J Forensic Sci, May 2009, Vol. 54, No. 3
doi: 10.1111/j.1556-4029.2009.01014.x
Available online at: www.blackwell-synergy.com

Christopher Howard,[1] Ph.D.; Simon Gilmore,[2] Ph.D.; James Robertson,[3] Ph.D.; and Rod Peakall,[1] Ph.D.

A *Cannabis sativa* STR Genotype Database for Australian Seizures: Forensic Applications and Limitations*

ABSTRACT: A genetic database was established with the aim of documenting the genetic diversity of *Cannabis sativa* in Australia for future utilization in forensic investigations. The database consisted of genotypes at 10 validated short tandem repeat loci for 510 plants representing drug seizures from across Australia and 57 fiber samples. A total of 106 alleles and 314 different genotypes were detected. All fiber samples exhibited unique genotypes while 55% of the drug samples shared a genotype with one or more samples. Shared genotypes were mostly found within seizures; however, some genotypes were found among seizures. Statistical analysis indicated that genotype sharing was a consequence of clonal propagation rather than a lack of genetic resolution. Thus, the finding of shared genotypes among seizures is likely due to either a common supplier, or direct links among seizures. Notwithstanding the potential intelligence information provided by genetic analysis of *C. sativa*, our database analysis also reveals some present limitations.

KEYWORDS: forensic science, DNA typing, *Cannabis sativa,* short tandem repeat, genotype database, ANUCS301, ANUCS302, ANUCS303, ANUCS304, ANUCS305, ANUCS501, B01-CANN1, B05-CANN1, B02-CANN2, C11-CANN1

Varieties of the plant, *Cannabis sativa* L., have long been associated with human exploitation. *Cannabis sativa* is thought to have originated in the central Asia region, and has since been distributed worldwide by humans who have cultivated the plant as a source of fiber, fodder, oils, medicines, and intoxicants for thousands of years (1–4). Leaves and inflorescences contain psychoactive compounds collectively deemed cannabinoids, with Δ^9-tetrahydrocannabinolic acid (THC) being the most common (5). Despite the wide range of possible uses for *C. sativa*, due to its intoxicant properties, the cultivation and possession of the plant is prohibited by law in many countries.

Notwithstanding its prohibition in many jurisdictions, drug varieties of *C. sativa* typically characterized by elevated levels of THC (6), remain the world's most frequently used illicit drug (7). It is widely presumed that organized crime groups largely supply the domestic black market for *C. sativa*. However, despite this presumption, law enforcement agencies are often limited by their inability to link producers operating in suspected syndicates.

In some jurisdictions, licensing arrangements are available and advanced breeding schemes are actively cultivating low-THC varieties for fiber and seed oil industries (8–10). However, from an agricultural perspective, the inability to readily distinguish between fiber and drug *C. sativa* varieties based on morphology poses a major impediment to further development of the crop. In addition, from a law enforcement perspective, the full-scale agriculture of *C. sativa* for fiber and seed oil poses a security problem, with the possibility of licensed crops being used as a cover for illegal drug crops and the potential for theft and subsequent fraudulent distribution of agricultural types as drug types. Also, as long-distance dispersal of *C. sativa* pollen has been documented (11), there is the possibility of undetected contamination of fiber crops with pollen of drug varieties.

A wide range of botanical evidence is being increasingly used in forensic investigations. Historically, this has centered on the use of distinctive morphological characters of seeds and pollen (12); however, more recently, genetic techniques are increasingly being adopted for the identification of species from botanical evidence (13,14). The most commonly used type of genetic markers for discrimination between individuals in human forensic investigations, short tandem repeat (STR) markers (15), have recently been developed for *C. sativa* (16–19). The first comprehensive study employing a number of these STR markers provided information on *C. sativa* agronomic type, and the geographical origin of *C. sativa* drug seizures (18).

In the first study, to validate STR markers for forensic use in plants, Howard et al. (20) demonstrated consistent amplification of 10 STR loci in four multiplex reactions and showed that air-dried leaf tissue (easily obtainable from drug seizure samples) was particularly suitable as a DNA source. Crucial to the advancement of DNA analysis of *C. sativa*, these validated markers will enable routine DNA analysis in both forensic (21), and fiber variety breeding contexts (10,22). With validated STR markers in hand for *C. sativa*, the next step before these genetic markers can be meaningfully employed operationally is to establish a genetic database (23). The purpose of such a database is to provide insight into the patterns of allelic and genotypic variation within and among seizures or other

[1] School of Botany and Zoology, The Australian National University, Canberra ACT 0200, Australia.

[2] Centre for Forensic Science, Canberra Institute of Technology, GPO Box 826, Canberra ACT 2601, Australia.

[3] Forensic and Technical, Australian Federal Police, GPO Box 401, Canberra ACT 2601, Australia.

*Financial support through the receipt of a grant from the National Drug Law Enforcement Research Fund (NDLERF) and funds from the Australian Federal Police (AFP) made this work possible.

Received 28 Jan. 2008; and in revised form 2 June 2008; accepted 2 Aug. 2008.

© 2009 American Academy of Forensic Sciences

sample groups. This knowledge is critical for understanding the capability and limitations of genetic analysis of *C. sativa* for forensic applications.

This study builds on the earlier work of Gilmore et al. (18), Gilmore and Peakall (17), and subsequently Howard et al. (20), and describes the development of an Australian national genetic database for the forensic investigation of *C. sativa* based on STR markers. The aim of this study was to document both the allelic and genotypic diversity found at the 10 validated *C. sativa* STR loci for some 500 *C. sativa* samples representing both fiber and drug varieties. Sampling for the database included drug seizures from five states and territories of Australia and fiber varieties currently being evaluated for the hemp industry in Australia. The forensic insights provided by the analysis of this database will be discussed in relation to the nature of these samples. To our knowledge, this is the first genetic database in the world to be produced for validated STR profiles of *C. sativa*.

Methods

Sample Collection, DNA Extraction, and STR Genotype Scoring

Cannabis sativa drug samples were obtained from seizures from the following states and territories of Australia: the Australian Capital Territory (ACT); Victoria (VIC); South Australia (SA); Western Australia (WA); and Tasmania (TAS). Samples of hemp/fiber varieties of *C. sativa* were obtained from EcoFibre Industries (Toowoomba, Queensland, Australia). Drug samples consisted of plants that were grown using three different known methods: "field" refers to samples grown in the ground and/or in fields; "pot" refers to samples grown in pots or containers using artificial media or soil; "hydroponic" refers to samples grown using hydroponic equipment. Among the drug samples, hydroponically grown samples were most numerous (41%), followed by field-grown (30%) and pot-grown (25%).

In addition to the above samples for which cultivar type, Australian state of origin, and growth type was known, two sets (designated below as set 1 and set 2) of *C. sativa* samples were obtained. Set 1: consisted of a set of drug samples from multiple seizures from within the ACT for which the growing conditions were unknown. The seizures from which these samples originated were subsequently symbolized by a "?" character. Set 2: consisted of 13 *C. sativa* seedlings of uncertain cultivar type and origin, obtained as seed from the Australian Federal Police. These ambiguous samples in set 2 were included in analyses of total *C. sativa* only, but excluded then from calculations where cultivar type or state of origin was required. The *C. sativa* samples in set 1 and set 2 provided the opportunity to explore the population assignment procedures described below.

A total of 510 individual *C. sativa* samples were analyzed for a set of 10 *C. sativa* STR loci originally characterized by Alghanim and Almirall (16) and Gilmore and Peakall (17). The samples consisted of 440 known drug samples from 100 independent seizures and 57 known hemp/fiber samples from 12 independent groups (Table 1). For all *C. sativa* samples, PCR amplification and genotype scoring followed the multiplex PCR and allele scoring procedures in Howard et al. (20), using DNA extracted following Miller Coyle et al. (24).

Statistical Analysis of Genetic Data

The first step in the statistical analysis was to determine the number of multilocus genotypes present and whether any multilocus genotype sharing was evident among samples (hereafter we refer to "multilocus genotypes" simply as "genotypes"). Some sharing of genotypes was revealed by this analysis. This sharing may be attributed to either insufficient resolution of the genetic markers or clonal propagation of plants such that shared genotypes reflect a common clonal source. For the statistical analysis that follows, it was assumed that sharing of genotypes within a seizure most likely reflects a common clonal source, given the high frequency of clonal propagation of *C. sativa* (25). In this case, only one representative of the genotype per seizure was included in subsequent allele frequency-based analyses. Furthermore, it was assumed that any sharing of genotypes among seizures was independent and unrelated, such that replicated shared genotypes were retained among seizures. All statistical analysis was performed using the population genetic analysis software, GENALEX 6 (26), version 6.1, unless indicated otherwise.

TABLE 1—*Summary of the state of origin and nature of* Cannabis sativa *samples used in this study.*

Region	Cultivar Type	Growing Type	Number of Samples	Number of Seizures
Australian Capital Territory	Drug	Hydroponic*	36	4
		Field†	46	13
		Pot‡	73	7
		Unknown§	15	12
South Australia	Drug	Hydroponic*	82	13
		Field†	25	4
Victoria	Drug	Hydroponic*	29	15
		Field†	34	4
Western Australia	Drug	Hydroponic*	34	12
		Field†	28	3
		Pot‡	29	12
Tasmania	Drug	Pot‡	9	1
Unknown	Uncertain§	Unknown¶	13	1
–	Fiber		57	12
		Total	510	113

Samples were obtained from both drug seizures and licensed fiber varieties.

*Refers to samples grown using hydroponic equipment.

†Refers to samples grown in the ground and/or in fields.

‡Refers to samples grown in pots or containers using artificial media or soil.

§Cultivar type uncertain. Referred to as set 2 in the text.

¶Growing conditions unknown. Referred to as set 1 in the text.

Allele Frequency-Based Statistical Analyses

Allele frequency-based statistical analyses were performed at five levels: (i) The total data set of all *C. sativa* samples. (ii) All drug and fiber samples. (iii) Drug samples divided into field- (F), hydroponic- (H) and pot-grown (P) groups. (iv) Drug samples divided into Australian state of origin groups. (v) Drug samples divided into individual seizure groups. For each analysis level a range of standard population genetic statistics were calculated including: the number of alleles (Na), the number of effective alleles (Ne), observed heterozygosity (Ho), expected heterozygosity (He), and the fixation index (FI) for all 10 STR loci. These allele frequency-based statistics provide estimates of genetic diversity that can be compared among loci, among groups, and among species.

Hardy–Weinberg Equilibrium (HWE), and Linkage Disequilibrium (LD) tests were performed for each locus on all of the levels listed above (except level 5) using the software GENEPOP (27). As noted in Howard et al. (20), unlike human forensic DNA analysis where the assumption of random mating is closely approximated, it cannot be assumed this will be the case for *C. sativa* due to the ability to clonally propagate plants. Consequently, Mendelian segregation is avoided, resulting in identical genotypes between plants of clonal origin. Furthermore, measures of LD in domesticated plants often prove unreliable for inferring linkage given that the targeted selection of some phenotypic characters often impose a bias (28). Clonal reproduction has been shown to further bias LD estimates (28).

Following Gilmore et al. (18), an Analysis of Molecular Variance (AMOVA) was performed to separately estimate the degree of genetic differentiation among fiber and drug samples, among state of origin of drug samples, and among growth type groups of drug samples.

Population Assignment

Population assignment tests were employed to assess the ability to correctly assign a sample of *C. sativa* to either drug or fiber type, based only on its genotype. Following the recommendation of Paetkau et al. (29) for predicting the statistical power of assignment tests, genotype log likelihood (log [*L*]) biplots were plotted for the drug and fiber sample groups. In such biplots, a strong indication of sufficient statistical power to correctly assign a sample is indicated when the two populations form discrete nonoverlapping clusters (29). Genotype likelihood biplots for *C. sativa* drug samples were also generated for drug growth type (hydroponically-, field-, or pot-grown) and the Australian state of origin. Generation of these plots and standard population assignment tests were performed using GENALEX.

Subsequently, using GENECLASS V2 (30) the novel Monte Carlo re-sampling method of Paetkau et al. (31) was employed to perform assignment tests and estimate probabilities of inclusion. Population assignment based on log (*L*) values and the simulation-based assignment tests were performed for two types of "unknown" samples: (i) A random subsample of 24 drug and five fiber samples was removed from the full database (representing *c.* 10% of the original group's size) and treated as the "unknown" group. With these samples excluded from frequency calculations, it was then determined whether they were correctly assigned as drug or fiber types based on the log (*L*) values and the outcomes of simulation testing. This process was repeated for five replicates (representing 145 "unknown" samples in total). (ii) The 13 *C. sativa* seeding samples of uncertain cultivar type (previously designated as "set 2"), with the remaining samples in the database as the reference drug and fiber populations.

Match Probabilities

Random match probability (RMP), probability of identity (PI) and probability of identity sibs (PIsibs) were calculated using GENALEX by the formulas shown below:

$$\mathrm{RMP} = \prod p_i^2 \times \prod 2p_ip_j$$

where for a specific multilocus genotype within a given population, Π indicates the chain multiplication across loci, p_i is the frequency of the *i*-th allele at homozygous loci and p_i and p_j are the frequencies of the *i*-th and *j*-th alleles at heterozygous loci.

$$\mathrm{PI} = 2(\sum p_i^2)^2 - \sum p_i^4$$

where for a single locus, p_i is the frequency of the *i*-th allele at the locus for the population in question. The PI over multiple loci is calculated as the product of the individual locus PI values.

$$\mathrm{PIsibs} = 0.25 + (0.5\sum p_i^2) + (0.5(\sum p_i^2)^2) - (0.25\sum p_i^4)$$

where for a single locus, p_i is the frequency of the *i*-th allele at the locus for the population in question. The PIsibs over multiple loci is calculated as the product of the individual locus PIsibs values.

Random match probablity provides an estimate of the probability of encountering a specific genotype in the population in question (32,33). PI estimates the probability that two unrelated individuals drawn at random from the population will have the same genotype, while PIsibs estimates the probability of identity taking into account the potential relatedness of samples (34,35). Note that PI and PIsibs estimate the average probability of a match for any genotype, rather than for a specific genotype as is the case for RMP. Despite the likely violation of the random mating assumption in *C. sativa*, these measures offer useful comparative statistics among *C. sativa* samples and populations.

Results

Genotype Recovery

A total of 314 genotypes were detected over the 10 STR loci examined for all *C. sativa* samples. All 57 fiber samples had a unique genotype while among the 440 known drug samples, 197 genotypes were unique, with 47 genotypes being shared across the remaining 243 samples (i.e., 440–197) (Fig. 1). The drug seizures from within the ACT from which growth type was unknown (set 1) included mostly unique genotypes but also some that were shared between these ACT seizures and among seizures from different states (see below). The 13 seedling samples from set 2 each had a unique genotype.

Figure 2 shows the number of different genotypes resolved for increasing combinations of loci, ordered from most to least informative. For the fiber samples, all 57 genotypes were resolved with the combination of only three loci. For the drug samples, including genotype matches within seizures, the number of unique genotypes that were resolved started to plateau with the combination of seven loci and did not change beyond nine loci (Fig. 2).

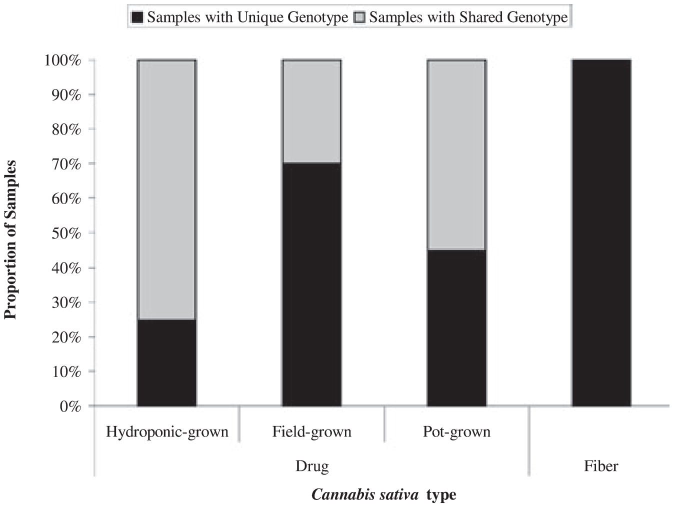

FIG. 1—*Patterns of genotype sharing among* Cannabis sativa *samples. The proportion of samples with unique versus shared genotypes for both* C. sativa *variety and drug growth type are shown.*

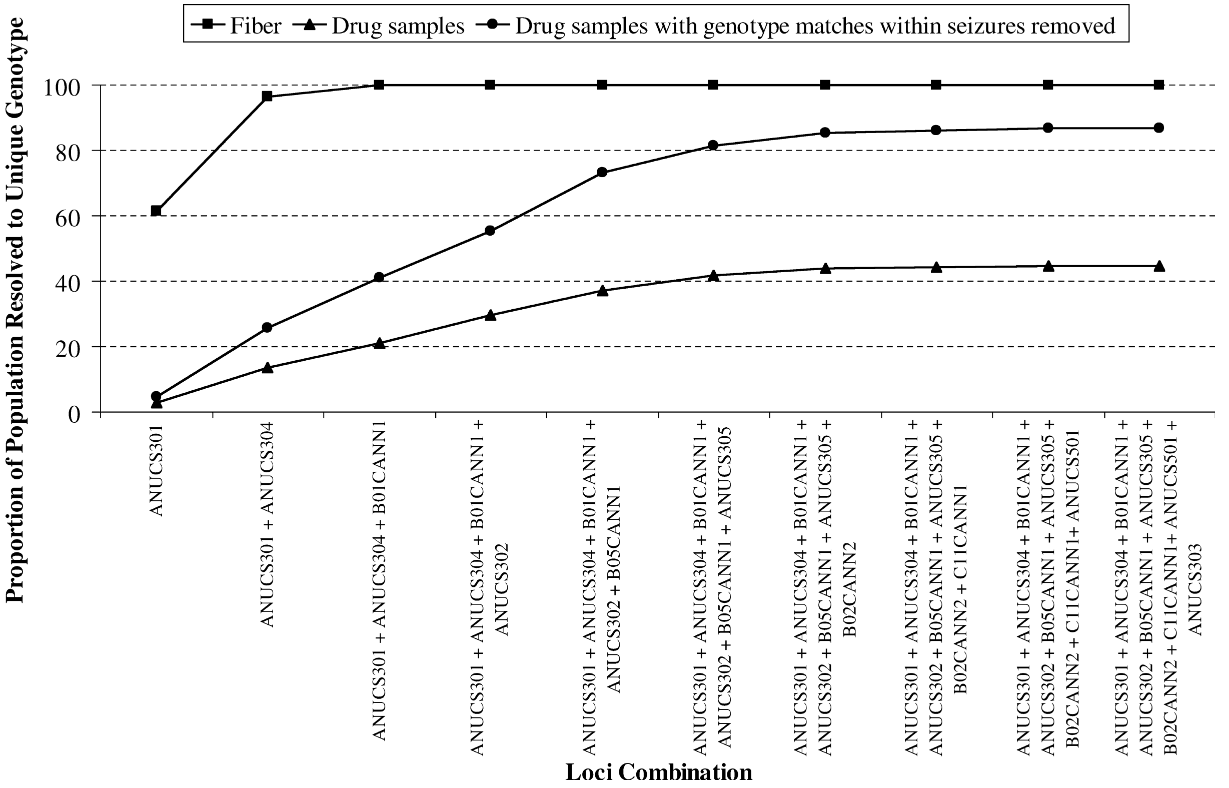

FIG. 2—*Multilocus genotype resolution over 10 short tandem repeat loci showing the proportion of fiber and drug samples resolved to a unique genotype for increasing combinations of loci.*

Genotypic Patterns

Multiple occurrences of the same genotype were common within seizures consisting of multiple plants and were more frequent within rather than among seizures. In total, 38 of the 47 shared genotypes were only found within a single seizure. Shared drug genotypes were most frequently found within hydroponically grown samples (57% of the total) while unique drug genotypes were mostly found in field-grown samples (49% of the total) (Fig. 1). Despite the removal of shared genotypes from the analysis, as expected, for most loci there was significant deviation from HWE, and some LD was evident (full data not shown).

Nine of the 47 shared genotypes were found among seizures, with three of these being present in seizures from two or more states, denoted genotypes *F*, *M*, and *N* (Fig. 3*a* and 3*b*). Seizures of hydroponically grown samples from SA had a high degree of genotype sharing, with seven of the 13 seizures of hydroponically grown samples from SA sharing the same genotype, denoted *P*. Five of these seven seizures were exclusively genotype *P*. Victorian hydroponic seizures also showed similar levels of genotype sharing within and among independent seizures, with six of the 15 independent hydroponic seizures consisting exclusively of the genotype *F*. Genotype *F* was also found in several independent hydroponic seizures from SA and in one unknown growth type seizure from the ACT. The remaining genotypes shared within states, including the two genotypes shared between states (*M*, shared between WA and an unknown growth type seizure from the ACT; *N* shared between VIC, WA, and an unknown growth type seizure from the ACT), were not found in as high abundance between independent seizures as that of genotypes *F* and *P*.

The average RMP estimate for all recovered drug genotypes was 5.4×10^{-8} with a range of 9.6×10^{-7} to 9.5×10^{-20}. The RMP estimate for all *C. sativa* genotypes recovered was 5.0×10^{-9} with a range of 9.6×10^{-8} to 3.1×10^{-25}. The RMP estimates for the shared genotypes: *BB*, *EE*, *K*, *N*, and *P*, were notably smaller than the average RMP for the drug samples, which suggests that rare alleles were present in these genotypes. The RMP estimates for the remaining shared genotypes: *B*, *F*, *M*, and *Z*, were larger than the average RMP for the drug samples, which suggests that these genotypes were composed of more common alleles. The PI and PIsibs for all drug genotypes recovered were estimated to be 2.4×10^{-8} and 5.5×10^{-4} respectively, and 2.3×10^{-9} and 3.1×10^{-4} respectively for all *C. sativa* genotypes recovered.

Allelic Diversity in Cannabis sativa

A total of 106 alleles were detected over all 10 STR loci for the 510 *C. sativa* samples. Within the drug samples, 76 alleles were detected of which 14 were unique to the drug type of *C. sativa*. Within the fiber samples, 92 alleles were detected with 30 being unique to only the fiber type of *C. sativa*. Overall, the number of alleles per locus ranged from 23 (ANUCS301) to 4 (ANUCS501 and B02-CANN1).

On average over the 10 STR loci, the fiber group revealed considerably more alleles than the drug sample group. In turn, unique alleles were more common in fiber samples. The average Na, average Ne, and the average number of unique alleles were similar for the field-, hydroponic-, and pot-grown drug growth type groups. However, the average He was considerably lower for the hydroponic drug group. Allelic diversity was also variable among the state drug growth groups. At a locus by locus level there was variation in the Na and the frequency of alleles among the drug growth groups, with the average Na for the ACT and WA drug groups being similar and higher than the average number of alleles for

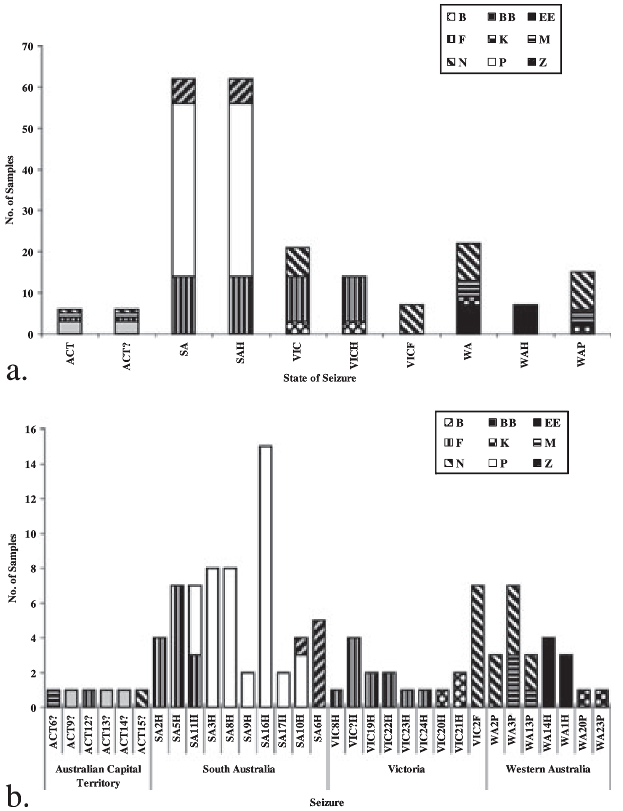

FIG. 3—*The distribution of shared multilocus genotypes among seizures. a) All except three of the genotypes shared among seizures were found within one state. b) Genotypes* F, N, *and* M *were shared between states.*

VIC and SA drug populations. The average He was the highest for the ACT and WA drug groups, with a considerable decrease in this measure within the SA, VIC, and TAS groups. For most loci, allelic distribution and frequency was uneven among the drug and fiber groups and also within drug growth type groups as well as among states.

Tables showing the full list of genotypes and summarizing the allele frequency data can be freely sourced from Howard et al. (36) at http://www.ndlerf.gov.au/pub/Monograph_29.pdf.

Ability to Distinguish between Fiber and Drug Samples

The AMOVA analysis revealed that there was modest, yet significant, genetic differentiation ($F_{ST} = 0.094$ $p > 0.001$) between the fiber and drug samples, with this difference accounting for 9% of the total genetic variance. This was higher than the level of differentiation detected between drug and fiber samples reported in Gilmore et al. (18), where a different subset of *C. sativa* STRs were used. Within the drug samples, the degree of genetic differentiation among the state of origin groups was similar to that among the fiber and drug groups ($F_{ST} = 0.077$, $p > 0.001$); however, the degree of genetic differentiation among the drug growth type groups was lower ($F_{ST} = 0.041$, $p > 0.001$).

Despite the modest differentiation among drug and fiber samples (representing only 9% of the total genetic variation), there was minimal overlap between the two types of *C. sativa* in the genotype likelihood biplot (Fig. 4). This indicated the potential for assignment tests to aid identification of unknown *C. sativa* as either drug or fiber type samples. Table 2 summarizes the outcomes of assignment tests for the subset of samples that were randomly extracted from the database and excluded from the frequency calculation underpinning the subsequent assignment tests. For the assignment test based on log (*L*) values, on average 92% of the drug subset samples were correctly identified as drug, while 100% of the fiber subset were correctly identified as fiber. When based on the simulation outcomes, at $p > 0.01$ for inclusion, 89% of drug samples and 92% of fiber samples were assigned correctly to their respective group. However, for the same set of samples, 65% of the drug samples could not be ruled out as possibly belonging to the fiber group. Similarly, 8% of the fiber samples could not be ruled out as belonging to the drug group (Table 2).

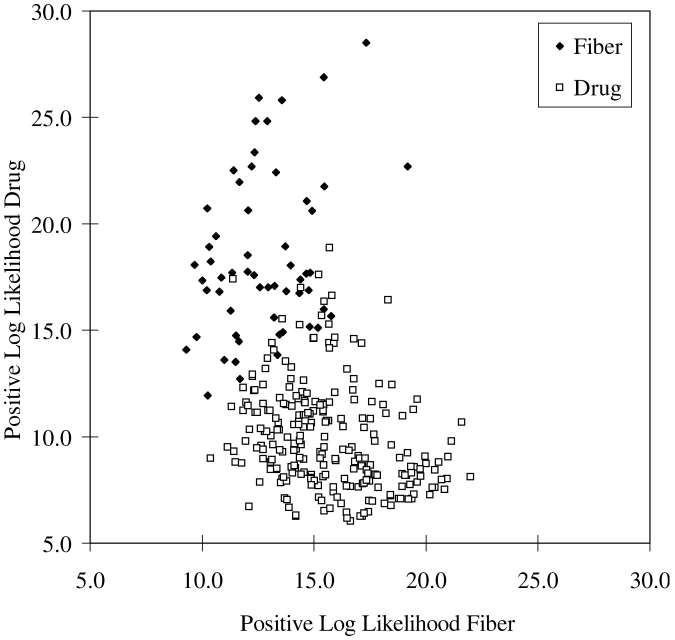

FIG. 4—*Genotype likelihood biplot showing the discrimination between drug and fiber samples.*

Population assignment tests based on log (*L*) values for the 13 *C. sativa* seedling samples from set 2 revealed that nine samples had a genotype indicative of drug type samples, with the remaining four indicative of fiber type samples. However, the simulation outcomes revealed that none of the samples could be ruled out as belonging to the drug type (at $p > 0.01$ for inclusion). This ambiguous assignment outcome for the set 2 samples that most likely originated from a drug seizure by the Australian Federal Police, likely reflects the genetic similarity of some drug and fiber samples (Fig. 4).

Despite some genetic differentiation, discrete clustering in the genotype likelihood biplots was not apparent among Australian state of origin groups. In addition, given the low level of genetic differentiation separating the drug growth type groups, genotype likelihood biplots among these groups did not show discrete nonoverlapping clusters (data not shown). As a result, the drug samples from set 1 could not be unambiguously assigned to a growth type. Therefore, using this current database, there is limited genetic power to distinguish among these groups with assignment tests.

Discussion

Genetic Diversity of Australian Cannabis sativa

To our knowledge, we have built the world's first *C. sativa* genetic database. Based on the genetic analysis of STR loci, the

TABLE 2—*Results of population assignment tests for drug and fiber samples of* Cannabis sativa.

C. sativa Type Population	Random *C. sativa* Sample Subset	log (*L*)—Placement in Actual Group (%)	Simulated Probability of Inclusion			
			p > 0.01—Drug (%)	*p* > 0.01—Fiber (%)	*p* > 0.001—Drug (%)	*p* > 0.001—Fiber (%)
Drug	1	92	79	71	92	75
	2	92	92	58	96	83
	3	100	96	58	100	63
	4	88	88	67	92	75
	5	88	92	71	92	79
	Average	92	89	65	94	75
Fiber	1	100	0	80	20	80
	2	100	20	100	20	100
	3	100	20	80	60	80
	4	100	0	100	0	100
	5	100	0	100	40	100
	Average	100	8	92	28	92

The proportion of samples placed in their correct population are indicated from log likelihood (log [*L*]) values and simulated probability of inclusion.

current standard in human forensic analysis (15), the database contains genotype data across 10 loci for some 500 *C. sativa* plants representing drug seizures from five Australian states and territories and a selection of fiber samples. While additional STR loci are available for *C. sativa*, and have been used successfully for population studies (18), the selection of the 10 loci used in this study was based on the need to use developmentally validated STR loci that most closely matched the standards in human forensic analysis and avoided many of the interpretive challenges common with STRs (37,38).

Concurring with the study of Gilmore et al. (18), the analysis of the present database revealed that fiber varieties were genetically more diverse than drug varieties of *C. sativa*. For example, while fiber samples represented only 11% of the 510 samples tested, these samples contained 86% of the total allelic diversity. Furthermore, 28% of the total of 106 alleles were only found in fiber samples. Moreover, all of the fiber samples tested had unique genotypes. This finding of high genetic diversity within the fiber samples is consistent with obligate outcrossing and long-distance wind-dispersed pollen that likely characterizes this dioecious plant (10). It is also apparent that a wide genetic base has been sourced by the hemp industry.

Despite their lower genetic diversity when compared with fiber samples, a high proportion of drug samples did exhibit a unique genotype across the 10 STR loci. These genetically distinct samples were found among field-, hydroponic-, and pot-grown drug samples, but were most frequent in field-grown samples. Of the total of 106 alleles, 13% of the alleles detected were unique to the drug samples.

Genotypic Patterns among Australian Cannabis sativa

Unique genotypes were common among the Australian *C. sativa* samples that were analyzed, with genotype sharing occurring only among the drug samples. The finding of genotype sharing among some drug samples, and the lack of any genotype sharing among the fiber samples is of interest. The challenge in the case of *C. sativa* (and many other plants) is that unlike humans (except identical twins), some genotype sharing due to clonal propagation can be expected. However, this genotype sharing may also be due to lack of sufficient resolution at the set of 10 STR loci used in the study.

One way to assess whether these 10 STR markers provide sufficient resolution is to empirically determine the rate at which unique genotypes are recovered with increasing combinations of loci within the database itself. This analysis revealed that for the genetically more diverse fiber samples the combination of three or four loci was more than sufficient to "individualize" all of the 57 genotypes (see Fig. 2). For the less diverse drug samples, most unique genotypes were recovered with 7 or 8 loci, with subsequent additional loci failing to find substantial numbers of extra genotypes.

Probability of identity estimates provides another way to assess whether the 10 STR loci provided adequate resolution. The PI estimates indicated that the chance of obtaining identical genotypes by sexual reproduction in a randomly mating population of *C. sativa* is approximately one in 400 million. However, given the violation of the random mating assumption the more conservative PIsibs is recommended. Estimates of PIsibs indicated that the probability of two samples, including genetically related samples, having the same identical genotype was in the order of one in 3000. Therefore, in this database of some 500 samples, encountering shared genotypes by chance, even allowing for closely related individuals, appears very unlikely. Consequently, the finding of shared genotypes is most likely due to a common genetic origin enabled by clonal propagation.

Further support for clonal propagation as the basis for genotype sharing is indicated by the patterns of genotype sharing. If genotype sharing was a consequence of insufficient genetic resolution it should be found evenly across the samples, irrespective of their growth type. However, the majority of samples with shared genotypes (57%) occurred within hydroponic seizures (Fig. 1), the growth type for which clonal propagation is known to be most frequent (25). In addition, the overwhelming majority of shared genotypes, 38 of 47 (81%), were detected within seizures. Of the remaining nine genotypes shared among seizures, all but three were exclusive to a single Australian state. On the weight of evidence it is therefore concluded that the genotype sharing detected in the database is predominantly, if not exclusively, a consequence of clonal propagation. Below the forensic implications of this finding are explored.

Forensic Applications and Limitations

The construction of the genetic database and associated analysis was completed "blind" with the only information provided with the samples being the varietal type of *C. sativa*, the state of origin, and (where known) the growth type of the drug samples (hydroponic-, pot-, or field-grown). Other information, such as known or suspected linkages among seizures, was not provided. Such additional knowledge would allow a better assessment of the forensic value of the database. However, in the absence of this information, the comments below on the forensic applications remain somewhat speculative.

The patterns of genotype sharing that were uncovered in the database suggest some variation in the form of drug production within Australia. It is inferred that the production consists of two types of perpetrator: (i) Small independent growers using a combination of field-, pot- and hydroponic-growth methods that mainly generate unique genotypes. (ii) Organized crime syndicates of a variety of operational size that largely employ hydroponic propagation, leading to the proliferation of shared genotypes that reflect either a common supplier, or direct links among seizures.

One shared genotype of interest was genotype *P* (Fig. 3*b*) that was exclusive to South Australian hydroponic samples and found among several seizures. The RMP value for this genotype was approximately two orders of magnitude lower than the average RMP, indicating that multiple occurrences of this genotype by sexual reproduction are particularly unlikely. Consequently, linkages among the seizures are implied. Similarly, other cases of potential linkage are implied by genotype sharing among the states (Fig. 3). If this genetic knowledge reinforces suspected linkages from other evidence, this combined knowledge may aid in prosecution.

Notwithstanding the potential intelligence information provided by genetic analysis of *C. sativa* drug seizures, it is presently not possible to categorically assign a state of origin to an Australian seizure. As already noted, there is some sharing of genotypes among states, and this likely underestimates the degree of human-assisted gene flow that occurs between the states. Nonetheless, there were state-by-state differences in alleles and allele frequency that may become even more pronounced as the database expands. It is possible that *C. sativa* drug seizures from other countries may exhibit more informative differences than among states within Australia (18) but this analysis was beyond the scope of the present study.

The genetic similarity that was identified among fiber and drug varieties undoubtedly reflects their common evolutionary origin, but poses several challenges for the law enforcement community. The combination of low genetic diversity within drug samples and the presence of unique fiber- and drug-specific alleles has the potential to provide a strong indication as to whether a sample is of drug or fiber origin. Furthermore, notwithstanding only moderate genetic differentiation, the assignment tests identified a large proportion of the samples correctly (on average >92% for drug samples and 100% for fiber samples, Table 2).

Ideally, a DNA test for drug versus fiber varieties of *C. sativa* would be based on the direct analysis of the gene/s responsible for THC regulation. Until such a test is available the combination of nuclear STR data with organelle DNA haplotype data may further enhance discrimination among fiber and drug varieties of *C. sativa* (39). A further solution to aid the identification of drug versus fiber plants may be a DNA profile register of fiber varieties, analogous to the DNA registers proposed to assist with the legal trafficking of wildlife (40).

Given the identified limitations, most of which reflect biological reality, rather than technical constraints, what practical recommendations can be made? The detection of genotype sharing among multiple drug seizures may provide objective and independent corroboration of suspected linkages. Equally, this genetic evidence may refute suspected linkages. With appropriate consideration, there will be a range of circumstances where genetic analysis of *C. sativa* seizures will be of forensic value, be it for prosecutor or defense assistance in drug-related crime or for intelligence gathering for other investigations. However, as in human forensics, genetic analysis must complement, rather than replace, other forms of evidence (41). With the establishment of this first *C. sativa* genetic database, the next step in the implementation of *C. sativa* DNA typing can now be handed to established forensic laboratories. Ultimately the final step will be realized when this technology is evaluated in the courtroom.

Acknowledgments

For assistance with *Cannabis sativa* sampling and DNA extractions, we thank the following people and organizations: Dennis Pianca, Dr. Julieanne Dougherty, and Daniel Andres of the ACT Government Analytical Laboratory, Toxicology, and Forensic Chemistry Unit, who assisted with sampling drug seizures from within the Australian Capital Territory, and provided laboratory access for DNA extractions from these samples; Susan Fiddian of the Botany Branch, Victoria Police Forensic Services Department, for access to *C. sativa* drug seizures from within Victoria; Tanya McKew of Forensic Science South Australia, for providing us with DNA from drug seizures within South Australia; Colin Priddis, Hannah Crisp, and Dr. Kevin Ho of the Forensic Science Laboratory, Chemistry Centre Western Australia, for the supply of DNA from drug seizures within Western Australia; Dr. Michael Manthey and Carl Grosser of the Forensic Science Service Tasmania, for the supply of DNA from seizures within Tasmania; The Management of EcoFibre Industries Ltd, Queensland, Australia, for the supply of fiber varieties, and the Australian Federal Police for laboratory access to extract DNA from these samples. We thank Elizabeth McKeown and Meg Malaika for providing technical assistance in the laboratory and Christine Hayes and Dr. Daniel Ebert for technical advice. Finally, we dedicate this work to the memory of Meg Malaika, who started this project but tragically was unable to see its completion.

References

1. Abel EL. Marihuana: the first twelve thousand years. New York: Plenum Press, 1980.
2. Grispoon L, Bakalar JB. Marihuana, the forbidden medicine. New Haven: Yale University Press, 1993.
3. Mercuri AM, Accorsi CA, Mazzanti MB. The long history of *Cannabis* and its cultivation by the Romans in central Italy, shown by pollen records from Lago Albano and Lago di Nemi. Veg Hist Archaeobot 2002;11(4):263–76.
4. Small E, Cronquist A. A practical and natural taxonomy for *Cannabis*. Taxon 1976;25(4):405–35.
5. de Zeeuw RA, Malingre TM, Merkus WHM. Tetrahydrocannabinolic acid, an important component in evaluation of *Cannabis* products. J Pharm Pharmacol 1972;24(1):1–6.
6. Pacifico D, Miselli F, Micheler M, Carboni A, Ranalli P, Mandolino G. Genetics and marker-assisted selection of the chemotype in *Cannabis sativa* L. Mol Breed 2006;17(3):257–68.
7. Anderson P. Global use of alcohol, drugs and tobacco. Drug Alcohol Rev 2006;25(6):489–502.
8. van der Werf HMG, Mathijssen E, Haverkort AJ. The potential of hemp (*Cannabis sativa* L.) for sustainable fibre production: a crop physiological appraisal. Ann Appl Biol 1996;129(1):109–23.
9. Struik PC, Amaducci S, Bullard MJ, Stutterheim NC, Venturi G, Cromack HTH. Agronomy of fibre hemp (*Cannabis sativa* L.) in Europe. Ind Crops Prod 2000;11(2–3):107–18.
10. Ranalli P. Current status and future scenarios of hemp breeding. Euphytica 2004;140(1-2):121–31.
11. Cabezudo B, Recio M, Sanchez-Laulhe JM, Del Mar Trigo M, Toro FJ, Polvorinos F. Atmospheric transportation of marijuana pollen from North Africa to the southwest of Europe. Atmos Environ 1997;31(20): 3323–8.
12. Miller Coyle H, Ladd C, Palmbach T, Lee HC. The green revolution: botanical contributions to forensics and drug enforcement. Croat Med J 2001;42(3):340–5.

13. Craft KJ, Owens JD, Ashley MV. Application of plant DNA markers in forensic botany: genetic comparison of *Quercus* evidence leaves to crime scene trees using microsatellites. Forensic Sci Int 2007;165(1):64–70.
14. Ward J, Peakall R, Gilmore SR, Robertson J. Molecular identification system for grasses: a novel technology for forensic botany. Forensic Sci Int 2005;152(2–3):121–31.
15. Butler JM. Genetics and genomics of core short tandem repeat loci used in human identity testing. J Forensic Sci 2006;51(2):253–65.
16. Alghanim HJ, Almirall JR. Development of microsatellite markers in *Cannabis sativa* for DNA typing and genetic relatedness analyses. Anal Bioanal Chem 2003;376(8):1225–33.
17. Gilmore S, Peakall R. Isolation of microsatellite markers in *Cannabis sativa* L. (marijuana). Mol Ecol Notes 2003;3(1):105–7.
18. Gilmore S, Peakall R, Robertson J. Short tandem repeat (STR) DNA markers are hypervariable and informative in *Cannabis sativa*: implications for forensic investigations. Forensic Sci Int 2003;131(1):65–74.
19. Hsieh H-M, Hou R-J, Tsai L-C, Wei C-S, Liu S-W, Huang L-H, et al. A highly polymorphic STR locus in *Cannabis sativa*. Forensic Sci Int 2003;131(1):53–8.
20. Howard C, Gilmore S, Robertson J, Peakall R. Developmental validation of a *Cannabis sativa* STR multiplex system for forensic snalysis. J Forensic Sci 2008;53(5):1061–7.
21. Miller Coyle H, Palmbach T, Juliano N, Ladd C, Lee HC. An overview of DNA methods for the identification and individualization of marijuana. Croat Med J 2003;44:315–21.
22. Mandolino G, Carboni A. Potential of marker-assisted selection in hemp genetic improvement. Euphytica 2004;140(1-2):107–20.
23. Foreman LA, Champod C, Evett IW, Lambert JA, Pope S. Interpreting DNA evidence: a review. Int Stat Rev 2003;71(3):473–95.
24. Miller Coyle H, Shutler G, Abrams S, Hanniman J, Neylon S, Ladd C, et al. A simple DNA extraction method for marijuana samples used in amplified fragment length polymorphism (AFLP) analysis. J Forensic Sci 2003;48(2):343–7.
25. ACC. Illicit drug data report 2005-06. Canberra: Australian Crime Commission, 2007.
26. Peakall R, Smouse PE. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes 2006;6(1):288–95.
27. Raymond M, Rousset F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Hered 1995;86(3):248–9.
28. Flint-Garcia SA, Thornsberry JM, Buckler ES. Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 2003;54:357–74.
29. Paetkau D, Slade R, Burden M, Estoup A. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. Mol Ecol 2004;13(1):55–65.
30. Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A. GENECLASS2: a software for genetic assignment and first-generation migrant detection. J Hered 2004;95(6):536–9.
31. Paetkau D, Calvert W, Stirling I, Strobeck C. Microsatellite analysis of population structure in Canadian polar bears. Mol Ecol 1995;4(3):347–54.
32. Samuels JE, Asplen C. The future of forensic DNA testing: predictions of the research and development working group. Washington, DC: National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, 2000.
33. National Research Council. The evaluation of forensic DNA evidence. Washington, DC: National Academy Press, 1996.
34. Waits LP, Luikart G, Taberlet P. Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. Mol Ecol 2001;10(1):249–56.
35. Buckleton J, Triggs CM. Relatedness and DNA: are we taking it seriously enough? Forensic Sci Int 2005;152(2–3):115–9.
36. Howard C, Gilmore S, Robertson J, Peakall R. Application of new DNA markers for forensic examination of *Cannabis sativa* seizures—developmental validation of protocols and a genetic database. Monograph 29. Hobart: National Drug Law Enforcement Research Fund, 2008.
37. Hoffman JI, Amos W. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Mol Ecol 2005;14(2):599–612.
38. Hauge XY, Litt M. A study of the origin of ''shadow bands'' seen when typing dinucleotide repeat polymorphisms by the PCR. Hum Mol Genet 1993;2(4):411–5.
39. Gilmore S, Peakall R, Robertson J. Organelle DNA haplotypes reflect crop-use characteristics and geographic origins of *Cannabis sativa*. Forensic Sci Int 2007;172(2–3):179–90.
40. Palsboll PJ, Berube M, Skaug HJ, Raymakers C. DNA registers of legally obtained wildlife and derived products as means to identify illegal takes. Conserv Biol 2006;20(4):1284–93.
41. Lynch M, McNally R. ''Science,'' ''common sense,'' and DNA evidence: a legal controversy about the public understanding of science. Public Underst Sci 2003;12(1):83–103.

Additional information and reprint requests:
Professor Rod Peakall, Ph.D.
School of Botany and Zoology
The Australian National University
Canberra ACT 0200
Australia
E-mail: rod.peakall@anu.edu.au