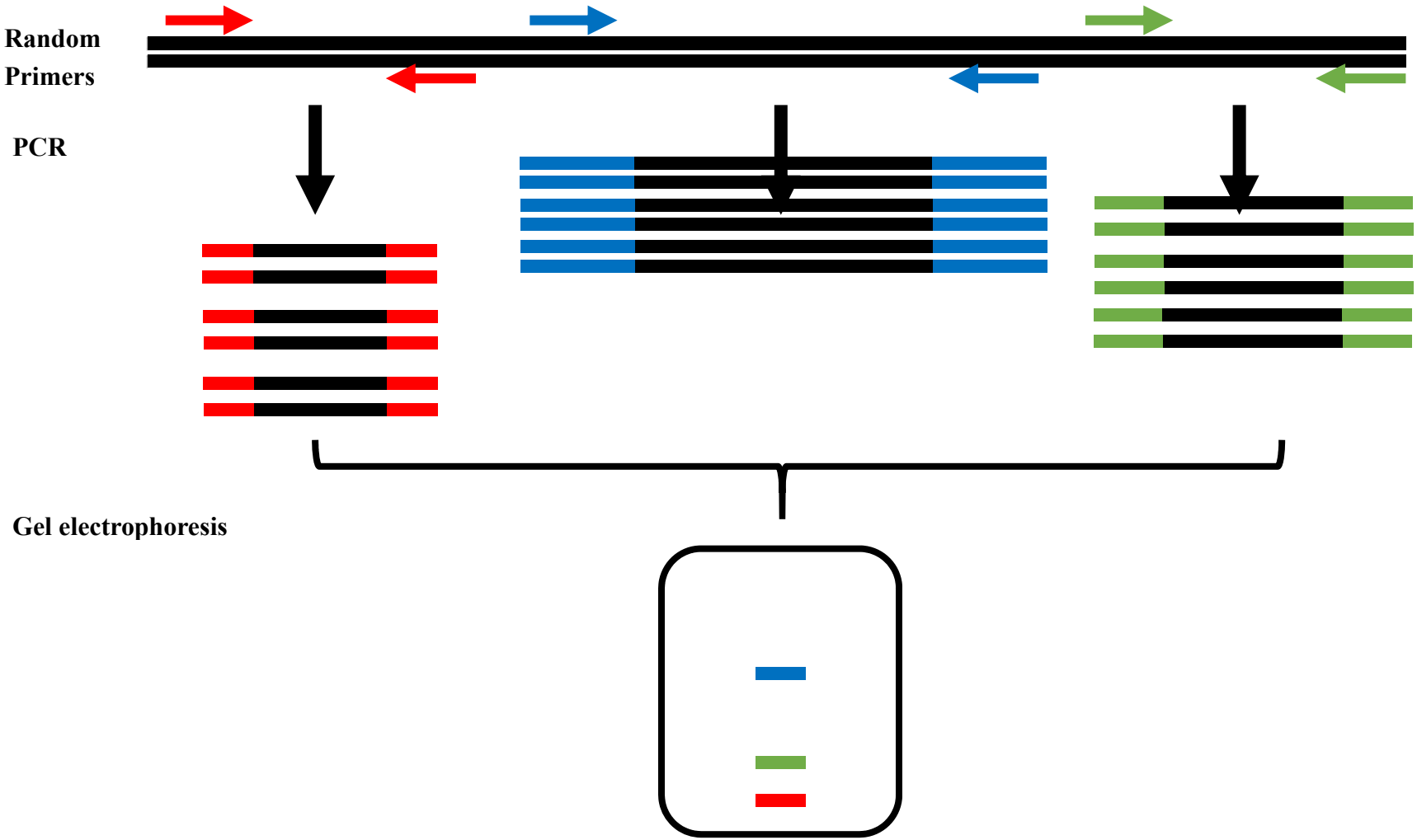


Fig. 1.4. Several sets of random primers are added to a PCR mix. PCR is performed, and several copies of fragments will amplify. The variable fragment length is then visualized via gel electrophoresis

In 1995, Gillan et al. demonstrated that RAPD could provide a higher discriminating power compared to HPLC [88]. Samples indistinguishable by HPLC could be differentiated using the RAPD technique with only three primers. Jagadish et al. further showed that RAPD could cluster samples from similar biogeographical areas while *H. lupulus* formed a separate cluster [89]. Faeti et al. demonstrated the ability to access variability amongst 13 hemp cultivars using ten random primers with a high level of polymorphism observed [90]. As Jagadish et al. observed, a grouping of cultivars was correlated to the geographical origin. In 1998, Shirota et al. demonstrated the capability of both and RAPD and RFLP to distinguish different chemotypes (fiber vs. drug) of *C. sativa* [91]. Other studies have further shown the utility of RAPD in individualizing marijuana samples [90, 92, 93]. Forapani et al. suggested that differentiation across all hemp varieties was possible using RAPD after evaluating six hemp varieties [92]. More recently, Pinarkara et al. successfully used RAPD analysis to distinguish samples based on geographical areas within Turkey [93]. Although RAPD is inexpensive and yields a moderate power of discrimination, the technique suffers from poor reproducibility and difficulty with interpretation.

Inter-simple sequence repeats (ISSRs)

ISSRs anneal directly to simple sequence repeats. Information on sequence variation is not necessary as the primers anchor to the simple repeats such as $(CA)_n$ (Fig. 1.5.).

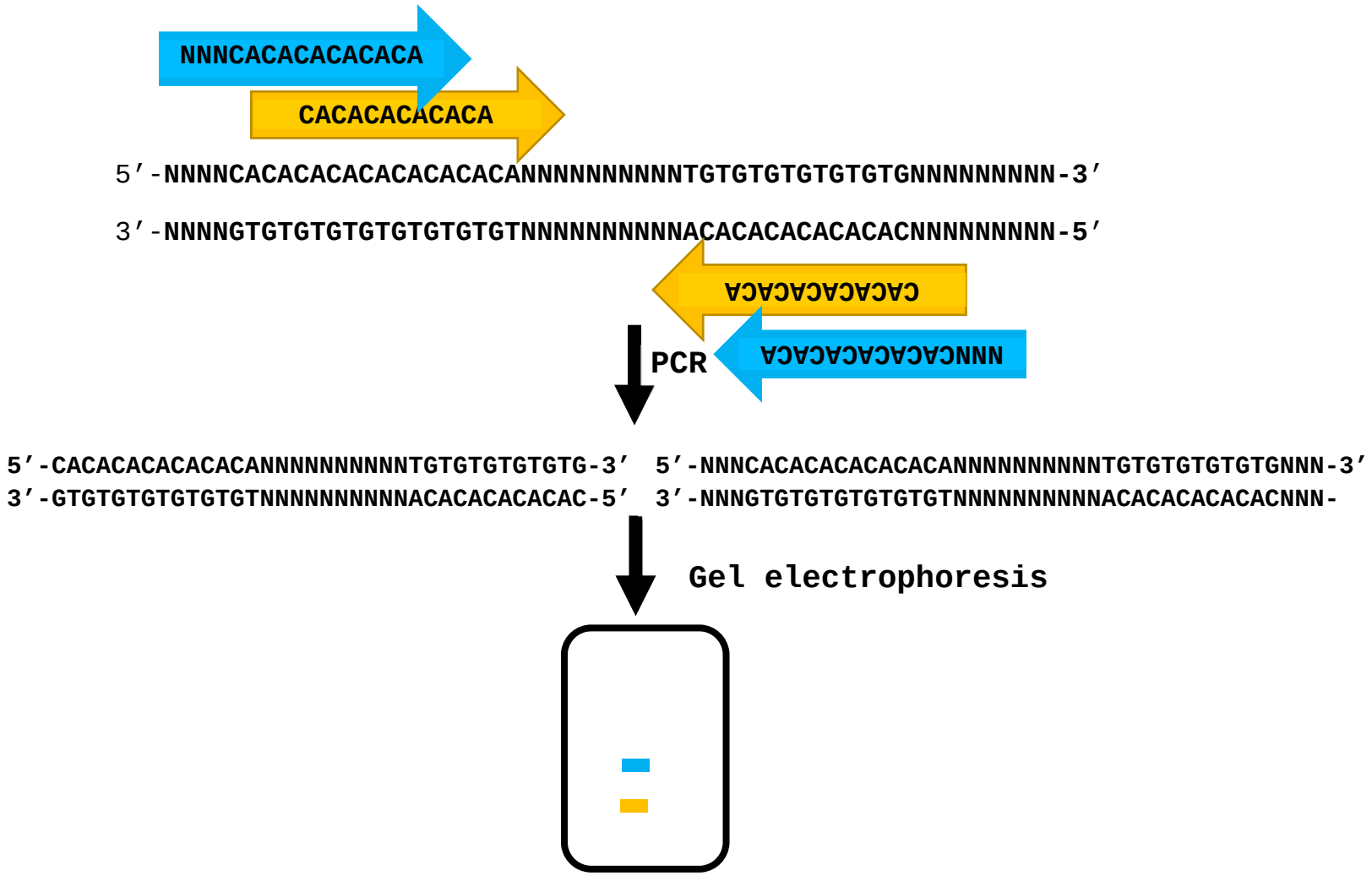


Fig. 1.5. PCR amplification using ISSR primers (anchors to simple sequence repeats) followed by visualization of fragments

Several groups showed the utility of ISSRs in estimating the genetic difference among several samples of *C. sativa* [94, 95]. Kojoma et al. demonstrated that ISSRs generate specific band patterns amongst nine different samples originating from three distinct hemp strains [94]. Hakki et al. established that ISSRs could distinguish both between and within drug and fiber types with the use of 18 primers [95]. Principal coordinate analysis (PCoA) was used to statistically visualize the drug and hemp types. Hakki also noted that the hemp samples showed higher variability as compared to the drug type samples. Pinarkara et al., demonstrated that ISSRs provide a slightly higher discriminating power compared to RAPD [93]. Recent studies have also evaluated the use of ISSRs to assess both inter- and intra- species relationships [96, 97]. As with RAPD typing, ISSRs encounter reproducibility problems and are not ideal for interpretation and comparison purposes.

Amplified fragment length polymorphisms (AFLP)

AFLP is a PCR based tool that has been used in genetic research and DNA fingerprinting [98, 99]. AFLP was developed in the early 1990s by KeyGene (Wageningen, Netherlands) and combines the techniques of RFLP and PCR [100, 101]. Briefly, restriction enzymes are used for digestion, oligonucleotide adapters are ligated to the digested products, and selective amplification via PCR is performed. Selective amplification is performed through primer design. Primers are designed to be complementary to the adapter sequence, restriction site sequence, and part of the restriction fragment. The amplified products are visualized via capillary electrophoresis, and scoring is performed based on the presence or absence of a polymorphism (Fig. 1.6.).

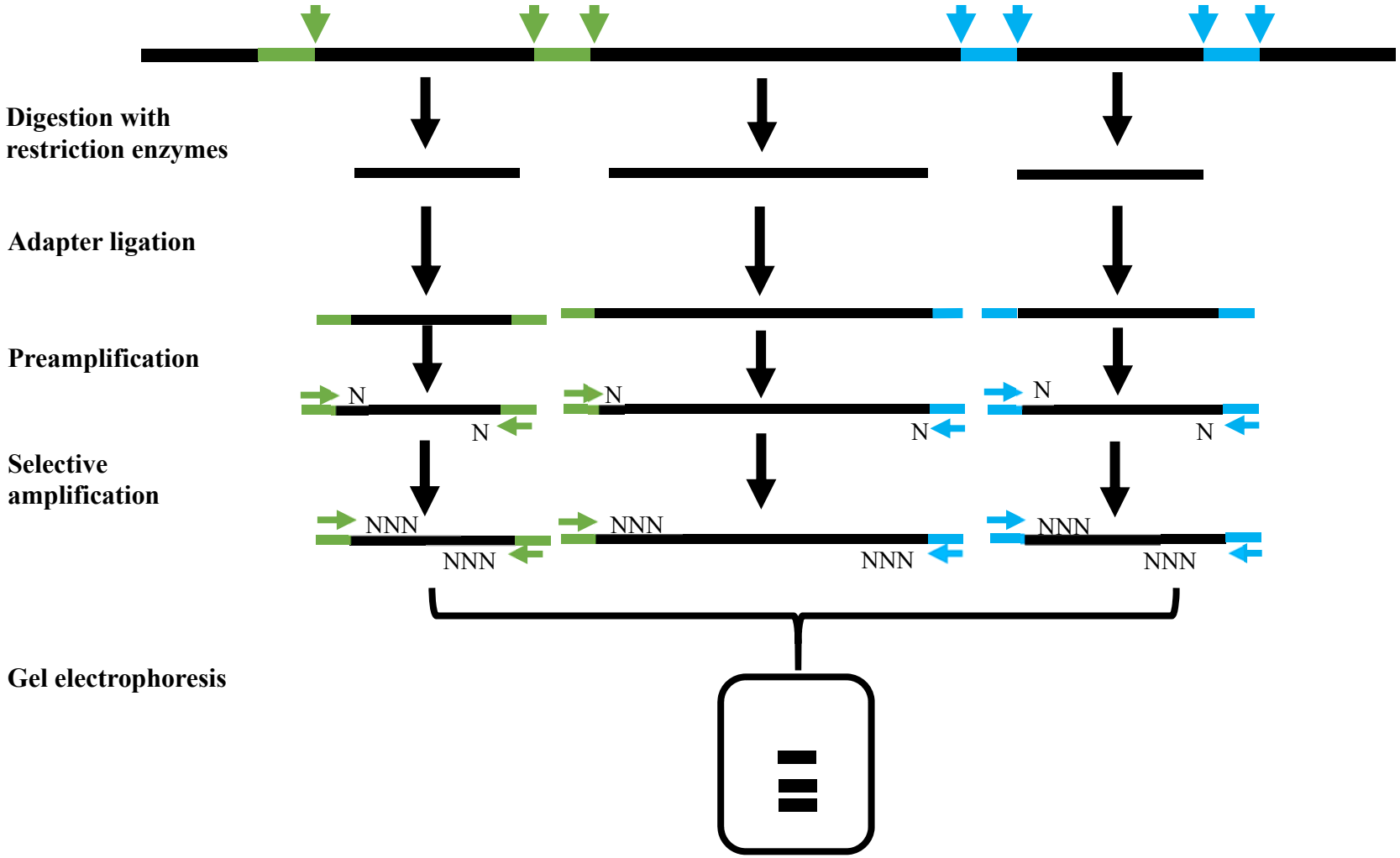


Fig. 1.6. Diagram of the AFLP procedure. The DNA fragment is digested with restriction enzymes and a series of amplification steps are performed to yield an AFLP profile

Historically, AFLP has been used as a technique for evaluating the genetic structure of *Cannabis*. In 2003, Coyle et al. evaluated AFLP patterns from American marijuana seizures [102]. Results demonstrated that AFLP profiles of marijuana could be generated from 100 mg of starting material. AFLP profiles were able to distinguish between individuals even with a single primer pair, and importantly, clones yielded identical AFLP profiles. Datwyler and Weiblen established that AFLP could be used as a tool to distinguish drug type *Cannabis* from hemp [103]. AFLP has also been used to evaluate the extent of genetic diversity of hemp in China [104]. Flachowksky et al. used AFLP to study the dioecious nature of *Cannabis* and was able to identify a male-specific AFLP band but no female-specific AFLP band [15]. Peil et al. studied male and female progenies of a single genetic cross using AFLP technology and observed a pseudoautosomal region on the sex chromosomes of *Cannabis*, which would allow for recombination events to take place between the X and Y chromosome [16].

There are advantages and disadvantages to using AFLP for DNA fingerprinting. Advantages include the relative abundance of fragment length polymorphisms (by restriction enzyme) in the genome, and no prior sequence knowledge is necessary for their design. However, AFLP has several disadvantages from a forensic standpoint. These fragment length polymorphisms may not be randomly distributed in the genome, instead clustering in certain genomic regions such as centromeres. Purified and high molecular weight DNA is needed for input. Lastly, the technique targets markers that are primarily dominant and thus, the resulting bands may not be independent of one another [105].

Short tandem repeats (STRs)

Short tandem repeat markers or microsatellites are defined as DNA sequences (two to seven bases) that are repeated in a tandem manner (e.g. (GAT)(GAT)(GAT)(GAT)) [106]. Short Tandem Repeat (STR) markers or microsatellites are the gold standard for human identification, and as such research has primarily focused on the development of STR panels to identify marijuana plants. STRs have the distinct advantages of codominance, reproducibility, multiplexing capability, and high power of discrimination [107].

In 2003, several *Cannabis* STRs were identified [62, 108, 109]. Gilmore and Peakall designed 15 primer pairs to isolate 15 microsatellite markers: six dinucleotide (ANUCS201, ANUCS202, ANUCS203, ANUCS204, ANUCS205, ANUCS206), eight trinucleotide (ANUCS301, ANUCS302, ANUCS303, ANUCS304, ANUCS305, ANUCS306, ANUCS307, ANUCS308), and one pentanucleotide (ANUCS501) [108]. Preliminary research revealed that all 15 microsatellites were reliably amplified and were hypervariable [108]. Alghanim and Almirall identified an additional 11 microsatellites: three dinucleotides (C08-CANN2, H11-CANN1, H09-CANN2) and eight trinucleotides (C11-CANN1, B01-CANN1, D02-CANN2, B02-CANN2, E07-CANN1, B05-CANN1, D02-CANN1, H06-CANN2) [109]. All 11 microsatellites were found to be useful in evaluating the genetic relatedness of seized *Cannabis* material [109]. Hsieh et al. identified one highly polymorphic hexanucleotide marker, CS1, with repeat numbers ranging from three to 40 [62]. CS1 was shown to be *Cannabis*-specific with no cross-reactivity observed when 20 species were tested including *Nicotiana tabacum* and *Humulus japonicus.*

Several studies have been performed evaluating the utility of these markers in a forensic setting. Gilmore et al. evaluated five (ANUCS201, ANUCS202, ANUCS301, ANUCS302, ANUCS303) of the original 15 microsatellites and demonstrated that the microsatellites were hypervariable and could prove promising in determining the geographical origin and classifying samples as drug or fiber [110]. Next, Howard et al. developed a multiplex STR system with ten STR loci for the genetic identification of *C. sativa* [78]. A combination of ten microsatellites originally described by Gilmore and Peakall [108] and Alghanim and Almirall [109] was used for this STR system. The ten microsatellites were amplified across four separate multiplexes, and a developmental validation was performed according to SWGDAM guidelines. Following validation, Howard et al. created an STR database for marijuana seizures in Australia [80]. Howard et al. noted the presence of identical genotypes in the marijuana seizures in the Australian STR database and statistical analysis showed that these identical genotypes were a result of clonal propagation rather than poor genetic resolution of the STR markers [80].

Mendoza et al. developed a multiplex with six previously described loci (ANUCS303, ANUCS305, E07-CANN1, D02-CANN1, H06-CANN2) amplified in one reaction [111]. The multiplex was able to differentiate 98 *Cannabis* samples with a calculated probability of finding the same genome in an unrelated population to be one in 9090. Although the multiplex was sufficient for individualizing samples, it was unable to differentiate between drug type and fiber type. Allgeier et al. used collection cards to start a US DNA database of marijuana samples using the highly polymorphic marker, CS1 [112]. Allgeier demonstrated the validity and feasibility of using DNA collection cards in the field to preserve *Cannabis* DNA for future analysis. Samples included fresh marijuana

leaves, dried material, and hashish. DNA typing success was variable depending on sample type with fresh samples yielding full profiles, dried materials generating partial profiles, and hashish producing no profiles. Results demonstrated that DNA collection cards could be used for marijuana databasing purposes when using fresh plant material. Due to its species specificity, Shirley et al. demonstrated that the CS1 marker may also be used to identify *Cannabis* seed [113]. Seeds from the same strain exhibited different genotypes while showing overall genetic similarities through shared alleles. Most importantly, due to its species specificity, this technique allows for the identification of *Cannabis* without having to grow the plant.

In 2012, Köhnemann et al., developed and validated a 15-locus STR multiplex consisting of previously described markers (D02-CANN1, C11-CANN1, H09-CANN2, B01, CANN1, E07-CANN1, ANUCS305, ANUCS308, B05-CANN1, H06-CANN2, ANUCS501, CS1, ANUCS302, B02-CANN2, ANUCS501) [114]. Validation studies for the multiplex included sensitivity, specificity, and reproducibility. Köhnemann et al. also found identical DNA profiles presumably from clonally propagated plants. In addition, polyploidy (3 or more alleles) was detected in five STR markers. Both polyploidy and clonally propagated *Cannabis* greatly affects the allele frequency estimates. In efforts to standardize genotyping of *Cannabis sativa,* Valverde et al. proposed nomenclature for the 15 STRs previously described [115]. A total of 130 alleles were sequenced across the 15 loci, with sequence variations within the motif and flanking regions noted. SNPSTR (single nucleotide polymorphisms (SNPs) in flanking region and/or STR repeat motif) haplotypes were reported for all 15 loci and demonstrated an increase in power of discrimination across all loci. The nomenclature proposed followed standard international guidelines

[116-118]. This standardized nomenclature for alleles is imperative for generating a uniform allelic ladder.

Valverde et al. proposed nomenclature for seven novel STR markers [119]. In accordance with ISFG recommendations to use tetranucleotide markers, six (3735, 9043, 9269, 5159, 4910, 1528) were tetranucleotide repeats while one (nH09) was a trinucleotide repeat [120]. The draft genome published in 2011 [3] was used to search for new STR markers through the use of a tandem repeat search tool, Phobos 3.3.12 [121]. Initial screening revealed 16 STR markers with nine markers being discarded due to low polymorphism or a flanking region with too much variability for primer design. Nomenclature and SNPSTR haplotypes were reported for the remaining seven STR markers.

Using an extensive database of 1,324 samples, Dufresnes et al. genotyped *Cannabis* samples from hemp and drug cultivars [79]. Five multiplexes were used to genotype 13 STR markers previously described by Gilmore and Peakall [108] and Alghanim and Almirall. [109]. The study yielded a large resource describing the genetic signature of cultivars. However, the data is size based, and not allele based as no allelic ladder was used. Principal component analysis (PCA) and Bayesian clustering of genotypes revealed that the STR markers captured the genetic diversity of cultivars.

Soler et al. evaluated genetic variability of 20 cultivars of *C. sativa* var. *indica* and two cultivars of *C. sativa* var. *sativa* [122]. Variability was assessed with six previously identified dinucleotide markers, and results revealed 14 genetic clusters with individuals from the same cultivar generally clustering together. Importantly, *C. sativa* var. *sativa* was

statistically differentiated from *C. sativa* var. *indica*. High variation was observed within cultivars, and Soler noted that this variation could be exploited for breeding purposes.

Only a relatively small number of STRs have been reported in *Cannabis*. In efforts to develop more STR markers, Gao et al. used the genome and transcriptome published in 2011 as a means to search for microsatellites [123]. Gao identified SSRs from expressed sequence tags (ESTs). This is a quick and efficient way to identify STRs and may help elucidate certain agronomic traits as ESTs are linked to genes. Though this may not be ideal in a forensic identification setting, it is an effective way to map the genome. Potential STRs were detected from 32,324 sequences available on NCBI using the AutoSSR software [124]. Primers were then developed for 3,442 EST-SSRs based on the sequences of the flanking regions. Data revealed that one STR occurs for every 8.49 kb sequenced. Furthermore, results showed that trinucleotides (50.99% of markers) represented most of the tandem repeats, with AAG/CTT (17.96%) being the most frequently observed motifs. In contrast, Alghanim and Almirall observed dinucleotides to be the most common motif [109]. This difference may be due to the methods used in mining and developing the STRs. After random screening of EST-SSRs, only 56 loci were used to evaluate the genetic diversity and relatedness of 115 *Cannabis* (hemp) varieties. PCoA based on the 56 loci revealed that the EST-SSRs could separate the 115 varieties into four distinct groups based on geography: Northern China, Europe, Central China, and Southern China. This study represented the first large-scale development of STR markers in *Cannabis* and presented potential loci that could be used in future studies.

Genotyping by Sequencing (GBS)

GBS is an alternative to array-based screening approaches for SNPs and offers a way to compare samples in the absence of a reference genome. Briefly, high molecular weight DNA is digested with a specific restriction enzyme, barcode adapters are ligated to the sticky ends at the cut site, barcoded fragments are amplified, and barcoded libraries are sequenced using massively parallel sequencing (MPS) strategies [125, 126]. With the advent of MPS platforms, GBS provides a low cost per samples and can compare samples in the absence of a reference genome. Additionally, GBS can be used to discover new STR markers.

In a study by Sawler et al., 81 drug-type and 43 fiber-type samples were genotyped using GBS [25]. The drug-type samples represented a broad range of commercial strains with a reported percentage of ancestry (% *C. sativa* var. *sativa* and % *C. sativa* var. *indica*) while the fiber-type samples embodied a diverse group of samples from European and Asian accessions. A total of 14,031 SNPs were reported. Principle component analysis (PCA) demonstrated that the SNPs separated the groups into drug-type and hemp-type samples. This distinction was confirmed using Bayesian clustering with the fastSTRUCTURE software with k=2 ancestral populations [127]. While previous studies have shown that marijuana and hemp differ in their ability to synthesize cannabinoids, specifically THC, this study demonstrated that there is a difference at the genome-wide level between drug-type and fiber-type *Cannabis* [21]. Fiber-type *Cannabis* or hemp is usually classified as *C.* sativa; however, Sawler found genetic evidence that hemp was more closely related to *C. sativa* var. *indica* ancestries [25]. This finding is consistent with Hilling's allozyme study [32] and a study using RAPD markers [128]. Additionally,

Sawler noted that reported strain ancestries often do not reflect a molecular or genetic structure [25].

Soorini et al. further investigated the use of GBS in *Cannabis* genetic mapping [26]. Samples were comprised of 70 samples from 35 locations in Iran, two samples from Afghanistan, and 26 accessions from the Center for Genetic Resources (CGN) in The Netherlands and Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) in Germany. CGN accessions were comprised of fiber germplasms, and IPK represented hemp germplasms. A total of 98 *Cannabis* samples were genotyped using an Illumina HiSeq with 24,710 SNPs identified after quality filtering [26]. Soorini observed that majority of SNPs (62.7%) were transitions (A/G or C/T). This ratio of transitions (62.7%) to transversions (37.3%) has been observed in other species including maize [129], *Escherichia coli* [130], and oil palm [131]. Soorini combined data with Sawler and found 13,325 SNPs across 209 samples. Using Nei's genetic distance [132], it was revealed that the CGN/IPK accessions and hemp samples (Sawler et al.) clustered together with a genetic distance of 0.00496 while the Iran samples most closely clustered with the drug-type samples (Sawler et al.) with a genetic distance of 0.00921. Structure analysis of the four populations (Iran, CGN/IPK, drug-type, and hemp) using discriminant analysis of principal components (DAPC) demonstrated that each population could be defined in a unique cluster [133, 134]. Furthermore, PCA and fastSTRUCTURE [127] analysis revealed that the Iran samples formed two distinct clusters separated based on location (east or west) in Iran.

Massively parallel sequencing (MPS)

Pre-MPS

Until recently (2004), Sanger Sequencing has been the prominent form/gold standard of sequencing [135]. Sanger Sequencing can only sequence one target (up to 1000 bp) and one sample per reaction [136]. The technique relies on chain terminator dideoxynucleotides (ddNTPs) that contain a different fluorescent label representative of a specific nucleotide (A, T, C, G) [136-138]. These ddNTPs are analogous to deoxyribonucleotides (dNTPs) except that ddNTPs lack the 3' hydroxyl group that is necessary extension of the product [137]. A Sanger Sequencing reaction has a mix of dNTPs and ddNTPs that promote both extension and selective termination of the sequence resulting in extension products of different lengths with a fluorescently labeled nucleotide at the 3' end [137, 138]. The extension products are then separated by size via capillary electrophoresis, and the last base (ddNTP) of the extension product is called based on a characteristic wavelength of the excited fluorescent tag [139, 140]. These base calls are then lined up based on the size of the fragment resulting in the base-by-base sequence of the target.

Overview of MPS and its advantages

In contrast to the low throughput of Sanger Sequencing, massively parallel sequencing (MPS) can sequence hundreds of samples and targets in a simultaneously [135]. Additionally, MPS provides comprehensive coverage of genetic markers through both targeted and full genome sequencing [135]. MPS technology has been successfully used in the field of medicine, microbiology, environmental, and forensic sciences [141-144]. Currently, no targeted MPS workflows have been used for the genetic identification of *C.*

sativa. Analogous to HID typing, capillary electrophoresis (CE) of STR markers is the gold standard when genetically identifying marijuana for forensic or intelligence purposes. While STR by CE is a reliable and robust technique, only 25 to 30 loci are configurable across five dye channels [145]. Additionally, MPS has the potential to provide a deeper understanding (size-based and sequence-based) of polymorphisms, which in turn allows for greater power of discrimination compared to size- based CE-STR genotyping. There are two MPS platforms widely used in the field of forensic genetics. the Ion S5 System (Thermo Fisher Scientific) and the MiSeq FGx™ (Verogen).

Sequencing platforms

Semi-conductor sequencing

The Ion S5 System (Thermo Fisher Scientific) is a semi-conductor platform that sequences through the detection of pH change (release of hydrogen ion (H^+)) when a nucleotide is incorporated and the signal is translated to a base call [146]. More specifically, the semi-conductor chip detects this pH change because it registers a change in voltage [146].

Prior to sequencing, samples are preped via library preparation. Briefly, library preparation reuslts in targetted or random fragments that contain adapters and barcodes ligated to the ends of the template. The adpaters facilate clonal amplification and the barcodes (unique sequences) allow for multiple samples to be sequenced simultaneously. Next, these libraries are clonally amplified via emulsion PCR on an Ion Sphere™ particle, which is then flowed across a semi-conductor chip [147]. The templated chip is then sequenced using the semi-conductor chemsitry. During sequencing, one nucleotide is washed over the chip at a time, and if the nucleotide is incorporated in that particular well

then a change in voltage is occurred [146]. This sequencing is happening across millions of wells simulataneously.

Reversible-terminator

MiSeq FGx™ (Verogen) platform uses a method known as sequencing by synthesis which closely mimics traditional Sanger Sequencing and uses fluorescently labeled reversible terminator nucleotides [135, 148]. Template DNA molecules are generated either by PCR for targeted amplification or random fragmentation. These targets or fragments are prepared for sequencing through end repair and ligation of universal adapters [148]. Additionally, indices may be added to the template to facilitate sequencing of multiple samples simultaneously. The adapter-fragments are ligated on a flow cell where clusters are generated through bridge amplification. The sequencing occurs in three stages: chain extension via DNA polymerase and the four reversible nucleotide terminator, washing of unincorporated nucleotides and imaging, and lastly, the dye and terminator group are cleaved, and the template is ready for incorporation of next base [148]. Images are taken across the entire flow cell at each cycle to visualize the base incorporated in each cluster [148].

Organelle DNA

DNA Barcoding

Historically, plant species have been identified by their morphological features such as shape, size, and color. This type of identification often requires an experienced taxonomist. However, if the plant material is damaged or immature, identifications may not be possible. The use of DNA for species identification was proposed in 2003 by Paul Hebert [149]. Hebert coined this type of identification as “DNA barcoding.” In a similar

manner that a barcode can identify a product at a store, short sequences within a plant's genome could also provide identification. DNA barcoding is also used to identify animals. In animal DNA barcoding, a 648 bp region of the mitochondrial gene Cytochrome c oxidase 1 (CO1) is used to identify almost all animal groups [150]. CO1 is not a useful DNA barcode in plants as it evolves too slowly; however, the chloroplast mutates at an estimated four times faster rate than the mitochondria [151]. Several regions have been proposed as DNA barcodes for plants including *rbcL, matK, trnH-psbA,* and nuclear ITS gene. Although still debated within the community, *rbcL* and *matK* are the preferred barcodes in the Barcode of Life database [149]. Barcode regions targeted with consensus primers can be used to amplify many species even if the species is unknown. Though these regions are largely conserved, there are still polymorphisms between species that may be due to evolutionary processes. This conservation allows for universal primers to be designed to amplify regions of interest among a large number of plants. A set of conserved primers were proposed by Weising and Gardner [152]. However, this conservation also makes it difficult to distinguish between similar chloroplast genomes. Therefore, to study plant phylogenetics between species, the *rbcL* gene is often targeted.

Origin determination

Organelle markers are relatively stable from generation to generation and may be used to predict the biogeographical origin of plants such as *Cannabis*. These stable markers can become fixed in particular biogeographic populations but will remain discriminatory for populations from different regions. Analysis of organelle DNA, including both mitochondrial and chloroplast DNA, has been shown to be a valuable tool in analyzing evolutionary and population diversity in plant species as it is inherited uniparentally [153-

155]. In *C. sativa,* chloroplast and mitochondrial DNA are both inherited maternally [156]. Like human mitochondrial DNA, this inheritance pattern reveals a genetic snapshot of the evolutionary and biogeographic information of a single *Cannabis* plant. Both the chloroplast [4] and mitochondrial [6] genomes have been mapped for *C. sativa* and are freely accessible. Several studies have evaluated phylogenetic relationships in angiosperms like *Cannabis* using regions of the chloroplast and mitochondrial genomes [36, 152, 154, 157]. Universal primer sets have been used to isolate polymorphic regions in the chloroplast and mitochondrial genomes [153, 158]. Chloroplast regions targeting *Cannabis* population structure include *rbc*L [159], *trn*L – *trn*F [154, 160], *trn*H – *trn*K [155, 158], ccmp2 [152] and ccmp6 [152] region of the chloroplast. In addition, nad4 and nad5 regions of the mitochondria have been identified as polymorphic regions for *Cannabis* [155]. These regions have been evaluated previously by Gilmore et al. and results have shown that these organelle loci can somewhat discriminate *Cannabis* samples based on geographic origin [155].

Chloroplast DNA

Genome

The chloroplast genome is a double-stranded circular genome, approximately 153,871 bp in length. The chloroplast genome has been completely sequenced and mapped for *Cannabis sativa*. [4, 5]. Complete annotated genomes for *Cannabis* include Korean hemp strain, cheongsam (KR184827), African drug type, yoruba nigeria (KR363961), carmagnola (KP274871), and dagestani (KR779995). The four annotated genomes contain 127 genes: 86 protein coding, four rRNA, and 37 tRNA [4, 5]. The *Cannabis* chloroplast genome is quadripartite like most land plants, meaning that there is a long single copy

region (LSC) and short single copy region (SSC) that is separated by two inverted regions (IRa and IRb) [161, 162]. Vergara et al. observed that 16 SNPs were present when comparing two hemp varieties (carmagnola and dagestani) [4]. Similarly, Oh et al. noted 18 insertion/deletion polymorphisms (InDels) and nine SNPs when comparing a hemp strain (cheongsam) to a drug strain (yoruba nigeria) [5]. Oh et al. remarked that all polymorphisms were in the LSC and SSC regions with all but three polymorphisms occurring outside gene coding areas. The three mutations were found within exons of three genes: *matK, rsp16,* and *rpoc1*. The mutations in *matK* and *rsp16* were nonsynonymous while the mutation in *rpoc1* was synonymous or silent.

trnL – trnF

Early research focused on using organelle markers for species identification of plant materials. In 1998, Linacre and Thorpe identified an intergenic sequence between two chloroplast tRNA genes (*trnL* and *trnF*) that was specific for *Cannabis* DNA [163]. Previously identified universal primers from conserved priming sites were used to initially amplify and confirm the sequence of the intergenic space [164]. In addition, internal PCR primers were designed for *Cannabis*-specific amplification to serve as a *Cannabis* confirmation test. The *Cannabis*-specific primers were later used in a study to demonstrate the sensitivity of the technique in detecting trace amounts of *Cannabis* DNA on skin [165]. Kohjyouma et al. used two primers (E and F) proposed by Linacre and Thrope to amplify a 353/354 bp portion of the intergenic space [166]. Results showed that the primer pairs were not *Cannabis*-specific as both *Cannabis sativa* and *Humulus lupulus* yielded amplicon products. Interestingly, when the amplicons were sequenced, two sequence variants, “type-1” and “type-2”, were found amongst 33 *Cannabis* populations. “Type-2”

variants were a result of a one bp deletion. Furthermore, ten base-pair substitutions were observed between "type-2" *Cannabis sativa* and *Humulus lupulus.* This work demonstrated that while the region was generally conserved, differentiation between *Cannabis* populations could be observed. A recent study in 2015 identified a 687 bp sequence from the same intergenic space that could discriminate *Cannabis sativa* from other members of the Cannabaceae family [160].

rbcL

When studying intra-species relationships, non-coding regions such as intergenic spacers are targeted as they tend to evolve at a quicker rate than coding sequences [167]. The *rbcL* gene codes for the large subunit of the enzyme ribulose-1,5-bisphosphate carboxylase/oxidase (RUBISCO). This enzyme is involved with carbon fixation during the photosynthetic reaction. Due to the enzyme's key role in plant survival, this region is well preserved, and universal primers can be used to amplify a wide range of species [168, 169]. Gilmore et al. targeted a 3,000 bp region in *rbc*L – orf106 [155]. Yang et al. evaluated a 1,000 bp region and was able to differentiate *Cannabis sativa* from other members of the Cannabaceae family [36]. Additionally, a close relatedness to *Humulus lupulus* was confirmed in this study. More recently, Mello et al. identified a short segment (~561 bp) within the *rbcL* gene that has a potential to discriminate *Cannabis* from different sources [159]. Polymorphisms were observed between the three populations (Rio de Janeiro, China, and the UK) that were tested. Specifically, two SNP locations were found. This region warrants future study to potentially determine haplotypes for biogeographical origin.

Mitochondrial DNA

Background

Mitochondria are ubiquitous throughout the eukaryotic domain and serve as the "powerhouse" of the cell. They are double-membraned organelles that are responsible for generating ATP through the coupling of electron transport and oxidative phosphorylation. Mitochondria have their own genome outside of the nucleus that is responsible for encoding the critical, energy-generating functions of the mitochondria. Though mitochondrial genome size varies by species, a set of universal primers have been developed to analyze mitochondrial variability [158]. In *Cannabis*, the mitochondrial genome is inherited uniparentally and unchanged from the mother plant [156]. Due to this inheritance pattern, studying mitochondrial DNA may help sort *Cannabis* by biogeographic origin or chemotype.

Genome

In 2011, van Bakel et al. published a partially assembled mitochondrial genome of the Purple Kush variety [3]. White et al. improved upon this partial genome and generated a complete, annotated mitochondrial genome. The whole-genome library of a female plant from the carmagnola variety of *Cannabis sativa* was sequenced using an Illumina Hiseq2500 platform (San Diego, CA) (Accession number KR059940). The genome length is 415,499 bp, which codes for 54 genes (38 protein-coding genes,15 tRNA genes, and three rRNA genes). Once fully sequenced, error corrected, and annotated, White et al. compared the annotated genome to two unannotated genomes (purple kush13 and LA Confidential) from medicinal genomics. Compared to the annotated carmagnola variety, 69 mismatches and 271 InDels were found with purple kush13, and 164 mismatches and

212 InDels were found relative to LA confidential. It is likely that the large number of base-pair discrepancies may be more representative of errors in sequencing or assembly rather than molecular differences between the strains. In addition, White et al. performed phylogenetic analysis using 17 shared mitochondrial genes from 11 species found in the NCBI public database. The 17 coding sequences were aligned with the Clustal X software [170], and a maximum likelihood analysis was done MEGA v. 6.0612 [171]. The results mirrored the currently accepted relationships between the orders within angiosperms.

Polymorphic regions

The mitochondrial genome in plants has received less attention that the chloroplast genome due to its low mutation rate. For *Cannabis*, it was observed that there is approximately one polymorphism per 1.7 kb sequenced [155]. Due to this low mutation rate and predicted low number of polymorphisms, only one group has targeted specific sites in the mitochondrial genome to study intra-species variation within *Cannabis*. Gilmore et al. evaluated polymorphic sites in the *Cannabis* mitochondrial genome [155]. Five sites (*cox* 2 exon1 to exon2, *nad* 1 exon4 to exon5, *nad4* exon3 to exon4, *nad*5 exon4 to exon5, *nad* 7 exon1 to exon2) were screened and only two contained polymorphisms (*nad4* and *nad5*) [155].

Standardization of non-human forensic genetics

Scientific Working Group on DNA Analysis Methods (SWGDAM)

The predecessor to the Scientific Working Group on DNA Analysis Methods (SWGDAM) was the Technical Working Group on DNA Analysis Methods (TWGDAM). TWGDAM began in 1988 when forensic DNA technology was first introduced in the United States. TWGDAM was sponsored by the FBI and consisted of 31 scientists from

16 laboratories in the US and Canada. The purpose of this working group was to set standards and quality control measures in the field. As a result, SWGDAM has published many standards and guidelines for the implementation of techniques in forensic genetics including developmental and internal validation guidelines for new methods. These guidelines assure the quality of the results and mirror the Quality Assurance Standards (QAS) put forth by the FBI [172]. Some critical guidelines for developmental validation include sensitivity studies, species specificity studies, and the evaluation of precision and accuracy [173]. Additionally, internal validations should also be performed before a technique can be used in a laboratory for casework. Internal validation studies should include sensitivity and stochastic studies to determine the analytical threshold, stochastic threshold, heterozygote balance, and stutter ratios [173]. Although the SWGDAM guidelines were written for HID purposes, in the absence of specific guidelines for non-human forensics they should also be followed as closely as possible in non-human genetics to ensure the robustness and standardization of the DNA method. Two canine STR assays have been published following SWGDAM guidelines: “DogFiler” [174] and “Mini-DogFiler” [175].

Organization of Scientific Area Committee for forensic science (OSAC)

In 2014, the OSAC was established under the National Institute of Standards and Technology (NIST) in the United States. It was created with the purpose of strengthening forensic science by developing discipline-specific standards and encouraging the adoption of those standards. Additionally, the OSAC is responsible for identifying research and development needs in forensic science. The OSAC consists of more than 550 members from government, academic, and private agencies. The organizational structure is divided

into five Scientific Area Committees (SACs) and 25 subcommittees. There is a Wildlife Forensics subcommittee within the Biology/DNA SAC that is responsible for developing and approving standards for non-human biological evidence (morphological and genetic).

International Society of Forensic Genetics (ISFG)

ISFG was founded in 1968 and is currently composed of approximately 1100 scientists from more than 60 countries. As specific needs arise in the community of forensic genetics, ISFG assembles a commission of experts in the field to make recommendations for the community. These recommendations are published and publicly available. In 2011, ISFG published a set of recommendations for the use of non-human DNA for forensic genetics investigations [120]. Non-human DNA has historically had very little standardization. ISFG specifically addressed animal DNA; however, the same recommendations should be applied to plant DNA. Thirteen recommendations were given, and all reflect standards implemented for HID testing. Some key recommendations include the use of tetranucleotide markers, sequenced allelic ladders, species specificity testing, and the use of an allele frequency database. Although ISFG recommendations have not been followed for *Cannabis* DNA typing specifically, efforts have been made to follow ISFG recommendations in the general field of non-human genetics [176, 177].

Statement of the problem

Marijuana is the most commonly used illicit drug in the United States. After a period of decline in the last decade, its use has been increasing amongst young people since 2007, corresponding to a diminishing perception of the drug's risks that may be associated with the increased public debate over the drug's legal status.

Although the federal government considers marijuana a Schedule I substance (having no medicinal uses and high risk for abuse), eight states – California, Colorado, Nevada, Maine, Massachusetts, Washington, Oregon, Alaska – have legalized marijuana for adult recreational use, and more than 20 additional states have passed laws allowing its use as a treatment for certain medical conditions.

On November 2012, Washington and Colorado passed legislation to legalize recreational marijuana sales to anyone age 21 or older, which prompted the opening of a stream of marijuana recreational dispensaries in Colorado beginning on January 1, 2014, while Washington stores opened their doors in Spring of 2014. Due to new legislation, law enforcement agencies are facing a new challenge: preventing the diversion of marijuana products bought in legalized states from being trafficked to other states where the drug is still illegal. Although security measurements have been implemented to monitor the commercial flow of the drug from the production to the final customer, no DNA registry was implemented to track the product due to limited funds. The development of a validated method using molecular techniques for genetic identification of *C. sativa* plants (with a corresponding DNA database) will allow not only the individualization of commercial specimens but will also identify illegal products.

The use of a DNA-based identification method (including lineage DNA markers for origin determination) will also allow law enforcement agencies (e.g., U.S. Customs & Border Protection, through their operations at the airport and the US border) to associate cases where *C. sativa* samples are involved (illegal traffic of *C. sativa* from Mexico).

A validated genetic method that enables the association of these drug cases is necessary to investigate illegal operations. Although several techniques have been

published and implemented to investigate the origin of marijuana samples (including palynology, chemical profiling, and isotopic analysis), none can provide information that could link growers [74-77]. Various approaches utilizing genetic information may provide even finer resolution than isotopic analysis, and DNA-based tools for *C. sativa* identification and population studies are being developed by multiple research groups around the world [79, 80, 112, 114].

Indeed, a multiplex short tandem repeat (STR) system was successfully developed for the molecular identification of *C. sativa* [78] as well as for the formation of an STR database for *Cannabis* seizures in Australia [80]. Additionally, efforts have been conducted to use DNA profiling to discriminate marijuana sources, including a combination of chloroplast DNA (cpDNA) and mitochondrial DNA (mtDNA) analysis to reflect crop-use and geographic origin [155]. The association between different *C. sativa* plants has been previously assessed using a combination of autosomal STR markers and statistical genetics tools [109].

In most dispensaries/growing operations, the quality and potency of THC is maintained by taking cuttings from a high-THC content "mother" plant and directly rooting them in the soil or hydroponic liquid. This clonal form of propagation results in plants contains identical DNA (like monozygotic twins in humans). DNA typing of marijuana in this situation would confirm linkage of growing operations as well as assess distribution patterns through the tracking of clonal material. Other growers cultivate their marijuana plants from seed. Each seed has its unique genetic composition, but seeds coming from the same mother can be traced back using lineage DNA markers.

None of the previously published reports using *Cannabis* STR profiling have followed three crucial ISFG recommendations for the use of non-human DNA in forensic genetic investigations [120]: a) avoiding the use of dinucleotides (instead, tetranucleotides are recommended), b) the use of sequenced allelic ladders for accurate designation of alleles and inter-laboratory STR profile sharing, and c) relevant population and forensic parameters studied in a representative homogeneous (low F_{ST}) population of *C. sativa* for random match probability estimations or verification of genetic relatedness. The combined use of organelle and autosomal DNA markers will allow the association of different cases (or group of samples) by a) detecting the presence of clones, b) the association between group of samples and fragments of the same plant, and c) determining the geographical origin of a sample or group of samples.

Moreover, the application of new technologies, such as massively parallel sequencing (MPS), will allow for an automated, high-throughput analysis with more comprehensive coverage of genetic markers to take full advantage of the increased power of discrimination afforded by sequencing.

Here, we propose the development of a comprehensive analytical tool that includes the combination of a newly developed multiplex STR method (following ISFG recommendations for non-human DNA testing), a set of lineage markers (cpDNA and mtDNA) for discriminating *C. sativa* sources, and an MPS approach for sequencing DNA in a massively parallel fashion with both high coverage and high throughput of specified targets.

References

1. Ainsworth C (2000) Boys and girls come out to play: the molecular biology of dioecious plants. Ann Bot 86:211-221. https://doi.org/10.1006/anbo.2000.1201
2. Sakamoto K, Akiyama Y, Fukui K, Kamada H, Satoh S (1998) Characterization; genome sizes and morphology of sex chromosomes in hemp (*Cannabis sativa* L.). Cytologia (Tokyo) 63:459-464. https:/doi.org/10.1508/cytologia.63.459
3. van Bakel H, Stout J, Cote A, Tallon C, Sharpe A, Hughes T, Page J (2011) The draft genome and transcriptome of *Cannabis sativa.* Genome Biol 12:R102. https://doi.org/10.1186/gb-2011-12-10-r102
4. Vergara D, White KH, Keepers KG, Kane NC (2016) The complete chloroplast genomes of *Cannabis sativa* and *Humulus lupulus*. Mitochondrial DNA Part A DNA Mapp Seq Anal 27:3793-3794. https://doi.org/10.3109/19401736.2015.1079905
5. Oh H, Seo B, Lee S, Ahn DH, Jo E, Park JK, Min GS (2016) Two complete chloroplast genome sequences of *Cannabis sativa* varieties. Mitochondrial DNA Part A DNA Mapp Seq Anal 27: 2835-2837. 10.3109/19401736.2015.1053117
6. White KH, Vergara D, Keepers KG, Kane NC (2016) The complete mitochondrial genome for *Cannabis sativa*. Mitochondrial DNA Part B 1:715-716. https://doi.org/10.1080/23802359.2016.1155083
7. Andre CM, Hausman JF, Guerriero G (2016) *Cannabis sativa*: the plant of the thousand and one molecules. Front Plant Sci 7:19. https://doi.org/10.3389/fpls.2016.00019

8. Shephard HL, Parker JS, Darby P, Ainsworth CC (2000) Sexual development and sex chromosomes in hop. New Phytol 148:397-411. https://doi.org/10.1046/j.1469-8137.2000.00771.x

9. Vyskot B, Hobza R (2004) Gender in plants: sex chromosomes are emerging from the fog. Trends Genet 20: 432-438. https://doi.org/10.1016/j.tig.2004.06.006

10. Sakamoto K, Shimomura K, Komeda Y, Kamada H, Satoh S (1995) A male-associated DNA sequence in a dioecious plant, *Cannabis sativa* L. Plant Cell Physiol 36:1549-1554

11. Sakamoto K, Ohmido N, Fukui K, Kamada H, Satoh S (2000) Site-specific accumulation of a line-like retrotransposon in a sex chromosome of the dioecious plant *Cannabis sativa*. Plant Mol Biol 44:723-732. https://doi.org/10.1023/A:1026574405717

12. Mandolino G, Carboni A, Forapani S, Faeti V, Ranalli P (1999) Identification of DNA markers linked to the male sex in dioecious hemp (*Cannabis sativa* L.). Theor and Appl Genet 98:86-92. https://doi.org/10.1007/s001220051043

13. Mandolino G, Carboni A, Bagatta M, Moliterni VMC, Ranalli P (2002) Occurrence and frequency of putatively Y chromosome linked DNA markers in *Cannabis sativa* L. Euphytica 126:211-218. https://doi.org/10.1023/a:1016382128401

14. Sakamoto K, Abe T, Matsuyama T, Yoshida S, Ohmido N, Fukui K, Satoh S (2005) RAPD markers encoding retrotransposable elements are linked to the male sex in *Cannabis sativa* L. Genome 48:931-936. https://doi.org/10.1139/g05-056

15. Flachowsky H, Schumann E, Weber WE, Peil A (2001) Application of AFLP for the detection of sex-specific markers in hemp. Plant Breeding 120:305-309. https://doi.org/10.1046/j.1439-0523.2001.00620.x

16. Peil A, Flachowsky H, Schumann E, Weber WE (2003) Sex-linked AFLP markers indicate a pseudoautosomal region in hemp (*Cannabis sativa* L.). Theor Appl Genet 107:102-109. https://doi.org/10.1007/s00122-003-1212-5

17. Techen N, Chandra S, Lata H, Elsohly MA, Khan IA (2010) Genetic identification of female *Cannabis sativa* plants at early developmental stage. Planta Med 76:1938-1939. https://doi.org/10.1055/s-0030-1249978

18. Shao H, Song SJ, Clarke RC (2003) Female-associated DNA polymorphisms of hemp (*Cannabis sativa* L.). J of Industrial Hemp 8:5-9. https://doi.org/10.1300/J237v08n01_02

19. Sirikantaramas S, Taura F, Tanaka Y, Ishikawa Y, Morimoto S, Shoyama Y (2005) Tetrahydrocannabinolic acid synthase, the enzyme controlling marijuana psychoactivity, is secreted into the storage cavity of the glandular trichomes. Plant Cell Physiol 46:1578-1582. https://doi.org/10.1093/pcp/pci166

20. Kojoma M, Seki H, Yoshida S, Muranaka T (2006) DNA polymorphisms in the tetrahydrocannabinolic acid (THCA) synthase gene in "drug-type" and "fiber-type" *Cannabis sativa* L. Forensic Sci Int 159:132-140. https://doi.org/10.1016/j.forsciint.2005.07.005

21. de Meijer EPM, Bagatta M, Carboni A, Crucitti P, Moliterni VMC, Ranalli P, Mandolino G (2003) The inheritance of chemical phenotype in *Cannabis sativa* L. Genetics 163:335-346

22. Rotherham D, Harbison SA (2011) Differentiation of drug and non-drug *Cannabis* using a single nucleotide polymorphism (SNP) assay. Forensic Sci Int 207:193-197. https://doi.org/10.1016/j.forsciint.2010.10.006

23. Cascini F, Passerotti S, Martello S (2012) A real-time PCR assay for the relative quantification of the tetrahydrocannabinolic acid (THCA) synthase gene in herbal cannabis samples. Forensic Sci Int 217:134-138. https://doi.org/10.1016/j.forsciint.2011.10.041

24. Cirovic N, Kecmanovic M, Keckarevic D, Keckarevic Markovic M (2017) Differentiation of *Cannabis* subspecies by THCA synthase gene analysis using RFLP. J of Forensic and Leg Med 51:81-84. https://doi.org/10.1016/j.jflm.2017.07.015

25. Sawler J, Stout JM, Gardner KM, Hudson D, Vidmar J, Butler L, Page JE, Myles S (2015) The genetic structure of marijuana and hemp. PLoS ONE 10:e0133292. https://doi.org/10.1371/journal.pone.0133292

26. Soorni A, Fatahi R, Haak DC, Salami SA, Bombarely A (2017) Assessment of genetic diversity and population structure in Iranian cannabis germplasm. Sci Rep 7:15668. https://doi.org/10.1038/s41598-017-15816-5

27. Piomelli D, Russo EB (2016) The *Cannabis sativa* versus *Cannabis indica* debate: An interview with ethan russo, md. Cannabis Cannabinoid Res 1:44-46. https://doi.org/10.1089/can.2015.29003.ebr

28. Small E (2015) Evolution and classification of *Cannabis sativa* (marijuana, hemp) in relation to human utilization. Bot Rev 81:189-294. 10.1007/s12229-015-9157-3

29. Linnaeus C (1753) Species Plantarum. Stockholm, Sweden

30. Lamarck JB (1785) Encyclopédie métodique. Botanique: Paris-Liege, 1783–1803

31. Small E, Cronquist A (1976) A practical and natural taxonomy for cannabis. Taxon 25:406-435

32. Hillig KW (2005) Genetic evidence for speciation in *Cannabis* (Cannabaceae). Genet Res and Crop Evol 52:161-180. https://doi.org/10.1007/s10722-003-4452-y

33. Hillig KW, Mahlberg PG (2004) A chemotaxonomic analysis of cannabinoid variation in *Cannabis* (Cannabaceae). Am J Bot 91:966-975. https://doi.org/10.3732/ajb.91.6.966

34. Small E (1978) A numerical and nomenclatural analysis of morpho-geographic taxa of Humulus. Syst Bot 3:37-76. https://doi.org/10.2307/2418532

35. Sytsma KJ, Morawetz J, Pires JC, Nepokroeff M, Conti E, Zjhra M, Hall JC, Chase MW (2002) Urticalean rosids: circumscription, rosid ancestry, and phylogenetics based on *rbc*L, *trn*L-F, and *ndh*F sequences. Am J Bot 89:1531-1546. https://doi/10.3732/ajb.89.9.1531

36. Yang MQ, van Velzen R, Bakker FT, Sattarian A, Li DZ, Yi TS (2013) Molecular phylogenetics and character evolution of cannabaceae. Taxon 62:473-485. https://doi.org/10.12705/623.9

37. Bell CD, Soltis DE, Soltis PS (2010) The age and diversification of the angiosperms re-revisited. Am J Bot 97:1296-1303. https://doi.org/10.3732/ajb.0900346

38. Li HL (1974) The origin and use of cannabis in eastern asia linguistic-cultural implications. Econ Bot 28: 293-301. https://doi.org/10.1007/bf02861426

39. Fleming MP, Clarke R (1998) Physical evidence for the antiquity of *Cannabis sativa* L. J of Int Hemp Association 5:80-93

40. Kung CT (1952) Archaeology in China. University of Toronto Press Toronto, ON

41. Chang KC (1963) The archaeology of ancient China. Yale University Press, New Haven, CT

42. Schultes R, Hofmann A (1980) The botany and chemistry of hallucinogens. Charles C Thomas, Springfield, IL

43. Small E (1979b) The species problem in *Cannabis*: science and semantics. Volume 2. Corpus, Toronto, ON

44. Bócsa I, Karus M (1998) The cultivation of hemp: botany, varieties, cultivation and harvesting. Hemptech, Sebastopol, CA

45. Schultes RE, Klein WM, Plowman T, Lockwood TE (1974) *Cannabis*: an example of taxonomic neglect. Botanical Museum Leaflets, Harvard University 23:337-367

46. Schultes RE (1973) Man and marijuana: thousands of years before it became the superstar of the drug culture, *Cannabis* was cultivated for fiber, food, and medicine. American Museum of Natural History

47. Schultes RE (1970) Random thoughts and queries on the botany of *Cannabis*. The botany and chemistry of *Cannabis* J. & A. Churchill, London, pp.11-33

48. Li HL (1974) An archaeological and historical account of *Cannabis* in China. Econ Bot 28:437-448.

49. Li HL (1978) Hallucinogenic plants in Chinese herbals. J of Psychedelic Drugs 10: 17-26. https://doi.org/10.1080/02791072.1978.10471863

50. Jiang HE, Li X, Zhao YX, Ferguson DK, Hueber F, Bera S, Wang YF, Zhao LC, Liu CJ, Li CS (2006) A new insight into C*annabis sativa* (Cannabaceae) utilization from 2500-year-old yanghai tombs, Xinjiang, China. J of Ethnopharmacology 108:414-422. https://doi.org/10.1016/j.jep.2006.05.034

51. Mukherjee A, Roy SC, De Bera S, Jiang HE, Li X, Li CS, Bera S (2008) Results of molecular analysis of an archaeological hemp (*Cannabis sativa* L.) DNA sample from north west China. Genet Resources and Crop Evol 55:481-485. https://doi.org/10.1007/s10722-008-9343-9

52. Aldrich MR (1977) Tantric *Cannabis* use in India. J of Psychedelic Drugs 9:227-233. https://doi.org/10.1080/02791072.1977.10472053

53. ElSohly MA (2007) Marijuana and the cannabinoids. Humana Press, Totowa

54. Elsohly MA, Slade D (2005) Chemical constituents of marijuana: the complex mixture of natural cannabinoids. Life sciences 78:539-548. https://doi.org/10.1016/j.lfs.2005.09.011

55. Grotenhermen F, Russo E (2002) *Cannabis* and cannabinoids pharmacology, toxicology, and therapeutic potential. The Haworth Integrative Healing Press, New York

56. Small E, Beckstead HD (1973) Cannabinoid phenotypes in *Cannabis sativa.* Nature 245:147. https://doi.org/10.1038/245147a0

57. Chandra S, Lata H, Khan IA, Elsohly MA (2008) Photosynthetic response of *Cannabis sativa* L. To variations in photosynthetic photon flux densities, temperature and CO_2 conditions. Physiol and Mol Biol of Plants 14:299-306. https://doi.org/10.1007/s12298-008-0027-x

58. Chandra S, Lata H, Khan IA, ElSohly MA (2011) Temperature response of photosynthesis in different drug and fiber varieties of *Cannabis sativa* L. Physiol and Mol Biol of Plants 17:297-303. https://doi.org/10.1007/s12298-011-0068-4

59. Miller Coyle H, Palmbach T, Juliano N, Ladd C, Lee HC (2003) An overview of DNA methods for the identification and individualization of marijuana. Croat Med J 44:315-321

60. Ranney T. (2006) Polyploidy: from evolution to new plant development. Combined Proceedings Int Plant Propagators' Soc 56

61. Mansouri H, Bagheri M (2017) Induction of polyploidy and its effect on cannabis sativa l. In: Chandra S, Lata H, ElSohly MA, eds. *Cannabis sativa* L. - botany and biotechnology. Springer International Publishing Cham. New York, pp. 365-383

62. Hsieh HM, Hou RJ, Tsai LC, Wei CS, Liu SW, Huang LH, Kuo YC, Linacre A, Lee JC (2003) A highly polymorphic STR locus in *Cannabis sativa*. Forensic Sci Int 131:53-58

63. Knight G, Hansen S, Connor M, Poulsen H, McGovern C, Stacey J (2010) The results of an experimental indoor hydroponic cannabis growing study, using the 'screen of green' (scrog) method-yield, tetrahydrocannabinol (THC) and DNA analysis. Forensic Sci Int 202:36-44. https://doi.org/10.1016/j.forsciint.2010.04.022

64. McKenna GJ (2014) The current status of medical marijuana in the United States. Hawai'i J of Med & Pub Health 73:105-108

65. Ooyen-Houben Mv, Kleemans E (2015) Drug policy: the “dutch model”. Crime and Justice 44:165-226. https://doi/10.1086/681551

66. Room R (2014) Legalizing a market for cannabis for pleasure: Colorado, Washington, Uruguay and beyond. Addiction (Abingdon, England) 109:345-351. https://doi.org/10.1111/add.12355

67. Marijuana and the controlled substances act (2014) Congressional Digest 93:2.

68. Mead A (2017) The legal status of *Cannabis* (marijuana) and cannabidiol (CBD) under U.S. Law. Epilepsy Behav 70:288-291. https://doi.org/10.1016/j.yebeh.2016.11.021

69. Carliner H, Brown QL, Sarvet AL, Hasin DS (2017) Cannabis use, attitudes, and legal status in the U.S.: a review. Prev Med 104:13-23. https://doi.org/10.1016/j.ypmed.2017.07.008

70. Pacula RL, Powell D, Heaton P, Sevigny EL (2015) Assessing the effects of medical marijuana laws on marijuana use: the devil is in the details. J Policy Anal Manag 34: 7-31. https://doi.org/10.1002/pam.21804

71. Applications to become registered under the controlled substances act to manufacture marijuana to supply researchers in the United States. Policy statement (2016) Federal register 81:53846-53848

72. Mikes F, Hofmann A, Waser PG (1971) Identification of (-)-delta 9-6a,10a-trans-tetrahydrocannabinol and two of its metabolites in rats by use of combination gas chromatography-mass spectrometry and mass fragmentography. Biochem Pharmacol 20:2469-2476

73. Mitosinka GT, Thornton JI, Hayes TL (1972) The examination of cystolithic hairs of cannabis and other plants by means of the scanning electron microscope. J Forensic Sci Soc 12:521-529

74. Bryant VM, Jones GD (2006) Forensic palynology: current status of a rarely used technique in the United States of America. Forensic Sci Int 163:183-197. https://doi.org/10.1016/j.forsciint.2005.11.021

75. Brenneisen R, elSohly MA (1988) Chromatographic and spectroscopic profiles of cannabis of different origins: Part i. J Forensic Sci 33:1385-1404

76. Shibuya EK, Souza Sarkis JE, Neto ON, Moreira MZ, Victoria RL (2006) Sourcing brazilian marijuana by applying irms analysis to seized samples. Forensic Sci Int 160:35-43. https://doi.org/10.1016/j.forsciint.2005.08.011

77. Shibuya EK, Sarkis JES, Negrini-Neto O, Martinelli LA (2007) Carbon and nitrogen stable isotopes as indicative of geographical origin of marijuana samples seized in the city of São Paulo (Brazil). Forensic Sci Int 167:8-15. https://doi.org/10.1016/j.forsciint.2006.06.002

78. Howard C, Gilmore S, Robertson J, Peakall R (2008) Developmental validation of a *Cannabis Sativa* STR multiplex system for forensic analysis. J Forensic Sci 53:1061-1067. https://doi.org/10.1111/j.1556-4029.2008.00792.x

79. Dufresnes C, Jan C, Bienert F, Goudet J, Fumagalli L (2017) Broad-scale genetic diversity of *Cannabis* for forensic applications. PLoS ONE 12:e0170522. https://doi.org/10.1371/journal.pone.0170522

80. Howard C, Gilmore S, Robertson J, Peakall R (2009) A *Cannabis sativa* STR genotype database for Australian seizures: forensic applications and limitations. J Forensic Sci 54:556-563. https://doi.org/10.1111/j.1556-4029.2009.01014.x

81. Srivastava AK, Schlessinger D (1991) Structure and organization of ribosomal DNA. Biochimie 73:631-638

82. Pillay M, Kenny ST (2006) Structural organization of the nuclear ribosomal RNA genes in *Cannabis* and *Humulus* (cannabaceae). Plant Systematics and Evolution: 97

83. Gigliano GS, Caputo P, Cozzolino S (1997) Ribosomal DNA analysis as a tool for the identification of *Cannabis sativa* L. Specimens of forensic interest. Sci Justice 37:171-174. https://doi.org/10.1016/s1355-0306(97)72170-1

84. Gigliano GS (1998) Identification of C*annabis sativa* L. (cannabaceae) using restriction profiles of the internal transcribed spacer II (ITS2). Sci Justice 38: 225-230

85. Gigliano GS (1999) Preliminary data on the usefulness of internal transcribed spacer I (ITS1) sequence in *Cannabis sativa* L. identification. J Forensic Sci 44:475-477

86. Daud Khaled AK, Neilan BA, Henriksson A, Conway PL (1997) Identification and phylogenetic analysis of lactobacillus using multiplex RAPD-PCR. FEMS Microbiol Lett 153:191-197

87. Yu YL, Lin TY (1997) Construction of phylogenetic tree fornicotiana species based on RAPD markers. J Plant Res 110:187-193. https://doi.org/10.1007/BF02509307

88. Gillan R, Cole MD, Linacre A, Thorpe JW, Watson ND (1995) Comparison of *Cannabis sativa* by random amplification of polymorphic DNA (RAPD) and HPLC of cannabinoids: a preliminary study. Sci Justice 35:169-177. https://doi.org/10.1016/s1355-0306(95)72658-2

89. Jagadish V, Robertson J, Gibbs A (1996) Rapd analysis distinguishes *Cannabis sativa* samples from different sources. Forensic Sci Int 79:113-121. https://doi.org/10.1016/0379-0738(96)01898-1

90. Faeti V, Mandolino G, Ranalli P (1996) Genetic diversity of *Cannabis sativa* germplasm based on RAPD markers. Plant Breeding 115:367-370. https://doi.org/10.1111/j.1439-0523.1996.tb00935.x

91. Shirota O, Watanabe A, Yamazaki M, Saito K, Shibano K, Sekita S, Satake N (1998) Random amplified polymorphic DNA and restriction fragment length polymorphism analyses of *Cannabis sativa*. Natural medicines 52: 160-166

92. Forapani S, Ranalli P, Mandolino G, Moliterni VMC, Carboni A, Paoletti C (2001) Comparison of hemp varieties using random amplified polymorphic DNA markers. Crop science 41:1682-1689

93. Pinarkara E, Kayis SA, Hakki EE, Sag A (2009) RAPD analysis of seized marijuana (*Cannabis sativa* L.) in turkey. Electronic J of Biotech 12:1-13. https://doi.org/10.2225/vol12-issue1-fulltext-7

94. Kojoma M, Iida O, Makino Y, Sekita S, Satake M (2002) DNA fingerprinting of cannabis sativa using inter-simple sequence repeat (ISSR) amplification. Planta Medica 68:60-63. https://doi.org/10.1055/s-2002-19875

95. Hakki EE, Kayis SA, Pinarkara E, Sag A (2007) Inter simple sequence repeats separate efficiently hemp from marijuana (*Cannabis sativa* L.). Electron J Biotechnol 10:570-581. https://doi.org/10.2225/vol10-issue4-fulltext-4

96. Punja ZK, Rodriguez G, Chen S (2017) Assessing genetic diversity in *Cannabis sativa* using molecular approaches. ln: Chandra S, Lata H, ElSohly M (eds) *Cannabis sativa* L. - Botany and Biotechnology. Springer International Publishing Cham. New York, pp. 395-418

97. Khatak S, Ghai M, Dahiya S (2016) ISSR marker based inter and intra-specific diversity analysis in different genotypes of *Cannabis sativa*. Int Conf on Innovative Res in Engg Sci and Mang 3:6-16

98. Paun O, Schönswetter P (2012) Amplified fragment length polymorphism (AFLP) - an invaluable fingerprinting technique for genomic, transcriptomic and epigenetic studies. Methods in Mol Biol (Clifton, NJ) 862:75-87. https://doi.org/10.1007/978-1-61779-609-8_7

99. Vuylsteke M, Peleman JD, van Eijk MJT (2007) AFLP technology for DNA fingerprinting. Nature Protocols 2:1387. https://doi.org/10.1038/nprot.2007.175

100. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M (1995) AFLP: A new technique for DNA fingerprinting. Nucleic Acids Res 23:4407-4414

101. Zabeau M, Vos P (1993) Selective restriction fragment amplification: a genreal method for DNA fingerprinting. European Patent Office, publication 0 534 858 A1, bulletin 93/13

102. Miller Coyle H, Shutler G, Abrams S, Hanniman J, Neylon S, Ladd C, Palmbach T, Lee HC (2003) A simple DNA extraction method for marijuana samples used in amplified fragment length polymorphism (AFLP) analysis. J Forensic Sci 48:343-347

103. Datwyler SL, Weiblen GD (2006) Genetic variation in hemp and marijuana (*Cannabis sativa* L.) according to amplified fragment length polymorphisms. J Forensic Sci 51:371-375. https://doi.org/10.1111/j.1556-4029.2006.00061.x

104. Hu ZG, Guo HY, Hu XL et al (2012) Genetic diversity research of hemp (*Cannabis sativa* L.) cultivar based on AFLP analysis. J Plant Genet Resources 13:555-561

105. Mueller UG, Wolfenbarger LL (1999) Aflp genotyping and fingerprinting. Trends Ecol Evol 14:389-394. https://doi.org/10.1016/S0169-5347(99)01659-6

106. Butler JM (2005) Forensic DNA typing: Biology, technology, and genetics of STR markers. Elseiver Academic Press, New York

107. Butler JM (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. J Forensic Sci 51:253-265. https://doi.org/10.1111/j.1556-4029.2006.00046.x

108. Gilmore S, Peakall R (2003) Isolation of microsatellite markers in *Cannabis sativa* L. (marijuana). Mol Ecol 3:105-107. https://doi.org/10.1046/j.1471-8286.2003.00367.x

109. Alghanim HJ, Almirall JR (2003) Development of microsatellite markers in *Cannabis sativa* for DNA typing and genetic relatedness analyses. Anal Bioanal Chem 376:1225-1233. https://doi.org/10.1007/s00216-003-1984-0

110. Gilmore S, Peakall R, Robertson J (2003) Short tandem repeat (STR) DNA markers are hypervariable and informative in *Cannabis sativa*: implications for forensic investigations. Forensic Sci Int 131:65-74

111. Mendoza MA, Mills DK, Lata H, Chandra S, ElSohly MA, Almirall JR (2009) Genetic individualization of *Cannabis sativa* by a short tandem repeat multiplex system. Anal Bioanal Chem 393:719-726. https://doi.org/10.1007/s00216-008-2500-3

112. Allgeier L, Hemenway J, Shirley N, LaNier T, Coyle HM (2011) Field testing of collection cards for cannabis sativa samples with a single hexanucleotide DNA marker. J Forensic Sci 56:1245-1249. https://doi.org/10.1111/j.1556-4029.2011.01818.x

113. Shirley N, Allgeier L, LaNier T, Coyle HM (2013) Analysis of the NMI01 marker for a population database of *Cannabis* seeds. J Forensic Sci 58:S176-182. https://doi.org/ 10.1111/1556-4029

114. Kohnemann S, Nedele J, Schwotzer D, Morzfeld J, Pfeiffer H (2012) The validation of a 15 STR multiplex PCR for cannabis species. Int J Leg Med 126:601-606. https://doi.org/10.1007/s00414-012-0706-6

115. Valverde L, Lischka C, Scheiper S, Nedele J, Challis R, de Pancorbo MM, Pfeiffer H, Kohnemann S (2014) Characterization of 15 STR cannabis loci: nomenclature proposal and SNPSTR haplotypes. Forensic Sci Int Genet 9:61-65. https://doi.org/10.1016/j.fsigen.2013.11.001

116. Gill P, Brinkmann B, d'Aloja E, Andersen J, Bar W, Carracedo A, Dupuy B, Eriksen B, Jangblad M, Johnsson V, Kloosterman AD, Lincoln P, Morling N, Rand

S, Sabatier M, Scheithauer R, Schneider P, Vide MC (1997) Considerations from the European DNA profiling group (EDNAP) concerning STR nomenclature. Forensic Sci Int 87:185-192

117. Olaisen B, Bar W, Brinkmann B, Budowle B, Carracedo A, Gill P, Lincoln P, Mayr WR, Rand S (1998) DNA recommendations 1997 of the international society for forensic genetics. Vox sang 74: 61-63.

118. DNA recommendations-1994 report concerning further recommendations of the DNA commission of the ISFH regarding pcr-based polymorphisms in str (short tandem repeat) systems (1995). Vox Sang 69:70-71.

119. Valverde L, Lischka C, Erlemann S, de Meijer E, de Pancorbo MM, Pfeiffer H, Köhnemann S (2014) Nomenclature proposal and SNPSTR haplotypes for 7 new *Cannabis sativa* L. STR loci. Forensic Sci Int Genet 13:185-186. http://dx.doi.org/10.1016/j.fsigen.2014.08.002

120. Linacre A, Gusmao L, Hecht W, Hellmann AP, Mayr WR, Parson W, Prinz M, Schneider PM, Morling N (2011) ISFG: recommendations regarding the use of non-human (animal) DNA in forensic genetic investigations. Forensic Sci Int Genet 5:501-505. https://doi.org/10.1016/j.fsigen.2010.10.017

121. Kraemer L, Beszteri B, Gäbler-Schwarz S, Held C, Leese F, Mayer C, Pöhlmann K, Frickenhaus S (2009) Stamp: extensions to the staden sequence analysis package for high throughput interactive microsatellite marker design. BMC Bioinformatics 10:41 https://doi.org/10.1186/1471-2105-10-41

122. Soler S, Gramazio P, Figàs MR, Vilanova S, Rosa E, Llosa ER, Borràs D, Plazas M, Prohens J (2017) Genetic structure of *Cannabis sativa* var. *indica* cultivars

based on genomic SSR (gSSR) markers: Implications for breeding and germplasm management. Ind Crops Prod 104:171-178. https://doi.org/10.1016/j.indcrop.2017.04.043

123. Gao C, Xin P, Cheng C, Tang Q, Chen P, Wang C, Zang G, Zhao L (2014) Diversity analysis in *Cannabis sativa* based on large-scale development of expressed sequence tag-derived simple sequence repeat markers. PLoS ONE 9:e110638. https://doi.org/10.1371/journal.pone.0110638

124. Wang C, Guo W, Zhang T, Li Y, Liu H (2009) AutoSSR: an improved automatic software for SSR analysis from large-scale est sequences. Cotton Science 21:243-247

125. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, Normandeau É, Laroche J, Larose S, Jean M, Belzile F (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS ONE 8:e54603. https://doi.org/10.1371/journal.pone.0054603

126. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379. https://doi.org/10.1371/journal.pone.0019379

127. Raj A, Stephens M, Pritchard JK (2014) fastStructure: variational inference of population structure in large SNP data sets. Genet 197:573-589. https://doi.org/10.1534/genetics.114.164350

128. Piluzza G, Delogu G, Cabras A, Marceddu S, Bullitta S (2013) Differentiation between fiber and drug types of hemp (*Cannabis sativa* L.) from a collection of wild and domesticated accessions. Genet Resour Crop Evol 60:2331-2342. https://doi.org/10.1007/s10722-013-0001-5

129. Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84-91. https://doi.org/10.1104/pp.102.019422

130. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in escherichia coli. Nature 274:775. https://doi.org/10.1038/274775a0

131. Pootakham W, Jomchai N, Ruang-areerate P, Shearman JR, Sonthirod C, Sangsrakru D, Tragoonrung S, Tangphatsornruang S (2015) Genome-wide SNP discovery and identification of qtl associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). Genomics 105:288-295. https://doi.org/10.1016/j.ygeno.2015.02.002

132. Nei M (1972) Genetic distance between populations. Am Nat 106:283-292

133. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 11:94. https://doi.org/10.1186/1471-2156-11-94

134. Jombart T (2008) Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403-1405. https://doi.org/10.1093/bioinformatics/btn129

135. Moorthie S, Mattocks CJ, Wright CF (2011) Review of massively parallel DNA sequencing technologies. The HUGO Journal 5:1-12. 10.1007/s11568-011-9156-3

136. Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107:1-8. https://doi.org/10.1016/j.ygeno.2015.11.003

137. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74:5463-5467

138. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238:336-341

139. Swerdlow H, Gesteland R (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res 18:1415-1419

140. Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, Smith LM (1990) High speed DNA sequencing by capillary electrophoresis. Nucleic Acids Res 18: 4417-4421

141. Seo SB, King JL, Warshauer DH, Davis CP, Ge J, Budowle B (2013) Single nucleotide polymorphism typing with massively parallel sequencing for human identification. Int J Leg Med 127:1079-1086. https://doi.org/10.1007/s00414-013-0879-7

142. Massart S, Olmos A, Jijakli H, Candresse T (2014) Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Research 188:90-96. http://doi.org/10.1016/j.virusres.2014.03.029

143. Vasan N, Yelensky R, Wang K, Moulder S, Dzimitrowicz H, Avritscher R, Wang B, Wu Y, Cronin MT, Palmer G, Symmans WF, Miller VA, Stephens P, Pusztai L (2014) A targeted next-generation sequencing assay detects a high frequency of therapeutically targetable alterations in primary and metastatic breast cancers: Implications for clinical practice. The Oncologist 19:53-458. https://doi.org/10.1634/theoncologist.2013-0377

144. Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R, Claverie J-M, Decelle J, Dolan John R, Dunthorn M, Edvardsen B, Gobet A, Kooistra Wiebe HCF, Mahé F, Not F, Ogata H, Pawlowski J, Pernice Massimo C, Romac S, Shalchian-Tabrizi K, Simon N, Stoeck T, Santini S, Siano R, Wincker P, Zingone A, Richards Thomas A, de Vargas C, Massana R (2014) Patterns of rare and abundant marine microbial eukaryotes. Current Biol 24:813-821. http://doi.org/10.1016/j.cub.2014.02.050

145. Lazaruk K, Walsh PS, Oaks F, Gilbert D, Rosenblum BB, Menchen S, Scheibler D, Wenz HM, Holt C, Wallin J (1998) Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument. Electrophoresis 19:86-93. https://doi.org/10.1002/elps.1150190116

146. Rothberg JM, Hinz W, Rearick TM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348. 10.1038/nature10242
https://www.nature.com/articles/nature10242#supplementary-information

147. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms:

Comparison of ion torrent, pacific biosciences and illumina miseq sequencers. BMC Genomics 13:341 https://doi.org/101186/1471-2164-13-341 13: 341

148. Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53. https://doi.org/10.1038/nature07517

149. Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings Biolog Sci 270:313-321. https://doi.org/10.1098/rspb.2002.2218

150. Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society B: Biolog Sci 270:S96-S99. https://doi.org/10.1098/rsbl.2003.0025

151. Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol 28:87-97

152. Weising K, Gardner RC (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome 42:9-19

153. Dumolin-Lapegue S, Pemonge MH, Petit RJ (1997) An enlarged set of consensus primers for the study of organelle DNA in plants. Mol Ecol 6:393-397

154. Kohjyouma M, Lee IJ, Iida O, Kurihara K, Yamada K, Makino Y, Sekita S, Satake M (2000) Intraspecific variation in *Cannabis sativa* L. Based on intergenic spacer region of chloroplast DNA. Biol Pharm Bull 23:727-730. https://doi.org/10.1248/bpb.23.727

155. Gilmore S, Peakall R, Robertson J (2007) Organelle DNA haplotypes reflect crop-use characteristics and geographic origins of *Cannabis sativa*. Forensic Sci Int 172:179-190. https://doi.org/10.1016/j.forsciint.2006.10.025

156. Zhang Q, Sodmergen, Liu Y (2003) Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. Plant and Cell Physiol 44:941-951

157. Drew BT, Ruhfel BR, Smith SA, Moore MJ, Briggs BG, Gitzendanner MA, Soltis PS, Soltis DE (2014) Another look at the root of the angiosperms reveals a familiar tale. Syst Biol 63:368-382

158. Demesure B, Sodzi N, Petit RJ (1995) A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Mol Ecol 4:129-131

159. Mello IC, Ribeiro AS, Dias VH, Silva R, Sabino BD, Garrido RG, Seldin L, de Moura Neto RS (2016) A segment of *rbc*L gene as a potential tool for forensic discrimination of *Cannabis sativa* seized at Rio de Janeiro, Brazil. Int J Leg Med 130:353-356. https://doi.org/10.1007/s00414-015-1170-x

160. Dias VH, Ribeiro AS, Mello IC, Silva R, Sabino BD, Garrido RG, Seldin L, Moura-Neto RS (2015) Genetic identification of *Cannabis sativa* using chloroplast *trn*L-F gene. Forensic Sci Int Genet 14:201-202. https://doi.org/10.1016/j.fsigen.2014.10.003

161. Wang S, Shi C, Gao L-Z (2013) Plastid genome sequence of a wild woody oil species, prinsepia utilis, provides insights into evolutionary and mutational patterns

of rosaceae chloroplast genomes. PLoS ONE 8:e73946. https://doi.org/10.1371/journal.pone.0073946

162. Li H, Cao H, Cai YF, Wang JH, Qu SP, Huang XQ (2014) The complete chloroplast genome sequence of sugar beet (*Beta vulgaris* ssp. *vulgaris*). Mitochondrial DNA 25:209-211. https://doi.org/10.3109/19401736.2014.883611

163. Linacre A, Thorpe J (1998) Detection and identification of *Cannabis* by DNA. Forensic Sci Int 91:71-76

164. Fangan BM, Stedje B, Stabbetorp OE, Jensen ES, Jakobsen KS (1994) A general approach for PCR-amplification and sequencing of chloroplast DNA from crude vascular plant and algal tissue. BioTechniques 16:484-494

165. Wilkinson M, Linacre AMT (2000) The detection and persistence of *Cannabis sativa* DNA on skin. Sci Justice 40:11-14. https://doi.org/10.1016/S1355-0306(00)71927-7

166. Kohjyouma M, Lee IJ, Iida O, Kurihara K, Yamada K, Makino Y, Sekita S, Satake M (2000) Intraspecific variation in *Cannabis sativa* L. Based on intergenic spacer region of chloroplast DNA. Biol Pharm Bull 23:727-730

167. Fitch WM, Ayala FJ, National Academy of S. (1995) Tempo and mode in evolution: genetics and paleontology 50 years after simpson. National Academies Press Washington, D.C.

168. Bafeel S, Arif I, A Bakir M, Khan H, H Al Farhan A, Al-Homaidan A, Ahamed A, Thomas J (2011) Comparative evaluation of PCR success with universal primers of maturase k (*mat*K) and ribulose-1, 5-bisphosphate carboxylase oxygenase large subunit (*rbc*L) for barcoding of some arid plants. Plant Omics J 4:195-198

169. Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in land plants: Evaluation of *rbc*L in a multigene tiered approach. Canadian J Bot 84:335-341. https://doi.org/10.1139/b06-047

170. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal w and clustal x version 2.0. Bioinformatics 23:2947-2948. https://doi.org/10.1093/bioinformatics/btm404

171. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) Mega6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725-2729. https://doi.org/10.1093/molbev/mst197

172. Quality assurance standards for forensic DNA testing laboratories (2009). FBI.

173. SWGDAM validation guidelines for DNA analysis methods. (2016) https://docs.wixstatic.com/ugd/4344b0_813b241e8944497e99b9c45b163b76bd.pdf

174. Wictum E, Kun T, Lindquist C, Malvick J, Vankan D, Sacks B (2013) Developmental validation of dogfiler, a novel multiplex for canine DNA profiling in forensic casework. Forensic Sci Int Genet 7:82-91. https://doi.org/10.1016/j.fsigen.2012.07.001

175. Kun T, Lyons LA, Sacks BN, Ballard RE, Lindquist C, Wictum EJ (2013) Developmental validation of mini-dogfiler for degraded canine DNA. Forensic Sci Int Genet 7:151-158. https://doi.org/10.1016/j.fsigen.2012.09.002

176. Berger B, Berger C, Hecht W, Hellmann A, Rohleder U, Schleenbecker U, Parson W (2014) Validation of two canine STR multiplex-assays following the isfg

recommendations for non-human DNA analysis. Forensic Sci Int Genet 8:90-100. https://doi.org/10.1016/j.fsigen.2013.07.002

177. Schury N, Schleenbecker U, Hellmann AP (2014) Forensic animal DNA typing: Allele nomenclature and standardization of 14 feline STR markers. Forensic Sci Int Genet 12:42-59. https://doi.org/10.1016/j.fsigen.2014.05.002

CHAPTER II

Evaluation of a 13-loci STR multiplex system for *Cannabis sativa* genetic identification[1]

This dissertation follows the style and format of *International Journal of Legal Medicine*.

[1] Houston R, Birck M, Hughes-Stamm S, Gangitano D (2016) Evaluation of a 13-loci STR multiplex system for *Cannabis sativa* genetic identification. Int J Legal Med 130:635-647. https://doi.org/10.1007/s00414-015-1296-x

Reprinted with permission from publisher.

Abstract

Marijuana (*Cannabis sativa*) is the most commonly used illicit substance in the USA. The development of a validated method using *Cannabis* short tandem repeats (STRs) could aid in the individualization of samples as well as serve as an intelligence tool to link multiple cases. For this purpose, a modified 13-loci STR multiplex method was optimized and evaluated according to ISFG and SWGDAM guidelines. A real-time PCR quantification method for *C. sativa* was developed and validated, and a sequenced allelic ladder was also designed to accurately genotype 199 *C. sativa* samples from 11 U.S. Customs and Border Protection seizures. Distinguishable DNA profiles were generated from 127 samples that yielded full STR profiles. Four duplicate genotypes within seizures were found. The combined power of discrimination of this multilocus system is 1 in 70 million. The sensitivity of the multiplex STR system is 0.25 ng of template DNA. None of the 13 STR markers cross-reacted with any of the studied species, except for *Humulus lupulus* (hops) which generated unspecific peaks. Phylogenetic analysis and case-to-case pairwise comparison of 11 cases using F_{ST} as genetic distance revealed the genetic association of four groups of cases. Moreover, due to their genetic similarity, a subset of samples (N=97) was found to form a homogeneous population in Hardy-Weinberg and linkage equilibrium. The results of this research demonstrate the applicability of this 13-loci STR system in associating *Cannabis* cases for intelligence purposes.

Keywords: Forensic DNA, Forensic botany, *Cannabis sativa,* Short tandem repeats, Reference population

Introduction

Cannabis sativa L. is a plant cultivated worldwide as a source of fiber (hemp), medicine, and intoxicant [1, 2]. Traditionally, *C. sativa* is divided into two main types: fiber type (hemp) and drug type (marijuana). Marijuana differs from hemp by the presence of a high quantity of the psychoactive drug, Δ9-tetrahydrocannabinol (THC) [3, 4]. In the USA, marijuana is the most commonly used illicit substance [5]. Consequently, marijuana is a highly trafficked drug to and within the USA by organized crime syndicates.

The federal government considers *C. sativa* a Schedule I controlled substance. However, it has become legalized for medical use in 23 states and for adult recreational use in four states (Colorado, Washington, Oregon, and Alaska) and the District of Columbia. Because of legalization, law enforcement faces a unique challenge in tracking and preventing the flow of legal marijuana to states where it is still illegal. Although security measures (barcodes) were implemented to monitor the commercial flow [6], no DNA registry was created due to the prohibitive expense.

The development of a validated method using molecular markers, such as short tandem repeats (STRs) for the genetic identification of *C. sativa* will aid in the individualization of *Cannabis* samples as well as serve as an intelligence tool to link *Cannabis* cases (e.g., illegal traffic at the USA-Mexico border). Specifically, the use of a DNA-based method for identification will allow federal law enforcement agencies (e.g., U.S. Customs and Border Protection (CBP) and Drug Enforcement Administration (DEA)) to form links between cases involving the cross-border trafficking of *Cannabis*.

When identifying marijuana for legal purposes, Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) recommendations require the confirmation of

THC via gas chromatography mass spectroscopy (GCMS), the microscopic confirmation of the presence of cystolithic hairs, and a positive Duquenois-Levine color test [7]. These tests are sufficient for prosecuting an individual for possession of marijuana but do not provide any meaningful intelligence as to the origin or individualization of the plant. However, there are many methods that can be used to individualize and determine the origin of a marijuana sample. These methods include, but are not limited to, palynology [8], chemical profiling [9], isotope ratio mass spectrometry (IRMS) [10, 11], and DNA analysis [12].

DNA has been shown to provide higher resolution for the individualization of marijuana plants as compared to the other techniques [13]. In the 1990s, DNA techniques were developed and evaluated for the purpose of individualizing marijuana, including random amplified polymorphic DNA (RAPD) [14], amplified fragment length polymorphism (AFLP) [15], intersimple sequence repeat amplification (ISSRs) [16], chloroplast and mitochondrial DNA [13], and short tandem repeats (STRs) [17–19]. As STRs are considered the gold standard for human identification, research has focused on the development of STR panels to identify marijuana plants [12]. In Australia, a multiplex STR system was successfully developed for the genetic identification of *C. sativa* [12] followed by a subsequent STR database for marijuana seizures [20]. Howard et al. noted the presence of identical genotypes in the marijuana seizures in the Australian STR database [20].

Identical genotypes occur due to cultivation via clonal propagation instead of sexual propagation. Most growers and dispensaries prefer clonal propagation to maintain consistent quality and potency of their products. For clonal propagation, clippings from the

desired female plants, which contain higher THC levels, are directly rooted in the soil. Clonal propagation results in plants that are genetically identical, while seed propagation results in plants with a unique genetic makeup [21]. In the case of clonal propagation, DNA typing will allow direct linkage of cases to a common grower or distributor.

In the USA, there have been attempts to create an STR database for *Cannabis* [22] as well as extensive research on a hypervariable STR marker, CS1 [23]. However, more comprehensive genetic tools need to be developed to provide a better insight into the genetic variation of marijuana. In addition, none of the previously published reports using *Cannabis* STR profiling have followed two important International Society of Forensic Genetics (ISFG) recommendations for the use of non-human DNA in forensic genetic investigations [24]: (a) the use of sequenced allelic ladders for accurate designation of alleles and interlaboratory STR profile sharing and (b) relevant population and forensic parameters studied in a reference population database of *C. sativa* for random match probability estimations or verification of genetic relatedness.

This study expands upon the earlier work of Köhnemann et al., which described a 15 STR multiplex for the individualization of marijuana [25]. We developed and validated an accurate real-time PCR DNA quantification method for *C. sativa* and evaluated a 13-loci STR multiplex method for genotyping marijuana following ISFG/SWGDAM guidelines (i.e., use of sequenced allelic ladder, sensitivity, species specificity). This STR panel could not only assist law enforcement agencies in verifying legal marijuana products but could also aid in the linkage of cases related to the illegal trade of *Cannabis*. Eventually, the genetic information contained within a sample may be used to link the marijuana to a grower or distributor. This DNA-based method could also be used as a complement to

previously established marijuana profiling programs for intelligence purposes in organizations such as CBP and DEA.

Materials and methods

Sample collection

Marijuana samples (N=199) were obtained from 11 previously processed case sets at the U.S. Customs and Border Protection LSS Southwest Regional Science Center. A minimum of ten specimens were randomly sampled from each case set. For collection, individual marijuana plant fragments (stem or flowers) were cut, with 10 mg of the plant tissue used for this study.

DNA extraction

Plant material was dissected into small pieces with a sterile blade and then homogenized using a Kimble-Chase Kontes™ Pellet Pestle™ (Fisher Scientific, Pittsburgh, PA, USA) with liquid nitrogen. DNA extraction was then performed using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) as per the manufacturer's protocol [26]. This extraction method was previously validated by Miller-Coyle et al. for forensic DNA extraction of C. sativa [15].

Preparation of a *Cannabis* DNA standard using UV spectrophotometry

A *C. sativa* DNA standard was prepared according to a previously published report [27]. Briefly, DNA extracted from five C. sativa samples was pooled and concentrated using a Microcon-100 filter (EMD Millipore, Billerica, MA, USA) by centrifugation at 3000×g. DNA concentration was assessed using an Evolution 60S UV/VIS spectrophotometer (Thermo Fisher Scientific, South San Francisco, CA) and measuring UV absorbance at 260 nm.

DNA quantitation by real-time PCR

DNA samples were quantified by real-time PCR on a StepOne™ Real-Time PCR System (Applied Biosystems, Carlsbad, CA, USA) using SYBR Green PCR Master Mix (Applied Biosystems) and C. sativa specific primers (ANUCS304) [12]. An aliquot of DNA extract (2 µL) was added to 23 µL of master mix. The master mix consisted of 12.5 µL of 2× SYBR Green Master Mix (Applied Biosystems), 0.5 µL ANUCS 304 primers (20 µM) (Integrated DNA Technologies, Coralville, IA, USA), 0.8 µL bovine serum albumin (Sigma-Aldrich, 8 mg/mL), and 9.2 µL deionized H2O. The real-time PCR cycling conditions were as follows: initial denaturation stage (10 min, 95 °C) and cycling stage (15 s at 95 °C followed by 1 min at 60 °C) for 40 cycles. The previously prepared *C. sativa* DNA standard was serially diluted (12.75 to 0.01 ng/µL) to generate a calibration curve. Cycle threshold (Ct) values were determined at 0.2 ΔRn using the automatic baseline algorithm. Linearity range was assessed by R^2 estimation, and a minimum correlation of 99 % was accepted for quantification.

Validation studies of the qPCR method for *C. sativa* DNA quantitation

The validation studies for the *Cannabis* DNA quantitation assay included (i) reproducibility and precision, (ii) sensitivity, and (iii) species specificity. For this purpose, eight *Cannabis* DNA standards (12.75, 3.19, 1.59, 0.40, 0.20, 0.10, 0.02, and 0.01 ng/µL) along with three *Cannabis* control DNA samples were run in duplicate in 15 separate real-time PCR runs. Real-time PCR amplification efficiencies were estimated using the slope of the standard plot regression line: efficiency=$[10^{(-1/\text{slope})}]-1$. To determine species specificity, the real-time PCR assay was used to amplify non-*C. sativa* DNA samples including *Ocimum basilicum* (basil), *Canis lupus familiaris* (dog), *Bos taurus* (beef),

Humulus lupulus (Hops), *Homo sapiens* (human), *Mentha* (mint), *Nicotiana tabacum* (tobacco), *Allium cepa* (onion), and *Felis catus* (cat).

Loci and multiplex amplification conditions

Cannabis STR profiling was conducted in a 13-loci multiplex format modified from a previous report [25]. Thirteen previously published *Cannabis* microsatellites (E07 CANN1, ANUCS 302, H09 CANN2, D02 CANN1, C11 CANN1, B01 CANN1, B05 CANN1, H06 CANN2, ANUCS 305, ANUCS 308, ANUCS 301, CS1, and ANUCS 501) were used in this study (Table 2.1.). Amplification of these markers was performed via PCR using the Type-it™ Microsatellite PCR Kit (Qiagen) on the Eppendorf Master Cycler Gradient (Eppendorf, Hamburg, Germany). The PCR reactions were prepared at a 12.5-µL volume using 0.5 ng of template DNA. An aliquot of DNA (2 µL) from each sample was added to 10.5 µL of PCR master mix. The PCR master mix consisted of 6.25 µL of 2× Type-it™ Multiplex PCR Master Mix (Qiagen), 1.25 µL 10× Primer mix, 1.25 µL 5 Q-Solution (Qiagen), 0.4 µL 8 mg/mL bovine serum albumin (BSA) (Sigma-Aldrich St. Louis, MO, USA), and 1.35 µL deionized H2O. Forward primers were labeled with four different fluorescent dyes (6-FAM™, PET™, NED™, and VIC™, Life Technologies), and final optimal concentrations of forward and reverse primers are displayed in Table 2.1. PCR cycling conditions were as follows: activation for 5 min at 95 °C, followed by 30 cycles of 30 s at 95 °C, 90 s at 60 °C, 30 s at 72 °C and a final extension of 30 min at 60 °C. Every set of PCR reactions included one negative and one positive control (sample #1-D1).

Table 2.1. Characteristics of 13 *Cannabis* STR markers used in this study

Marker	Dye	STR Motif	Type of Repeat	Observed alleles	Primer concentration (μM)	Genbank Accession No.
D02	6-FAM™	(GTT)	Simple	6, 7, 8	0.04	KT203591-2
C11	6-FAM™	$(TGG)_x$ $(TGG)_y$	Compound/Indel	13, 14, 15, 21	0.05	KT203583-5
H09	6-FAM™	(GA)	Simple	11, 13, 16, 18, 19, 21,23,24, 25	0.08	KT203598-602
B01	6-FAM™	$(GAA)_x$ (A)$(GAA)_y$	Complex	11, 13, 14, 15	0.09	KT203579-80
E07	VIC™	(ACT)	Simple	7, 8, 9	0.30	KT203593-5
305	VIC™	(TGG)	Single	4, 6, 8, 11	0.08	KT203571-3
308	VIC™	(TA)	Simple	5, 8, 9, 12	0.13	KT203574-6
B05	VIC™	(TTG)	Simple	3, 7, 8, 9, 10	0.03	KT203581-2
H06	VIC™	(ACG)	Simple	7, 8, 9	0.07	KT203596-7
501	NED™	(TTGTG)	Simple	4, 5, 6	0.10	KT203577-8
CS1	NED™	(CACCAT)	Simple	10, 12, 13, 16, 17, 23, 24, 25, 26, 27, 28, 29, 32	0.14	KT203586-90
302	PET™	$(ACA)_x$ $(ACA)_y$ $(ACA)_z$	Compound	29, 31, 33, 34, 35, 36, 37	0.08	KT203569-70
301	PET™	(TTA)	Simple	15, 16, 17, 24, 25	0.30	KT203566-8

Capillary electrophoresis and genotyping

Fragment separation and detection of PCR products was performed on the 3500 Genetic Analyzer (Applied Biosystems). An aliquot (1 μL) of PCR product was added to 10 μL of cocktail (9.5 μL Hi-Di Formamide® and 0.5 μL LIZ® 500 Size Standard, Applied Biosystems). Samples were then denatured for 5 min and loaded on the 3500 Genetic Analyzer (Applied Biosystems) and run using the following conditions: oven: 60 °C; prerun 15 kV, 180 s; injection 1.6 kV, 8 s; run 19.5 kV, 1330 s; capillary length 50 cm; polymer: POP-7™; and dye set G5. A customized bin set was designed, and an allelic ladder (generated from sequence data for each marker) was included with each injection to ensure accurate genotyping. Genotyping was performed using GeneMapper v. 4.1 software (Applied Biosystems). The analytical threshold was set at 150 relative fluorescence units (RFUs).

Allelic ladder design

Fifty C. sativa samples were screened initially to determine the variability of alleles observed in the population. Using the most common alleles observed, an allelic ladder was generated according to previous reports [28, 29]. Briefly, these samples were amplified in single PCR, and then the concentration of all amplicons was balanced, diluted approximately 1:1000, and subsequently reamplified with 20 cycles. These reamplified products represented the allelic ladder for each STR marker. Each of these single STR allelic ladders was amplified to attain high RFU values (approximately 24,000 RFUs). These amplified single allelic ladders were then diluted 1:1000 in TE buffer for future use as a second-generation ladder. All of these high RFU single STR marker allelic ladders

were then combined prior to electrophoresis to attain a combined allelic ladder for all 13 loci tested.

Allele sequencing

Two to five homozygous samples representing the most common alleles were selected for sequencing. Indeed, single alleles selected from heterozygous samples were previously isolated by electrophoresis on a 2 % high-resolution agarose gel (Sigma-Aldrich) and purified using the MinionElute Gel Extraction Kit (Qiagen) according to the manufacturer's instructions [30]. Homozygote samples were preferentially chosen for simplicity. However, for marker CS1, heterozygote samples were selected due to the highly polymorphic nature of CS1. PCR amplification and cycling sequencing was performed on the Veriti® Thermal Cycler (Applied Biosystems) using the BigDye® Direct Cycle Sequencing Kit (Applied Biosystems) as per the manufacturer's protocol with the exception of a 60 °C annealing temperature (instead of 62 °C) [31]. Samples were loaded on the 3500 Genetic Analyzer (Applied Biosystems) and run using the following conditions: oven 60 °C; prerun 18 kV, 60 s; injection 1.6 kV, 8 s; run 19.5 kV, 1020 s; capillary length 50 cm; polymer: POP-7™; and dye set Z. Data analysis was performed using the Sequencing Analysis software v.5.4 (Applied Biosystems). Sequences were then aligned and proofread using the Geneious Pro Software R7.1.9 (Biomatters, Auckland, New Zealand). Previous research from Valverde et al. and ISFG recommendations from human-specific STR loci were followed when determining the nomenclature of the alleles [32–35]. Sequences were submitted to Genbank (accession no. KT203566 to KT203602).

Sensitivity study

To determine the sensitivity range of this PCR multiplex, dilutions of four DNA samples were prepared to generate template DNA amounts of 1.0 ng, 500, 250, 125, 62.5, and 31.2 pg. The 24 dilutions were amplified in triplicate with the 13-loci STR multiplex developed in this study to measure the lowest amount of template DNA that reproducibly produced full profiles.

Specificity study

To assess specificity, the 13 STR markers were used to amplify non-*C. sativa* DNA. Samples tested included *Ocimum basilicum* (basil), *Bos taurus* (beef), *Daucus carota* (carrot), *Felis catus* (cat), *Gallus domesticus* (chicken), *Canis lupus familiaris* (dog), *Allium sativum* (garlic), *Humulus lupulus* (Hops), *Homo sapiens* (human), *Ilex paraguariensis* (mate), *Mentha* (mint), *Allium cepa* (onion), *Origanum vulgare* (oregano), *Petroselinum crispum* (parsley), *Pinus echinata* (pine), *Sus scrofa domesticus* (pork), *Rosmarinus officinalis* (rosemary), *Origanum vulgare ssp. Hirtum* (spicy oregano), *Nicotiana tabacum* (tobacco), and *Solanum lycopersicum* (tomato). Plant samples were extracted using the Qiagen DNeasy Plant Mini Kit as per the manufacturer's protocol [26]. Animal samples were extracted using the QIAamp DNA Investigator Kit as per the manufacturer's protocol [36]. For human DNA, TaqMan® control genomic human DNA (Applied Biosystems) was used. The DNA concentration was determined using a UV spectrophotometer by measuring absorbance at 260 nm, and the quality of the DNA extraction was assessed via electrophoresis on a 2 % agarose gel. Extracts were then amplified (2–10 ng) in duplicate using the developed STR multiplex to detect cross-reaction amplification across the various species.

Additional studies with loci ANUCS301, ANUCS302, ANUCS308, and B01-CANN1

Annealing temperatures were determined for primers of loci ANUCS301, ANUCS302, ANUCS308, and B01 CANN1 using an Eppendorf Master Cycler Gradient (Eppendorf). PCR reactions were prepared at a 12.5-µL volume using 1.0 ng of template DNA. An aliquot of DNA (2 µL) from each sample was added to 10.5 µL of PCR master mix. The PCR master mix consisted of 6.25 µL of 2× HotStarTaq *Plus* Master Mix (Qiagen), 1.25 µL 2 µM Primer mix, 1.25 µL 5× Q-Solution (Qiagen), 0.4 µL 8 mg/mL BSA (Sigma-Aldrich), and 1.35 µL deionized H2O. Gradient PCR cycling conditions were as follows: activation for 5 min at 95 °C, followed by 30 cycles of 30 s at 94 °C, 30 s at a gradient of (60±10 °C; 12 wells), 30 s at 72 °C, and a final extension of 30 min at 60 °C. The optimal annealing temperature was determined via electrophoresis on a 2 % agarose gel. Ten previously genotyped homozygote *Cannabis* samples were amplified with markers ANUCS301, ANUCS302, ANUCS308, and B01 CANN1 at their corresponding annealing temperatures using the previously described HotStarTaq *Plus* protocol (Qiagen). PCR products were run and genotyped as described in the "Capillary electrophoresis and genotyping" section.

Statistical analysis

For STR markers, the number of multi-locus genotypes and the genotype sharing among samples were determined. Phylogenetic analysis of different seizures using the unweighted pair group method using arithmetic averaging (UPGMA) and coefficient of coancestry F_{ST} as genetic distance were estimated with the Genetic Data Analysis (GDA) software [37]. Evaluation of population differentiation between seizures was assessed

using a case-to-case pairwise comparison using F_{ST} as genetic distance with the Arlequin v. 3.5 software [38]. The p value for statistically significant differences was set at 0.05.

For the reference population database (N=97), allele frequencies and parameters of forensic interest were estimated using the PowerStats v.12 software [39]. In addition, population genetic statistics (number of alleles, observed heterozygosity, expected heterozygosity) as well as Hardy-Weinberg equilibrium and linkage disequilibrium tests were performed on this reference population with the GDA software. Null allele analysis was performed using the Genepop v.4.2 software [40]; corrected allele frequencies were also reported. p values for statistically significant differences were set at 0.05, and Bonferroni correction for multiple comparisons was applied when applicable.

Results and discussion

DNA extraction and quantitation

DNA was successfully extracted from all *C. sativa* samples (N=199). The average amount (±standard deviation) of DNA extracted was 34.7±60.6 ng/mg of plant tissue. The amount of DNA extracted from flower and stem tissues was 47.46± 73.90 and 6.92±6.54 ng/mg, respectively. The greater amount of DNA from the flowering part of the marijuana plant suggests that flower should be the preferential target for STR genotyping. An adequate amount of DNA was still extracted from the stem, but it should be noted that the pulverization step in the extraction procedure was more difficult with the stem due to its high cellulose content. The SWGDAM standards 9.4 and 9.5 state that the amount of human DNA should be quantified with quantitation standards in forensic samples prior to nuclear DNA amplification [41]. These SWGDAM standards should also be followed for

non-human DNA. However, to date. this is the first publication regarding *Cannabis* STR typing using a real-time PCR method for DNA quantitation.

Validation studies of the *Cannabis* qPCR quantitation method

Data generated for all eight quantification standards (Table 2.2.) and linear regression of the standard curve (Table 2.3.) from 15 separate real-time PCR assays demonstrated high reproducibility, precision, and sensitivity. The inter-run precision, expressed as the percent coefficient of variation of Ct (%CV=100×(standard deviation/mean)) had an average of 2.6 %. Among 15 individual assays, 12.75 ng/μL of the purified standard exhibited a Ct value of 20.3 (range 19.3–21.27). The subsequent fourfold dilution (3.19 ng/μL) exhibited a value of Ct of 21.86 (range 20.81–22.92). The difference in Ct values between each successive dilution was 1.89 and 1.12 for fourfold and twofold dilutions, respectively. The sensitivity of the quantitation assay was 10 pg/μL with an average Ct of 30.45 (range 28.77–31.65). As expected, standards 7 and 8 (0.02 and 0.01 ng/μL, respectively) exhibited the highest degree of variation with an average Ct (±standard deviation) of 29.10±0.73 and 30.45±0.72, respectively. The three *Cannabis* samples, included as a positive control during each real-time PCR run, tested the functionality of the assay and monitored reproducibility and precision. All three controls exhibited low Ct and quantity estimate variation between runs. As expected, all non-C. sativa samples produced negative results.

Table 2.2. Standard Ct data among 15 separate real-time PCR assays

Standard	*Cannabis* DNA (ng/µL)	Average	Standard deviation	Minimum	Maximum	Range
1	12.75	20.30	0.60	19.30	21.27	1.98
2	3.19	21.86	0.62	20.81	22.92	2.10
3	1.59	22.90	0.66	21.79	23.95	2.16
4	0.40	25.09	0.63	23.85	26.08	2.22
5	0.20	26.24	0.61	25.22	27.39	2.17
6	0.10	27.19	0.63	26.04	28.45	2.41
7	0.02	29.10	0.73	28.02	30.94	2.92
8	0.01	30.45	0.72	28.77	31.65	2.89

Table 2.3. Linear regression data from 15 separate real-time PCR runs

Run	Slope	Amplification efficiency (%)	R^2	Y-intercept
1	-3.35	98.80	0.996	23.89
2	-3.44	95.49	0.995	24.40
3	-3.37	98.19	0.992	23.66
4	-3.42	96.06	0.993	24.77
5	-3.49	93.58	0.993	23.82
6	-3.45	95.03	0.991	23.04
7	-3.34	99.17	0.994	22.77
8	-3.35	98.96	0.993	22.93
9	-3.32	100.25	0.993	23.57
10	-3.39	97.08	0.997	24.52
11	-3.33	99.87	0.992	24.32
12	-3.36	98.56	0.995	24.23
13	-3.40	96.96	0.994	23.76
14	-3.49	93.32	0.992	23.02
15	-3.35	99.01	0.997	23.89
Average	-3.39	97.36	0.994	23.77
Standard deviation	0.06	2.22	0.002	0.62
Coefficient of variation	0.02	0.02	0.002	0.03
Minimum	-3.49	93.32	0.991	22.77
Maximum	-3.32	100.25	0.997	24.77
Range	0.18	6.92	0.006	2.00

STR multiplex

A *Cannabis* multiplex STR system previously reported by Köhnemann et al. was used as reference for this study [25] with the following modifications: (a) primer concentrations optimized with Type-it™ Microsatellite PCR Kit (Qiagen), (b) use of 13 out of the 15 STR loci with a different combination of fluorescent dyes, and (c) use of a sequenced allelic ladder for accurate STR genotyping. After initial evaluation, two of the original 15 STR markers, B02 and H11, were removed due to close proximity to ANUCS302 and inefficient PCR amplification, respectively. All samples (N=199) were successfully amplified under the optimized multiplex conditions. However, only 127 out of 199 samples (64 %) resulted in full DNA profiles. The remaining 37 % of samples resulted in partial DNA profiles with maximum locus drop-out of two STR loci in any one sample. The loci most affected by locus drop-out were ANUCS301, ANUCS302, ANUCS308, and B01-CANN1 (22, 5, 11, and 9 %, respectively). Locus drop-out was most likely due to primer-primer interaction and/or weak primer binding. Primer-primer interaction analysis was performed using the Multiplex Manager software v.1.2 [42], and interactions were detected for the following pairs: 302/D02, 302/C11, 302/308, B05/308, B02/H11, and B02/301. This primer-primer interaction may also explain the severe inter-locus imbalance observed in STR markers B01, 308, and 301 (Fig. 2.1.).

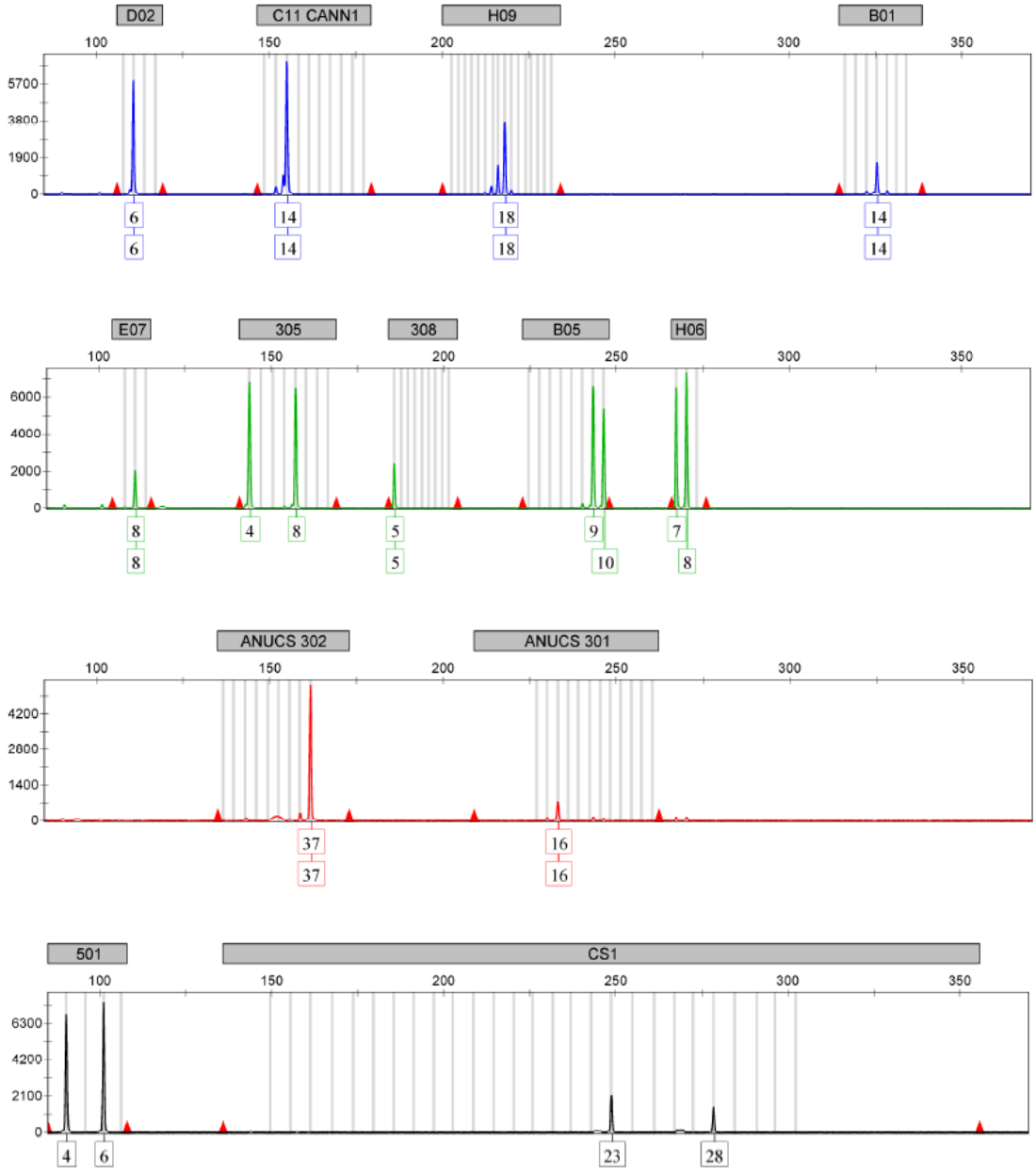


Fig. 2.1. Multiplex profile of 13 *Cannabis* STR loci using 0.5 ng of control template DNA (sample #1-D1)

To determine if weak primer binding and eventually primer-primer interaction were the use of allele drop-out, we experimentally determined the annealing temperatures of these four problematic markers. The annealing temperatures of markers ANUCS301, ANUCS302, ANUCS308, and B01 CANN1 were 53, 53, 55, and 55 °C, respectively. From ten previously genotyped homozygote *Cannabis* samples (at 60 °C) five, two, one, and eight individuals resulted to be heterozygotes for loci ANUCS301, ANUCS302, ANUCS308, and B01, respectively (Fig. 2.2.). Only one STR marker, H09, exhibited some difficulties for automatic allele calling due to high stutter peaks.

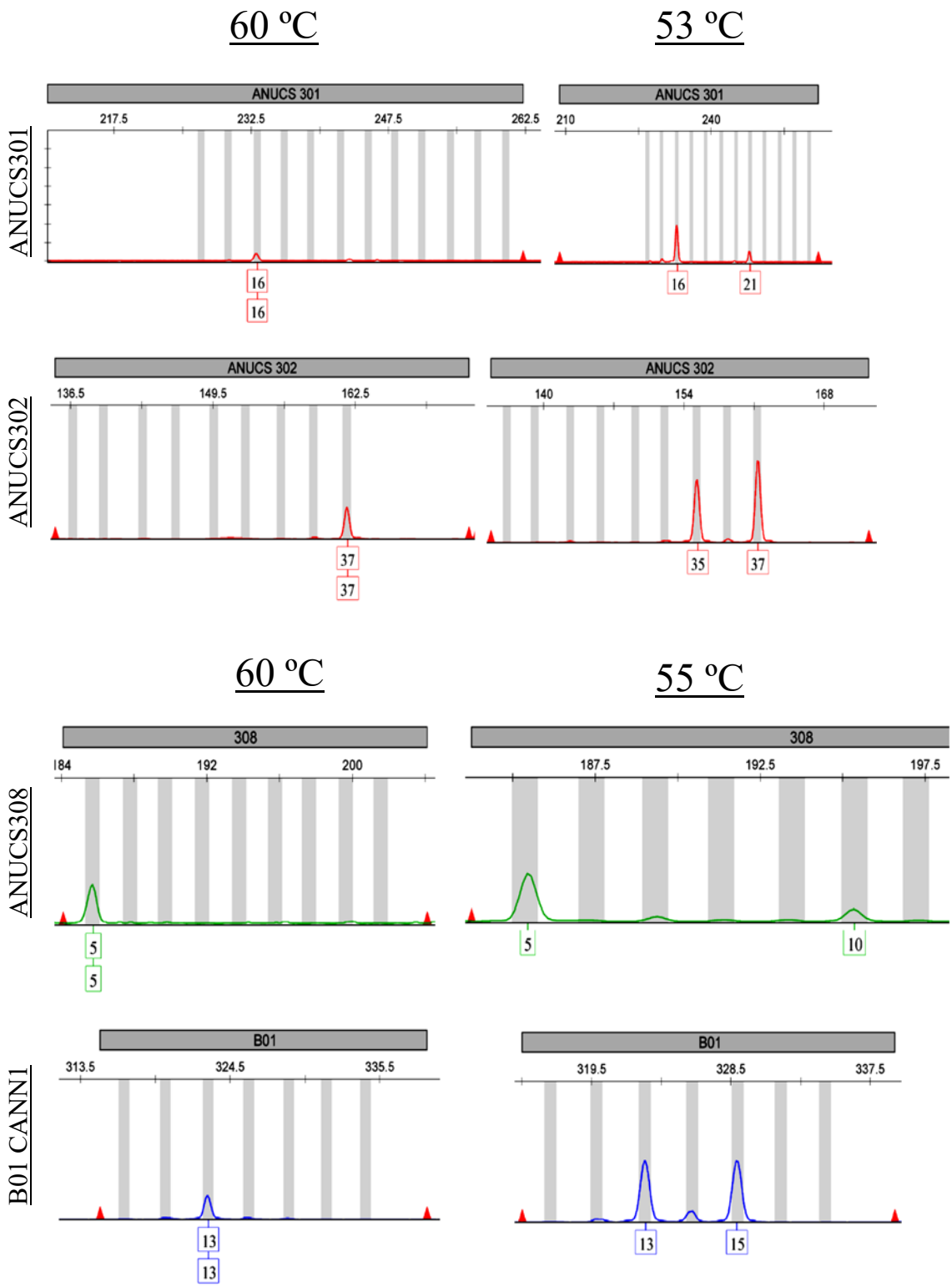


Fig. 2.2. Electropherograms of homozygote *Cannabis* samples (at 60 °C, left) displaying the recovery of sister alleles when amplified at their specific annealing temperatures (53 or 55 °C, right)

Allelic ladder and sequencing

For all 13-STR loci, an allelic ladder was developed with the most frequently observed alleles in the sample population (Fig. 2.3). The allelic ladder contained 56 alleles across the 13 STR loci (Fig. 2.3). Nomenclature following international guidelines was used to designate the allele calls [32]. In addition, the number of repeats for two to eight alleles per marker was confirmed via sequencing to ensure accurate nomenclature of the allelic ladder and confirmation of published sequencing results [32]. The sequencing results from the previous study were confirmed with the most commonly observed repeat motifs reported in Table 1. To date, this is the first publication reporting the use of an allelic ladder for *Cannabis* STR genotyping. The use of an allelic ladder is necessary for accurate DNA genotyping as well as for sharing STR data between labs. In addition, the use of an allelic ladder for STR genotyping is one of the ISFG guidelines for application in non-human DNA testing in a forensic setting [24].

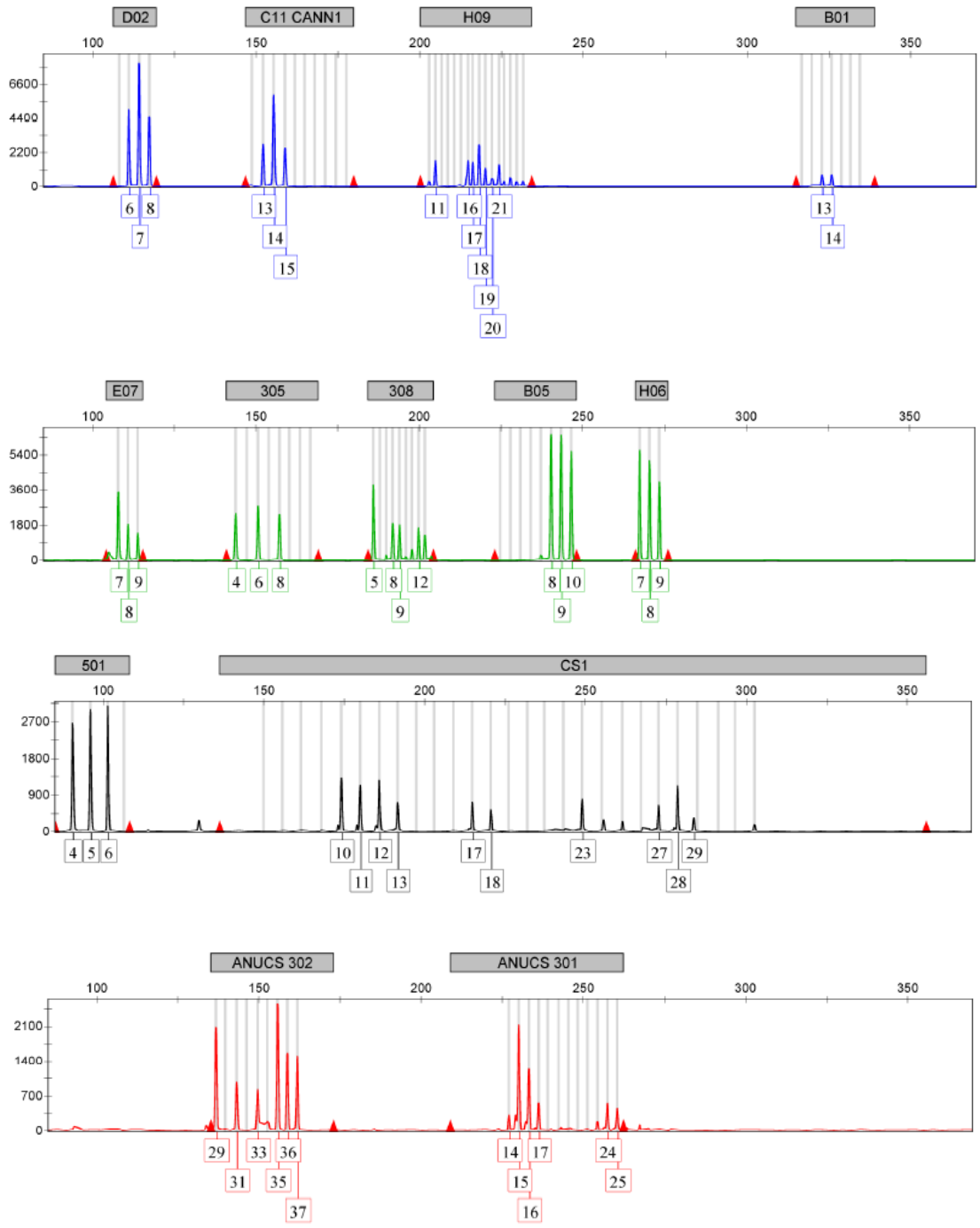


Fig. 2.3. Allelic ladder for 13 *Cannabis* STR loci with design based on sequence data obtained from most commonly observed alleles

STR multiplex validation studies

The sensitivity of the 13 loci STR multiplex was determined to be 0.25 ng by amplifying amounts of template DNA ranging from 0.03 to 1 ng. Allele drop-out and severe peak imbalance was observed when the template DNA was at, or below, 0.06 ng (Fig. 2.4). For the STR multiplex, the optimal input amount of DNA was determined to be 0.5 ng. Split peaks and off ladder peaks were observed for input amounts above 1.0 ng. Due to this narrow optimal range, it is critical to use an accurate DNA quantitation method (such as real-time PCR) to ensure an accurate input amount of DNA for PCR. When testing species specificity, STR genotyping showed that none of the 13 STR markers cross-reacted with any of the species tested except for *H. lupulus*, which generated unspecific peaks (106, 142, and 165 bp in the green dye channel). This unspecific cross-reactivity of H. lupulus was previously reported [12]. *H. lupulus* is closely related genetically to *C. sativa* as they both belong to the same family, Cannabaceae [43].

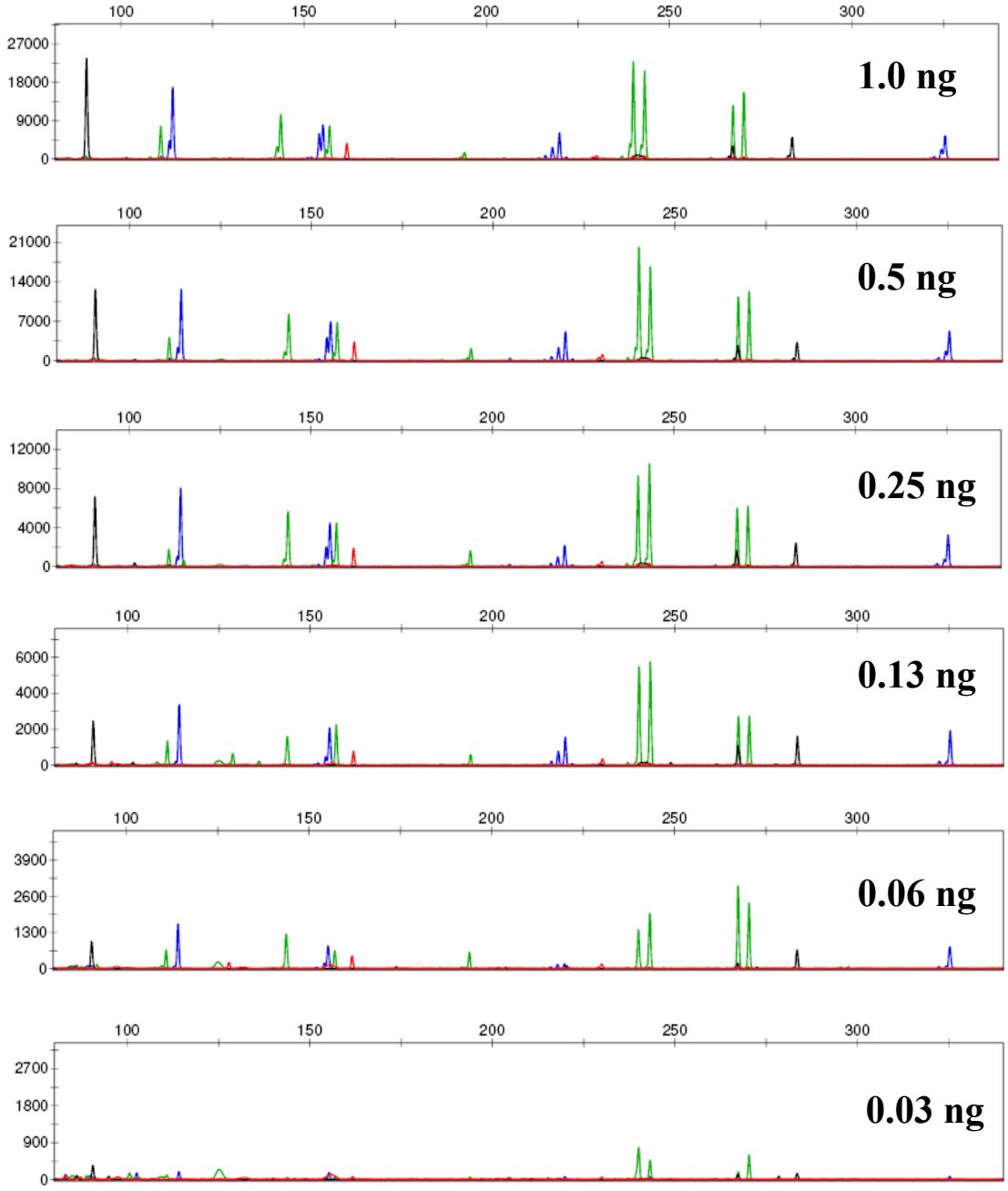


Fig. 2.4. Representative electropherograms from the sensitivity study using the Qiagen Type-it™ Microsatellite PCR Kit protocol overlaying the *blue*, *green*, *yellow*, and *red* dye channels for different amounts of template DNA

Statistical analysis

Distinguishable DNA profiles were generated from the 127 samples that generated full STR profiles. Four duplicate genotypes within seizures were found. From this general analysis of the STR profiles, and the lack of clonal material, it can be concluded that the analyzed *Cannabis* samples from Mexico were propagated from seeds. Nevertheless, other studies have reported a high incidence of *Cannabis* clonal material in seizures in Germany [25] and Australia [20]. Duplicate genotypes within seizures are not unexpected due to the sample collection method used. In addition, the presence of eight mixed-DNA samples was also detected; this may be due to the fact that some of the samples were previously ground and mixed.

Phylogenetic analysis and case-to-case pairwise comparison of 11 cases using F_{ST} as genetic distance revealed the genetic association of four pairs of cases (Fig. 2.5., Table 2.4.). Using the UPGMA method with F_{ST} as genetic distance, it was determined that genetic similarities exist between the following cases: 2 and 5, 6 and 7, 8 and 9, and cases 11, 3, and 4. No statistical significant differences were detected for any of these pair of cases ($p>0.05$) (Table 2.4.).

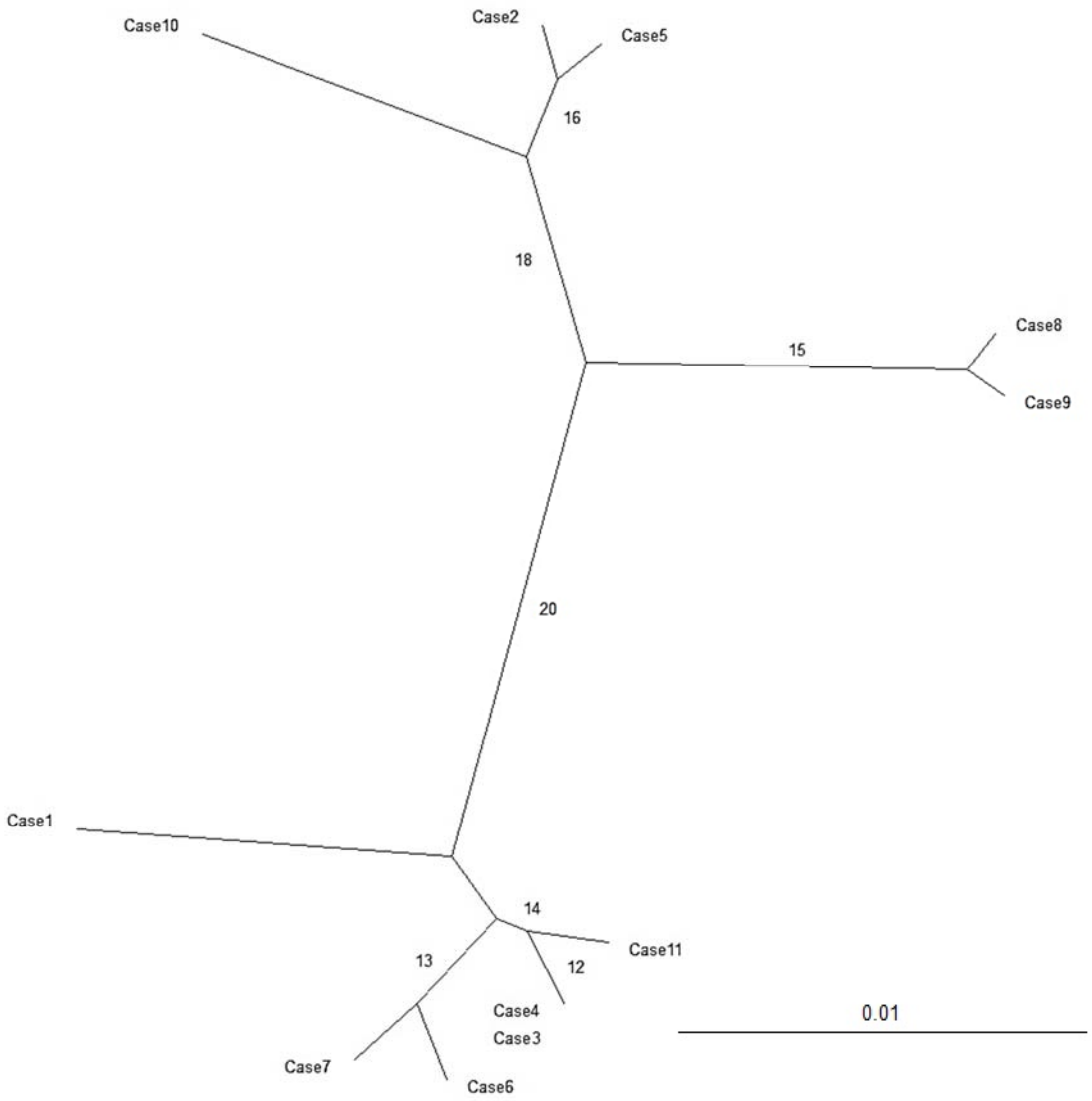


Fig. 2.5. UPGMA tree depicting genetic distances among 11 *Cannabis* sample sets (*N*=199) seized at the Mexico-US border, FST was set as genetic distance

Table 2.4. Case-to-case comparison among 11 *Cannabis* sample sets seized at the Mexico-US border by pair-wise genetic-distance analysis based on FST

Case ID	1	2	3	4	5	6	7	8	9	10
2	0.08960 (0.00000**)									
3	-0.01167 (0.51062)	0.04611 (0.00691**)								
4	0.02742 (0.02963*)	0.03911 (0.00247**)	- 0.01466 (0.95210)							
5	0.02423 (0.11605)	-0.00312 (0.50469)	0.02468 (0.08346)	0.01788 (0.11358)						
6	0.06433 (0.00691**)	0.0273 (0.11358)	0.02259 (0.11556)	0.02296 (0.07111)	0.02926 (0.18123)					
7	0.04027 (0.00444**)	0.01916 (0.04198*)	0.02325 (0.01432*)	-0.00617 (0.85728)	0.00201 (0.31358)	0.01104 (0.19704)				
8	0.11683 (0.00099**)	0.0468 (0.06222)	0.06553 (0.02963*)	0.04909 (0.00049**)	0.0578 (0.05235)	0.03854 (0.13827)	0.06101 (0.00741**)			
9	0.06475 (0.00741**)	0.04489 (0.02617*)	-0.00285 (0.33975)	-0.00187 (0.65481)	0.02722 (0.11901)	0.01005 (0.41432)	0.00419 (0.26469)	0.01702 (0.22815)		
10	0.06811 (0.00000**)	0.03594 (0.03259*)	0.04528 (0.00444**)	0.04068 (0.00198**)	0.0198 (0.16741)	0.04399 (0.03704*)	0.04071 (0.00346**)	0.04764 (0.06519)	0.03377 (0.05136)	
11	0.03821 (0.00790**)	0.03894 (0.00840**)	-0.01068 (0.76247)	0.00663 (0.15012)	0.0257 (0.04938*)	0.02752 (0.05580)	0.00865 (0.14963)	0.03904 (0.01580*)	0.00379 (0.39753)	0.01252 (0.15160**)

As shown in Fig. 2.5., the majority of cases exhibit genetic similarities. Moreover, due to this genetic similarity (common origin) determined by phylogenetic analysis, a subset of samples (cases 3, 4, 11; N=97) was determined to have an F_{ST} close to zero, confirmed by evaluation of 95 % confidence interval bootstrap analysis. This evidence strongly suggests that cases 3, 4, and 11 (N=97) belong to the same population.

No departures from linkage disequilibrium were detected for any of the STR markers studied in this reference population. However, departures from Hardy-Weinberg equilibrium were detected for STR markers B01, 308, 301, and 302 ($p<0.0038$) (Table 2.5.). Moreover, further analysis showed that these departures were due to the presence of null alleles. These findings are consistent with the severe inter-locus imbalance observed for these four markers suggesting a primer-binding and/or primer-primer interaction issue. Allele frequencies corrected for the presence of null alleles are reported in Table 2.5. These observed allele frequencies in the reference population could then be used to calculate parameters of forensic interest (Table 2.6) as well as random match probability estimations. The combined power of discrimination for the 13-locus multiplex is 1 in 70 million. To date, this is the first report of a *Cannabis* STR reference population for forensic purposes.

Table 2.5. Allele frequencies and Hardy-Weinberg evaluation of 13 *Cannabis* STR loci in a population sample of cases seized (Cases #3, #4 and #11) at the Mexico-US border (97 individuals, n = 194 chromosomes)

Allele	D02	C11	H09	B01	E07	305	308	B05	H06	501	CS1	302	301
3								0.005					
4						0.375				0.542			
5							*0.706*			0.016			
6	0.468					0.109				0.443			
7	0.490				0.247			0.021	0.330				
8	0.072				0.389	0.510	*0.012*	0.584	0.608				
9					0.363		*0.121*	0.326	0.062				
10								0.063			0.027		
11			0.087	*0.005*		0.005							
12							*0.012*				0.005		
13		0.297	0.011	*0.344*							0.011		
14		0.688		*0.441*									
15		0.010		*0.011*									*0.440*
16			0.125								0.005		*0.167*
17											0.086		
18			0.489										
19			0.082										
20													
21		0.005	0.109										
22													
23			0.087								0.226		*0.006*
24			0.005								0.011		*0.164*

(continued)

Allele	D02	C11	H09	B01	E07	305	308	B05	H06	501	CS1	302	301
25			0.005								0.022		*0.006*
26											0.022		
27											0.269		
28											0.237		
29											0.070	*0.011*	
30													
31												*0.547*	
32											0.011		
33												*0.027*	
34													
35												*0.188*	
36												*0.067*	
37												*0.038*	
HWE	0.556	0.758	0.083	0.00001*	0.102	0.461	0.00001*	0.576	0.133	0.253	0.173	0.0013*	0.0001*

HWE: Hardy-Weinberg equilibrium probability values of exact test (3200 shufflings).
Allele frequencies corrected for the presence of null alleles in italics (B01, 308, 302 and 301 loci).
* Statistically significant differences at 0.0038 levels (Bonferroni correction).

Table 2.6. Parameters of forensic interest of 13 analyzed *Cannabis* STR loci

	D02	C11	H09	B01	E07	305	308	B05	H06	501	CS1	302	301
Ho	0.557	0.417	0.630	0.226	0.547	0.563	0.155	0.516	0.515	0.438	0.753	0.389	0.269
He	0.566	0.441	0.715	0.522	0.656	0.590	0.312	0.551	0.520	0.513	0.811	0.566	0.609
PIC	0.470	0.360	0.690	0.410	0.580	0.500	0.280	0.470	0.430	0.400	0.780	0.520	0.540
PD	0.718	0.610	0.887	0.664	0.820	0.748	0.431	0.733	0.676	0.671	0.932	0.738	0.746

Ho: observed heterozygosity, He: expected heterozygosity, PIC: polymorphic information content, PD: power of discrimination.

Conclusion

In summary, a real-time PCR method for *Cannabis* DNA quantitation was developed and validated, and a 13-locus. *Cannabis* STR multiplex system was optimized and evaluated. In addition, an allelic ladder was developed for accurate genotyping. The system was determined to be specific for marijuana, and its sensitivity was as low as 0.25 ng. A reference *Cannabis* population database with associated allele frequencies for forensic purposes was also developed. In order to implement this STR system in a crime laboratory, an internal validation is required before its use, with particular attention to determining the stutter thresholds due to the dinucleotide markers (e.g., H09). Caution should be taken regarding interlocus balance as primers need to be redesigned or cycling conditions needs to be optimized to ensure optimal annealing for all 13 STR markers.

Future studies will include the development of an STR multiplex that includes tetranucleotide markers (replacing dinucleotide markers) and the use of massive parallel sequencing (MPS) with *Cannabis* STR panels.

Acknowledgments

We would like to thank all staff and personnel at the U.S. Customs and Border Protection LSS Southwest Regional Science Center for their great assistance and help with this project.

References

1. Small E, Cronquist A (1976) A practical and natural taxonomy for *Cannabis*. Taxon 25:405–435
2. Adams IB, Martin BR (1996) *Cannabis*: pharmacology and toxicology in animals and humans. Addiction 91:1585–1614
3. Werf HVD, Mathussen EWJM, Haverkort AJ (1996) The potential of hemp (*Cannabis sativa* L.) for sustainable fibre production: a crop physiological appraisal. Ann Appl Biol 129:109–123. https://doi.org/10.1111/j.1744-7348.1996.tb05736.x
4. Anderson P (2006) Global use of alcohol, drugs and tobacco. Drug Alcohol Rev 25:489–502
5. Center for Behavioral Health Statistics and Quality (2014) Results from the 2013 National Survey on Drug Use and Health: summary of national findings. U.S. Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. http://www.samhsa.gov/data/sites/default/files/NSDUHresultsPDFWHTML2013/Web/NSDUHresults2013.pdf. Accessed April 29 2015
6. Lucas AA (2014) Colorado—world's first fully regulated recreational marijuana market—collects $2M in recreational pot taxes. International Business Times. http://au.ibtimes.com/coloradoworlds-first-fully-regulated-recreational-marijuana-marketcollects-2m-recreational-pot. Accessed April 29 2015

7. Drug Enforcement Administration's Special Testing and Research Laboratory (2005) Monograph: marijuana. http://www.swgdrug.org/Monographs/MARIJUANA.pdf. Accessed July 2 2015

8. Bryant VM, Jones GD (2006) Forensic palynology: current status of a rarely used technique in the United States of America. Forensic Sci Int 163:183–197. https://doi.org/10.1016/j.forsciint.2005.11.021

9. Brenneisen R, elSohly MA (1988) Chromatographic and spectroscopic profiles of *Cannabis* of different origins: part I. J Forensic Sci 33:1385–1404

10. Shibuya EK, Souza Sarkis JE, Neto ON, Moreira MZ, Victoria RL (2006) Sourcing Brazilian marijuana by applying IRMS analysis to seized samples. Forensic Sci Int 160:35–43. https://doi.org/10.1016/j.forsciint. 2005.08.011

11. Shibuya EK, Sarkis JES, Negrini-Neto O, Martinelli LA (2007) Carbon and nitrogen stable isotopes as indicative of geographical origin of marijuana samples seized in the city of São Paulo (Brazil). Forensic Sci Int 167:8–15. https://doi.org/ 10. 1016/j.forsciint.2006.06.002

12. Howard C, Gilmore S, Robertson J, Peakall R (2008) Developmental validation of a *Cannabis sativa* STR multiplex system for forensic analysis. J Forensic Sci 53:1061–1067. https://doi.org/10.1111/j.1556-4029.2008.00792.x

13. Gilmore S, Peakall R, Robertson J (2007) Organelle DNA haplotypes reflect crop-use characteristics and geographic origins of *Cannabis sativa*. Forensic Sci Int 172:179–190. https://doi.org/10.1016/j. forsciint.2006.10.025

14. Gillan R, Cole MD, Linacre A, Thorpe JW, Watson ND (1995) Comparison of *Cannabis sativa* by random amplification of polymorphic DNA (RAPD) and HPLC

of cannabinoids: a preliminary study. Sci Justice 35:169–177. https://doi.org/10.1016/s1355-0306(95)72658-2

15. Miller Coyle H, Shutler G, Abrams S, Hanniman J, Neylon S, Ladd C, Palmbach T, Lee HC (2003) A simple DNA extraction method for marijuana samples used in amplified fragment length polymorphism (AFLP) analysis. J Forensic Sci 48:343–347

16. Kojoma M, Iida O, Makino Y, Sekita S, Satake M (2002) DNA fingerprinting of *Cannabis sativa* using inter-simple sequence repeat (ISSR) amplification. Planta Med 68:60–63. https://doi.org/10.1055/s-2002-19875

17. Alghanim HJ, Almirall JR (2003) Development of microsatellite markers in *Cannabis sativa* for DNA typing and genetic relatedness analyses. Anal Bioanal Chem 376:1225–1233. https://doi.org/10.1007/s00216-003-1984-0

18. Gilmore S, Peakall R (2003) Isolation of microsatellite markers in *Cannabis sativa* L. (marijuana). Mol Ecol Notes 3:105–107. https://doi.org/10.1046/j.1471-8286.2003.00367.x

19. Hsieh HM, Hou RJ, Tsai LC, Wei CS, Liu SW, Huang LH, Kuo YC, Linacre A, Lee JC (2003) A highly polymorphic STR locus in *Cannabis sativa*. Forensic Sci Int 131:53–58

20. Howard C, Gilmore S, Robertson J, Peakall R (2009) A *Cannabis sativa* STR genotype database for Australian seizures: forensic applications and limitations. J Forensic Sci 54:556–563. https://doi.org/10.1111/j.1556-4029.2009.01014.x

21. Miller Coyle H, Palmbach T, Juliano N, Ladd C, Lee HC (2003) An overview of DNA methods for the identification and individualization of marijuana. Croat Med J 44:315–321

22. Mendoza MA, Mills DK, Lata H, Chandra S, ElSohly MA, Almirall JR (2009) Genetic individualization of *Cannabis sativa* by a short tandem repeat multiplex system. Anal Bioanal Chem 393:719–726. https://doi.org/10.1007/s00216-008-2500-3

23. Shirley N, Allgeier L, LaNier T, Coyle HM (2013) Analysis of the NMI01 marker for a population database of *Cannabis* seeds. J Forensic Sci 58:S176–82. https://doi.org/10.1111/1556-4029.12005

24. Linacre A, Gusmao L, Hecht W, Hellmann AP, Mayr WR, Parson W, PrinzM, Schneider PM, Morling N (2011) ISFG: recommendations regarding the use of non-human (animal) DNA in forensic genetic investigations. Forensic Sci Int Genet 5:501–505. https://doi.org/10. 1016/j.fsigen.2010.10.017

25. Köhnemann S, Nedele J, Schwotzer D, Morzfeld J, Pfeiffer H (2012) The validation of a 15 STR multiplex PCR for *Cannabis* species. Int J Legal Med 126:601–606. https://doi.org/10.1007/s00414-012-0706-6

26. DNeasy® Plant handbook. (2013). Qiagen®, Hilden, Germany

27. Kline MC, Duewer DL, Travis JC, SmithMV, Redman JW, Vallone PM, Decker AE, Butler JM (2009) Production and certification of NIST standard reference material 2372 human DNA quantitation standard. Anal Bioanal Chem 394:1183–1192. https://doi.org/10.1007/s00216-009-2782-0

28. Baechtel FS, Smerick JB, Presley KW, Budowle B (1993) Multigenerational amplification of a reference ladder for alleles at locus D1S80. J Forensic Sci 38:1176–1182

29. Sajantila A, Puomilahti S, Johnsson V, Ehnholm C (1992) Amplification of reproducible allele markers for amplified fragment length polymorphism analysis. Biotechniques 12:16, 18, 20–22

30. MinElute® Handbook. (2008). Qiagen®, Hilden, Germany

31. BigDye® Direct Cycle Sequencing Kit protocol, version dated 02/ 2011. Life Technologies™, Carlsbad, CA

32. Valverde L, Lischka C, Scheiper S, Nedele J, Challis R, de Pancorbo MM, Pfeiffer H, Köhnemann S (2014) Characterization of 15 STR *Cannabis* loci: nomenclature proposal and SNPSTR haplotypes. Forensic Sci Int Genet 9:61–65. https://doi.org/10.1016/j.fsigen.2013.11.001

33. DNA recommendations—1994 report concerning further recommendations of the DNA Commission of the ISFH regarding PCR-based polymorphisms in STR (short tandem repeat) systems (1995) Vox Sang 69:70–71

34. Gill P, Brinkmann B, d'Aloja E, Andersen J, Bar W, Carracedo A, Dupuy B, Eriksen B, Jangblad M, Johnsson V, Kloosterman AD, Lincoln P, Morling N, Rand S, Sabatier M, Scheithauer R, Schneider P, Vide MC (1997) Considerations from the European DNA profiling group (EDNAP) concerning STR nomenclature. Forensic Sci Int 87:185–192

35. Olaisen B, Bär W, Brinkmann B, Budowle B, Carracedo A, Gill P, Lincoln P, Mayr WR, Rand S (1998) DNA recommendations 1997 of the international society for forensic genetics. Vox Sang 74:61–63

36. QIAamp® (2012) DNA investigator handbook. Qiagen®, Hilden, Germany

37. Lewis PO, Zaykin D (2001) Genetic data analysis: computer program for the analysis of allelic data. Version 1.0 (d16c). Free program distributed by the authors over the internet from http://lewis.eeb.uconn.edu/lewishome/software.html

38. Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

39. Tereba A (1999) Tools for analysis of population statistics. Profiles in DNA 3. Promega Corporation

40. Raymond M, Rousset F (1995) GENEPOP (Version 1.2): population genetics software for exact tests and ecumenicism. J Hered 86: 248–249

41. Quality assurance standards for forensic DNA testing laboratories (2009) FBI. http://www.fbi.gov/about-us/lab/biometric-analysis/codis/qas_testlabs. Accessed April 29 2015

42. Holleley CE, Geerts PG (2009) Multiplex manager 1.0: a crossplatform computer program that plans and optimizes multiplex PCR. Biotechniques 46:511–517

43. Sytsma KJ, Morawetz J, Pires JC, Nepokroeff M, Conti E, Zjhra M, Hall JC, Chase MW (2002) Urticalean rosids: circumscription, rosid ancestry, and phylogenetics based on rbcL, trnL-F, and ndhF sequences. Am J Bot 89:1531–1546. https://doi.org/10.3732/ajb.89.9.1531

CHAPTER III

Developmental validation of a novel 13 loci STR multiplex method for *Cannabis sativa* DNA profiling[1]

This dissertation follows the style and format of *International Journal of Legal Medicine*.

[1] Houston R, Birck M, Hughes-Stamm S, Gangitano D (2017) Developmental and internal validation of a novel 13 loci STR multiplex method for *Cannabis sativa* DNA profiling. Legal Med (Tokyo, Japan) 26:33-40. https://doi.org/10.1016/j.legalmed.2017.03.001

Abstract

Marijuana (*Cannabis sativa* L.) is a plant cultivated and trafficked worldwide as a source of fiber (hemp), medicine, and intoxicant. The development of a validated method using molecular techniques such as short tandem repeats (STRs) could serve as an intelligence tool to link multiple cases by means of genetic individualization or association of *Cannabis* samples. For this purpose, a 13-locus STR multiplex method was developed, optimized, and validated according to relevant ISFG and SWGDAM guidelines. The STR multiplex consists of 13 previously described *C. sativa* STR loci: ANUCS501, 9269, 4910, 5159, ANUCS305, 9043, B05, 1528, 3735, CS1, D02, C11, and H06. A sequenced allelic ladder consisting of 56 alleles was designed to accurately genotype 101 *C. sativa* samples from three seizures provided by a U.S. Customs and Border Protection crime lab. Using an optimal range of DNA (0.125 – 0.5 ng), validation studies revealed well-balanced electropherograms (inter-locus balance range: 0.500 – 1.296), relatively balanced heterozygous peaks (average peak height ratio of 0.83 across all loci) with minimal artifacts and stutter (average stutter of 0.021 across all loci). This multi-locus system is relatively sensitive (0.13 ng of template DNA) with a combined power of discrimination of 1 in 55 million. The 13 STR panel was found to be species specific for *C. sativa*; however, non-specific peaks were produced with *Humulus lupulus*. The results of this research demonstrate the robustness and applicability of this 13 loci STR system for forensic DNA profiling of marijuana samples.

Keywords: Forensic plant science, DNA typing, *Cannabis sativa,* Short tandem repeats, Reference population

Introduction

Forensic DNA typing is typically performed on human DNA samples. However, the molecular analysis of plant DNA is increasingly being studied and considered for use in criminal justice systems around the world [1-3]. In the field of forensic plant science, plant DNA can be used to link a suspect or a victim to a location (crime scene) or in the case of marijuana, can be used to aid in the investigation of drug cases. In the United States, marijuana is the most commonly used illicit controlled substance [4]. Consequently, it is a highly trafficked drug to and within the United States by organized crime syndicates. The development of a validated method using molecular techniques such as short tandem repeats (STRs) for the genetic identification of *C. sativa* may aid in the individualization and origin determination of *Cannabis* samples as well as serving as an intelligence tool to link *Cannabis* cases (e.g., illegal traffic at the US-Mexico border).

In 2003, the first polymorphic STR markers were published for *C. sativa* [5-7] and research has shown the utility of these markers in individualizing marijuana samples [8]. However, the technique has been rarely used in crime labs due to lack of standardization and validation. An analytical method should be easy to use, standardized, and validated before it can be fully utilized by a forensic laboratory.

In order to develop a reliable STR method for *Cannabis* identification, the best markers currently available were chosen as a measure of continuity within the field. In choosing markers, dinucleotide repeat markers were avoided. All markers chosen have been previously described using IUPAC nomenclature [9, 10]. Based upon our previous research [11], we improved upon a STR multiplex method by discarding STR loci that

performed poorly and incorporating six new tetranucleotide markers recently described by Valverde et al. [9].

This paper describes the development and optimization of a *C. sativa* STR multiplex in addition to a comprehensive developmental validation following guidelines established by the Scientific Working Group on DNA Analysis Methods (SWGDAM) [12]. Development validation studies included: sensitivity, species specificity, precision and accuracy, and genetic variation in a reference population. Additionally, internal validation studies were performed to provide detailed assessments of stutter ratios, peak height ratios (PHRs), and inter-locus balance of the assay.

Materials and methods

DNA samples

DNA from marijuana samples (*N*=101) was extracted from three seizures at the U.S. Customs and Border Protection LSSD Southwest Regional Science Center. A minimum of 10 specimens were randomly sampled from each case set. For collection, individual marijuana plant fragments (stem or flowers) were isolated and DNA was extracted and quantified according to Houston et al. [11]. Briefly, plant fragments (10 mg) were homogenized using liquid nitrogen and DNA was extracted using the DNeasy® Plant Mini Kit (Qiagen, Hilden, Germany) [13]. The amount of DNA was estimated via real-time PCR on the StepOne™ Real-Time PCR System (Thermo Fisher Scientific, South San Francisco, CA) using SYBR Green PCR Master Mix (Thermo Fisher Scientific) and *C. sativa* specific primers. DNA extracts were stored at -80 ºC until further analysis.

STR multiplex design and annealing temperature determination

The Multiplex Manager software v.1.2 [1] was used to evaluate any primer-primer interaction. Using a minimum distance of 20 bp between loci on the same dye channel, the 13 STR loci were configured across four dye channels. Annealing temperatures were determined for primers of each loci using an Eppendorf Master Cycler Gradient (Eppendorf, Hauppauge, NY). PCR reactions were prepared with a 12.5 µL volume using 1.0 ng of template DNA. An aliquot of DNA (2 µL) from each sample was added to 10.5 µL of PCR master mix. The PCR master mix consisted of 6.25 µL of 2x HotStarTaq® *Plus* Master Mix (Qiagen), 1.25 µL 2 µM Primer mix, 1.25 µL 5x Q-solution (Qiagen), 0.4 µL 8 mg/mL bovine serum albumin (Sigma-Aldrich, St. Louis, MO), and 1.35 µL deionized H_2O. Gradient PCR cycling were as follows: activation for 5 min at 95˚C, followed by 30 cycles of 30 s at 94˚C, 30 s at a gradient of (60±10 ˚C; 12 wells), 30 s at 72˚C, and a final extension of 30 min at 60˚C. The optimal annealing temperature for each primer set was determined via electrophoresis on a 2% agarose gel.

Loci and multiplex amplification conditions

Cannabis STR profiling was conducted in a 13 loci multiplex format modified from a previous study [2]. The multiplex consisted of previously published *Cannabis* STRs including seven markers from a previous panel (ANUCS501, ANUCS305, B05 CANN1, CS1, D02 CANN1, C11 CANN1, and H06 CANN2) [2] and six newly proposed tetranucleotide markers (9269, 4910, 5159, 9043, 1528, and 3735) [3]. PCR amplification was performed using the Type-it™ Microsatellite PCR Kit (Qiagen) on the Eppendorf Master Cycler Gradient. PCR reactions were prepared in 12.5 µL using 0.5 ng of template DNA. An aliquot of DNA (2 µL) from each sample was added to 10.5 µL of PCR master

mix. The PCR master mix consisted of 6.25 µL of 2x Type-it™ Multiplex PCR Master Mix (Qiagen), 1.25 µL 10X primer mix, 1.25 µL 5x Q-Solution, 0.4 µL 8 mg/mL bovine serum albumin, and 1.35 µL deionized H_2O. Forward primers were labeled with one of four different fluorescent dyes (6-FAM™, VIC™, NED™, or PET™, Thermo Fisher Scientific), with final optimal concentrations of forward and reverse primers displayed in Table 3.1. PCR cycling conditions were as follows: activation for 5 min at 95 ˚C, followed by 29 cycles of 30 s at 95 ˚C, 90 s at 57 ˚C, 30 s at 72 ˚C, and a final extension of 30 min at 60 ˚C. Every set of PCR reactions included one negative (deionized H_2O) and one positive control (sample #1-D1).

Table 3.1. Characteristics of 13 *Cannabis* STR markers used in this study

Marker	Dye	STR motif	Repeat type	Observed Alleles	Primer concentration (μM)	Annealing Temperature	Genbank accession no.
ANUCS501	6-FAM™	$(TTGTG)_x(CTGTG)_y$	Compound	4,5,6	0.10	58 °C	KT203577-8
9269	6-FAM™	(ATAA)	Simple	5.3,6,7	0.10	58 °C	KX668131-2
4910	6-FAM™	$(AAGA)_x(TAGA)_y(AAAA)_z$	Compound	4,10,14,15	0.20	58 °C	KX668123-4
5159	6-FAM™	$(AGAT)_x$	Simple with non-consensus allele	3,4,4.2,5.3,8,10	0.30	63 °C	KX668125-7
ANUCS305	VIC™	$(TGA)_x(TGG)_y$ (GGG)	Compound	4,6,8,11	0.10	55 °C	KT203571-3
9043	VIC™	$(TCTT)_x(CCTT)_y(TCTT)_z$	Compound	3,5,6	0.15	63 °C	KX668128-9
B05	VIC™	(TTG)	Simple	3,7,8,9,10	0.15	66 °C	KT203581-2
1528	VIC™	(ATTA)	Simple	6,7	0.30	63 °C	KX668119-20
3735	NED™	(TATG)	Simple	3,4,5,6,7	0.10	60 °C	KX668121-2
CS1	NED™	(ATCACC)*	Compound	10,12,13,16,17,23,24 25,26,27,28,29,32	0.25	58 °C	KT203586-90
D02	PET™	(GTT)	Simple	6,7,8	0.15	60 °C	KT203591-2
C11	PET™	$(TGG)_v(TTA)_w(TGG)_x$ N_{48} $(TGA)_y$ N_6 $(TGG)_z$	Compound/indel	13,14,15,21	0.15	60 °C	KT203583-5
H06	PET™	$(AAC)_v$ $(GAC)_w$ $(GAT)_x$ $(AAT)_y(GAC)_z$	Compound	7,8,9	0.15	63 °C	KT203596-7

* Most common motif observed

Capillary electrophoresis and genotyping

Separation and detection of PCR products was performed on the 3500 Genetic Analyzer (Thermo Fisher Scientific). An aliquot (1 µL) of PCR product was added to 9.5 µL Hi-Di™ Formamide and 0.5 µL GeneScan™ 600 LIZ® Size Standard v2.0 (Thermo Fisher Scientific). Samples were then denatured for 5 min and run on the 3500 Genetic Analyzer using the following conditions oven: 60°C; prerun 15 kV, 180 s; injection 1.6 kV, 8 s; run 19.5 kV, 1330 s; capillary length 50 cm; polymer: POP-7™; and dye set G5. A customized bin set was designed, and an allelic ladder (generated from sequence data for each marker) was included every third injection to ensure accurate genotyping. Genotyping was performed using a customized bin/panel on the GeneMapper v.5 software (Thermo Fisher Scientific). The analytical and stochastic thresholds were set at 100 and 700 relative fluorescence units (RFUs), respectively.

Allelic ladder design

For the six new tetranucleotide markers (9269, 4910, 5159, 9043, 1528, 3735), 40 random *Cannabis* samples were screened to determine the variability of the alleles observed in the population. Variability for the other seven markers was studied and published previously [2]. Using the most common alleles observed for all markers, an allelic ladder was generated according to previous reports [2, 4]. Briefly, these samples were amplified in single-plex PCR, then the concentration of all the amplicons was balanced, diluted approximately 1:1000, and subsequently re-amplified with 20 cycles. Each of these single STR allelic ladders were amplified to attain sufficiently high RFU values (~24,000 RFUs). These amplified allelic ladders were then diluted 1:1000 in TE buffer for future use as a second-generation ladder. All of these high RFU single STR

marker allelic ladders were then combined prior to capillary electrophoresis to attain a combined allelic ladder for all 13 loci genotyped.

Allele sequencing

For the six new tetranucleotide markers, at least two homozygous samples representing the most common alleles, were selected for sequencing. Sequence data for the remaining markers were previously reported [2]. PCR amplification and cycling sequencing were performed on the Veriti® thermal cycler (Thermo Fisher Scientific) using the BigDye® Direct Cycle Sequencing Kit (Thermo Fisher Scientific) as per the manufacturer's protocol with the exception of the annealing temperature (specific annealing temperature was used for each marker, Table 1). Samples were loaded on the 3500 Genetic Analyzer and capillary electrophoresis was performed using the following conditions: over 60 ˚C; prerun 18 kV, 60 s; injection 1.6 kV, 8 s; run 19.5 kV, 1020 s; capillary length 50 cm; polymer: POP-7™; and dye set Z. Data analysis was performed using the Sequencing Analysis software v.5.4 (Thermo Fisher Scientific). Sequences were then aligned and proofread using the Geneious Pro Software R8 (Biomatters, Auckland, New Zealand). Previous research from Valverde et al. and ISFG recommendations from human-specific STR loci were followed when determining the nomenclature of the alleles [3, 5]. Sequences were submitted to Genbank (accession numbers shown in Table 3.1).

Developmental validation

Species specificity

To assess specificity, the 13 STR markers were used to amplify non-*C. sativa* DNA. Animal samples tested included: *Ocimum basilicum* (basil), *Bos taurus* (beef), *Daucus carota* (carrot), *Felis catus* (cat), *Gallus domesticus* (chicken), *Canis lupus familiaris*

(dog), and *Homo sapiens* (human). Animal samples were extracted using the QIAamp DNA Investigator Kit (Qiagen) as per manufacturer's protocol [6]. For human DNA, TaqMan® control genomic human DNA (Thermo Fisher Scientific) was used. Plant samples tested included: *Allium sativum* (garlic), *Humulus lupulus* (Hops), *Ilex paraguariensis* (mate), *Mentha sp.* (mint), *Allium cepa* (onion), *Origanum vulgare* (oregano), *Petroselinum crispum* (parsley), *Pinus echinata* (pine), *Sus scrofa domesticus* (pork), *Rosmarinus officinalis* (rosemary), *Origanum vulgare ssp. Hirtum* (spicy oregano), *Nicotiana tabacum* (tobacco), and *Solanum lycopersicum* (tomato). Plant samples were extracted using the Qiagen DNeasy Plant Mini Kit as per the manufacturer's protocol [7]. The DNA concentration except for the human DNA was determined using a UV spectrophotometer by measuring absorbance at 260 nm, and the quality of the DNA extract was assessed via electrophoresis on a 2% agarose gel. Extracts were then amplified (2 – 10 ng) in duplicate using the developed STR multiplex to detect cross-reaction amplification across the various species.

Sensitivity and stochastic effects

To determine the sensitivity of this STR multiplex, dilutions of five different *Cannabis* DNA samples were prepared to generate template DNA amounts of 1, 0.5, 0.25, 0.13, 0.06, 0.03, and 0.016 ng for each DNA sample. The 35 dilutions were amplified in triplicate with the 13-loci STR multiplex to determine the lowest amount of template DNA that reproducibly produced a full STR profile. Data from the sensitivity study were also used to identify any stochastic effects and to establish a stochastic threshold.

Precision and Accuracy

To access precision of the assay, the fragment size of each allele in the allelic ladder was recorded across seven separate injections. The average size in base pairs and SD were calculated for each allele. Accuracy of the assay was estimated by amplifying and genotyping the positive control on five separate injections. The average size in base pairs and SD were calculated for each allele in the positive control.

Concordance Study

All samples (*N*=101) have been processed using a previous multiplex STR method [2]. The genotypes of the seven STR loci (ANUCS501, ANUCS305, B05, CS1, D02, C11, H06) that overlapped with the new 13-loci STR system were recorded and compared.

Internal validation

Stutter ratio determination

Stutter ratios were determined for each of the 13 STR loci using 25 samples (~0.5 ng of template DNA). DNA samples were amplified using the developed 13-loci STR multiplex. The stutter ratio was calculated by dividing height of the stutter peak by height of the associated allele peak. The mean, standard deviation (SD), range, and mean plus 3 SD values were determined.

Heterozygous peak height ratio and inter-loci balance

Heterozygous peak height ratios (PHR) were determined using 25 samples (~0.5 ng of template DNA). The samples were amplified using the newly developed STR multiplex method. PHR was determined by dividing the height of the smaller peak by the height of the larger peak in a heterozygous pair. Mean, SD, and minimum PHR (mean

minus 3 SD) were calculated. The inter-loci balance was also assessed by dividing the average peak height at one locus by the average peak height across all loci.

Statistical analysis

For all STR markers, the number of multi-locus genotypes and the genotype sharing among samples were determined. For the reference population database (N=95), allele frequencies and parameters of forensic interest were estimated using the PowerStats v.12 software [8]. In addition, exact tests for Hardy-Weinberg equilibrium and linkage disequilibrium were performed on this reference population with the Genetic Data Analysis v.1.0 (GDA) software [18]. The p value for statistically significant differences was set at 0.05 levels. Bonferroni correction for multiple comparisons was applied when applicable.

Results and discussion

Optimization of PCR reaction and cycling conditions

The *Cannabis* multiplex STR system was optimized using the Type-it™ Microsatellite PCR Kit (Qiagen). Primer concentrations were titrated to ensure inner-locus balance across the 13 STR markers (Table 3.1.). An example of an electropherogram of the novel 13 loci STR multiplex system is displayed in Fig. 3.1. Annealing temperatures were determined for each marker (primer set) to avoid the occurrence of null alleles. Annealing temperatures ranged from 63 ˚C to 55 ˚C. The optimal annealing temperature was determined to be 57 ˚C. Cycle number experiments were performed to determine the cycle number that yielded the most consistent STR profiles.

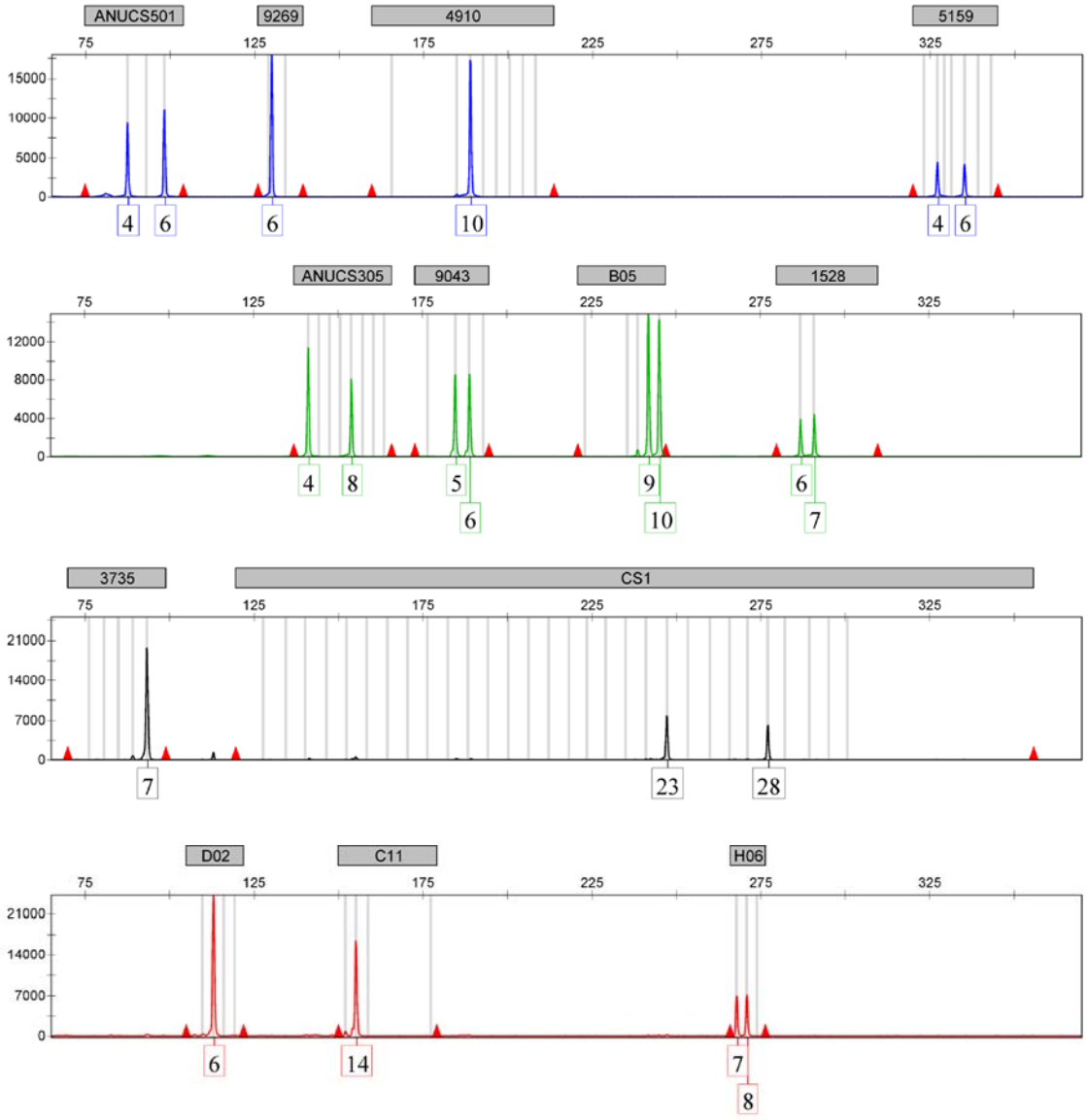


Fig. 3.1. Multiplex profile of 13 *Cannabis* STR loci using 0.5 ng of control template DNA (sample #1-D1)

Allelic ladder design

For all 13 loci, an allelic ladder was developed with the most common alleles observed in the sample population (Fig. 3.2.). The allelic ladder contained 56 alleles across the 13 STR loci. Nomenclature following international guidelines (ISFG) was used to designate the allele calls [9]. In addition, the number of repeats for two to four alleles per

tetranucleotide marker was determined via sequencing to ensure accurate nomenclature of the allelic ladder and confirmation of published sequencing results from Valverde et al. [3, 5]. The use of an allelic ladder is necessary for STR data sharing and is recommended in the ISFG guidelines for the application of non-human DNA testing for forensic applications [9].

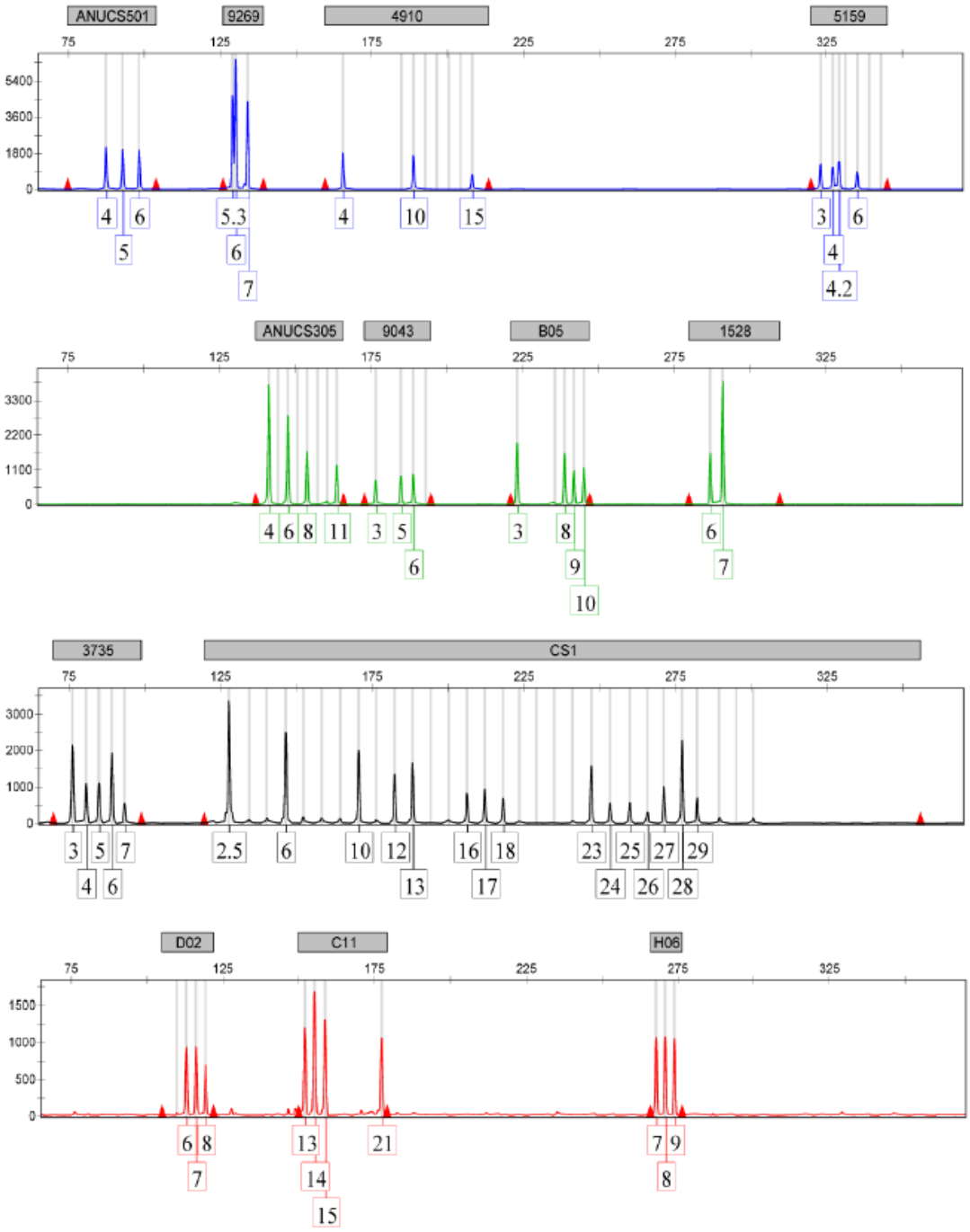


Fig. 3.2. Allelic ladder for 13 *Cannabis* STR loci which design was based on sequence data obtained from most common observed alleles

Validation experiments

When testing species specificity, results showed that none of the 13 STR markers displayed cross-reactivity with any of the species tested except for *H. lupulus,* which generated non-specific peaks (209 bp, 215 bp, and 255 bp in the green dye channel). This non-specific cross-reactivity of *H. lupulus* has been previously reported [2]] and was not unexpected as *H. lupulus* belongs to the same family, Cannabaceae, as *C. sativa* [10]. Nevertheless, these non-specific peaks cannot create any problems for data interpretation because their respective locations are off the ladder bins.

The sensitivity of the 13 loci multiplex was estimated to be 0.13 ng of DNA by amplifying amounts of template ranging from 1 ng to 0.016 ng. It was determined that allele drop-out and severe peak imbalance occurred when the template DNA was at or below 0.06 ng (Fig. 3.3). All alleles were correctly identified when amplifying 0.13 to 1 ng of template DNA (Fig. 3.3). The sensitivity study revealed the optimal input of DNA to be 0.5 ng. The stochastic threshold was determined to be 700 RFUs by examining the heterozygous loci where one of the sister alleles fell below the established analytical threshold. The stochastic threshold established represents the average peak heights of the false homozygotes plus 3 SD.

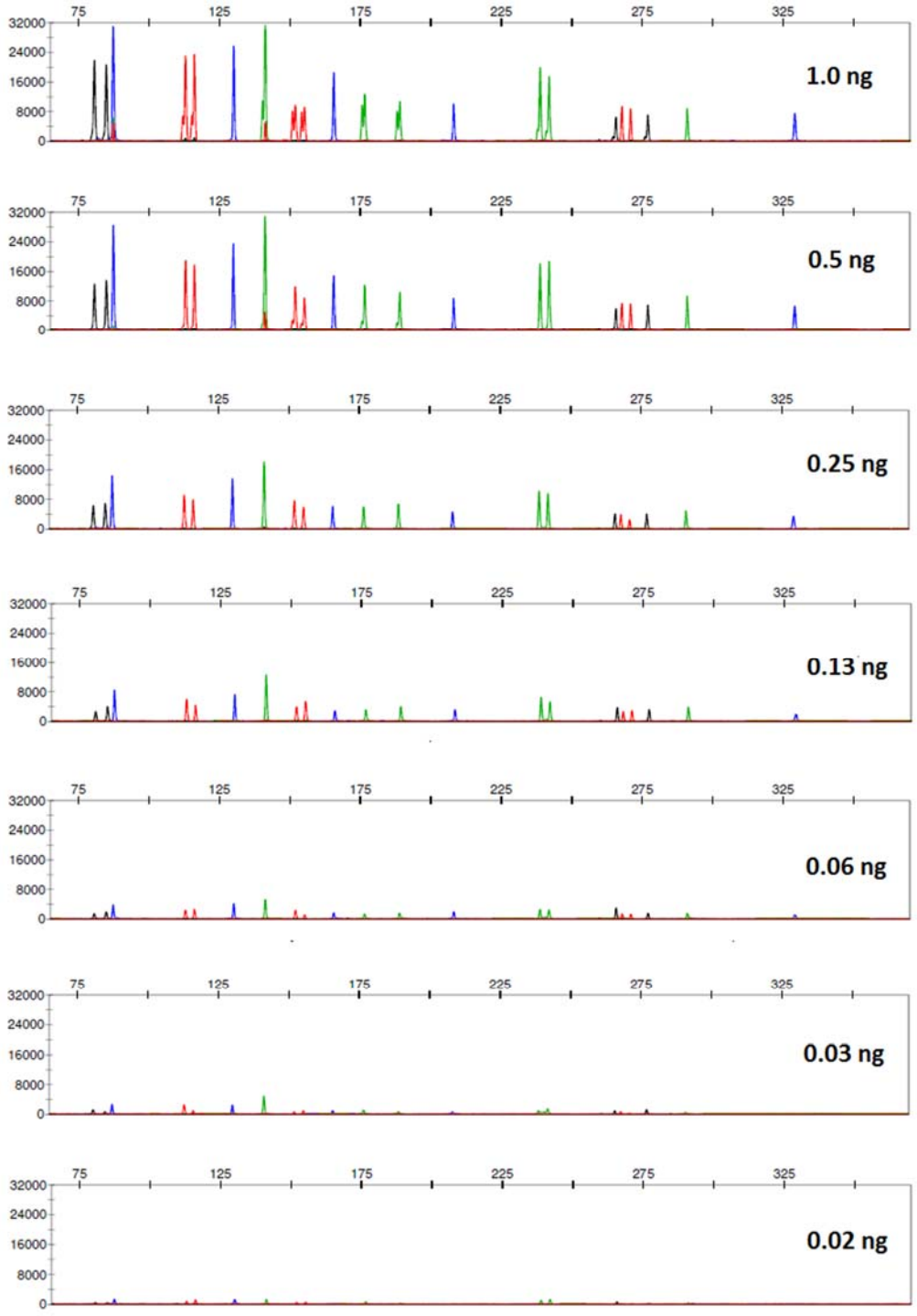


Fig. 3.3. *Cannabis* 13-loci multiplex DNA profiles obtained from serially diluted single-source template DNA ranging from 1 ng to 20 pg. One microliter of the amplification was analyzed on a 3500 Genetic Analyzer with a 1.6 kV, 8-s injection

Stutter ratios and the mean and standard deviation for the heterozygous PHRs were calculated for each of the 13 loci using 25 samples (~0.5 ng of template DNA) with the optimized protocol. The average stutter ratio across all loci was between 0.006 and 0.052 for the 13 loci multiplex protocol (Table 3.2.). The average PHR ranged from 0.689 to 0.895 with the median PHR above 0.70 for all loci except for CS1 (0.660) (Table 3.3.), which is a widely accepted PHR threshold [21]. Additionally, inter-locus balance was observed with a range from 0.500 (5159) to 1.671 (B05) (Table 3.3.).

Table 3.2. Observed stutter ratios, range, mean, standard deviation and upper range at each locus included in the 13 loci *Cannabis* STR multiplex system for samples (*N*=25) amplified using 0.5 ng of template DNA

Marker	Range	Mean	Standard deviation	Mean + 3SD
ANUCS501	0.000 – 0.105	0.035	0.036	0.141
9269	0.001 – 0.018	0.009	0.003	0.018
4910	0.001 – 0.062	0.016	0.016	0.064
5159	0.002 – 0.034	0.011	0.009	0.038
ANUCS305	0.003 – 0.035	0.012	0.007	0.033
9043	0.000 – 0.031	0.006	0.007	0.027
B05	0.002 – 0.078	0.035	0.014	0.077
1528	0.000 – 0.021	0.007	0.005	0.021
3735	0.001 – 0.054	0.024	0.014	0.067
CS1	0.012 – 0.062	0.030	0.015	0.074
D02	0.000 – 0.057	0.015	0.011	0.048
C11	0.029 – 0.233	0.052	0.038	0.166
H06	0.000 – 0.201	0.026	0.041	0.149

Table 3.3. Observed peak height ratios (PHR) mean, median, minimum, and maximum at each locus included in the 13 loci *Cannabis* STR multiplex system for samples (*N*=25) amplified using 0.5 ng of template DNA

Marker	Observations	Mean PHR	Median PHR	Minimum PHR	Maximum PHR	Inter-loci balance
ANUCS501	16	0.884	0.893	0.581	0.998	1.076
9269	10	0.694	0.720	0.328	0.872	1.019
4910	19	0.853	0.921	0.402	0.974	0.886
5159	19	0.823	0.879	0.389	0.978	0.500
ANUCS305	13	0.845	0.862	0.563	0.988	1.296
9043	14	0.889	0.928	0.620	0.994	1.063
B05	16	0.861	0.875	0.548	0.999	1.671
1528	10	0.838	0.871	0.537	0.941	0.617
3735	22	0.895	0.906	0.638	0.996	0.759
CS1	23	0.689	0.660	0.117	0.993	0.812
D02	14	0.861	0.889	0.472	0.986	1.561
C11	16	0.831	0.859	0.402	0.980	1.011
H06	15	0.820	0.813	0.472	0.967	0.728

Results of the accuracy and precision studies indicated that the value of three standard deviations was less than 0.5 bp for every allele in the allelic ladder as well as the positive control. STR profiles of 95 samples, previously amplified using reported multiplex conditions for markers 305, 501, B05, D02, H06, C11 and CS1 [11], were compared to the genotyping results using the new multiplex protocol to assess the concordance between the two STR systems. Full profile concordance was observed in all samples for the markers studied.

Multiplex PCR method performance and population studies

All samples (*N*=101) were successfully amplified under the optimized conditions. Samples that were deemed to be mixtures (*N*=5) were discarded from further analysis. Two duplicate samples from the same seizure were detected. This was not unexpected due to the sampling method used in this study. Distinguishable DNA profiles were generated from the 95 samples that generated full STR profiles.

No departures from Hardy-Weinberg equilibrium were detected for any of the STR markers studied in the reference population. A linkage disequilibrium test was performed to detect any correlations between alleles at any of the pair-wise comparisons of the 13 loci. For this database, there was a total of 78 pair-wise comparisons performed. Nine significant departures were observed (11.5% of the pair-wise tests) at a p-value of 0.05. However, after Bonferroni correction one departure survived between STR loci 9269 and H06. This might be attributed to the effects of population substructure [22]. Based on these observations, with little evidence of association between loci, the assumption of independence is valid, and a multiple-locus profile frequency can be estimated using the product rule.

Allele frequencies were determined and used to calculate parameters of forensic interest (Table 3.4.) as well as random match probability estimations. The combined power of discrimination for the 13 loci multiplex is 1 in 55 million.

Table 3.4. Allele frequencies and Hardy-Weinberg equilibrium evaluation of six new *Cannabis* STR markers in a reference population of cases seized at the Mexico-US border (95 individuals, n=190 chromosomes)

Allele	9269	4910	5159	9043	1528	3735
3			0.1540	0.2140		0.1280
4		0.5050	0.1220			0.1540
4.2			0.3940			
5				0.3330		0.1060
5.3	0.0570					
6	0.8960		0.2980	0.4530	0.1510	0.1810
7	0.0470				0.8490	0.4310
8			0.0110			
9						
10		0.4320	0.0210			
11						
12						
13						
14		0.0050				
15		0.0570				
HWE	0.1394	0.9053	0.8125	0.1813	0.4234	0.1434

HWE: Hardy-Weinberg equilibrium probability values of exact test (3200 shufflings)

Conclusions

The goal of this study was to develop a 13 loci *Cannabis* STR multiplex system for forensic DNA profiling that could approach the robustness of standard STR systems used for human identification (HID). This study was able to demonstrate that this new multiplex can produce high-quality STR profiles that are comparable with standard STR HID

systems. These technologies may assist the forensic community as the demand for *Cannabis* studies either for genetic identification or intelligence purposes increases.

By incorporating more recently discovered STR tetranucleotides and using a comprehensive approach to multiplex design, this 13-loci STR method was able to generate high quality DNA profiles with template input as low as 0.13 ng. STR success rates were improved when compared to a previous version of this method (100% vs 64%) [11]. This improvement is due to the implementation of a comprehensive strategy for multiplex STR design and optimization. The average stutter ratio across all loci ranged from 0.006 – 0.052; the maximum stutter upper range was estimated to be 0.166 for STR marker C11. Additionally, the mean PHR ranged from 0.689 – 0.895 across all loci.

In summary, this study demonstrates a robust and reliable 13 loci *Cannabis* STR multiplex can be used for forensic DNA profiling of marijuana samples. However, suitable data interpretation guidelines should be developed through internal validation studies prior to implementation.

Funding information

This study was partially funded by a Graduate Research Fellowship Award #2015-R2-CX-0030 (National Institute of Justice, Office of Justice Programs, U.S. Department of Justice). The opinions, findings, conclusions, or recommendations expressed in this presentation are those of the authors and do not necessarily reflect those of the National Institute of Justice.

References

1. Schield C, Campelli C, Sycalik J, Randle C, Hughes-Stamm S, Gangitano D (2016) Identification and persistence of *Pinus* pollen DNA on cotton fabrics: a forensic application. Sci Justice 56:29-34. https://doi.org/10.1016/j.scijus.2015.11.005
2. Bock JH, Norris DO, Forensic Plant Science, first ed. Elsevier Academic Press, London
3. Zaya DN, Ashley MV (2012) Plant genetics for forensic applications. Methods Mol Biol 862:35-52. https://doi.org/10.1007/978-1-61779-609-8_4
4. Center for Behavioral Health Statistics and Quality (2014) Results from the 2013 National Survey on Drug Use and Health: summary of national findings. U.S. Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. http://www.samhsa.gov/data/sites/default/files/NSDUHresultsPDFWHTML2013/Web/NSDUHresults2013.pdf. Accessed September 2016
5. Alghanim HJ, Almirall JR (2003) Development of microsatellite markers in *Cannabis sativa* for DNA typing and genetic relatedness analyses. Anal Bioanal Chem 376:1225-1233. https://doi.org/10.1007/s00216-003-1984-0
6. Gilmore S, Peakall R (2003) Isolation of microsatellite markers in *Cannabis sativa* L. (marijuana). Mol Ecol Notes 3:105-107. https://doi.org/10.1046/j.1471-8286.2003.00367.x
7. Hsieh HM, Hou RJ, L.C. Tsai, LC, Wei CS, Liu SW, Huang LH, Kuo YC, Linacre A, Lee JC (2003) A highly polymorphic STR locus in *Cannabis sativa*. Forensic Sci Int 131:53-58

8. Howard C, Gilmore S, Robertson J, Peakall R (2009) A *Cannabis sativa* STR genotype database for Australian seizures: forensic applications and limitations. J Forensic Sci 54:556-563. https://doi.org/10.1111/j.1556-4029.2009.01014.x

9. Valverde L, Lischka C, Erlemann S, de Meijer E, de Pancorbo MM, Pfeiffer H, Köhnemann S (2014) Nomenclature proposal and SNPSTR haplotypes for 7 new *Cannabis sativa* L. STR loci. Forensic Sci Int Genet 13:185-186. https://doi.org/10.1016/j.fsigen.2014.08.002

10. Valverde L, Lischka C, Scheiper S, Nedele J, Challis R, de Pancorbo MM, Pfeiffer H, Kohnemann S (2014) Characterization of 15 STR *Cannabis* loci: nomenclature proposal and SNPSTR haplotypes. Forensic Sci Int Genet 9:61-65. https:/doi.org/10.1016/j.fsigen.2013.11.001

11. Houston R, Birck M, Hughes-Stamm S, Gangitano D (2016) Evaluation of a 13-loci STR multiplex system for *Cannabis sativa* genetic identification. Int J Legal Med 130:635-647. https://doi.org/10.1007/s00414-015-1296-x

12. Scientific Working Group on DNA Analysis Methods. Validation Guidelines for DNA Analysis Methods (2012). http://media.wix.com/ugd/4344b0_cbc27d16dcb64fd88cb36ab2a2a25e4c.pdf Accessed September 2016

13. DNeasy® Plant Handbook (2012) Qiagen, Hilden, Germany

14. Holleley CE, Geerts PG (2009) Multiplex Manager 1.0: a crossplatform computer program that plans and optimizes multiplex PCR. BioTechniques 46:511-517. https://doi.org/10.2144/000113156

15. Griffiths RAL, Barber MD, Johnson PE, Gillbard SM, Haywood MD, Smith CD, Arnold J, Burke T, Urquhart AJ, Gill P (1998) New reference allelic ladders to improve allelic designation in a multiplex STR system. Int J Legal Med 111:267-272

16. QIAamp DNA Investigator® Handbook (2012) Qiagen, Hilden, Germany

17. Tereba A (1999) Tools for analysis of population statistics. Profiles in DNA 3. Promega Corporation

18. Lewis PO, Zaykin D (2001) Genetic data analysis: computer program for the analysis of allelic data, v. 1.0 (d16c). http://hydrodictyon.eeb.uconn.edu/people/plewis/downloads/gda-1.1.win32.zip Accessed July 2016

19. Linacre A, Gusmao L, Hecht W, Hellmann AP, Mayr WR, Parson W, Prinz M, Schneider PM, Morling N (2011) ISFG: recommendations regarding the use of non-human (animal) DNA in forensic genetic investigations. Forensic Sci Int Genet 5:501-505. https://doi.org/10.1016/j.fsigen.2010.10.017

20. Sytsma KJ, Morawetz J, Pires JC, Nepokroeff M, Conti E, Zjhra M, Hall JC, Chase MW (2002) Urticalean rosids: circumscription, rosid ancestry, and phylogenetics based on rbcL, trnL-F, and ndhF sequences. Am J Bot 89:1531-1546. https://doi:org/10.3732/ajb.89.9.1531

21. Butler JM (2005) Forensic DNA typing: Biology, technology, and genetics of STR markers. Elseiver Academic Press, New York

22. Budowle B, Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM (1999) Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. J Forensic Sci 44:1277-1286

CHAPTER IV

Nuclear, chloroplast, and mitochondrial data of a US *Cannabis* DNA database[1]

This dissertation follows the style and format of *International Journal of Legal Medicine*.

[1] Houston R, Birck M, LaRue B, Hughes-Stamm S, Gangitano D (2018) Nuclear, chloroplast, and mitochondrial data of a US *Cannabis* DNA database. Int J Legal Med https://doi.org/10.1007/s00414-018-1798-4

Reprinted with permission from publisher.

Abstract

As *Cannabis sativa* (marijuana) is a controlled substance in many parts of the world, the ability to track biogeographical origin of *Cannabis* could provide law enforcement with investigative leads regarding its trade and distribution. Population substructure and inbreeding may cause *Cannabis* plants to become more genetically related. This genetic relatedness can be helpful for intelligence purposes. Analysis of autosomal, chloroplast, and mitochondrial DNA allows not only for prediction of biogeographical origin of a plant, but also discrimination between individual plants.

A previously validated 13-autosomal STR multiplex was used to genotype 510 samples. Samples were analyzed from four different sites: 21 seizures at the US-Mexico border, Northeastern Brazil, hemp seeds purchased in the US, and the Araucania area of Chile. In addition, a previously reported multi-loci system was modified and optimized to genotype five chloroplast and two mitochondrial markers. For this purpose, two methods were designed: a homopolymeric STR pentaplex and a SNP triplex with one chloroplast (Cscp001) marker shared by both methods for quality control. For successful mitochondrial and chloroplast typing, a novel real-time PCR quantitation method was developed and validated to accurately estimate the quantity of the chloroplast DNA (cpDNA) using a synthetic DNA standard. Moreover, a sequenced allelic ladder was also designed for accurate genotyping of the homopolymeric STR pentaplex.

For autosomal typing, 356 unique profiles were generated from the 425 samples that yielded full STR profiles and 25 identical genotypes within seizures were observed. Phylogenetic analysis and case-to-case pairwise comparisons of 21 seizures at the US-

Mexico border, using Fixation Index (F_{ST}) as genetic distance, revealed the genetic association of nine seizures that formed a reference population.

For mitochondrial and chloroplast typing, subsampling was performed, and 134 samples were genotyped. Complete haplotypes (STRs and SNPs) were observed for 127 samples. As expected, extensive haplotype sharing was observed; five distinguishable haplotypes were detected. In the reference population, the same haplotype was observed 39 times and two unique haplotypes were also detected. Haplotype sharing was observed between the US border seizures, Brazil, and Chile, while the hemp samples generated a distinct haplotype.

Phylogenetic analysis of the four populations was performed and results revealed that both autosomal and lineage markers could discern population sub-structure.

Keywords: Forensic plant science, *Cannabis sativa*, DNA database, Chloroplast DNA, Mitochondrial DNA, Short tandem repeats

Introduction

Cannabis sativa (marijuana) is a plant used for various purposes, namely as an intoxicant, fiber, or medicine [1, 2]. As a result, *Cannabis* is a commodity highly trafficked around the world. The intoxicant properties of *C. sativa*, specifically the presence of the psychoactive cannabinoid Δ^9 – tetrahydrocannabinol (Δ^9-THC), make it a plant of interest to law enforcement. Several genotyping methods have been suggested as a means of tracking and individualizing marijuana plants [3-5]. As with human identification, autosomal STR typing can be used as a means of individualizing *Cannabis* samples. In the case of clonal propagation, these samples will have identical DNA profiles, allowing for direct associations. However, with sexually propagated plants, population substructure and inbreeding can occur within a growing field or an isolated geographical area. In this instance, the sub-structure and subsequent genetic relatedness can be helpful for intelligence purposes.

In addition, biogeographical tracking could provide law enforcement insight on its trade and distribution patterns. To predict the biogeographical origin of plants such as *Cannabis*, organelle markers are targeted due to their non-recombining inheritance and inherently low mutation rate. These organelle markers may become fixed in certain biogeographic populations but will remain discriminatory for populations from different regions. Analysis of organelle DNA, including both mitochondrial DNA (mtDNA) and chloroplast DNA (cpDNA) has been shown to be a valuable tool in analyzing evolutionary and population diversity in plant species as it is inherited uniparentally [6-8]. In *C. sativa*, cpDNA and mtDNA are both inherited maternally [9]. Like human mtDNA, this inheritance pattern reveals a genetic snapshot of the evolutionary and biogeographic

information of a single *Cannabis* plant. Both the chloroplast [10] and mitochondrial [11] genomes have been mapped for *C. sativa* and are freely accessible. Several studies have evaluated phylogenetic relationships in angiosperms, such as *Cannabis*, using regions of the chloroplast and mitochondrial genomes [7, 12-14]. Universal primer sets have been used to isolate polymorphic regions in the chloroplast and mitochondrial genomes [6, 15]. Chloroplast regions targeting *Cannabis* population structure include *rbcL* [16], *trnL – trnF* [7, 17], *trn H – trnK* [8, 15], ccmp2 (5' to *trn*S) [12] and ccmp6 (*orf*77 – *orf*82*)* [12] region of the chloroplast. In addition, *nad4 and nad5* regions of the mitochondria have been identified as polymorphic regions for *Cannabis* [8].

These regions have been evaluated previously by Gilmore et al. and results have shown that these organelle loci can to some extent discriminate *Cannabis* samples based on geographic origin [8]. Although a method was proposed to determine the haplotypes based on the organelle genetic data [8], two important International Society of Forensic Genetics (ISFG) recommendations for the use of non-human DNA [18] were not followed: (a) the use of a sequenced allelic ladder for accurate allele designation and inter-laboratory profile sharing, and (b) the use of an analytical method to accurately quantify cpDNA prior to downstream DNA analysis. The reported assay genotyped seven *Cannabis* organelle markers: five chloroplast markers (Cscp001, Cscp002, Cscp003, Cscp004, and Cscp005) and two mitochondrial markers (Csmt001 and Csmt002). Gilmore et al. genotyped five markers in single-plex (Cscp001, Cscp002, Cscp003, Cscp004, and Csmt001) by size using capillary electrophoresis without an allelic ladder. All loci are homopolymeric repeats except for Cscp001, which is an insertion-deletion polymorphism (INDEL). Inter-run variation of one base pair can affect genotyping of homopolymeric repeats if an allelic

ladder is not used. In addition, Gilmore et al. genotyped three markers in single-plex (Cscp001, Cscp005, and Csmt002) using an amplification refractory mutation system (ARMS) based assay [19]. However, more standardized technologies such as mini-sequencing (SNaPshot®, Thermo Fisher Scientific, South San Francisco, CA, USA) are more reliable and robust for SNP genotyping.

In this work, a DNA database consisting of 510 samples was used to genotype both autosomal and organelle DNA. A previously validated 13-autosomal STR multiplex [5] was used to genotype 510 samples from four different sites: 21 seizures at the US-Mexico border, Northeastern Brazil, hemp seeds purchased in the US, and the Araucania region of southern Chile. For organelle typing, the previously reported multi-loci system from Gilmore et al. was modified and optimized to genotype five chloroplast and two mitochondrial markers from a subsampling of the 510 samples [8]. For this purpose, two methods were designed: a homopolymeric STR pentaplex and a SNP triplex with one chloroplast (Cscp001) marker shared by both methods for quality control. For successful downstream organelle typing, a novel assay for the real-time PCR quantification of *Cannabis* cpDNA using synthetic DNA standards was developed, optimized and validated according to the Scientific Working Group on DNA Analysis (SWGDAM) guidelines [20]. In addition, a sequenced allelic ladder was designed for accurate genotyping of the homopolymeric STR pentaplex.

Materials and methods

DNA extraction

THC-containing (or THC-positive) *Cannabis* samples were obtained from three sources: U.S. Customs and Border Protection (CBP) (N=422), Northeast Brazil (N=8), and

the Araucania region of southern Chile (*N*=50). Additionally, three brands of hemp seeds were purchased: Navitas™ Organics (Novato, CA, USA) (*N*=10), Badia Spices Inc. (Doral, FL, USA) (*N*=10), and Manitoba Harvest (Winnipeg, MB, CA) (*N*=10).

For CBP and hemp samples, plant fragments/seeds were homogenized with liquid nitrogen followed by DNA extraction using the DNeasy® Plant Mini Kit (Qiagen, Hilden, Germany) as per manufacturer's protocol [21]. DNA was extracted on-site at U.S. Customs and Border Protection LSSD Southwest Regional Science Center from 21 separate seizures. From each seizure, at least ten individual specimens were randomly sampled. Individual *Cannabis* plant fragments consisting of stem, flowers, seeds, or leaves (10 mg) were isolated during collection. Additionally, all four tissue types (stem, flower, seed, and leaf) were specifically targeted in four individual *Cannabis* plants to compare the relative abundance of cpDNA. DNA was extracted from the hemp seeds at Sam Houston State University. Individual seeds (*N*=10) were randomly chosen from each of the three brands of hemp seed.

For the Brazilian and Chilean samples, DNA extracts were provided by the Federal University of Rio Grande do Sul in Brazil and by the Policia de Investigaciones in Chile, respectively. DNA extracts from Brazil consisted of eight unrelated samples while DNA extracts from Brazil consisted of eight unrelated samples while DNA extracts from Chile consisted of ten separate seizures with five DNA extracts from each seizure.

Autosomal DNA typing

The amount of nuclear DNA was previously estimated according to Houston et al. [22] via real-time PCR on the StepOne™ Real-Time PCR System (Thermo Fisher Scientific, South San Francisco, CA, USA) with SYBR™ Green PCR Master Mix (Thermo

Fisher Scientific) and *Cannabis*-specific primers (ANUCS304). DNA extracts were stored at -80 ºC until further analysis. *Cannabis* STR profiling was performed via a 13-loci multiplex using a previously validated method according to Houston et al. [5].

Chloroplast DNA quantitation

DNA synthetic standards

The DNA standards were comprised of two complementary, PAGE–purified synthetic oligonucleotides (Ultramers®, Integrated DNA Technologies, Coralville, IA, USA) (Table 4.1.). The oligonucleotides correspond to Cscp001 region of *C. sativa* cpDNA (GenBank accession AY958392.1). The forward and reverse oligonucleotides were reconstituted in TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH: 8.0) to generate a 10 µM stock for both solutions. Then, a diluted stock (2 µM) was generated for both forward and reverse oligonucleotides, which were then mixed in equal parts to create a 1 µM double-stranded, primary standard stock. Using Avogadro's constant (6.02×10^{23} copies per mol) to determine copies per µL (6.02×10^{11}copies per µL) and the molecular weight of the entire *Cannabis* cpDNA genome (1.67×10^{-4} pg/copy), the primary stock was diluted to generate the following standards: 1000, 200, 100, 10, 2, 1, 0.1, and 0.02 pg/µL.

Table 4.1. Sequences of cpDNA synthetic standard and primers

cpDNA standard (forward strand):
5’ – ATT TAT CCT CTC ATT CCG TTA GTG GTT TCT AAT TTG TTA TGT TTC TCG TTC ATT CTA ACT TTA CAA CCG GAC CTG AAT GAC CCT TTT TTT TAT TAT CAC AAG CCT TGT GAT ATA TAT GAA AGA CCT ACA AAT GAA CAT AAG GAA TCC CAA TGT GCA ATT GGA AT – 3’
cpDNA standard (reverse strand):
5’ – ATT CCA ATT GCA CAT TGG GAT TCC TTA TGT TCA TTT GTA GGT CTT TCA TAT ATA TCA CAA GGC TTG TGA TAA TAA AAA AAA GGG TCA TTC AGG TCC GGT TGT AAA GTT AGA ATG AAC GAG AAA CAT AAC AAA TTA GAA ACC ACT AAC GGA ATG AGA GGA TAA AT – 3’
Forward primer:
5’ – TCCTCTCATTCCGTTAGTGGT – 3’
Reverse primer:
5’ – AATTGCACATTGGGATTCC – 3’

Real-time PCR parameters for cpDNA quantitation

Quantification of chloroplast DNA was performed via real-time PCR on a StepOne™ Real-Time PCR System (Thermo Fisher Scientific) using SYBR™ Green PCR Master Mix (Thermo Fisher Scientific) and *Cannabis*-specific chloroplast primers (Cscp001) (Integrated DNA Technologies) (Table 1). An aliquot of DNA extract (2 µL) was added to a master mix (23 µL) consisting of 12.5 µL of 2X SYBR Green Master Mix, 0.5 µL Cscp001 primers (20 µM), 0.8 µL bovine serum albumin (8 mg/mL, Sigma-Aldrich, St. Louis, MO, USA), and 9.2 µL deionized H_2O. Standard real-time PCR cycling conditions were used with an initial denaturation (10 min, 95 °C) and cycling (40 cycles;

15 s at 95 °C followed by 1 min at 60 °C). Serial dilutions (1000 to 0.02 pg/µL) of the reconstituted synthetic DNA standard were used to generate a calibration curve. Linearity was evaluated using an R^2 estimation with a minimum correlation of 0.99 for acceptance.

Chloroplast DNA quantitation validation studies

Sensitivity studies (SWGDAM 3.3)

Ten standards were examined in a 1:10 dilution series (1000 pg/µL, 100 pg/µL, 10 pg/µL, 1 pg/µL, 0.1 pg/µL, 0.01 pg/µL, 0.001 pg/µL, 0.0001 pg/µL, 0.00001 pg/µL, and 0.000001 pg/µL) in triplicate. The standard curve was assessed to examine the limit of linearity and limit of detection. The limit of detection was determined to be the smallest DNA standard where the calibration curve still generated an R^2 value above 0.99.

Specificity study (SWGDAM 3.2)

To evaluate species specificity, the real-time PCR assay was used to amplify non-*Cannabis* species. The following plant species were evaluated: *Ocimum basilicum* (basil), *Allium sativum* (garlic), *Humulus lupulus* (Hops), *Origanum vulgare hirtum* (Italian oregano), *Ilex paraguariensis* (mate), *Mentha* (mint), *Origanum vulgare* (oregano), *Petroselinum crispum* (parsley), *Pinus echinata* (pine), *Rosmarinus officinalis* (rosemary), *Nicotiana tabacum* (tobacco), and *Solanum lycopersicum* (tomato). Animal species consisting of *Felis catus* (cat) and *Homo sapiens* (human) were also evaluated for species specificity. Plant samples were extracted with the DNeasy Plant Mini Kit (Qiagen) as per manufacturer's protocol [21]. The cat sample was extracted with the QIAamp DNA Investigator Kit (Qiagen) as per manufacturer's protocol [4]. TaqMan™ Control Genomic DNA (Thermo Fisher Scientific) was used for human DNA. For all extracts, DNA concentration was assessed at 260 nm with UV spectrophotometry. Extract quality was

evaluated via electrophoresis on a 2% agarose gel. Extracts were then assayed (~1 – 5 ng) in duplicate with the real-time PCR method to detect any cross-reactivity between the various species. In addition, melt curve analysis was performed to ensure the specificity of the amplification signal as non-specific PCR products and primer-dimers can also generate a fluorescent signal.

Precision and accuracy (SWGDAM 3.5)

Eight *Cannabis* DNA standards (1000, 200, 100, 10, 2, 1, 0.1, and 0.02 pg/µL) along with three control *Cannabis* extracts and a no template control were run in duplicate across 18 separate real-time PCR runs. Amplification efficiencies were estimated using the slope of the standard plot regression line: efficiency = $[10^{(-1/\text{slope})}]-1$. In addition, the coefficient of variation (%CV) was accessed for linearity, slope, y-intercept, and amplification efficiency across the 18 runs.

Chloroplast and mitochondrial STR typing

STR multiplex design and annealing temperature determination

The Multiplex Manager software v.1.2 [11] was used to assess any primer – primer interactions. The five STR loci were configured across two dye channels (blue and green) with a minimum distance of 20 bp between loci on the same dye channel. Forward and reverse PCR primer sequences can be found in Table 4.2. Annealing temperatures were experimentally determined for each primer set using the HotStar Taq *Plus* Master Mix (Qiagen) on an Eppendorf Master Cycler Gradient (ramp rate: 3 ˚C/s) (Eppendorf, Hauppauge, NY, USA) as per Houston et al. [5].

Table 4.2. Chloroplast and mitochondrial primers and regions targeted in this study

Locus	Primers	Primer Reference	Region of DNA
STR based			
Cscp001	F: 5' – tcctctcattccgttagtggt – 3' R: 5' – aattgcacattgggattcct – 3'	[8]	*trnL – trnF*
Cscp002	F: 5' – tcatttgatgaagtggggta – 3' R: 5' – gcatggggaacctactattt – 3'	[8]	*rbcL – orf*106
Cscp003	F: 5' – gatcccggacgtaatcctg – 3' R: 5' – atcgtaccgagggttcgaat – 3'	[12]	ccmp2 (5' to *trn*S)
Cscp004	F: 5' – cgatgcatatgtagaaagcc – 3' R: 5' – cattacgtgcgactatctcc – 3'	[12]	ccmp6 (*orf77 – orf82*)
Csmt001	F: 5' – atggcagagaagtttccata – 3' R: 5' – ttggctccctaaagactaaa – 3'	[8]	*nad* 4 exon 3 to exon 4
SNP based			
Cscp001	F: 5' – tccctctatccccaaaaagg – 3' R: 5' – attgcacattgggattcctt – 3' SBE F: 5' – ttttttttttttacaaccggacctgaatgacc – 3'	This study	*trnL – trnF*
Cscp005	F: 5' – tccactgccttgatccactt – 3' R: 5' – ccctctagacttagctgctct – 3' SBE R: 5' – cttttatcttgtctaaaattgaaat – 3'	This study	*trnH – trnK*
Csmt002	F: 5' – tgtgcgaagagtgcgttatg – 3' R: 5' – acttcactcgctaggggatg – 3' SBE F: 5' – ttttttttttttttttttttatgacctgtggccgcctg – 3'	This study	*nad* 5 exon 4 to exon 5

Homopolymeric STRs and multiplex amplification conditions

Cannabis STR profiling was conducted in a five–loci multiplex format modified from a previous study [8]. The multiplex included previously published *Cannabis* chloroplast (Cscp001, Cscp002, Cscp003, Cscp004) and mitochondrial (Csmt001) STR markers [8]. PCR amplification was performed with the Type-it™ Microsatellite PCR Kit (Qiagen) on a T100™ Thermal Cycler (ramp rate: 3 ˚C/s) (Bio-Rad, Hercules, CA, USA). PCR reactions were prepared at a volume of 12.5 µL using 20 – 80 pg of template DNA. An aliquot of DNA (2 µL) from each sample was added to 10.5 µL of PCR master mix. The PCR master mix consisted of 6.25 µL of 2X Type-it™ Multiplex PCR Mix (Qiagen), 1.25 µL 10X primer mix, 1.25 µL 5X Q-solution (Qiagen), and 1.75 µL deionized H_2O. Forward primers were labeled with a fluorescent dye (6-FAM™ or VIC™) with the optimal final concentrations of forward and reverse primers shown in Supplemental Table 2. PCR cycling conditions were performed using the following touchdown format: activation for 5 min at 95 ˚C followed by 1 cycle of 30 s at 95 ˚C, 90 s at 61 ˚C, 30 s at 72 ˚C, 1 cycle of 30 s at 95 ˚C, 90 s at 55 ˚C, 30 s at 72 ˚C, 29 cycles of 30 s at 95 ˚C, 90 s at 51 ˚C, 30 s at 72 ˚C, and a final extension of 30 min at 60 ˚C.

Capillary electrophoresis and genotyping

PCR products were separated and detected on a 3500 Genetic Analyzer (Thermo Fisher Scientific) using the parameters described in Houston et al. [5]. An allelic ladder was included with each injection and a customized bin set was designed to facilitate automated genotyping with the Genemapper ID v.5 software (Thermo Fisher Scientific). The analytical threshold was set to 100 Relative Fluorescence Units (RFUs).

Allelic ladder design

Forty *Cannabis* samples from various sources were selected to determine allelic variability. Using all alleles detected, an allelic ladder was generated according to previous reports [22, 23]. Briefly, the samples were amplified in single-plex and then the concentration (peak height) of all the amplicons was balanced. Due to the small number of alleles, there was no need to generate an individual ladder for each locus. Instead, the balanced samples (alleles) were combined to obtain a complete allelic ladder for all five loci genotyped

Allele sequencing

For the five homopolymeric STRs, at least two samples (alleles) were selected for sequencing. PCR amplification and cycle sequencing were performed on the Veriti® Fast thermal cycler (Thermo Fisher Scientific) using the BigDye® Direct Cycle Sequencing Kit (Thermo Fisher Scientific) as per the manufacturer's protocol [24] except for the annealing temperature (specific annealing temperature was used for each marker, Table 4.3.). Samples were sequenced on the 3500 Genetic Analyzer (Thermo Fisher Scientific) using the parameters described in Houston et al. (5). Alignment and proofreading was performed using the Geneious Pro Software R7.1.9 (Biomatters, Auckland, New Zealand). Sequences were submitted to Genbank (accession numbers shown in Table 4.3.).

Table 4.3. Characteristic of chloroplast and mitochondrial markers used in this study

Marker	Dye	Type of repeat	Final primer (SBE for SNP based) concentration (µM)	Annealing temperature (˚C)	Observed alleles	Genbank accession no.
STR Based						
Cscp001	6-FAM™	Single base INDEL	0.04	60.8	10, 11	MG196001 – 2
Cscp002	6-FAM™	homopolymer	0.01	55.1	10, 11, 12	MG196003 – 5
Cscp003	6-FAM™	homopolymer	0.02	51.4	7, 8	MG196006 – 7
Cscp004	6-FAM™	homopolymer	0.04	51.4	10, 11, 12	MG196008 – 10
Csmt001	VIC™	homopolymer	0.03	55.1	24, 27	MG196013 – 14
SNP Based						
Cscp001	n/a	Single base indel	0.10	58.1	T/C	MG196001 – 2
Cscp005	n/a	SNP	0.13	60.8	C/A	MG196011 – 12
Csmt002	n/a	SNP	0.30	63.5	C/T	MG196015 – 16

Dynamic range analysis

To assess the dynamic range of the multiplex assay, dilutions of three different *Cannabis* DNA samples were prepared to generate template cpDNA amounts of 140, 120, 100, 80, 60, 40, 20, 10, 5, and 2 pg for each DNA sample. The 21 dilutions were amplified and processed in three separate runs using the multiplex method.

Chloroplast and mitochondrial SNP typing

SNP triplex design and annealing temperature determination

Using the default parameters, the Primer3 software [25] was used to design three PCR primer pairs. In addition, the Autodimer software [26] was utilized to detect any primer–primer interactions. Forward and reverse sequences are displayed in Supplemental Table 4.1. For the PCR primer pairs, annealing temperatures were experimentally determined for each primer set using the HotStar Taq *Plus* Master Mix (Qiagen) on using an Eppendorf Master Cycler Gradient (ramp rate: 3 ˚C/s) (Eppendorf) as per Houston et al. [5]. A mini-sequencing method (SNaPshot®, Thermo Fisher Scientific) was chosen for SNP genotyping. Therefore, single base extension (SBE) primers were designed (Table 4.2). Poly-T tails of different sizes were added to the 5’ ends of the three SBE primers to ensure effective size separation during capillary electrophoresis. For a balanced SNP profile, primer titrations were performed with the SBE primers until optimization (Table 4.3.).

Multiplex PCR and SNaPshot of SNP triplex

Cannabis SNP profiling was conducted in a three–loci multiplex format modified from a previous study [8]. The multiplex consisted of previously published *Cannabis* chloroplast (Cscp001 and Cscp005) and mitochondrial (Csmt002) SNP markers [8]. PCR

amplification was performed using the Type-it™ Microsatellite PCR Kit (Qiagen) on a T100™ Thermal Cycler (ramp rate: 3 ˚C/s) (Bio-Rad). PCR reactions were prepared at a volume of 12.5 µL using 20 – 80 pg of template DNA. An aliquot of DNA (2 µL) from each sample was added to 10.5 µL of PCR master mix. The PCR master mix consisted of 6.25 µL of 2X Type-it™ Multiplex PCR Mix (Qiagen), 1.25 µL 10X primer mix, 1.25 µL 5X Q-solution (Qiagen), and 1.75 µL deionized H_2O. Both forward and reverse primers were unlabeled and equimolar at a final concentration of 0.2 µM. PCR cycling parameters were as follows: activation for 5 min at 95 ˚C followed by 30 cycles of 30 s at 95 ˚C, 90 s at 60 ˚C, 30 s at 72 ˚C, and a final extension of 30 min at 60 ˚C. PCR products were then purified to remove unincorporated primers and deoxynucleotides (dNTPs). For purification, 5 µL of calf alkaline phosphatase (CIAP) (1U/µL, Promega Corporation, Madison, WI, USA) and 2 µL of Exonuclease I (10U/µL, Invitrogen) were added to the PCR product. The samples were then incubated for 1.5 h at 37 ˚C followed by 30 min at 75 ˚C. Next, the SBE assay was performed using the SNaPshot™ Multiplex Kit (Thermo Fisher Scientific) as per manufacturer's instructions [27]. The SBE products were then purified with 1 µL of CIAP followed by incubation for 1.5 h at 37 ˚C and 30 min at 75 ˚C.

Capillary electrophoresis and genotyping

Separation and detection of purified SBE products was performed on a 3500 Genetic Analyzer (Thermo Fisher Scientific) with the following parameters: oven 60˚C; prerun 15 kV, 180s; injection 1.6 kV, 8 s; run 15 kV, 560 s; capillary length 50 cm; polymer: POP-7™; and dye set E5. Customized bins were designed to analyze the SNPs using the Genemapper ID v.5 software (Thermo Fisher Scientific). An analytical threshold of 100 RFUs was applied during analysis.

Sequencing of SNPs

For Cscp005 and Csmt002, at least one sample per SNP, was selected for allele confirmation via Sanger sequencing. PCR amplification was performed on the Eppendorf Master Cycler Gradient (Eppendorf) using the Type-it™ PCR Amplification Kit (Qiagen). PCR reactions were performed in single-plex following the same reaction and cycling parameters described in the "Multiplex PCR and SNaPshot of SNP triplex" section. Cycle sequencing were performed on the Veriti® Fast thermal cycler (Thermo Fisher Scientific) using the BigDye® Terminator v.3.1 (Thermo Fisher Scientific) as per manufacturer's protocol [28]. Separation and detection was performed following the parameters described in Houston et al. (5) and sequences were submitted to Genbank (accession numbers displayed in Table 4.3.).

Statistical analysis

Autosomal STR typing

For autosomal typing, the number of multi-locus genotypes and genotype sharing amongst samples was determined. Phylogenetic analysis of the 21 seizures at the US-Mexico border was performed with the Genetic Data Analysis (GDA) software [29] using the UPGMA method, with coancestry identity as genetic distance. To obtain the best phylogenetic tree, parsimony analysis was performed with the *PAUP* 4.0a* (build 157) software using a heuristic search [30]. Finally, case-to-case pairwise comparisons with F_{ST} as genetic distance and bootstrapping over loci to obtain 95% confidence interval for F_{ST} were performed with Arlequin v. 3.5 and GDA software, respectively [29] to determine a reference population from the 21 seizures. $P<0.05$ was accepted as the level of significance.

Phylogenetic analysis was assessed among the reference population (US-Mexico seizures), Brazil, Chile, and hemp samples. A distance matrix was assessed with the GDA software using the Neighbor Joining method with coancestry as genetic distance. Next, the *PAUP* 4.0a* (build 157) was invoked to perform parsimony analysis. An exhaustive search with hemp designated as an outgroup was performed to examine the genetic structure among the four populations. In addition, the Arlequin v. 3.5 software was used to perform pair-wise comparisons among the four populations using *Fst* as genetic distance [31]. To further examine population structure, the *STRUCTURE* software was used to evaluate the Bayesian clustering of genotypes from the four populations [32]. The parameters were as follows: admixture model without prior on sample origin, clusters from 1 to 12 groups (K), and ten replicates per K used. Each run consisted of 100,000 iterative steps after an initial burn-in of 100,000 steps. Next the Evanno method was assessed in STRUCTURE HARVESTER to predict the most likely number of clusters that explained the population structure [33]. The CLUMPAK package (Clustering Markov Packager Across K) was used to invoke two software: CLUster Matching and Permutation Program (CLUMPP) and DISTRUCT [34]. CLUMPP was used to permute and align the ten replicates as closely as possible while DISTRUCT was used to obtain the graphical display of the bar plots. Finally, the individual genotypes were visualized using Principal Component Analysis (PCA) with the R based software, *Adegenet* [35].

Organelle typing

For tissue type quantitation, data were tested for statistical significance by Analysis of variance (ANOVA) with Neumann-Keuls *post-hoc* comparisons, or *Student*'s t-test when appropriate. $P < 0.05$ was chosen as the level of significance.

For mitochondrial and chloroplast typing, subsampling was performed, and 134 samples were genotyped. The number haplotypes and haplotype sharing amongst samples was determined. Concordance between the two methods (STR and SNP) was evaluated using the Cscp001 marker.

Phylogenetic analysis was assessed between the reference population (US-Mexico seizures), Brazil, Chile, and hemp samples. A distance matrix was calculated with the GDA software using the Neighbor Joining method with coancestry as genetic distance. Next, the *PAUP* 4.0a* was invoked to perform parsimony analysis. An exhaustive search with hemp designated as an outgroup was performed to examine the genetic structure among the four populations.

Results and discussion

Validation studies of the *Cannabis* cpDNA real-time PCR quantitation method

The limit of detection of the real-time PCR assay was determined to be 0.02 pg/μL by running 10 standards (1000 to 0.00001 pg/μL) in triplicate. At 0.01 pg/μL and below, the linearity of the standard curve consistently dropped below an R^2 value of 0.99.

Forensic DNA evidence may contain a mixture of DNA from different species, and DNA extraction methods are not species-specific. *Cannabis* seizures may contain a mixture of plant types and/or contaminating human DNA. The real-time PCR primers may bind to and amplify non-*Cannabis* DNA and yield unreliable quantification values of *Cannabis* DNA in the sample. To avoid any non-*Cannabis* amplification, we selected a region of *Cannabis* cpDNA that is minimally homologous with other species (animal and plant). The Cscp001 was chosen because of its specificity for *Cannabis* and represents a single base insertion-deletion located within the *trn*L – *trn*F region of cpDNA [8]. The specificity of

this region was demonstrated by amplifying DNA from 14 non-*Cannabis* species. Minimal cross-reactivity was observed in 11 of the 14 non-*Cannabis* species. However, all cross-reactivity yielded quantification results below the limit of detection (< 0.01 pg/µL). As expected, the most significant cross–reactivity was observed in *Humulus lupulus* (Hops) as it is the closest genetic relative to *Cannabis* (0.003 pg/µL).

Data analysis from 18 separate runs confirmed the high sensitivity, reproducibility, and precision of the assay. The inter-run precision, expressed as the percent coefficient of variation of cycle threshold (Ct) (%CV = 100 x (standard deviation/mean)) had an average of 3.14%. Among 18 separate assays, 1000 pg/µL of the synthetic standard exhibited a Ct value 13.26 (range 11.66 – 13.96) (Table 4.4.). The subsequent five-fold dilution (200 pg/µL) exhibited a value of Ct of 16.07 (range 14.58 – 16.92). As expected, standards #1 and #2 (1000 and 200 pg/µL, respectively) exhibited the highest degree of variation with an average %CV of 5.27% and 3.95%, respectively.

Table 4.4. Quantification standard cycle threshold (Ct) data from 18 separate real-time PCR runs

Standard	*Cannabis* DNA (pg/uL)	Average	Standard Deviation	%CV	Minimum	Maximum	Range
1	1000	13.26	0.70	5.27	11.66	13.96	2.30
2	200	16.07	0.64	3.95	14.58	16.92	2.34
3	100	16.96	0.58	3.45	15.49	17.84	2.35
4	10	20.24	0.59	2.92	18.80	20.94	2.14
5	2	22.88	0.59	2.56	21.30	23.57	2.27
6	1	23.62	0.60	2.56	22.01	24.71	2.70
7	0.1	26.33	0.64	2.43	25.13	27.11	1.98
8	0.02	28.86	0.58	2.01	27.56	29.85	2.29

Reproducibility and precision were further demonstrated by compiling standard curves from 18 separate assays (Fig. 4.1.). Table 4.5. displays the consistently high amplification efficiencies as well as reproducible linear regression data from each of the 18 runs. In addition, the three *Cannabis* samples (positive controls) tested the functionality of the assay by monitoring reproducibility and precision. Low Ct and quantity estimate variation was observed for all three controls across the 18 runs.

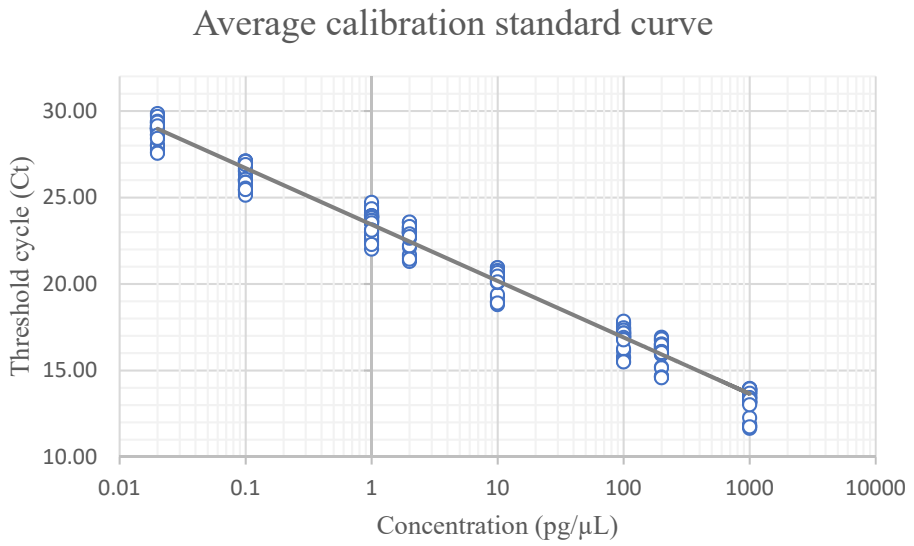


Fig. 4.1. Reproducibility of the standard calibration curve. The plot represents an average calibration standard curve generated from Ct values, corresponding to the quantity of the standard. Ct values are from 18 runs where each standard was amplified in duplicate. The trend line representing the average Ct values, has an R2 of 0.9829 and a slope of -3.26, corresponding to an amplification efficiency of 99.83%

Table 4.5. Linear regression data from 18 separate real-time PCR runs

Run	Slope	Amplification efficiency (%)	R^2	Y-intercept
1	-3.372	101.57%	0.998	22.314
2	-3.309	99.67%	0.995	22.600
3	-3.354	101.02%	0.998	23.887
4	-3.232	97.35%	0.996	23.958
5	-3.337	100.51%	0.997	24.033
6	-3.258	98.13%	0.996	24.158
7	-3.352	100.96%	0.998	23.726
8	-3.283	98.89%	0.996	23.480
9	-3.294	99.22%	0.997	23.497
10	-3.31	99.70%	0.996	23.625
11	-3.326	100.18%	0.998	22.071
12	-3.239	97.56%	0.998	23.884
13	-3.161	95.21%	0.996	23.472
14	-3.137	94.49%	0.996	23.504
15	-3.313	99.79%	0.998	23.636
16	-3.122	94.04%	0.996	23.664
17	-3.145	94.73%	0.994	23.353
18	-3.132	94.34%	0.994	23.024
Average	-3.260	98.19%	0.997	23.438
Standard Deviation	0.086	2.58%	0.001	0.583
Coefficient of Variation (%)	2.62%	2.62%	0.13%	2.49%
Minimum	-3.372	94.04%	0.994	22.071
Maximum	-3.122	101.57%	0.998	24.158
Range	0.25	7.53%	0.004	2.087

To date, this is the first publication concerning *Cannabis* organelle typing using a real-time PCR method for DNA quantitation. The Federal Bureau of Investigations Quality Assurance Standard 9.4 states that the amount of human DNA should be estimated using quantitation standards prior to DNA amplification [36]. Accordingly, an equivalent standard should be applied prior to amplification of non-human DNA. Although a method to quantify nuclear *Cannabis* DNA has been previously published, [22] a quantification assay specific to *Cannabis* organelle DNA has yet to be reported. No predictable ratio of nuclear DNA/organelle DNA is possible due to copy number variation. Indeed, the amount of cpDNA is variable depending on the type of plant tissue used for DNA extraction and the growth cycle in which the plant was harvested [37]. Real-time PCR quantification is a fast and reliable method to calculate the DNA concentration of a sample and may predict downstream PCR success. Nevertheless, the development of this quantification method requires the use of reference DNA standards. For human DNA, these reference materials are available through the National Institute of Standards and Technology (NIST) [7]. In the case of genomic *Cannabis* DNA, a primary DNA standard can be generated from a pool of concentrated extracts followed by quantification via UV absorbance reading at 260 nm [38]. However, this method cannot be easily applied to produce organelle DNA reference standards due to the difficulty in isolating organelle DNA during DNA extraction [39, 40]. Instead, a previous report with human mitochondrial DNA showed that an organelle DNA reference standard could be developed using synthesized DNA [41]. Using synthetic DNA as calibration standards allows the method to be reproducible between laboratories as NIST reference standards are not available for *Cannabis* nuclear or cpDNA. In this work, an analytical assay for the real-time PCR quantification of *Cannabis* cpDNA was developed,

optimized, and validated, according to the SWGDAM guidelines using synthetic DNA standards.

Cannabis cpDNA and mtDNA typing design

Chloroplast and mitochondrial STR multiplex design

Chloroplast and mitochondrial *Cannabis* STR markers described by Gilmore et al. were used as the reference for this study [5]. However, the following modifications were made: (a) a multiplex format, (b) primer concentrations optimized with the Type – it® Microsatellite PCR Kit (Qiagen), (c) annealing temperature determination for each marker (primer set), and (d) allelic ladder design. An example of an electropherogram of the five loci STR multiplex system is shown in Fig. 4.2. Annealing temperatures ranged from 51 ˚C to 61 ˚C (Table 4.2.). Due to the wide range of annealing temperatures, a touchdown PCR method was employed to amplify all five organelle markers in a single reaction.

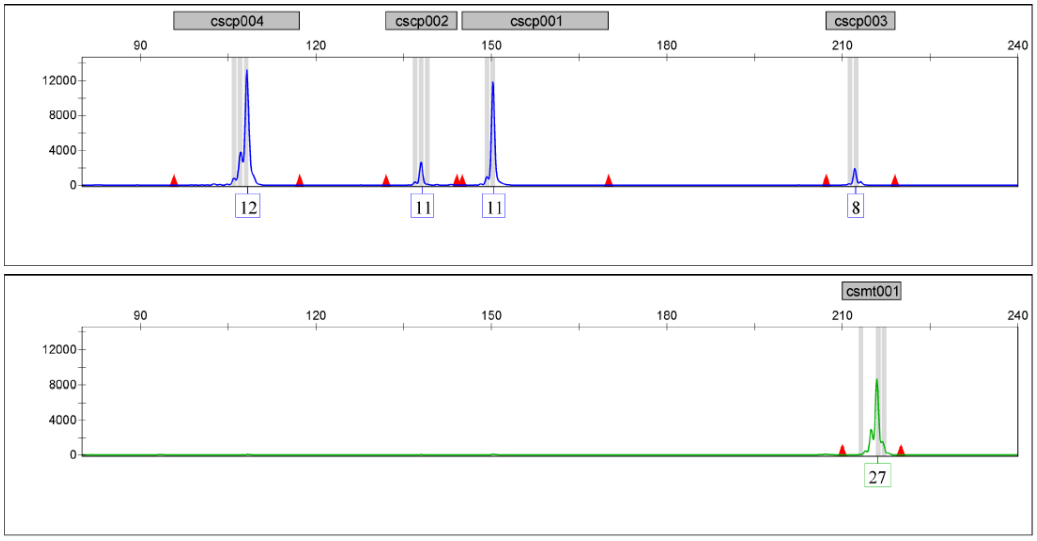


Fig. 4.2. Chloroplast and mitochondrial haplotype of *Cannabis* sample #11-D2 (homopolymer STR profile)

Allelic ladder and sequencing

For the homopolymeric pentaplex, an allelic ladder was designed for the alleles observed in the populations genotyped (Fig. 4.3.). The allelic ladder consisted of 12 alleles across the five homopolymeric loci. Allele nomenclature following the international guidelines (ISFG) was used to designate the alleles. The proposed nomenclature and detailed sequencing results can be found in Figs. 4.4. – 4.8. All alleles were confirmed by sequencing to ensure accurate allele designation. The use of an allelic ladder is critical for homopolymeric genotyping due to the inter-allelic single nucleotide difference.

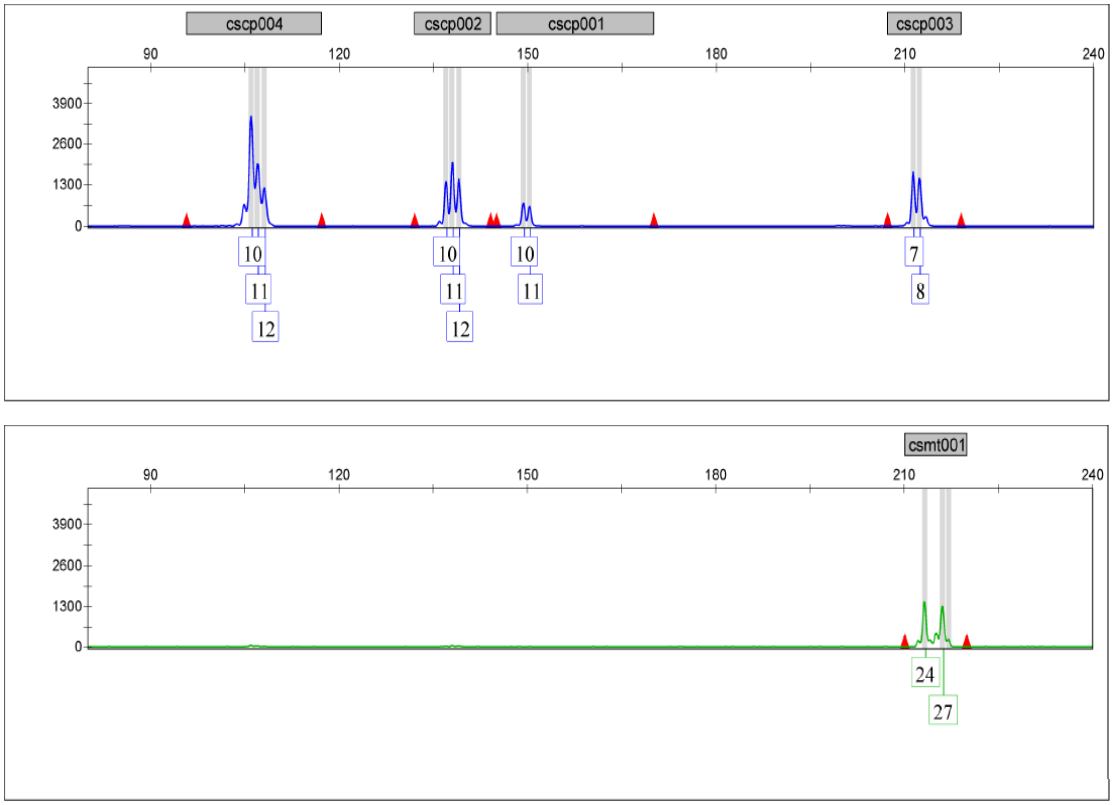


Fig. 4.3. Homopolymeric pentaplex STR allelic ladder

This is a note for Figs. 4.4. through 4.8. In the consensus sequences, the FW and RV primer binding sites are **underlined. The location of the repeat structure is indicated in the consensus sequence as **[REPEAT].** The accession numbers of the reference contigs are also **referenced** in the table. **N** refers to the total number of alleles found bearing the haplotype described.

TCCTCTCATTCCGTTAGTGGTTTCTAATTTGTTATGTTTCTCGTTCATTCTAACTT
TACAACCGGACCTGAATGA**[REPEAT]**ATTATCACAAGCCTTGTGATATATATGA
AAGACCTACAAATGAACATAAGGAATCCCAATGTGCAATT

Allele	[REPEAT]		N	Genbank Accesion Number
	C	T		
10	2	8	1	MG196001
11	3	8	2	MG196002

Fig. 4.4. Consensus sequence of Cscp001 locus, haplotypes found and allele nomenclature proposal.

TCATTTGATGAAGTGGGGTACTGAAAA**[REPEAT]**CTTTTTTGAGAACCCGTAG
TATCGTTTTGCTATATATGCTAAAATAGGATGAAACCCACTTTTCAATTATAAAT
AATTAATGTGAAATAGTAGGTTCCCCATGC

Allele	[REPEAT]	N	Genbank Accesion Number
	T		
10	10	1	MG196003
11	11	2	MG196004
12	12	1	MG196005

Fig. 4.5. Consensus sequence of Cscp002 locus, haplotypes found and allele nomenclature proposal

GATCCCGGACGTAATCCTGGACGTGAAGAATAAAAAATAAAGAAGATTTTTTG
[REPEAT]GCTTGATTTTAAAAAGTTCTTAGTAGGGTTTTAGCTATTTCCCACTTT
TAACTATAAGAAAATAACTAAAAAAAAGGGAACTCGCGAAAAATTCGAAAGG
AAATACAAGGTTATTGACGAAAACGGAAAGAGAGGGATTCGAACCCTCGGTA
CGAT

Allele	[REPEAT]	N	Genbank Accesion Number
	T		
7	7	1	MG196006
8	8	1	MG196007

Fig. 4.6. Consensus sequence of Cscp003 locus, haplotypes found and allele nomenclature proposal

CGATGCATATGTAGAAAGCCTA**[REPEAT]**CGAGTATTTATTAATGGATTCACTCT
TTTTTTTCTTTTCACTTTTTATTTCTATAGTGGAGATAGTCGCACGGTAATG

Allele	[REPEAT]	N	Genbank Accesion Number
	T		
10	10	1	MG196008
11	11	1	MG196009
12	12	2	MG196010

Fig. 4.7. Consensus sequence of Cscp004 locus, haplotypes found and allele nomenclature proposal

ATGGCAGAGAAGTTTCCATATTTATACCTTTTCTTGTTGGAGGGGCGACCGTCCGTTG
AACTACC**[REPEAT]**GATCCATTTCTTTAGTCTTTAGGGAGCCAA

Allele	[REPEAT]				N	Gen Bank Accession Number
	A	N96	C	T		
24	7	GGGTAAACCAATGTGATCATGACA TTGTAGGTGCTTGCGATGGGACGG ATGCGACTTTCCTCAGTTGGTTTG GGTGGCATAGCCCGTTGCAGAAGT	9	8	2	MG196013
27	8		11	8	2	MG196014

Fig. 4.8. Consensus sequence of csmt001 locus, haplotypes found and allele nomenclature proposal

Dynamic range

The dynamic range (optimal input range) of the five loci STR multiplex, when using the cpDNA quantitation method developed in this paper, was determined to be from 40 to 80 pg of template DNA (Fig. 4.9.). Some drop-out was observed below 40 pg and minor pull-up was observed above 80 pg. Due to the narrow range of optimal input DNA, it is essential to use an accurate and reproducible cpDNA quantitation method to ensure optimal downstream results.

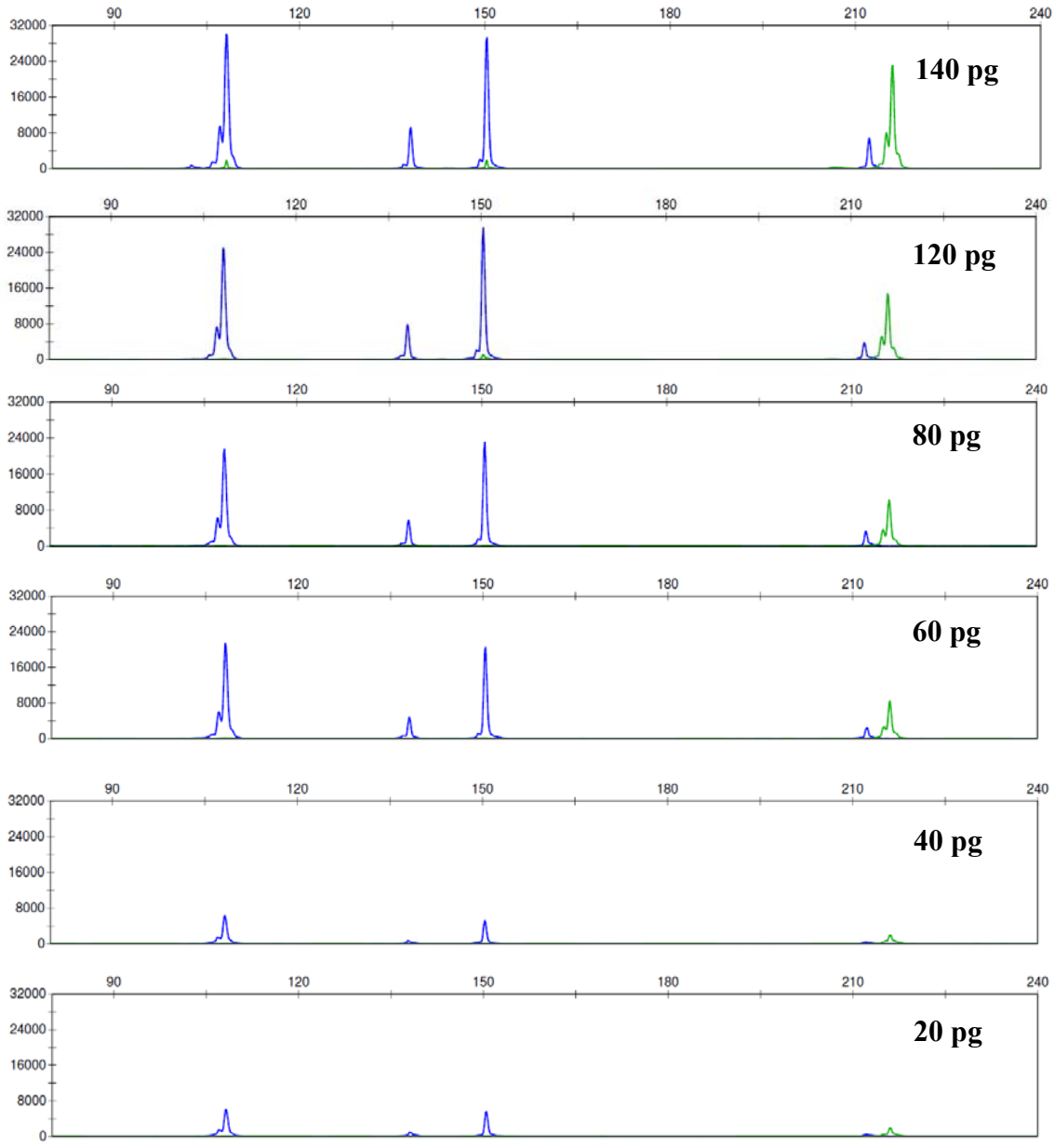


Fig. 4.9. Representative electropherograms overlaying the blue and green channels for the different amounts of template cpDNA using the multiplex organelle STR assay. The amount of DNA template tested was determined using the *Cannabis* real-time PCR quantitation method. The optimal input amount of the STR multiplex was determined to be from 40 to 80 pg of cpDNA

Chloroplast and mitochondrial SNP multiplex design and sequencing

Chloroplast and mitochondrial *Cannabis* SNP markers described by Gilmore et al. were used as the reference for this study [5]. However, the following modifications were made: (a) a multiplex format, (b) use of a SNaPshot-based assay for genotyping, and (c) annealing temperature determination for each marker (primer set). An example electropherogram of the three loci SNP profile is shown in Fig. 4.10. Annealing temperatures ranged from 58 ˚C to 63.5 ˚C (Table 4.3.). SNPs were confirmed via Sanger sequencing using the Big Dye Terminator v.3.1; detailed sequencing results are displayed in Figs. 4.11, 4.12.

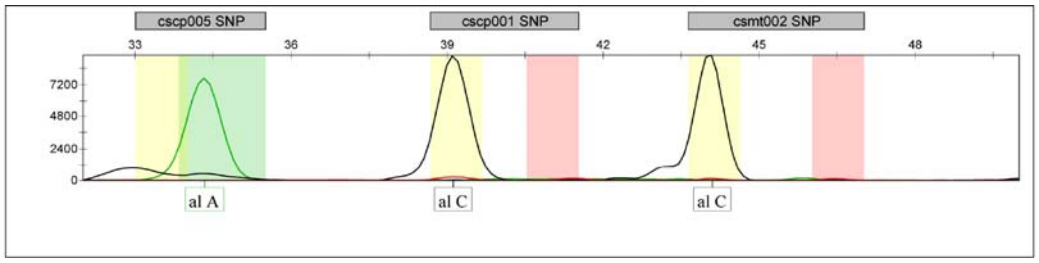


Fig. 4.10. Chloroplast and mitochondrial haplotype of *Cannabis* sample #11-D2 (SNP profile)

This is a note for Figs. 4.11. and 4.12. In the consensus sequence, the FW and RV primer binding sites are **underlined and the Single Base Extension (SBE) primer bind sites are **highlighted**. The **SNP** is indicated by its **nucleotide ambiguity code**. The accession numbers of the reference contigs are also **referenced** in the table. **N** refers to the total number of alleles found bearing the haplotype described.

TCCACTGCCTTGATCCCTTGGCTACATCCGCCCCTATATTAAATATTAAACAAAA
TTTTTTTAGTTTATTTGAATA**K**ATTTCAATTTTAGACAAGATAAAAGAAATTGAA
ACCTTTATTTTTATTTAATATCGAAATAATAAAATAAAAAAGAGAAGGATAAACT
GATAGAAATGAATATATTAATTATAAAAATATATTGAATCTTGAAGGAAAGAAAA
AAACTTATGTAACTAAAAAAAAAAAAAAAATACGAAATAATAAAAGGAGCAAT
ACTAAACTTCTTGATAGAAGTTTGGTATTGCTCCTTTAGCTTTATTTTCAATAAC
TACTCATATAGACTAATACCGAAGTTTTATCCATTTGTAGATGGAACTTCTAGAG
CAGCTAAGTCTAGAGGG

SNP		N	Genbank Accession Number
K (forward)	**R (reverse)**		
G	C	1	MG196011
T	A	1	MG196012

Fig. 4.11. Consensus sequence of cscp005 locus, haplotypes found, and allele nomenclature proposal. Reverse strand SNP is shown here because SBE primer used was a reverse primer and the SNP sequenced in the SBE reaction was the reverse strand

TTGTGCGAAGAGTGCGTTATGACCTGTGGCCGCCTGYCTGGTGGGGGCGGCT
CCTCCGTTGTGGGTAAACGGGAAACCCGACTCTACGAACCCGAGGAAAGGCT
GCACAGCAGTAGTAGGGGCGTTAAGACCGGAGCTTTTTGTAGTGCTAGCAGG
AGTGCAAGTGAATGAATCCCATCCCCTAGCGAGTGAAGT

SNP	N	Genbank Accession Number
Y		
C	1	MG196015
T	1	MG196016

Fig. 4.12. Consensus sequence of csmt002 locus, haplotypes found, and allele nomenclature proposal

Statistical analysis

Autosomal statistical analysis

All samples (N=510) were successfully amplified using the 13-loci multiplex format. However, only 425 out of 510 samples (83%) yielded full STR profiles. Majority of partial profiles were due to mixtures or low template DNA (<100pg). A full breakdown of STR success and number of genotypes can be found in Table 4.6.

Table 4.6. STR success and sample breakdown of four *Cannabis* populations

Source	Sample Number	Partial Profiles	Mixed Profiles	Full Profiles	Unique Genotypes	Duplicate Genotypes
US-Mexico	422	23	32	367	326	18 (41 samples)
Brazil	8	2	0	6	6	0
Chile	50	18	0	32	4	7 (28 samples)
Hemp	30	9	1	20	20	0
	510	52	33	425	356	25 (69 samples)

From the full profiles, 356 distinguishable genotypes were generated and 25 identical genotypes within seizures were observed. Genotype duplication within seizures was most likely due to sampling of same plant twice either when tissue sub-sampling was performed on an individual *Cannabis* plant or during inadvertent sub-sampling of the same plant. From the Chile samples, seven identical genotypes were observed. When looking at partial and full profiles from the Chile samples, it was noted that nine out of the ten seizures contained identical genotypes for all five samples within the seizures. It may be hypothesized that the nine seizures or cases contained marijuana that was clonally propagated.

Phylogenetic analysis and subsequent parsimony analysis of the 21 seizures at the US-Mexico revealed a genetic relatedness between all samples. Case-to-case pairwise comparisons of 21 seizures at the US-Mexico border, using F_{ST} as genetic distance, revealed the genetic association of nine seizures (N=157 samples) that formed a reference population. The F_{ST} between these nine seizures was calculated to be close to zero and relatedness was confirmed using 95% confidence interval bootstrap analysis.

Phylogenetic and parsimony analysis between the reference population (US-Mexico seizures), Brazil, Chile, and hemp samples could discriminate the four populations (Fig. 4.13.). Pair-wise comparisons with the Arlequin v. 3.5 software using F_{ST} as genetic distance revealed that all populations were different at a statistically significant level (p<0.01) (Table 4.7.). Interestingly, the THC-positive samples (CBP, Chile and Brazil) form a different cluster when compared to hemp.

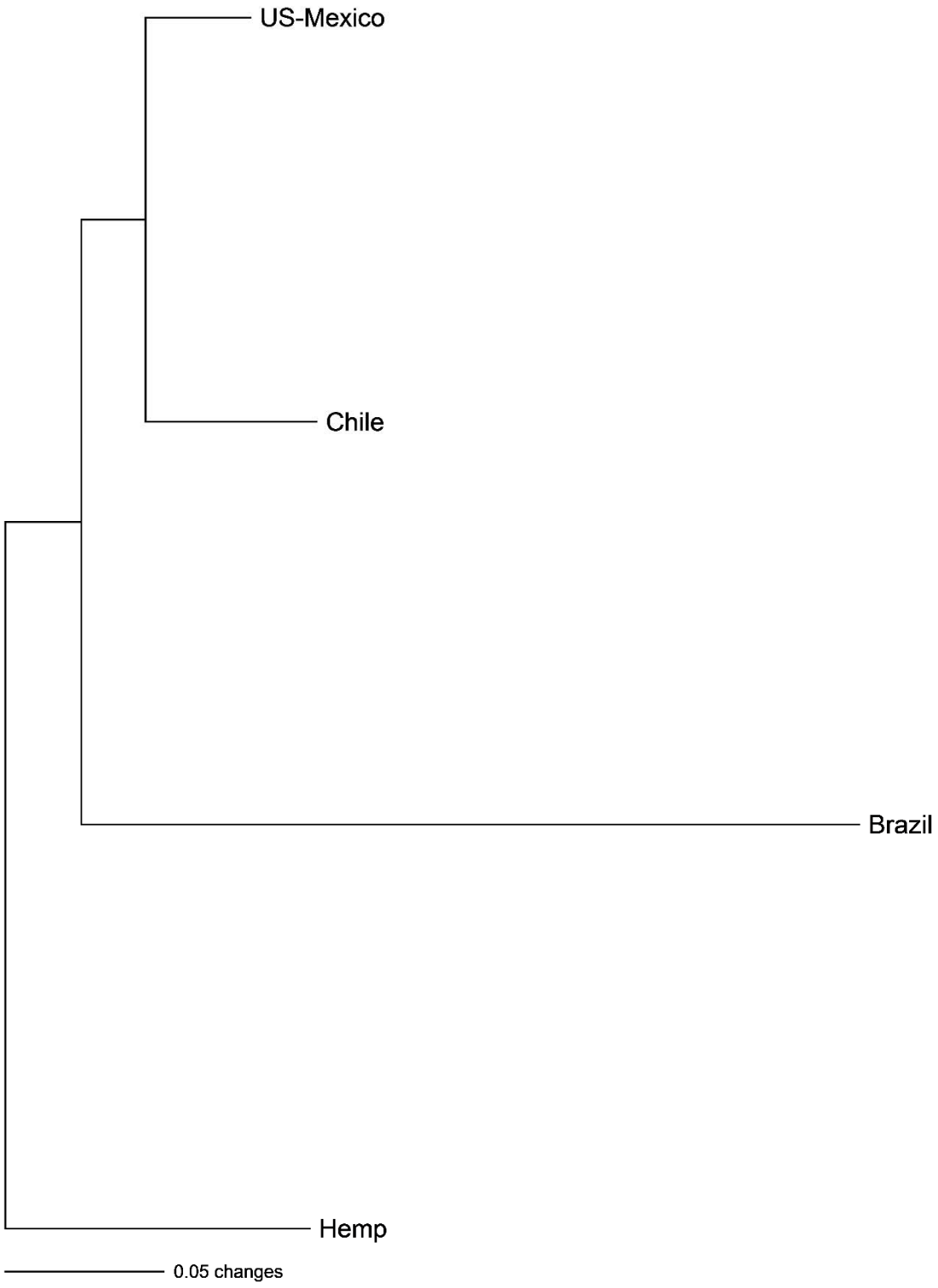


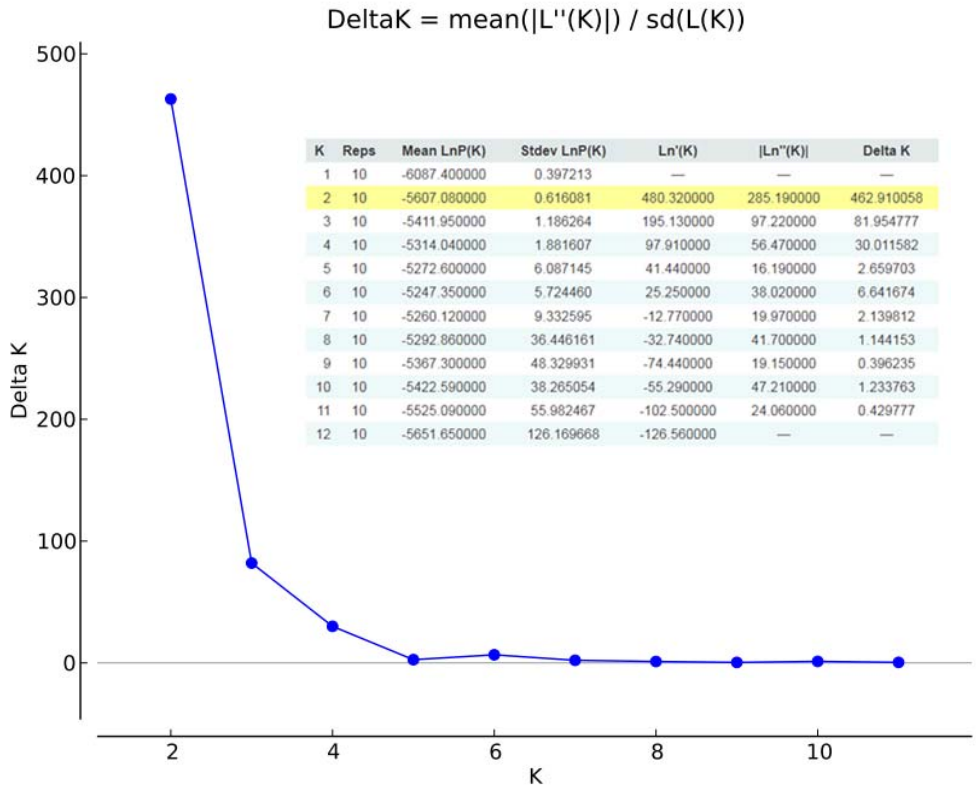
Fig. 4.13. Neighbor joining tree depicting genetic distances among four *Cannabis* population sets using autosomal genotypes; coancestry as genetic distance. Parsimony analysis using exhaustive search was performed

Table 4.7. Population-to-population comparison among four *Cannabis* populations using pairwise genetic-distance analysis based on F_{ST}

Population	US-Mexico	Brazil	Hemp
Brazil	0.29906 (0.00000[a])		
Hemp	0.16445 (0.00000[a])	0.37381 (0.00000[a])	
Chile	0.08506 (0.00000[a])	0.32181 (0.00000[a])	0.19731(0.00000[a])

Probability values of F_{ST} are displayed in parentheses
[a] Statistically significant differences at 0.01 levels

The *STRUCTURE* software was used to evaluate Bayesian clustering of the four populations. Structure Harvester results using the Evanno method revealed that K=2 was the maximum delta k (Fig. 4.14.).



K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
1	10	-6087.400000	0.397213	—	—	—
2	10	-5607.080000	0.616081	480.320000	285.190000	462.910058
3	10	-5411.950000	1.186264	195.130000	97.220000	81.954777
4	10	-5314.040000	1.881607	97.910000	56.470000	30.011582
5	10	-5272.600000	6.087145	41.440000	16.190000	2.659703
6	10	-5247.350000	5.724460	25.250000	38.020000	6.641674
7	10	-5260.120000	9.332595	-12.770000	19.970000	2.139812
8	10	-5292.860000	36.446161	-32.740000	41.700000	1.144153
9	10	-5367.300000	48.329931	-74.440000	19.150000	0.396235
10	10	-5422.590000	38.265054	-55.290000	47.210000	1.233763
11	10	-5525.090000	55.982467	-102.500000	24.060000	0.429777
12	10	-5651.650000	126.169668	-126.560000	—	—

Fig. 4.14. Structure Harvester results (graph and table) for maximum delta K calculation using the Evanno Method. K=2 was determined to be the maximum delta K according to Structure Harvester

Although Structure Harvester indicated that K=2 was the number of clusters that best described the data, the resulting bar plots for K=3 describe the datasets better based on phylogenetic data. Resulting bar plots (K=2 – 4) from the CLUMPAK software are shown in Fig. 4.15.

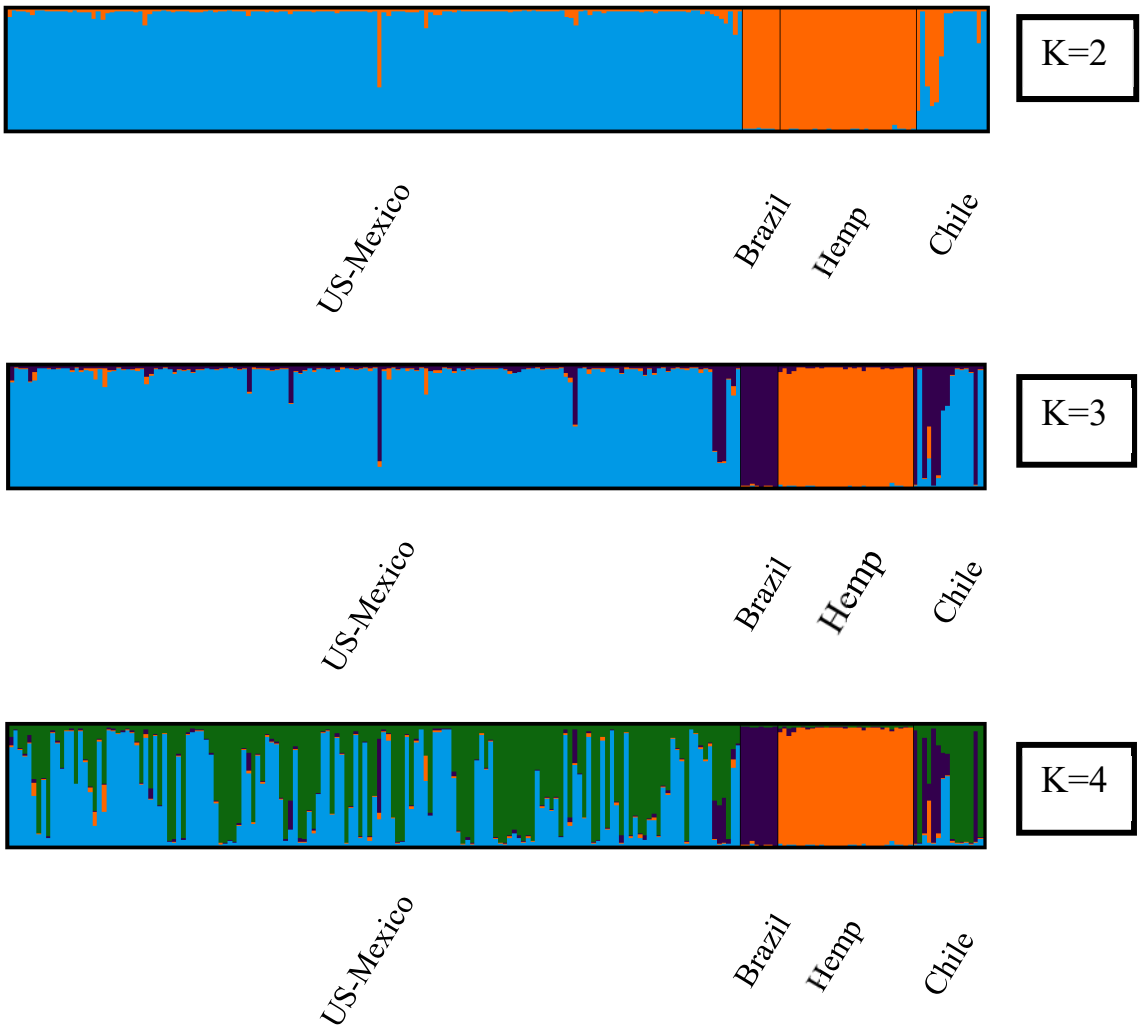


Fig. 4.15. Bayesian clustering based on autosomal genotypes from four *Cannabis* datasets using the STRUCTURE software. Results for K=2, K=3, and K=4 are shown. Iterations were combined and visualized using the CLUMPAK software. Colors in the bar plot depict the probability of assignment to each cluster

From the K=3 bar plot, there is a clear distinction between the hemp and the other three groups. This genetic difference was previously reported by Gilmore et al. using organelle data [8] and by Dufresnes et al. based on autosomal PCA analysis [42]. There is also some genetic sharing amongst US-Mexico, Chile, and Brazil samples. Using K=3, the Chilean population shows a genetic admixture between US-Mexico and Brazil. This is not unexpected since the samples share a similar biogeographical origin. Moreover, organelle genetic data confirms this hypothesis as major haplotype sharing was observed among these three groups. Finally, a PCA plot is displayed in Fig. 4.16. with the *Adegenet* software. The PCA plot shows a genetic relatedness among the three drug types (US-Mexico, Brazil, and Chile) and a distinction from the fiber type (Hemp). However, it is still possible to differentiate the three drug datasets. This differentiation of *Cannabis* samples from different origins could be useful in tracking the flow of marijuana.

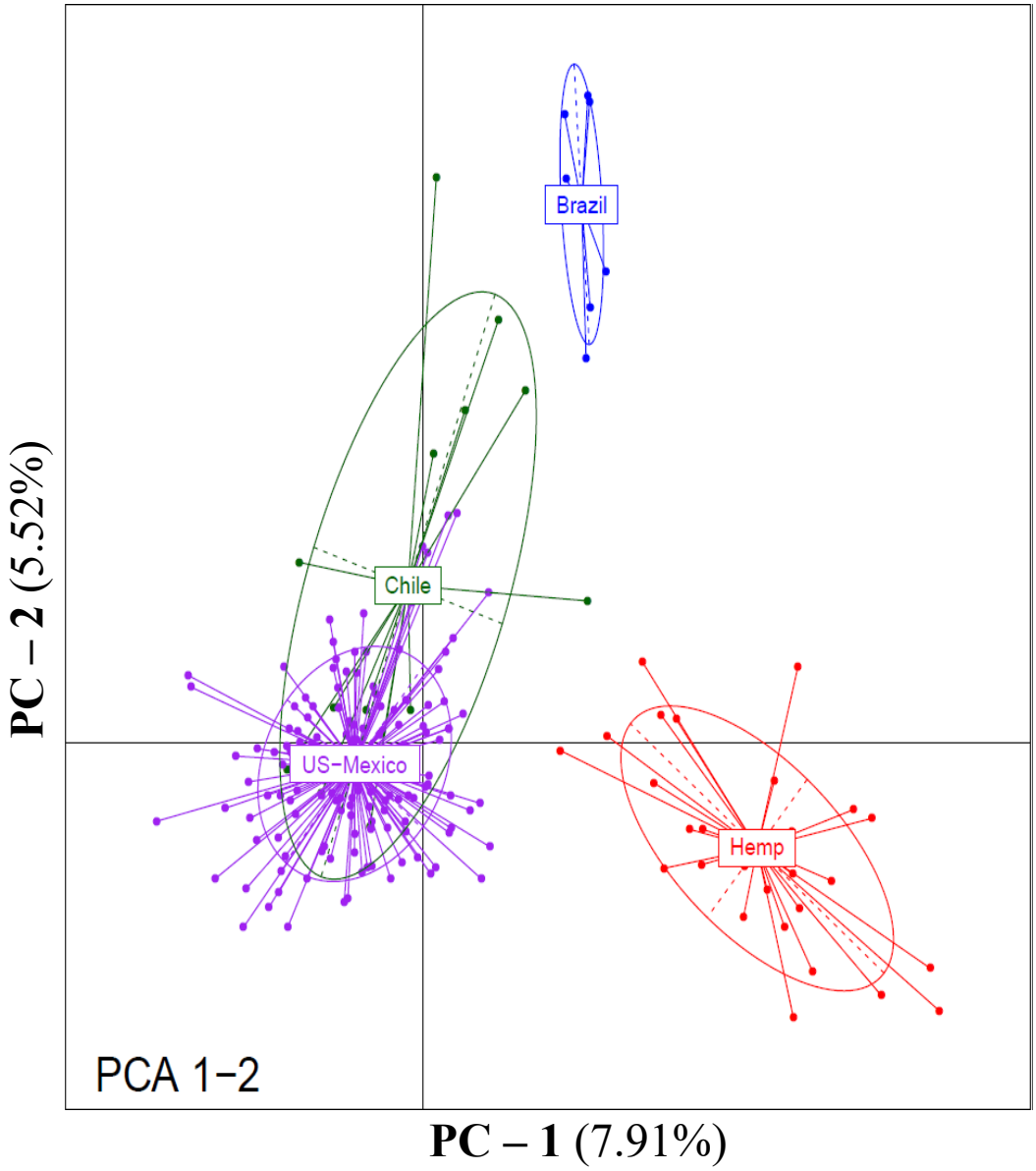


Fig. 4.16. Principal component analysis (PCA) on autosomal genotypes from four *Cannabis* datasets. The ellipses illustrate 95% inertia of each dataset while the dots represent individuals in the dataset. The eigenvalues for the first three principal components are 6.484, 4.523, and 3.580, respectively. The corresponding relative variance of principal component 1 and 2 are shown as a percent on the axes

Chloroplast and mitochondrial statistical analysis

CpDNA was successfully extracted from all *Cannabis* samples (*N*=510). The average amount (± standard deviation) of DNA extracted was 2.56 ± 4.18 ng/mg of plant tissue.

Four tissue types (stem, flower, seed, and leaf) were targeted in four different *Cannabis* plants to determine relative quantity of cpDNA. Seeds, followed by leaf, were shown to have the highest concentration of chloroplast DNA (Fig. 4.17.). The high concentration of cpDNA in the seed may be due to the high density of cells and its role in reproduction in plants. One-way ANOVA analysis showed that tissue type ($F_{3,12}$=9.4, p<0.01) had a statistically significant effect on the amount of cpDNA extracted. Statistically significant differences were found between seed and flower tissues (p<0.01) as well as between seed and stem tissues (p<0.01).

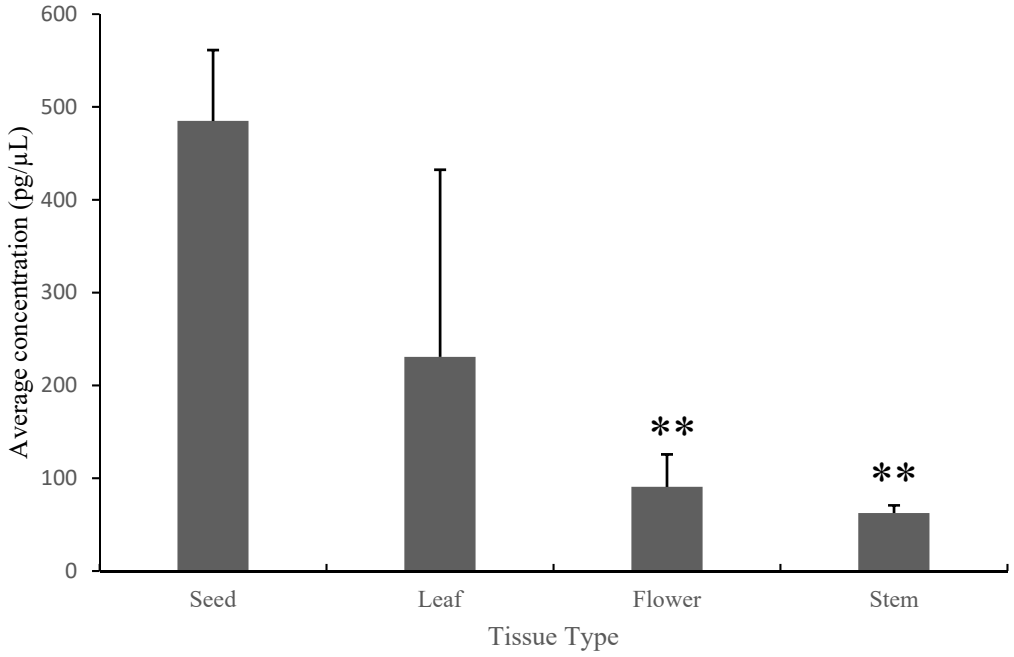


Fig. 4.17. Relative cpDNA quantitation (pg/µL) by *Cannabis* tissue type (*N*=4). Error bars represent standard deviations

** p-value < 0.01 when compared to seed tissue.

Due to predicted haplotype sharing, subsampling was performed for mitochondrial and chloroplast typing, and 134 samples were genotyped. Complete haplotypes (STRs and SNPs) were observed for 127 samples (Table 4.8.). As expected, extensive haplotype sharing was observed; only five distinguishable haplotypes were detected. In the reference population, the same haplotype was observed 39 times and two unique haplotypes were also detected. Haplotype sharing was observed between the US border seizures, Brazil, and Chile while the hemp samples generated a distinct haplotype. Complete allele concordance was observed for chloroplast marker Cscp001 using both typing methods (STR and SNP).

Table 4.8. Chloroplast and mitochondrial haplotypes of samples from Mexico, Brazil, Chile, and Canada observed in this study

Population	Code	N	Country of Origin	Haplotypes
Drug				
CBPCASE1	DC1	3	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE2	DC2	2	Mexico	(11)(11)(8)(12)(A)(27)(C) - 1
				(11)(10)(8)(12)(A)(27)(C) - 1
CBPCASE3	DC3	3	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE4	DC4	13	Mexico	(11)(11)(8)(12)(A)(27)(C) - 12
				(10)(10)(8)(12)(A)(27)(C) - 1
CBPCASE5	DC5	3	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE6	DC6	2	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE7	DC7	5	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE8	DC8	3	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE9	DC9	3	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE10	DC10	2	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE11	DC11	7	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE12	DC12	5	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE13	DC13	3	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE14	DC14	4	Mexico	(11)(11)(8)(12)(A)(27)(C) - 3
				(11)(10)(8)(12)(A)(27)(C)
CBPCASE15	DC15	3	Mexico	(11)(11)(8)(12)(A)(27)(C) - 2
				(11)(10)(8)(12)(A)(27)(C) - 1
CBPCASE16	DC16	3	Mexico	(11)(11)(8)(12)(A)(27)(C) - 1
				(10)(11)(8)(10)(A)(24)(C) - 1
CBPCASE17	DC17	9	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE18	DC18	7	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE19	DC19	7	Mexico	(11)(11)(8)(12)(A)(27)(C)
CBPCASE20	DC20	7	Mexico	(11)(11)(8)(12)(A)(27)(C) - 6
				(11)(10)(8)(12)(A)(27)(C) - 1
CBPCASE21	DC21	9	Mexico	(11)(11)(8)(12)(A)(27)(C)
BRZ1	DB1	2	Brazil	(10)(11)(8)(10)(A)(24)(C)
CHL1	DCH1	1	Chile	(11)(11)(8)(12)(A)(27)(C)
CHL2	DCH2	1	Chile	(11)(11)(8)(12)(A)(27)(C)
CHL3	DCH3	2	Chile	(11)(11)(8)(12)(A)(27)(C)
CHL4	DCH4	2	Chile	(11)(11)(8)(12)(A)(27)(C)
CHL5	DCH5	2	Chile	(10)(10)(8)(10)(A)(24)(C)
CHL6	DCH6	2	Chile	(11)(11)(8)(12)(A)(27)(C)
CHL8	DCH8	2	Chile	(11)(11)(8)(12)(A)(27)(C)
CHL9	DCH9	2	Chile	(10)(11)(8)(10)(A)(24)(C)
CHL10	DCH10	2	Chile	(11)(11)(8)(12)(A)(27)(C)
Fiber				
Navitas	FN1	2	Canada	(11)(12)(7)(11)(C)(27)(C)
Badia	FB1	2	Canada	(11)(12)(7)(11)(C)(27)(C)
Manitoba	FM1	2	Canada	(11)(12)(7)(11)(C)(27)(C)

The phylogenetic and parsimony analysis among the reference population (US-Mexico seizures), Brazil, Chile, and hemp samples is displayed in Fig. 4.18. The phylogenetic analysis of the organelle haplotypes between the four populations yielded similar results to autosomal genotypes.

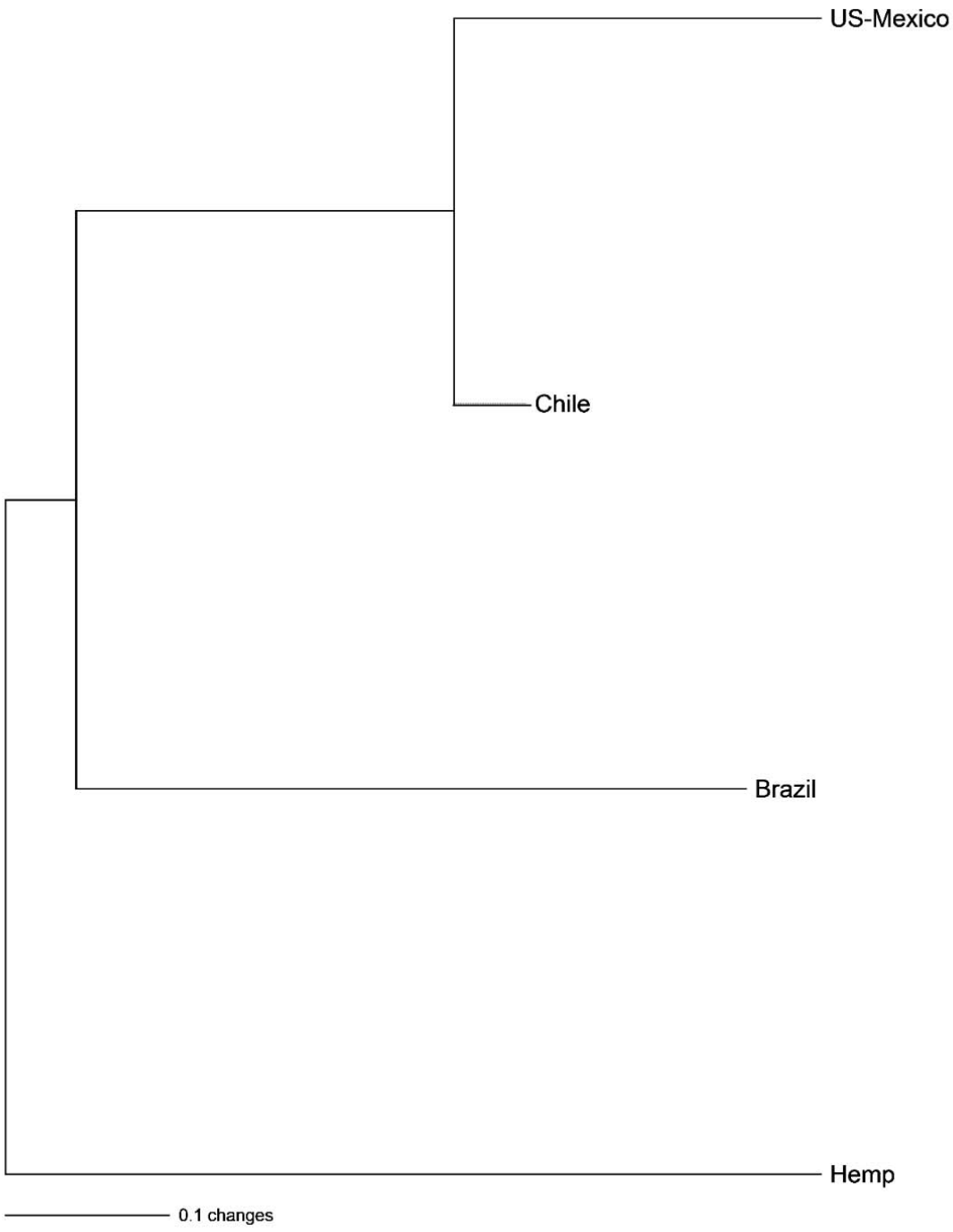


Fig. 4.18. Neighbor joining tree depicting genetic distances among four *Cannabis* population sets using chloroplast and mitochondrial haplotypes; coancestry as genetic distance. Parsimony analysis using exhaustive search was performed

Conclusions

The goal of this study was to genotype both autosomal and organelle DNA from a marijuana DNA database to elucidate population structure. Autosomal typing was accomplished using a previously validated method. For organelle typing, a novel real-time PCR quantification method was developed for determining the amount of cpDNA in *Cannabis* samples prior to downstream PCR-based analysis in accordance with SWGDAM guidelines. Organelle typing was performed by modifying and optimizing a previously reported system to genotype five chloroplast and two mitochondrial markers. Two novel methods were developed: a homopolymeric STR pentaplex and a SNP triplex with one marker (Cscp001) shared by both methods for quality control. Results revealed that both autosomal and lineage markers could discern population sub-structure and may be useful in classifying seized *Cannabis* samples.

In summary, this study demonstrates the applicability of genotyping both autosomal and organelle DNA for *Cannabis* samples and presents, for the first time, a US DNA database of *Cannabis* samples for nuclear, chloroplast, and mitochondrial DNA.

Funding information

This study was partially funded by a Graduate Research Fellowship Award #2015-R2-CX-0030 (National Institute of Justice, Office of Justice Programs, U.S. Department of Justice). The opinions, findings, conclusions, or recommendations expressed in this presentation are those of the authors and do not necessarily reflect those of the National Institute of Justice.

Acknowledgments

The authors would like to thank all staff and personnel at the U.S. Customs and Border Protection LSSD Southwest Regional Science Center for their great assistance and help with this project. The authors would also like to thank Roberta Marriot and Alejandra Figueroa for their kind donation of marijuana DNA extracts. Lastly, the authors greatly appreciate Haleigh Agot for her assistance with the chloroplast quantitation method.

References

1. Small E, Cronquist A (1976) A practical and natural taxonomy for *Cannabis*. *Taxon* 25:406-435
2. Adams IB, Martin BR (1996) *Cannabis*: pharmacology and toxicology in animals and humans. Addiction 91:1585-1614
3. Howard C, Gilmore S, Robertson J, Peakall R (2008) Developmental validation of a C*annabis Sativa* STR multiplex system for forensic analysis. J of Forenic Sci 53:1061-1067. https://doi.org/10.1111/j.1556-4029.2008.00792.x
4. Köhnemann S, Nedele J, Schwotzer D, Morzfeld J, Pfeiffer H (2012) The validation of a 15 STR multiplex PCR for *Cannabis* species. Int J Legal Med 126:601-606. https://doi.org/10.1007/s00414-012-0706-6
5. Houston R, Birck M, Hughes-Stamm S, Gangitano D (2017) Developmental and internal validation of a novel 13 loci STR multiplex method for *Cannabis sativa* DNA profiling. Legal Med (Tokyo, Japan) 26:33-40. https://doi.org/10.1016/j.legalmed.2017.03.001
6. Dumolin-Lapegue S, Pemonge MH, Petit RJ (1997) An enlarged set of consensus primers for the study of organelle DNA in plants. Mol Ecol 6:393-397
7. Kohjyouma M, Lee IJ, Iida O et al (2000) Intraspecific variation in *Cannabis sativa* L. based on intergenic spacer region of chloroplast DNA. Biol Pharm Bull 23:727-730. https://doi.org/10.1248/bpb.23.727
8. Gilmore S, Peakall R, Robertson J (2007) Organelle DNA haplotypes reflect crop-use characteristics and geographic origins of *Cannabis sativa*. Forensic Sci Int 172:179-190. https://doi.org/10.1016/j.forsciint.2006.10.025

9. Zhang Q, Sodmergen, Liu Y (2003) Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. Plant and Cell Physiol 44:941-951

10. Vergara D, White KH, Keepers KG, Kane NC (2016) The complete chloroplast genomes of *Cannabis sativa* and *Humulus lupulus*. Mitochondrial DNA Part A, DNA mapping, sequencing, and analysis 27:3793-3794. https://doi.org/10.3109/19401736.2015.1079905

11. White KH, Vergara D, Keepers KG, Kane NC (2016) The complete mitochondrial genome for *Cannabis sativa*. Mitochondrial DNA Part B 1:715-716. https://doi.org/10.1080/23802359.2016.1155083

12. Weising K, Gardner RC (1999) A set of conserved pcr primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome 42:9-19

13. Drew BT, Ruhfel BR, Smith SA, Moore MJ, Briggs BG, Gitzendanner MA, Soltis PS, Soltis DE (2014) Another look at the root of the angiosperms reveals a familiar tale. Syst Biol 63:368-382

14. Yang MQ, van Velzen R, Bakker FT, Sattarian A, Li DZ, Yi TS (2013) Molecular phylogenetics and character evolution of Cannabaceae. Taxon 62:473-485. https://doi.org/10.12705/623.9

15. Demesure B, Sodzi N, Petit RJ (1995) A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Mol Ecol 4:129-131

16. Mello IC, Ribeiro AS, Dias VH, Silva R, Sabino BD, Garrido RG, Seldin L, de Moura Neto RS (2016) A segment of rbcl gene as a potential tool for forensic discrimination of *Cannabis sativa* seized at rio de janeiro, brazil. Int J Legal Med 130:353-356. https://doi.org/10.1007/s00414-015-1170-x

17. Dias VH, Ribeiro AS, Mello IC, Silva R, Sabino BD, Garrido RG, Seldin L, Moura-Neto RS (2015) Genetic identification of *Cannabis sativa* using chloroplast trnl-f gene. Forensic Sci Int Genet 14:201-202. https://doi.org/10.1016/j.fsigen.2014.10.003

18. Linacre A, Gusmao L, Hecht W, Hellmann AP, Mayr WR, Parson W, Prinz M, Schneider PM, Morling N (2011) ISFG: Recommendations regarding the use of non-human (animal) DNA in forensic genetic investigations. Forensic Sci Int Genet 5:501-505. https://doi.org/10.1016/j.fsigen.2010.10.017

19. Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC, Markham AF (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res 17:2503-2516

20. Scientific Working Group on DNA Analysis Methods: Validation guidelines for DNA analysis methods. (2016) https://docs.wixstatic.com/ugd/4344b0_813b241e8944497e99b9c45b163b76bd.pdf. Accessed 10 September 2017

21. DNeasy® Plant Mini Kit Handbook. (2012) Qiagen, Hilden, Germany

22. Houston R, Birck M, Hughes-Stamm S, Gangitano D (2016) Evaluation of a 13-loci STR multiplex system for *Cannabis sativa* genetic identification. Int J Legal Med 130:635-647. https://doi.org/10.1007/s00414-015-1296-x

23. Griffiths RAL, Barber MD, Johnson PE, Gillbard SM, Haywood MD, Smith CD, Arnold J, Burke T, Urquhart AJ, Gill P (1998) New reference allelic ladders to improve allelic designation in a multiplex STR system. Int J of Legal Med 111:267-272.

24. Bigdye® Direct Cycle Sequencing Kit. (2011) Thermo Fisher Scientific, South San Francisco, CA

25. Koressaar T, Remm M (2007) Enhancements and modifications of primer design program primer3. Bioinformatics 23:1289-1291. https://doi.org/10.1093/bioinformatics/btm091

26. Vallone PM, Butler JM (2004) Autodimer: A screening tool for primer-dimer and hairpin structures. BioTechniques 37:226-231.

27. Abi Prism® Snapshot™ Multiplex Kit Protocol 4323357b. (2010) Thermo Fisher Scientific, South San Francisco, CA

28. Bigdye™ Terminator v3.1 Cycle Sequencing Kit. (2016) Thermo Fisher Scientific, South San Francisco, CA

29. Lewis P, Zaykin D (2001) Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c)

30. Swofford D (2002) PAUP*: phylogenetic Analysis Using Parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, MA

31. Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564-567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

32. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959.

33. Earl DA, vonHoldt BM (2012) Structure harvester: A website and program for visualizing structure output and implementing the evanno method. Conserv Genet Resour 4:359-361. https://doi.org/10.1007/s12686-011-9548-7

34. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour 15:1179-1191. https://doi.org/10.1111/1755-0998.12387

35. Jombart T (2008) Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403-1405. https://doi.org/10.1093/bioinformatics/btn129

36. Quality Assurance Standards for Forensic DNA Testing Laboratories (2009). https://www.fbi.gov/file-repository/quality-assurance-standards-for-forensic-dna-testing-laboratories.pdf. Accessed 10 September 2017

37. Fulgosi H, Jezic M, Lepedus H, Stefanic PP, Curkovic-Perica M, Cesar V (2012) Degradation of chloroplast DNA during natural senescence of maple leaves. Tree Physiol 32:346-354. https://doi.org/10.1093/treephys/tps014

38. Kline MC, Duewer DL, Travis JC, Smith MV, Redman JW, Vallone PM, Decker AE, Butler JM (2009) Production and certification of nist standard reference

material 2372 human DNA quantitation standard. Anal Bioanal Chem 394:1183-1192. https://doi.org/10.1007/s00216-009-2782-0

39. Mourad GS (1998) Chloroplast DNA isolation. In: Martinez-Zapater JM, Salinas J (eds) Arabidopsis Protocols. Humana Press, Totowa, NJ, pp 71-77

40. Diekmann K, Hodkinson TR, Fricke E, Barth S (2008) An optimized chloroplast DNA extraction protocol for grasses (*Poaceae*) proves suitable for whole plastid genome sequencing and snp detection. PLoS ONE 3:e2813. https://doi.org/10.1371/journal.pone.0002813

41. Kavlick MF, Lawrence HS, Merritt RT, Fisher C, Isenberg A, Robertson JM, Budowle B (2011) Quantification of human mitochondrial DNA using synthesized DNA standards. J of Forenisc Sci 56:1457-1463. https://doi.org/10.1111/j.1556-4029.2011.01871.x

42. Dufresnes C, Jan C, Bienert F, Goudet J, Fumagalli L (2017) Broad-scale genetic diversity of *Cannabis* for forensic applications. PLoS ONE 12:e0170522. https://doi.org/10.1371/journal.pone.0170522

CHAPTER V

Massively parallel sequencing of 12 autosomal STRs in *Cannabis sativa*[1]

This dissertation follows the style and format of *International Journal of Legal Medicine*.

[1] Houston R, Mayes C, King JL, Hughes-Stamm S, Gangitano D (2018). Submitted to *Electrophoresis*.

Abstract

Massively parallel sequencing (MPS) is an emerging technology in the field of forensic genetics that provides distinct advantages compared to capillary electrophoresis CE. This study offers a proof of concept that MPS technologies can be applied to genotype autosomal short tandem repeats (STRs) in *Cannabis sativa*. A custom panel for MPS was designed to interrogate 12 *Cannabis*-specific STR loci by sequence rather than size. A simple workflow was implemented to integrate the custom PCR multiplex into a workflow compatible with the Ion Plus Fragment Library Kit, Ion™ Chef, and Ion™ S5 System. For data sorting and sequence analysis, a custom configuration file was designed for STRait Razor v3 to parse and extract STR sequence data. This study represents a preliminary investigation of sequence variation for 12 autosomal STR loci in 16 *Cannabis* samples from three different countries. Full concordance was observed between the MPS and CE data. Results revealed intra-repeat variation in eight loci where the nominal or size-based allele was identical, but variances were also discovered in the sequence of the flanking region. Although only a small number of *Cannabis* samples were evaluated, this study demonstrates that more informative STR data can be obtained via MPS.

Keywords: *Cannabis sativa*, Forensic plant science, Ion™ S5, Massively parallel sequencing, Short tandem repeats

Introduction

Massively parallel sequencing (MPS) technology provides a platform for more comprehensive coverage of genetic markers. MPS technologies can sequence DNA in a massively parallel fashion with high coverage and high throughput of specified targets. In recent years sequencing costs and run times of the MPS systems have dropped substantially and now offer a potentially cost-effective approach to genetically characterize samples for genetic identification purposes [1, 2]. MPS technology has been successfully tested in the fields of medicine, microbiology, environmental, and forensic science [2-5] and offers an invaluable opportunity to expand its applications to the field of forensic plant science, specifically the genetic identification of *Cannabis sativa* samples. Previous studies have shown the value of STR typing for the genetic identification of marijuana [6-10]. As with human identification (HID), capillary electrophoresis (CE) of STR markers is the gold standard for the DNA-based identification of marijuana for forensic or intelligence purposes. While CE offers a reliable and robust technique, it has disadvantages such as limited multiplexing capability with a maximum of 25 to 30 loci configurable across five dye channels [11]. In addition, MPS has the potential to provide deeper interrogation of sequence-based polymorphisms, which in turn allows for a greater power of discrimination compared to size-based STR genotyping by CE.

Currently, no targeted MPS workflows have been used for *C. sativa*. Instead, *Cannabis* studies have focused on using Genotyping by Sequencing (GBS) strategies [12, 13]. GBS provides researchers an alternative to array-based screening of single nucleotide polymorphisms (SNPs) and offers a way to compare samples in the absence of a reference genome. While this type of sequencing may be useful for agricultural and medicinal

purposes, targeted sequencing is needed for forensic comparisons. Targeted sequencing without a commercial MPS panel can be complicated due to the difficulty in integrating a customized panel with a commercial library preparation kit. Custom panels can be designed by manufacturers (i.e. Thermo Fisher Scientific and Illumina); however, *Cannabis* is not currently a supported species. PCR is a common method of targeting the DNA to only sequence the regions of interest and ensure adequate coverage for those regions. Nevertheless, other studies have reported success using customized MPS panels for human identification including a 10-loci STR multiplex [14], 13-loci Y-STR multiplex [15], and 23-loci Y-STR multiplex [16].

Another difficulty with targeted sequencing of a custom panel is creating a bioinformatic pipeline to compile and analyze the sequence data. Only a draft genome currently exists for *C. sativa* [17] making alignment-based analyses difficult. STRait Razor is a parsing and analysis tool that does not rely on alignment for analysis [18]. Instead, STRait Razor uses an algorithm to search for 5' and 3' anchor sequences within the sequencing data to target the locus of interest. The current version of STRait Razor (v3) is compatible with Microsoft Windows and is a free, adaptable bioinformatics suite [19]. Although originally designed for HID MPS panels, this tool is easily customizable for targeted sequencing of any loci (e.g., STR/SNP) or species.

In this study, a multiplex PCR assay was designed for the amplification and subsequent sequencing of 12 previously reported *Cannabis*-specific STRs [20]. A multiplex PCR system was successfully utilized for MPS analysis of 12 STR markers from a previously validated STR multiplex for *Cannabis* genetic identification [20]. MPS performance including read depth, heterozygote balance, noise, and CE concordance was

assessed. Results demonstrated that MPS technologies can be used to genotype autosomal STRs in *C. sativa.* In addition, this study reveals a workflow that can be used to integrate any custom PCR multiplex into a MPS pipeline.

Materials and methods

DNA Samples

THC-positive *Cannabis* samples were obtained from three sources: U.S. Customs and Border Protection (N=8), Northeast Brazil (N=2), and the Araucarian region of Chile (N=3). Two samples from Chile (Chile 47 and Chile 48) were previously identified to be clones. Additionally, Canadian-grown hemp seeds (N=3) were purchased from three brands: Navitas™ Organics, Badia Spices Inc., and Manitoba Harvest Hemp Food. DNA extraction was performed according to Houston et al. [8]. DNA concentrations were estimated using the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific) on the Qubit® 2.0. Fluorometer (Thermo Fisher Scientific) [21]. Five nanograms of input DNA was used for MPS typing.

MPS panel design

The autosomal loci analyzed in this study consisted of 12 *Cannabis*-specific STR markers (ANUCS501, 9269, 4910, 5159, ANUCS305, 9043, B05-CANN1, 1528, 3735, D02-CANN1, C11-CANN1, H06-CANN2) from a previously validated multiplex [20]. Allele and sequence variation was obtained from Valverde et al. [7, 22] and Houston et al. [20, 23]. A custom AmpliSeq™ panel could not be designed as *Cannabis sativa* is not currently a supported species and a reference genome is not available. Instead, primer sequences (non-fluorescent) from a previous CE method were used [20]. Primer sequences and PCR parameters from the previously validated multiplex were used to ensure adequate

amplification efficiency of all amplicons. In addition, primer concentrations were titrated according at Houston et al. [20] to ensure a more balanced sequencing profile of the amplicons. To note, the hexanucleotide marker, CS1, was not included in MPS analysis due to the sequencing and data analysis challenges posed by the markers highly variable amplicon length and complexity of sequence.

Multiplex PCR amplification and quantitation

Multiplex PCR amplification was performed using the Type-it™ Microsatellite PCR kit (Qiagen, Hilden, Germany) on a T100™ Thermal Cycler (Bio-Rad, Hercules, CA, USA). Primer sequences and concentrations are displayed in Table 5.1. PCR reactions were prepared at a 20 µL volume using 5 ng of template DNA. The reaction consisted of 10 µL of 2X Type-it™ Multiplex PCR Mix (Qiagen), 2 µL 10X primer mix, 2 µL 5X Q-solution (Qiagen), 0.67 µL bovine serum albumin (8 mg/mL, Sigma-Aldrich, St. Louis, MO, USA) and 5.33 µL of template DNA/deionized H_2O. PCR cycling conditions were as follows: activation for 5 min at 95 ˚C followed by 29 cycles of 30 s at 95 ˚C, 90 s at 57 ˚C, 30 s at 72 ˚C, and a final extension of 30 min at 60 ˚C. Amplified products were purified using the MinElute® PCR Purification Kit (Qiagen) with an elution volume of 30 µL [24]. The quality of purified PCR product was assessed using the Agilent DNA 1000 Kit (Agilent Technologies, Santa Clara, CA, USA) on the Agilent 2100 Bioanalyzer (Agilent Technologies) as per manufacturers protocol [25]. Quantity of purified products was determined using the Qubit® dsDNA HS Assay kit (Thermo Fisher Scientific) as per manufacturer's recommendations [21].

Table 5.1. Primer information for the 12 loci in the multiplex system

Marker	Forward Primer	Reverse Primer	Primer conc. (µM)	Size Range (bp)
ANUCS501	AGCAATAATGGAGTGAGTGAAC	AGAGATCAAGAAATTGAGATTCC	0.10	80 – 95
9269	CCCAAACTACTGTTTGTGCC	ACTTGCACGTGATGTTAGATCC	0.10	131 – 139
4910	TCTCCAAAGACATTATTGAACAAA	GGTATCAAGAGCCAGGTTTCA	0.20	170 – 214
5159	CCAGAGCTTGTGGATCTCCT	AGTACGAAAGGGCACTGAGG	0.30	327 – 339
ANUCS305	AAAGTTGGTCTGAGAAGCAAT	CCTAGGAACTTTCGACAACA	0.10	141 – 162
9043	AAAGCTCGATGTCATCTCTACAC	TGCTCAATGCCTTATTCATGCT	0.15	179 – 195
B05-CANN1	TTGATGGTGGTGAAACGGC	CCCCAATCTCAATCTCAACCC	0.15	235 – 244
1528	TTGTCTAGTGCCTTTGTCATGC	AGGATGACCAAATTTGCTCCA	0.30	280 – 310
3735	TGATTCTGTGTTTGTGTGCAAT	CATCGCACCCACAGGTTAGT	0.10	79 – 99
D02-CANN1	GGTTGGGATGTTGTTGTTGTG	AGAAATCCAAGGTCCTGATGG	0.15	105 – 111
C11-CANN1	GTGGTGGTGATGATGATAATGG	TGAATTGGTTACGATGGCG	0.15	150 – 175
H06-CANN2	TGGTTTCAGTGGTCCTCTC	ACGTGAGTGATGACACGAG	0.15	266 – 273

Library preparation

Sequencing libraries were prepared using the Ion Plus Fragment Library Kit (Thermo Fisher Scientific) following the amplicon libraries without fragmentation protocol [26]. Based on the concentration from the Qubit assay, 100 ng of purified PCR product (79 µL) was added to the end repair reaction. End-repaired amplicons were purified using Agencourt™ AMPure™ XP Reagent (Beckman Coulter, Indianapolis, IN, USA) (1.8X sample volume). Adapters with barcodes were ligated using Ion Xpress™ Barcode Adapters one to 16 (Thermo Fisher Scientific). Barcode-ligated libraries were purified with the Agencourt™ AMPure™ XP Reagent (1.5X sample volume due to small size of some amplicons, ~ 80 bp before ligation). Library concentration was assessed using the Ion Library TaqMan® Quantitation Assay (Thermo Fisher Scientific) [27].

Templating and sequencing

Libraries (*N*=16) were normalized to 50 pM and pooled to a 25 µL volume for templating. The pooled libraries were templated with the Ion Chef™ System (Thermo Fisher Scientific) using the Ion 520™ and Ion 530™ Kit (Thermo Fisher Scientific) and then loaded onto one Ion 530™ chip (Thermo Fisher Scientific). Sequencing was performed on the Ion™ S5 System (Thermo Fisher Scientific) with 850 flows.

Data sorting and analysis

Sequencing data (base calls and quality scoring) was generated on the Torrent Suite Software v. 5.2.2. The reads were filtered by quality and separated by barcode within the Torrent Suite Software. Barcode separated FASTQ files were exported using the file exporter plugin. STRait Razor v3 was used for data analysis of the FASTQ files [19]. For this, a custom configuration file was designed to detect and extract autosomal STR data.

For backwards compatibility with a previous CE based assay [20], the configuration file was designed to extract STR data in a manner to capture the size-based allele determined by capillary electrophoresis. Custom anchors (5' and 3') and motif sequences were assigned for each locus. To note, sequences that differ by one base pair are tolerated within the STRait Razor algorithm. STRait Razor results were imported into a customized STRait Razor Excel workbook to collate and visualize data by sequence and allele call. Allele calls with a read depth greater than 50x coverage were considered for analysis. The read depth for each allele and heterozygote balance was calculated at each STR locus per sample. Heterozygote balance was calculated for heterozygote loci with the lower coverage allele divided by the higher coverage allele. In addition, relative noise percentage at each locus was assessed. Noise can be placed into three distinct categories: stutter (-2 repeat, -1 repeat, +1 repeat), noise at allele position, and artifacts [28, 29]. For this study, all three noise categories were combined when measuring relative noise at each locus. Percent noise was calculated at loci that were homozygous or heterozygotes that differed by at least four repeats.

STR typing by CE

The amount of nuclear DNA was previously estimated according to Houston et al. via real-time PCR on the StepOne™ Real-Time PCR System (Thermo Fisher Scientific) with SYBR™ Green PCR Master Mix (Thermo Fisher Scientific) and *Cannabis*-specific primers (ANUCS304) [23]. *Cannabis* STR profiling was performed in a 13-loci multiplex format using a previously validated method according to Houston et al. [20]. All loci were identical to the MPS method with the addition of the highly complex marker, CS1. Concordance was assessed for all samples between the CE method and MPS method.

Results and discussion

MPS library preparation

A relatively low amount of input DNA (5 ng) was used for the initial PCR reaction as per the suggestion of the library preparation kit (Ion Plus Fragment Library Kit) used. Recommendations for input into PCR are 10-100 ng of input DNA. While this input amount may seem high for forensic purposes, *Cannabis* cases typically consist of large seizures, where extracting sufficient amounts of DNA is not a concern. Indeed, only 10 mg of plant material is needed to yield sufficient DNA. To note, all samples required a dilution (~1:10) before input into the end-repair reaction. Thus, future studies may include reducing input DNA to test the tolerance of the Ion Plus Fragment Library Kit.

Data sorting and analysis

Results from the Torrent Suite Software revealed templating and sequencing was successful with 63% chip loading and 31% polyclonal for a final of 43% usable reads with an average read length of 207 bp. The final sequencing yield was 2.1 GB of data consisting of 10,124,641 quality-filtered reads that were obtained from sequencing 16 samples on an Ion 530™ chip. An average of 608,000 reads was obtained for each sample. Locus specific sequences were parsed and extracted using STRait Razor v3. Sequences were imported into a customized Excel workbook with Fig. 5.1. displaying an example histogram output of one sample (18-A5).

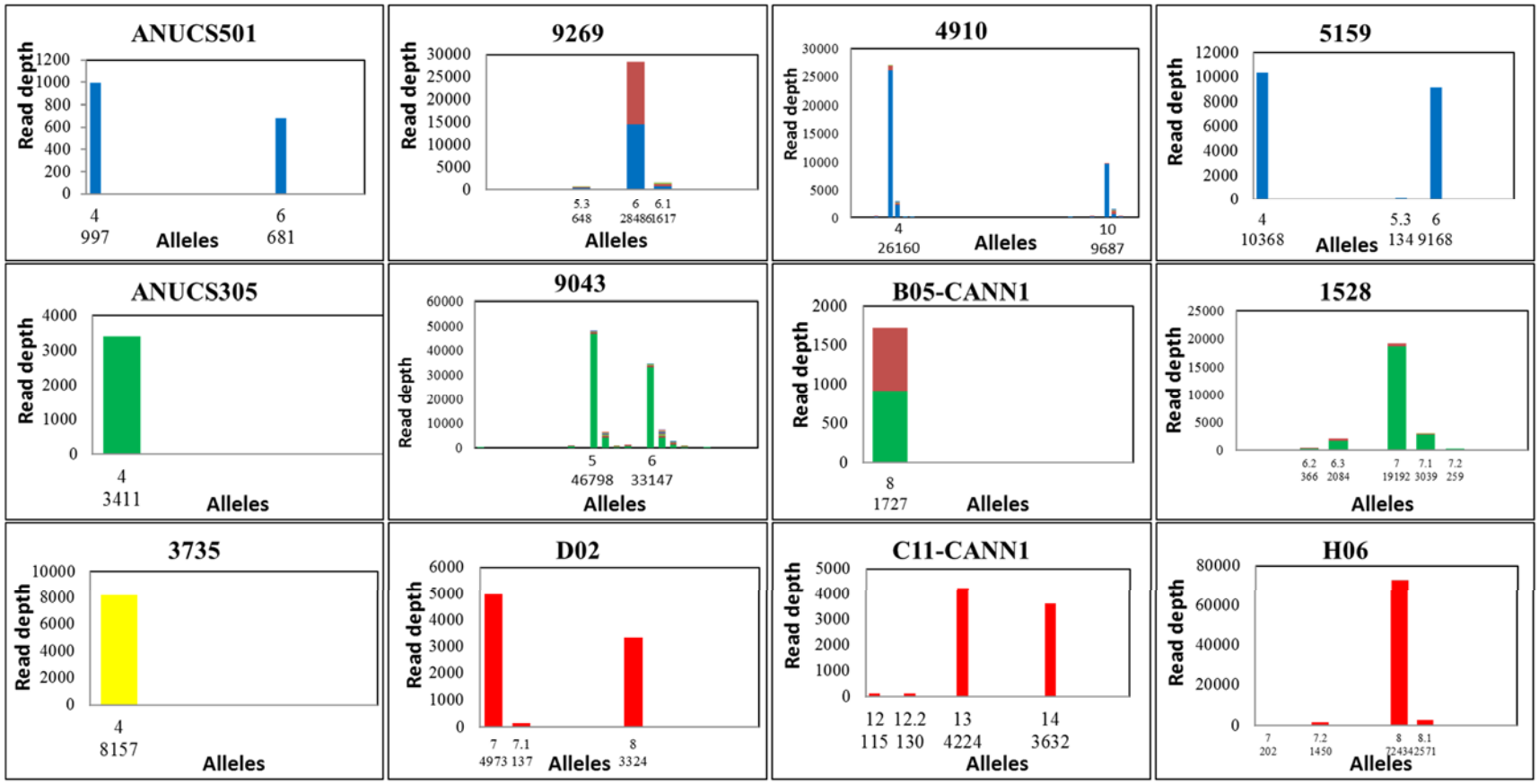


Fig. 5.1. A histogram portrayal of the allele calls and read depth for barcode 5 (18-A5). Nominal alleles with sequence variations (such as B05) are stacked on top of one another with a different color distinguishing the other allele

Sequence analysis

One distinct advantage afforded by MPS as compared to CE is the deeper interrogation of amplicons. This information can greatly increase the power of discrimination through intra-repeat variation and SNPs within the flanking region. Among the 16 samples sequenced, eight loci (9269, 5159, ANUCS305, B05-CANN1, 1528, C11-CANN1, D02-CANN1) that were genotyped as the same nominal allele by size could be differentiated by sequence (Table 5.2.). Intra-repeat variability has been previously reported by Valverde et al. [7, 22]; however, some sequence variations were novel to this study and were submitted to Genbank (Accession no. MH136628 – 40). A summary of sequencing variability amongst the 12 loci can be found in Figs. 5.2. – 5.13. To note, for continuity in the field of *Cannabis* genetics, the format for reporting sequence variability mirrors Valverde et al. [7, 22]. With new alleles discovered by sequencing variation (isoalleles), allele frequencies in reference population databases will need to be updated to take full advantage of the increased power of discrimination afforded by MPS. Given the relatively low number of *Cannabis* samples currently sequenced, it is likely that new isoalleles will be reported in the future, and is also possible that more sequencing data may necessitate new allele nomenclature, such as adjusting where to begin calling the STR.

Table 5.2. Allelic variation observed in this study by sequence. Count refers to the number of observations in this study only

Locus	Allele call by size	Repeat Motif	MPS allele	Count
ANUCS501	2	$[TTGTG]_2$	$[TTGTG]_2$ TGG	2
ANUCS501	4	$[TTGTG]_4$	$[TTGTG]_4$ TGG	13
ANUCS501	6	$[TTGTG]_6$	$[TTGTG]_6$ TGG	8
9269	4	$[ATAA]_4$	CTTTCTATCAGTATCTATAAAATATTAACAAGAAAAGAGCATT $[ATAA]_4$	1
9269	5.3	$[ATAA]_6$	CTT-CTATCAGTATCTATAAAATATTTACAAGAAAAGAGCATT $[ATAA]_6$	3
9269	6	$[ATAA]_6$	CTTTCTATCAGTATCTATAAAATATTAACAAGAAAAGAGCATT $[ATAA]_6$	10
9269	6	$[ATAA]_6$	CTTTCTATCAGTATCTATAAAATATTTACAAGAAAAGAGCATT $[ATAA]_6$	5
9269	6	$[ATCA]$ $[ATAA]_5$	CTTTCTATCAGTATCTATAAAATATTTACAAGAAAAGAGCATT $[ATCA]$ $[ATAA]_5$	6
4910	4	$[AAGA]_4$	AAAT $[AAGA]_4$ AAAACTTATGGCCAGTAAGCGTTTCCCTTGCTGGTTACCTTTCTTCAGT CTTTGAGGAATTCATTCGAACACTCTGTCAATCTCAACTGGTTTCTTCA AACTCTAATC	9
4910	5	$[AAGA]_5$	AAAT $[AAGA]_5$ AAAACTTATGGCCAGTAAGCGTTTCCCTTGCTAATTTCCTTTCTTCAGT CTTTGAGGAATTCATTCGAACACTCTGTCAACCTCAACTGGTTTCTTCA AACTCTAATC	2
4910	5.3	$[AAGA]_6$	AAAT $[AAGA]_6$ AAAACTTATGGCCAGTAAGCGTTTCCCTTGCTGGTTACCTT-CTTCAGTCTTTGAGGAATTCATTCGAACACGCTGTCAACCTCAACTGGT TTCTTCAAACTCTAATC	1

(continued)

Locus	Allele call by size	Repeat Motif	MPS allele	Count
4910	10	$[AAGA]_{10}$	AAAT $[AAGA]_{10}$ AAAACTTATGGCCAGTAAGCGTTTCCCTTGTTGGTTACCTTTCTTCAGT CTTTGAGGAATTCATTCGAACACTCTGTCAACCTCAACTGGTTTCTTCA AACTCTAATC	11
5159	3	$[AGAT]_3$	GAAAATGACACAATAACAAGATCACTACAAAACAAACTTTTATG $[AGAT]_3$ACTTACAAATCCCCAT	1
5159	4	$[AGAT]_4$	GAAAATGACACAATAACAAGATCACTACAAAACAAACTTTTATG $[AGAT]_4$ ACTTACAAATCCCCAT	3
5159	4.2	$[AGAT]_4$	[AT]TAAAATGACACAATAACAAGATCACTACAAAACAAACTTTTATG $[AGAT]_4$ ACTTACAAATCCCCAT	4
5159	4.2	[AGAT] [CGAT] [AGAT] [AT] [AGAT]	GAAAATGACACAATAACAAGATCACTACAAAACAAACTTCTATG[AGA T] [CGAT] [AGAT] [AT] [AGAT] ACTTACAAATCCCCAT	2
5159	6	$[AGAT]_6$	GAAAATGACACAATAACAAGATCACTACAAAACAAACTTTTATG $[AGAT]_6$ ACTTACAAATCCCCAT	13
5159	7	$[AGAT]_7$	GAAAATGACACAATAACAAGATCACTACAAAACAAACTTTTATG $[AGAT]_7$ ACTTACAAATCCCCAT	2
ANUCS305	6	$[TGA]$ $[TGG]_5$	TTTGAATTGTGACTATCTTGATGT $[TGA]$ $[TGG]_5$	3
ANUCS305	7	$[TGA]$ $[TGG]_6$	TTTGAATTGTGACTATCTTGATGT $[TGA]$ $[TGG]_6$	1
ANUCS305	8	$[TGA]$ $[TGG]_6$ [GGG]	TTTGAATTGTGACTATCTTGATGT $[TGA]$ $[TGG]_6$ [GGG]	9
ANUCS305	9	$[TGA]$ $[TGG]_7$ [GGG]	TTTGAATTGTGACTATCTTGATGT $[TGA]$ $[TGG]_7$ [GGG]	1
ANUCS305	9	$[TGA]$ $[TGG]_5$ $[TGA]$ $[TGG]_2$	TTTGAATTGTGACTATCTTGATGT $[TGA]$ $[TGG]_5$ $[TGA]$ $[TGG]_2$	1
9043	3	$[TCTT]_3$	TTTTGTG $[TCTT]_3$	6
9043	5	$[TCTT]_5$	TTTTGTG $[TCTT]_5$	6

(continued)

Locus	Allele call by size	Repeat Motif	MPS allele	Count
9043	5	$[TCTT]_3$ [TCCT] [TCTT]	TTTTGTG $[TCTT]_3$ [TCCT] [TCTT]	4
9043	6	$[TCTT]_6$	TTTTGTG $[TCTT]_6$	8
B05-CANN1	7	$[TTG]_7$	TTAGGGTTTTGAGAATTTGGGT $[TTG]_7$ TGGGTTTAAAGAAAGAT	1
B05-CANN1	8	$[TTG]_8$	TTAGGGTTTTGAGAATTTGGGT $[TTG]_8$ TGGGTTTAAAGAAAGAT	13
B05-CANN1	8	[TTG] [GTG] $[TTG]_6$	TTAGGGTTTTGAGAATTTGGGT [TTG] [GTG] $[TTG]_6$ TGGGTTTAAAGAAAGAT	5
B05-CANN1	9	$[TTG]_9$	TTAGGGTTTTGAGAATTTGGGT $[TTG]_9$ TGGGTTTAAAGAAAGAT	7
1528	6	$[ATTA]_6$	GGAATAACTTG $[ATTA]_6$TATTTTATCCAAATAAACAGATTAAGGTAATGTTATTTATTATTACAACTCGCCATCATCAGCCAAGTACTCATGATTGAATAATTTCTCTTAAGCTCAAGTGCTTTAAAAGTGATCTCTCAGTCTCACTGATCTATATAGTAG	5
1528	7	$[ATTA]_7$	GGAATAACTTG $[ATTA]_7$ TATTTTATCCAAATAAACAGATTAAGGTAATGTTATTTATTATTACAACTCGCCATCATCAGCCAAGTACTCATGATTGAATAATTTCTCTTAAGCTCAAGTGCTTTAAAAGTGATCTCTCAGTCTCACTGATCTATATAGTAG	12
1528	7	$[ATTA]_6$	GGAATAACTTG $[ATTA]_6$ TATTTTATCCAAATAAACAGATTAAGGTAATGTTATTTATTATTACAACTCGCCATCATCAGCCAAGTACTCATGATTGAATAATTTCTCTTAAGCTCAAGTGCTTTAAAAGTGATCTCTCAGTCTCACTGATCTATATAGTAG[CTAG]	3

(continued)

Locus	Allele call by size	Repeat Motif	MPS allele	Count
3735	3	$[TATG]_3$	GCA $[TATG]_3$TATA	4
3735	4	$[TATG]_4$	GCA $[TATG]_4$TATA	6
3735	5	$[TATG]_5$	GCA $[TATG]_5$ TATA	5
3735	6	$[TATG]_6$	GCA $[TATG]_6$ TATA	7
3735	7	$[TATG]_7$	GCA $[TATG]_7$ TATA	4
3735	8	$[TATG]_8$	GCA $[TATG]_8$ TATA	1
D02-CANN1	5	$[GTT]_5$	GTA $[GTT]_5$ ATTT	1
D02-CANN1	6	$[GTT]_6$	GTA $[GTT]_6$ ATTT	9
D02-CANN1	6	[ATT] $[GTT]_5$	GTA [ATT] $[GTT]_5$ ATTT	3
D02-CANN1	7	$[GTT]_7$	GTA $[GTT]_7$ ATTT	8
D02-CANN1	8	$[GTT]_8$	GTA $[GTT]_8$ ATTT	4
C11-CANN1	15	[TGG] [TTA] $[TGG]_4$ N48 $[TGA]_5$ N6 $[TGG]_4$	[TGG] [TTA] $[TGG]_4$ TTATGATTAATATGGCTATTATGTTTATGGTGGTTATGGTTGTGATGG $[TGA]_5$ TGGTGT $[TGG]_4$	2
C11-CANN1	16	[TGG] [TTA] $[TGG]_4$ N48 $[TGA]_6$ N6 $[TGG]_4$	[TGG] [TTA] $[TGG]_4$ TTATGATTAATATGGCTATTATGTTTATGGTGGTTATGGTTGTGATGG $[TGA]_6$ TGGTGT $[TGG]_4$	2

(continued)

Locus	Allele call by size	Repeat Motif	MPS allele	Count
H06-CANN2	7	$[AAC]_2$ $[GAC]_5$	AAA $[AAC]_2$ $[GAC]_5$ GCC	4
H06-CANN2	8	$[AAC]_2$ [GAC] [GAT] $[GAC]_4$	AAA $[AAC]_2$ [GAC] [GAT] $[GAC]_4$ GCC	16
H06-CANN2	9	$[AAC]_3$ $[GAT]_2$ $[GAC]_4$	AAA $[AAC]_3$ $[GAT]_2$ $[GAC]_4$ GCC	2

****This is a note for Figs. 5.2. through 5.13.** In the consensus sequences, the forward and reverse primer binding sites are underlined, the nucleotide substitutions are signaled in bold and the indels between brackets. The location of the repeat structure is indicated in the consensus sequence as **[REPEAT]** and its variable structure is individually described for every haplotype in the table. The nucleotide variations or indels of the flanking region are reported in the table in the same order of appearance in the consensus sequence, and they appear organized as pre-SNPs (the SNPs before the repeat region) and post-SNPs (after the repeat region). The sequence data taken from the literature is also referenced in the table. N refers to the total number of alleles found in this study bearing the haplotype described.

AGCAATAATGGAGTGAGTGAACTCTTCTC**[REPEAT]**TGGTTGTGGGAGCCATT
GGGAATCTCAATTTCTTGATCTCT

Allele	[REPEAT]		N	Accession No.	Reference
	TTGTG	CTGTG			
2	2		2	MH136637	This study
4	4		13	KT203577, AY167013.1	[10], [22], [23]
5	4	1			[22]
5	5			AGQN01317157.1	[17], [22]
6	6		8	KT203578	[22], [23]
7	7				[22]

Fig. 5.2. Consensus sequence of the ANUCS501 locus, allele nomenclature, and haplotypes observed in this and previous studies

CCCAAACTACTGTTTGTGCCATTTCACGTGTTCCTTTGTCATTTTCTT**[T]**CTATC
AGTATCTATAAAATATT**W**ACAAGAAAAGAGC**R**TT**[REPEAT]**TGGATCTAACAT
CACGTGCAAGT

Allele	pre-SNPs			[REPEAT]		N	Accession Numer	Reference
	[T]	W	R	ATCA	ATAA			
4	T	A	A		4	1	MH136634	This study
5	T	A	A		5			[7]
5.3	–	T	A		6	3		[7]
6	T	T	A		6	5		[7]
6	T	A	A		6	10	KX668131 – 2 AGQN01009269.1	[7] , [20]
6	T	T	A	1	5	6	MH136633	This study
7	T	A	G		7			[7]

Fig. 5.3. Consensus sequence of the 9269 locus, allele nomenclature, and haplotypes observed in this and previous studies

TCTCCAAAGACATTATTGAACAAAT**[REPEAT]**AAAAC**W**TATGGCCAGTAAGCG
TTTCCCTTG**Y**T**RR**TT**W**CCTTTCTTCAGTCTT[T]GAGGAATTCATTCGAACAC**K**
CTGTCAA**Y**CTCAACTGGTTTCTTCAAACTCTAATCTGAAACCTGGCTCTTGATA
CC

Allele	[REPEAT]			post-SNPs								N	Accession Number	Reference
	AAGA	TAGA	AAAA	W	Y	R	R	W	[T]	K	Y			
4	4			T	C	G	G	A	T	T	T	9	KX668123	[7], [20]
5	5			T	T	G	G	A	T	T	C			[7]
5	5			A	C	A	A	T	T	T	C			[7]
5	5			T	C	A	A	T	T	T	C	2		[7]
5.3	6			T	C	G	G	A	-	G	C	1	MH136628	This study
7	3	1	3	T	C	G	G	A	T	T	T			[7]
10	10			T	T	G	G	A	T	T	C	11	KX668124	[7], [20]
10	10			T	C	G	G	A	T	T	T		AGQN01174910.1	[7]
10	9		1	T	T	G	G	A	T	T	C			[7]
14	14			T	T	G	G	A	T	T	C			[7]
15	15			T	T	G	G	A	T	T	C			[7]

Fig. 5.4. Consensus sequence of the 4910 locus, allele nomenclature, and haplotypes observed in this and previous studies

CCAGAGCTTGTGGATCTCCTGAAGTTTTCCAGTCTCAGAAAGTTTCAAGATTG
CTGTTGACATGTCCACAGCCAGAGGAGAATCTCTTTGAAAAGCCTG**R**TATGGT
AAATTAAAAGTAA[AT]**K**AAAATGACACAATAACAAGATCACTACAAAACAAA
CT**KY**TATG**[REPEAT][ACTTACAAATCCCCAT]**CCACCCCTAGTGAATGGCTGT
CCAATGATCCCGAAATCAGTTTGGTCTGATAGGAACAATTCAACATAAGGAAG
CTCATCTATGATGGCTGCCACACCTCCAGCA**Y**TTGGGCCCAGCCTCAGTGCCC
TTTCGTACT

Allele	pre-SNPs					[REPEAT]					post-SNPs		N	Accession Number	Reference
	R	[AT]	K	K	Y	AGAT	CGAT	AGAT	AT	AGAT	[ACTTACAAATCCCCAT]	Y			
3	A	-	G	T	T	3					ACTTACAAATCCCCAT	T		KX6681271	[7], [20]
4	A	-	G	T	T	4					ACTTACAAATCCCCAT	T	3	KX668126	[7], [20]
4.2	*	AT	T	T	T	4					ACTTACAAATCCCCAT	*	4	MH136629	This study
4.2	A	-	G	T	C	3			1	1	ACTTACAAATCCCCAT	C			[7]
4.2	*	-	G	T	C	1	1	1	1	1	ACTTACAAATCCCCAT	*	2	MH136630	This study
6	G	-	G	G	T	10					–	T		AGQN01195159.1, KX668125	[7], [20]
6	A	-	G	T	T	6					ACTTACAAATCCCCAT	T	13	AGQN01269836.1	[7]
7	*	-	G	T	T	7					ACTTACAAATCCCCAT	*	2	MH136631	This study

Fig. 5.5. Consensus sequence of the 5159 locus, allele nomenclature, and haplotypes observed in this and previous studies
* Indicates that this SNP was not sequenced in this study

AAAGTTGGTCTGAGAAGCAATA**M**TGTTTGCTTTGTAGTATTTGAATTGTGACT
ATCTTGATGT**[REPEAT]**TGATTGTTGGAGGG**R**TTTTCGTATTCGAGGA**Y**TCCAG
CAACGGTGGTGTTGTCGAAAGTTCCTAGG

Allele	pre-SNPs	[REPEAT]					post-SNPs		N	Accession Nb / Reference	Reference
	M	TGA	TGG	TGA	TGG	GGG	R	Y			
4	C		4				G	T	11	KT203572	[22], [23]
6	C	1	5				G	T	6	KT203571	[22], [23]
7	*	1	6				*	*	1	MH136634	This study
8	C	1	6			1	G	T	9		[22]
8	A	1	6			1	G	T		AGQN01198374.1 [21] KT203573	[17], [22], [23]
8	C	2	6				A	T			[22]
9	*	1	7			1	*	*	1	MH136635	This study
9	*	1	5	1	2		*	*	1	MH136636	This study
11	C	1	10				G	C		AY167009.1 [4]	[10], [22]

Fig. 5.6. Consensus sequence of the ANUCS305 locus, allele nomenclature, and haplotypes observed in this and previous studies
* Indicates that this SNP was not sequenced in this study

AAAGCTCGATGTCATCTCTACACTTTGCAAGAAAAGAA**Y**TTCTATATTTACAT
GAGAAGTTACTATGTTTTGTG**[REPEAT]**CTGATTTAAGCATATGAGTAGTTAG
TGAACAA**R**ATAGATAGTCAACCTGGTGTCTGCCCACAAAGGATGAAAGCATG
AATAAGGCATTGAGCA

Allele	pre-SNPs	[REPEAT]				post-SNPs	N	Reference contig	Reference
	Y	TCTT	TCCT	CCTT	TCTT	R			
3	T	3				G	6	KX668128	[7], [20]
3	T	1		1	1	G			[7]
5	T	5				G	6		[7]
5	T	3	1		1		4	KX668129	[20]
6	T	6				A	8	AGQN01009043.1	[7]
7	C	7				A			[7]

Fig. 5.7. Consensus sequence of the 9043 locus, allele nomenclature, and haplotypes observed in this and previous studies

TTGATGGTGGTGAAACGGCGACGTTTGAGGTGGGTAAGAGAAATTGGCGGCG
GAGGAGGAAGAAGAAGGTGGTGGGTATGAAGATGGTGGTG**R**AATTGGGAAA
TGGGTTGTTGAAGAAGAAGAATTATTATTAGGGTTTTGAGAATTTGGGT**[REPE**
AT]TGGGTTTAAAGAAAGATTGCATATCGAAGGTCTGTTGTTGGGGTTGAGATT
GAGATTGGGG

Allele	pre-SNPs	[REPEAT]			N	Accession Nb / Reference	Reference
	R	TTG	GTG	TTG			
7	G	7			1		[22]
8	G	8			13	KT203581, AGQN01328984.1	[17], [22], [23]
8	*	1	1	6	5	MH136638	This study
9	A	9			7	KT203582	[22], [23]
10	G	10					[22]

Fig. 5.8. Consensus sequence of the B05 locus, allele nomenclature, and haplotypes observed in this and previous studies
* Indicates that this SNP was not sequenced in this study

TTGTCTAGTGCCTTTGTCATGCATGTC**W**TACGTAACGGGCGATGGTGGTGGTG
GAA**S**TATGTGGCCTAATT**M**ACTACAGTACT**R**GAA**Y**AACTTG**[REPEAT]**TATTTT
ATCCAAATAAACAGATTAAGGTAATGTTATTTATTAT**[AT]**TACAACTCGCCATCA
TCAGCCAAGTACTCATGATTGAATAATTTCTCTTAAGCTCAAGTGCTTTAAAAG
TGATCTCTCA**[GTCTCA]**CTGATCTATATAGTAG**[CTAG]**TTTAATGGAGCAAATT
TGGTCATCCT

Allele	pre-SNPs					[REPEAT]	post-SNPs			N	Accesion Number	Reference
	W	S	M	R	Y	ATTA	[AT]	[GTCTCA]	[CTAG]			
2	T	G	A	G	C	3	AT	–	–			[7]
6	A	C	A	G	T	6	–	GTCTCA	–			[7]
6	A	G	A	G	T	6	–	GTCTCA	–	5	KX668119, AGQN01001528.1	[7], [20]
7	T	G	A	G	T	7	–	GTCTCA	–	12	KX668120	[7], [20]
7	T	G	C	G	T	6	–	GTCTCA	CTAG	3		[7]

Fig. 5.9. Consensus sequence of the 1528 locus, allele nomenclature, and haplotypes observed in this and previous studies

TGATTCTGTGTTTGTGTGCAATGCA**[REPEAT]**TATAGTGAAAGTTGTTTG**W**AG
TACTAACCTGTGGGTGCGATG

Allele	[REPEAT]	post-SNPs	N	Accesion Number	Reference
	TATG	W			
3	3	A	4	KX668121, AGQN01216044.1	[7], [20]
4	4	T	6		[7]
5	5	T	5		[7]
6	6	T	7		[7]
7	7	T	4	KX668122, AGQN01123735.1	[7], [20]
8	8	T	1		[7]

Fig. 5.10. Consensus sequence of the 3735 locus, allele nomenclature, and haplotypes observed in this and previous studies

GGTTGGGATGTTGTTGTTGTGTCAGAATAGGTTTGTACGTA**[REPEAT]**ATTTGG
GTTCATGAGATAAAGGGTATGCAACCCATCAGGACCTTGGATTTCT

Allele	[REPEAT]		N	Accession Number	Reference
	ATT	GTT			
5		5	1		[22]
6		6	9	AGQN01120802.1, KT203591	[17], [22], [23]
6	1	5	3	MH136640	This study
7		7	8	KT203592	[22], [23]
8		8	4		[22]

Fig. 5.11. Consensus sequence of the D02-CANN1 locus, allele nomenclature, and haplotypes observed in this and previous studies

GTGGTGGTGATGATGATAATGG**[REPEAT]**TGACAATT**R**TGTTTTCGCCATCGTAACCAATTCA

Allele	[REPEAT]							post-SNPs	N	Accession Number	Reference
	TGG	TTA	TGG	N48	TGA	N6	TGG	R			
13	1			TTATGATTAATATGGCTATTATGTT TATGGTGGTTATGGTTGTGATGG	8	TGGTGT	4	G	7	KT203583	[22], [23]
14	1				8		5	G	7		[22]
14	1				8		5	A			[22]
14	1				9		4	G	5	KT203584	[22], [23]
14			1		8		5	*	1	MH136639	This study
15	1	1	4		5		4	G	2	KT203585	[22], [23]
15	1				9		5	G		AGQN01087310.1, AGQN01053545.1	[22]
16	1	1	4		6		4	G	2		[22]
17	1	1	4		7		4	G			[22]
18	1	1	4		8		4	G			[22]
18	2	1	4		7		4	G			[22]

Fig. 5.12. Consensus sequence of the C11-CANN1 locus, allele nomenclature, and haplotypes observed in this and previous studies
* Indicates that this SNP was not sequenced in this study

TGGTTTCAGTGGTCCTCTCGAAATGAGTAAAAACAATCACAACAGTAAA**[REP EAT]**GCCTACGTTGAGGTCACTCTGGACATCCACGACGA**Y**ACAGTGGCCGTTC A**Y**AGCGTCCAAGC**R**AC**H**ACGACAGGGAACGAGGACCCTGAGCTTGCTCTGC TCACCAAG**M**AGACTCTTCACGACATCAACAAG**Y**C**YK**CTAAATCCTCTTCCTT CGGCTCATCTCGTTTCCGTACAGCTTCATCTCGTGTCATCACTCACGT

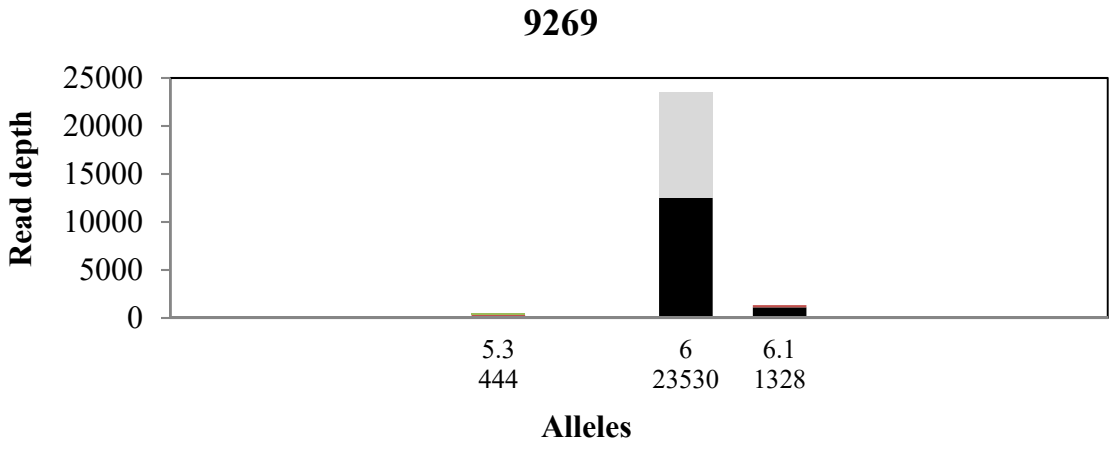
Allele	[REPEAT]					post-SNPs								N	Accession Number	Reference
	AAC	GAC	GAT	AAT	GAC	Y	Y	R	H	M	Y	Y	K			
7	2	5				C	T	A	C	A	C	T	T	4	KT203596 AGQN01201155.1	[17], [22], [23]
8	2	6				C	T	A	C	C	T	C	T			[22]
8	2	1	1		4	C	C	A	T	A	T	C	T	16	KT203597	[22], [23]
9	2	1	1	1	4	T	C	A	A	A	T	C	G			[22]
9	3		1	1	4	T	C	A	A	A	T	C	G			[22]
9	3		2		4	T	C	G	C	A	T	C	T	2	AGQN01141370.1	[17], [22], [23]

Fig. 5.13. Consensus sequence of the H06-CANN2 locus, allele nomenclature, and haplotypes observed in this and previous studies

Concordance

For each sample, allele calls identified by MPS were compared to those determined by CE typing. As MPS technology increasingly reveals the complexity of STR motifs and SNPs within the flanking regions, there is a pressing need to establish standards for defining and reporting STRs by MPS. The International Society for Forensic Genetics (ISFG) has proposed nomenclature requirements for massively parallel sequencing of STRs [30, 31]. For MPS typing of human STRs this relies on aligning string sequences to a reference genome. Because there was no reference genome for *Cannabis sativa* at the time of this study, MPS alleles were reported using the string sequence and nominal allele number extracted with STRait Razor. Additionally, MPS technology must be backwards compatible with STR typing by CE to ensure concordance. For this study, the nomenclature previously established by Valverde et al. [7, 22] and Houston et al. [20] was used. Complete concordance was observed between the two methods when comparing the length-based

alleles extracted by STRait Razor and the allele number observed by CE. The clonal samples (Chile 47 and Chile 48) were also determined to be identical by sequence. Interestingly, there were 13 instances where loci previously believed to be homozygous were determined to be heterozygote by sequence (Fig. 5.14.).



Nominal Allele	String sequence	Read depth
6	CTTTCTATCAGTATCTATAAAATATTTACAAGAAAAGAGCATT [ATCA] $[ATAA]_5$	12580
6	CTTTCTATCAGTATCTATAAAATATTTACAAGAAAAGAGCATT $[ATAA]_6$	10950

Fig. 5.14. Example of previously classified homozygote peak determined to be heterozygous by sequence. Histogram visualization isoalleles is shown as well as sequence variation between the two "6" alleles

Coverage

Average read depth was calculated for each locus across 16 samples (Fig. 5.15.). Homozygote peaks were regarded as two peaks when calculating average read depth. On average, 13,000x coverage was observed at each locus with coverage ranging from 1,011x for ANUCS305 and 41,378x for 9043. A generally balanced locus-to-locus read depth was detected between seven of the loci (9269, 4910, 5159, 1528, 3735, C11-CANN1, D02-CAN1). A relatively low read depth was observed for three loci (ANUCS501, ANUCS305, B05-CANN1) while a high read depth was noted for 9043 and H06-CANN2.

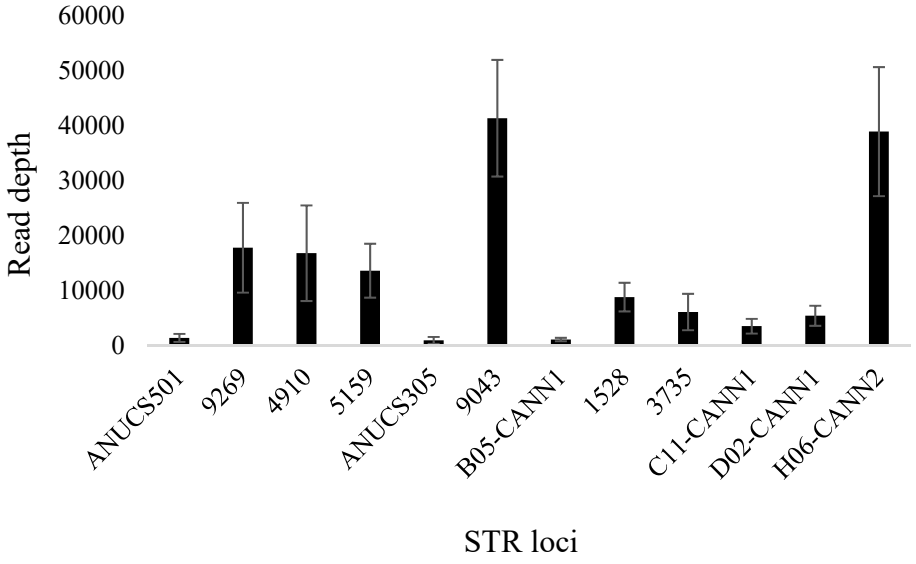


Fig. 5.15. Average read depth across all loci for 16 samples with 5 ng of input DNA. The error bars represent standard deviation

There are some potential explanations to explain this disparity. Strand bias, a common phenomenon seen even in commercial kits during sequencing [32-34] was also observed for ANUCS305, 5159, 4910, and B05-CANN1. Strand bias and allele specific bias across the four loci can be visualized in Figs. 5.16. – 5.19. For ANUCS305, negative strand bias was observed for all alleles except one (allele eight). Although negative strand bias was observed, only the more balanced forward strand was analyzed for ANUCS305 to accommodate the allele specific bias at allele eight. For 5159, positive strand bias was only observed in one allele (seven). Although only observed in one allele, data analysis settings had to be adjusted to ensure that sister heterozygote peaks were called in the case that the seven allele was one of the sister peaks. For 4910, negative strand bias was observed for one allele (ten). In this case, no reads were observed in the forward orientation. This observation may be due to secondary structure in the forward strand of this allele inhibiting single-base elongation [34]. For B05, high negative strand bias was observed in all alleles; approximately eight times more reads were observed on the reverse strand as compared to the reverse strand. To ensure accurate allele designation, only the reverse strand was analyzed for 4910 and B05 while only the forward strand was analyzed for 5159 and ANUCS305.

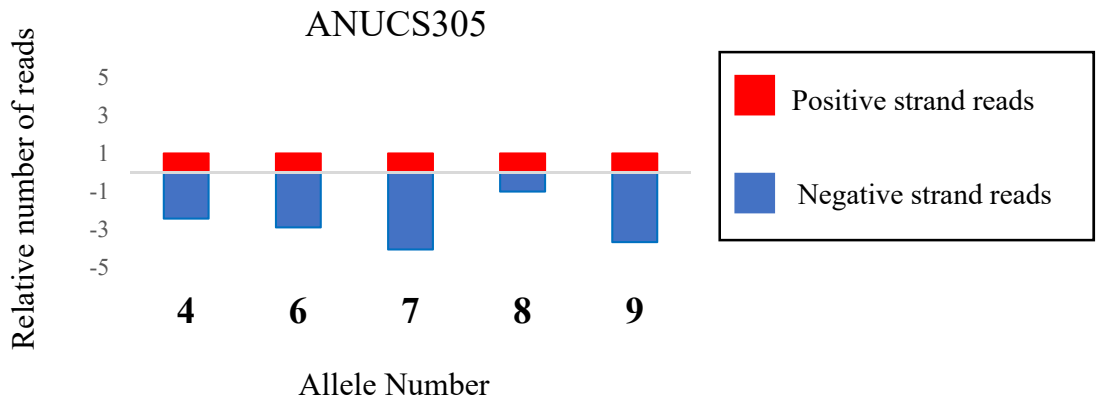


Fig. 5.16. Strand bias for ANUCS305. The bar chart represents the average relative percentage of reads in each direction based on the allele

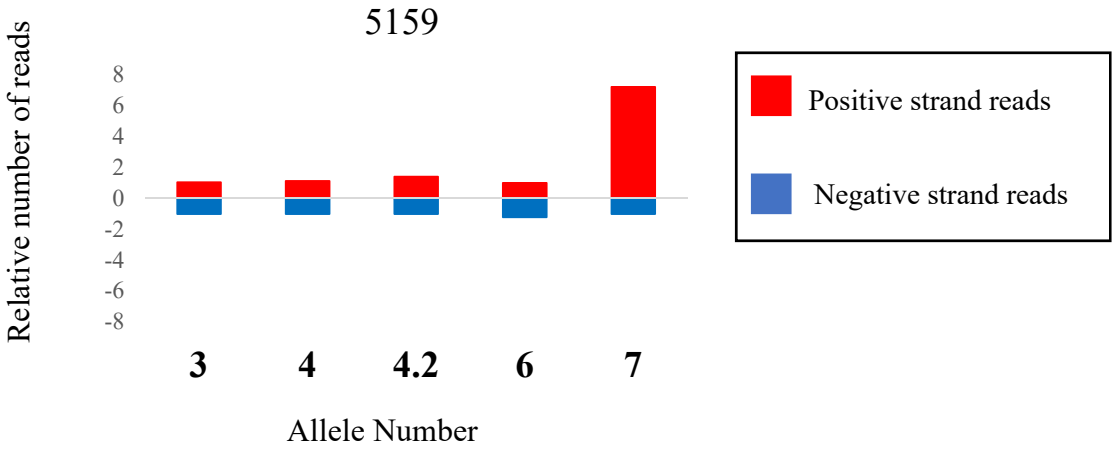


Fig. 5.17. Strand bias for 5159. The bar chart represents the average relative percentage of reads in each direction based on the allele

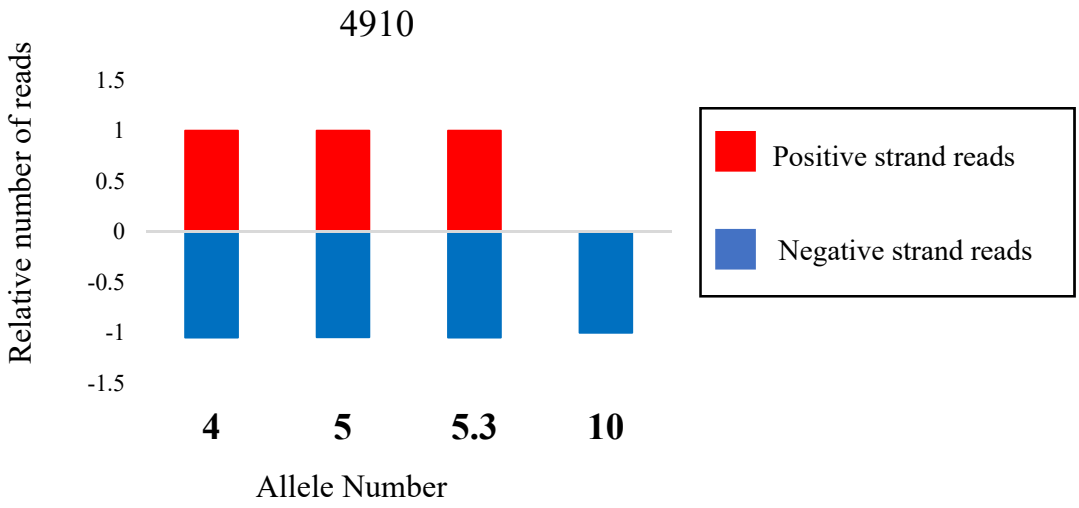


Fig. 5.18. Strand bias for 4910. The bar chart represents the average relative percentage of reads in each direction based on the allele

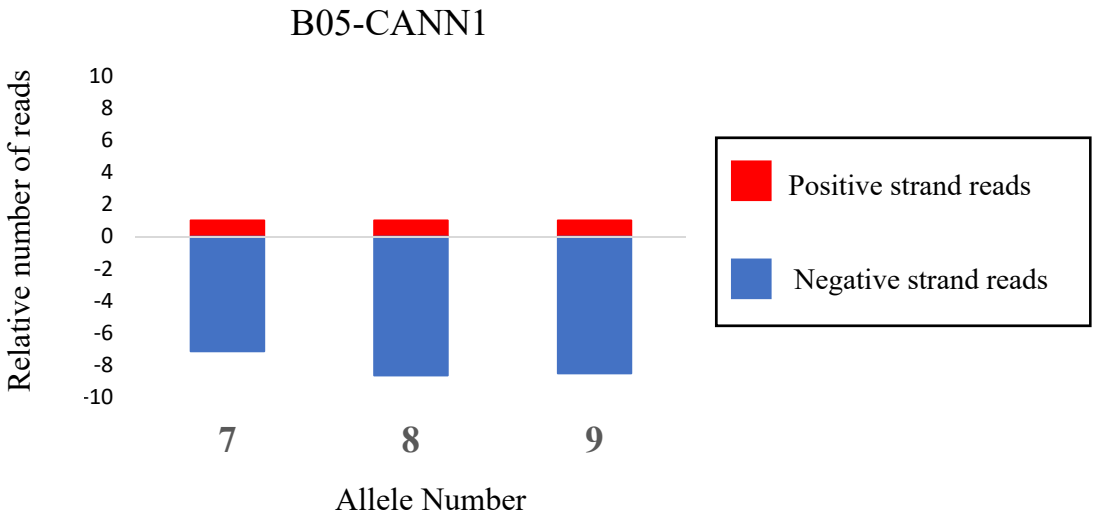


Fig. 5.19. Strand bias for B05-CANN1. The bar chart represents the average relative percentage of reads in each direction based on the allele

It can be hypothesized that ANUCS501 performed poorly due to its small amplicon size (79 – 95 bp) and may have been partially removed in the size selection process during library preparation. Primer concentrations may also be adjusted; however, amplification efficiency of ANUCS501 based on CE data is not an issue. Therefore, future designs would benefit from redesigning the primer sequences to increase amplicon length for this locus. In turn, 9043 and H06-CANN2 performed exceptionally well due their amplicon size (180 – 275 bp). While the library preparation kit used can perform size selection across a wide range of amplicon sizes, 200 – 300 bp was indicated as being the ideal amplicon size. The disparity in read depth indicates that the multiplex may not be completely optimized for MPS. Further studies will need to be performed to assess the minimum read depth required for accurate allele calling and in turn the maximum number of samples that can be sequenced simultaneously.

For heterozygotes, average heterozygote balance was calculated for each locus across 16 samples (Fig. 5.20.). The average heterozygote balance was greater than 0.4 across all loci with an overall average of 0.73 ± 0.16. Two loci, 4910 and ANUCS305, had relatively lower heterozygote balance than other loci (0.41 ± 0.07 and 0.42 ± 0.21, respectively). This may be due to the wide range of allelic variation observed within these loci with the larger alleles consistently yielding lower coverage during MPS (Figs. 5.21. – 5.22.)

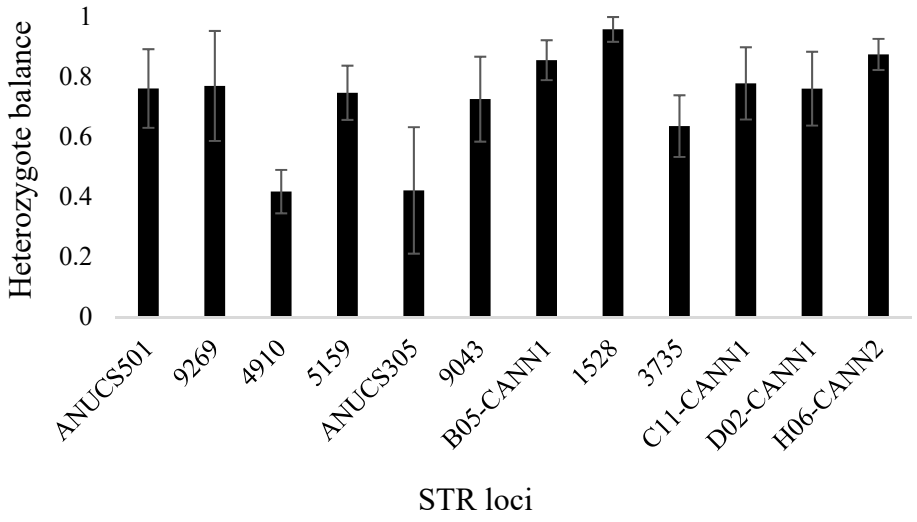


Fig. 5.20. Heterozygote balance across all loci for 16 samples with 5 ng of input DNA. The error bars represent standard deviation

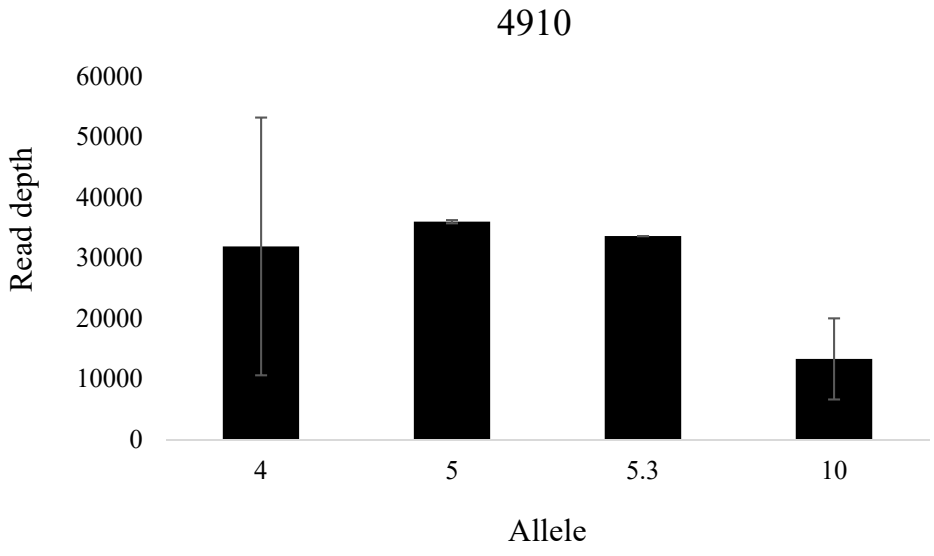


Fig. 5.21. Relative read depth across alleles at the 4910 locus. The error bars represent standard deviation

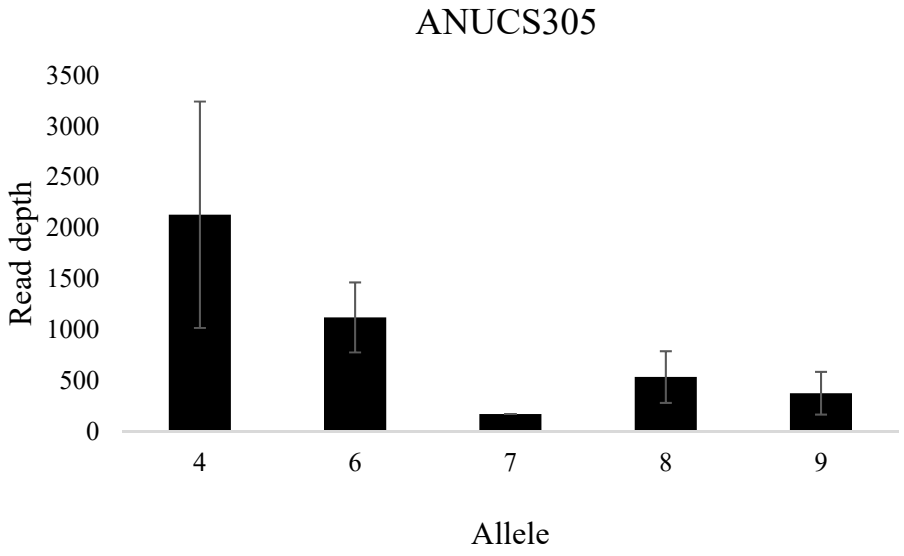


Fig. 5.22. Relative read depth across alleles at the ANUCS305 locus. The error bars represent standard deviation

Noise analysis

Noise was observed at all loci with an overall average of 0.15 ± 0.10 (Fig. 5.23.). Ten loci had average noise percentages less than 20%. Two loci, 4910 and 1528, had an average noise percentage > 30%, at 32.6% and 35.0% respectively. Most noise at these loci was due to sequence error in homopolymeric regions (insertions, deletions, or base substitutions). This is a well-documented problem in all sequencing platforms, especially semiconductor sequencing platforms such as the one used in this study [35-37]. Additionally, both loci are highly variable within the flanking region (Figs 5.4. and 5.9.) making bioinformatic sequence extraction difficult. Even with a high percentage of noise, the true alleles were readily identified as the noise was distributed across multiple locations.

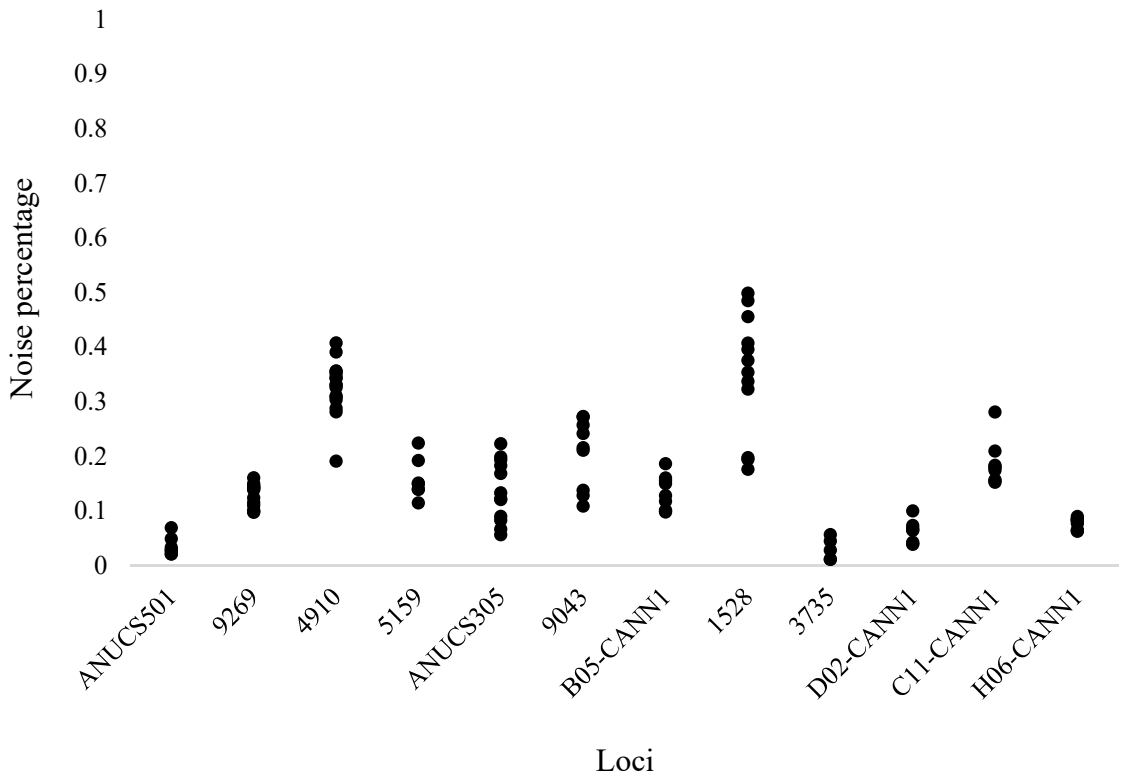


Fig. 5.23. Noise percentages of STRs from 16 *Cannabis* samples

Conclusions

This study investigates the sequence variation of 12 autosomal STR loci in 16 *Cannabis* samples from three countries. This study also provides a simple workflow for STR sequencing using the Ion™ S5 that allows for the easy integration of custom non-human PCR multiplexes into MPS workflows. Given the successful proof of concept, future research may include expanding the number of loci, redesigning PCR primers where possible, sensitivity studies, and a larger, more variable sample database. *Cannabis* genotyping would benefit from the addition of more loci, and MPS is an ideal platform for expanding and assessing new loci as well as updating nomenclature and allele frequencies. Indeed, a single multiplex could be designed to sequence hundreds of *Cannabis* specific

STRs and/or SNPs across hundreds of samples simultaneously. Moreover, an ideal *Cannabis* MPS panel would include autosomal STRs, chloroplast, and mitochondrial markers to interrogate identification and biogeographical origin.

Role of funding

This study was partially funded by a Graduate Research Fellowship Award #2015-R2-CX (National Institute of Justice, Office of Justice Programs, U.S. Department of Justice). The opinions, findings, conclusions, or recommendations expressed in this presentation are those of the authors and do not necessarily reflect those of the National Institute of Justice.

References

1. Guo F, Yu J, Zhang L, 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. https://doi.org/10.1016/j.fsigen.2017.09.003

2. Seo SB, King JL, Warshauer DH, Davis CP, Ge J, Budowle B (2013) Single nucleotide polymorphism typing with massively parallel sequencing for human identification. Int J Leg Med 127:1079-1086. https://doi.org/10.1007/s00414-013-0879-7

3. Massart S, Olmos A, Jijakli H, Candresse T (2014) Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188:90-96. https://doi.org/10.1016/j.virusres.2014.03.029

4. Vasan N, Yelensky R, Wang K, Moulder S, Dzimitrowicz H, Avritscher R, Wang B, Wu Y, Cronin MT, Palmer G, Symmans WF, Miller VA, Stephens P, Pusztai L (2014) A targeted next-generation sequencing assay detects a high frequency of therapeutically targetable alterations in primary and metastatic breast cancers: Implications for clinical practice. Oncologist 19:453-458. https://doi.org/10.1634/theoncologist.2013-0377

5. Logares R, Audic S, Bass D et al (2014) Patterns of rare and abundant marine microbial eukaryotes. Curr Biol 24:813-821. https://doi.org/10.1016/j.cub.2014.02.050

6. Howard C, Gilmore S, Robertson J, Peakall R (2009) A *Cannabis sativa* STR genotype database for Australian seizures: forensic applications and limitations. J Forensic Sci 54:556-563. https://doi.org/10.1111/j.1556-4029.2009.01014.x

7. Valverde L, Lischka C, Scheiper S et al (2014) Characterization of 15 STR *Cannabis* loci: Nomenclature proposal and SNPSTR haplotypes. Forensic Sci Int Genet 9:61-65. https://doi.org/10.1016/j.fsigen.2013.11.001

8. Houston R, Birck M, LaRue B, Hughes-Stamm S, Gangitano D (2018) Nuclear, chloroplast, and mitochondrial data of a US *Cannabis* DNA database. Int J of Leg Med. https://doi.org/10.1007/s00414-018-1798-4

9. Dufresnes C, Jan C, Bienert F, Goudet J, Fumagalli L (2017) Broad-scale genetic diversity of *Cannabis* for forensic applications. PLoS ONE 12:e0170522. https://doi.org/10.1371/journal.pone.0170522

10. Gilmore S, Peakall R (2003) Isolation of microsatellite markers in *Cannabis sativa* L. (marijuana). Mol Ecol 3:105-107. https://doi.org/10.1046/j.1471-8286.2003.00367.x

11. Lazaruk K, Walsh PS, Oaks F, Oaks F, Gilbert D, Rosenblum BB, Menchen S, Scheibler D, Wenz HM, Holt C, Wallin J (1998) Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument. Electrophoresis 191:86-93. https://doi.org/10.1002/elps.1150190116

12. Sawler J, Stout JM, Gardner KM et al (2015) The genetic structure of marijuana and hemp. PLoS ONE 10:e0133292. https://doi.org/10.1371/journal.pone.0133292

13. Soorni A, Fatahi R, Haak DC, Salami SA, Bombarely A (2017) Assessment of genetic diversity and population structure in iranian *Cannabis* germplasm. Sci Rep 7:15668. https://doi.org/10.1038/s41598-017-15816-5

14. Zhao X, Li H, Wang Z, Ma K, Cao Y, Liu W (2016) Massively parallel sequencing of 10 autosomal STRs in Chinese using the Ion Torrent personal genome machine (PGM). Forensic Sci Int Genet 25:34-38. https://doi.org/10.1016/j.fsigen.2016.07.014

15. Zhao X, Ma K, Li H, Cao Y, Liu W, Zhou H, Ping Y (2015) Multiplex Y-STRs analysis using the Ion Torrent personal genome machine (PGM). Forensic Sci Int Genet 19:192-196. https://doi.org/10.1016/j.fsigen.2015.06.012

16. Kwon SY, Lee HY, Kim EH, Lee EY, Shin KJ (2016) Investigation into the sequence structure of 23 Y chromosomal STR loci using massively parallel sequencing. Forensic Sci Int Genet 25:132-141. https://doi.org/10.1016/j.fsigen.2016.08.010

17. van Bakel H, Stout J, Cote A, Tallon C, Sharpe A, Hughes T, Page J (2011) The draft genome and transcriptome of *Cannabis sativa*. Genome Biol 12:R102. https://doi.org/10.1186/gb-2011-12-10-r102

18. Warshauer DH, Lin D, Hari K, Jain R, David C, LaRue B, King JL, Budowle B (2013) Strait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci Int Genet 7:409-417. https://doi.org/10.1016/j.fsigen.2013.04.005

19. Woerner AE, King JL, Budowle B (2017) Fast STR allele identification with STRait Razor 3.0. Forensic Sci Int Genet 30:18-23. https://doi.org/10.1016/j.fsigen.2017.05.008

20. Houston R, Birck M, Hughes-Stamm S, Gangitano D (2017) Developmental and internal validation of a novel 13 loci STR multiplex method for *Cannabis sativa* DNA profiling. Leg Med (Tokyo, Japan) 26:33-40. https://doi.org/10.1016/j.legalmed.2017.03.001

21. Qubit® dsDNA HS assay kits. (2015). Thermo Fisher Scientific, South San Francisco, CA

22. Valverde L, Lischka C, Erlemann S, deMeijer E, de Pancorbo M, Pfeiffer H, Kohnemann S (2014) Nomenclature proposal and SNPSTR haplotypes for 7 new *Cannabis sativa* L. STR loci. Forensic Sci Int Genet 13:185-186. http://doi.org/10.1016/j.fsigen.2014.08.002

23. Houston R, Birck M, Hughes-Stamm S, Gangitano D (2016) Evaluation of a 13-loci STR multiplex system for *Cannabis sativa* genetic identification. Int J Leg Med 130:635-647. https://doi.org/10.1007/s00414-015-1296-x

24. Minelute® handbook. (2008). Qiagen, Hilden, Germany

25. Agilent DNA 1000 kit guide. (2016). Agilent Technologies, Santa Clara, CA

26. Prepare amplicon libraries without fragmentation using the ion plus fragment library kit. (2016). Thermo Fisher Scientific, South San Francsico, CA

27. Ion library taqman® quantitation kit user guide. (2017). Thermo Fisher Scientific, South San Francsico, CA

28. Zeng X, King JL, Budowle B (2017) Investigation of the STR loci noise distributions of PowerSeq™ Auto System. Croat Med J 58:214-221. https://doi.org/10.3325/cmj.2017.58.214

29. Elwick K, Zeng X, King J, Budowle B, Hughes-Stamm S (2017) Comparative tolerance of two massively parallel sequencing systems to common PCR inhibitors. Int J Leg Med. https://doi.org/10.1007/s00414-017-1693-426

30. Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmão L, Hares DR, Irwin JA, King JL, Knijff Pd, Morling N, Prinz M, Schneider PM, Neste CV, Willuwit S, Phillips C (2016) Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the international society for forensic genetics (ISFG) on minimal nomenclature requirements. Forensic Sci Int Genet 22:54-63. https://doi.org/10.1016/j.fsigen.2016.01.009

31. Phillips C, Gettings KB, King JL, Ballard D, Bdner M, Borsuk L, Parson W (2018) "The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR sequence guide. Forensic Sci Int Genet 34:262-169. https://doi.org/10.1016/j.fsigen.2018.02.017

32. Churchill JD, Chang J, Ge J, Rajagopalan N, Wootton SC, Chang CW, Lagacé R, Liao W, King JL, Budowle B (2015) Blind study evaluation illustrates utility of the Ion PGM™ system for use in human identity DNA typing. Croat Med J 56:218-229. https://doi.org/10.3325/cmj.2015.56.218

33. Fordyce SL, Mogensen HS, Borsting C, Lagacé RE, Chang CW, Rajagopalan N, Morling N (2015) Second-generation sequencing of forensic strs using the Ion

Torrent HID STR 10-plex and the Ion PGM. Forensic Sci Int Genet 14:132-140. https://doi.org/10.1016/j.fsigen.2014.09.020

34. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S (2011) Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39:e90. https://doi.org/10.1093/nar/gkr344

35. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: Characterising errors in Ion Torrent PGM data. PLoS Comput Biol 9:e10003031. https://doi.org/doi:101371/journalpcbi1003031

36. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:241. https://doi.org/101186/1471-2164-13-341

37. Guo Y, Li J, Li C-I, Long J, Samuels DC, Shyr Y (2012) The effect of strand bias in Illumina short-read sequencing data. BMC Genomics 13:666. https://doi.org/10.1186/1471-2164-13-666

CHAPTER VI

CONCLUSIONS

Cannabis sativa L. (marijuana) is the most commonly used illicit controlled substance in the United States. As a result, it is highly trafficked to and within the United States by organized crime syndicates. Additionally, law enforcement faces a unique challenge in tracking and preventing flow of the legal marijuana to states where it is still illegal. Moreover, significant illegal *C. sativa* traffic from Mexico exists at the US border. The development of a validated method using molecular techniques such as short tandem repeats (STRs) for the genetic identification of *C. sativa* may aid in the individualization and origin determination of *Cannabis* samples as well as serve as an intelligence tool to link *Cannabis* cases (e.g., illegal traffic at the US-Mexico border). To date, no DNA typing method for *Cannabis* using short tandem repeat (STR) markers following ISFG or SWGDAM recommendations has been reported (e.g., use of sequenced allelic ladder, use of tetra-nucleotide STR markers). This project explores the utility of forensic genetic tools for the identification and origin determination of *C. sativa*. Results provide the forensic genetic community with a comprehensive genetic tool (STR, cpDNA, mtDNA, and MPS) that allows for the individualization of *Cannabis* samples, the association of different cases as well as origin determination of samples for forensic and intelligence purposes.

Prior to downstream STR typing, a real-time PCR method for *Cannabis* DNA quantitation was developed and validated according to SWGDAM guidelines. A previously described 15-loci STR multiplex was evaluated and modified. In addition, an allelic ladder was developed for accurate genotyping. The system was determined to be specific for *Cannabis*, and its sensitivity was as low as 250 pg. A reference *Cannabis*

population database (*N*=97 samples) with associated allele frequencies for forensic purposes was also established. Results revealed that the multiplex was not suitable for forensic testing due to high heterozygote peak imbalance in some markers, high stutter peaks in dinucleotide markers, inter-loci peak imbalance, and the presence of null alleles in four of the loci.

Based on the previously evaluated multiplex, a novel 13-locus multiplex was designed. Poorly performing STR markers were replaced with more recently discovered tetranucleotide markers, and a more comprehensive strategy for multiplex STR design and optimization was implemented. Both developmental and internal validation studies were performed following ISFG/SWGDAM guidelines. STR success rates were improved when compared to the previous multiplex (100% vs 64%). Indeed, high quality DNA profiles were generated with template input as low as 0.13 ng. Low average stutter was observed across all loci (0.006 – 0.052) with the maximum stutter upper range estimated at the C11 locus (0.166). Additionally, high mean PHR were observed across all loci (0.689 – 0.895). Results support a 13-locus *Cannabis* STR multiplex system for forensic DNA profiling that could approach the robustness of standard STR systems used for HID as the multiplex yielded high-quality STR profiles comparable to commercial HID systems. Given the robustness of this assay, this technology may assist the forensic community as the demand for *Cannabis* profiling either for genetic identification or intelligence purposes increases. However, appropriate data interpretation guidelines should be established via internal validation studies prior to implementation.

To test the robustness and validity of the technique, over 500 *Cannabis* samples from four distinct sources were obtained: US-Mexico border (*N*=21 seizures), Chile,

Brazil, and hemp. Samples were genotyped using both autosomal and organelle DNA genotyping techniques. For autosomal typing, the previously validated real-time PCR quantitation and 13-locus STR method were utilized. For organelle typing, a novel real-time PCR quantification method was developed and validated to calculate the concentration of cpDNA in *Cannabis* samples prior to downstream typing. Organelle typing was performed by modifying and optimizing a previously reported system to genotype five chloroplast and two mitochondrial markers. Two novel assays were developed: a homopolymeric STR pentaplex and a SNP triplex with one marker (Cscp001) shared by both assays as a quality control measure. Initial phylogenetic and case-to-case comparisons revealed a larger homogenous sub-population consisting of nine seizures (N=157 samples). These samples formed a reference population that was used to represent a homogenous population from the US-Mexico border. Based on the genotypes obtained, phylogenetic analysis was assessed among the US-Mexico reference population, Brazil, Chile, and hemp samples. Population sub-structure was initially evaluated using a Neighbor Joining method followed by parsimony analysis. To further examine population sub-structure, *STRUCTURE* software was used to evaluate the Bayesian clustering of genotypes from the four populations, and finally the individual genotypes were visualized using PCA. Both autosomal and organelle markers could elucidate population sub-structure and may be suitable for categorizing seized *Cannabis* samples. All phylogenetic methods were able to clearly distinguish marijuana from hemp. Interestingly, organelle genotyping revealed a unique haplotype associated with fiber-type samples. Although this population study would benefit from a wider range of samples, the results reveal the applicability of genotyping both autosomal and organelle DNA for *Cannabis* samples.

Additionally, this study presents, for the first time, a database of US *Cannabis* samples consisting of nuclear, chloroplast, and mitochondrial DNA.

Lastly, as a proof of concept the previously validated STR method was integrated into a MPS pipeline. Importantly, a simple workflow for STR sequencing using the Ion™ S5 was established allowing for the easy integration of custom non-human PCR multiplexes into MPS platforms. For sequencing analysis and bioinformatic processing, a custom configuration file was designed for STRait Razor v3 to parse and extract STR sequence data. Sixteen samples were processed using the designed MPS pipeline, and results revealed full concordance for the size-based STR between the MPS and CE platforms. Interestingly, intra-repeat variation was observed eight loci where the nominal or size-based allele was identical, but sequence variances were discovered in the flanking region. Given the successful proof of concept and increased power of discrimination observed with MPS, future research would benefit from expanding the number of loci as well as including chloroplast and mitochondrial markers. As MPS is an ideal platform for expanding and assessing new loci, a single MPS multiplex could be designed to sequence hundreds of *Cannabis*-specific autosomal and organelle markers (STRs and/or SNPs) across hundreds of samples simultaneously.

In summary, the techniques and results of this research provide the forensic DNA community with a comprehensive genetic tool (STR, cpDNA, mtDNA, and MPS) that allows for the individualization of cannabis samples and the association of different cases for forensic and intelligence purposes. Given the ever-changing legal environment surrounding *Cannabis*, the methods and findings from this research have the potential to expand into fields beyond forensic science, including medicine and commercial industry.

REFERENCES

Abi Prism® Snapshot™ Multiplex Kit Protocol 4323357b. (2010) Thermo Fisher Scientific, South San Francisco, CA

Adams IB, Martin BR (1996) *Cannabis*: pharmacology and toxicology in animals and humans. Addiction 91:1585-1614

Agilent DNA 1000 kit guide. (2016). Agilent Technologies, Santa Clara, CA

Ainsworth C (2000) Boys and girls come out to play: the molecular biology of dioecious plants. Ann Bot 86:211-221. https://doi.org/10.1006/anbo.2000.1201

Aldrich MR (1977) Tantric *Cannabis* use in India. J of Psychedelic Drugs 9:227-233. https://doi.org/10.1080/02791072.1977.10472053

Alghanim HJ, Almirall JR (2003) Development of microsatellite markers in *Cannabis sativa* for DNA typing and genetic relatedness analyses. Anal Bioanal Chem 376:1225-1233. https://doi.org/10.1007/s00216-003-1984-0

Allgeier L, Hemenway J, Shirley N, LaNier T, Coyle HM (2011) Field testing of collection cards for *Cannabis sativa* samples with a single hexanucleotide DNA marker. J Forensic Sci 56:1245-1249. https://doi.org/10.1111/j.1556-4029.2011.01818.x

Anderson P (2006) Global use of alcohol, drugs and tobacco. Drug Alcohol Rev 25:489-502

Andre CM, Hausman JF, Guerriero G (2016) *Cannabis sativa*: the plant of the thousand and one molecules. Front Plant Sci 7:19. https://doi.org/10.3389/fpls.2016.00019

Applications to become registered under the controlled substances act to manufacture marijuana to supply researchers in the United States. Policy statement (2016) Federal register 81:53846-53848

Baechtel FS, Smerick JB, Presley KW, Budowle B (1993) Multigenerational amplification of a reference ladder for alleles at locus D1S80. J Forensic Sci 38:1176-1182

Bafeel S, Arif I, A Bakir M, Khan H, H Al Farhan A, Al-Homaidan A, Ahamed A, Thomas J (2011) Comparative evaluation of PCR success with universal primers of maturase k (matK) and ribulose-1, 5-bisphosphate carboxylase oxygenase large subunit (rbcL) for barcoding of some arid plants. Plant Omics J 4:195-198

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84-91. https://doi.org/10.1104/pp.102.019422

Bell CD, Soltis DE, Soltis PS (2010) The age and diversification of the angiosperms re-revisited. Am J Bot 97:1296-1303. https://doi.org/10.3732/ajb.0900346

Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53. https://doi.org/10.1038/nature07517

Berger B, Berger C, Hecht W, Hellmann A, Rohleder U, Schleenbecker U, Parson W (2014) Validation of two canine STR multiplex-assays following the isfg recommendations for non-human DNA analysis. Forensic Sci Int Genet 8:90-100. https://doi.org/10.1016/j.fsigen.2013.07.002

Bigdye® Direct Cycle Sequencing Kit. (2011) Thermo Fisher Scientific, South San Francisco, CA

Bigdye™ Terminator v3.1 Cycle Sequencing Kit. (2016) Thermo Fisher Scientific, South San Francisco, CA

Bock JH, Norris DO, Forensic Plant Science, first ed. Elsevier Academic Press, London

Bócsa I, Karus M (1998) The cultivation of hemp: botany, varieties, cultivation and harvesting. Hemptech, Sebastopol, CA

Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: Characterising errors in Ion Torrent PGM data. PLoS Comput Biol 9:e10003031. https://doi.org/doi:101371/journalpcbi1003031

Brenneisen R, elSohly MA (1988) Chromatographic and spectroscopic profiles of *Cannabis* of different origins: Part i. J Forensic Sci 33:1385-1404

Bryant VM, Jones GD (2006) Forensic palynology: current status of a rarely used technique in the United States of America. Forensic Sci Int 163:183-197. https://doi.org/10.1016/j.forsciint.2005.11.021

Budowle B, Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM (1999) Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. J. Forensic Sci. 44:1277-1286

Butler JM (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. J Forensic Sci 51:253-265. https://doi.org/10.1111/j.1556-4029.2006.00046.x

Butler JM (2005) Forensic DNA typing: Biology, technology, and genetics of STR markers. Elseiver Academic Press, New York

Carliner H, Brown QL, Sarvet AL, Hasin DS (2017) *Cannabis* use, attitudes, and legal status in the U.S.: a review. Prev Med 104:13-23. https://doi.org/10.1016/j.ypmed.2017.07.008

Cascini F, Passerotti S, Martello S (2012) A real-time PCR assay for the relative quantification of the tetrahydrocannabinolic acid (THCA) synthase gene in herbal *Cannabis* samples. Forensic Sci Int 217:134-138. https://doi.org/10.1016/j.forsciint.2011.10.041

Center for Behavioral Health Statistics and Quality (2014) Results from the 2013 National Survey on Drug Use and Health: summary of national findings. U.S. Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. http://www.samhsa.gov/data/sites/default/files/NSDUHresultsPDFWHTML2013/Web/NSDUHresults2013.pdf. Accessed April 29 2015

Chandra S, Lata H, Khan IA, Elsohly MA (2008) Photosynthetic response of *Cannabis sativa* L. To variations in photosynthetic photon flux densities, temperature and CO2 conditions. Physiol and Mol Biol of Plants 14:299-306. https://doi.org/10.1007/s12298-008-0027-x

Chang KC (1963) The archaeology of ancient China. Yale University Press, New Haven, CT

Churchill JD, Chang J, Ge J, Rajagopalan N, Wootton SC, Chang CW, Lagacé R, Liao W, King JL, Budowle B (2015) Blind study evaluation illustrates utility of the Ion PGM™ system for use in human identity DNA typing. Croat Med J 56:218-229. https://doi.org/10.3325/cmj.2015.56.218

Cirovic N, Kecmanovic M, Keckarevic D, Keckarevic Markovic M (2017) Differentiation of *Cannabis* subspecies by THCA synthase gene analysis using RFLP. J of Forensic and Leg Med 51:81-84. https://doi.org/10.1016/j.jflm.2017.07.015

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in escherichia coli. Nature 274:775. https://doi.org/10.1038/274775a0

Datwyler SL, Weiblen GD (2006) Genetic variation in hemp and marijuana (*Cannabis sativa* L.) according to amplified fragment length polymorphisms. J Forensic Sci 51:371-375. https://doi.org/10.1111/j.1556-4029.2006.00061.x

Daud Khaled AK, Neilan BA, Henriksson A, Conway PL (1997) Identification and phylogenetic analysis of lactobacillus using multiplex RAPD-PCR. FEMS Microbiol Lett 153:191-197

de Meijer EPM, Bagatta M, Carboni A, Crucitti P, Moliterni VMC, Ranalli P, Mandolino G (2003) The inheritance of chemical phenotype in *Cannabis sativa* L. Genetics 163:335-346

Demesure B, Sodzi N, Petit RJ (1995) A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Mol Ecol 4:129-131

Dias VH, Ribeiro AS, Mello IC, Silva R, Sabino BD, Garrido RG, Seldin L, Moura-Neto RS (2015) Genetic identification of *Cannabis sativa* using chloroplast trnL-F gene. Forensic Sci Int Genet 14:201-202. https://doi.org/10.1016/j.fsigen.2014.10.003

Diekmann K, Hodkinson TR, Fricke E, Barth S (2008) An optimized chloroplast DNA extraction protocol for grasses (Poaceae) proves suitable for whole plastid genome sequencing and snp detection. PLoS ONE 3:e2813. https://doi.org/10.1371/journal.pone.0002813

DNA recommendations-1994 report concerning further recommendations of the DNA Commission of the ISFH regarding PCR-based polymorphisms in STR (short tandem repeat) systems (1995) Vox Sang 69:70-71

DNeasy® Plant Mini Kit Handbook. (2012) Qiagen, Hilden, Germany

Drew BT, Ruhfel BR, Smith SA, Moore MJ, Briggs BG, Gitzendanner MA, Soltis PS, Soltis DE (2014) Another look at the root of the angiosperms reveals a familiar tale. Syst Biol 63:368-382

Drug Enforcement Administration's Special Testing and Research Laboratory (2005) Monograph: marijuana. http://www.swgdrug.org/Monographs/MARIJUANA.pdf. Accessed July 2 2015

Dufresnes C, Jan C, Bienert F, Goudet J, Fumagalli L (2017) Broad-scale genetic diversity of *Cannabis* for forensic applications. PLoS ONE 12:e0170522. https://doi.org/10.1371/journal.pone.0170522

Dumolin-Lapegue S, Pemonge MH, Petit RJ (1997) An enlarged set of consensus primers for the study of organelle DNA in plants. Mol Ecol 6:393-397

Earl DA, vonHoldt BM (2012) Structure harvester: A website and program for visualizing structure output and implementing the evanno method. Conserv Genet Resour 4:359-361. https://doi.org/10.1007/s12686-011-9548-7

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379. https://doi.org/10.1371/journal.pone.0019379

ElSohly MA (2007) Marijuana and the cannabinoids. Humana Press, Totowa

Elsohly MA, Slade D (2005) Chemical constituents of marijuana: the complex mixture of natural cannabinoids. Life sciences 78:539-548. https://doi.org/10.1016/j.lfs.2005.09.011

Elwick K, Zeng X, King J, Budowle B, Hughes-Stamm S (2017) Comparative tolerance of two massively parallel sequencing systems to common PCR inhibitors. Int J Leg Med. https://doi.org/10.1007/s00414-017-1693-426

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564-567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Faeti V, Mandolino G, Ranalli P (1996) Genetic diversity of *Cannabis sativa* germplasm based on RAPD markers. Plant Breeding 115:367-370. https://doi.org/10.1111/j.1439-0523.1996.tb00935.x

Fangan BM, Stedje B, Stabbetorp OE, Jensen ES, Jakobsen KS (1994) A general approach for PCR-amplification and sequencing of chloroplast DNA from crude vascular plant and algal tissue. BioTechniques 16:484-494

Fitch WM, Ayala FJ, National Academy of S. (1995) Tempo and mode in evolution: genetics and paleontology 50 years after simpson. National Academies Press Washington, D.C.

Flachowsky H, Schumann E, Weber WE, Peil A (2001) Application of AFLP for the detection of sex-specific markers in hemp. Plant Breeding 120:305-309. https://doi.org/10.1046/j.1439-0523.2001.00620.x

Fleming MP, Clarke R (1998) Physical evidence for the antiquity of *Cannabis sativa* L. J of Int Hemp Association 5:80-93

Forapani S, Ranalli P, Mandolino G, Moliterni VMC, Carboni A, Paoletti C (2001) Comparison of hemp varieties using random amplified polymorphic DNA markers. Crop science 41:1682-1689

Fordyce SL, Mogensen HS, Borsting C, Lagacé RE, Chang CW, Rajagopalan N, Morling N (2015) Second-generation sequencing of forensic strs using the Ion Torrent HID STR 10-plex and the Ion PGM. Forensic Sci Int Genet 14:132-140. https://doi.org/10.1016/j.fsigen.2014.09.020

Fulgosi H, Jezic M, Lepedus H, Stefanic PP, Curkovic-Perica M, Cesar V (2012) Degradation of chloroplast DNA during natural senescence of maple leaves. Tree Physiol 32:346-354. https://doi.org/10.1093/treephys/tps014

Gao C, Xin P, Cheng C, Tang Q, Chen P, Wang C, Zang G, Zhao L (2014) Diversity analysis in *Cannabis sativa* based on large-scale development of expressed sequence tag-derived simple sequence repeat markers. PLoS ONE 9:e110638. https://doi.org/10.1371/journal.pone.0110638

Gigliano GS (1998) Identification of *Cannabis sativa* L. (cannabaceae) using restriction profiles of the internal transcribed spacer II (ITS2). Sci Justice 38: 225-230

Gigliano GS (1999) Preliminary data on the usefulness of internal transcribed spacer I (ITS1) sequence in *Cannabis sativa* L. identification. J Forensic Sci 44:475-477

Gigliano GS, Caputo P, Cozzolino S (1997) Ribosomal DNA analysis as a tool for the identification of *Cannabis sativa* L. Specimens of forensic interest. Sci Justice 37:171-174. https://doi.org/10.1016/s1355-0306(97)72170-1

Gill P, Brinkmann B, d'Aloja E, Andersen J, Bar W, Carracedo A, Dupuy B, Eriksen B, Jangblad M, Johnsson V, Kloosterman AD, Lincoln P, Morling N, Rand S, Sabatier

M, Scheithauer R, Schneider P, Vide MC (1997) Considerations from the European DNA profiling group (EDNAP) concerning STR nomenclature. Forensic Sci Int 87:185-192

Gillan R, Cole MD, Linacre A, Thorpe JW, Watson ND (1995) Comparison of *Cannabis sativa* by random amplification of polymorphic DNA (RAPD) and HPLC of cannabinoids: a preliminary study. Sci Justice 35:169-177. https://doi.org/10.1016/s1355-0306(95)72658-2

Gilmore S, Peakall R (2003) Isolation of microsatellite markers in *Cannabis sativa* L. (marijuana). Mol Ecol 3:105-107. https://doi.org/10.1046/j.1471-8286.2003.00367.x

Gilmore S, Peakall R, Robertson J (2003) Short tandem repeat (STR) DNA markers are hypervariable and informative in *Cannabis sativa*: implications for forensic investigations. Forensic Sci Int 131:65-74

Gilmore S, Peakall R, Robertson J (2007) Organelle DNA haplotypes reflect crop-use characteristics and geographic origins of *Cannabis sativa*. Forensic Sci Int 172:179-190. https://doi.org/10.1016/j. forsciint.2006.10.025

Griffiths RAL, Barber MD, Johnson PE, Gillbard SM, Haywood MD, Smith CD, Arnold J, Burke T, Urquhart AJ, Gill P (1998) New reference allelic ladders to improve allelic designation in a multiplex STR system. Int J Legal Med 111:267-272

Grotenhermen F, Russo E (2002) *Cannabis* and cannabinoids pharmacology, toxicology, and therapeutic potential. The Haworth Integrative Healing Press, New York

Guo F, Yu J, Zhang L, 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. https://doi.org/10.1016/j.fsigen.2017.09.003

Guo Y, Li J, Li C-I, Long J, Samuels DC, Shyr Y (2012) The effect of strand bias in Illumina short-read sequencing data. BMC Genomics 13:666. https://doi.org/10.1186/1471-2164-13-666

Hakki EE, Kayis SA, Pinarkara E, Sag A (2007) Inter simple sequence repeats separate efficiently hemp from marijuana (*Cannabis sativa* L.). Electron J Biotechnol 10:570-581. https://doi.org/10.2225/vol10-issue4-fulltext-4

Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107:1-8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings Biolog Sci 270:313-321. https://doi.org/10.1098/rspb.2002.2218

Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society B: Biolog Sci 270:S96-S99. https://doi.org/10.1098/rsbl.2003.0025

Hillig KW (2005) Genetic evidence for speciation in *Cannabis* (Cannabaceae). Genet Res and Crop Evol 52:161-180. https://doi.org/10.1007/s10722-003-4452-y

Hillig KW, Mahlberg PG (2004) A chemotaxonomic analysis of cannabinoid variation in *Cannabis* (Cannabaceae). Am J Bot 91:966-975. https://doi.org/10.3732/ajb.91.6.966

Holleley CE, Geerts PG (2009) Multiplex manager 1.0: a crossplatform computer program that plans and optimizes multiplex PCR. Biotechniques 46:511-517

Houston R, Birck M, Hughes-Stamm S, Gangitano D (2016) Evaluation of a 13-loci STR multiplex system for *Cannabis sativa* genetic identification. Int J Leg Med 130:635-647. https://doi.org/10.1007/s00414-015-1296-x

Houston R, Birck M, Hughes-Stamm S, Gangitano D (2017) Developmental and internal validation of a novel 13 loci STR multiplex method for *Cannabis sativa* DNA profiling. Leg Med (Tokyo, Japan) 26:33-40. https://doi.org/10.1016/j.legalmed.2017.03.001

Houston R, Birck M, LaRue B, Hughes-Stamm S, Gangitano D (2018) Nuclear, chloroplast, and mitochondrial data of a US *Cannabis* DNA database. Int J of Leg Med. https://doi.org/10.1007/s00414-018-1798-4

Howard C, Gilmore S, Robertson J, Peakall R (2008) Developmental validation of a *Cannabis Sativa* STR multiplex system for forensic analysis. J Forensic Sci 53:1061-1067. https://doi.org/10.1111/j.1556-4029.2008.00792.x

Howard C, Gilmore S, Robertson J, Peakall R (2009) A *Cannabis sativa* STR genotype database for Australian seizures: forensic applications and limitations. J Forensic Sci 54:556-563. https://doi.org/10.1111/j.1556-4029.2009.01014.x

Hsieh HM, Hou RJ, Tsai LC, Wei CS, Liu SW, Huang LH, Kuo YC, Linacre A, Lee JC (2003) A highly polymorphic STR locus in *Cannabis sativa*. Forensic Sci Int 131:53-58

Hu ZG, Guo HY, Hu XL et al (2012) Genetic diversity research of hemp (*Cannabis sativa* L.) cultivar based on AFLP analysis. J Plant Genet Resources 13:555-561

Ion library taqman® quantitation kit user guide. (2017). Thermo Fisher Scientific, South San Francsico, CA

Jagadish V, Robertson J, Gibbs A (1996) Rapd analysis distinguishes *Cannabis sativa* samples from different sources. Forensic Sci Int 79:113-121. https://doi.org/10.1016/0379-0738(96)01898-1

Jiang HE, Li X, Zhao YX, Ferguson DK, Hueber F, Bera S, Wang YF, Zhao LC, Liu CJ, Li CS (2006) A new insight into *Cannabis sativa* (Cannabaceae) utilization from 2500-year-old yanghai tombs, Xinjiang, China. J of Ethnopharmacology 108:414-422. https://doi.org/10.1016/j.jep.2006.05.034

Jombart T (2008) Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403-1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 11:94. https://doi.org/10.1186/1471-2156-11-94

Kavlick MF, Lawrence HS, Merritt RT, Fisher C, Isenberg A, Robertson JM, Budowle B (2011) Quantification of human mitochondrial DNA using synthesized DNA standards. J of Forenisc Sci 56:1457-1463. https://doi.org/10.1111/j.1556-4029.2011.01871.x

Khatak S, Ghai M, Dahiya S (2016) ISSR marker based inter and intra-specific diversity analysis in different genotypes of *Cannabis sativa*. Int Conf on Innovative Res in Engg Sci and Mang 3:6-16

Kline MC, Duewer DL, Travis JC, Smith MV, Redman JW, Vallone PM, Decker AE, Butler JM (2009) Production and certification of nist standard reference material 2372 human DNA quantitation standard. Anal Bioanal Chem 394:1183-1192. https://doi.org/10.1007/s00216-009-2782-0

Knight G, Hansen S, Connor M, Poulsen H, McGovern C, Stacey J (2010) The results of an experimental indoor hydroponic *Cannabis* growing study, using the 'screen of green' (scrog) method-yield, tetrahydrocannabinol (THC) and DNA analysis. Forensic Sci Int 202:36-44. https://doi.org/10.1016/j.forsciint.2010.04.022

Kohjyouma M, Lee IJ, Iida O, Kurihara K, Yamada K, Makino Y, Sekita S, Satake M (2000) Intraspecific variation in *Cannabis sativa* L. Based on intergenic spacer region of chloroplast DNA. Biol Pharm Bull 23:727-730. https://doi.org/10.1248/bpb.23.727

Kohnemann S, Nedele J, Schwotzer D, Morzfeld J, Pfeiffer H (2012) The validation of a 15 STR multiplex PCR for *Cannabis* species. Int J Leg Med 126:601-606. https://doi.org/10.1007/s00414-012-0706-6

Kojoma M, Iida O, Makino Y, Sekita S, Satake M (2002) DNA fingerprinting of *Cannabis sativa* using inter-simple sequence repeat (ISSR) amplification. Planta Med 68:60-63. https://doi.org/10.1055/s-2002-19875

Kojoma M, Seki H, Yoshida S, Muranaka T (2006) DNA polymorphisms in the tetrahydrocannabinolic acid (THCA) synthase gene in "drug-type" and "fiber-type" *Cannabis sativa* L. Forensic Sci Int 159:132-140. https://doi.org/10.1016/j.forsciint.2005.07.005

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour 15:1179-1191. https://doi.org/10.1111/1755-0998.12387

Koressaar T, Remm M (2007) Enhancements and modifications of primer design program primer3. Bioinformatics 23:1289-1291. https://doi.org/10.1093/bioinformatics/btm091

Kraemer L, Beszteri B, Gäbler-Schwarz S, Held C, Leese F, Mayer C, Pöhlmann K, Frickenhaus S (2009) Stamp: extensions to the staden sequence analysis package for high throughput interactive microsatellite marker design. BMC Bioinformatics 10:41 https://doi.org/10.1186/1471-2105-10-41

Kun T, Lyons LA, Sacks BN, Ballard RE, Lindquist C, Wictum EJ (2013) Developmental validation of mini-dogfiler for degraded canine DNA. Forensic Sci Int Genet 7:151-158. https://doi.org/10.1016/j.fsigen.2012.09.002

Kung CT (1952) Archaeology in China. University of Toronto Press Toronto, ON

Kwon SY, Lee HY, Kim EH, Lee EY, Shin KJ (2016) Investigation into the sequence structure of 23 Y chromosomal STR loci using massively parallel sequencing. Forensic Sci Int Genet 25:132-141. https://doi.org/10.1016/j.fsigen.2016.08.010

Lamarck JB (1785) Encyclopédie métodique. Botanique: Paris-Liege, 1783-1803

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal w and clustal x version 2.0. Bioinformatics 23:2947-2948. https://doi.org/10.1093/bioinformatics/btm404

Lazaruk K, Walsh PS, Oaks F, Gilbert D, Rosenblum BB, Menchen S, Scheibler D, Wenz HM, Holt C, Wallin J (1998) Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument. Electrophoresis 19:86-93. https://doi.org/10.1002/elps.1150190116

Lewis PO, Zaykin D (2001) Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c)

Li H, Cao H, Cai YF, Wang JH, Qu SP, Huang XQ (2014) The complete chloroplast genome sequence of sugar beet (Beta vulgaris ssp. vulgaris). Mitochondrial DNA 25:209-211. https://doi.org/10.3109/19401736.2014.883611

Li HL (1974) An archaeological and historical account of *Cannabis* in China. Econ Bot 28:437-448.

Li HL (1974) The origin and use of *Cannabis* in eastern asia linguistic-cultural implications. Econ Bot 28: 293-301. https://doi.org/10.1007/bf02861426

Li HL (1978) Hallucinogenic plants in Chinese herbals. J of Psychedelic Drugs 10: 17-26. https://doi.org/10.1080/02791072.1978.10471863

Linacre A, Gusmao L, Hecht W, Hellmann AP, Mayr WR, Parson W, Prinz M, Schneider PM, Morling N (2011) ISFG: recommendations regarding the use of non-human (animal) DNA in forensic genetic investigations. Forensic Sci Int Genet 5:501-505. https://doi.org/10.1016/j.fsigen.2010.10.017

Linacre A, Thorpe J (1998) Detection and identification of *Cannabis* by DNA. Forensic Sci Int 91:71-76

Linnaeus C (1753) Species Plantarum. Stockholm, Sweden

Logares R, Audic S, Bass D et al (2014) Patterns of rare and abundant marine microbial eukaryotes. Curr Biol 24:813-821. https://doi.org/10.1016/j.cub.2014.02.050

Lucas AA (2014) Colorado-world's first fully regulated recreational marijuana market-collects $2M in recreational pot taxes. International Business Times.

http://au.ibtimes.com/coloradoworlds-first-fully-regulated-recreational-marijuana-marketcollects-2m-recreational-pot. Accessed April 29 2015

Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, Smith LM (1990) High speed DNA sequencing by capillary electrophoresis. Nucleic Acids Res 18: 4417-4421

Mandolino G, Carboni A, Bagatta M, Moliterni VMC, Ranalli P (2002) Occurrence and frequency of putatively Y chromosome linked DNA markers in *Cannabis sativa* L. Euphytica 126:211-218. https://doi.org/10.1023/a:1016382128401

Mandolino G, Carboni A, Forapani S, Faeti V, Ranalli P (1999) Identification of DNA markers linked to the male sex in dioecious hemp (*Cannabis sativa* L.). Theor and Appl Genet 98:86-92. https://doi.org/10.1007/s001220051043

Mansouri H, Bagheri M (2017) Induction of polyploidy and its effect on *Cannabis sativa* l. In: Chandra S, Lata H, ElSohly MA, eds. *Cannabis sativa* L. - botany and biotechnology. Springer International Publishing Cham. New York, pp. 365-383

Marijuana and the controlled substances act (2014) Congressional Digest 93:2.

Massart S, Olmos A, Jijakli H, Candresse T (2014) Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188:90-96. https://doi.org/10.1016/j.virusres.2014.03.029

McKenna GJ (2014) The current status of medical marijuana in the United States. Hawai'i J of Med & Pub Health 73:105-108

Mead A (2017) The legal status of *Cannabis* (marijuana) and cannabidiol (CBD) under U.S. Law. Epilepsy Behav 70:288-291. https://doi.org/10.1016/j.yebeh.2016.11.021

Mello IC, Ribeiro AS, Dias VH, Silva R, Sabino BD, Garrido RG, Seldin L, de Moura Neto RS (2016) A segment of rbcl gene as a potential tool for forensic discrimination of *Cannabis sativa* seized at rio de janeiro, brazil. Int J Legal Med 130:353-356. https://doi.org/10.1007/s00414-015-1170-x

Mendoza MA, Mills DK, Lata H, Chandra S, ElSohly MA, Almirall JR (2009) Genetic individualization of *Cannabis sativa* by a short tandem repeat multiplex system. Anal Bioanal Chem 393:719-726. https://doi.org/10.1007/s00216-008-2500-3

Mikes F, Hofmann A, Waser PG (1971) Identification of (-)-delta 9-6a,10a-trans-tetrahydrocannabinol and two of its metabolites in rats by use of combination gas chromatography-mass spectrometry and mass fragmentography. Biochem Pharmacol 20:2469-2476

Miller Coyle H, Palmbach T, Juliano N, Ladd C, Lee HC (2003) An overview of DNA methods for the identification and individualization of marijuana. Croat Med J 44:315-321

Miller Coyle H, Shutler G, Abrams S, Hanniman J, Neylon S, Ladd C, Palmbach T, Lee HC (2003) A simple DNA extraction method for marijuana samples used in amplified fragment length polymorphism (AFLP) analysis. J Forensic Sci 48:343-347

Minelute® handbook. (2008). Qiagen, Hilden, Germany

Mitosinka GT, Thornton JI, Hayes TL (1972) The examination of cystolithic hairs of *Cannabis* and other plants by means of the scanning electron microscope. J Forensic Sci Soc 12:521-529

Moorthie S, Mattocks CJ, Wright CF (2011) Review of massively parallel DNA sequencing technologies. The HUGO Journal 5:1-12. 10.1007/s11568-011-9156-3

Mourad GS (1998) Chloroplast DNA isolation. In: Martinez-Zapater JM, Salinas J (eds) Arabidopsis Protocols. Humana Press, Totowa, NJ, pp 71-77

Mueller UG, Wolfenbarger LL (1999) Aflp genotyping and fingerprinting. Trends Ecol Evol 14:389-394. https://doi.org/10.1016/S0169-5347(99)01659-6

Mukherjee A, Roy SC, De Bera S, Jiang HE, Li X, Li CS, Bera S (2008) Results of molecular analysis of an archaeological hemp (*Cannabis sativa* L.) DNA sample from north west China. Genet Resources and Crop Evol 55:481-485. https://doi.org/10.1007/s10722-008-9343-9

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S (2011) Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39:e90. https://doi.org/10.1093/nar/gkr344

Nei M (1972) Genetic distance between populations. Am Nat 106:283-292

Newmaster SG, Fazekas AJ, Ragupathy S (2006) DNA barcoding in land plants: Evaluation of rbcL in a multigene tiered approach. Canadian J Bot 84:335-341. https://doi.org/10.1139/b06-047

Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC, Markham AF (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res 17:2503-2516

Oh H, Seo B, Lee S, Ahn DH, Jo E, Park JK, Min GS (2016) Two complete chloroplast genome sequences of *Cannabis sativa* varieties. Mitochondrial DNA Part A DNA Mapp Seq Anal 27: 2835-2837. 10.3109/19401736.2015.1053117

Olaisen B, Bär W, Brinkmann B, Budowle B, Carracedo A, Gill P, Lincoln P, Mayr WR, Rand S (1998) DNA recommendations 1997 of the international society for forensic genetics. Vox Sang 74:61-63

Ooyen-Houben Mv, Kleemans E (2015) Drug policy: the "dutch model". Crime and Justice 44:165-226. https://doi/10.1086/681551

Pacula RL, Powell D, Heaton P, Sevigny EL (2015) Assessing the effects of medical marijuana laws on marijuana use: the devil is in the details. J Policy Anal Manag 34: 7-31. https://doi.org/10.1002/pam.21804

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol 28:87-97

Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmão L, Hares DR, Irwin JA, King JL, Knijff Pd, Morling N, Prinz M, Schneider PM, Neste CV, Willuwit S, Phillips C (2016) Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the international society for forensic genetics (ISFG) on minimal nomenclature requirements. Forensic Sci Int Genet 22:54-63. https://doi.org/10.1016/j.fsigen.2016.01.009

Paun O, Schönswetter P (2012) Amplified fragment length polymorphism (AFLP) - an invaluable fingerprinting technique for genomic, transcriptomic and epigenetic studies. Methods in Mol Biol (Clifton, NJ) 862:75-87. https://doi.org/10.1007/978-1-61779-609-8_7

Peil A, Flachowsky H, Schumann E, Weber WE (2003) Sex-linked AFLP markers indicate a pseudoautosomal region in hemp (*Cannabis sativa* L.). Theor Appl Genet 107:102-109. https://doi.org/10.1007/s00122-003-1212-5

Phillips C, Gettings KB, King JL, Ballard D, Bdner M, Borsuk L, Parson W (2018) "The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR sequence guide. Forensic Sci Int Genet 34:262-169. https://doi.org/10.1016/j.fsigen.2018.02.017

Pillay M, Kenny ST (2006) Structural organization of the nuclear ribosomal RNA genes in *Cannabis* and Humulus (cannabaceae). Plant Systematics and Evolution: 97

Piluzza G, Delogu G, Cabras A, Marceddu S, Bullitta S (2013) Differentiation between fiber and drug types of hemp (*Cannabis sativa* L.) from a collection of wild and domesticated accessions. Genet Resour Crop Evol 60:2331-2342. https://doi.org/10.1007/s10722-013-0001-5

Pinarkara E, Kayis SA, Hakki EE, Sag A (2009) RAPD analysis of seized marijuana (*Cannabis sativa* L.) in turkey. Electronic J of Biotech 12:1-13. https://doi.org/10.2225/vol12-issue1-fulltext-7

Piomelli D, Russo EB (2016) The *Cannabis sativa* versus *Cannabis* indica debate: An interview with ethan russo, md. *Cannabis* Cannabinoid Res 1:44-46. https://doi.org/10.1089/can.2015.29003.ebr

Pootakham W, Jomchai N, Ruang-areerate P, Shearman JR, Sonthirod C, Sangsrakru D, Tragoonrung S, Tangphatsornruang S (2015) Genome-wide SNP discovery and identification of qtl associated with agronomic traits in oil palm using genotyping-

by-sequencing (GBS). Genomics 105:288-295. https://doi.org/10.1016/j.ygeno.2015.02.002

Prepare amplicon libraries without fragmentation using the ion plus fragment library kit. (2016). Thermo Fisher Scientific, South San Francsico, CA

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238:336-341

Punja ZK, Rodriguez G, Chen S (2017) Assessing genetic diversity in *Cannabis sativa* using molecular approaches. ln: Chandra S, Lata H, ElSohly M (eds) *Cannabis sativa* L. - Botany and Biotechnology. Springer International Publishing Cham. New York, pp. 395-418

QIAamp DNA Investigator® Handbook (2012) Qiagen, Hilden, Germany

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:241. https://doi.org/101186/1471-2164-13-341

Quality assurance standards for forensic DNA testing laboratories (2009) FBI. http://www.fbi.gov/about-us/lab/biometric-analysis/codis/qas_testlabs. Accessed April 29 2015

Qubit® dsDNA HS assay kits. (2015). Thermo Fisher Scientific, South San Francisco, CA

Raj A, Stephens M, Pritchard JK (2014) fastStructure: variational inference of population structure in large SNP data sets. Genet 197:573-589. https://doi.org/10.1534/genetics.114.164350

Ranney T. (2006) Polyploidy: from evolution to new plant development. Combined Proceedings Int Plant Propagators' Soc 56

Raymond M, Rousset F (1995) GENEPOP (Version 1.2): population genetics software for exact tests and ecumenicism. J Hered 86: 248-249

Room R (2014) Legalizing a market for *Cannabis* for pleasure: Colorado, Washington, Uruguay and beyond. Addiction (Abingdon, England) 109:345-351. https://doi.org/10.1111/add.12355

Rothberg JM, Hinz W, Rearick TM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348. 10.1038/nature10242

Rotherham D, Harbison SA (2011) Differentiation of drug and non-drug *Cannabis* using a single nucleotide polymorphism (SNP) assay. Forensic Sci Int 207:193-197. https://doi.org/10.1016/j.forsciint.2010.10.006

Sajantila A, Puomilahti S, Johnsson V, Ehnholm C (1992) Amplification of reproducible allele markers for amplified fragment length polymorphism analysis. Biotechniques 12:16, 18, 20-22

Sakamoto K, Abe T, Matsuyama T, Yoshida S, Ohmido N, Fukui K, Satoh S (2005) RAPD markers encoding retrotransposable elements are linked to the male sex in *Cannabis sativa* L. Genome 48:931-936. https://doi.org/10.1139/g05-056

Sakamoto K, Akiyama Y, Fukui K, Kamada H, Satoh S (1998) Characterization; genome sizes and morphology of sex chromosomes in hemp (*Cannabis sativa* L.). Cytologia (Tokyo) 63:459-464. https:/doi.org/10.1508/cytologia.63.459

Sakamoto K, Ohmido N, Fukui K, Kamada H, Satoh S (2000) Site-specific accumulation of a line-like retrotransposon in a sex chromosome of the dioecious plant *Cannabis sativa*. Plant Mol Biol 44:723-732. https://doi.org/10.1023/A:1026574405717

Sakamoto K, Shimomura K, Komeda Y, Kamada H, Satoh S (1995) A male-associated DNA sequence in a dioecious plant, *Cannabis sativa* L. Plant Cell Physiol 36:1549-1554

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74:5463-5467

Sawler J, Stout JM, Gardner KM et al (2015) The genetic structure of marijuana and hemp. PLoS ONE 10:e0133292. https://doi.org/10.1371/journal.pone.0133292

Schield C, Campelli C, Sycalik J, Randle C, Hughes-Stamm S, Gangitano D (2016) Identification and persistence of Pinus pollen DNA on cotton fabrics: a forensic application. Sci Justice 56:29-34. https://doi.org/10.1016/j.scijus.2015.11.005

Schultes R, Hofmann A (1980) The botany and chemistry of hallucinogens. Charles C Thomas, Springfield, IL

Schultes RE (1970) Random thoughts and queries on the botany of *Cannabis*. The botany and chemistry of *Cannabis* J. & A. Churchill, London, pp.11-33

Schultes RE (1973) Man and marijuana: thousands of years before it became the superstar of the drug culture, *Cannabis* was cultivated for fiber, food, and medicine. American Museum of Natural History

Schultes RE, Klein WM, Plowman T, Lockwood TE (1974) *Cannabis*: an example of taxonomic neglect. Botanical Museum Leaflets, Harvard University 23:337-367

Schury N, Schleenbecker U, Hellmann AP (2014) Forensic animal DNA typing: Allele nomenclature and standardization of 14 feline STR markers. Forensic Sci Int Genet 12:42-59. https://doi.org/10.1016/j.fsigen.2014.05.002

Scientific Working Group on DNA Analysis Methods. Validation Guidelines for DNA Analysis Methods (2012). http://media.wix.com/ugd/4344b0_cbc27d16dcb64fd88cb36ab2a2a25e4c.pdf Accessed September 2016

Seo SB, King JL, Warshauer DH, Davis CP, Ge J, Budowle B (2013) Single nucleotide polymorphism typing with massively parallel sequencing for human identification. Int J Leg Med 127:1079-1086. https://doi.org/10.1007/s00414-013-0879-7

Shao H, Song SJ, Clarke RC (2003) Female-associated DNA polymorphisms of hemp (*Cannabis sativa* L.). J of Industrial Hemp 8:5-9. https://doi.org/10.1300/J237v08n01_02

Shephard HL, Parker JS, Darby P, Ainsworth CC (2000) Sexual development and sex chromosomes in hop. New Phytol 148:397-411. https://doi.org/10.1046/j.1469-8137.2000.00771.x

Shibuya EK, Sarkis JES, Negrini-Neto O, Martinelli LA (2007) Carbon and nitrogen stable isotopes as indicative of geographical origin of marijuana samples seized in the city

of São Paulo (Brazil). Forensic Sci Int 167:8-15. https://doi.org/ 10. 1016/j.forsciint.2006.06.002

Shibuya EK, Souza Sarkis JE, Neto ON, Moreira MZ, Victoria RL (2006) Sourcing Brazilian marijuana by applying IRMS analysis to seized samples. Forensic Sci Int 160:35-43. https://doi.org/10.1016/j.forsciint. 2005.08.011

Shibuya EK, Souza Sarkis JE, Neto ON, Moreira MZ, Victoria RL (2006) Sourcing brazilian marijuana by applying irms analysis to seized samples. Forensic Sci Int 160:35-43. https://doi.org/10.1016/j.forsciint.2005.08.011

Shirley N, Allgeier L, LaNier T, Coyle HM (2013) Analysis of the NMI01 marker for a population database of *Cannabis* seeds. J Forensic Sci 58:S176-182. https://doi.org/ 10.1111/1556-4029

Shirota O, Watanabe A, Yamazaki M, Saito K, Shibano K, Sekita S, Satake N (1998) Random amplified polymorphic DNA and restriction fragment length polymorphism analyses of *Cannabis sativa*. Natural medicines 52: 160-166

Sirikantaramas S, Taura F, Tanaka Y, Ishikawa Y, Morimoto S, Shoyama Y (2005) Tetrahydrocannabinolic acid synthase, the enzyme controlling marijuana psychoactivity, is secreted into the storage cavity of the glandular trichomes. Plant Cell Physiol 46:1578-1582. https://doi.org/10.1093/pcp/pci166

Small E (1978) A numerical and nomenclatural analysis of morpho-geographic taxa of Humulus. Syst Bot 3:37-76. https://doi.org/10.2307/2418532

Small E (1979b) The species problem in *Cannabis*: science and semantics. Volume 2. Corpus, Toronto, ON

Small E (2015) Evolution and classification of *Cannabis sativa* (marijuana, hemp) in relation to human utilization. Bot Rev 81:189-294. 10.1007/s12229-015-9157-3

Small E, Beckstead HD (1973) Cannabinoid phenotypes in *Cannabis sativa*. Nature 245:147. https://doi.org/10.1038/245147a0

Small E, Cronquist A (1976) A practical and natural taxonomy for *Cannabis*. Taxon 25:405-435

Soler S, Gramazio P, Figàs MR, Vilanova S, Rosa E, Llosa ER, Borràs D, Plazas M, Prohens J (2017) Genetic structure of *Cannabis sativa* var. indica cultivars based on genomic SSR (gSSR) markers: Implications for breeding and germplasm management. Ind Crops Prod 104:171-178. https://doi.org/10.1016/j.indcrop.2017.04.043

Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, Normandeau É, Laroche J, Larose S, Jean M, Belzile F (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS ONE 8:e54603. https://doi.org/10.1371/journal.pone.0054603

Soorni A, Fatahi R, Haak DC, Salami SA, Bombarely A (2017) Assessment of genetic diversity and population structure in Iranian *Cannabis* germplasm. Sci Rep 7:15668. https://doi.org/10.1038/s41598-017-15816-5

Srivastava AK, Schlessinger D (1991) Structure and organization of ribosomal DNA. Biochimie 73:631-638

Swerdlow H, Gesteland R (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res 18:1415-1419

SWGDAM validation guidelines for DNA analysis methods. (2016) https://docs.wixstatic.com/ugd/4344b0_813b241e8944497e99b9c45b163b76bd.pdf

Swofford D (2002) PAUP*: phylogenetic Analysis Using Parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, MA

Sytsma KJ, Morawetz J, Pires JC, Nepokroeff M, Conti E, Zjhra M, Hall JC, Chase MW (2002) Urticalean rosids: circumscription, rosid ancestry, and phylogenetics based on rbcL, trnL-F, and ndhF sequences. Am J Bot 89:1531-1546. https://doi.org/10.3732/ajb.89.9.1531

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) Mega6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725-2729. https://doi.org/10.1093/molbev/mst197

Techen N, Chandra S, Lata H, Elsohly MA, Khan IA (2010) Genetic identification of female *Cannabis sativa* plants at early developmental stage. Planta Med 76:1938-1939. https://doi.org/10.1055/s-0030-1249978

Tereba A (1999) Tools for analysis of population statistics. Profiles in DNA 3. Promega Corporation

Vallone PM, Butler JM (2004) Autodimer: A screening tool for primer-dimer and hairpin structures. BioTechniques 37:226-231.

Valverde L, Lischka C, Erlemann S, de Meijer E, de Pancorbo MM, Pfeiffer H, Köhnemann S (2014) Nomenclature proposal and SNPSTR haplotypes for 7 new *Cannabis sativa* L. STR loci. Forensic Sci Int Genet 13:185-186. http://dx.doi.org/10.1016/j.fsigen.2014.08.002

Valverde L, Lischka C, Scheiper S et al (2014) Characterization of 15 STR *Cannabis* loci: Nomenclature proposal and SNPSTR haplotypes. Forensic Sci Int Genet 9:61-65. https://doi.org/10.1016/j.fsigen.2013.11.001

Valverde L, Lischka C, Scheiper S, Nedele J, Challis R, de Pancorbo MM, Pfeiffer H, Kohnemann S (2014) Characterization of 15 STR *Cannabis* loci: nomenclature proposal and SNPSTR haplotypes. Forensic Sci Int Genet 9:61-65. https://doi.org/10.1016/j.fsigen.2013.11.001

van Bakel H, Stout J, Cote A, Tallon C, Sharpe A, Hughes T, Page J (2011) The draft genome and transcriptome of *Cannabis sativa*. Genome Biol 12:R102. https://doi.org/10.1186/gb-2011-12-10-r102

Vasan N, Yelensky R, Wang K, Moulder S, Dzimitrowicz H, Avritscher R, Wang B, Wu Y, Cronin MT, Palmer G, Symmans WF, Miller VA, Stephens P, Pusztai L (2014) A targeted next-generation sequencing assay detects a high frequency of therapeutically targetable alterations in primary and metastatic breast cancers: Implications for clinical practice. The Oncologist 19:53-458. https://doi.org/10.1634/theoncologist.2013-0377

Vergara D, White KH, Keepers KG, Kane NC (2016) The complete chloroplast genomes of *Cannabis sativa* and Humulus lupulus. Mitochondrial DNA Part A DNA Mapp Seq Anal 27:3793-3794. https://doi.org/10.3109/19401736.2015.1079905

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M (1995) AFLP: A new technique for DNA fingerprinting. Nucleic Acids Res 23:4407-4414

Vuylsteke M, Peleman JD, van Eijk MJT (2007) AFLP technology for DNA fingerprinting. Nature Protocols 2:1387. https://doi.org/10.1038/nprot.2007.175

Vyskot B, Hobza R (2004) Gender in plants: sex chromosomes are emerging from the fog. Trends Genet 20: 432-438. https://doi.org/10.1016/j.tig.2004.06.006

Wang C, Guo W, Zhang T, Li Y, Liu H (2009) AutoSSR: an improved automatic software for SSR analysis from large-scale est sequences. Cotton Science 21:243-247

Wang S, Shi C, Gao L-Z (2013) Plastid genome sequence of a wild woody oil species, prinsepia utilis, provides insights into evolutionary and mutational patterns of rosaceae chloroplast genomes. PLoS ONE 8:e73946. https://doi.org/10.1371/journal.pone.0073946

Warshauer DH, Lin D, Hari K, Jain R, David C, LaRue B, King JL, Budowle B (2013) Strait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci Int Genet 7:409-417. https://doi.org/10.1016/j.fsigen.2013.04.005

Weising K, Gardner RC (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome 42:9-19

Werf HVD, Mathussen EWJM, Haverkort AJ (1996) The potential of hemp (*Cannabis sativa* L.) for sustainable fibre production: a crop physiological appraisal. Ann Appl Biol 129:109-123. https://doi.org/10.1111/j.1744-7348.1996.tb05736.x

White KH, Vergara D, Keepers KG, Kane NC (2016) The complete mitochondrial genome for *Cannabis sativa*. Mitochondrial DNA Part B 1:715-716. https://doi.org/10.1080/23802359.2016.1155083

Wictum E, Kun T, Lindquist C, Malvick J, Vankan D, Sacks B (2013) Developmental validation of dogfiler, a novel multiplex for canine DNA profiling in forensic casework. Forensic Sci Int Genet 7:82-91. https://doi.org/10.1016/j.fsigen.2012.07.001

Wilkinson M, Linacre AMT (2000) The detection and persistence of *Cannabis sativa* DNA on skin. Sci Justice 40:11-14. https://doi.org/10.1016/S1355-0306(00)71927-7

Woerner AE, King JL, Budowle B (2017) Fast STR allele identification with STRait Razor 3.0. Forensic Sci Int Genet 30:18-23. https://doi.org/10.1016/j.fsigen.2017.05.008

Yang MQ, van Velzen R, Bakker FT, Sattarian A, Li DZ, Yi TS (2013) Molecular phylogenetics and character evolution of cannabaceae. Taxon 62:473-485. https://doi.org/10.12705/623.9

Yu YL, Lin TY (1997) Construction of phylogenetic tree fornicotiana species based on RAPD markers. J Plant Res 110:187-193. https://doi.org/10.1007/BF02509307

Zabeau M, Vos P (1993) Selective restriction fragment amplification: a genreal method for DNA fingerprinting. European Patent Office, publication 0 534 858 A1, bulletin 93/13

Zaya DN, Ashley MV (2012) Plant genetics for forensic applications. Methods Mol Biol 862:35-52. https://doi.org/10.1007/978-1-61779-609-8_4

Zeng X, King JL, Budowle B (2017) Investigation of the STR loci noise distributions of PowerSeqTM Auto System. Croat Med J 58:214-221. https://doi.org/10.3325/cmj.2017.58.214

Zhang Q, Sodmergen, Liu Y (2003) Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. Plant and Cell Physiol 44:941-951

Zhao X, Li H, Wang Z, Ma K, Cao Y, Liu W (2016) Massively parallel sequencing of 10 autosomal STRs in Chinese using the Ion Torrent personal genome machine (PGM). Forensic Sci Int Genet 25:34-38. https://doi.org/10.1016/j.fsigen.2016.07.014

Zhao X, Ma K, Li H, Cao Y, Liu W, Zhou H, Ping Y (2015) Multiplex Y-STRs analysis using the Ion Torrent personal genome machine (PGM). Forensic Sci Int Genet 19:192-196. https://doi.org/10.1016/j.fsigen.2015.06.012

VITA

Rachel Houston

Forensic molecular biologist with extensive experience on molecular HID and Non-human DNA forensics. Strong background in forensic biology, SNPs, massive parallel sequencing, biostatistics, population genetics, and forensic plant science.

Relevant Professional Experience

Sam Houston State University January 2014 – Present

- Graduate Assistant
- Aided in laboratory preparation, inventory, administrative duties, and troubleshooting instruments (ABI 3500 Genetic Analyzer and 7500 Real-Time PCR)
- Teaching Assistant for Forensic Biology Lab, Advanced Forensic DNA, and Non-human Forensics
- Assistant instructor for Forensic Science Educator Training class at SHSU (July 2015 and July 2016)
- Exhibitor for High School Tours of SHSU department of Forensic Science
- Volunteer and presenter at Houston Hispanic Forum Career and Education Fair

US Customs and Border Protection Southwest Regional Lab June 2014 – October 2015

- Student Trainee with experience in both drug and latent print units
- Experience using Gas Chromatography/Mass Spectrometry (GC/MS) and Fourier Transform Infrared Spectroscopy (FTIR)
- Project analyzing the application of autosomal DNA profiling of marijuana samples with official MOU collaboration between SHSU and Department of Homeland Security

Education

Sam Houston State University, Huntsville, TX August 2013 – Present

- Pending Doctor of Philosophy in Forensic Science
- GPA: 4.0
- Graduation: May 2018

University of Texas at Dallas, Richardson, TX August 2009 – May 2013

- B.S. in Biology with minor in Criminology
- GPA: 3.793
- Graduated Cum Laude May 2013

Relevant Education Experience

Sam Houston State University

- Crime Scene Investigation, Forensic Biology, Advanced Forensic DNA, Non-human Forensics, Behavioral Genetics, Pharmacogenomics, Forensic Statistics and Interpretation, Statistical Genetics, Forensic Toxicology, Forensic Instrumental Analysis, Pattern and Physical Evidence Concepts, Trace Evidence and Microscopic Analysis, Controlled Substances

University of Texas at Dallas

- Forensic Biology, Biochemistry 1 and 2, Genetics, Molecular and Cell Biology

Skills and Qualifications

Molecular Techniques

- DNA extraction: Proficient with extraction using PCIA, Chelex®, and Qiagen® Investigator Kit, Plant DNeasy kit
- Quantification: Quantifiler® Trio, PowerQuant® System, Investigator® Quantiplex® Pro Kit, InnoQuant® HY; Qubit HS DNA ds
- Amplification: Identifiler® Plus, GlobalFiler® PCR, Investigator® 24plex QS Kit, PowerPlex® Fusion 6C System, Canine Genotypes 1.1 and 1.2 Kits (including participation in ISAG proficiency test)
- Sequencing using Big Dye Direct PCR and Big Dye Terminator v.3.1
- Snapshot minisequencing

Instruments

- Qubit® 2.0 Fluorometer, Robotic Extraction Platforms (QIAcube®, EZ1 Advanced xL, and Automate Express), ABI Real-Time 7500 PCR System, StepOne Real-Time PCR System, GeneAmp® 9700, Eppendorf Mastercycler Gradient Thermal Cycler, T100

Bio-Rad Thermal Cycler, Veriti Thermal Cycler, ProFlex PCR system, Applied Biosystems® 3500 Genetic Analyzer, Ion™ PGM System, Ion™ S5 Sequencing System, Ion™ Chef Instrument

Software

- Applied Biosystems Genemapper ID-X, Microsoft Office, R Statistical Software, Geneious, Primer3, Autodimer, STRmix, Genemapper v5 (design of bins and panels) Structure, GDA, Genepop, Arlequin, PAUP, PowerStats, Past3, and *Adegenet*

Research Grant Funding

- **NIJ** – Graduate Research Fellowship (2015-R2-CX-0030)
  Development of a Comprehensive Genetic Tool for Identification of Cannabis Sativa Samples for Forensic and Intelligence Purposes
  PI: Rachel Houston, CO.PI: David Gangitano
  Award amount: **$46,008/yr ($138,024 total)**
  Grant Period: Jan 2015 – Sept 2018

Publications in Peer Reviewed Journals

- Houston R., LaRue B., Birck M., Hughes-Stamm S., Gangitano D. Nuclear, chloroplast, and mitochondrial data of a US Cannabis DNA Database. Int J Legal Med. 2018. https://doi.org/10.1007/s00414-018-1798-4

- Holmes AS., Houston R., Elwick K., Gangitano D., Hughes-Stamm S. Evaluation of four commercial quantitative real-time PCR kits with inhibited and degraded samples. Int J Legal Med. 2017. https://doi.org/10.1007/s00414-017-1745-9.

- Houston R., Birck M., Hughes-Stamm S., Gangitano D. Developmental and internal validation of a novel 13 loci STR multiplex for Cannabis sativa DNA profiling. Legal Med. 2017 May; 26; 33 – 40.

- Houston R., Birck M., Hughes-Stamm S., Gangitano D. Evaluation of a 13-loci STR multiplex System for Cannabis sativa genetic Identification. Int J Legal Med. 2016 May;130(3);635 – 47.

Peer-Review Presentations/Posters

- "Nuclear, Chloroplast, and Mitochondrial Data of a US Cannabis DNA Database". **Rachel Houston BS**, Sheree Hughes-Stamm PhD, David Gangitano PhD. Pittcon. Orlando, FL. February 2018. (Poster Presentation)

- "Nuclear, Chloroplast, and Mitochondrial Data of a US Cannabis DNA Database". **Rachel Houston BS**, Sheree Hughes-Stamm PhD, David Gangitano PhD. American Academy of Forensic Sciences. Seattle, WA. February 2018. (Oral Presentation)

- "Nuclear, Chloroplast, and Mitochondrial Data of a US Cannabis DNA Database". **Rachel Houston BS**, Sheree Hughes-Stamm PhD, David Gangitano PhD. International Symposium on Human Identification. Seattle, WA. October 2017. (Poster Presentation)

- "Alternate methods for the collection, preservation, & processing of DNA samples from decomposing human cadavers; A DVI strategy". Amy Sorensen, MS; Rachel Houston, BS; Kyleen Elwick, BS; Carrie Mayes, BS; Kayla Ehring, BA; David Gangitano, PhD; **Sheree Hughes-Stamm, PhD.** 6th QIAGEN Investigator Forum. Prague, Czech Republic (2017). Oral Presentation. (Co-author)

- "Developmental Validation of a Novel 13-loci STR multiplex System for Cannabis sativa DNA Profiling". **Rachel Houston BS**, Sheree Hughes-Stamm PhD, David Gangitano PhD. American Academy of Forensic Sciences. New Orleans, LA. February 2017. (Oral Presentation)

- "HID & MPS for Post-blast bomb fragments and highly inhibited samples". Esiri Tasker, Kyleen Elwick, Bobby LaRue, Charity Beherec, Rachel Houston, David Gangitano, **Sheree Hughes-Stamm**. November 2016. Summit Forum of Forensic Technology and Applications, China Association for Forensic Science and Technology. Foshan, Guangzhou, China. Invited Speaker. (Co-author)

- "Developmental Validation of a Novel 13-loci STR multiplex System for Cannabis sativa DNA Profiling". **Rachel Houston BS**, Matthew Birck PhD, Sheree Hughes-Stamm PhD, David Gangitano PhD. International Symposium on Human Identification. Minneapolis, MN. September 2016. (Poster Presentation)

- "Bodies, Bones and Bombs; Human Identification". Esiri Tasker, Charity Beherec, Rachel Houston, **Sheree Hughes-Stamm**. Human Identification University Series. Office of the Chief Medical Examiner. NYC, NY. July 2016. (Co-author)

- "Bodies, Bones and Bombs; Human Identification". Esiri Tasker, Charity Beherec, Rachel Houston, **Sheree Hughes-Stamm**. 2nd Human Identification Solutions (HIDS) Conference. Barcelona, Spain. May 2016. (Co-author)

- "Evaluation of a 13-loci STR multiplex System for Cannabis sativa genetic Identification". **Rachel Houston BS**, Sheree Hughes-Stamm PhD, David Gangitano PhD. American Academy of Forensic Sciences. Las Vegas, NV. February 2016. (Oral Presentation)

- "Evaluation of a 13-loci STR multiplex System for Cannabis sativa genetic Identification". **Rachel Houston BS,** Matthew Birck PhD, Sheree Hughes-Stamm PhD, David Gangitano PhD. International Symposium on Human Identification. Grapevine, TX. October 2015. (Poster Presentation)

- "Evaluation of a 13-loci STR multiplex System for Cannabis sativa genetic Identification**". Rachel Houston BS**, Sheree Hughes-Stamm PhD, David Gangitano PhD. Association of Forensic DNA Analysts and Administrators meeting. Dallas, TX. July 2015. (Oral Presentation)

Other Products

- **Amy Sorensen**, **Rachel Houston**, Kyleen Elwick, **Sheree Hughes-Stamm**. How do modern quantification kits STACK-UP? June 2017. Forensic Magazine webinar sponsored by QIAGEN.

- Presentation at Customs and Border Protection Newark lab

- Webinar for Customs and Border Protection

Professional Affiliations

- American Academy of Forensic Sciences (AAFS) – General Member (2013 - current): Member number 147372

- Association of Forensic DNA Analysts and Administrators (AFDAA) – Student Member (2015 – current)

Awards

- 3 Minute Thesis – People's Choice Award (SHSU)
  2017

- LTC Michael A. Lytle '77 Academic Prize in Forensic Science Scholarship Fund (SHSU) 2015

Continuing Education

- Bloodborne and Airborne Pathogens

- OSHA Certification in Blood Bourne Pathogens and Laboratory Standard
- Globalfiler® and Quantifiler Trio® Training with Applied Biosystems
- Ion™ S5 Sequencing System Training with ThermoFisher Scientific
- RTI Training: Induction to Uncertainty in Forensic Chemistry and Toxicology
- RTI Training: SOP Writing for ISO 17025 Accreditation
- RTI Training: Answering the NAS: The Ethics of Leadership and the Leadership of Ethics
- RTI Training: To Hell and Back: The Ethics of Stewardship and the Stewardship of Ethics
- Advanced Word and Excel Training (SHSU)
- ASCLD DNA Mixtures Webinar Series: Managers overview
- Digital Next-Generation Sequencing for Targeted Enrichment, an Introduction to Technology