Center for Law & Human Behavior

The University of Texas at El Paso

DNA BARCODING, HIGH THROUGHPUT DNA SEQUENCING, AND FORENSIC SCIENCE:

RECENT ADVANCES AND FUTURE PROSPECTS

DHS SYMPOSIUM SERIES NO. 3

June 23, 2016

Research in Brief

CBTIR

A Department of Homeland Security Center of Excellence

Center for Law & Human Behavior
The University of Texas at El Paso
500 West University Avenue
Prospect Hall, Room 226
El Paso, Texas 79968
Tel: (915) 747-5920
Email: clhb@utep.edu
Website: http://clhb.utep.edu
Follow us on Twitter @NCBSI

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

This symposium is supported by the U.S. Department of Homeland Security, Science and Technology, Office of University Programs under Grant Award Number DHS-14-ST-061-COE-00.

ABOUT THIS REPORT

This report provides a brief overview of some recent applications of DNA barcoding, metabarcoding, high-throughput DNA sequencing, and next-generation DNA sequencing in forensics and details some preliminary work looking at the application of DNA metabarcoding in pollen forensics. As DNA sequencing has become faster, cheaper, and more accurate, its applications have become more widespread and varied. DNA sequencing is commonly used in scientific fields such as ecology, genetics, and evolution, but some of the newer techniques used in these fields have yet to be applied to forensic science. Here we discuss the use of high-throughput DNA sequencing method, known as DNA metabarcoding, to identify plant material that may be of use to the law enforcement and forensic community.

What Are The Research Findings?

Our group was able to successfully identify multiple plant species found in soil samples taken from multiple geographic areas to the genus and, in some cases, species level. Additionally, the sequencing results were consistent with our knowledge of what plants are generally found in each geographic area. This suggests that DNA metabarcoding may be a valid methodology to identify pollen in environmental samples and would be quicker and potentially more sensitive than current visual methods of identification.

What Were the Study's Limitations?

Presented here are preliminary analyses for a subset of samples used in a larger study on the effectiveness of DNA metabarcoding for pollen identification in environmental samples. Further analyses and follow-up studies are currently being conducted to determine how broadly applicable this methodology is in terms of matrices and substrates. Additionally, while some effort has been made to compile a botanical list suitable for forensic geolocation, at this time such a list does not exist (though see Goodman et al. 2015). So, while this study does suggest that DNA metabarcoding is capable of identifying what plant material found in a soil originated from, it has not been used to do so in a forensic setting or for law enforcement purposes.

Who Should Read This Report?

Crime lab personnel and federal, state, and local law enforcement agents who handle crime and evidence materials.

Introduction

Most of us are familiar with the use of genetic methods, particularly DNA fingerprinting, for the identification of human individuals in forensic science. Genetic methods are also routinely used in wildlife forensics, drug enforcement, and food security labs for the identification of targeted non-human species (Galimberti et al. 2012, Ogden & Linacre 2015, Mello et al. 2015). Some currently used methods are only applicable to a specific species. In some cases, heritable patterns found in highly variable repeat regions of the target species' genome, known as microsatellites, SSRs, or more commonly "DNA fingerprints", can allow scientists to identify what population an individual or sample comes from. For example, microsatellites can be used to determine whether timber is coming from a legally or an illegally harvested population (Degen et al. 2013).

Other genetic methods, such as DNA barcoding can be applied to a broad taxonomic range of species. A DNA barcode is a short sequence of DNA that is found in a broad range of species, meaning it is taxonomically conserved. Because sequences used for DNA barcodes are present in many species, and vary between species, but not within species, ideally each species should have its own DNA barcode. Therefore, species can be identified based on their species' sequence at that barcoding locus. DNA barcoding can be used to identify what species a sample comes from even when morphological identification would be difficult or even impossible. For example, DNA barcoding can be used to determine whether butchered meat originated from a domestic species, such as cattle, or a protected species susceptible to poaching, such as a rhinoceros (Dalton & Kotze 2011).

Recent innovations in sequencing technology can allow us to identify multiple species present in a single sample. DNA metabarcoding combines next-generation DNA sequencing, DNA barcodes, and bioinformatics to identify and quantify diversity within a sample. There are many potential applications for DNA metabarcoding in forensic science. For example, this process has been used to identify the plants present in mixtures sold as traditional Chinese medicine (Coughlan et al. 2012).

DNA barcoding can also be applied to the field of pollen forensics (Bell et al. 2016). Pollen is one of a plant's most unique structures. Individual pollen grains have long been used to identify plant species present in an ecosystem. Often, as in the case of paleoecology, these species are identified long after the rest of the plant is gone. Pollen can also provide forensic evidence of where a person or object has been because different locations will have distinct sets of pollen species and those will often vary depending on the time of year. Here we report on the use of DNA metabarcoding to identify plant material in environmental samples.

Project Overview

Despite the wealth of information which the analysis of pollen can provide and its potential applications in many different aspects of both border security, food safety, and ecosystem services, pollen analysis and identification is still done mainly using microscopy. There are very few facilities in the United States that routinely conduct forensic pollen analysis for government agencies (Hwang & Masters 2013). Because accurate pollen classification requires specialized training and experience in identifying and correctly classifying the morphological traits of individual pollen grains and relies heavily on multiple pollen image databases, it is often a time-consuming and highly specialized task.

However, with the advent of newer, more sensitive, and cheaper next-generation sequencing technologies, we will be able to not only speed up the identification process, but also provide more comprehensive results with less taxonomically specialized expertise. Recently, several research groups have begun work developing molecular and bioinformatics protocols that will allow for the rapid and accurate identification of pollen species in mixed pollen samples (Bell et al. 2016). Our group has recently completed work using DNA metabarcoding to identify plant material found in soil samples using two DNA barcoding primers. We present some of our preliminary analyses here for one of those primers.

Objectives

Broadly, this project demonstrates the utility of DNA metabarcoding in quickly and accurately identifying plant material to the genus and sometimes the species level from environmental samples. We used multiple DNA extraction protocols to extract DNA from mixed DNA samples, specifically from forensic soil samples taken from multiple geographic areas. The DNA extractions from these samples were then sequenced using the high-throughput next generation sequencing technique, DNA metabarcoding.

Taken as a whole, this project demonstrates the ease with which accurate plant identification can be obtained using slightly modified off-the-shelf DNA extraction kits and high-throughput next generation sequencing techniques. This opens the door for the faster and more efficient identification of plant and pollen samples collected from a variety of forensic settings.

Methodology

For this study we chose to use two DNA barcode primers. The first primer amplifies DNA from the nuclear internal transcribed sequence or ITS region, a region that is conserved across the tree of life. The second primer used amplifies a region of the chloroplast genome which codes for the Ribulose-1, 5-bisphosphate carboxylase/ oxygenase enzyme known as rbcL. This sequence only amplifies in organisms with chloroplasts, such as plants and algae. Here we present some preliminary analyses for the ITS primer developed during a summer 2014 collaboration with Drs. Michael Moody and Kyle Johnson at UTEP sponsored by the DHS and ORISE.

Utility and accuracy of ITS region for plant identification: In order for our proposed technique to be utilized more broadly, it was critical that we first determine whether the nuclear plant ITS region would 1) be informative as a locus (i.e. enough plants have their ITS regions sequenced to allow for the identification of useful species) and 2) allow for the accurate identification of unknown samples to at least the level of genus. To do this we first compiled plant species lists for Dona Anna County, NM and the metro Atlanta, GA counties (Cobb, Dekalb, Fulton, and Gwinnett) from publically available herbarium lists (SEINet: http://swbiodiversity.org/portal/index.php, Valdosta State University Herbarium: http://herb.valdosta.edu/). We then searched for the key words "internal transcribed sequence", "complete", and the genus and species name of each plant in the publically available nucleotide database of NCBI's GenBank (http://www.ncbi.nlm.nih.gov/genbank/).

To determine how informative the ITS region would be, we calculated the percentage of geographically identified species with complete ITS regions available in GenBank, as well as the number of potential geographically informative species between the two regions. Additionally to test the utility of phylogenetic trees built using ITS regions to identify the genus of unknown, unsequenced samples, we used these previously collected sequences to build phylogenetic trees (Tamura et al. 2013). We then used these trees to determine the accuracy of the ITS region in identifying the genus of unknown plant samples.

Sample collection & DNA extractions: A total of four soil samples were used in this preliminary. The four soil samples represent environmental samples with a mix of plant and non-plant material. One soil sample was collected from an El Paso Arroyo (Lat 31.46 N/ Long 106.29 W), another from an Arroyo at the UTEP Indio Mountain Research Station (Lat 30.46N/ Long 105.0 W), the third from metro Atlanta, GA backyard (Lat 33.45 N/ Long 84.16 W), and the fourth was from metro Raleigh, VA plot (Lat 35.46 N/ Long 78.38 W). These samples were chosen in order to allow for multiple comparisons across different geographic scales. Two of the comparisons are on a smaller scale (between 120-300 miles), and four are long distance comparisons of approximately 1,500 miles between the two sampling points. Samples were collected using a modified "pinch" sampling method (Adams, 1975). Briefly, for each sample, 20 random pinches of topsoil were collected and combined from a 100 x 100 meter area.

DNA from the environmental soil samples was extracted using the standard extraction protocol in the commercially available MoBio Soil Extraction Kit.

DNA metabarcoding at the ITS region: DNA metabarcode sequencing is done in a two step process. First the samples are amplified at the specified primer (ITS primer sequence) using single-step 30 cycle hot start PCR (Qiagen, Valencia, CA). All PCR amplicons are purified using Agencourt Ampure beads (Agencourt Biosci Corp, MA, USA). Samples were then sequenced using a Roche 454 FLX instrument and reagents (following manufacturer's guidelines). The sequence data is processed using a proprietary analysis pipeline (www.mrdnalab.com, MR DNA, Shallowater, TX). Similar to other freely available pipelines, this process removes primer sequences, short reads, sequences with ambiguous base calls, homopolymer runs exceeding 6bp, and chimeras. Reads are taxonomically classified using BLASTn (www.ncbi.nlm.nih.gov) and compiled into both "counts" and "percentage" files. Data is provided in tab-delimited text files that can be further analyzed using excel and excel compatible programs.

Major Study Findings

This pilot project asked two basic questions about the applicability of ITS primers and DNA metabarcoding to forensic pollen identification and geolocation:

1. Do we have enough plant ITS regions sequenced to allow for sequence identification to the genus level?

2. Can we identify plant material in soil samples using DNA metabarcoding?

Utility and accuracy of ITS region for plant identification: Our research into the utility of the ITS barcoding region for identifying plant species to the genus level suggests that it would indeed work as a method for plant identification and eventually geolocation. Of the 1,723 herbarium specimens listed in the Dona Ana herbarium list 633 (37%) had complete ITS1 and/or complete ITS 2 sequences, while of the 138 metro Atlanta herbarium specimens, 67 (49%) had complete ITS1 and/or complete ITS 2 sequences. In both of these cases, this is a conservative estimate of the number of plants which we would be able to identify to the species level since we included only complete ITS sequences in our analysis, and many species only have partial sequences available. Many times species identification would be possible with a partial sequence, but for these analyses we included only those with complete sequences.

Additionally, we compiled a single neighbor-joining phylogenetic tree based on the ITS sequences from two closely related plant families (the *Cyperaceae* and the *Poaceae*) which were comprised of 6 genera and 49 species and 2 genera and 12 species respectively. After compiling the tree, we ran several simulations in which an "unknown" sequence of the same genus (a sequence that was not included in the original building of the tree) was included in the building of the new tree. In all cases, the "unknown" sequence grouped with its expected genus. This means that in cases where we have a true unknown sample (a sample which has not had it's ITS region sequenced and deposited in GenBank), we would still be able to identify it to the genus level. Current morphological identification can often only identify pollen to the genus. That means that even with missing information, this methodology is as useful as the current one, and when all the information is available, this method is superior in its taxonomic level of identification.

Plant identification using DNA metabarcoding for four soil samples: We successfully extracted DNA from all four soil samples without further modification of the standard soil DNA extraction kit. We also successfully amplified the ITS barcode region in all four soil DNA samples. Preliminary analyses using the ITS2 bioinformatics pipeline recently designed by Sickel et al. (2015), identified a total of eight plant species in the four samples.

In the El Paso soil sample, 100% of the plant sequences were identified to the species level and were found to have all originated from monkey comb (*Pithecoctenium cynanchoides*). This is a plant common in the El Paso, TX and New Mexico region. All of the plant sequences from the soil sample taken at Indio Mountain Research Station in Texas were from the genus *Solanum*, commonly known as the nightshades. Based on herbarium lists for the region, the sequences most likely originated from the silverleaf nightshade (*Solanum elaeagnifolium*), a plant commonly found in this dry desert area. Three plants made up of the majority (90%) of the sequences from the Atlanta, GA sample. One of the three samples, American Pokeweed (*Phytolacca americana*) was identified to the species level, while the other two the ornamental *Veronica sp.* and the invasive *Achyranthes sp.* were identified to the genus level. The Raleigh, NC soil sample was primarily a grass from the genus *Cynodon*, a mustard from the genus *Lepidium*, and English plaintain *Plantago lanceolata* (identified to the species level).

Importantly, none of the four samples showed any overlap in the species identified, suggesting that there was no common contaminant. Additionally, all species and genera identified are consistent with the plants commonly found in the geographic regions sampled. This suggests that this method for pollen identification is one that could easily work for quickly and efficiently identifying pollen to the genus and sometimes species level. Additionally, pollen profiles generated using this methodology were geographically distinct and therefore could have possible applications in forensic geolocation.

Future Areas of Research

While DNA barcodes have been applied to fields such as wildlife forensics, food safety, and forensic entomology, high-throughput and next-generation sequencing technologies have yet to be applied in many of these cases. Research in other fields, such as genetics and ecology, frequently uses these cutting edge techniques to answer basic research questions.

As demonstrated in this report, high-throughput DNA sequence technology can allow for quicker and more accurate identification of species from environmental samples containing a mixture of DNA. The broad applicability and low cost of this and other emerging DNA sequencing technologies make them prime candidates for use in forensic fields. However, proper validation is necessary before these techniques can be used for law enforcement purposes.

Implications for Practice

Currently pollen forensic analyses are rarely performed during DHS investigations due to the extensive training needed. The project described here outlines a more efficient method of pollen identification that requires minimal training and is possible with only slight modifications of commercially available technologies. Using simple DNA extraction techniques and relatively inexpensive high-throughput DNA sequencing technology, pollen forensics would become more readily available to a larger number of forensic labs, could be utilized in a larger number of forensic investigations (including drug, transnational criminal, trade and tariff investigations), and could be completed in a fraction of the current time. The adoption of these cutting edge techniques can and should be made a priority in forensic and law enforcement labs due to their potential to greatly increase the speed and efficiency with which samples can be processed and utilized in criminal investigations.

Adams DP & PJ Mehringer Jr. (1975) Modern pollen surface samples—an analysis of subsamples. Journal of Research of the U. S. Geological Survey, Vol. 3 (733–736)

Bell KL, KS Burgess, KC Okamoto, R Aranda, & BJ Brosi (2016) Review and future prospects for DNA barcoding methods in forensic palynology. Forensic Science International: Genetics, Vol 21 (110-116)

Coghlan ML, J Haile, J Houston, DC Murray, NE White, P Moolhuijzen, MI Bellgard, & M Bunce (2012) Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. PLoS Genetics, Vol. 8:4 (e1002657)

Dalton D & A Kotze (2011) DNA barcoding as a tool for species identification in three forensic wildlife cases in South Africa. Forensic Science International, Vol 207:1-3 (e51-e54)

Degen B, SE Ward, MR Lemes, C Navarro, S Cavers, & AM Sebbenn (2013) Verifying the geographic origin of mahogany (*Swietenia macrophylla* King) with DNA-fingerprints. Forensic Science International: Genetics, Vol 7:1 (55-62)

Galimberti A, F De Mattia, A Losa, I Bruni, S Federici, M Casiraghi, S Martellos, & M Labra (2013) DNA barcoding as a new tool for food traceability, Food Research International, Vol. 50:1 (55-63)

Goodman FJ, JW Doughty, C Gary, CT Christou, BB Hu, EA Hultman, DG Deanto, & D Masters (2015) PIGLT: A pollen identification and geolocation system for forensic applications, Conference Paper, 2015 IEEE International Symposium on Technologies for Homeland Security

Hwang GM & D Masters (2013) Forensic geolocation challenge: is pollen analysis the answer? AASP The Palynological Society Newsletter, Special Issue July 2013

Mello ICT, ASD. Ribeiro, VHG. Dias, R Silva, BD Sabino, RG Garrido, & L Seldin (2016) A segment of rbcL gene as a potential tool for forensic discrimination of *Cannabis sativa* seized at Rio de Janeiro, Brazil, International Journal of Legal Medicine, Vol. 130:2 (353-356)

Ogden R & A Linacre (2015) Wildlife forensic science: a review of genetic geographic origin assignment, Forensic Science International: Genetics, Vol. 18 (152-159)

Sickel W. MJ Ankenbrand, G Grimmer, A Holzschuh, S Hartel, J Lanzen, I Steffan-Dewenter, & A Keller (2015) Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach, BMC Ecology, Vol 15:20

Tamura K, G Stecher, D Peterson, A Filipski A, & S Kumar (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution, Vol. 30 (2725-2729)

About the author

Dr. Jennifer Kovacs is an Assistant Professor of Biology at Spelman College.

For further information related to this work please contact Dr. Kovacs at (404) 270-5724 or at jkovacs@spelman.edu

CBTIR

A Department of Homeland Security Center of Excellence

http://www.uh.edu/cbtir/

Twitter: @CBTIR_UH

CENTER FOR LAW AND HUMAN BEHAVIOR

http://clhb.utep.edu

Twitter: @NCBSI