FauxDIS: A Searchable Forensic DNA Database to Support Experiential Learning

Jacqueline Baranski[1], Karen Davalos-Romero BS[1], Melanie Blum BS[2], Nichole Burke BS[2], Ashley Foster MS[2], Ashley Hall PhD[1*]

[1]University of Illinois at Chicago, Department of Biopharmaceutical Sciences, 833 S. Wood St, Chicago, IL 60612 USA.*corresponding author: amhall7@uic.edu

[2]University of Nebraska-Lincoln, Forensic Science Degree Program, Department of Entomology, PO Box 830816, Lincoln NE 68583-0816 USA

Abstract: CODIS (Combined DNA Index System) is the generic term used to describe the system of U.S. criminal justice DNA databases administered at the local, state and national levels. As of December 2019, the national database contained over 14 million genetic profiles. Access is restricted, however, to authorized government agencies and the database cannot be used in learning exercises. Therefore, we have initiated the construction of a DNA profile database modeled after CODIS and intended for use by educators. The FauxDIS DNA Database is a tool that can be used as part of experiential learning exercises in which students apply the scientific method to solve mock crimes. The DNA profiles generated from collected evidence are searched against the known profiles contained in FauxDIS and statistics applied to quantify the power of an identification. The database currently contains 151 DNA profiles. To generate these profiles, we have developed a work flow analogous to those employed in U.S. operational forensic laboratories. The use of expensive commercial kits has been avoided, making the methods cost-effective and easily transferrable to other laboratories. The FauxDIS DNA Database is available for use by educators in exchange for the submission of novel profiles or unique samples to be profiled. The goal is to encourage national and international collaboration leading to the establishment of a learning network.

Keywords: forensic, DNA, CODIS, database, experiential learning

Introduction

The collection and analysis of evidence from a crime scene is the scientific method in action. Investigators make observations, formulate hypotheses about the probative value of potential evidence, and test these educated guesses by submitting samples to an operational forensic laboratory for analysis. In the case of biological evidence, analysts extract DNA from an unknown (crime scene) sample, quantify the nucleic acid, amplify it by polymerase chain reaction (PCR), and generate a profile. The unknown DNA profiles can be searched against a database of known (reference) samples for the purpose of identification. Translating these processes to the teaching laboratory would create opportunities in experiential learning. To meet this goal, the objectives of the work described here are: 1) construct a tool that educators can apply to reinforce the scientific method through experiential learning; 2) challenge students to demonstrate proficiency in the molecular biological techniques central to forensic DNA analysis; and 3) connect educators in a learning network.

The utility of forensic DNA profiling-based modules in undergraduate teaching laboratories has been demonstrated in earlier work (1-4). In a series of exercises, students generated two- to three-locus forensic profiles using commercial kit-based assays coupled with native agarose or capillary electrophoresis for PCR product visualization. The modules not only introduced students to the biochemical principles applied in forensic analysis but also instructed them in the meticulous laboratory technique required for biological investigations. We have developed an educational module that expands upon this to include thirteen CODIS STR loci plus a sex-informative locus. DNA profiling is completed on a capillary electrophoresis unit, and the resulting genotype is searched against our newly developed DNA database, the function of which was inspired by the U.S. national DNA database.

CODIS (Combined DNA Index System) is the general term used to describe the system of U.S. criminal justice DNA databases administered at the local, state and national level (5, 6). The national arm of this database, NDIS (National DNA Index System), contains over 14 million offender profiles as of December 2019 (https://www.fbi.gov/services/laboratory/biometric-analysis/codis/ndis-statistics). An unknown (crime scene) sample is searched against the database with the goal of identifying a known (reference) sample. These activities are restricted to authorized government labs and cannot be

used as experiential learning tools for students. Therefore, we have initiated the construction of a DNA profile database modeled after CODIS and intended for use by educators - FauxDIS.

The FauxDIS DNA Database currently contains 151 genetic profiles, each comprising fifteen short tandem repeat (STR) markers and one sex-informative insertion/deletion. It is a tool that can be used as part of experiential learning exercises in which students apply the scientific method to solve a mock crime by generating DNA profiles and searching the database of known samples. Statistics can be applied to quantify the power of identifications. The reagents and instrumentation used to construct and maintain the FauxDIS database are analogous to those employed in U.S operational crime laboratories, challenging students to demonstrate proficiency in the molecular biological techniques central to forensic DNA analysis.

With the FauxDIS established, the final project objective is the creation of a learning network of educators, both domestic and international. Collaborating educators will be given access to FauxDIS in exchange for the submission of complete, novel, single-source DNA profiles to the database. The source DNA samples can then be incorporated into the educators' on-site experiential learning activities, culminating in a profile search in the FauxDIS database that is successful due to the prior deposit of those genotypes. Increasing the size of the database by including collaborator contributions will not only foster this alliance and increase the overall impact of the project, but will strengthen individual probability calculations based on direct observation (described below) by decreasing allele frequencies.

In this report, we introduce the FauxDIS DNA Database and demonstrate its function. We describe the development of a cost-effective analysis procedure, the goal of which is to increase the accessibility of the laboratory exercises to other colleges and universities. We have avoided the use of expensive commercial kits by: 1) adapting a procedure for the expression and purification of *Taq* DNA polymerase; 2) extracting DNA with a standard phenol:chloroform protocol; 3) quantifying DNA using a published SYBR Green method; and 4) developing an in-house multiplex PCR primer mix.

Methods

Sample Collection

Samples from human subjects were collected with informed consent using the University of Illinois at Chicago protocol 2016_0431. Sterile cotton-tipped swabs (Puritan, Guilford, ME) were used to collect buccal samples from donors. The swabs were dried in a room temperature hood overnight and stored in a sealed paper envelope at 4°C until use.

Taq DNA Polymerase

Taq is a heat-resistant enzyme produced by the thermophilic microorganism *Thermus aquaticus*. This microbe thrives in extreme environments such as hot springs and hydrothermal vents, and its enzymes have evolved to be stable at high temperatures. Commercially, it is one of the most expensive reagents used in the DNA profiling process. It can, however, be expressed and purified in-house at a greatly reduced cost. To engineer the starting material, the gene encoding *Taq* is inserted into *E. coli*. When induced, the *Taq* gene is overexpressed, producing large quantities of the protein.

Stocks of *E. coli* containing the plasmid, pAKTaq, have been deposited in the Addgene Plasmid Repository (https://www.addgene.org/) by David Engelke (7), and are offered to educators at minimal cost. To produce *Taq* polymerase for our database, we adapted the procedure described by Bellin et al (8), which can be completed in nine sessions: 1) preparation of starting material: induction of *Taq* expression in large-scale cultures, 2) lysis of bacterial cells and heat precipitation, 3) polyethylenimine (PEI) precipitation, salt washes and dialysis, 4) purification by chromatography: BioRex 70 columns (Bio-Rad, Hercules CA); 5) ultrafiltration and dialysis, 6) protein determination by fluorometry, 7) visualization by polyacrylamide gel electrophoresis, an 8) determination of enzyme activity by PCR.

Isolation and Purification of DNA

DNA was extracted from samples using a standard phenol:chloroform method (9). Briefly, the cotton tip was removed from a swab and incubated overnight (12 – 18 hours) at 56°C in 400 µl DNA extraction buffer (100 mM NaCl, 10 mM Tris-HCl pH 8.0, 25 mM EDTA, 0.5% SDS, 0.1 mg/mL Proteinase K). The swab was removed to a Spin-X filter (Corning, Tewksbury MA) and the tube was centrifuged. Four hundred microliters (an equal volume) of 25:24:1 phenol/chloroform/isoamyl alcohol (Fisher, Norcross GA) were added. DNA was precipitated for at least 1 hour in 1 mL (2.5 volumes) absolute ethanol at-20°C and pelleted by centrifugation. The pellet was washed twice with 1 mL (2.5 volumes) 70% ethanol and dried in a 56°C incubator. The DNA was re-solubilized in 30 µl sterile water by overnight incubation in a 56°C water bath (12 – 18 hours).

Quantification

Human DNA was quantified by *Alu*-specific real-time PCR (10). Ten microliter reactions were prepared containing: 2 µl of purified DNA, 2.6 µl of nuclease-free water, 5 µl of 2X iTaq Universal SYBR Green Supermix (Bio-Rad), and 0.4 µM each primer (forward-GTCAGGAGATCGAGACCATCCC;reverse-TCCTGCCTCAGCCTCCCAAG) (Sigma Aldrich, St. Louis, MO). The standard curve ranged from 0.0077 ng/µl to 16.7 ng/µl with a total of eight data points and was

generated using a human genomic DNA standard (Bioline USA Inc., Taunton, MA). The cycling conditions were: 95ºC for 2 minutes; then 35 cycles of 95ºC for 15 seconds, 68ºC for 1 minute. A melt curve was generated from 65ºC to 95ºC to confirm the single PCR product.

DNA Profiling: Multiplex PCR

Forensic PCR multiplexes in the U.S. typically contain between 15 – 24 short tandem repeat (STR) loci and at least one sex-informative locus. We have developed an in-house version of a 16-locus system, PowerPlex 16 (Promega, Madison WI). Using the publicly-available primer sequences (11), we optimized the system for use in our hands, resulting in an informative forensic assay at a fraction of the cost of the commercial kit. The 25 µl reaction contained: 5 µl 5X Colorless GoTaq® Flexi Buffer, 2.5 mM MgCl$_2$, 2.5 µM dNTPS, 2.5 units GoTaq® Flexi DNA Polymerase (Promega Corporation, Madison WI), and 10 mg/ml BSA (ThermoFisher Scientific, Waltham MA). The primers were combined as a master mix in the following concentrations: DSS51358 – 0.15 µM; THO1 – 0.15 µM; D21S11 - 0.30 µM; D18S51 – 0.41 µM; PentaE – 1.10 µM; D5S818 – 0.19 µM; D13S317 – 0.13 µM; D7S820 – 0.36 µM; D16S539 – 0.17 µM; CSF1PO – 0.20 µM; PentaD – 0.70 µM; Amelogenin – 0.25 µM; vWA – 0.12 µM; D8S1179 – 0.51 µM; TPOX – 0.61 µM; FGA – 0.75 µM. The cycling conditions were: 1) 96°C for 2 minutes; 2) 10 cycles of: 94°C 30 seconds, ramp 0.5°/s to 60°C for 30 seconds, ramp 0.2°/s to 70°C for 45 seconds; 3) 22 cycles of: 90°C for 30 seconds, ramp 0.5°/s to 60°C for 30 seconds, ramp 0.2°/s to 70°C for 45 seconds; 4) 60°C for 30 minutes.

DNA Profiling: Genetic Analysis

PCR product was analyzed using the 3130 Genetic Analyzer (ThermoFisher Scientific, Waltham MA). One microliter of PCR product was combined with 0.5 µl GS LIZ 600 Lane Standard (ThermoFisher Scientific) and 9.5 µl deionized formamide (ThermoFisher Scientific). Samples were electrophoresed on the 3130 Genetic Analyzer with the following run parameters: G5 dye set; Oven Temperature: 60°C; Polymer_Fill_Volume: 6500 steps; Current Stability: 5µA; PreRun Voltage: 15 kV; Pre Run Time: 180 s; InjectionVoltage: 1.2 kV; Injection Time: 10 s; Voltage Number of Steps: 40 nk; Voltage Step Interval 15 s; Data Delay Time: 1 s; Run Voltage: 15 kV; Run Time 1500 s.

Hazards and Safety Precautions

Laboratory personnel maintain current safety, chemical, and bloodborne pathogen training in accordance with University of Illinois standards. Material safety data sheets (MSDS) containing information about potential hazards and how to work safely with chemicals are maintained in the laboratory. Personnel are completely trained under the supervision of experienced researchers in the use of all reagents and instrumentation. Universal precautions are practiced when handling all samples of human origin, and phenol:chloroform:isoamyl alcohol is only opened inside a chemical safety hood. Any waste generated from its use is deposited in a red biohazard sharps container in the hood. All biological waste is deposited in an appropriate biohazard or sharps container for collection and disposal by UIC Environmental Health and Safety personnel.

Results

The size of the peaks in the DNA profile, defined in base pairs, need to be converted to allele number for transfer to the FauxDIS Database. Genemapper ID-X software (ThermoFisher Scientific) can translate size to allele number, but the exercise could also be a learning tool. The complete allelic ladder can be defined by running a known DNA standard and using the allele(s) determined at each locus as benchmarks. The size of the repeat unit is defined for each locus (https://strbase.nist.gov/), and that number can be added or subtracted from the benchmark allele. For example, the locus CSF1PO has a repeating four-base sequence (AGAT). The genotype of the 2800M control DNA at CSF1PO is 12, 12. In our hands, this allele is 341 bp. The size of allele 13 is calculated as 341 + 4 = 345 bp. Using the allelic ladder, the number of base pairs can be converted to allele calls for all of the loci, and student understanding of the structure and function of the alleles in the DNA profiling system are reinforced.

The prototype FauxDIS DNA Database is contained in an Excel file; we are currently converting it to a searchable, online format that can be easily shared and accessed by collaborating users.

We can search for a match to the genotype of an unknown sample by querying the reference samples in the FauxDIS DNA Database. In the Excel file, select "Data" from the ribbon at the top of the spreadsheet and select "Filter" (Figure 1A). Clicking on the arrow in one of the cells opens a drop-down menu from which the locus-specific genotype that matches the unknown can be selected. In Figure 1B, we selected the D3 locus.

With all choices except "14, 15" de-selected from the D3 drop-down menu, the query returned 7 profiles with the genotype 14, 15 (Figure 2A). To further refine the search, selecting the genotype "6, 7"at THO1 returned a single reference profile (Figure 2B).

Once a database reference profile matching the crime scene sample is identified, FauxDIS can further be used as a tool to teach the principles fundamental to the calculation of allele frequencies and genotype probabilities. We

Baranski

describe two alternative models, the choice of which can be based upon the educational background of the students involved and/or on the depth to which the professor wishes to explore the subject. The probability that the DNA profile of a random, unrelated person in the population will match the profile generated from an unknown (crime scene) sample is the Random Match Probability (RMP). The RMP can be calculated based on either observed or expected allele frequencies.

Allele frequency can be estimated by direct observation using the counting method (12). The number of times a DNA profile is observed in the database is compared to the total number of profiles, e.g. sample DB0470 (Figure 2B) has a frequency of 1 in 151, or 0.66%. Determination of genotype frequencies by counting does not rely on theoretical assumptions, therefore it is a simpler method. However, it does not take advantage of the power of the genetic approach.

Theoretical models based on the principles of population genetics can be applied to calculate expected allele frequencies (13). There are two basic assumptions: 1) independence between loci (linkage equilibrium); and 2) independence between alleles (Hardy-Weinberg equilibrium). Linkage equilibrium indicates that the loci are independent and associate randomly. From a forensic standpoint, this means that each matching allele is assumed to provide statistically independent evidence and the frequencies across all of the tested loci can be multiplied to calculate the RMP using the product rule. For a population in Hardy-Weinberg equilibrium, allele frequency can be correlated with genotype frequency. For heterozygotes, frequency is calculated by: $2p_ip_j$, where p_i = the frequency of the one allele and p_j = the frequency of the other allele. Homozygote frequency is calculated by: $p^2 + p(1-p)\theta$, where p = allele frequency and $\theta = 0.01$ in a typical population or $\theta = 0.03$ in an isolated population. The theta correction is a measure of the effects of population substructure i.e. co-ancestry of alleles (14). A table of expected allele frequencies that can be used in calculations of the RMP is available in the literature (15) and online at: (https://www.promega.com/products/pm/genetic-identity/population-statistics/allele-frequencies/).

Figure 1A-B The first 20 profiles in the FauxDIS Database are shown. The genetic loci are listed across the top of the Excel spreadsheet, and the genotypes are entered in the cells. To search the Database: A) select Data from the top ribbon and click on Filter (indicated by arrows); B) click on the arrow in one of the cells to open a drop-down menu and select the genotype (D3 is indicated by the arrow).

A

	Sample	Analyst	D3	THO1	D21	D18	Penta E	D5	D13	D7	D16	CSF	Penta D	Amel	vWA	D8	TPOX	FGA
16	DB0093	NL	14, 15	8, 9.3	30, 32	15, 16	14, 17	9, 12	11, 13	8, 11	9, 10	21, 24	9, 12	X, X	14, 19	10, 13	8, 11	14, 28
22	DB0117	NL	14, 15	7, 8	30, 32.2	15, 21	7, 7	9, 13	10, 13	8, 11	8, 9	11, 13	8, 12	X, Y	18, 19	13, 14	8, 11	22, 24
36	DB0173	AF	14, 15	6, 8	27, 32	16, 16	7, 11	9, 11	13, 14	8, 10	17,17	11, 11	11, 12	X, X	17, 19	10, 14	8, 8	23, 23
45	DB0213	NL	14, 15	8, 9.3	28, 29	11, 14	12, 15	11, 12	8, 11	9, 10	11, 14	11, 12	9, 10	X, X	14, 16	11, 12	8, 9	20, 21
86	DB0470	NL	14, 15	6, 7	28, 30	12, 15	8, 13	7, 11	8, 11	9, 11	11, 11	10, 10	14, 15	X, X	17, 18	12, 13	8, 12	19, 22
102	DB0571	NL	14, 15	8, 9.3	29, 32	17, 17	7, 15	9, 12	8, 11	8, 8	12, 12	11, 12	10, 14	X, X	15, 16	13, 14	8, 8	24, 26
111	DB0670	AF	14, 15	6, 8	28, 28	15, 16	15, 18	11, 11	13, 13	9, 10	10, 13	10, 10	9, 12	X, X	18, 18	11, 12	9, 11	24, 25

B

	Sample	Analyst	D3	THO1	D21	D18	Penta E	D5	D13	D7	D16	CSF	Penta D	Amel	vWA	D8	TPOX	FGA
86	DB0470	NL	14, 15	6, 7	28, 30	12, 15	8, 13	7, 11	8, 11	9, 11	11, 11	10, 10	14, 15	X, X	17, 18	12, 13	8, 12	19, 22

Figure 2A-B Refining the FauxDIS Database search: A) from the profile list generated for Figure 1, select the genotype 14, 15 at the D3 locus to return seven profiles; B) designate the genotype 6,7 at THO1 and identify a single profile.

Baranski

Discussion and Conclusion

We report the development of a DNA profile database modeled after CODIS (16) and available for use as a teaching tool. FauxDIS currently contains 151 DNA profiles and is searchable as an Excel spreadsheet. We have defined a set of analysis procedures analogous to those employed in U.S. operational forensic laboratories for the generation of DNA profiles from biological samples, populating the database with reference (known) samples. In experiential exercises encouraging the application of the scientific method to crime scene investigation, students can generate DNA profiles from unknown (crime scene) samples. The unknown profiles are searched against FauxDIS, and statistics applied to calculate the Random Match Probability. To our knowledge, there is no similar tool available to educators at this time.

FauxDIS is in the early stages of development. We report a fully functional spreadsheet-based database here, but this format will become unsustainable with the continued addition of DNA profiles. With this proof-of-concept in place, we are developing an interactive online tool that can easily be made available to contributors. In either format, we anticipate collaborations with educators both domestically and internationally resulting in a learning network.

We recognize that many colleges and universities will be limited by the availability of the necessary instrumentation to generate a DNA profile. To extend the experiential learning opportunity to as many students as possible, we will also accept single-source samples for in-house analysis. In exchange for a certain number of unique samples, we will generate profiles and deposit them in the database, as if they were collected and submitted to an operational forensic laboratory.

The introduction of this tool as a part of experiential exercises designed to reinforce the practice of the scientific method is expected to be of great benefit to students. First, students who participate in experiential learning activities develop a better understanding of basic scientific principles and are more likely to be retained in a STEM discipline. Next, the experience gained by participation in these exercises may be attractive to potential employers, as the work flow and statistical analyses are analogous to those used in an operational crime laboratory. With additional profiles and collaboration in a network of educators, we believe the FauxDIS DNA Database will be a dynamic learning tool.

Acknowledgements

References

1. McNamara-Schroeder K, Olonan C, Chu S, Montoya MC, Alviri M, Ginty S, et al. DNA fingerprint analysis of three short tandem repeat (STR) loci for biochemistry and forensic science laboratory courses. Biochem Mol Biol Educ 2006;34(5):378-83.

2. Lounsbury KM**.** Crime scene investigation: An exercise in generating and analyzing DNA evidence. Biochem Mol Biol Educ.2003;31:37-41.

3. DeLong Frost L, Peart ST**.** DNA isolation from a dried blood sample, PCR amplification, and population analysis: Making the most of commercially available kits. Biochem Mol Biol Educ 2006;31(6):418-21.

4. Millard JT, Pilon AM**.** Identification of Forensic Samples via Mitochondrial DNA in the Undergraduate Biochemistry Laboratory. J Chem Educ 2003;80(4):444-6.

5. Reeder DJ**.** Impact of DNA typing on standards and practice in the forensic community. Arch Pathol Lab Med 1999;123(11):1063-5.

6. Baechtel FS, Monson KL, Forsen GE, Budowle B, Kearney JJ**.** Tracking the violent criminal offender through DNA typing profiles--a national database system concept. EXS 1991;58:356-60.

7. Engelke DR, Krikos A, Bruck ME, Ginsburg D**.** Purification of *Thermus aquaticus* DNA polymerase expressed in Escherichia coli. Anal Biochem 1990;191(2):396-400.

8. Bellin RM, Bruno MK, Farrow MA**.** Purification and characterization of Taq polymerase: A 9-week biochemistry laboratory project for undergraduate students. Biochem Mol Biol Educ 2010;38(1):11-6.

9. Comey C, Koons B, Presley K., Smerick J, Sobieralski C, Stanley D, and Baechtel F. DNA extraction strategies for amplified fragment length polymorphism analysis. J Forensic Sci 1994;39(5):1254-69.

10. Nicklas JA, Buel E**.** Development of an Alu-based, real-time PCR method for quantitation of human DNA in forensic samples. J Forensic Sci 2003;48(5):936-44.

11. Krenke BE, Tereba A, Anderson SJ, Buel E, Culhane S, Finis CJ, et al. Validation of a 16-locus fluorescent multiplex system. J Forensic Sci 2002;47(4):773-85.

12. NRC**.** DNA Technology in Forensic Science. Washington (DC): National Academy of Sciences; 1992.

13. NRC**.** The Evaluation of Forensic DNA Evidence. Washington (DC); 1996.

14. Butler J**.** Forensic DNA Typing, 2nd Edition: Academic Press; 2005.

15. Steffen CR, Coble MD, Gettings KB, Vallone PM**.** Corrigendum to 'U.S. Population Data for 29

Autosomal STR Loci' [Forensic Sci Int Genet 7 (2013) e82-e83]. Forensic Sci Int Genet 2017;31:e36-e40.

16. Butler J**.** DNA Databases: Uses and Issues. Advanced Topics in Forensic DNA Typing: Methodology: Elsevier; 2012:214-70.