Application of DNA metabarcoding and high-throughput sequencing for enhanced forensic soil DNA analysis

Jennifer M. Young

Thesis submitted in fulfilment of the requirements for

Degree of Doctor of Philosophy

University of Adelaide

July 2014

TABLE OF CONTENTS

THESIS ABSTRACT

The complex and variable soil matrix can support a wide range of biota that can provide information about local vegetation, soil conditions (e.g. soil acidity) and habitat type. As the combination of microbes, plants and animals within a soil is often specific to a given site, identification of the soil biota can narrow the likely source of a soil sample. DNA fingerprinting analysis of soil microbes has been used as forensic evidence in court to establish a link between a suspect and a site, victim or object. However, previous genetic analyses have relied on patterns of fragment length variation produced by amplification of unidentified taxa in the soil extract, particularly bacteria. In contrast, the development of advanced DNA sequencing technologies now provides the ability to generate a detailed picture of soil communities and the taxa present, allowing for improved discrimination between samples. This thesis examines the use of DNA metabarcoding combined with high-throughput sequencing (HTS) technology to distinguish between soils from different locations in a forensic context. Specifically, I review the DNA extraction protocols available for soils and recommend best practice for successful analysis (Chapter 2). Following this, I examine the reproducibility and discriminatory power of four different genetic markers for forensic soil discrimination using HTS (Chapter 3). Non-bacterial DNA, particularly fungi, were found to be the most promising target for soil discrimination and additionally showed consistent PCR amplification and low contamination risk. It is known that DNA extraction protocols can introduce discrepancies in soil community profiles, and the optimal sample size for an accurate and representative survey of soil diversity has been debated. Therefore I used various soil types to test the robustness of modified DNA extraction protocols (Chapter 4) and trace, or limited, amounts of soil (Chapter 5). I make recommendations about the optimal DNA extraction method and sample size given soil properties such as clay content, soil pH and texture. To assess the application of this method in forensic casework, I then designed a mock case scenario. DNA profiles of six soil samples recovered from a suspect’s belongings were compared to those collected from seven reference sites around Adelaide, South Australia. This study demonstrated that the soil from the suspect’s belongings had eukaryote diversity more similar to those collected from the crime scene than to any other sample collected at random. This suggested the presence of the suspect at the crime scene. This result was compared to that from a soil analysis method currently accepted in court. In this case example, both methods successfully established a link between the suspect’s belongings and the crime scene; however, DNA analysis improved resolution between reference locations. This thesis demonstrates the first practical application of DNA metabarcoding and high-throughput sequencing (HTS) to forensic soil analysis. I show that this approach is consistently able to distinguish between soil samples taken from different localities, and consequently may be employed as an additional line of evidence or investigation in forensic casework.

THESIS DECLARATION

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution to Jennifer M. Young and, to the best of my knowledge and belief, contains no material previously published or written by any other person, except where due reference has been made in the text. I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis (as listed below) resides with the copyright holder(s) of those works.

I also give my permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, and also through web research engines, unless permission has been granted by the University to restrict access for a period of time.

....

Jennifer M. Young

July 2014

Publications

Young, Jennifer M., et al. "Limitations and recommendations for successful DNA extraction from forensic soil samples: A review." *Science & Justice* 54.3 (2014): 238-244.

Young, Jennifer M., Laura S. Weyrich, and Alan Cooper. “Forensic soil DNA analysis using high-throughput sequencing: A comparison of four molecular markers.” *Forensic Science International: Genetics* 13 (2014): 176-184.

RAWLENCE, N. J., LOWE, D. J., WOOD, J. R., **YOUNG, J. M**., CHURCHMAN, G., HUANG, Y. T., & COOPER, A. (2014). Using palaeoenvironmental DNA to reconstruct past environments: progress and prospects. *Journal of Quaternary Science*, *29*(7), 610-626.

Oral Presentations

22nd International Symposium on the Forensic Sciences (ANZFSS), Adelaide, August 2014. High-throughput sequencing of soil eukaryotes links a suspect to a crime scene: a mock case scenario.

Forensic Science South Australia (FSSA) Seminar, Adelaide, July 2014, Overview of PhD Thesis

25th World Congress of the International Society for Forensic Genetics (ISFG), Melbourne, September 2013. The use of NGS for enhanced forensic soil DNA analysis.

Three minute thesis competition, University of Adelaide, July 2013, Forensic soil analysis using metagenomic analysis.

AFP-UC Forensic R&D Workshop, Canberra, March 2012. DNA metabarcoding for forensic soil analysis.

SMANZFL Meeting (Senior Managers of Australian and New Zealand Forensic Labs), November 2011

Post Graduate Initial Seminar, University of Adelaide, November 2011, Environmental Genomics for soil Forensics.

Upcoming presentations

Australian Federal Police Research and Development Team, Canberra, December 2014, *Overview of PhD Thesis*

Poster Presentations

22nd International Symposium on the Forensic Sciences (ANSFSS), Hobart, September 2012. *Forensic soil analysis: Using DNA metabarcoding and next generation sequencing to differentiate between habitat types.*

2012 Post Graduate Poster Day, University of Adelaide, July 2012. *DIGGING THE DIRT: validation of DNA metabarcoding to differentiate forensic soil samples* from different habitat types.

CANQUA-CGRG Conference, Edmonton, August 2013, *Using palaeoenvironmental DNA to reconstruct past environments: progress, problems, and prospects.* Rawlence, N.R., Lowe, D.J., Wood, J.R., **Young, J**., Churchman, G.J., Huang, Y.-T., Cooper, A.

Awards

ANZFFS Travel Award, 22nd International Symposium on the Forensic Sciences, Adelaide, August

ACKNOWLEDGEMENTS

First, massive thanks to Laura Weyrich without whom this would not have been possible. I am so grateful for all the time, effort, support and encouragement that you gave me throughout this process. You are a superstar! Many thanks to Alan Cooper, firstly for giving me the opportunity to advance my career in Forensic Science, but also for enabling me to fulfil my dream of living in Australia. For this is I am very grateful.

To all members of ACAD for their help, advice and friendship over the last three years, you helped make my move to Australia awesome. In particular, I would like to thank CSI ACAD for their encouragement on Friday mornings and for providing memorable conference trips; many thanks to Jeremy Austin, Denice Higgins, Eleanor Tobe, Jennifer Templeton and Janette Edson. I am privileged to have had the opportunity to work with you all.

To all my brilliant friends in Adelaide, thanks for putting up with me when I was stressed and rambling on about dirt. In particular, to my partner in crime Jennifer Shaw, I cannot thank you enough for always being there to pick me up and pull me through the tough times. Huge thanks to my partner and best friend, Dale Bragg, for always believing in me and making me smile when times were rough. You gave me the strength to see this through, I am lucky to have to you by my side.

Thank you to all my Scottish folks back home for your continued support. Special thanks to my mum and dad, I could never have done this without all your love and support over the years. I missed you all heaps but your trips to Adelaide really made it all worthwhile.

CHAPTER 1

General Introduction

General Introduction

Forensic analysis of soils

Forensic science receives intense media attention and public interest due to the revolutionised ability of investigators to solve crimes. Analysis of physical evidence is commonly a deciding factor in casework by establishing what transpired at a scene or who was involved. In particular, soil can serve as powerful, nearly ‘ideal’ contact trace evidence, as it is highly individualistic, easy to characterise, has a high transfer and retention probability and is often overlooked in attempts to conceal evidence (1, 2). The concept of soil evidence was first introduced in the late 1880’s by Sir Arthur Conan Doyle in the famous series ‘The Adventures of Sherlock Holmes.’ Subsequently, Hans Gross, an Austrian criminal investigator, stressed the importance of soil science to forensic investigations in his book *Handbuch fur Untersuchungrichler als System dur Kriminalistik* (Handbook for Examining Magistrates, 1893), and in 1904, German chemist Georg Popp was the first forensic scientist to utilize soil evidence in court (3). Since these early cases, soil analysis has been applied to forensic casework in two ways: as evidence to link a suspect, location or object to a crime scene, or as intelligence to provide information on the likely origin of a soil sample in the absence of reference samples. The use of soil for evidential purposes was demonstrated by Concheri *et al.* (2) who used both chemical and biological analysis to link soil particles found in a suspect’s car to soil in a corn field where a body had been recovered. In contrast, mineralogical analysis of soil from a shovel found in a suspect’s car provided forensic intelligence by directing the investigation towards the Oakbank Quarry in the Adelaide Hills, where two bodies were later recovered (4). On the back of this success, the Centre for Australian Forensic Soil Science (CAFSS) was established as the first formal worldwide network of forensic soil scientists in 2003, and is now actively involved in routine casework.

Soil is conceived as an organised natural body at the surface of the earth that serves as a medium for plant growth; however, most engineers and geologists tend to regard soils mainly as weathered rock or regolith. Soils form on the surface of a parent material or underlying substrate by physical, chemical and biological processes. Such processes result in variable soil characteristics including horizonation, colour, presence of pedality, texture and/or consistency. The Australian Soil Classification scheme aims to categorise soils based on such variations. In particular, many soil samples encountered in forensic work are classified as Anthroposols, i.e. human made soils. Human activities can be responsible for the creation of 'non-natural' parent materials as well as 'non-natural' alteration processes such as the addition of anthropogenic materials to surface soils. Anthroposols are classified into suborders based on the modifications made to the parent material. For example, hortic soils include those subject to additions of organic residues such as organic wastes, composts, mulches. Garbic describes mineral soil or regolithic materials that are underlain by land fill of manufactured origin (domestic or industrial) predominantly of an organic nature, whereas Urbic describes those underlain by land fill of predominantly a mineral nature (e.g. manufactured glass, plastics, concrete, etc.) or contain a mixture of manufactured materials and materials of pedogenic origin. Other Anthroposol classes include Cumulic, Dredgic, Spolic and Scalpic, each of which the geology, mineralogy, chemistry and biology will reflect the specified human activity. As a result, such soils can provide information regarding the likely source of a forensic soil sample.

Due to the variation and complex compositions of earth materials, forensic soil analysis is a multidisciplinary field spanning pedology, geochemistry, mineralogy, biology, geophysics, archaeology and forensic science (5). The application of earth evidence in forensic science has been referred to as forensic pedology; forensic geology; forensic geoscience; geoforensics; and soil forensics. Forensic pedology describes the study of soil, both *in situ* as a natural material possibly disrupted by unusual events, or as a transferred material on suspects, victims and associated items (e.g. tyres, vehicles)(6). Forensic geology refers to the use of geological methods (e.g. geophysics, petrography, geochemistry, microscopy, micropalaeontology) in the analysis of samples and places that may be connected with criminal behaviour and disasters. Characteristics examined include grain size, sorting, grain shape, grain surface and can also include analysis of mineralogy (thin section petrography, X-ray diffraction), particle size (by sieving and weighing) and chemistry (including XRF, ICP, FTIR, microprobe, amongst others). Forensic geoscience encompasses forensic pedology, geology as well as statistics and bedrock geology. Similarly, Geoforensics also encompasses bedrock geomorphology (origin and evolution of topographic and bathymetric features created by physical or chemical processes), GIS, remote sensing, human geography (including sociology), geostatistics, as well as some specialist analytical techniques including bone taphonomy, isotope analysis and the SEM-microprobe method of QemScan. Geomorphology is typically used for searching the land for surface or buried objects and can include criminal behaviour and landform analysis. Soil forensics includes all the above methods available applied specifically to soil, as opposed to bedrock geology, for instance.

Forensic soil analysis typically follows a systematic approach developed by CAFSS (4). This step-wise analysis scheme enables grossly different samples to be excluded in the initial stages of analysis. Stage 1 involves characterisation of visual and physical properties such as soil texture, consistency, particle size, pH, and colour (4, 7-9). Following this, microscopy and spectroscopy techniques are applied to examine the grain size and shape, elemental composition, inorganic and/or organic compounds within soils (10-13). For example, the inorganic fraction of soil can be analysed using infra-red spectroscopy (14), a non-destructive technique that can identify the presence of rare minerals. Similarly, atomic absorption spectroscopy can detect levels of trace elements, such as lead content which was used to discriminate garden and allotment soil samples from urban and rural areas in England (15). Inductively Coupled Plasma Mass Spectrometry (ICP-MS) has also been used to link a sample to a specific origin based on chemical elemental distribution (2). Discrimination of soils based on organics is also possible using UV-VIS absorbance spectroscopy (16) and pyrolysis gas chromatography (PyGC) (17, 18). Although physical and chemical analyses allow simple, rapid screening of complex materials (19), many of these techniques are destructive and require a large soil mass (>1g) that is often unavailable for forensic casework. Furthermore, some of these features can be semi-subjective. For example, colour analyses are based on visual comparison to the Munsell colour chart (20-22), and interpretation of spectra relies on expert examination that may not be readily available (23).

Identification followed by individualisation are the key drivers of forensic science. Comparison of soil chemical characteristics to a reference database, such as the Australian Soil Resource Information System (ASRIS), or the Australian Soil Classification (ASC) database, can identify the likely origin(s) of a sample and allow efficient comparison of a forensic soil sample to nationwide soil distributions (5, 24); however, soil mineralogy typically varies at regional rather than local scales, limiting geographic precision of these techniques (25). However, the soil matrix supports a wide range of organisms with specialised habitat requirements and restricted geographical distributions, resulting in a unique combination of soil biota in a given area. As the combination of microbes, plants and animals within a soil is often specific to a given site, individualisation of samples based on the soil biota can discriminate soils with similar chemical and mineralogical properties.

Plant fragments, including roots, pollen, seeds, or leaves, are often found during investigations and have the potential to link a suspect, victim, or vehicle to a crime scene (26-30). In addition, invertebrates, including mites, beetles, arachnids and nematodes commonly associated with leaf litter and organic matter on the soil surface can also aid in forensic soil analysis. For example, Calliphorid (carrion fly) maggots isolated from soil have proved to be useful in determining the post-mortem interval (PMI) (31). Nevertheless, morphological identification can be challenging when poor quality specimens are available. For instance, it may not be possible to identify a plant species if reproductive (e.g. flowers) or other diagnostic structures are not available (32), and larvae and eggs from invertebrates are extremely difficult to identify morphologically compared to adult insects. In addition, many soil bacteria and fungi cannot be cultured and therefore cannot be characterised using traditional methods.

Soil DNA analysis

An alternative to morphological identification of soil biota is DNA analysis (33-35). To date microbes, particularly bacteria, have been the target for soil DNA analyses as they can provide discriminative DNA profiles (36); however, the taxonomic resolution of bacterial DNA fingerprint methods is limited. DNA fingerprint techniques rely on differences in fragment lengths between species in a sample to generate a profile and so individual species present are not identified. In contrast, the recent development of DNA metabarcoding (PCR amplification of DNA mixtures using universal primers) and high-throughput sequencing (HTS) enables rapid species identification and offers the potential to improve soil discrimination by targeting non-culturable microorganisms (i.e. those that cannot be cultured on routine microbiological media) and alternative soil taxa such as eukaryotes.

Forensic discrimination using microbial T-RFLP analysis

The majority of forensic soil DNA studies to date have examined microbial diversity (reviewed in 37) using DNA fingerprinting methods (35) including denaturing gradient gel electrophoresis (DGGE; 38), amplicon length heterogeneity PCR (LH-PCR; 39), and terminal restriction fragment length polymorphism (T-RFLP; 40, 41, 42). T-RFLP is extensively used in forensic soil science and is done by amplifying a region of the 16S ribosomal RNA encoding gene (rRNA) and digesting it with restriction endonucleases. The 16S rRNA fragments of varying length are separated by gel electrophoresis and analysed to provide a distinct profile (fingerprint) dependent upon the species composition within the sample. This method was introduced by Liu *et al.* in 1997 to characterise bacterial communities in activated sludge, bioreactor sludge, aquifer sand and termite guts (43).

Shortly after Horswell, J. (36) demonstrated that T-RFLP could generate discriminative microbial DNA profiles from soil. The benefits of T-RFLP include small soil sample sizes (≤ 1 g), use of common forensic equipment, ease of automation, and low cost, allowing rapid and reproducible soil community fingerprinting (43-45). Furthermore, T-RFLP analysis has been shown to provide higher discriminatory power between sites than elemental analysis (46) and allows a more powerful analysis than culture-based techniques that account for only 2% of the total bacteria present in a sample (47). Although T-RFLP is a useful forensic tool (36, 48-51), resolution and taxonomic identification are limited by co-migration of multiple species appearing as a single species during electrophoresis (38, 52). In contrast, DNA metabarcoding can provide a standardised species identification method from complex soil communities, whilst increasing the discriminatory power between locations for forensic application.

DNA barcoding

DNA barcoding is a common molecular biology tool used for rapid species identification (53) and relies on PCR amplification of a variable fragment (DNA template) that is flanked by a conserved sequence. For species identification, barcode regions require sufficient divergence to differentiate between species but with little or no intra-specific variation (54-56). DNA barcoding relies on comparative data being available for the barcode region from identified organisms, such as voucher specimens. Barcode regions that can be amplified with universal primers that are compatible with a wide range of species are the most useful (55), such as 16S rRNA (bacteria), *rbcL* (plants), and 18S rRNA (eukaryotes). The sequences generated from voucher specimens are then entered into online databases that contain millions of genetic sequences from additional voucher specimens, e.g. GenBank. Such databases provide a reference set for identification of unidentified DNA sequences. For animals, the protein coding *CO1* mitochondrial region, (~650 bp region near the 5' end of the cytochrome oxidase subunit) is recommended by International Barcode of Life (iBOL) as the standard universal barcode (57). In contrast, for plants, it is widely recognised that no single locus will achieve the same level of discrimination as the animal *CO1* thus no such universal barcode region has been established. The Consortium for the Barcode of Life (CBOL) have recommended the use of *rbcL* and *matK* as the standard plant barcodes (58) and have been supported by a number of studies (59-63). There is increasing interest in the use of the nuclear ribosomal DNA region known as the internal transcribed spacer (ITS) to target fungal taxa as it is highly variable at low taxonomic levels (56, 64).

DNA metabarcoding and high-throughput sequencing of soil communities

DNA analysis from soils is far more complex than analysis from a single animal or plant, as a soil contains DNA from multiple specimens. The application of DNA barcoding to assess a DNA mixture is termed DNA metabarcoding (65). DNA metabarcoding involves PCR amplification of a target gene using universal primers to extract genetic information from a DNA mixture that ideally represents the diversity of a particular group of taxa within a sample.

DNA metabarcoding can be used to profile soil communities by targeting specific taxonomic groups. Currently, the most commonly utilised markers for environmental samples are 16S ribosomal RNA gene region (bacteria), 18S ribosomal RNA gene region (eukaryotes), and the internal transcribed spacer 1 (ITS1) (fungi)(56). For each of these markers, curated reference databases are regularly updated to enable robust taxonomic identification of the sequences present in a sample: Greengenes (66), SILVA (67) and

UNITE (68-70) are curated databases for 16S, 18S and ITS, respectively. Traditional barcoding of single taxa using a single specimen means that high quality DNA extracts can be used, and therefore long barcode regions (>500 bp) can be amplified with high discrimination capacity. In contrast, metabarcoding generally utilises shorter regions since DNA is more degraded in environmental samples. For example, the recommended *matk* region for plants is not used as large fragment lengths (760 bp) are not easily amplified from degraded DNA in soil. Instead, the identification of plant material from soils most commonly utilises the *rbcL* and *trnL* gene regions (71); however, unfortunately no curated database has yet been developed. Curated databases are advantageous for assigning accurate taxonomic identifications as the sequences are regularly monitored to ensure that all entries are reliable. However, for environmental samples such as soils, many sequences may return as 'unknown' as many will not previously have been sequenced and so show no match to the curated database entries.

DNA metabarcoding coupled with high-throughput sequencing (HTS) offers the potential to drastically increase taxonomic resolution within soil samples compared to DNA fingerprinting and thus increase the discriminatory power between different geographic locations. Multivariate analysis methods based on a distance matrix are commonly used to visualise the similarity between different metagenomic samples. Multidimensional Scaling (MDS) used throughout this thesis is an example of unconstrained multivariate analysis and illustrates the similarity between all samples. As a result, sample variation within a single site can be observed and Analysis of Similarity (ANOSIM) statistics can be applied to determine significant differences between samples. Recent advances in HTS have revolutionized the field of genomics, making it possible to rapidly generate large amounts of sequence data at a substantially lower cost (72). Many sequencing platforms are available, utilising different combinations of template preparation, sequencing and data analysis (73-75). In all of these different approaches, platform-specific adapters and unique tags (termed indexes) are incorporated into the sequences during PCR amplification; the unique tag enables multiple samples to be sequenced simultaneously. This technology has been successfully used to identify communities from soils (34, 76-78) and has shown to detect higher diversity than traditional DNA barcoding. For example, HTS of nematode diversity in tropical rainforest identified 7700 individuals, whereas traditional barcoding from individual specimens only identified 360 individuals (79, 80). This indicates the potential of HTS to generate a more detailed composite picture of a source area. However, it is important to ensure that bioinformatic analysis does not increase diversity estimates due to sequencing errors (81, 82). Despite the clear potential to explore alternative soil communities and increase taxonomic resolution, DNA metabarcoding and HTS cannot be utilised in casework without prior validation and consideration of potential limiting factors.

Limitations of DNA metabarcoding in a forensic context

Prior to DNA analysis, there are several important considerations. For example, sufficient number and size of samples should be collected to ensure reproducible results. As primers are designed to target a wide range of taxa, it is possible to detect low levels of background DNA in laboratory reagents. Contamination should be considered and appropriate controls should be incorporated to ensure the signal detected did originate from a specific sample and was not introduced during analysis. In the laboratory, the initial DNA extraction step is crucial because biases due to incomplete cell lysis are transferred to downstream processing steps, and the loss of genetic information due to inefficient extraction could potentially be detrimental to a case. Soil DNA fingerprint profiles are strongly dependent on the DNA extraction method used (83), yet the effect of DNA extraction methods on the ability to discriminate between forensic soil samples has not yet been explored.

A specific target and gene region must be selected for PCR amplification. Although DNA barcodes often target DNA fragments >500 bp to maximise taxonomic resolution, soil DNA tends to be degraded by bacteria, and so PCR amplification from soil extracts is generally restricted to shorter genetic regions (<300 bp). As a result, the taxonomic resolution for some target organisms is often limited by the short fragment length. In addition, PCR amplification can also introduce bias as a result of primer design, polymerase enzyme choice, cycle number, copy number of a gene in different species and PCR inhibition (84-88). PCR amplification efficiency can be influenced by intrinsic differences in the amplification efficiency of templates (23) or by the self-annealing of the most abundant templates in the late stages of amplification (31). PCR inhibitors such as salts, proteases, organic solvents, humic acids and plant polysaccharides can influence the activity of the DNA polymerase and should be removed during DNA extraction. As HTS increases the taxonomic diversity detected from a single sample, such biases may be more pronounced than in traditional soil DNA fingerprinting methods. As a result, potential genetic targets for forensic application must be evaluated in terms of discriminatory power, reproducibility, spatial variability and contamination risk.

Other limitations associated with the use of DNA metabarcoding for soil forensics will be specific to a case. In many circumstances the evidence sample will be transferred to an object and removed from the crime scene. As a result, such evidence samples are not recovered for some time following the crime, during which period the soil will be exposed to various storage and environmental conditions, such as drying, which could influence the soil community (89, 90). In addition, such samples often come into contact with soils from other locations prior to recovery (91, 92). This could introduce mixing or layering of soil particles (4). Separation of soil particles based on visual properties is possible, therefore chemical analysis of soils remains feasible (93). However, mixed DNA cannot be physically separated; therefore the DNA signal of interest could be obscured or confounded. In order to determine a match between an evidence sample and a crime scene sample, reference samples are required for comparison. However, reference samples can be collected months after the crime has occurred, introducing temporal and seasonal variation in genetic signals, such as rainfall, temperature and humidity (40, 94, 95), potentially resulting in false negatives or false positives. Furthermore, the soil storage method in the laboratory following collection can alter the DNA profile (96, 97). As a result, each of these factors must be examined to determine how robust DNA metabarcoding and HTS might be in a forensic context. Knowledge of the circumstances under which soil DNA evidence would be most robust, and potential limitations should also be factored in when evaluating the strength of a match.

Validation of forensic methods for use in court

Forensic laboratories performing analysis for court can be scrutinised during cross-examination. As a result, the forensic science community, initially through the Senior Managers of Australian and New Zealand Forensic Laboratories (SMANZFL) and subsequently through SMANZFL and NIFS has developed a national laboratory accreditation program. The program is jointly managed by the National Association of Testing Authorities (NATA) and the American Association of Crime Laboratory Director's Laboratory Accreditation Board (ASCLD/LAB). Forensic laboratories accredited under this initiative must demonstrate robust, reliable and reproducible methods, and scientists must complete six-monthly proficiency tests overseen by external agencies (e.g. DNA Advisory Board standard 13.1, or National Association of Testing Authorities, Australia). Such validation ensures good laboratory practice in terms of instrument maintenance, documented Standard Operation Procedures (SOPs), chain of custody, evidence handling and casework reporting in accordance to international standards (e.g. ISO/IEC 17025:2005). However, government bodies, such as the Australian Federal Police (AFP), often collaborate with non-accredited academic and research institutes to initiate new technologies for potential use in the legal system.

Technological advance is important in ensuring the best possible scientific evidence is available to the courts. However, the introduction of new techniques into the court system requires appropriate validation, quality management as well as awareness of the underlying methodological strengths and weaknesses. Ultimately it is the court that rules on acceptance following the point of precedence (based on common law from England), i.e. the principle by which courts are obliged to follow past decisions as justification for subsequent similar situations. Therefore it is important that the Judge or Magistrate has adequate background knowledge of the basic principles associated with the evidence and methodology to successfully cross-examine or evaluate evidence in relation to a case. However, the NAS report (Strengthening Forensic Science in the US: a path forward) published in 2009 had major implications for acceptance of methods in forensic science world-wide. The overarching theme of this report was the need to only use validated scientific methods in forensic practice where there is an underlying scientific principle. Human DNA profiling was regarded as the Gold Standard due to the extensive science, and validation that underpinned its use.

The introduction of new evidence into the South Australian or Australian justice system requires general acceptance from the scientific community as well as the legal community. SWGDAM (Scientific Working Group for DNA analysis methods) makes use of criteria including Sensitivity, Stability, Specificity, Accuracy and Precision, blind trials and publication for general acceptance of new technology. As a result, three stages of method validation must be considered before DNA metabarcoding can be implemented into casework (98, 99). First, *developmental validation* is required to address the specificity, sensitivity, reproducibility and precision of a method, assessing associated biases, appropriate use of controls and reference databases. This stage should be tested using well characterised samples, and any modifications to previously validated methods should be documented. Next, *preliminary validation* should demonstrate the ability of a method to support an investigation, or corroborate information in a specific case. This usually involves the use of mock and reference samples and includes comparison of results to those achieved by alternative methods. Beyond this, any method used to generate evidence must also undergo *inferential validation* to demonstrate accurate documentation of technical procedures and robust statistical data interpretation. Validating DNA metabarcoding would involve systematic assessment of factors that may influence the result, such as environmental factors, as well as identifying the aspects of the procedure that must be controlled, such as sample storage and background contamination. The Australian justice system also requires the use of the Daubert standard i.e. the rate/chance of a false positive or negative result to provide an indication of the weight of the evidence presented. Therefore, robust statistical analysis must be developed to ascertain the confidence level associated with a result; particularly problematic in cases involving a single evidence sample.

Overview of thesis

In this thesis, I explore the use of DNA metabarcoding and HTS for forensic soil discrimination. To achieve this, five manuscripts have been compiled to explore best practice and assess the robustness of this methodology given potential limitations associated with casework.

Chapter 2

Limitations and recommendations for successful DNA extraction from forensic soil samples: A review

For forensic soil DNA analysis, it is crucial to optimise DNA extraction efficiency to maximise the information retrieved from a limited forensic sample. As soils are highly variable in composition, different DNA extraction protocols are required to efficiently extract DNA from different soil types. Chapter 2 reviews issues surrounding soil DNA extraction, with particular reference to the key interactions between DNA molecules and soil components. This review article published in *Science and Justice* discusses possible extraction modifications and highlights the considerations required prior to soil DNA extraction, as well as the potential limitations associated with forensic DNA metabarcoding.

Chapter 3

Forensic soil DNA analysis using high-throughput sequencing: a comparison of four molecular markers

Different molecular markers can be used to target specific taxonomic groups within soils. However, different groups can be more abundant, more ubiquitous and show different degrees of spatial variation. Identification of the advantages and disadvantages associated with different taxa is a necessary validation step before HTS can be applied in forensic soil analysis. Chapter 3, currently under revision at *Forensic Science International: Genetics*, compares the reproducibility and discriminatory power of four molecular markers targeting different taxa (bacterial 16S rRNA, eukaryotic18S rRNA, plant *trn*L intron and fungal internal transcribed spacer I (ITS1) rRNA) to distinguish two sites. This study demonstrates the potential use of multiple DNA markers for forensic soil analysis using HTS, and identifies some of the standardisation and evaluation steps necessary before this technique can be applied in casework.

Chapter 4

Extended cell lysis and residual soil DNA extraction detect additional fungal diversity from trace quantities of soil

Inefficient DNA extraction can markedly alter the measured abundance of taxa, or prevent some taxa from being detected altogether, reducing the genetic information recovered. Chapter 4 compares the DNA yield and fungal diversity recovered using a commercial DNA extraction protocol compared to three DNA extraction modifications of the standard protocol. This study indicates that the standard DNA extraction protocol fails to detect the full fungal diversity in a sample, and that the DNA extraction protocol should be adjusted depending on soil pH and clay content to maximise fungal diversity detected from a limited quantity of material.

Chapter 5

High-throughput sequencing of trace quantities of soil provides discriminative and reproducible fungal DNA profiles

The quantity of soil available for analysis is often limited during forensic casework. The aim of this study was to determine the minimum soil quantity required to obtain reproducible and discriminative profiles using HTS of fungal ITS rRNA. Chapter 4 examines the reproducibility and discriminatory power of DNA profiles from sample sizes as little as 50 mg, and provides a guide on the optimal mass for a given soil type based on soil texture and pH.

Chapter 6

Predicting the origin of soil evidence: high-throughput eukaryote sequencing and MIR spectroscopy applied to a crime scene scenario

Prior to this thesis, bacterial DNA was the focus of forensic soil DNA analysis. Chapters 3, 4 and 5 demonstrate the potential of soil eukaryotes for discrimination. Chapter 6 demonstrates the use of non-bacterial soil DNA to link a suspect to a mock crime scene. This experimental case study has been designed to include factors likely to be encountered during the course of an investigation. Temporal variation was incorporated by including a six week time lapse before collecting reference samples, and storage effect on evidence samples was included by storing the suspect’s belongings in a car boot for the same time span. In addition, air-drying reference samples prior to DNA extraction explored the effect of sample moisture. To determine the performance of this novel technique, this chapter compares HTS metabarcoding to a soil analysis method currently accepted in court. In addition, two multivariate analysis methods are applied to determine the most appropriate statistical approach given the forensic question and sample availability to maximise the power of soil evidence.

Chapter 7

General discussion and conclusion

This concluding chapter consolidates the most significant outcomes and highlights the contribution of my thesis to forensic soil DNA analysis. I highlight additional factors that must be tested to further validate this technique, and provide recommendations for optimising the method based on soil properties. To improve the reliability of the data analysis, I detail future bioinformatics development and stress the need for standardisation, robust reference databases to strengthen such evidence. Furthermore, I suggest that raising awareness of soil DNA evidence to investigation personnel will improve best practice from sample collection and thus increase the reliability of soil DNA evidence in court.

References

1. **Morgan, R. M., and P. A. Bull.** 2007. The philosophy, nature and practice of forensic sediment analysis. Progress in Physical Geography **31:**43-58.
2. **Concheri, G.** 2011. Chemical elemental distribution and soil DNA fingerprints provide the critical evidence in murder case investigation. PloS ONE **6:**e20222.
3. **Murray, R. C.** 2004. Evidence from the earth: forensic geology and criminal investigation. Mountain Press Publishing.
4. **Fitzpatrick, R. W., M. D. Raven, and S. T. Forrester.** 2009. A systematic approach to soil forensics: criminal case studies involving transference from crime scene to forensic evidence. Criminal and Environmental Soil Forensics**:**105-127.
5. **Fitzpatrick, R. W.** 2009. Soil: Forensic Analysis. Wiley Encyclopedia of Forensic Science**:**2377-2388.
6. **Brooks, M., and K. Newton.** 1969. Forensic pedology. Police J. **42:**107.
7. **Guedes, A.** 2011. Characterisation of soils from the Algarve region (Portugal): A multidisciplinary approach for forensic applications. Science & Justice **51:**77-82.
8. **Wanogho, S., G. Gettinby, and B. Caddy.** 1987. Particle size distribution analysis of soils using laser diffraction. Forensic Science International **33:**117-128.
9. **Pye, K., and S. J. Blott.** 2004. Particle size analysis of sediments, soils and related particulate materials for forensic purposes using laser granulometry. Forensic Science International **144:**19-27.
10. **Ruffell, A., and P. Wiltshire.** 2004. Conjunctive use of quantitative and qualitative X-ray diffraction analysis of soils and rocks for forensic analysis. Forensic science international **145:**13-23.
11. **Pye, K., and D. J. Croft.** 2007. Forensic analysis of soil and sediment traces by scanning electron microscopy and energy-dispersive x-ray analysis: An experimental investigation. Forensic Science International **165:**52-63.
12. **Dawson, L. A., and S. Hillier.** 2010. Measurement of soil characteristics for forensic applications. Surface and Interface Analysis **42:**363-377.
13. **Pye, K., S. J. Blott, D. J. Croft, and S. J. Witton.** 2007. Discrimination between sediment and soil samples for forensic purposes using elemental data: An investigation of particle size effects. Forensic Science International **167:**30-42.
14. **Cox, R., H. Peterson, J. Young, C. Cusik, and E. Espinoza.** 2000. The forensic analysis of soil organic by FTIR. Forensic Science International **108:**107-116.

15. **Chaperlin, K.** 1981. Lead content and soil discrimination in forensic science. Forensic Science International **18:**79-84.

16. **Thanasoulias, N. C., E. T. Piliouris, M. S. E. Kotti, and N. P. Evmiridis.** 2002. Application of multivariate chemometrics in forensic soil discrimination based on the UV-Vis spectrum of the acid fraction of humus. Forensic Science International **130:**73-82.

17. **Leinweber, P., and H. R. Schulten.** 1999. Advances in analytical pyrolysis of soil organic matter. J. Anal. Appl. Pyrolysis **49:**359-383.

18. **Nakayama, M., Y. Fujita, K. Kanbara, N. Nakayama, N. Mitsuo, H. Matsumoto, and T. Satoh.** 1992. Forensic chemical study on soil. (1) Discrimination of area by pyrolysis products of soil. Japanese Journal of Toxicology and Environmental Health **38:**38-44.

19. **Petraco, N., T. A. Kubic, and N. D. K. Petraco.** 2008. Case studies in forensic soil examinations. Forensic Science International **1778:**E23-E27.

20. **Pye, K., and D. J. Croft.** 2004. Forensic geoscience: introduction and overview. Geological Society, London, Special Publications **232:**1-5.

21. **Guedes, A., H. Ribeiro, B. Valentim, A. Rodrigues, H. Sant'Ovaia, I. Abreu, and F. Noronha.** 2011. Characterization of soils from the Algarve region (Portugal): A multidisciplinary approach for forensic applications. Science & Justice **51:**77-82.

22. **Sugita, R., and Y. Marumo.** 1996. Validity of color examination for forensic soil identification. Forensic Science International **83:**201-210.

23. **Fitzpatrick, R. W.** 2011. Getting the dirt: The value of soil in criminal investigations. Gazette **73:**22-23.

24. **Fitzpatrick, R. W.** 2003. Demands on soil classification in Australia. Soil Classification: a Global Desk Reference**:**77-100.

25. **McKenzie, N., D. Jacquier, R. F. Isbell, and K. Brown.** 2004. Australian Soils and Landscapes: An illustrated compendium. CSIRO Publishing, Collingwood VIC 3066 Australia.

26. **Bruce, R. G., and M. E. Dettmann.** 1996. Palynological analysis of Australian surface soild and their potential in forensic science. Forensic Science International **81:**77-94.

27. **Brown, A. G., A. Smith, and O. Elmhurst.** 2002. The combined use of pollen and soil analyses in a search and subsequent murder investigation. Journal of Forensic Sciences **47:**614-618.

28. **Mildenhall, D. C., P. E. J. Wiltshire, and V. M. Bryant.** 2006. Forensic palynology: Why do it and how it works. Forensic Science International **163:**163-172.

29. **Wiltshire, P. J.** 2009. Forensic Ecology, Botany, and Palynology: Some Aspects of Their Role in Criminal Investigation, p. 129-149. *In* K. Ritz, L. Dawson, and D. Miller (ed.), Criminal and Environmental Soil Forensics. Springer Netherlands.

30. **Coyle, H. M., C. L. Lee, W. Y. Lin, H. C. Lee, and T. M. Palmbach.** 2005. Forensic botany: Using plant evidence to aid in forensic death investigation. Croatian Medical Journal **46:**606-612.

31. **Harvey, M. L., M. W. Mansell, M. H. Villet, and I. R. Dadour.** 2003. Molecular identification of some forensically important blowflies of southern Africa and Australia. Medical and Veterinary Entomology **17:**363-369.

32. **Ferri, G.** 2009. Forensic botany: species identification of botanical trace evidence using a multigene barcoding approach. international Journal of Legal Medicine **123:**395-401.

33. **Gangneux, C., M. Akpa-Vinceslas, H. Sauvage, S. Desaire, S. Houot, and K. Laval.** 2011. Fungal, bacterial and plant dsDNA contributions to soil total DNA extracted from silty soils under different farming practices: Relationships with chloroform-labile carbon. Soil Biology & Biochemistry **43:**431-437.

34. **Taberlet, P., E. Coissac, M. Hajibabaei, and L. H. Rieseberg.** 2012. Environmental DNA. Mol. Ecol. **21:**1789-1793.

35. **Sensabaugh, G. F.** 2009. Microbial Community Profiling for the Characterisation of Soil Evidence: Forensic Considerations. Springer, Dordrecht.

36. **Horswell, J.** 2002. Forensic comparison of soils by bacterial community DNA profiling. Journal of Forensic Sciences **47:**350-353.

37. **Cano, R. J.** 2010. Molecular Microbial Forensics. Royal Soc Chemistry, Cambridge.

38. **Lerner, A., Y. Shor, A. Vinokurov, Y. Okon, and E. Jurkevitch.** 2006. Can denaturing gradient gel electrophoresis (DGGE) analysis of amplified 16S rDNA of soil bacterial populations be used in forensic investigations? Soil Biology & Biochemistry **38:**118-1192.

39. **Moreno, L. I., D. Mills, J. Fetscher, K. John-Williams, L. Meadows-Jantz, and B. McCord.** 2011. The application of amplicon length heterogeneity PCR (LH-PCR) for monitoring the dynamics of soil microbial communities associated with cadaver decomposition. J. Microbiol. Methods **84:**388-393.

40. **Meyers, M. S., and D. R. Foran.** 2008. Spatial and temporal influences on bacterial profiling of forensic soil samples. Journal of Forensic Sciences **53:**652-660.

41. **Lenz, E. J., and D. R. Foran.** 2010. Bacterial profiling of soil using genus-specific markers and multidimensional scaling. Journal of Forensic Sciences **55:**1437-1442.

42. **Macdonald, C. A.** 2011. Discrimination of soils at regional and local levels using bacterial and fungal t-RFLP profiling. Journal of Forensic Sciences **56:**61-69.

43. **Liu, W. T., T. L. Marsh, H. Cheng, and L. J. Forney.** 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl. Environ. Microbiol. **63:**4516-4522.

44. **Moeseneder, M. M., J. M. Arrieta, G. Muyzer, C. Winter, and G. J. Herndl.** 1999. Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. Appl. Environ. Microbiol. **65:**3518-3525.

45. **Osborn, A. M., E. R. B. Moore, and K. N. Timmis.** 2000. An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. Environ. Microbiol. **2:**39-50.

46. **Moreno, L. I., D. K. Mills, J. Entry, R. T. Sautter, and K. Mathee.** 2006. Microbial metagenome profiling using amplicon length heterogeneity-polymerase chain reaction proves more effective than elemental analysis in discriminating soil specimens. Journal of Forensic Sciences **51:**1315-1322.

47. **Kirk, J. L., L. A. Beaudette, M. Hart, P. Moutoglis, J. N. Klironomos, H. Lee, and J. T. Trevors.** 2004. Methods of studying soil microbial diversity. J. Microbiol. Methods **58:**169-188.

48. **Macdonald, C. A., R. Ang, S. J. Cordiner, and J. Horswell.** 2011. Discrimination of Soils at Regional and Local Levels Using Bacterial and Fungal T-RFLP Profiling. Journal of Forensic Sciences **56:**61-69.

49. **Macdonald, C. A., R. Ang, S. J. Cordiner, and J. Horswell.** 2011. Discrimination of Soils at Regional and Local Levels Using Bacterial and Fungal T-RFLP Profiling*. Journal of Forensic Sciences **56:**61-69.

50. **Heath, L. E., and V. A. Saunders.** 2008. Spatial variation in bacterial DNA profiles for forensic soil comparisons. Canadian Society of Forensic Science Journal **41:**29-37.

51. **Quaak, F. C. A., and I. Kuiper.** 2011. Statistical data analysis of bacterial t-RFLP profiles in forensic soil comparisons. Forensic Science International **210:**96-101.

52. **Thies, J. E.** 2007. Soil microbial community analysis using terminal restriction fragment length polymorphisms. Soil Sci. Soc. Am. J. **71:**579-591.

53. **Hebert, P. D., A. Cywinska, and S. L. Ball.** 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences **270:**313-321.

54. **Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. DeWaard.** 2003. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. Ser. B-Biol. Sci. **270:**313-321.

55. **Newmaster, S. G., A. J. Fazekas, and S. Ragupathy.** 2006. DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. Canadian Journal of Botany **84:**335-341.

56. **Epp, L. S., S. Boessenkool, E. P. Bellemain, J. Haile, A. Esposito, T. Riaz, C. Erseus, V. I. Gusarov, M. E. Edwards, A. Johnsen, H. K. Stenoien, K. Hassel, H. Kauserud, N. G. Yoccoz, K. Brathen, E. Willerslev, P. Taberlet, E. Coissac, and C. Brochmann.** 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Mol. Ecol. **21:**1821-1833.

57. **Rougerie, R., T. Decaens, L. Deharveng, D. Porco, S. W. James, C. H. Chang, B. Richard, M. Potapov, Y. Suhardjono, and P. D. N. Hebert.** 2009. DNA barcodes for soil animal taxonomy. Pesqui. Agropecu. Bras. **44:**789-802.

58. **Hollingsworth, P. M., L. L. Forrest, J. L. Spouge, M. Hajibabaei, S. Ratnasingham, M. van der Bank, M. W. Chase, R. S. Cowan, D. L. Erickson, A. J. Fazekas, S. W. Graham, K. E. James, K.-J. Kim, W. J. Kress, H. Schneider, J. van AlphenStahl, S. C. H. Barrett, C. van den Berg, D. Bogarin, K. S. Burgess, K. M. Cameron, M. Carine, J. Chacon, A. Clark, J. J. Clarkson, F. Conrad, D. S. Devey, C. S. Ford, T. A. J. Hedderson, M. L. Hollingsworth, B. C. Husband, L. J. Kelly, P. R. Kesanakurti, J. S. Kim, Y.-D. Kim, R. Lahaye, H.-L. Lee, D. G. Long, S. Madrinan, O. Maurin, I. Meusnier, S. G. Newmaster, C.-W. Park, D. M. Percy, G. Petersen, J. E. Richardson, G. A. Salazar, V. Savolainen, O. Seberg, M. J. Wilkinson, D.-K. Yi, D. P. Little, and C. P. W. Grp.** 2009. A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the United States of America **106:**12794-12797.

59. **Burgess, K. S., A. J. Fazekas, P. R. Kesanakurti, S. W. Graham, B. C. Husband, S. G. Newmaster, D. M. Percy, M. Hajibabaei, and S. C. H. Barrett.** 2011. Discriminating plant species in a local temperate flora using the *rbcL + matK* DNA barcode. Methods in Ecology and Evolution **2:**333-340.

60. **von Crautlein, M., H. Korpelainen, M. Pietilainen, and J. Rikkinen.** 2011. DNA barcoding: a tool for improved taxon identification and detection of species diversity. Biodiversity and Conservation **20:**373-389.

61. **Hollingsworth, P. M., S. W. Graham, and D. P. Little.** 2011. Choosing and using a plant DNA barcode. PloS ONE **6:**e19254, doi:10.1371/journal.pone.0019254.

62. **Yu, J., J.-H. Xue, and S.-L. Zhou.** 2011. New universal *matK* primers for DNA barcoding angiosperms. Journal of Systematics and Evolution **49:**176-181.

63. **Gao, T., Z. Sun, H. Yao, J. Song, Y. Zhu, X. Ma, and S. Chen.** 2011. Identification of Fabaceae plants using the DNA barcode *matK*. Planta Medica **77:**92-94.

64. **Blaalid, R., S. Kumar, R. Nilsson, K. Abarenkov, P. Kirk, and H. Kauserud.** 2013. ITS1 versus ITS2 as DNA metabarcodes for fungi. Molecular ecology resources **13:**218-224.

65. **Valentini, A., F. Pompanon, and P. Taberlet.** 2008. DNA barcoding for ecologists. Trends Ecol. Evol. **24:**110-117.

66. **McDonald, D., M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight, and P. Hugenholtz.** 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME journal **6:**610-618.

67. **Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner.** 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research **41:**D590-D596.

68. **Kõljalg, U., K. H. Larsson, K. Abarenkov, R. H. Nilsson, I. J. Alexander, U. Eberhardt, S. Erland, K. Høiland, R. Kjøller, and E. Larsson.** 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytol. **166:**1063-1068.

69. **Kõljalg, U., R. H. Nilsson, K. Abarenkov, L. Tedersoo, A. F. S. Taylor, M. Bahram, S. T. Bates, T. D. Bruns, J. Bengtsson-Palme, T. M. Callaghan, B. Douglas, T. Drenkhan, U. Eberhardt, M. Dueñas, T. Grebenc, G. W. Griffith, M. Hartmann, P. M. Kirk, P. Kohout, E. Larsson, B. D. Lindahl, R. Lücking,

M. P. Martín, P. B. Matheny, N. H. Nguyen, T. Niskanen, J. Oja, K. G. Peay, U. Peintner, M. Peterson, K. Põldmaa, L. Saag, I. Saar, A. Schüßler, J. A. Scott, C. Senés, M. E. Smith, A. Suija, D. L. Taylor, M. T. Telleria, M. Weiss, and K.-H. Larsson. 2013. Towards a unified paradigm for sequence-based identification of fungi. Mol. Ecol. **22:**5271-5277.

70. **Abarenkov, K., R. Henrik Nilsson, K. H. Larsson, I. J. Alexander, U. Eberhardt, S. Erland, K. Høiland, R. Kjøller, E. Larsson, and T. Pennanen.** 2010. The UNITE database for molecular identification of fungi–recent updates and future perspectives. New Phytol. **186:**281-285.

71. **Hollingsworth, M. L., A. A. Clark, L. L. Forrest, J. Richardson, R. T. Pennington, D. G. Long, R. Cowan, M. W. Chase, M. Gaudeul, and P. M. Hollingsworth.** 2009. Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Molecular Ecology Resources **9:**439-457.

72. **Logares, R., T. H. Haverkamp, S. Kumar, A. Lanzén, A. J. Nederbragt, C. Quince, and H. Kauserud.** 2012. Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. J. Microbiol. Methods **91:**106-113.

73. **Magi, A., M. Benelli, A. Gozzini, F. Girolami, F. Torricelli, and M. L. Brandi.** 2010. Bioinformatics for Next Generation Sequencing Data. Genes **1:**294-307.

74. **Metzker, M. L.** 2010. Applications of Next Generation Sequencing - the next generation. Nat. Rev. Genet. **11:**31-46.

75. **Jünemann, S., F. J. Sedlazeck, K. Prior, A. Albersmeier, U. John, J. Kalinowski, A. Mellmann, A. Goesmann, A. von Haeseler, and J. Stoye.** 2013. Updating benchtop sequencing performance comparison. Nat. Biotechnol. **31:**294-296.

76. **Epp, L. S., S. Boessenkool, E. P. Bellemain, J. Haile, A. Esposito, T. Riaz, C. ErsÉUs, V. I. Gusarov, M. E. Edwards, A. Johnsen, H. K. StenØIen, K. Hassel, H. Kauserud, N. G. Yoccoz, K. A. BrÅThen, E. Willerslev, P. Taberlet, E. Coissac, and C. Brochmann.** 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Mol. Ecol. **21:**1821-1833.

77. **Andersen, K., K. L. Bird, M. Rasmussen, J. Haile, H. Breuning-Madsen, K. H. KjÆR, L. Orlando, M. T. P. Gilbert, and E. Willerslev.** 2011. Meta-barcoding

of ‘dirt’ DNA from soil reflects vertebrate biodiversity. Mol. Ecol.**:**doi: 10.1111/j.1365-294X.2011.05261.x.

78. **Taberlet, P., S. Prud'homme, E. Campione, J. Roy, C. Miquel, W. Shehzad, L. Gielly, D. Rioux, P. Choler, J. Clément, C. Melodelima, F. Pompanon, and E. Coissac.** 2011. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Dryad Data Repository *(unpublished data)*.

79. **Porazinska, D. L., R. M. Giblin-Davis, A. Esquivel, T. O. Powers, W. Sung, and W. K. Thomas.** 2010. Ecometagenetics confirms high tropical rainforest nematode diversity. Mol. Ecol. **19:**5521-5530.

80. **Porazinska, D. L., R. M. Giblin-Davis, W. Sung, and W. K. Thomas.** 2010. Linking operational clustered taxonomic units (OCTUs) from parallel ultra-sequencing (PUS) to nematode species. Zootaxa **2427:**55-63.

81. **Unterseher, M., A. Jumpponen, M. Opik, L. Tedersoo, M. Moora, C. F. Dormann, and M. Schnittler.** 2011. Species abundance distributions and richness estimations in fungal metagenomics - lessons learned from community ecology. Mol. Ecol. **20:**275-285.

82. **Kunin, V., A. Engelbrektson, H. Ochman, and P. Hugenholtz.** 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. **12:**118-123.

83. **Terrat, S., R. Christen, S. Dequiedt, M. Lelièvre, V. Nowak, T. Regnier, D. Bachar, P. Plassart, P. Wincker, and C. Jolivet.** 2012. Molecular biomass and MetaTaxogenomic assessment of soil microbial communities as influenced by soil DNA extraction procedure. Microbial biotechnology **5:**135-141.

84. **Martin-Laurent, F., L. Philippot, S. Hallet, R. Chaussod, J. C. Germon, G. Soulas, and G. Catroux.** 2001. DNA extraction from soils: Old bias for new microbial diversity analysis methods. Appl. Environ. Microbiol. **67:**2354-2359.

85. **Egert, M., and M. W. Friedrich.** 2003. Formation of pseudo-terminal restriction fragments, a PCR-related bias affecting terminal restriction fragment length polymorphism analysis of microbial community structure. Appl. Environ. Microbiol. **69:**2555-2562.

86. **Sipos, R., A. J. Szekely, M. Palatinszky, S. Revesz, K. Marialigeti, and M. Nikolausz.** 2007. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol. Ecol. **60:**341-50.

87. **Hong, S., J. Bunge, C. Leslin, S. Jeon, and S. S. Epstein.** 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. ISME J **3:**1365-1373.

88. **Klappenbach, J. A., J. M. Dunbar, and T. M. Schmidt.** 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl. Environ. Microbiol. **66:**1328-33.

89. **Bainard, L. D., J. N. Klironomos, and M. M. Hart.** 2010. Differential effect of sample preservation methods on plant and arbuscular mycorrhizal fungal DNA. J. Microbiol. Methods **82:**124-130.

90. **Macdonald, L. M., B. K. Singh, N. Thomas, M. J. Brewer, C. D. Campbell, and L. A. Dawson.** 2008. Microbial DNA profiling by multiplex terminal restriction fragment length polymorphism for forensic comparison of soil and the influence of sample condition. J. Appl Microbiol **105:**813-821.

91. **Bull, P. A., A. Parker, and R. M. Morgan.** 2006. The forensic analysis of soils and sediment taken from the cast of a footprint. Forensic Science International **162:**6-12.

92. **Croft, D. J., and K. Pye.** 2004. Multi-technique comparison of source and primary transfer soil samples: an experimental investigation. Science & Justice **44:**21-28.

93. **Morgan, R. M., J. Robertson, C. Lennard, K. Hubbard, and P. A. Bull.** 2010. Quartz grain surface textures of soils and sediments from Canberra, Australia: A forensic reconstruction tool. Aust. J. Forensic Sci. **42:**169-179.

94. **Darby, B. J., D. A. Neher, D. C. Housman, and J. Belnap.** 2011. Few apparent short-term effects of elevated soil temperature and increased frequency of summer precipitation on the abundance and taxonomic diversity of desert soil micro- and meso-fauna. Soil Biology & Biochemistry **43:**1474-1481.

95. **Baker, K. L., S. Langenheder, G. W. Nicol, D. Ricketts, K. Killham, C. D. Campbell, and J. I. Prosser.** 2009. Environmental and spatial characterisation of bacterial community composition in soil to inform sampling strategies. Soil Biol. Biochem. **41:**2292-2298.

96. **Lee, Y. B., N. Lorenz, L. K. Dick, and R. P. Dick.** 2007. Cold storage and pretreatment incubation effects on soil microbial properties Soil Sci. Soc. Am. J. **71:**1299-1305.

97. **Pesaro, M., F. Widmer, G. Nicollier, and J. Zeyer.** 2003. Effects of freeze–thaw stress during soil storage on microbial communities and methidathion degradation. Soil Biol. Biochem. **35:**1049-1061.

98. **Butler, J. M.** 2005. Forensic DNA typing: biology, technology, and genetics of STR markers. Academic Press.

99. **Bruce Budowle, S. E. S., R G. Breeze, P S. Keim, and S A . Morse.** 2010. Validation of microbial forensics in scientific, legal, and policy context. Microbial Forensics. *In* B. Budowle (ed.), Second ed. Elsevier. 649-663.

CHAPTER 2

Limitations and recommendations for successful DNA extraction from forensic soil samples: A review

Young, J. M., N. J. Rawlence, L. S. Weyrich, and A. Cooper**.** (2014). Limitations and recommendations for successful DNA extraction from forensic soil samples: A review. *Science & Justic*e. 54.3: 238-244.

Statement of authorship

Limitations and recommendations for successful DNA extraction from forensic soil samples: A review

Published in Science and Justice, February 2014

Jennifer M. Young (Candidate)

Reviewed literature, created the table, and wrote the paper.

I hereby certify that the statement of contribution is accurate

Signed. ...... Date.....23/06/2014

Nicholas J. Rawlence

Provided advice on content and edited manuscript

I hereby certify that the statement of contribution is accurate

Signed Date.....23/06/2014

Laura S. Weyrich

Provided advice on structure and edited manuscript

I hereby certify that the statement of contribution is accurate

Signed ... Date.....23/06/2014

Alan Cooper

Edited the manuscript

I hereby certify that the statement of contribution is accurate

Signed Date…..23/06/2014

Young, J.M., Rawlence, N.J., Weyrich, L.S. & Cooper, A. (2014) Limitations and recommendations for successful DNA extraction from forensic soil samples: a review.
Science and Justice, v. 54(3), pp. 238-244

CHAPTER 3

Forensic soil DNA analysis using high-throughput sequencing: a comparison of four molecular markers

Young, J.M., L.S. Weyrich, A. Cooper. (2014). Forensic soil DNA analysis using high-throughput sequencing: a comparison of four molecular markers. *Forensic Science International: Genetics.* 13: 176-184.

Statement of authorship

Forensic soil DNA analysis using high-throughput sequencing: a comparison of four molecular markers

In review at *Forensic Science International: Genetics*

Jennifer M. Young (Candidate)

Performed sample collection, DNA extractions, PCR amplifications, and sequencing, selected molecular markers, conducted downstream processing and analysis of data, interpreted the results, created tables, and wrote the paper.

I hereby certify that the statement of contribution is accurate

Signed. .. Date.....04/07/2014.....

Laura S. Weyrich

Provided advice on data analysis and edited manuscript

I hereby certify that the statement of contribution is accurate

Signed .... Date.....04/07/2014

Alan Cooper

Edited the manuscript

I hereby certifv that the statement of contribution is accurate

Signed........ Date.....04/07/2014

Young, J.M., Weyrich, L.S. & Cooper, A. (2014) Forensic soil DNA analysis using high-throughput sequencing: a comparison of four molecular markers.
Forensic Science International: Genetics, v. 13(November), pp. 176-184

CHAPTER 4

Extended cell lysis and residual soil DNA extraction detect additional fungal diversity from trace quantities of soil

Young, J.M., L.S. Weyrich, L. J. Clarke, A. Cooper. DNA extraction modifications and soil re-extraction enhance fungal diversity estimates for forensic analysis of trace samples.
(In review at Forensic Science, Medicine and Pathology)

Statement of authorship

Extended cell lysis and residual soil DNA extraction detect additional fungal diversity from trace quantities of soil

In preparation for *Australian Journal of Forensic Sciences*

Jennifer M. Young (Candidate)

Designed experiment, DNA extractions, PCR amplifications, and sequencing, conducted downstream processing and analysis of data, interpreted the results, created tables, and wrote the paper.

I hereby certify that the statement of contribution is accurate

Signed. . Date.....04/07/2014

Laura S. Weyrich

Provided advice on interpretation of results, content, structure, edited manuscript

I hereby certify that the statement of contribution is accurate

Signed .... Date.....04/07/2014

Laurence J. Clarke

Edited the manuscript

I hereby certify that the statement of contribution is accurate

Signed... .... Date.....04/07/2014

Alan Cooper

Edited the manuscript

I hereby certifv that the statement of contribution is accurate

Signed. Date.….04/07/2014

Abstract

High-throughput sequencing technology provides a means to generate detailed surveys of the soil microbial community. However, inefficient DNA extraction can markedly alter the measured abundance of taxa, or prevent some taxa from being detected altogether. During forensic soil analysis, maximising the genetic information recovered and capturing an accurate representation of the diversity from limited quantities of soil is vital to produce robust, reproducible comparisons between forensic samples. In this study, we use High-throughput sequencing (HTS) of the internal transcribed spacer I (ITS1) ribosomal DNA, to compare the performance of a standard commercial DNA extraction kit (MOBIO PowerSoil DNA Isolation kit) and three modified protocols of this kit: soil pellet re-extraction (RE); an additional 24-hour lysis incubation step at room temperature (RT); and 24-hour lysis incubation step at 55°C (55). We show that DNA yield is not correlated with the resulting fungal diversity detected and that the optimal DNA extraction protocol varies depending on soil pH and clay content. The four DNA extraction methods displayed distinct fungal community profiles for individual samples, with many phyla detected exclusively using the modified methods. This suggests that standardization of these protocols for forensic analysis may involve specific extractions for different soil types. Furthermore, our results indicate that the application of multiple DNA extraction methods will provide a more complete inventory of fungal biodiversity, and in particular, that re-extraction of the residual soil pellet offers a novel tool for forensic soil analysis when only trace quantities are available.

Keywords: soil, forensics, metagenomics, DNA extraction, high-throughput sequencing

Introduction

Soil is a powerful form of contact trace evidence that can link a suspect to a location, object or victim when reference samples are available (1), or alternatively, provide information on the likely geographical origin of a forensic sample in the absence of a reference (2,3). DNA fingerprinting methods, such as T-RFLP, have limited resolution and individual taxa cannot be identified. This approach relies on differences in DNA fragment lengths between different taxa and does not identify individual sequences: multiple species can co-migrate and be represented by a single peak. High-throughput sequencing (HTS) offers a means to detect a more detailed picture of the soil community. However, obtaining an accurate representation of microbial soil communities has proven problematic due to difficulties in recovering DNA from complex soil matrices (4-7).

Although many DNA extraction protocols have been developed to improve DNA recovery from soils, studies have shown significant biases in bacterial DNA profiles when different protocols are applied (8-11). Commercial soil DNA extraction kits provide an easy and effective means for studying soil diversity; the PowerSoil DNA Isolation kit (MOBIO, Carlsbad, CA, USA,) is widely used because it efficiently removes polymerase chain reaction (PCR) inhibitors using a patented Inhibitor Removal Technology (IRT) which causes flocculation of humic acids, proteins and polysaccharides which can be separated from the DNA supernatant upon centrifugation. However, studies have shown that portions of the endogenous DNA are not captured using the standard extraction kit protocols (12). For example, by examining the band patterns from PCR-DGGE (Denaturing Gradient Gel Electrophoresis) of the 16S ribosomal DNA, Jiang *et al*. (13) showed that some protocols successfully extract DNA from Archaea and certain bacterial phyla, while failing to detect fungal taxa. The efficiency of the initial lysis step is crucial in releasing intracellular DNA into the solution, and thus it is important to identify the treatment that provides the most complete diversity. An alternative extraction approach to increase DNA yield and profile diversity is to perform successive extractions of the residual soil pellet (12, 14). Jones *et al.* (14) described quantitative differences in bacterial diversity when successive DNA extractions were completed and sequenced. Furthermore, Feinstein *et al.* (12) used high-throughput sequencing (HTS) to demonstrate that successive extractions of a soil sample were dominated by different bacterial phyla, although no variation in fungal community composition was observed. These studies demonstrate that methods incorporating re-extraction can potentially increase DNA yield and provide a more accurate representation of soil bacterial diversity. However, it would be desirable to establish standardized extraction procedures that detect a reproducible and accurate measure of soil fungal diversity.

Fungal profiles are of particular interest for forensic applications because fungal diversity have been shown to provide better discrimination between soil samples than bacterial profiles (15). MacDonald *et al.* (15) showed that fungal DNA profiles generated using Terminal Restriction Fragment Length Polymorphism (T-RFLP) were more discriminative than bacterial or archaeal DNA profiles. In addition, the effects of air-drying (30°C for 5 days) soil prior to DNA extraction were negligible for fungi but significantly impacted bacterial DNA profiles. This suggests that fungal DNA profiles are more resilient to desiccation, and therefore can provide a more robust target for forensic soil analysis due to the protein coat. Incomplete detection of soil fungal DNA and biases associated with different DNA extraction methods are current limitations of soil DNA analysis. Therefore, to establish fungal DNA profiling as a method for forensic soil analysis, we need to ensure the chosen DNA extraction method provides a complete and unbiased profile of the fungal DNA.

Using limited soil material (250 mg) to mimic forensic casework, this study examined the impact of different DNA extraction methods to enhance DNA recovery, fungal diversity, and the ability to discriminate between soil samples. Using HTS of the internal transcribed spacer I (ITS1) ribosomal DNA, we compared the performance of a standard commercial DNA extraction kit (MOBIO PowerSoil DNA Isolation kit) and three modified DNA extraction methods applied to MOBIO: soil pellet re-extraction (RE); an additional 24-hour lysis incubation step at room temperature (RT); and 24-hour lysis incubation step at 55°C (55). To identify which modified DNA extraction method maximized DNA recovery, DNA yield and fungal diversity were assessed using five soil samples with distinct physical and chemical properties. The five soil samples varied in soil pH and clay content as these are known drivers of DNA interactions with soil particles. We demonstrate that a standard DNA extraction using the PowerSoil DNA Isolation kit fails to detect specific fungal taxa and that the optimal DNA extraction protocol involves modifications and variations depending on soil pH and clay content.

Materials and Methods

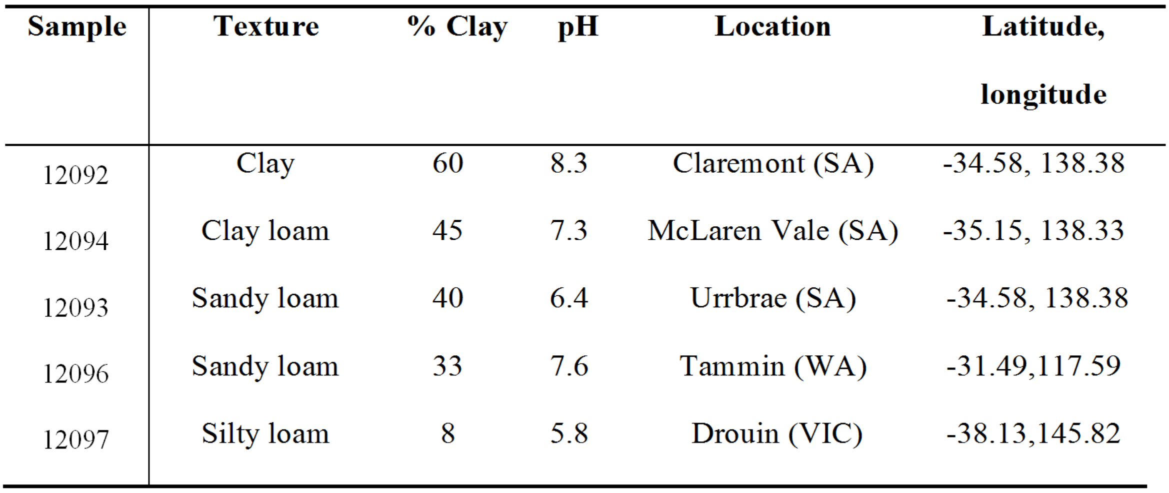
Sample collection and DNA extraction

Soil samples with different chemical properties (Table 1) were collected from the upper layer of soil (0-20 cm) at five sites across Australia using a sterile screw cap plastic container and stored at 4 ºC prior to extraction. For each soil, pH was measured using 1:5 soil/water extracts and the particle size distribution was determined using a sedimentation approach. Based on percentage of sand, silt and clay present each sample was classified according to the soil textural triangle. This chemical analysis was performed by Alla

Marchuk (PhD Candidate) from School of Agriculture, Food and Wine, University of Adelaide.

From each bulk sample, 250 mg was sub-sampled directly into a bead tube provided in the PowerSoil DNA Isolation kit (MOBIO, Carlsbad, CA, USA) using DNA-free disposable spatula and the manufacturer's protocol was followed (MB). The Precellys 24 homogeniser (Bertin Technologies, Saint-Quentin-en-Yvelines, France) was used for the bead-beating lysis step (two rounds of 30 seconds at 5500 rpm). The residual soil pellet from MB was re-extracted (RE) following the same protocol. In addition, two modified versions of the protocol were performed involving a 24 hour lysis incubation step at room temperature (RT) or 55°C (55). An extraction blank was processed in parallel to samples for each protocol. DNA yield of each extract was quantified using the NanoDrop 2000 Spectrophotometer (Thermo Scientific, Wilmington, USA) by taking an average of two measurements.

Table 1: Sampling location and chemical properties of the soils used in this study.



Sample	Texture	% Clay	pH	Location	Latitude, longitude
12092	Clay	60	8.3	Claremont (SA)	-34.58, 138.38
12094	Clay loam	45	7.3	McLaren Vale (SA)	-35.15, 138.33
12093	Sandy loam	40	6.4	Urrbrae (SA)	-34.58, 138.38
12096	Sandy loam	33	7.6	Tammin (WA)	-31.49,117.59
12097	Silty loam	8	5.8	Drouin (VIC)	-38.13,145.82

PCR amplification and library preparation

PCR was used to amplify the internal transcribed spacer I (ITS1) using universal fungal primers ITS5 (5’-CCTCTCTATGGGCAGTCGGTGATGGAAGTAAAAGTCGTAACAAGG-3’) and 5.8S_fungi (5’-CCATCTCATCCCTGCGTGTCTCCGACTCAG**nnnnnnn**CAAGAGATCCGTTGTTGAAAGTT-3’) (16). The primers were modified to include Ion Torrent sequencing adapters (underlined: P1 adapter on the ITS5 primer; A adapter on the 5.8S_fungi primer), and a unique seven base pair multiplex identifier (in bold) (17) which was used to separate sequences by sample and extraction method during data analysis. PCR amplification of each DNA extract was performed in triplicate in a 25 µl reaction mix containing 2.5 mM $MgCl_2$, 0.24 mM dNTPs, 0.24 µm of each primer, 0.4 mg/µl BSA, 0.5 U Amplitaq Gold DNA polymerase in 10x reaction buffer (Applied Biosystems, Melbourne, Australia), and 1 µl DNA extract. Reactions were PCR-amplified using an initial denaturation of 9 mins at 94 ºC, followed by 35 cycles of 94 ºC for 30 sec, 54 ºC for 30 sec, and 72 ºC for 45 sec, and a final extension at 72 ºC for 7 mins. A no-template control was additionally included for each MID tag. Agarose gel electrophoresis revealed no PCR products in the no-template or extraction blank controls and were omitted from further analysis. Triplicate PCR products were pooled to minimise PCR bias (18) and purified using Agencourt AMPure XP PCR Purification kits (Beckman Coulter Genomics, Australia). Purified PCR products were quantified using the HS dsDNA Qubit Assay on a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and pooled to equimolar concentration. The amplicon library concentration was measured using the HS D1K Tapestation (Agilent Technologies) and diluted to 11.6 pM. Emulsion PCR and Ion Sphere Particle enrichment were performed on the Ion OneTouch system^{TM} (Life Technologies) using the Ion

OneTouch ^{TM} 200 Template Kit v2 DL. Sequencing was carried out on the Ion Torrent Personal Genome Machine^{TM} (Life Technologies) using the Ion PGM ^{TM} 200 Sequencing Kit and an Ion 316^{TM} semiconductor chip (Life Technologies).

Data Analysis

Following base calling on the Torrent Suite v3.4.2 (Life Technologies), sequence reads were exported and de-multiplexed using the fastx_barcode_splitter tool (FASTX-toolkit v0.0.12; http://hannonlab.cshi.edu/fastx_toolkit). Cutadapt v1.1 (19) was used to trim primer sequences, and fastx_clipper tool was used to remove sequences <100 bp (parameters; –Q33 –l 100). We used a strict zero mismatch threshold for both the MID tag and primer sequences. Reads which had a Phred score less than 20 for 90% of the sequences were removed using fastq_quality_filter tool (FASTX-toolkit v0.0.12). The resulting fastq files were converted into .fna formatted files (script available from http://genomics.azcc.arizona.edu/help.php3) for analysis in QIIME v.1.5.0 (20). Reads were *de novo* clustered at 97% identity to create Operational Taxonomic Units (OTUs) using UCLUST (21), with the most abundant read in each cluster used as the representative sequence. Since de-noising of PGM sequence reads is problematic (22), we aimed to minimise the effect of sequencing error by applying stringent quality filtering, clustering at 97% similarity and describing OTUs at high taxonomic levels (class or above). To compare OTU count and composition between samples, the OTU table was rarefied to exclude differences as a result of sequencing depth.

To visualise OTU overlap between different extraction methods, a Venn diagram was generated for each sample (*http://bioinfogp.cnb.csic.es/tools/venny/index.html)*. The number of OTUs detected was plotted against DNA yield (ng/mg) to explore association of these two factors. Using non-rarefied data, we assigned taxonomy to OTUs using the UNITE database (23) for molecular identification of fungi (*unite_ref_seqs_21nov2011.fasta, accessed 2011*). Differences in fungal taxonomic composition between extracts were visualised both at phyla and class level.

The OTUs detected with each extraction method were compared to determine differences in discriminatory power. A Bray-Curtis cluster dendrogram was generated in PRIMER6 (PRIMER-E, Plymouth Routines in Multivariate Ecological Research v. 6, *PRIMER-E Ltd*, Luton, UK) with default parameters to visualise the similarity between samples using each protocol. To compare the discriminatory power between samples using each method, pair-wise Bray-Curtis distances were calculated for each pair of soils. Differences in discriminatory power between protocols were measured as the mean pairwise Bray-Curtis distance, and statistical significance was determined using one-way ANOVA in SPSS Statistics 21 software package (IBM, USA).

Results

DNA yield increased with extraction modifications

DNA yield from a standard extraction kit was compared to the DNA concentrations recovered from three modified protocols of the MOBIO PowerSoil DNA Isolation kit (re-extraction of the residual soil pellet (RE), 24 hour lysis at room temperature (RT) or 24 hour lysis at 55°C (55). This was completed for five different soil samples with contrasting chemical properties to assess the efficiency of each modification across a broad range of soil types, as a likely scenario for a forensic case study. Each alternative method increased the DNA yield compared to the standard kit extraction (MB). Although re-extraction of the residual soil pellet substantially increased the DNA yield from all samples compared to the standard method, both lysis incubation protocols (RT and 55) also increased yields relative to the standard protocol (Fig. 1A). This demonstrates that a standard DNA extraction kit protocol fails to recover all DNA present within a sample, and that DNA recovery can be enhanced by applying a 24 hour incubation step. Of the two protocols that modified incubation temperatures, the highest DNA yield appeared to reflect soil pH. For soils with a pH >7.5 (12092 and 12096), the highest DNA yield was observed with lysis at 55°C, whereas soils with pH <7.5 (12093, 12094 and 12097) yielded more DNA with lysis at room temperature. In addition, the influence of lysis temperature on DNA recovery was more pronounced in high clay content soils (12092 and 12094), as the DNA yield was increased by > 20% by altering the temperature.

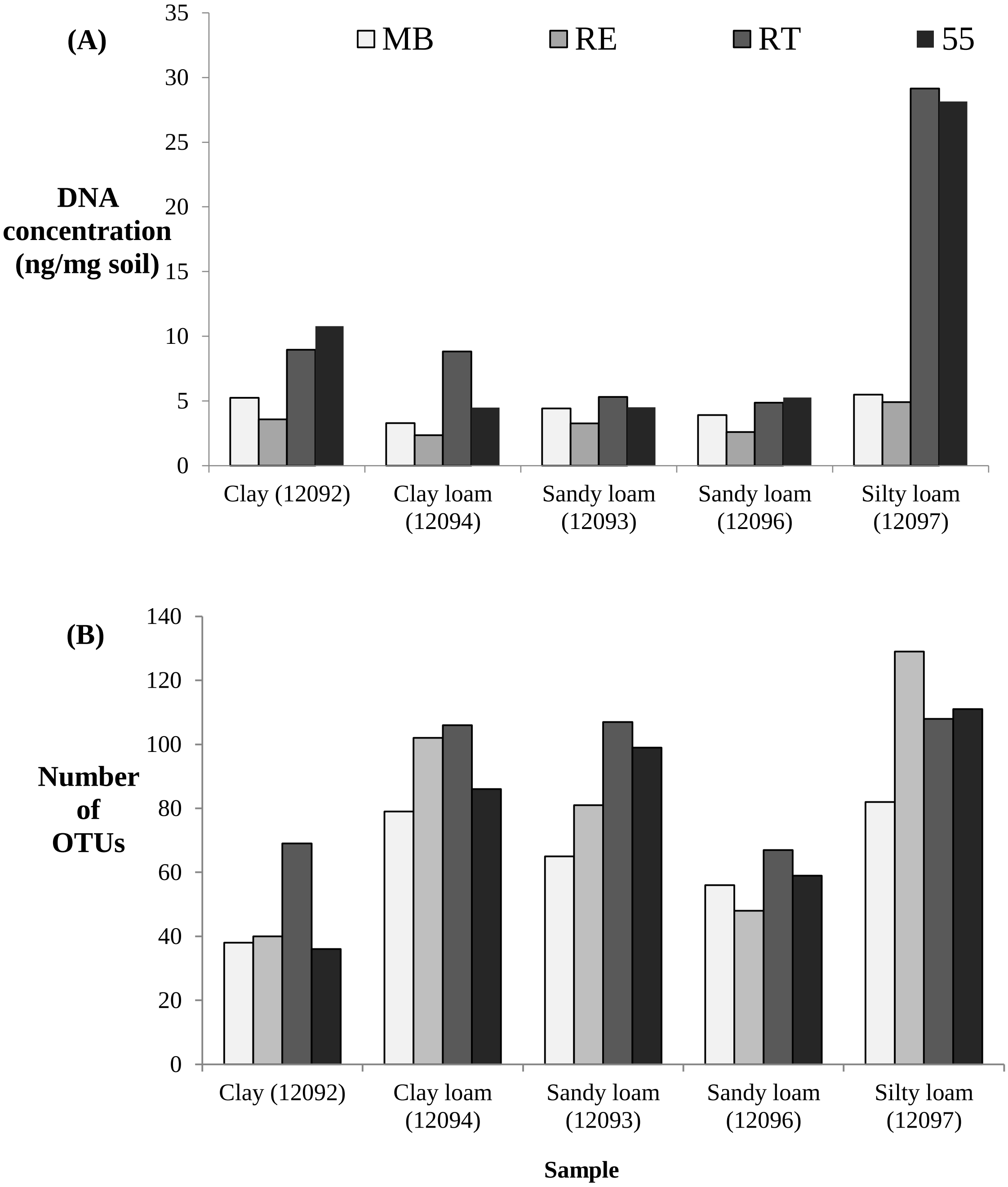


Fig. 1: (A) DNA yield and (B) number of OTUS detected using a commercial DNA extraction kit (MOBIO PowerSoil DNA Isolation kit, MB) and three modified protocols: extraction of the residual soil pellet from MB (RE); 24 hr lysis step at room temperature (RT); and 24 hr lysis step at 55°C (55). Number of OTUs based on rarefied data set / OTU table.

Fungal diversity increased with extraction modifications

HTS of the ITS amplicons yielded 2,099,417 sequences, ranging from 27,639 to 231,488 sequences per sample (Table S1). Of these, 229,738 sequences were retained following primer and MID trimming, and quality filtering (10.9% of raw reads); the number of final sequences per sample ranged from 1,217 (12093-MB) to 24,220 (12092-55). All samples were rarefied to 1217 sequences to exclude differences due to sequencing depth prior to comparing OTU count and composition between samples. Rarefaction at an even sequencing depth excluded differences in OTU count as a result of variations in the number of reads per sample (Fig. S1) and enabled a standardised approach for data analysis across all samples, a feature essential for the validation of a novel forensic technique.

The number of OTUs detected by each protocol was compared at a sequencing depth of 1217 to determine which extraction method detected the highest fungal diversity (Fig. 1B). Interestingly, the number of fungal OTUs was not correlated with total DNA yield (Fig. S2), and a core set of OTUs was detected for a given sample regardless of extraction method, even though this OTU overlap represented only 18.7 ± 3.1 % (mean ± SD) of the total OTUs detected from each sample when all four methods were combined (Fig. 2). Instead, the method yielding the greatest number of OTUs depended on soil clay content. For the soil with low clay content (8%, 12097), re-extraction detected a higher number of OTUs and more unique OTUs than either the RT and 55 (Fig.1B and Fig. S3) methods, whereas other soils (>30% clay) yielded more OTUs when the RT method was applied. These results suggest that soil clay content contributes to the ability to obtain fungal diversity from a given soil type, and demonstrates that a high proportion of OTUs remain undetected when the standard MB extraction method is applied.

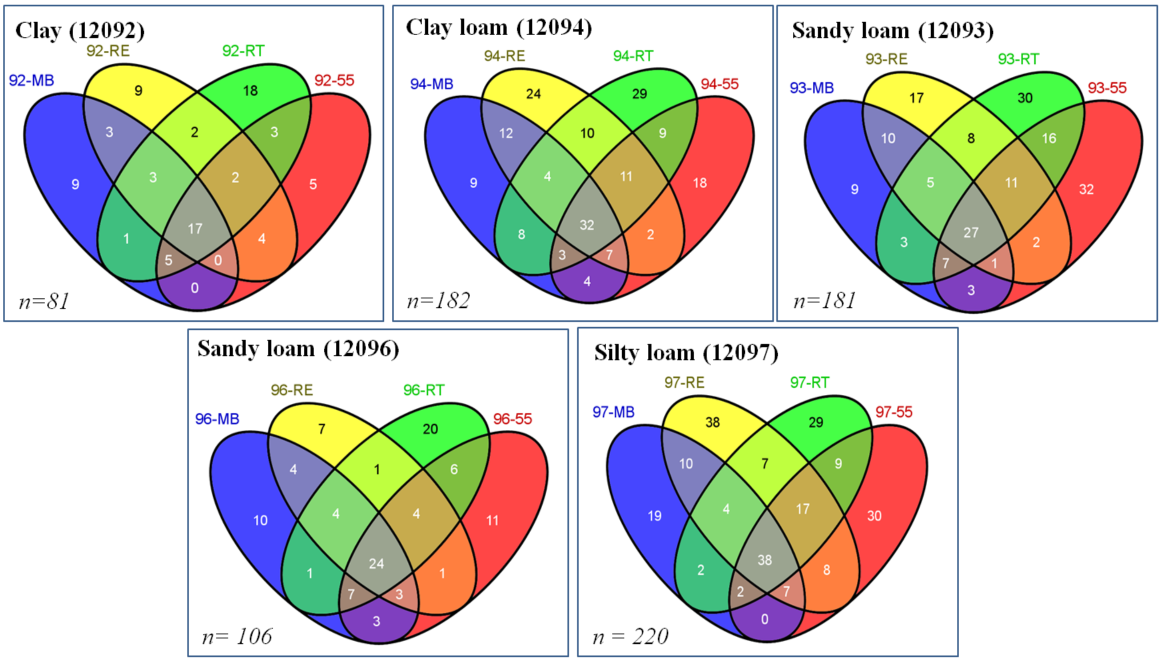


Fig. 2: Fungal OTU overlap between DNA extracts generated using four DNA extraction protocols using rarefied OTU table. For five soil samples, the MOBIO PowerSoil DNA Isolation kit (MB) and three modified protocols were used: extraction of the residual soil pellet from MB (RE); 24 hr lysis step at room temperature (RT); 24hr lysis step at 55°C (55).

**n = number of OTUs detected using the rarefied OTU table.*

OTU composition varied between extraction methods

We investigated whether specific fungal taxa were preferentially extracted with a particular method, and if this varied with soil type, using the non-rarefied data so that less abundant taxa were not excluded during the sub-sampling involved in rarefaction. Although no specific trend was observed across all soils, different DNA extraction methods revealed distinct fungal communities depending on soil type (Fig. 3 and table S3). When specific fungal taxa were investigated, trends were observed in relationship with the clay content of soil. For example, relative abundance of Basidiomycota and Zygomycota decreased using the modified methods in the low clay content soil (12097), whereas the relative proportion of Zygomycota and Ascomycota increased using the modified protocols for soils with higher clay content (12094 and 12096). In addition, several fungal phyla were also identified exclusively in the modified extraction protocols. For example, Blastocladiomycota, Chytridiomycota and Glomeromycota were detected with all three modified protocols in 12097 despite being undetected using MB (Table S2). More specifically, 24 hour lysis incubation identified Diversisporales and Rhizophylicidales (12093 and 12096), and re-extraction detected Lecanoromycetes (12093) and Scolecobasidium (12094) exclusively. This result suggests that the fungal diversity from a sample can be markedly enhanced by modifying the standard DNA extraction protocol.

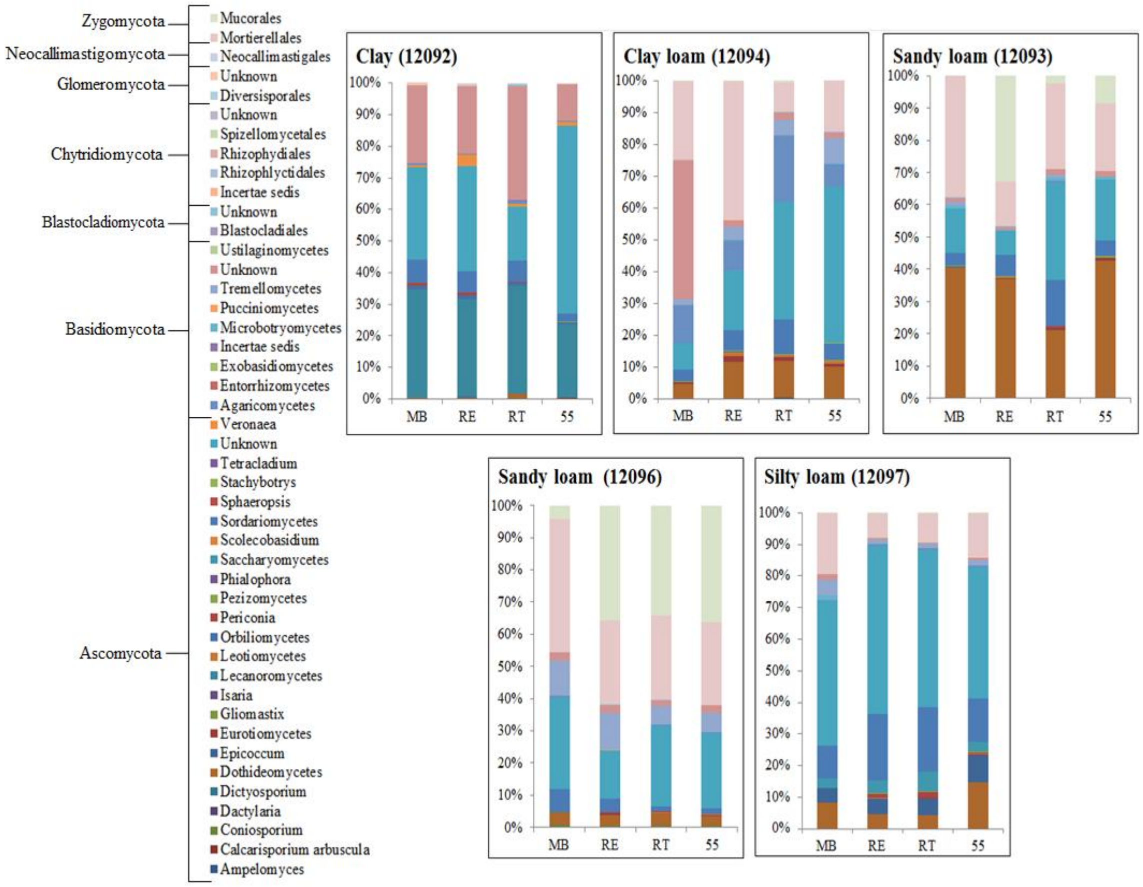


Fig. 3: Relative proportions of fungal OTUs within each phylum using four different DNA extraction methods on five soil types. See legend of Fig. 1 for abbreviations.

Discriminatory power

Forensic soil science is based on the ability of soil DNA methods to discriminate between samples or sites, allowing investigators to identify samples from the same location. Such result should be supported by a match probability i.e. how many other locations will have a similar genetic profile by chance. This is achieved by comparing the soil DNA profile from a sample of unknown origin to those collected from multiple reference sites. In this study, variation in OTU composition due to DNA extraction method did not prevent sample differentiation and discrimination (Fig. 4), indicating that extraction bias altered OTU composition less than the naturally occurring differences between samples. No significant difference in the mean Bray-Curtis distance was observed between the four different methods (*one-way ANOVA*, $F_{3,\,36}=0.190$, $p=0.856$), indicating that no single DNA extraction method consistently provided better discrimination between the different soils.

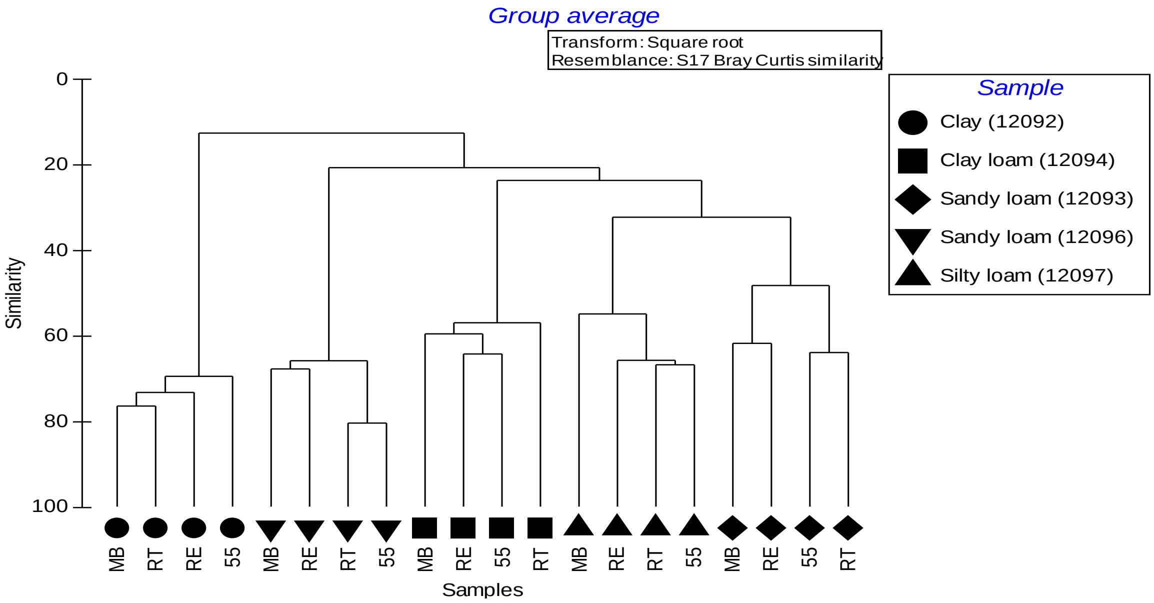


Fig. 4: Bray-Curtis cluster dendrogram of five soil samples based on fungal OTU composition detected using four DNA extraction methods: commercial MOBIO PowerSoil DNA Isolation kit (MB), extraction of the residual soil pellet from MB (RE), 24 hr lysis step at room temperature (RT) and 24hr lysis step at 55°C (55).

Discussion

Our results show that the optimal DNA extraction method for analysis of forensic soil samples varies depending on soil type. All modified DNA extraction methods increased the DNA yield and diversity compared to the standard DNA extraction kit protocol. However, total DNA yield was not correlated with fungal diversity or the ability to discriminate between samples. For all soil types, 24 hour lysis incubation step increased the DNA yield; however, the optimal incubation temperature was related to soil pH. For soils with a high pH (>7.5), incubation at 55°C increased DNA yield, whereas room temperature incubation increased DNA yield from soils with a low pH (<7.5). In contrast, optimal fungal diversity was related to the clay content of the soil. For soils with a low clay content (<30%), re-extraction of the soil pellet produced the highest number of OTUs, whereas 24 hour lysis incubation at room temperature detected the highest number of OTUs from soils with a high clay content (>30%). Interestingly, each method detected unique OTUs not detected with other methods for each soil (Fig. 2). Therefore, we suggest that multiple extraction methods can provide a more complete inventory of the fungal diversity and thus increased the robustness of the soil comparison. However, in cases with limited soil available, we recommend 24 hour lysis at room temperature (or re-extraction for low clay content soils), followed by re-extraction of the soil pellet to maximise the fungal diversity from a single sample.

For all soil types DNA yield increased with 24 hour lysis incubation; however, the optimal temperature varied with soil pH (Table 2). Double layer repulsion dictates the strength of extracellular DNA interactions with soil particles and thus how readily DNA is released during extraction (24). Studies have shown that DNA, which carries a negative charge, is more readily adsorbed onto soils with low pH, i.e. more positively charged, than high pH, i.e. more negatively charged (25-27). In addition, Cai *et al.* (25) also demonstrated that an increase in temperature promoted DNA absorption to low pH soils. Our finding that incubation at 55°C decreased the DNA yield from low pH samples compared to room temperature lysis suggests that increasing the lysis temperature in low pH soils may strengthen the interaction between DNA and the soil particles.

Table 2: Summary of the optimal DNA extraction methods for DNA yield and OTU count for each soil type. Extraction modifications: extraction of the residual soil pellet from MB (RE); 24 hr lysis step at room temperature (RT); and 24 hr lysis step at 55°C (55).

Soil texture classification	Clay content (%)	pH (1:5)	Highest DNA concentration (ng/mg soil)	Highest number of OTUs detected
Clay	**60**	**8.3**	**55***	**RT**
Clay loam	**45**	**7.3**	**RT***	**RT**
Sandy loam	**40**	**6.4**	**RT**	**RT**
Sandy loam	**33**	**7.6**	**55**	**RT**
Silty loam	**8**	**5.8**	**RT**	**RE**

**Greater than 20% difference in DNA yield between RT and 55.*

For forensic soil discrimination, the DNA extraction method should aim to detect the highest fungal diversity from a limited quantity of soil. In this study, the DNA extraction method that generated the highest fungal diversity varied with soil clay content. As clay soild consist of smaller particles, a high clay content increases the surface area of a soil, thus creating more DNA-binding sites via cation bridging (24, 28-30). As a result, during the extraction process, extracellular DNA molecules can be more readily adsorbed onto the soil surface, thus decreasing DNA yield and diversity. It is possible that soils with higher clay content may benefit from an extended lysis period in a chelating reagent, such as ethylenediaminetetraacetic acid (EDTA), which can strip the cations from the clay surface and reduce extracellular DNA binding. As a result, less DNA might remain bound to the soil during the lysis step, and increase DNA yield and diversity detected from a sample. In contrast, DNA is more readily removed from low clay content soils which have less potential binding sites, so increasing the lysis period has less impact on the soil fungal communities obtained. Instead, re-extraction of the low clay content soil yielded the highest fungal diversity. The results from this study provide a basic guide to efficiently increase the fungal diversity detected from soils with different clay contents (Table 2).

The quantity of soil available is often limited in forensic casework, which highlights the importance of maximising DNA yield and fungal diversity from minimal material. However, detection of fungal diversity is further complicated by the fact that individual fungal taxa have varying degrees of resistance to lysis, thus influencing the release of intracellular DNA into solution. For example, many soil fungi have melanised cell walls which provide resistance to lysis (31); or can form resting structures, such as sclerotia, that allow fungi to survive in extreme conditions (32-34). We demonstrate that many taxa, including entire phyla, remain undetected when only one DNA extraction method is applied. Therefore, we suggest two approaches for improving the fungal diversity detected from forensic soil samples. First, multiple extraction methods should be applied to generate a more complete inventory of the diversity present within each sample, providing a more robust comparison than any single method. Second, completing a re-extraction of the residual soil pellet is likely to provide a more detailed picture of the soil diversity. In this study, 24 hour incubation at room temperature (RT) outperformed the commercial kit (MB) for all soil types, so re-extraction of the soil pellet, following an initial extraction with a longer RT lysis step, could further increase DNA yield and OTU diversity from a single sample of limited quantity. In addition, our study demonstrates that re-extraction of the soil pellet can provide a profile comparable to that recovered from the initial material, despite subtle differences in diversity (Fig. 4). As a result, if the residual pellet is retained and stored at -20^{o}C then a re-extraction could be utilised, if a later re-examination of a sample was required.

This study confirms that much of the DNA and fungal taxa present in single soil samples is not extracted using a single application of a standard DNA extraction kit, and that the optimal DNA extraction protocol varies depending on soil pH and clay content. Our results suggest that applying multiple extraction methods would enable more robust comparisons between soil samples during forensic case study; however, this would increase the processing time and cost. In particular, extraction of the residual soil pellet following an initial extraction with 24 hour lysis incubation may offer a novel tool for forensic soil analysis. Re-extraction is a simple and efficient method to increase the detection of fungal diversity from a limited quantity of soil.

Acknowledgements

This work was supported by an ARC Linkage Grant between the University of Adelaide and the Australian Federal Police. We thank members of the Australian Centre for Ancient DNA for helpful comments on previous versions of the manuscript.

References

1. **Concheri, G.** 2011. Chemical elemental distribution and soil DNA fingerprints provide the critical evidence in murder case investigation. PloS ONE **6:**e20222.
2. **Fitzpatrick, R. W., M. D. Raven, and S. T. Forrester.** 2009. A systematic approach to soil forensics: criminal case studies involving transference from crime scene to forensic evidence. Criminal and Environmental Soil Forensics**:**105.
3. **Sensabaugh, G. F.** 2009. Microbial Community Profiling for the Characterisation of Soil Evidence: Forensic Considerations. Springer Netherlands, 49-60.
4. **Delmont, T. O., P. Robe, I. Clark, P. Simonet, and T. M. Vogel.** 2011. Metagenomic comparison of direct and indirect soil DNA extraction approaches. J. Microbiol. Methods **86:**397-400.
5. **Robe, P., R. Nalin, C. Capellano, T. A. Vogel, and P. Simonet.** 2003. Extraction of DNA from soil. Eur. J. Soil Biol. **39:**183-190.
6. **Courtois, S., A. Frostegard, P. Goransson, G. Depret, P. Jeannin, and P. Simonet.** 2001. Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. Environ. Microbiol. **3:**431-439.
7. **Frostegard, A., S. Courtois, V. Ramisse, S. Clerc, D. Bernillon, F. Le Gall, P. Jeannin, X. Nesme, and P. Simonet.** 1999. Quantification of bias related to the extraction of DNA directly from soils. Appl. Environ. Microbiol. **65:**5409-5420.
8. **Knauth, S., H. Schmidt, and R. Tippkötter.** 2012. Comparison of commercial kits for the extraction of DNA from paddy soils. Lett. Appl. Microbiol. **56:**222-228.
9. **Holmsgaard, P. N., A. Norman, S. C. Hede, P. H. B. Poulsen, W. A. Al-Soud, L. H. Hansen, and S. J. Sørensen.** 2011. Bias in bacterial diversity as a result of Nycodenz extraction from bulk soil. Soil Biol. Biochem. **43:**2152-2159.
10. **Inceoglu, O., E. F. Hoogwout, P. Hill, and J. D. van Elsas.** 2010. Effect of DNA extraction method on the apparent Microbial Diversity of Soil. Appl. Environ. Microbiol. **76:**3378-3382.
11. **Martin-Laurent, F., L. Philippot, S. Hallet, R. Chaussod, J. C. Germon, G. Soulas, and G. Catroux.** 2001. DNA extraction from soils: Old bias for new microbial diversity analysis methods. Appl. Environ. Microbiol. **67:**2354-2359.

12. **Feinstein, L. M., W. J. Sul, and C. B. Blackwood.** 2009. Assessment of bias associated with incomplete extraction of microbial DNA from soil. Appl. Environ. Microbiol. **75:**5428-5433.

13. **Jiang, Y. X., J. G. Wu, K. Q. Yu, C. X. Ai, F. Zou, and H. W. Zhou.** 2011. Integrated lysis procedures reduce extraction biases of microbial DNA from mangrove sediment. J. Biosci. Bioeng. **111:**153-157.

14. **Jones, M. D., D. R. Singleton, W. Sun, and M. D. Aitken.** 2011. Multiple DNA Extractions Coupled with Stable-Isotope Probing of Anthracene-Degrading Bacteria in Contaminated Soil. Appl. Environ. Microbiol. **77:**2984-2991.

15. **Macdonald, L. M., B. K. Singh, N. Thomas, M. J. Brewer, C. D. Campbell, and L. A. Dawson.** 2008. Microbial DNA profiling by multiplex terminal restriction fragment length polymorphism for forensic comparison of soil and the influence of sample condition. Journal of Applied Microbiology **105:**813-821.

16. **Epp, L. S., S. Boessenkool, E. P. Bellemain, J. Haile, A. Esposito, T. Riaz, C. Erseus, V. I. Gusarov, M. E. Edwards, A. Johnsen, H. K. Stenoien, K. Hassel, H. Kauserud, N. G. Yoccoz, K. Brathen, E. Willerslev, P. Taberlet, E. Coissac, and C. Brochmann.** 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Mol. Ecol. **21:**1821-1833.

17. **Meyer, M., and M. Kircher.** 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb. Protoc. **2010:**pdb.prot5448.

18. **Berry, D., K. B. Mahfoudh, M. Wagner, and A. Loy.** 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl. Environ. Microbiol. **77:**7846-7849.

19. **Martin, M.** 2012. Cutadapt removes adapter sequences from high-throughput sequencing reads. Bioinformatics in Action **17:**10-12.

20. **Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight.** 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Meth. **7:**335-336.

21. **Edgar, R. C.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:**2460-2461.

22. **Bragg, L. M., G. Stone, M. K. Butler, P. Hugenholtz, and G. W. Tyson.** 2013. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. PloS Comput. Biol. **9:**e1003031.

23. **Kõljalg, U., R. H. Nilsson, K. Abarenkov, L. Tedersoo, A. F. S. Taylor, M. Bahram, S. T. Bates, T. D. Bruns, J. Bengtsson-Palme, T. M. Callaghan, B. Douglas, T. Drenkhan, U. Eberhardt, M. Dueñas, T. Grebenc, G. W. Griffith, M. Hartmann, P. M. Kirk, P. Kohout, E. Larsson, B. D. Lindahl, R. Lücking, M. P. Martín, P. B. Matheny, N. H. Nguyen, T. Niskanen, J. Oja, K. G. Peay, U. Peintner, M. Peterson, K. Põldmaa, L. Saag, I. Saar, A. Schüßler, J. A. Scott, C. Senés, M. E. Smith, A. Suija, D. L. Taylor, M. T. Telleria, M. Weiss, and K.-H. Larsson.** 2013. Towards a unified paradigm for sequence-based identification of fungi. Mol. Ecol. **22:**5271-5277.

24. **Young, J. M., N. J. Rawlence, L. S. Weyrich, and A. Cooper.** 2014. Limitations and recommendations for successful DNA extraction from forensic soil samples: A review. Science & Justice **54:**238-244.

25. **Cai, P., Q. Huang, D. Jiang, X. Rong, and W. Liang.** 2006. Microcalorimetric studies on the adsorption of DNA by soil colloidal particles. Colloids and Surfaces B: Biointerfaces **49:**49-54.

26. **Saeki, K., M. Sakai, and S. I. Wada.** 2010. DNA adsorption on synthetic and natural allophanes. Applied Clay Science **50:**493-497.

27. **Shen, Y., H. Kim, M. P. Tong, and Q. Y. Li.** 2011. Influence of solution chemistry on the deposition and detachment kinetics of RNA on silica surfaces. Colloid Surf. B-Biointerfaces **82:**443-449.

28. **Nguyen, T. H., and K. L. Chen.** 2007. Role of divalent cations in plasmid DNA adsorption to natural organic matter-coated silica surface. Environ. Sci. Technol. **41:**5370-5375.

29. **Nguyen, T. H., and M. Elimelech.** 2007. Plasmid DNA adsorption on silica: Kinetics and conformational changes in monovalent and divalent salts. Biomacromolecules **8:**24-32.

30. **Levy-Booth, D. J., R. G. Campbell, R. H. Gulden, M. M. Hart, J. R. Powell, J. N. Klironomos, K. P. Pauls, C. J. Swanton, J. T. Trevors, and K. E. Dunfield.** 2007. Cycling of extracellular DNA in the soil environment. Soil Biology & Biochemistry **39:**2977-2991.

31. **Eisenman, H. C., and A. Casadevall.** 2012. Synthesis and assembly of fungal melanin. Appl. Microbiol. Biotechnol. **93:**931-40.

32. **Cooke, R.** 1986. The Fifth Kingdom, p. 689-690, Transactions of the British Mycological Society, vol. 86.

33. **Gwynne-Vaughan, H. C. I., and B. Barnes.** 1930. The Structure and Development of the Fungi. CUP Archive.

34. **Kendrick, B.** 2001. Fungi and the History of Mycology. John Wiley & Sons Ltd, Chichester. http://www.els.net.

Supplementary Material

Table S1: Overview of the sequence counts following trimming and data filtering steps.

Sample	Extraction method	Number of sequences per sample	Primer, barcode and length trimming (% of raw reads)	Quality filtering (% of raw reads)
Clay (92)	MB	40359	29902 (74.1)	6196 (20.7)
	RE	91039	66071 (72.6)	19906 (30.1)
	RT	79043	60519 (76.6)	13390 (22.1)
	55	121767	91157 (74.9)	24220 (26.6)
Sandy loam (93)	MB	28881	21273 (73.7)	1217 (5.7)
	RE	86285	64015 (74.2)	12866 (20.1)
	RT	98883	66945 (67.7)	14565 (21.8)
	55	89032	59241 (66.5)	11330 (19.1)
Clay loam (94)	MB	146604	47755 (32.6)	2623 (5.5)
	RE	173467	67113 (38.7)	3100 (4.6)
	RT	231488	94538 (40.8)	2258 (2.4)
	55	185542	81601 (44.0)	4658 (5.7)
Sandy loam (96)	MB	27693	21554 (77.8)	3038 (14.1)
	RE	89150	71537 (80.2)	18909 (26.4)
	RT	91516	67996 (74.3)	13985 (20.6)
	55	89953	69033 (76.7)	12084 (17.5)
Silty loam (97)	MB	43023	32288 (75.0)	4798 (14.9)
	RE	108490	78191 (72.1)	21640 (27.7)
	RT	85665	55881 (65.2)	15337 (27.4)
	55	191537	132072 (69.0)	23618 (17.9)

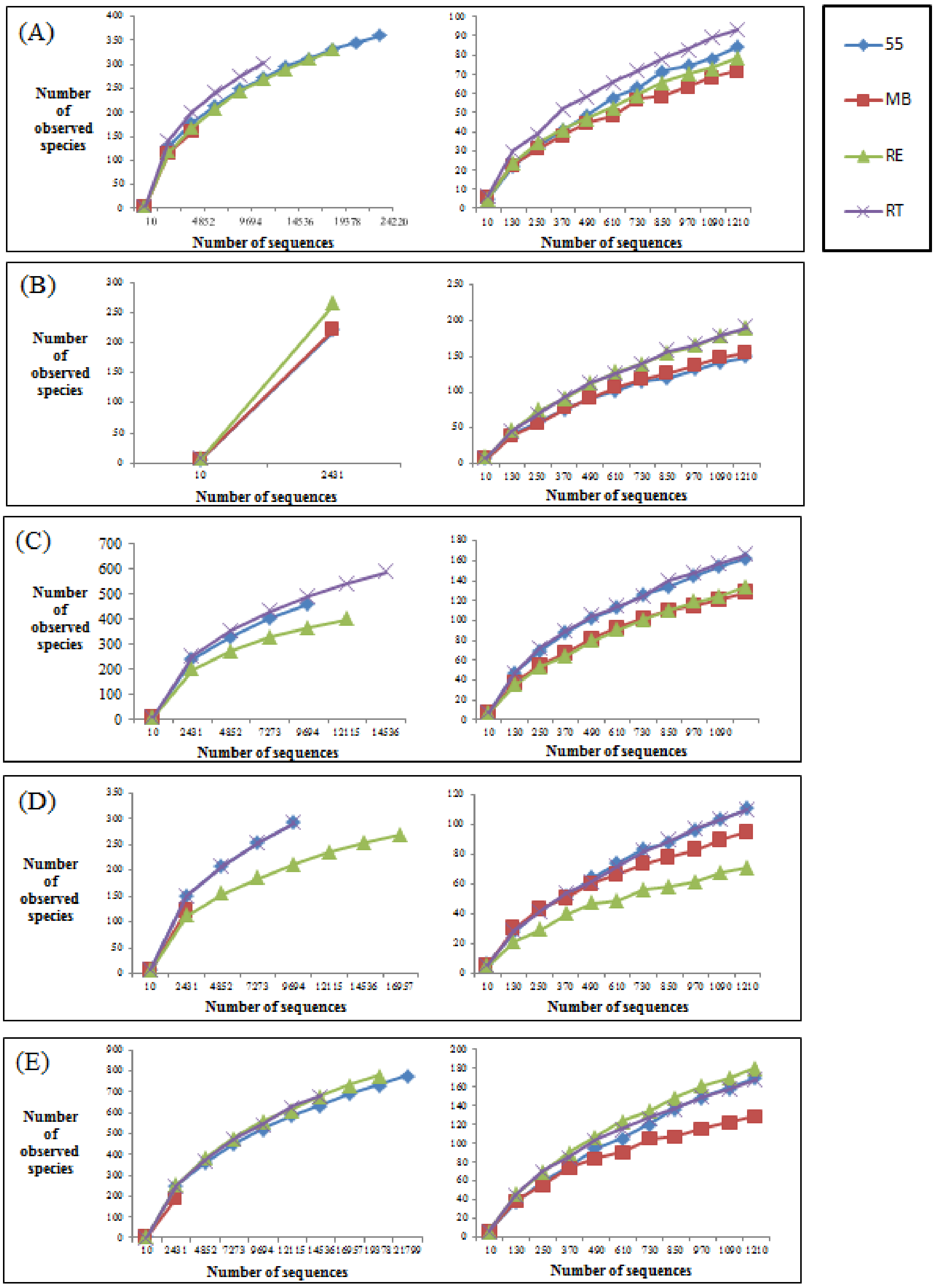


Fig. S1: Accumulation curves of observed species generated using non-rarefied data and rarefied data for samples (A) Clay soil (12092), (B) Clay loam soil (12094), (C) Sandy loam soil (12093), (D) Sandy loam soil (12094), and (E) Silty loam soil (12097).

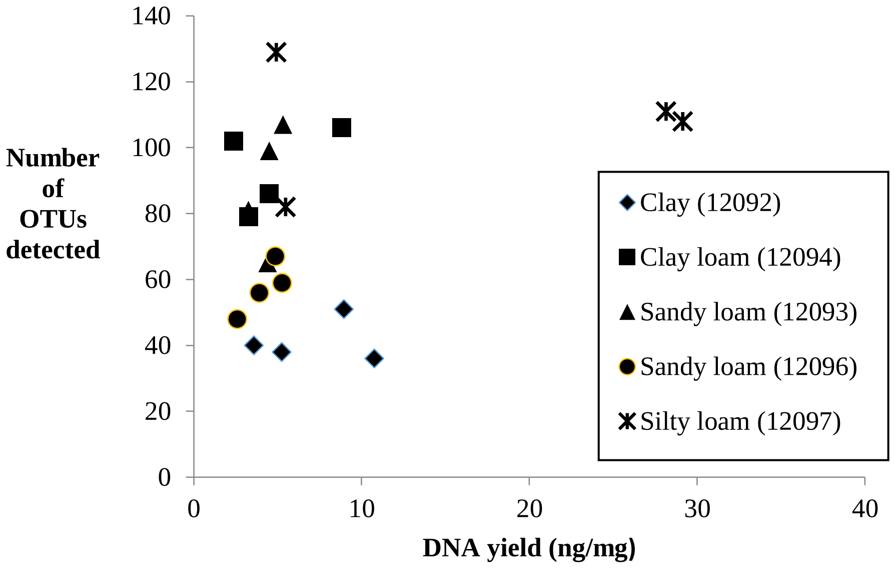


Fig. S2: Relationship between DNA yield (ng/mg soil) and the number of OTUs based on the rarefied OTU table.

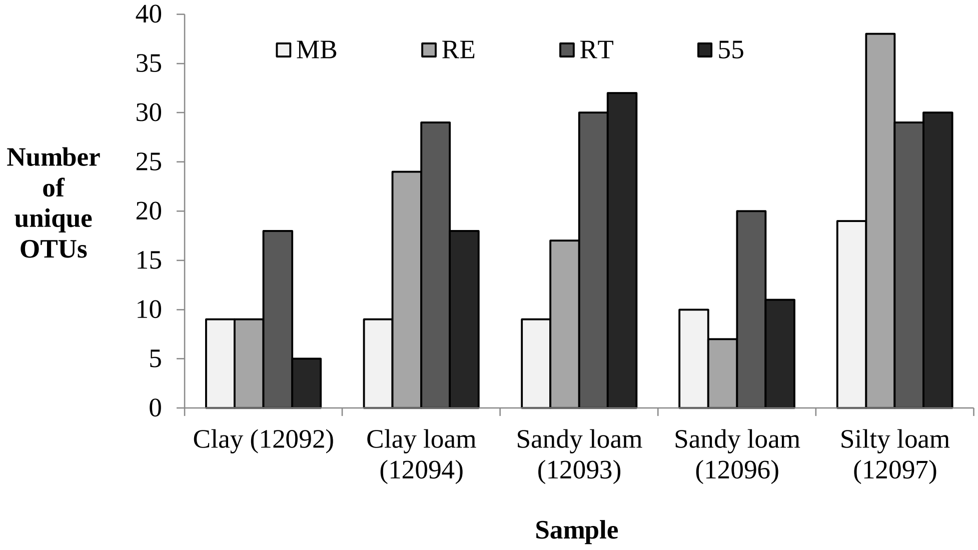


Fig. S3: The number of unique fungal OTUS detected from a standard DNA extraction protocol (MB) and three modified protocols: extraction of the residual soil pellet from MB (RE); 24 hr lysis step at room temperature (RT); and 24hr lysis step at 55°C (55).

Table S2: Relative percentage of each fungal Phyla detected using four DNA extraction protocols on five different soil types. The four methods included; commercial MOBIO PowerSoil DNA Isolation kit (MB) and three modified protocols were used: RE, extraction of the residual soil pellet from MB; RT, 24 hr lysis step at room temperature; 55, 24hr lysis step at 55^{o}C. Values highlighted in RED represent phyla for each sample that were undetected using the commercial kit but detected using a modified method.

Sample	Method	**Ascomycota**	**Basidiomycota**	**Blastocladiomycota**	**Chytridiomycota**	**Glomeromycota**	**Neocallimastigomycota**	**Zygomycota**
Clay (12092)	MB	74.134	25.196	0.000	0.000	0.559	0.000	0.112
	RE	77.208	21.945	0.000	0.000	0.436	0.000	0.411
	RT	61.838	37.266	0.421	0.000	0.000	0.000	0.474
	55	87.809	11.666	0.011	0.011	0.043	0.000	0.461
Clay loam (12094)	MB	17.788	57.197	0.000	0.189	0.063	0.063	24.701
	RE	40.522	15.606	0.000	0.000	0.201	0.067	43.603
	RT	61.947	28.192	0.000	0.126	0.126	0.126	9.482
	55	66.799	17.026	0.000	0.114	0.170	0.227	15.664
Sandy loam (12093)	MB	58.861	3.165	0.316	0.000	0.000	0.000	37.658
	RE	51.829	1.430	0.037	0.037	0.000	0.000	46.667
	RT	66.852	4.191	0.014	0.027	0.041	0.000	28.876
	55	67.695	2.693	0.016	0.033	0.049	0.000	29.514
Sandy loam (12096)	MB	40.252	14.016	0.180	0.000	0.090	0.000	45.463
	RE	23.657	14.194	0.281	0.000	0.200	0.000	61.668
	RT	31.514	8.043	0.014	0.043	0.129	0.000	60.257
	55	29.188	8.649	0.016	0.016	0.144	0.000	61.987
Silty loam (12097)	MB	72.200	8.448	0.000	0.000	0.000	0.000	19.352
	RE	89.610	2.439	0.036	0.181	0.325	0.000	7.409
	RT	88.359	1.801	0.380	0.127	0.279	0.000	9.054
	55	82.753	3.036	0.081	0.061	0.445	0.000	13.623

Table S3: Relative proportions of fungal OTUs using four different DNA extraction methods on five soil types: The four methods included; commercial MOBIO PowerSoil DNA Isolation kit (MB) and three modified protocols were used: RE, extraction of the residual soil pellet from MB; RT, 24 hr lysis step at room temperature; 55, 24hr lysis step at 55^{o}C. Values highlighted in RED represent phyla for each sample that were undetected using the commercial kit but detected using a modified method.

	Sample		**Clay**	**12092**			**Clay loam**	**12094**			**Sandy loam**	**12093**			**Sandy loam**	**12096**			**Silty laom**	**12097**	
Phyla	**Taxon**	**MB**	**RE**	**RT**	**55**	**MB**	**RE**	**RT**	**55**	**MB**	**RE**	**RT**	**55**	**MB**	**RE**	**RT**	**55**	**MB**	**RE**	**RT**	**55**
Ascomycota	Ampelomyces	0.000	0.000	0.000	0.000	0.000	0.019	0.014	0.016	0.000	0.000	0.253	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Calcarisporium arbuscula	0.000	0.000	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Coniosporium	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.898	0.541	0.600	0.562	0.000	0.000	0.000	0.000
	Dactylaria	0.000	0.000	0.000	0.000	0.000	0.000	0.054	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Dictyosporium	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.126	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Dothideomycetes	0.335	0.508	1.607	0.268	40.506	36.936	21.023	42.654	4.651	11.520	11.631	10.159	3.774	3.288	3.657	3.017	8.350	4.698	4.362	14.696
	Epicoccum	0.056	0.097	0.000	0.000	0.316	0.037	0.122	0.082	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	4.420	4.644	5.199	8.441
	Eurotiomycetes	0.000	0.048	0.000	0.054	0.000	0.223	0.420	0.718	0.503	1.875	1.138	0.681	0.000	0.441	0.029	0.064	0.000	0.217	0.152	0.324
	Gliomastix	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.067	0.000	0.114	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Isaria	0.112	0.024	0.000	0.064	0.000	0.000	0.000	0.033	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Lecanoromycetes	34.078	30.898	34.238	23.321	0.000	0.093	0.000	0.000	0.000	0.000	0.000	0.000	0.180	0.020	0.014	0.016	0.000	0.000	0.025	0.061
	Leotiomycetes	0.000	0.000	0.000	0.000	0.000	0.260	0.244	0.016	0.189	1.273	0.506	1.022	0.000	0.060	0.400	0.305	0.000	0.235	0.025	0.162
	Orbiliomycetes	1.341	1.113	1.001	0.536	0.000	0.037	0.000	0.082	0.000	0.067	0.126	0.000	0.000	0.040	0.043	0.096	0.000	0.054	0.076	0.061
	Periconia	0.726	1.065	0.132	0.139	0.000	0.037	0.353	0.065	0.000	0.000	0.000	0.000	0.000	0.301	0.157	0.128	0.098	1.211	1.725	0.324
	Pezizomycetes	0.000	0.097	0.079	0.129	0.316	0.093	0.190	0.343	0.189	0.469	0.379	0.227	0.000	0.000	0.000	0.016	0.000	0.108	0.203	0.364
	Phialophora	0.000	0.000	0.079	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.020	0.029	0.000	0.000	0.000	0.000	0.020
	Saccharyomycetes	0.000	0.000	0.026	0.000	0.000	0.019	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	3.045	4.138	6.366	3.077
	Scolecobasidium	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.067	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Sordariomycetes	7.542	6.702	6.558	2.539	3.797	6.834	14.255	4.913	3.646	6.162	10.746	5.165	7.008	4.150	1.457	1.605	10.413	21.088	20.238	13.684
	Sphaeropsis	0.000	0.000	0.026	0.011	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.018	0.025	0.000
	Stachybotrys	0.000	0.000	0.000	0.021	0.000	0.000	0.000	0.000	0.000	0.134	0.000	0.170	0.000	0.000	0.000	0.000	0.000	0.054	0.025	0.101
	Tetracladium	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.057	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Unknown	29.330	33.148	16.961	59.282	13.924	7.242	30.178	18.772	8.548	18.888	37.042	49.205	28.392	14.796	25.129	23.379	45.874	53.144	49.937	41.437
	Veronaea	0.615	3.508	1.106	1.446	0.000	0.000	0.000	0.000	0.063	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Basidiomycota	Agaricomycetes	0.503	0.411	1.001	0.193	0.000	0.093	0.651	0.082	11.565	9.377	20.860	7.037	0.809	0.341	0.300	0.497	0.098	0.398	0.431	0.486
	Entorrhizomycetes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.051	0.000
	Exobasidiomycetes	0.000	0.000	0.000	0.000	0.000	0.000	0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.000	0.000	0.000	0.000	0.000	0.000
	Incertae sedis	0.000	0.000	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Microbotryomycetes	0.000	0.000	0.026	0.000	0.949	0.223	0.760	0.751	0.000	0.067	0.126	0.284	0.000	0.040	0.014	0.000	1.473	0.289	0.025	0.061
	Pucciniomycetes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.054	0.051	0.000
	Tremellomycetes	0.056	0.000	0.000	0.021	0.949	0.223	1.004	0.229	1.823	4.086	4.678	7.832	10.512	11.548	5.700	5.745	5.108	0.976	0.710	1.903
	Unknown	24.637	21.534	36.213	11.452	1.266	0.891	1.763	1.616	43.683	2.009	2.402	1.759	2.695	2.225	2.029	2.407	1.768	0.488	0.431	0.506
	Ustilaginomycetes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.126	0.067	0.126	0.114	0.000	0.000	0.000	0.000	0.000	0.235	0.101	0.081
Blastocladiomycota	Blastocladiales	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.254	0.000
	Unknown	0.000	0.000	0.421	0.011	0.316	0.037	0.014	0.016	0.000	0.000	0.000	0.000	0.180	0.281	0.014	0.016	0.000	0.036	0.127	0.081
Chytridiomycota	Incertae sedis	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.033	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Rhizophlyctidales	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.029	0.016	0.000	0.000	0.000	0.000
	Rhizophydiales	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.189	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Spizellomycetales	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.126	0.076	0.020
	Unknown	0.000	0.000	0.000	0.011	0.000	0.037	0.027	0.000	0.000	0.000	0.126	0.114	0.000	0.000	0.014	0.000	0.000	0.054	0.051	0.040
Glomeromycota	Diversisporales	0.112	0.073	0.000	0.021	0.000	0.000	0.000	0.000	0.000	0.000	0.126	0.057	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Unknown	0.447	0.363	0.000	0.021	0.000	0.000	0.041	0.049	0.063	0.201	0.000	0.114	0.090	0.200	0.129	0.144	0.000	0.325	0.279	0.445
Neocallimastigomycota	Neocallimastigales	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.063	0.067	0.126	0.227	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Zygomycota	Mortierellales	0.112	0.411	0.474	0.461	37.658	13.723	26.475	20.992	24.576	43.537	9.102	15.664	41.060	25.962	25.943	25.690	19.253	7.355	9.029	13.583
	Mucorales	0.000	0.000	0.000	0.000	0.000	32.943	2.401	8.521	0.126	0.067	0.379	0.000	4.403	35.706	34.314	36.297	0.098	0.054	0.025	0.040
	Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

CHAPTER 5

High-throughput sequencing of trace quantities of soil provides discriminative and reproducible fungal DNA profiles

Young, J.M., L.S. Weyrich, L. J. Clarke, A. Cooper. High-throughput sequencing of trace quantities of soil provides reproducible and discriminative fungal DNA profiles.

(In review at Journal of Forensic Sciences)

Statement of authorship

High-throughput sequencing of trace quantities of soil provides discriminative and reproducible DNA profiles

In preparation for *Australian Journal of Forensic Sciences*

Jennifer M. Young (Candidate)

Designed experiment, DNA extractions, PCR amplifications, and sequencing, conducted downstream processing and analysis of data, interpreted the results; created tables, and wrote the paper.

I hereby certify that the statement of contribution is accurate

Signed... Date.....04/07/2014

Laura S. Weyrich

Provided advice on interpretation of results, content, structure, edited manuscript

I hereby certify that the statement of contribution is accurate

Signed .... Date.....04/07/2014

Laurence J. Clarke

Edited the manuscript

I hereby certify that the statement of contribution is accurate

Signed.... .... Date.....04/07/2014

Alan Cooper

Edited the manuscript

I hereby certify that the statement of contribution is accurate

Signed Date.....04/07/2014

Abstract

High-throughput sequencing (HTS) technology provides a means to generate detailed surveys of soil microbial communities, which have the potential to facilitate traditional approaches in forensic soil science. However, robust soil DNA analysis relies on an accurate and reproducible representation of the biodiversity in a sample which may be problematic if a limited quantity of material is available as is common in forensic cases. In this study, we applied HTS to varying masses of five soil types to assess the effect of soil mass on fungal DNA community profiles and the ability to differentiate between samples. Our results show that an increase in DNA yield with larger sample size was not always correlated with an increase in fungal diversity, and that the five different soils could be successfully differentiated regardless of soil mass. These results demonstrate that DNA profiles recovered from minimal soil quantities (50 mg) are comparable to those obtained using the recommended mass in a commercial DNA extraction kits (250 mg). We also found that reproducibility of DNA profiles from different sample sizes varied depending on soil texture For soils with a very fine texture (>% clay) or a very coarse texture (<% clay) , duplicate extracts were most similar using larger sample sizes, whereas soils with moderate texture were more similar with smaller sample sizes. Given that HTS of soil fungal communities was robust to the quantity of starting material used, we conclude that trace samples can provide valuable forensic evidence and can be sub-sampled for independent analysis, whilst maintaining a reliable soil DNA profile.

Keywords: soil, DNA, fungi, forensic, metagenomics, high-throughput sequencing (HTS), sample size, internal transcribed spacer (ITS).

Introduction

Soil is commonly used in forensic casework to link a suspect to a crime scene. Standard analyses examine intrinsic properties of soil including mineralogy, geophysics, texture and colour (1-4). However, DNA profiling of organisms within the soil matrix can provide a site-specific signal for use in forensic soil discrimination (5). Previous molecular methods (reviewed in 6) have relied on patterns of fragment length variation produced by amplification of unidentified microbial taxa in the soil extract; however, such methods provide little resolution of spatially and temporally variable microbial communities (7, 8). In contrast, high-throughput sequencing (HTS) technologies can drastically increase the power and resolution to distinguish soil samples by generating a detailed picture of soil microbial communities from thousands of DNA sequences in a single run (9-12). However, the quantity of soil available in forensic casework is often limited (1, 13) and varies depending on the type of contact, materials involved, soil properties and persistence (3). As a result, soil DNA analysis is not widely utilised in forensic science, due to concerns that trace samples may not provide an accurate representation of a particular location (13).

The structure of the soil matrix results in fine-scale heterogeneity in the distribution of taxa, particularly microorganisms within soil aggregates (14). This leads to debate about the optimal sample mass for community analysis. Previous studies have suggested that larger sample sizes (up to 5 kg) provide a more accurate representation of soil diversity (15-17) and other studies suggest small samples typically used in commercial DNA extraction kits (250 mg) may only detect a fraction of the overall diversity (18). By culturing single genus (Nitrobacter) from three different volumes of an agricultural soil, Grundmann and Gourbiere (14) suggested that small sample masses (<1 g) permit the detection of rare taxa by accessing micro-spatial niches. However, bacterial culturing did not provide a detailed picture of the overall bacterial diversity, nor did the study consider differences across soil types. Similarly, Ellingsoe and Johnsen (16) assessed the most representative sample size for studying soil bacterial structure by culturing and by visually comparing 16S rDNA denaturing gradient gel electrophoresis (DGGE) community fingerprints from a single forest soil type. In both analyses, small sample sizes (0.01 g and 0.1 g) showed more variation between replicates than larger sample sizes (10 g). Following this, Ranjard *et al.* (15) used Automated Ribosomal Intergenic Spacer Analysis (ARISA) of three soil types to suggest that the reproducibility of profiles varied with soil texture, with fine clay soil exhibiting the greatest variation between replicates. However, the range of sample sizes examined was 0.125 g to 4 g, and often the mass of soil available in forensic casework is <50 mg (19). In addition, the effect of sample size on the discriminatory power between different soils was not considered. As a result, the effect of sample size on the microbial diversity detected in a range of soil types requires further examination before HTS analysis of trace quantities of soils can be used in casework.

The effect of soil mass on fungal DNA profiles is of particular interest for forensic application because fungi reportedly show better discrimination between soil samples than bacteria (20). However, Ranjard *et al*. (15) recommend that sample masses <1 g were adequate for reproducible bacterial profiles, and that >1 g soil is required for reproducible fungal profiles. This is a potential limitation of fungal DNA profiling in forensic science, because the soil quantity available in casework is often less than 1 g. However, this recommendation was based solely on visual comparisons of DNA profiles generated using fingerprinting techniques (15, 16, 18), which provide a relatively coarse scale resolution and can underestimate diversity due to co-migration of DNA fragments from distinct species. In contrast, HTS can generate thousands of reads per sample and detect individual taxa, including those present at low abundance. The potential to detect low abundance fungal taxa using HTS warrants further examination of the effect of soil mass on fungal diversity using this advanced DNA sequencing technology.

This study provides a detailed comparison of fungal DNA profiles generated using HTS from trace quantities of soil (50, 150 and 250 mg) to determine whether sample size influences: (1) the number of taxa detected; (2) the reproducibility of fungal DNA profiles and (3) the power to discriminate between soil samples.

Materials and Methods

Sample collection and DNA extraction

To examine the effect of soil mass on a broad range of soil types, five Australian soils with different chemical properties were collected (0-20 cm depth) using a sterile plastic screw cap container, and all samples were stored at 4 ºC prior to extraction (Table 1). For each soil, pH was measured using 1:5 soil/water extracts and the particle size distribution was determined using a sedimentation approach. Based on percentage of sand, silt and clay present each sample was classified according to the soil textural triangle. This chemical analysis was performed by Alla Marchuk (PhD Candidate) from School of Agriculture, Food and Wine, University of Adelaide.

From each bulk sample, three sample sizes (50 mg, 150 mg and 250 mg) were sub-sampled and directly processed using the PowerSoil DNA Isolation kit (MOBIO, Carlsbad, CA, USA) following the manufacturer's protocol and DNA was eluted in 100 µL of elution buffer. Duplicate extracts were processed for each sample mass, and extraction blank controls were performed in parallel to samples using the same protocol, excluding the addition of any soil material. DNA yield of each extract was quantified using the NanoDrop 2000 Spectrophotometer (Thermo Scientific, Waltham, MA, USA) by taking an average of two measurements.

Table 1: Sampling location and chemical properties of the soils used in this study.

Sample	Texture	% Clay	pH	Location	Latitude, longitude
12092	Clay	60	8.3	Claremont (SA)	-34.58, 138.38
12094	Clay loam	45	7.3	McLaren Vale (SA)	-35.15, 138.33
12093	Sandy loam	40	6.4	Urrbrae (SA)	-34.58, 138.38
12096	Sandy loam	33	7.6	Tammin (WA)	-31.49,117.59
12097	Silty loam	8	5.8	Drouin (VIC)	-38.13,145.82

PCR amplification and library preparation

The internal transcribed spacer I (ITS1) was PCR-amplified using universal fungal primers ITS5 (5'-CCTCTCTATGGGCAGTCGGTGATGGAAGTAAAAGTCGTAACAAGG-3') and 5.8S_fungi (5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**nnnnnnn**CAAGAGATCCGTTGTTGAAAGTT-3') (21). The primers were modified to include Ion Torrent sequencing adapters, as underlined above (P1 adapter on the ITS5 primer; A adapter on the 5.8S_fungi primer). A unique seven base pair multiplex identifier (MID) tag (22) was incorporated into the 5.8S_fungi primer to separate sequences by sample and extraction method during data analysis. Each DNA extract was PCR-amplified in triplicate; each reaction was performed in a final volume of 25 µl comprising 2.5 mM $MgCl_2$, 0.24 mM dNTPs, 0.24 µm of each primer, 0.4 mg/ml BSA, 0.5 U Amplitaq Gold DNA polymerase in 1x reaction buffer (Applied Biosystems, Melbourne, Australia) and 1 µl DNA extract. Cycling conditions were as follows: 9 min at 94 ºC, followed by 35 cycles of 94 ºC for 30 s, 54 ºC for 30 s, and 72 ºC for 45 s, and a final extension at 72 ºC for 7 min. A no-template control was included for each MID tag. No PCR products were detected in the no template or extraction blank controls by agarose gel electrophoresis so these were omitted from further analysis. Triplicate PCR products were pooled and purified using Agencourt AMPure XP PCR Purification kit (Beckman Coulter Genomics, Lane Cove, NSW, Australia). Purified PCR products were quantified on an Agilent 2200 TapeStation using High Sensitivity D1K ScreenTape and reagents (Agilent Technologies, Santa Clara, CA, USA) and pooled to equimolar concentration. The amplicon library was quantified using the HS D1K Tapestation (Agilent Technologies) and diluted to 11.6 pM. Emulsion PCR and Ion Sphere Particle enrichment were performed on the Ion OneTouch system^{TM} using the Ion OneTouch^{TM} 200 Template Kit v2 DL (Life Technologies), before sequencing on the Ion Torrent Personal Genome Machine^{TM} using the Ion PGM^{TM} 200 Sequencing Kit and an Ion 316^{TM} semiconductor chip (Life Technologies).

Data Analysis

After sequencing, base calling was performed using Torrent Suite v3.4.2 (Life Technologies). Sequencing reads were de-multiplexed by MID tag using the fastx_barcode_splitter tool (FASTX-toolkit v0.0.12; http://hannonlab.cshi.edu/fastx_toolkit). Cutadapt v1.1 (23) was used to trim primer sequences and the fastx_clipper tool was used to remove reads <100 bp (parameters; –Q33 –l 100). We used a strict zero mismatch threshold for both the MID tag and primer sequences. Reads with a Phred score less than 20 for 90% of the sequence, were removed using fastq_quality_filter tool (FASTX-toolkit v0.0.12). The number of reads per sample following each of the data processing steps is summarised in Table S2. The resulting files were converted (.fastq to .fna, available from http://genomics.azcc.arizona.edu/help.php3) for analysis in QIIME (v.1.5.0.) (24). Reads were *de novo* clustered at 97% identity to create Operational Taxonomic Units (OTUs) using UCLUST (25), with the most abundant read in each cluster used as the representative sequence and only clusters with two or more reads retained. To compare differences between samples, an OTU table was created with all samples rarefied to 2642 sequences to account for differences in total read abundance between samples; 2642 was the minimum number of reads obtained for a single sample. Rarefaction enabled a standardised approach for data analysis across all samples, a feature essential for the validation of a novel forensic technique

To determine the relationship between DNA yield and soil mass, total DNA yield (ng) was plotted against soil mass (mg), and the efficiency of the DNA extraction at each soil mass was examined by comparing the DNA yield per mg of material. The relationship between OTU count and soil mass was examined by plotting the number of OTUs against soil mass (mg); the number of OTUs was then plotted against total DNA yield to determine the relationship between these two variables.

For each sample size, a Bray-Curtis (BC) distance resemblance matrix was generated from a rarefied OTU table in PRIMER6 (PRIMER-E, Plymouth Routines in Multivariate Ecological Research v. 6, *PRIMER-E Ltd*, Luton, UK) . BC distance provides a measure of dissimilarity between pairs of samples based on OTU composition and abundance: BC =100, samples are completely different; BC = 0 samples are identical. Sample reproducibility was measured using the inverse of Bray-Curtis distance (100-BC) as a measure of similarity between duplicate extracts at each sample mass. The reproducibility of duplicate extracts at each soil mass was visualised using multidimensional scaling (MDS) plots based on Bray-Curtis distance with default parameters.

The power to discriminate between soils was measured using the BC distance between extracts from different soils. A rarefied OTU table including all samples and extracts was imported into PRIMER6, and the data were square root transformed before generating a resemblance matrix using Bray-Curtis distance. ANOSIM (analysis of similarities) was used to determine significant differences in OTU composition due to sample size. To determine which sample size provided the highest discriminatory power, BC distance was calculated using extracts generated from the same soil mass. Significant differences in BC distance at each sample size were examined using one-way ANOVA in the SPSS Statistics 21 software package (IBM, USA). Furthermore, to stimulate a limited quantity of sample in forensic casework, the discriminatory power (BC distances) between extracts generated from different starting masses were also compared.

Results

DNA yield increased with soil mass.

DNA yield from five soil samples using three different starting masses was compared to determine the most efficient sample size for DNA recovery. All soil types showed a linear decrease in total DNA yield (ng) with a decrease in sample size (Fig. 1A: $R^2 =0.73$ *to* 0.96). On average, total DNA yield was reduced by 33 ± 7% (mean ± SD) when 150 mg was used instead of 250 mg, and decreased by 71.7 ± 4.3% (mean ± SD) when the mass was reduced to 50 mg soil (Fig. S1). All samples showed significantly less DNA yield using smaller sample sizes (Table S1, one-way ANOVA, $F_{(2,3)}$; 12092, $p =0.01$, 12094, $p =0.03$, 12093, $p <0.01$, 12092, $p =0.13$, 12097, $p =0.02$) with the sandy loam soil (12096) as the exception; a significant difference was observed between 50 mg and 250 mg only ($p= 0.05$). Interestingly, the effect of sample size on extraction efficiency (ng DNA /mg soil) varied between soil types (Fig. 1B) and was likely influenced by soil texture. For example, DNA yield (ng DNA /mg soil) recovered for the fine clay soil (12092) significantly decreased between 50 mg and 250 mg (*t-test;* $p= 0.026$), and for the coarse sandy loam soil (12093), the DNA extraction efficiency was statistically higher at 50 mg than 150 or 250 mg (Fig. S2; one-way ANOVA, $F_{(2.3)} = 24.6$, $p= 0.014$). In contrast, sample size had no significant effect on DNA extraction efficiency from soils with moderate texture (12094 and 12096) or the sandy loam soil (12096).

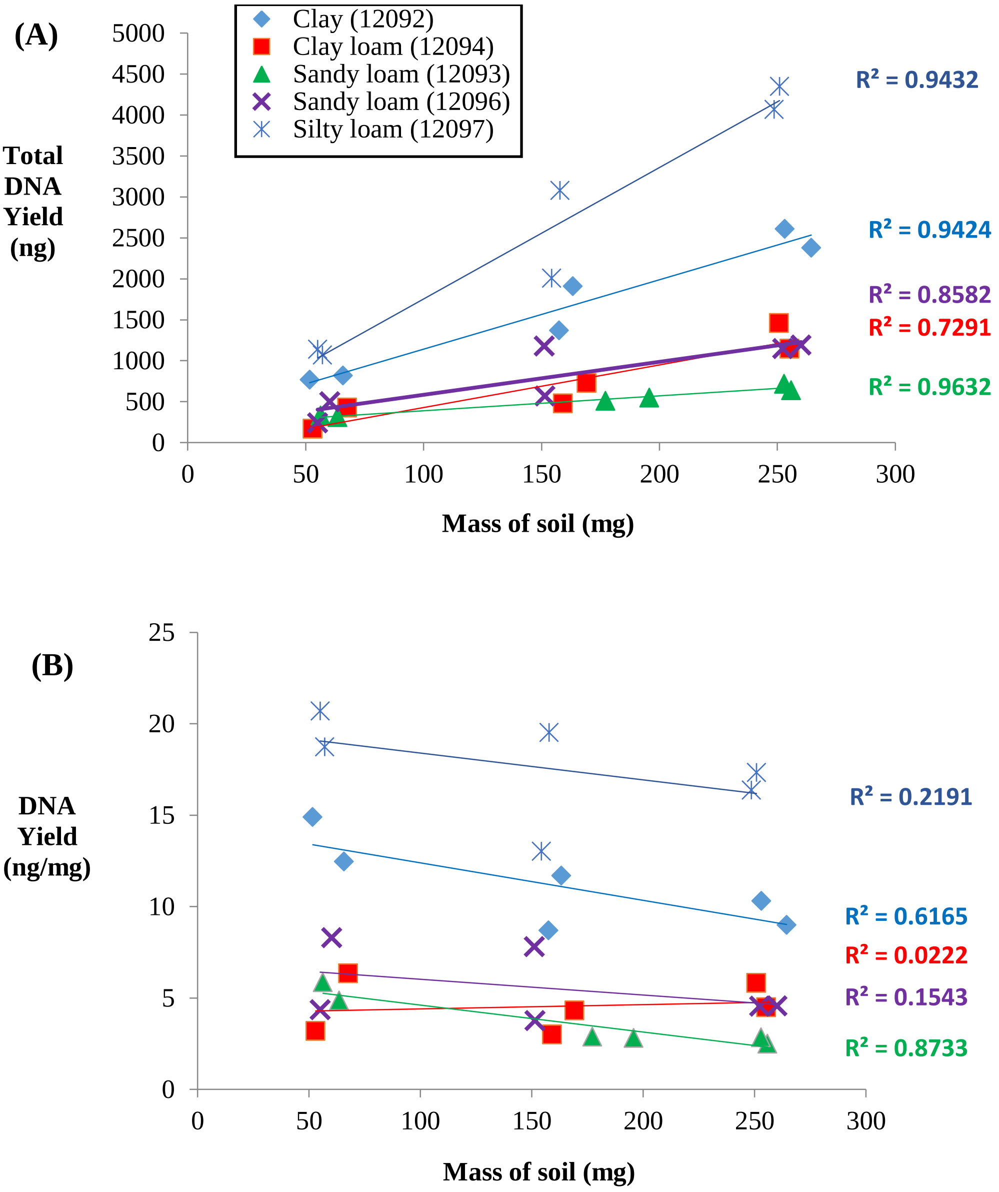


Fig. 1: Effect of soil mass on (A) Total DNA yield and (B) DNA extraction efficiency from five soil samples as a result of soil mass (50 mg, 150 mg and 250 mg) used in the DNA extraction.

OTU counts varied according to soil texture

HTS of the ITS amplicons yielded 8,355,404 sequences, ranging from 184809 to 369297 sequences per sample (Table S2). A total of 5,275,627 sequences (6% of raw reads) contained zero mismatches to both the primers and MID tags and were >100bp in length. Of these, 911,576 sequences (11 % of raw reads) met the quality filter threshold and were used in subsequent analysis; the number of final sequences per sample ranged from 2,642 (12093_250_ B) to 82,037 (12092_50_A). The relationship between total DNA yield and OTU count was examined using rarefied data to exclude differences in OTU count as a result of variations in the number of reads per sample (Fig. S3).

The effect of sample size on OTU count varied depending on soil texture (Fig. 2). A significant difference in OTU count was observed between 50 mg and 250 mg of soil for the fine clay soil (12092) and the coarse sandy soils (12093 and 12096)(Table S3, one-way ANOVA, $F_{(2,3)}$; 12092, *p =0.03*, 12093, *p =0.04*, 12096, *p <0.05*). The number of OTUs detected in the fine clay soil (12092) decreased by 8% and 31% when the sample size was reduced from 250 mg to 150 mg and 50 mg, respectively (Fig. S4). Similarly, the number of OTUs in the coarse sandy loam soils (12093 and 12096) decreased by 22% and 5%, respectively, when sample size was reduced from 250 mg to 150 mg soil, and OTU count further decreased by 31% (12093) and 22% (12096), when 50 mg soil was used. For soils with a moderate texture (12094 and 12097), no significant difference in the number of OTUs was observed between different sample sizes (Table S3, one-way ANOVA, $F_{(2,3)}$; 12094, *p =0.10.*, 12097, *p =0.94*). This could be related to sampling bias associated with heterogeneous distribution of taxa in both fine and coarse soil samples.

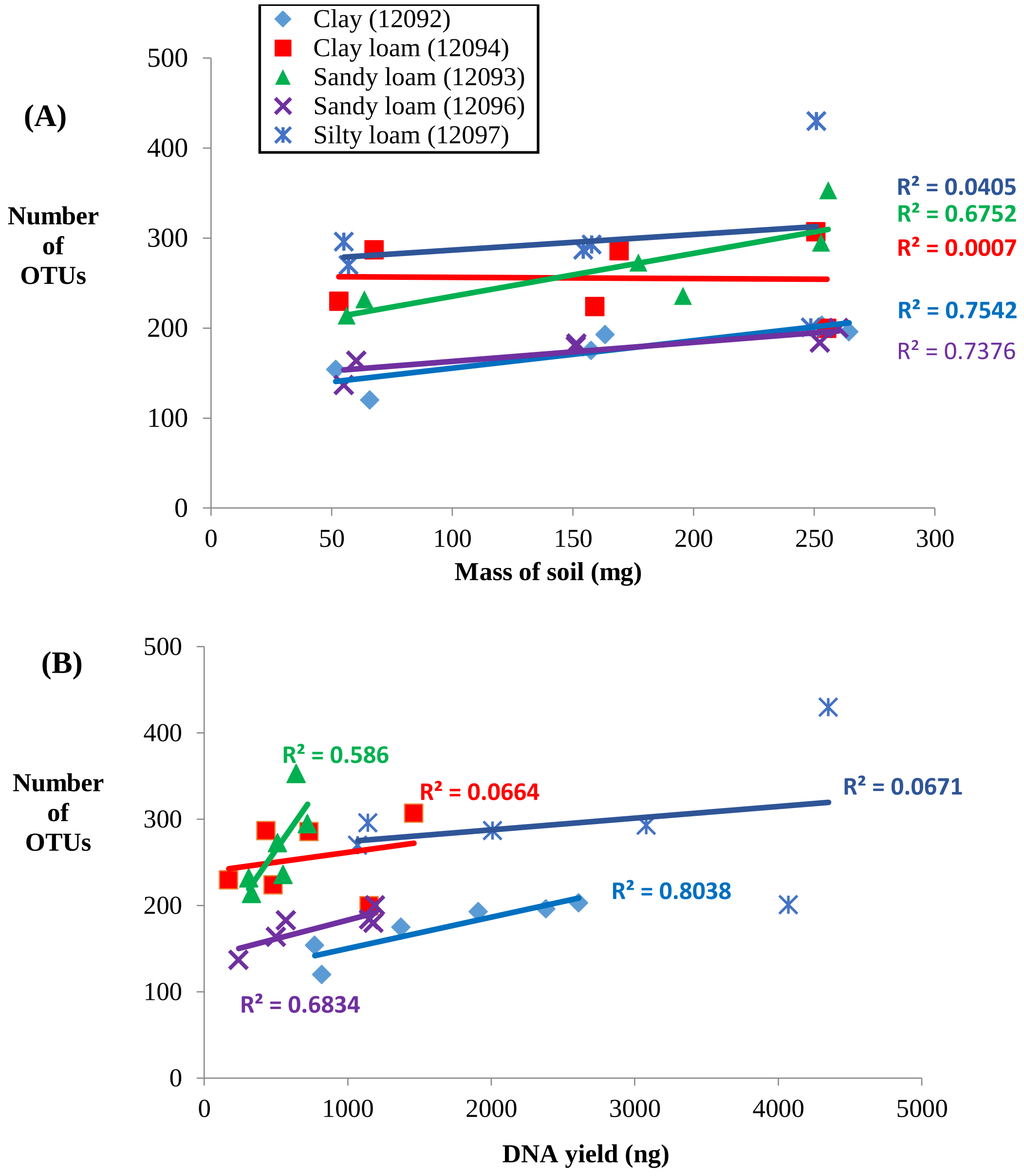
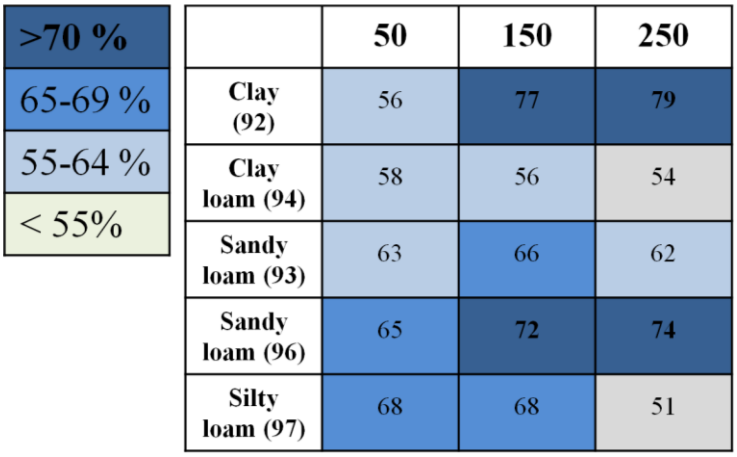


Fig. 2: (A) Effect of soil mass on number of OTUs detected and (B) Correlation between DNA yield and number of OTUs detected from five soil samples as a result of soil mass (50 mg, 150 mg and 250 mg) used in the DNA extraction.

Reproducibility within a sample varied with soil texture and pH

The OTU composition of duplicate DNA extracts was compared to determine which soil mass provided the most reproducible profile and examine if reproducibility was dependent on soil type (Table 2 and Fig. 3). DNA profile reproducibility was influenced by soil texture. First, the similarity between duplicate extracts of fine (12092) and coarse (12093 and 12096) soils increased with larger sample sizes, which could be related to an overall increase in the number of total OTUs obtained from these samples (Fig 2B). For soils with a moderate texture (12094 and 12097), optimal reproducibility was observed at 50 mg. Interestingly, soil pH also appeared to impact DNA profile reproducibility. Soils with pH > 7.5 (12092 and 12096) exhibited optimal reproducibility at 250 mg, whereas soils with pH < 7.5 (12093, 12094 and 12097) were most reproducible using 150 mg soil. This suggests that a low pH limits the reproducibility of DNA profile at larger sample sizes by promoting heterogeneity.

Table 2: Reproducibility of fungal OTU composition from duplicate extracts using three different starting masses in DNA extraction. Values represent the inverse of based on Bray-Curtis distance (100-BC).



>70 %
65-69 %
55-64 %
< 55%

	50	150	250
Clay (92)	56	77	79
Clay loam (94)	58	56	54
Sandy loam (93)	63	66	62
Sandy loam (96)	65	72	74
Silty loam (97)	68	68	51

To further examine the effect of sample size on reproducibility, we compared the fungal taxa detected in duplicate extracts based on the non-rarefied data (Table S4). The most consistent abundance of fungal phyla mirrored the reproducibility findings for each sample type. For example, duplicate extracts of the fine (12092) and coarse (12093 and 12096) soils, which were more reproducible when 250 mg of soil was used, contained similar levels of Basidiomycetes and unknown fungi in the 250 mg extracts whereas the levels of these were variable when 150 mg or 50 mg of soil was used. The lack of reproducibility at each sample size was also attributed to taxa disparities between duplicate extracts. For example, Chytridiomycota was detected in both extracts of the moderate textured soils using 50 mg, while this phylum was only identified in one of the duplicate extracts of 150 mg and 250 mg (12094). Similarly, the 150 and 250 mg samples were also the least reproducible for this sample type. These results provide an indication that soil texture and pH should dictate the most reproducible sample size for DNA extraction.

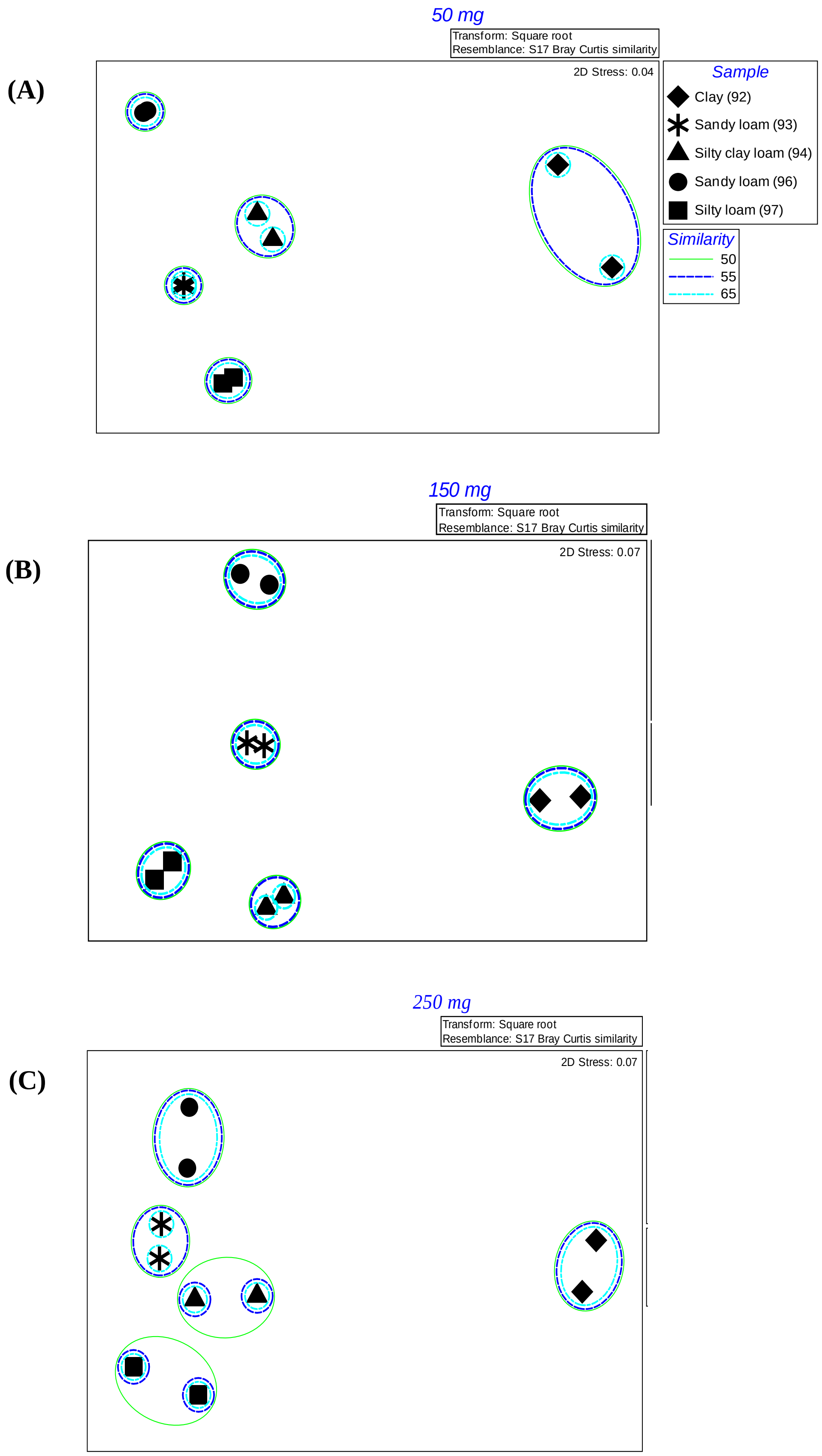


Fig. 3: MDS plot based on Bray-Curtis distance using different sample sizes in the DNA extraction: (A) 50 mg soil, (B) 150 mg, and (C) 250 mg.

Effect of sample mass on discriminatory power

As the quantity of soil available in casework can vary, we examined the effect of sample size on the ability to discriminate between different soils. When all DNA extracts were analysed together, sample size did not prevent differentiation of the five soils (Fig. 4, *two-way ANOSIM, ($p<0.01$): $R=1$ between samples, $R=0.46$ between masses*). Interestingly, duplicate extracts with the same sample mass did not necessarily cluster together, suggesting that sample size bias can be less than within sample variation. When each sample size was analysed independently, no significant difference was observed in the mean BC distance between samples (Table 3 and Fig. S5; *one-way ANOVA, $F_{(2,12)}=1.6$, $p=0.23$*). The relative differences between the soils varied depending on the sample size (Fig. S6) suggesting that reproducibility can alter the level of discrimination between individual samples; however, different soils were still distinguishable using any soil mass. We also found that the optimal sample size for soil discrimination was also in agreement with that which produced the most reproducible DNA profile (Table 3). For example, when comparing the two sandy loam soils (12093 and 12096), the optimal sample sizes for discrimination analyses were 150 mg and 250 mg respectively, both of which generated the most reproducible DNA profile for that individual sample. From this, we can conclude that the maximum discrimination between different soils is obtainable by minimizing within sample variation.

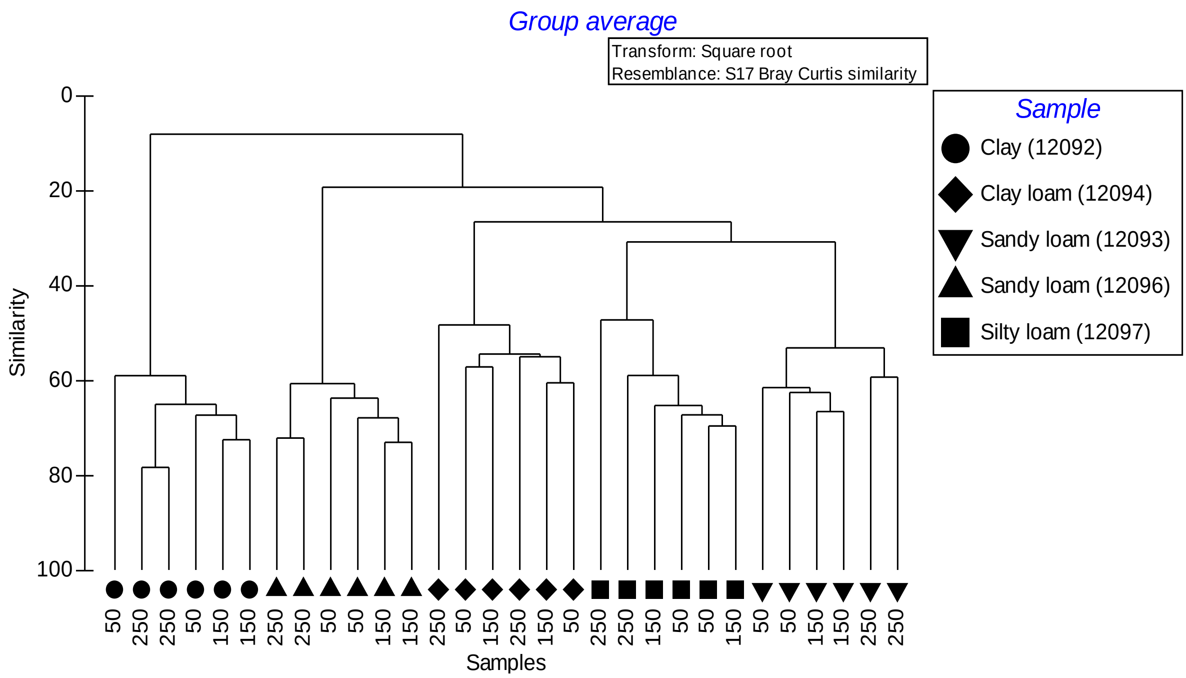
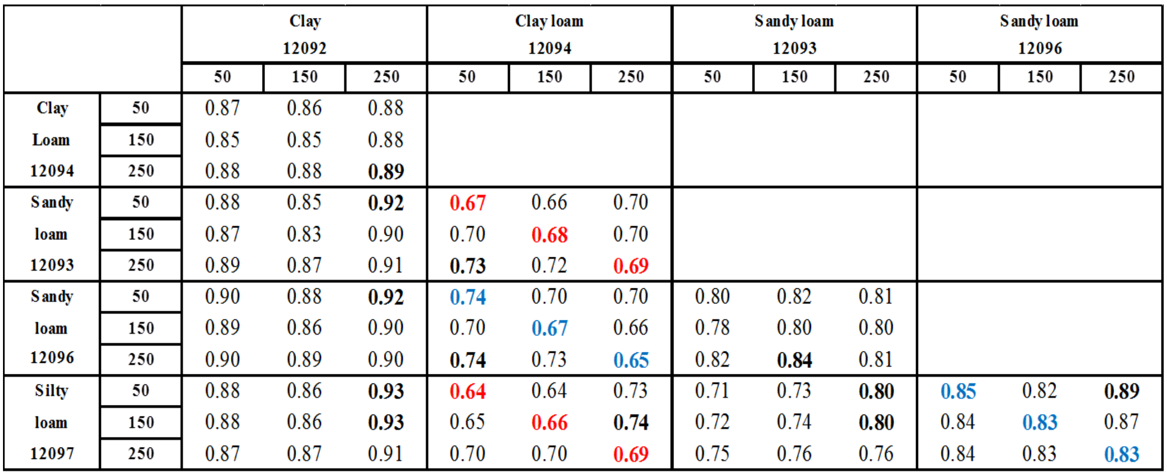


Fig. 4: Bray-Curtis cluster dendrogram based on fungal ITS profiles from five soil samples. For each sample, a total of six extracts were obtained using three different soil masses in the DNA extraction protocol; 50 mg, 150 mg and 250 mg.

Table 3: Distance matrix based on Bray-Curtis (BC) distance between pairs of DNA extracts. Values in bold highlight the highest BC distance between the two soils being compared. Red indicates comparisons for which the BC distance between the samples increases with soil mass. Blue indicates comparisons for which the BC distance between the samples decreases with soil mass.



		Clay 12092			Clay loam 12094			Sandy loam 12093			Sandy loam 12096		
		50	150	250	50	150	250	50	150	250	50	150	250
Clay	50	0.87	0.86	0.88									
Loam	150	0.85	0.85	0.88									
12094	250	0.88	0.88	**0.89**									
Sandy	50	0.88	0.85	**0.92**	0.67	0.66	0.70						
loam	150	0.87	0.83	0.90	0.70	0.68	0.70						
12093	250	0.89	0.87	0.91	**0.73**	0.72	0.69						
Sandy	50	0.90	0.88	**0.92**	0.74	0.70	0.70	0.80	0.82	0.81			
loam	150	0.89	0.86	0.90	0.70	0.67	0.66	0.78	0.80	0.80			
12096	250	0.90	0.89	0.90	**0.74**	0.73	0.65	0.82	**0.84**	0.81			
Silty	50	0.88	0.86	**0.93**	0.64	0.64	0.73	0.71	0.73	**0.80**	0.85	0.82	**0.89**
loam	150	0.88	0.86	**0.93**	0.65	0.66	**0.74**	0.72	0.74	**0.80**	0.84	0.83	0.87
12097	250	0.87	0.87	0.91	0.70	0.70	0.69	0.75	0.76	0.76	0.84	0.83	0.83

Discussion

This study demonstrates that sample size influences fungal diversity and thus reproducibility of the DNA profile, and provides an indication of the optimal soil mass for DNA analysis of soils with different texture and pH (Table 4). For fine clay soils and coarse sandy soils, a larger sample size (250 mg) produced the most reproducible DNA profile due to an increase in total OTU count. For soils with low pH, the most reproducible DNA profile was obtained using smaller sample sizes. Nevertheless, variations in fungal DNA profiles attributed to soil mass did not prevent discrimination between the soils. From these results, we demonstrate that sample size can be reduced to conserve limited quantities of soil; however, in doing so DNA profile reproducibility will vary depending on the soil type.

Table 4: Summary of the findings of this study.

Soil texture Ordered fine to coarse	**pH**	**DNA extraction efficiency**	**Number of OTUs**	**Optimal DNA profile reproducibility**
Clay soil (12092)	>7.5	Sample size dependent	Sample size dependent	250 mg
Clay loam (12094)	<7.5	n/a	n/a	50 mg
Silty loam (12097)	<7.5	n/a	n/a	50 mg
Sandy loam (12093)	>7.5	Sample size dependent	Sample size dependent	250 mg
Sandy loam (12096)	<7.5	Sample size dependent	Sample size dependent	150 mg

This study confirms that trace quantities of soil can provide valuable DNA profiles for use in forensic soil analysis. Ranjard *et al.* (15) recommended sample masses >1 g soil for reproducible fungal profiles; however, our results demonstrate that variation in OTU composition as a result of sample mass did not prevent the differentiation of soil samples and HTS profiles of trace soil samples can be robustly compared to larger reference samples. This is in agreement with previous DNA fingerprinting studies that successfully distinguished between soil samples regardless of mass, despite subtle variations in microbial DNA fingerprints (15, 16, 18). Furthermore, this indicates that sufficient material for forensic analysis could be obtained by sub-sampling (i.e. <250 mg) to provide replicate samples for DNA extraction or provide material for analysis by other methods, such as mid-infrared (MIR) spectrometry, while still maintaining a reliable and discriminative DNA profile.

Reducing the sample size to conserve limited quantities of material in casework, may jeopardised DNA profile reproducibility, particularly for fine or coarse soils. We found that the effect of sample size on fungal DNA profile reproducibility was related to soil texture. This is supported by Ranjard *et al.* (15) who showed bacterial profile variation between replicates from fine clay soil was greater than that from sand and silt soils. The current study demonstrates that sample size of fine clay soils, and coarse sandy loam soils, also influences DNA profiles reproducibility. Coarse soils consist of few large particles with a relatively small surface area for DNA to bind, so smaller sample sizes likely under-sample the fungal diversity. High clay content soils consisted of fine particles (<0.002 mm) with a high affinity for DNA molecules; therefore, the density of DNA molecules in a given mass of soil is much higher than that for soils with moderate texture. As a result, subsampling bias from high clay content soils is evident with reduced mass, whereas sampling bias is less pronounced in moderate textured soils, likely because the latter contain a more homogenous distribution of taxa (14). This study shows that soil pH can also influence DNA profile reproducibility but the effect was less pronounced than soil texture. Soils with lower pH have increased positive charge and a higher affinity for DNA molecules (26-28). Therefore, random adsorption of extracellular DNA molecules onto the low pH soils is more pronounced with larger sample sizes, generating more variable DNA profiles. Although, sample size bias did not prevent soil discrimination, where possible, it would be beneficial to use larger sample sizes for both fine and coarse textured soils to maximise the reproducibility of the sample.

DNA profile reproducibility will be more important depending on the stage of the forensic investigation, i.e. evidential stages or investigative stages. Evidential stages rely on establishing a link between a suspect and an object, victim, or location, given prior knowledge of the case. For example, soil DNA analysis for evidential value would compare unknown samples from shoes, or a shovel, to reference samples from known locations, to suggest the presence of a suspect at a particular location. For this purpose, the reproducibility of the DNA profile is less of a concern in terms of dissimilarity values. Instead, the reproducibility of a sample is taken into context; two extracts from the same sample or location, must be more similar to one another, than to any extract obtained from a different sample, or location. In this context, small sample sizes will provide a reliable result. In contrast, investigative stages rely on analysis that will direct the focus of a case to a particular location. With regards to soil DNA analysis, this would require capturing the complete diversity within a sample to identify individual taxa with specific habitat requirements or restricted geographical distribution. For a reliable result in this context, the DNA profile should be highly reproducible, so that particular taxa are identified with confidence. In this context, capturing a complete representation of the fungal diversity is crucial therefore for this purpose, large sample sizes are recommended, especially for fine clay soils and coarse soil samples.

Forensic soil discrimination relies on detection of site specific signals to discriminate between geographical locations therefore a method capable of isolating the less ubiquitous taxa from soil is of interest. Early reports using culturing methods indicated that large sample masses (>1g) detect the most abundant taxa within a soil, whereas small soil masses (<1 g) increase the detection of rare taxa (14). This suggested that the discriminatory power between forensic samples could potentially be increased using smaller sample sizes. However, our results indicate that this was not the case as the mean discriminatory power was comparable at each sample size. Instead, we indicate that discriminatory power between pairs of soils can be improved by tailoring the sample mass according to soil particle size; by reducing within sample variation between sample variation increased. However, as forensic analysis favours standardisation of methods this may be difficult to achieve when analysing across different soil types.

Analysis of soil fungal DNA is a promising target for forensic soil discrimination, based on high discriminatory power and reproducible DNA profiles (20). However, a potential limitation of this method was that only a small quantity may be available for analysis. Here, we demonstrate that HTS can provide reliable fungal DNA profiles from as little as 50 mg of soil and show that trace samples can be robustly compared to larger samples for evidential stages of an investigation. We show that soil texture can hinder DNA profile reproducibility at small sample sizes, particularly for very fine clay soils and coarse soils. As a result, soil DNA analysis for investigative purposes would benefit from larger sample sizes to minimise sample bias and capture a more complete picture of the fungal diversity in a sample. Increased DNA profile reproducibility would enable identification of key taxa with confidence. Although this study addresses sample size as a potential pitfall of HTS of soil biota for forensic application, other factors such as transfer effects, storage conditions and temporal variation are also commonly encountered during an investigation, each of which may influence the fungal DNA profiles generated. Therefore further studies are required to establish the robustness of soil fungal DNA analysis using HTS before this method can be routinely applied in forensic science.

Acknowledgements

This work was supported by the Australian Research Council, and the Australian Federal Police. We thank Alla Marchuk for providing the samples and details of soil properties. We thank members of the Australian Centre for Ancient DNA for helpful comments on the manuscript.

References

1. **Ruffell, A.** 2010. Forensic pedology, forensic geology, forensic geoscience, geoforensics and soil forensics. Forensic Science International 202:9-12.

2. **Pye, K.** 2007. Geological and soil evidence. Forensic applications.

3. **Morgan, R. M., and P. A. Bull.** 2007. The philosophy, nature and practice of forensic sediment analysis. Progress in Physical Geography 31:43-58.

4. **Ritz, K., L. Dawson, and D. Miller.** 2008. Criminal and environmental soil forensics. Springer.

5. **Concheri, G**. 2011. Chemical elemental distribution and soil DNA fingerprints provide the critical evidence in murder case investigation. PloS ONE 6:e20222.

6. **Cano, R. J.** 2010. Molecular Microbial Forensics. Royal Soc Chemistry, Cambridge.

7. **Martiny, J. B. H., B. J. M. Bohannan, J. H. Brown, R. K. Colwell, J. A. Fuhrman, J. L. Green, M. C. Horner-Devine, M. Kane, J. A. Krumins, C. R. Kuske, P. J. Morin, S. Naeem, L. Ovreas, A. L. Reysenbach, V. H. Smith, and J. T. Staley.** 2006. Microbial biogeography: putting microorganisms on the map. Nat. Rev. Microbiol. 4:102-112.

8. **Dequiedt, S., N. P. A. Saby, M. Lelievre, C. Jolivet, J. Thioulouse, B. Toutain, D. Arrouays, A. Bispo, P. Lemanceau, and L. Ranjard.** 2011. Biogeographical patterns of soil molecular microbial biomass as influenced by soil characteristics and management. Glob. Ecol. Biogeogr. 20:641-652.

9. **Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. A. Gilbert, G. Smith, and R. Knight.** 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621-1624.

10. **Metzker, M. L.** 2010. Applications of Next Generation Sequencing - the next generation. Nat. Rev. Genet. 11:31-46.

11. **Metzker, M. L.** 2010. Next Generation Technologies: Basics and Applications. Environ. Mol. Mutagen. 51:691-691.

12. **Magi, A., M. Benelli, A. Gozzini, F. Girolami, F. Torricelli, and M. L. Brandi.** 2010. Bioinformatics for Next Generation Sequencing Data. Genes 1:294-307.

13. **Pye, K., and D. J. Croft.** 2004. Forensic geoscience: introduction and overview. Geological Society, London, Special Publications 232:1-5.

14. **Grundmann, L. G., and F. Gourbiere**. 1999. A micro-sampling approach to improve the inventory of bacterial diversity in soil. Applied Soil Ecology 13:123-126.

15. **Ranjard, L., D. P. H. Lejon, C. Mougel, L. Schehrer, D. Merdinoglu, and R. Chaussod.** 2003. Sampling strategy in molecular microbial ecology: influence of soil sample size on DNA fingerprinting analysis of fungal and bacterial communities. Environ. Microbiol. 5:1111-1120.

16. **Ellingsøe, P., and K. Johnsen.** 2002. Influence of soil sample sizes on the assessment of bacterial community structure. Soil Biol. Biochem. 34:1701-1707.

17. **Taberlet, P., S. M. Prud'homme, E. Campione, J. Roy, C. Miquel, W. Shehzad, L. Gielly, D. Rioux, P. Choler, and J. C. Clement.** 2012. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Mol. Ecol. 21:1816-1820.

18. **Kang, S., and A. L. Mills**. 2006. The effect of sample size in studies of soil microbial community structure. J. Microbiol. Methods 66:242-250.

19. **Fitzpatrick, R. W., M. D. Raven, and S. T. Forrester.** 2009. A systematic approach to soil forensics: criminal case studies involving transference from crime scene to forensic evidence. Criminal and Environmental Soil Forensics:105.

20. **Macdonald, C. A.** 2011. Discrimination of soils at regional and local levels using bacterial and fungal t-RFLP profiling. Journal of Forensic Sciences 56:61-69.

21. **Epp, L. S., S. Boessenkool, E. P. Bellemain, J. Haile, A. Esposito, T. Riaz, C. Erseus, V. I. Gusarov, M. E. Edwards, A. Johnsen, H. K. Stenoien, K. Hassel, H. Kauserud, N. G. Yoccoz, K. Brathen, E. Willerslev, P. Taberlet, E. Coissac, and C. Brochmann.** 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Mol. Ecol. 21:1821-1833.

22. **Meyer, M., and M. Kircher**. 2010. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb. Protoc. 2010:pdb.prot5448.

23. **Martin, M.** 2012. Cutadapt removes adapter sequences from high-throughput sequencing reads. Bioinformatics in Action 17:10-12.

24. **Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J.**

Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Meth. 7:335-336.

25. **Edgar, R. C.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460-2461.

26. **Cai, P., Q. Huang, D. Jiang, X. Rong, and W. Liang.** 2006. Microcalorimetric studies on the adsorption of DNA by soil colloidal particles. Colloids and Surfaces B: Biointerfaces 49:49-54.

27. **Saeki, K., M. Sakai, and S. I. Wada.** 2010. DNA adsorption on synthetic and natural allophanes. Applied Clay Science 50:493-497.

28. **Shen, Y., H. Kim, M. P. Tong, and Q. Y. Li**. 2011. Influence of solution chemistry on the deposition and detachment kinetics of RNA on silica surfaces. Colloid Surf. B-Biointerfaces 82:443-449.

Supplementary Material

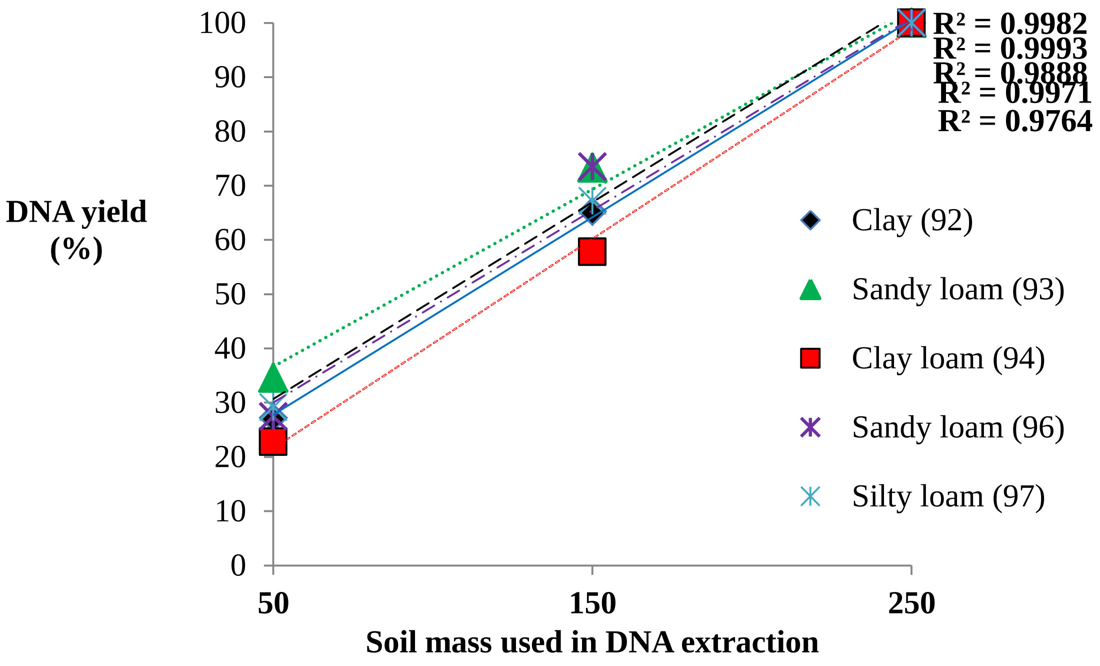


Fig. S1: Percent decrease in DNA yield with small sample size.

Table S1: Pair-wise one-way ANOVA significance values for the DNA yield (ng) detected using reduced mass of soil (50 mg, 150 mg and 250mg) in DNA extraction.

(I) Mass		12092	12094	12093	12096	12097
Between	groups	0.013	0.005	0.03	0.127	0.015
50	150	.039	.214	.011	.159	.050
	250	.006	.014	.002	.060	.006
150	50	.039	.214	.011	.159	.050
	250	.038	.037	.028	.356	.035
250	50	.006	.014	.002	.060	.006
	150	.038	.037	.028	.356	.035

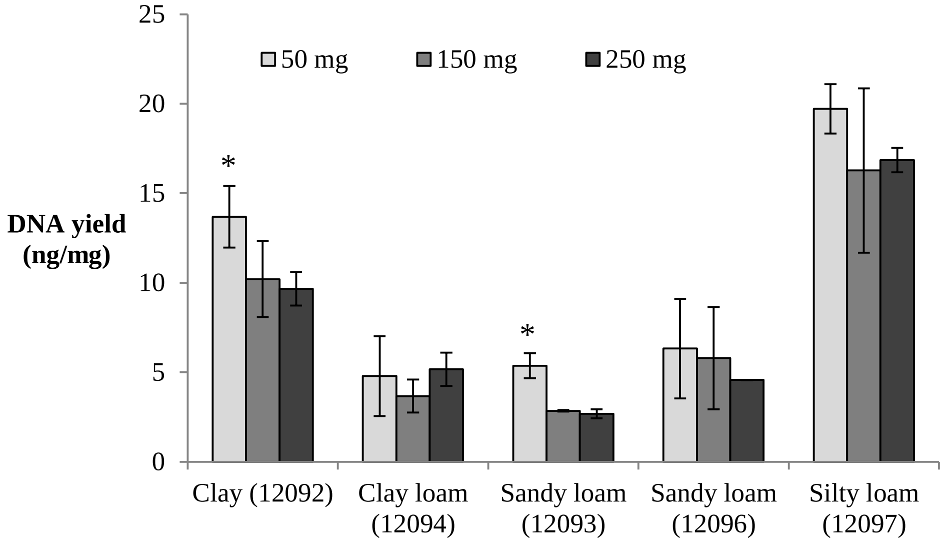


Fig. S2: DNA yield (ng/mg soil) using reduced mass of soil (50 mg, 150 mg and 250mg) in DNA extraction. Values are means ± SD from duplicate extracts (*n=2*) at each mass.

* indicates a significant difference (*one-way ANOVA, p<0.05*) in DNA yield per mg between 50 mg and 250 mg for a given soil.

Table S2: Summary of the number of sequences per sample.

Sample	Soil mass used in DNA extraction (mg)	Sequences per MID	Trim primers, MIDs and length filter	Quality filter
92	50	319641	234116	82037
		212987	162686	37293
	150	341028	237470	40944
		312992	225876	39588
	250	188507	123651	9863
		203889	132642	9691
93	50	324785	202165	53236
		313255	196476	51358
	150	369297	252767	39474
		355052	232151	35049
	250	**184809**	**68555**	**2642**
		224310	105420	4549
94	50	303327	168700	39481
		277189	159101	14261
	150	325805	200837	27399
		335488	190353	17076
	250	224310	73743	3331
		209497	87954	4292
96	50	337101	233503	35562
		308543	224717	55846
	150	332139	248252	31642
		347355	241404	36845
	250	225909	131332	6989
		215889	136469	7936
97	50	330810	222297	63919
		304337	195319	54749
	150	279470	195048	49366
		249431	179509	42561
	250	211416	111501	4816
		186836	101613	9781

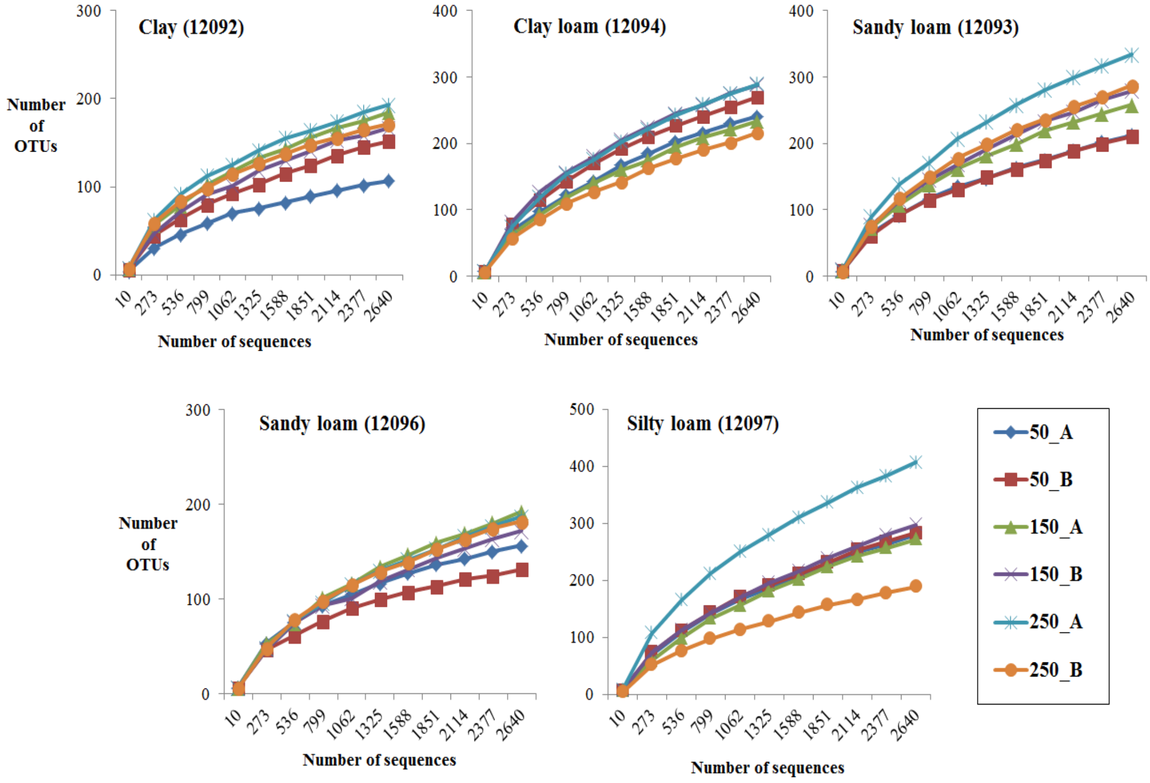


Fig. S3: Rarefaction curves for each of the five soil samples.

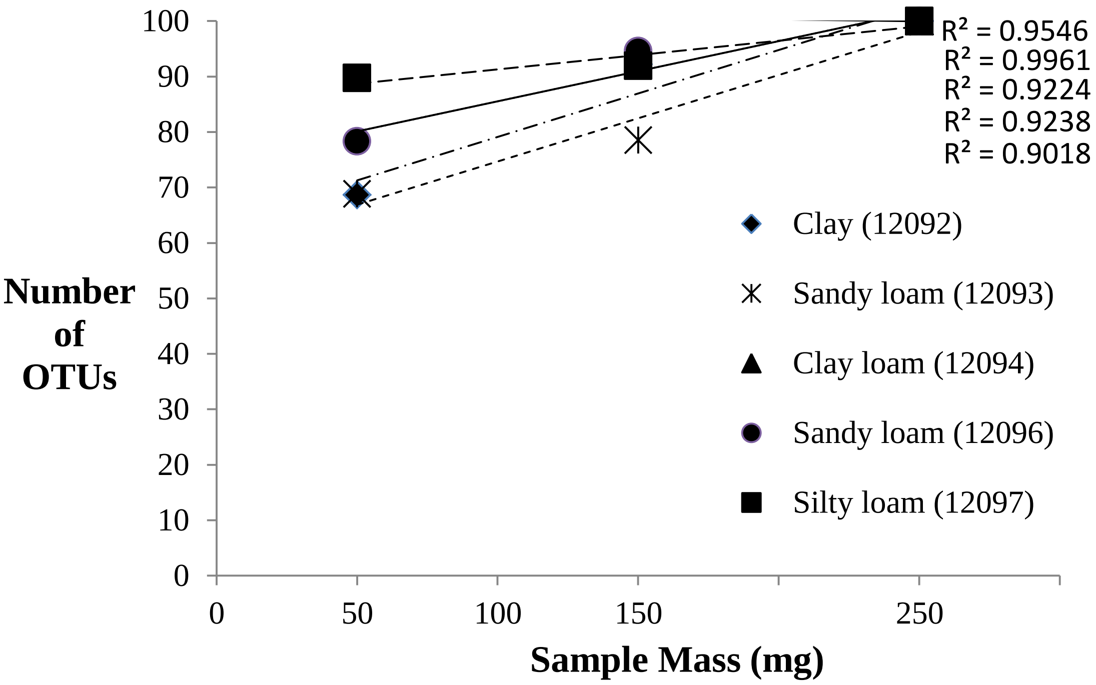


Fig. S4: Percent decrease in OTU count with reduced soil mass used in DNA extraction of five soil types.

Table S3: Pair-wise one-way ANOVA significance values for the number of OTUs detected using reduced mass of soil (50 mg, 150 mg and 250mg) in DNA extraction.

Mass		12092	12094	12093	12096	12097
50	150	0.06	0.95	0.36	0.09	0.95
	250	**0.03**	0.93	**0.04**	**0.05**	0.75
150	50	0.06	0.95	0.36	0.09	0.95
	250	0.40	0.98	0.10	0.48	0.80
250	50	**0.03**	0.93	**0.04**	**0.05**	0.75
	150	0.40	0.98	0.10	0.48	0.80

Table S4: Relative abundance of each Phylum detected in duplicate extracts of five soil samples using three different sample sizes in the DNA extraction.

Extract	Ascomycota	Basidiomycota	Blastocladiomycota	Chytridiomycota	Glomeromycota	Neocallimastigomycota	Zygomycota	Unknown Fungi	Total
92_50_A	8.509	60.679	0.000	0.002	0.015	0.000	0.161	30.634	100
92_50_B	22.491	1.364	0.000	0.006	0.086	0.000	0.690	75.363	100
92_150_A	25.686	7.696	0.000	0.000	0.070	0.000	0.591	65.958	100
92_150_B	17.490	22.584	0.000	0.005	0.042	0.000	0.406	59.472	100
92_250_A	10.540	17.098	0.000	0.000	0.033	0.000	2.490	69.840	100
92_250_B	7.344	11.605	0.012	0.000	0.023	0.000	0.304	80.712	100
93_50_A	18.773	2.480	0.023	0.035	0.010	0.000	28.098	50.581	100
93_50_B	17.006	2.588	0.006	0.038	0.006	0.000	23.854	56.502	100
93_150_A	28.635	4.844	0.194	0.032	0.035	0.000	16.404	49.856	100
93_150_B	26.529	3.955	0.000	0.015	0.009	0.000	15.489	54.003	100
93_250_A	18.069	6.170	0.601	0.040	0.000	0.000	33.534	41.587	100
93_250_B	21.761	12.700	0.047	0.141	0.023	0.000	35.141	30.188	100
94_50_A	35.222	2.732	0.000	0.113	0.193	0.000	2.437	59.302	100
94_50_B	21.975	7.837	0.000	0.069	0.023	0.000	3.534	66.562	100
94_150_A	31.063	2.808	0.000	0.000	0.173	0.000	3.422	62.533	100
94_150_B	34.299	5.323	0.000	0.042	0.090	0.000	5.108	55.138	100
94_250_A	26.269	7.704	0.000	0.031	0.093	0.093	6.838	58.973	100
94_250_B	28.592	2.411	0.000	0.000	0.073	0.000	9.230	59.693	100
96_50_A	41.909	10.890	0.000	0.013	0.063	0.000	18.325	28.800	100
96_50_B	20.716	5.855	0.000	0.000	0.019	0.000	23.167	50.243	100
96_150_A	28.704	7.467	0.000	0.000	0.008	0.000	42.587	21.233	100
96_150_B	27.628	10.970	0.000	0.000	0.000	0.000	35.260	26.142	100
96_250_A	6.065	10.051	0.000	0.000	0.000	0.000	36.116	47.768	100
96_250_B	7.025	12.398	0.014	0.000	0.000	0.000	31.626	48.938	100
97_50_A	31.320	0.486	0.018	0.062	0.033	0.000	1.813	66.268	100
97_50_B	33.696	0.625	0.019	0.056	0.026	0.000	3.199	62.380	100
97_150_A	23.091	0.459	0.020	0.025	0.033	0.000	6.240	70.133	100
97_150_B	27.984	0.533	0.005	0.033	0.024	0.000	3.656	67.765	100
97_250_A	25.147	1.438	0.022	0.000	0.087	0.000	14.121	59.185	100
97_250_B	21.848	0.686	0.000	0.000	0.000	0.000	5.193	72.273	100

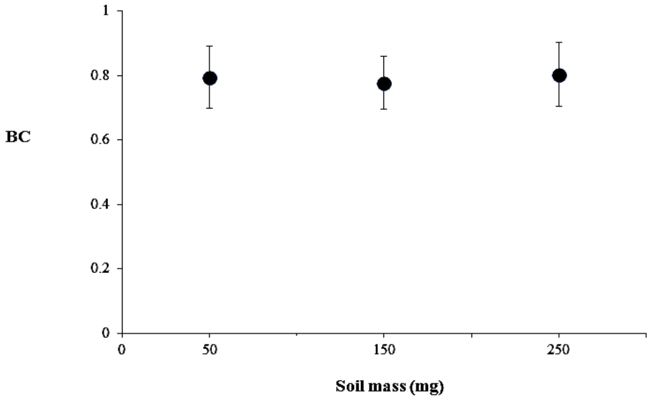


Fig. S5: Discriminatory power (Bray-Curtis distance) between five soil samples using three different starting soil masses (250 mg, 150mg and 50 mg) in the DNA extraction. Values are means ± SD of pairwise comparisons (n=10); a high dissimilarity value shows high discriminatory power.

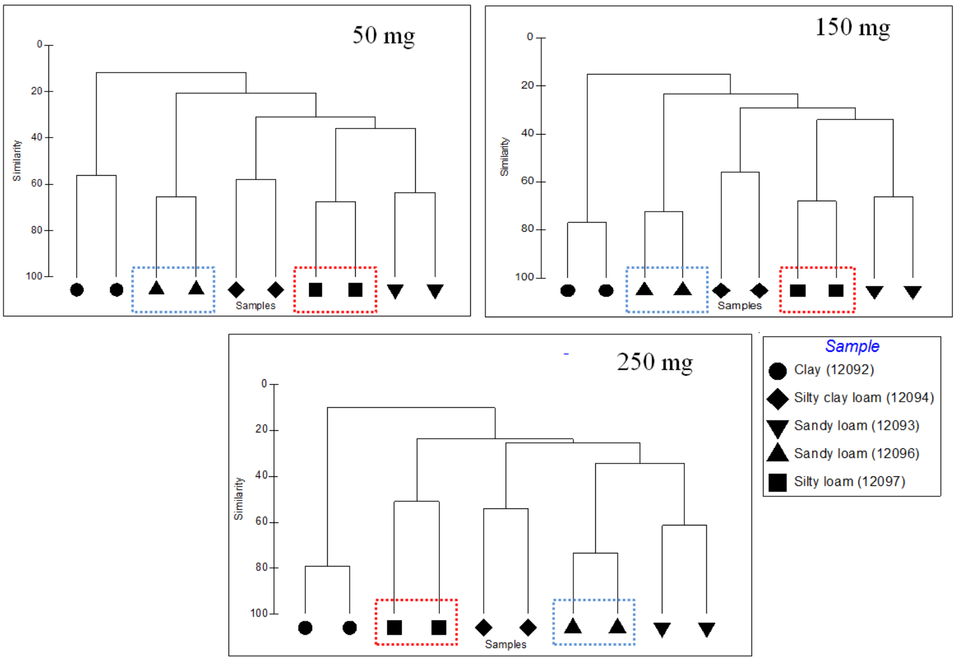


Fig. S6: Bray Curtis cluster dendrogram based on fungal ITS profiles from five soil samples using three different soil sample sizes in the DNA extraction. BLUE shows the increased resolution of sample 12096 due to increased reproducibility with 250 mg soil. RED shows the decreased resolution of sample 12097 due to decreased reproducibility with 250 mg soil.

CHAPTER 6

Predicting the origin of soil evidence: high-throughput eukaryote sequencing and MIR spectroscopy applied to a crime scene scenario

Young, J.M., L.S. Weyrich, J. Breen, L. MacDonald, A. Cooper. High-throughput sequencing of soil eukaryotes links a suspect to a crime scene: an experimental case study
(In review at *Forensic Science International*)

Statement of authorship

Predicting the origin of soil evidence: high-throughput eukaryote sequencing and MIR spectroscopy applied to a crime scene scenario

In preparation for *Forensic Science International: Genetics*

Jennifer M. Young (Candidate)

Designed experiment, DNA extractions, PCR amplifications, and library preparation, conducted downstream processing and analysis of data, interpretation of results; created tables, and wrote the paper.

I hereby certify that the statement of contribution is accurate

Signed.. ...... Date.....04/07/2014

Laura S. Weyrich

Provided advice on experimental design, interpretation of results, content, structure and edited manuscript

I hereby certify that the statement of contribution is accurate

Signed ..... Date.....04/07/2014

Lynne Macdonald

Performed MIR spectroscopy and spectral analysis, interpretation of results and edited the manuscript

I hereby certify that the statement of contribution is accurate

Signed... .... Date.....4/07/2014

James Breen

Performed 18S sequencing, data filtering, OTU picking, edited the manuscript

I hereby certify that the statement of contribution is accurate

Signed... Date.....4/07/2014

Alan Cooper

Edited the manuscript

I hereby certifv that the statement of contribution is accurate

Signed.... Date.....04/07/2014

Abstract

Soil can serve as powerful trace evidence in forensic casework, because it is highly individualistic and can be characterised using a number of techniques. Complex soil matrixes can support a vast number of organisms that can provide a site-specific signal for use in forensic soil discrimination. Previous DNA fingerprinting techniques rely on variations in fragment length to distinguish between soil profiles and focus solely on microbial communities. However, the recent development of high throughput sequencing (HTS) has the potential to provide a more detailed picture of the soil community by accessing non-culturable microorganisms and by identifying specific bacteria, fungi, and plants within soil. To demonstrate the application of HTS to forensic soil analysis, 18S ribosomal RNA profiles of six forensic mock crime scene samples were compared to those collected from seven reference locations across South Australia. Our results demonstrate the utility of non-bacterial DNA to discriminate between different sites, and were able to link a soil to a particular location. In addition, HTS complemented traditional Mid Infrared (MIR) spectroscopy soil profiling, but was able to provide statistically stronger discriminatory power at a finer scale. Through the design of an experimental case scenario, we highlight the considerations and potential limitations of this method in forensic casework. We show that HTS analysis of soil eukaryotes was robust to environmental variation, e.g. rainfall and temperature, transfer effects, storage effects and spatial variation. In addition, this study utilizes novel analytical methodologies to interpret results for investigative purposes and provides prediction statistics to support soil DNA analysis for evidential stages of a case.

Introduction

Soil presents an ideal form of trace evidence in criminal investigations due to individual characteristics that relate to provenance, including underlying mineralogy and vegetation cover. Soil particles have a high transfer and retention probability, and soil adhered to objects, such as footwear and car tyres, is commonly overlooked by a suspect in attempts to conceal evidence. Consequently, soil particles recovered from crime scenes can potentially provide a wealth of information. Both chemical and biological signals generated from soils can indicate the provenance of a sample for investigative purposes (1), or establish a link between a suspect and an object, site, or victim for evidential value (2, 3). Traditional soil analyses include colour, and particle size, as well as mineralogical and chemical analysis, including Mid Infrared (MIR) spectroscopy (1, 4, 5). Such techniques can be useful for initial screening of samples (1, 6); however, often more intricate analyses are required to discriminate samples. These techniques can also be limited by regional scale resolution as variation is driven by underlying geology (7). In contrast, biological signals afforded from DNA fingerprinting methods (8) offer an alternative method with fine scale variation (9), as demonstrated by Young *et al.* who examined reproducibility of high-throughput sequencing from different soil types (10). Although soil bacterial DNA profiling has been previously accepted in court (2), the most extensively used DNA fingerprinting method is *Terminal Restriction Fragment Length Polymorphism* (T-RFLP) analysis, even though this method cannot identify specific microorganisms. High-throughput sequencing (HTS) has previously been applied to link microbial communities on keyboards (11) and bite marks (12) to specific individuals. This approach remains a promising new methodology for forensic soil analysis (10, 13-15). However, the robustness of this technique must be validated as microbial soil profiles are continuously developing and adapting and have no distinct boundary.

During forensic soil discrimination analysis, an ideal analytical target will be endemic in both the forensic samples and the environmental site of interest. Bacteria can generate a highly site specific DNA profile, as the community structure can be influenced by soil type, seasonal variation, site management, vegetation cover and environmental conditions (16-19). Therefore, the focus of soil DNA fingerprinting analysis for forensic application has been microbes, specifically targeting the bacterial 16S ribosomal RNA (rRNA) gene region. However, soil fungal DNA profiles have been reported to be more discriminative than bacterial DNA profiles (10, 20). In addition, the fungal profile appears to be more robust under changing soil conditions, such as drying (21). Recently, Young *et al.* demonstrated that the internal transcribed spacer (ITS1) region for fungal specific profiles and the 18S rRNA gene region for general eukaryote diversity were less susceptible to contamination, in comparison to bacterial DNA profiles (10). Although, soil eukaryotes appear promising for soil discrimination, the differential resolution of sites within similar locations has been questioned, notably a recent study using T-RFLP analysis which described similar fungal compositions from similar soil types despite pronounced geographic separation (22). Although this conservatism could be valuable for identifying particular soil types at investigative stages of a case, it may become a significant issue when soil is used to discriminate between similar locations. As HTS analysis is capable of identifying individual taxa present in soils, unlike T-RFLP, there is a requirement for further examination of fungal markers in these contexts.

In addition to providing resolving power, a genetic target must also be robust to practical factors that could influence the soil DNA profile. Often, a temporary gap exists between the time of a crime and retrieval of a forensic soil sample, so the effects of seasonal variation, drying, and sample transfer, i.e. removal of soil from the crime scene, must be considered before forensic soil analysis can be robustly utilized in court. Seasonal variation is not regarded as an issue for mineralogical analysis as the underlying geology is not affected; however, the soil DNA profile can potentially be considerably altered. Initial studies using 16S rRNA T-RFLP showed monthly fluctuations in bacterial community structure (23). Similarly, the presence of micro-fauna can fluctuate with elevated temperature and frequency of summer precipitation, although this effect has not been shown to be statistically significant for most groups using ANOVA analysis (24). These issues clearly require further investigation. Soil removed from the environment during the crime event, as a result of contact with materials (25), will also be subjected to varying conditions (e.g. temperature, sunlight, humidity, moisture) depending on the circumstances of a case and relocation. For example, soils adhering to shoes or shovels often dry out during storage, potentially altering the DNA profile. Transfer of soil to an object can also be biased according to particle size, which is dependent upon soil properties, mineralogy, and the type of contact, e.g. footwear (26, 27). Furthermore, layers or mixtures of soil are commonly encountered on objects, such as shoes or shovels, due to sequential use before and after the crime. Therefore it is unlikely that the evidence samples will contain soil exclusively from the crime scene, potentially complicating DNA analysis using highly sensitive HTS technology. Although the alteration of morphological composition raises concern for physical analysis methods, the effect of transfer on DNA profiles using HTS is unknown. Evaluation of these factors is imperative for the implementation of HTS to forensic science.

In this study, we assessed two forensic soil profiling techniques to predict the provenance of soil collected during a mock case study, which included soil evidence from shoes, a shovel, and a car boot. To achieve this, eukaryotic microorganism DNA profiles and mid-infrared (MIR) spectra from evidential samples were compared to several ‘crime scene’ and ‘alibi sites’ at varying distances from this location, including similar and different soil types. For both methods, we predict the most likely location of each evidential sample, and report the degree of confidence in correctly identifying the crime scene as the source. Reference and evidential samples were collected six weeks after the ‘crime’ and reference samples were dried prior to analysis to assess the effect of seasonal variation, sample transfer and desiccation on the ability to discriminate between the sites.

Materials and Methods

Experimental case study

To provide appropriate forensic context for this study, a fictional murder and body dump scenario was developed. The scenario locations, and distances are presented in Fig. 1A, and the timeline of events and weather conditions are provided in Figure 1B with Day 1 representing the disposal of the body. In brief, a woman was reported missing (Day 5), and her body was found approximately three weeks later (Day 27) in a roadside verge in the Tooperang area of South Australia (Fig. 1A, 0 m). The pathologists report indicated that the woman was most likely killed before being taken to this location, as limited blood was observed where the body was found. The forensic report indicated a DNA database match with a male DNA profile that was obtained from a semen stain on the woman’s clothing (Day 33). Subsequent questioning of the suspect confirmed that they had had intercourse the day she was last seen alive (Day 1), but the suspect claimed the woman was alive when he left her apartment that evening. The suspect claims he had never been to the Tooperang area and agreed to a search of his vehicle, where the shoes and shovel were recovered.

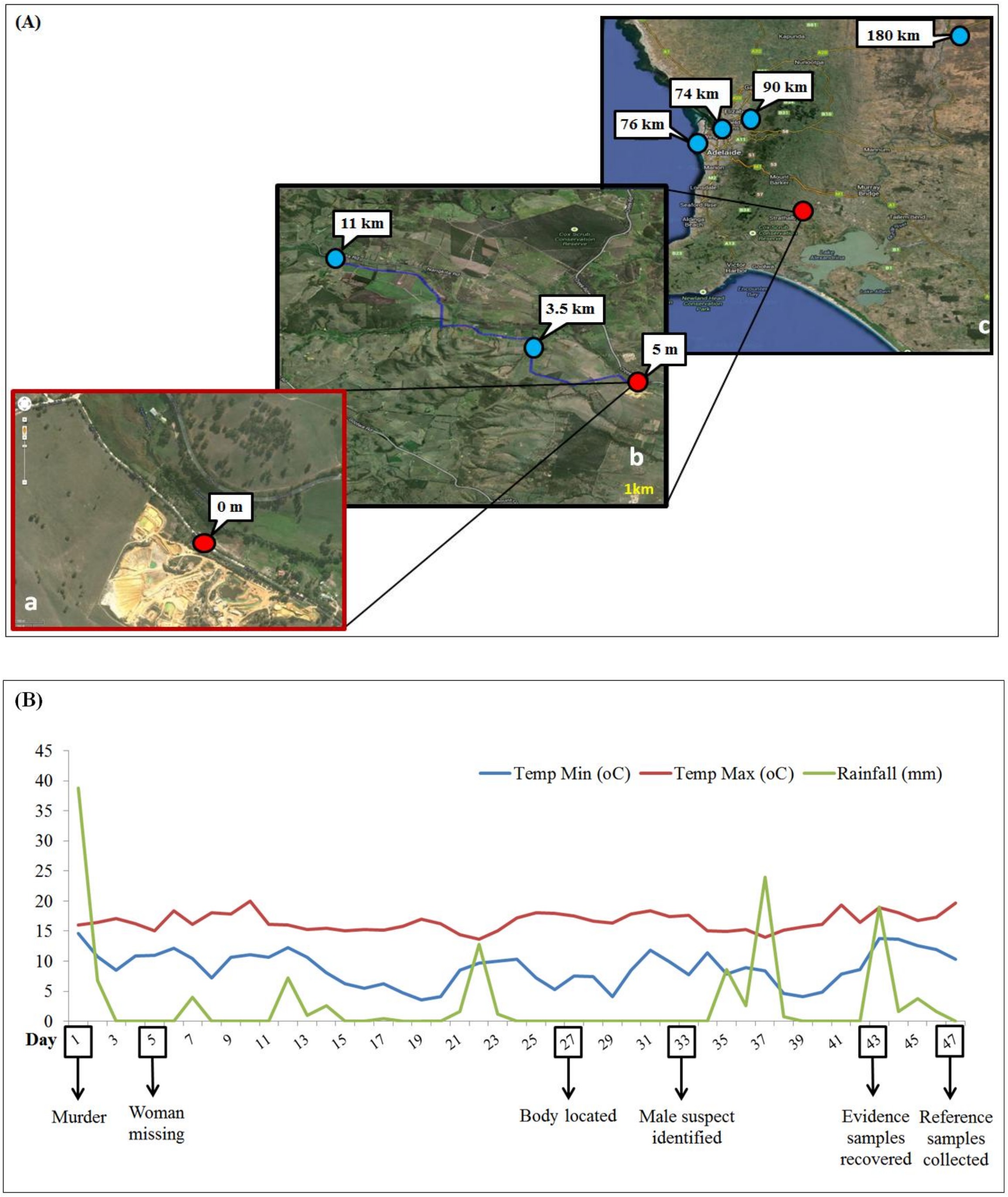


Fig. 1: (A) Aerial view of the crime scene and the sites from which reference samples were collected at increasing distance, and (B) the series of events from criminal activity to collection of soil samples. The minimum and maximum daily temperature (^{O}C) and the average daily rainfall (as recorded by the Bureau of Meteorology Australian Government) show the temporal variation over the course of the investigation.

Sample collection

In order to represent criminal activities at the disposal site, an unsealed road side verge in the Tooperang area was visited and a shovel was used to dig a shallow hole (depth of two feet) (Fig. 1A), which did not disturb the underlying clay layer that is typical of the area. It was not raining at the time; however, a significant rainfall event (37 mm) had occurred in the preceding days (Fig. 1B). The used shovel and the shoes (trainers) worn by the 'suspect' were placed unwrapped in the car boot of a Ford Falcon Wagon at the scene. The shovel and shoes were left undisturbed in the car boot for six weeks. In addition, shoes and the shovel used in the scenario were not new or cleaned prior to crime scene set-up to reflect a real-life scenario. Six weeks later (Day 43), soil samples were collected from the shoes, shovel, and car boot using sterile 15 ml CELLSTAR® tubes and placed at 4 ^{o}C, including two shoe samples (right shoe and left shoe), three shovel samples (A, B and C) and one car boot sample (Table 1).

To collect the comparison soil reference samples, approximately six grams was collected from the top two cm of the soil using sterile 15 ml CELLSTAR® tubes and stored at 4 ^{o}C (Day 47). First, soil samples were collected from the exact location where the body was found (0 m; Fig. 1A, Table 1), and three additional samples were collected 5 meters from the site of the body (5 m; Fig. 1A, Table 1). Second, samples from three locations of similar location type, i.e. roadside verges, were collected at increasing distances from the crime scene (3.5 km, 11 km and 90 km) (Fig. 1A). In addition, samples from three different locations with various soil types were also collected: city park area (74 km), coastal area (76 km) and an arid area (180 km). To capture a representative diversity and account for spatial variation at each site, three soil samples were taken 5 meters apart at each reference location and each was analysed separately. To capture a representative diversity and account for spatial variation at each site, three soil samples were taken 5 meters apart along a transect at each reference location and each was analysed separately. The crime scene samples, as well as the reference samples, were collected at 0 m, 5 m, 3.5 km, 11 km, 90 km and 74 km were all dark brown in colour (high organic content) and of moderate texture. These samples also had a high moisture content due to recent rainfall events (Figure 1B). In contrast, the samples collected from the coastal site (76 km) and the arid site (180 km) were pale yellow and orange in colour respectively, and both of a fine sandy texture. Although such samples would likely be excluded upon initial screening in practice, the sandy samples were included in this study as outliers to demonstrate the variation in eukaryote diversity detected by the method.

Table 1: Details of the soil samples collected for this study.

Sample ID	Sample Type	Distance from crime scene	Location of sample	Origin of sample	Location Type
EBC1	Extraction blank control	n/a	n/a	n/a	n/a
E1R	Evidence	Unknown	Right shoe sole	Unknown	Unknown
E1L	Evidence	Unknown	Left shoe sole	Unknown	Unknown
E2A	Evidence	Unknown	Shovel	Unknown	Unknown
E2B	Evidence	Unknown	Shovel	Unknown	Unknown
E2C	Evidence	Unknown	Shovel	Unknown	Unknown
E3	Evidence	Unknown	Car boot	Unknown	Unknown
EBC2	Extraction blank control	n/a	n/a	n/a	n/a
X	Reference	0	Location of body	-35.389,138.748	Roadside verge
R1	Reference	5m	Crime scene	-35.389,138.748	Roadside verge
R2	Reference	3.5km	Olsen Road	-35.379,138.722	Roadside verge
R3	Reference	11km	Willowburn Drive	-35.347,138.665	Roadside verge
EBC3	Extraction blank control	n/a	n/a	n/a	n/a
R4	Reference	90km	Barker Road	-34.779,138.672	Roadside verge
R5	Reference	74km	Pioneer Womans Memorial Gardens	-34.916,138.598	City Parkland
R6	Reference	76km	Henley Beach	-34.917,138.494	Coastal area
R7	Reference	180km	Swan Reach	-34.546,139.601	Arid area
EBC4	Extraction blank control	n/a	n/a	n/a	n/a

MIR spectrometry analysis

Soil sub-samples (1 g) were air-dried at 50^{o}C for 48 hours and finely ground in a Retsch MM400 grinding mill (28 Hz, 180 s) before diffuse reflectance mid-infrared (MIR) spectra (Nicolet 6700 FTIR spectrometer equipped with a KBr beam-splitter and a DTGS detector, Thermo Fisher Scientific Inc, MA, USA) were acquired over 8000 – 4000 cm^{-1} (resolution of 8 cm^{-1}), as described in Baldock *et al.* (25). The background signal intensity of the silicon carbide disk was corrected following 240 background scans. For each soil sample, 60 scans were acquired and averaged to produce a reflectance spectrum, which was converted to an absorbance spectra using Omics software (v8). Spectral data were means centred, baseline corrected, and truncated to include wavenumber region 6000 – 1029 nm. Within Excel, data were smoothed by a factor of five to allow reduction of data points without a significant loss of information. The spectra were converted to the second differential to enhance the detection of small peaks, which can have greater chemical importance compared to the overlying signal intensity and provide greater separation of overlapping peaks (26-28).

18S rRNA analysis using HTS

DNA Extraction

Within 24 hours of collection, DNA was extracted from 250 mg of each soil sample using the PowerSoil DNA Isolation kit (MOBIO, Carlsbad, CA, USA), following manufacturer's instructions. To reduce cross-contamination, evidential and reference samples were extracted separately, and evidential samples were processed first. Two extraction blank controls (EBC) were also included in parallel when processing the evidential samples (one prior to samples and one after samples), and four EBCs were included in parallel with reference samples (Table S1). In addition, a subset of the reference samples were air-dried at 25 ^{o}C for 72 hours in the laboratory prior to DNA extraction to examine the effects of temporal variation and desiccation during the course of the investigation (Fig. 1B). Samples 0 m (A, B and C), 5 m (A, B and C), 3.5 km (A) and 11 km (A) were examined in dried and wet forms (Table 3).

PCR amplification and library preparation

All DNA extracts were PCR amplified using the same protocol, and evidence samples were amplified independently from the reference samples. DNA extracts and extraction blank controls were amplified using universal eukaryote 18S rRNA primers that were modified to include Illumina sequencing adapters (underlined) and unique 12 bp Golay barcodes (Table S1):1391F_Euk forward primer, 5’- AATGATACGGCGACCACCGAGATCTACAC TATCGCCGTT CG GTACACACCGCCCGTC-3’; EukBr reverse primer 3’-; CAAGCAGAAGACGGCATACGAGATnnnnnnnnnnnTCCCTTGTCTCCAGTCAGTC AGCATGATCCTTCTGCAGGTTCACCTAC -5’)(29). PCR amplifications were performed in a 25 μl reaction mix containing 2.5 mM $MgCl_2$, 0.24 mM dNTPs, 0.24 μm of each primer, 0.4 mg/μl bovine serum albumin, 0.5 U Amplitaq Gold DNA polymerase in 10x reaction buffer (Applied Biosystems, Melbourne Australia), and 1 μl DNA extract. The PCR protocol including the following parameters: 9 mins at 94 ^{o}C, followed by 35 cycles of 94 ^{o}C for 30 sec, 62 ^{o}C for 20 sec, and 72 ^{o}C for 45 sec, and a final extension at 72^{o}C for 7 mins. PCR amplifications were performed in triplicate and pooled to minimise PCR bias, and a no-template PCR amplification control was included for each barcode to monitor background DNA levels in PCR reagents. No PCR product was visible for the no-template PCR amplifications so these were not sequenced. Triplicate PCR products were pooled and purified using an Agencourt AMPure XP PCR Purification kit (Beckman Coulter Genomics, NSW), and each was quantified using the HS dsDNA Qubit Assay on a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). Purified PCR products from all samples (*n=43*) were pooled to equimolar concentration, and the library was diluted to 2nM and sequenced using a 300 cycle Illumina MiSeq kit.

All individually indexed 18S rRNA libraries were de-multiplexed from raw bcl files using CASAVA version 1.8.2 (http://support.illumina.com/sequencing/sequencing_software/casava.ilmn), allowing for one mismatch in the index because indices were separated by ≥2 bp. Samples were then processed to remove sequencing adapters using Cutadapt v.1.1 (30), which removed reads less than 100bp, and trimmed reads were filtered for sequence quality of less than Q20 over 90% of each sequence using fastx toolkit v.0.0.14 (http://hannonlab.cshi.edu/fastx_toolkit).

Processed sequences were then formatted for use with QIIME v.1.8.0 (http://genomics.azcc.arizona.edu/help.php3), where sequences with greater than 97% similarity to the SILVA v104 reference database (31) were binned into Operational Taxonomic Units (OTUs) using closed reference clustering in UCLUST (32). A set of representative sequences was generated by collapsing identical sequences and then selecting the most abundant sequence to represent that OTU. The number of sequences per sample ranged from 16,465 (E1R) to 346,285 (R6C), so all samples were rarefied to 16,465 to exclude differences due to sequencing depth prior to comparing OTU count and compositions between samples. Rarefaction at an even sequencing depth enabled a standardised approach for data analysis across all samples, a feature essential for the validation of a novel forensic technique. Any OTUs detected in the EBC extracts were removed from the experimental samples to ensure only OTUs native to the samples were included. The number of OTUs was determined from the OTU table both before and after removal of EBC OTUs.

Statistical analysis

To determine the most appropriate statistical analysis method, two different non-parametric multivariate analyses were carried out in PRIMER (PRIMER-E, Plymouth Routines in Multivariate Ecological Research v. 6, *PRIMER-E Ltd,* Luton, UK). Multivariate analysis can be either non-constrained (all samples are analysed independently) or constrained (samples are analysed base on *a priori* groupings). Non-constrained non-metric multi-dimensional scaling (nMDS) analysis treats each data point independently based on distance measures between samples, indicating within-site variation and providing a visualisation of the general relationships between samples. The level of confidence in the 2D representation of the multi-dimensional relationships is indicated by the MDS associated 'stress,' i.e. <0.2 provides good representation (33). Analysis of Similarity (ANOSIM) was used to determine significant differences between groups of soil based on the site of origin. ANOSIM reports both the level of dissimilarity between sample groups (global R) and the associated level of significance (*P*) to provide statistical pair-wise comparisons between designated groups.

In contrast to nMDS, constrained Canonical Analysis of Principal Coordinates (CAP) analysis can provide a more powerful sample comparison as differences among *a priori* groups are maximized, whilst differences within the groups are minimised (44, 45), therefore increasing the overall resolution between sites. CAP analysis also enables more robust statistics to be applied by utilising the 'leave one out' cross validation procedure as a classification analysis. The "leave-one-out" procedure removes a single sample of known origin from the dataset and then attempts to place it into the multivariate space (48). This is repeated with all data points, and the success rate of correct classification is termed the misclassification error. Following this, samples of unknown origin (evidential samples in this context) are placed into the canonical space and classified into one of the *a priori* groups by observing which group centroid was the closest. The error statistics associated with sample prediction provides an indication of how reliable the predictions were, and are supported by a visual ordination plot. For statistical analysis by both methods, a Manhattan based resemblance matrix was generated from the MIR spectral data and a Bray-Curtis based resemblance matrix was generated from the 18S rRNA data.

To explain the underlying differences between different MIR signals, one-way SIMPER (Similarity Percentages) analysis was performed using the Euclidean Distance resemblance matrix. SIMPER analysis was also used to calculate the pair-wise similarity (inverse Bray-Curtis distance) between the OTU composition of each evidence sample and each reference site.

Results

Spectral chemistry and 18S rRNA community structure of crime scenario soil samples

MIR analysis

nMDS was used to examine which sites had the most similar MIR spectra, variation within each reference site and how reference sites and evidential samples compare. The crime scene reference samples (0 m and 5 m samples) and the evidence samples formed a distinct cluster, separate from all other reference sites (Fig 2A). As a result, the spectra from the three evidence samples were statistically more similar to those collected from the crime scene (0 m and 5 m) than to any reference sample collected at random (Table S2). This discrepancy was due to the presence of carbonate peaks at 2500 cm^{-1} and 1800 cm^{-1} (34) present in the evidential samples only (Fig. S1A and S1B); this peak was absent in all other reference samples (Fig. S1C and S1D). As a result of this feature, the crime scene soil was easily distinguished from the other reference samples; however, MIR analysis may have been less successful had the crime occurred at another location.

From the nMDS plot (Fig. 2A) created from MIR spectra, the sandy soils collected from the coastal (76 km) and arid (180 km) sites showed little within-site variation; the average squared Euclidean distance between the three replicate samples was 0.01 for both sites. In contrast, the organic rich reference locations showed more within-site variation; the average square Euclidean distances were 0.02 (74 km references), 0.03 (11 km reference) and 0.05 (3.5 km reference). As a result, the arid and coastal sites formed distinct clusters; however, the resolution between the organic rich soils was poor with this traditional soil profiling technique (Fig. 2A).

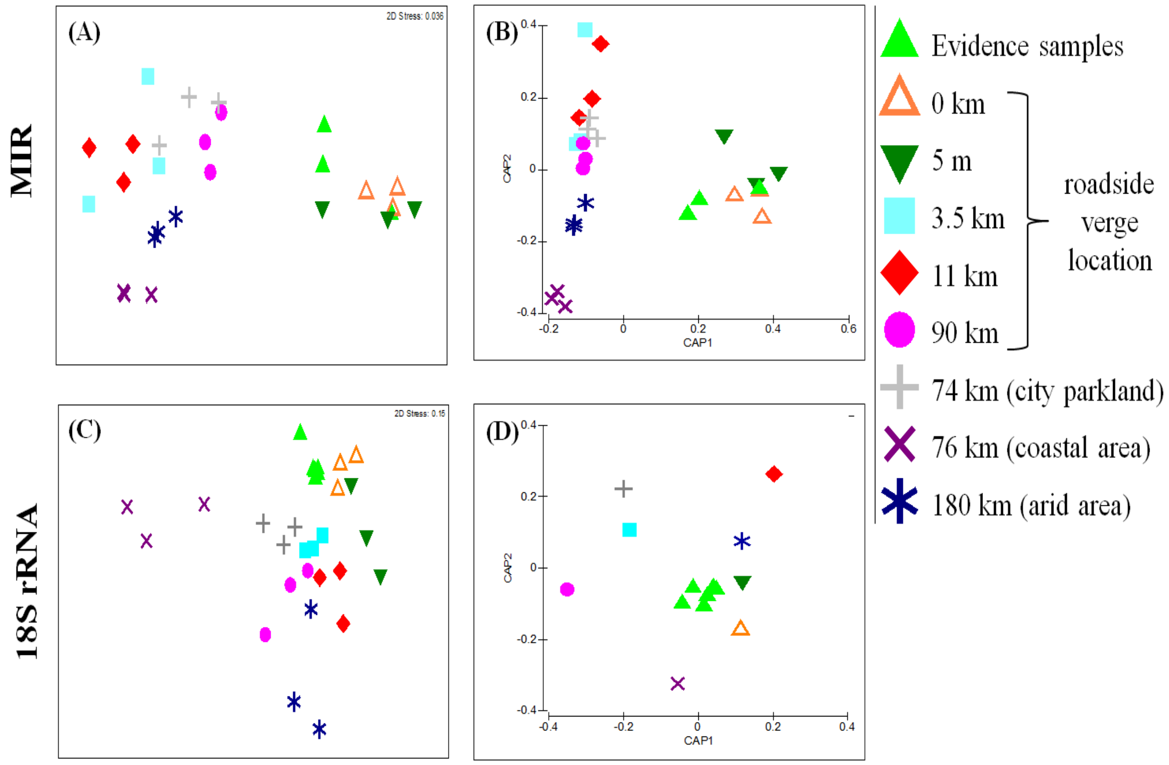


Fig. 2: MIR analysis using (A) MDS plots and (B) CAP analysis, and 18S rRNA analysis using (C) MDS plot and (D) CAP analysis.

18S rRNA analysis

HTS of the 18S rRNA was performed to compare the results of this novel technique to the more traditional MIR soil DNA profiling method. To ensure only sequences native to the samples were included in the analysis, short sequences, low quality sequences and any OTUs detected in the extraction blank controls were excluded. In total, 13,877,950 sequences were obtained from two sequencing runs, ranging from 85,576 to 745,168 sequences per sample (Table S1). A total of 10,797,393 sequences (77.8% of raw reads) were retained after length trimming. From this, 7,969,759 sequences (57.4% of raw reads) met the quality filter threshold and were used in the subsequent analysis in QIIME. 3,724,007 (26.8% of raw reads) sequences remained following closed reference OTU at 97% similarity against the SILVA database, indicating that large portions of microscopic eukaryotic soil diversity remain unresolved. This study used highly conservative closed reference OTU picking to exclude unidentified OTUs which discards any sequences that are not 97% similar to the reference sequences; however, more sequences could have been retained by applying open reference OTU picking which retains sequences that do not match to a reference database entry and clusters these unknown sequences to each other (*de novo*). Removal of the extraction blank control sequences (EBCs) decreased the number of OTUs detected in both the evidence and reference samples by 34.3 ± 7.6% (Fig. S3), highlighting the importance of monitoring background DNA levels in this type of analysis.

Forensic soil DNA analysis relies on evidential samples having a more similar diversity to samples from the crime scene, than to any other samples collected at random. The diversity captured by 18S rRNA sequencing was explored to determine both the proportion of eukaryotic DNA present and variation in taxa detected. On average, 92.9 ± 6.5% of the diversity per sample was identified as eukaryotic, while the remainder was attributed to bacterial DNA (Fig. S4A). Some eukaryotic groups were common across all sample locations (Fig. S4B), namely metazoan (25.4 ± 18.0%) and fungi (18.0 ± 12.6%). Some fungal phyla were consistently detected across all locations (Fig. S4C), e.g. Ascomycota (49.6 ± 30.0%), Basidiomycota (44.9 ± 28.1%), and Chytridiomycota (5.0 ±4.6%). However, no metazoan phyla were detected across all locations; instead, different sample locations were dominated by Arthropoda, Nematoda, Rotifera or Porifera (Fig. S4D). This range and distribution of shared and unique eukaryotic taxa detected across the different sample locations indicates the extensive eukaryote variation detected, even at high taxonomic level, and thus highlights the potential of non-bacterial DNA for use in soil discrimination.

Within-site variation of the18S rRNA profiles differed from the within-site variation observed using MIR analysis. The sandy soils from the coastal and arid locations showed similar within-site variation (BC distance 0.62 and 0.55, respectively) to the organic rich locations: BC distances of triplicate samples within a site were 0.48 (3.5 km), 0.63 (11 km), 0.62 (90 km), and 0.66 (74 km). Pair-wise comparisons of Bray-Curtis Distance showed that the shoe samples and shovel samples were significantly more similar to the location of the body (0 m) than to any of the other reference sites (Fig. 3 and Table S3). Although the car boot sample showed a similar trend, this sample was least similar to the location of the body than was the shoe or shovel samples (Fig. 3), and represented a single sample case scenario for which no statistical analysis, i.e. ANOVA, was possible. ANOSIM analysis demonstrated that the OTU composition of the evidential samples was significantly different from four of the eight reference groups (Table S2), and there was no distinct cluster observed for the crime scene samples in the nMDS plot using DNA analysis (Fig. 2C). 18S rRNA profiles were more variable within a given are than MIR spectra and the resolution between sites using nMDS analysis was limited. Although this analysis narrowed the possible origin of the evidence samples, it did not establish a specific link to the crime scene. This suggests that nMDS of 18S rRNA should be used for investigative stages of a case, rather than evidential stages.

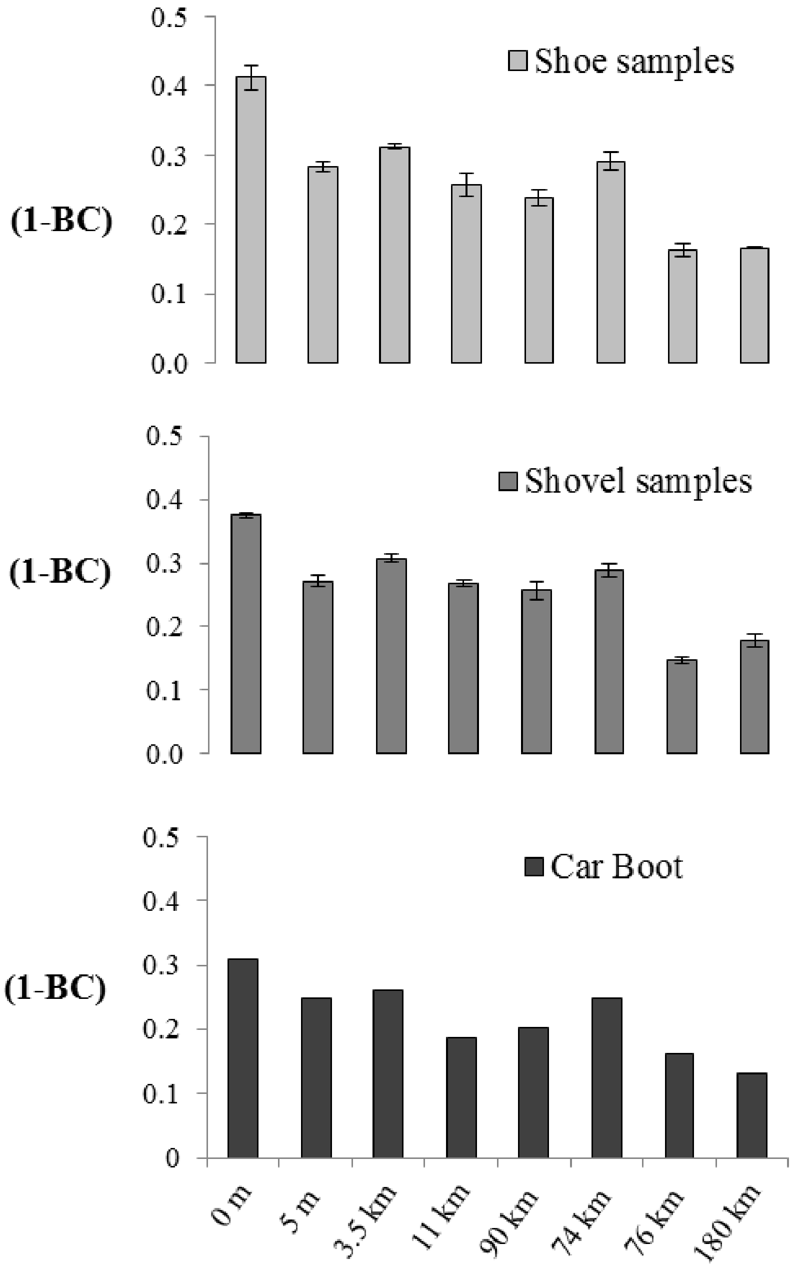


Fig. 3: OTU compositional similarity between each reference location and (A) the shoe samples (*n=2*), (B) the shovel samples (*n=3*), and (C) the car sample (*n=1*).

Predicted origin of evidence samples using CAP analysis

MIR analysis

Using MIR analysis, the specific origin of the different evidential samples within the crime scene differed depending on the item that the soil was recovered from (Table 2). For example, the predicted origin of the shoe samples was the exact location of the body (0 m), whereas the shovel sample was linked more generally to the crime scene. This suggests that MIR would be useful for identifying the general locality of an unknown sample; however, this method may not be useful for indicating the exact origin. To identify the error associated with this traditional soil analysis method 'leave-one out' cross validation was applied. In this instance, the misclassification error associated with MIR analysis was 29.2%, i.e. only 17/24 reference samples were correctly classified to the correct site (Table S4). This was due to spectral variation within some reference sites and poor resolution between the organic rich soils. Poor resolution between the reference sites will increase the error rate using the leave-one out cross validation approach and thus weaken the strength of the evidence. Although all three evidence samples were successfully predicted to have originated from the crime scene in this study, this MIR analysis may have been less informative had the body been recovered at one of the four reference sites with organic rich soil.

Table 2: Predicted origin of unknown samples using CAP analysis**.**

		Predicted origin from CAP analysis	
Unknowns	Evidence sample	**MIR**	**18S**
E1R	Right Shoe	0 m	0 m
E1L	Left Shoe	0m	0 m
E2A	Shovel A	5 m	0 m
E2B	Shovel B	n/a	0 m
E2C	Shovel C	n/a	0 m
E3	Car Boot	n/a	0 m

18S rRNA analysis

In contrast to the MIR data analysis, CAP analysis of the 18S rRNA predicted the origin of all six evidence samples to be the specific location of the body (Table 2). This finding illustrates the power of constrained analysis (CAP) to extract unique site specific signals that were masked by ubiquitous OTUs in the nMDS analysis. In addition, the misclassification error was only 9.4%, i.e. 29/32 reference samples were correctly classified (Table S4). Although the organic rich reference sites showed some within-site DNA profile variation in the nMDS, CAP analysis allowed each of the organic soil type locations to be distinguished (Fig. 2D). Overall, within-site variation of all reference samples was reduced with CAP analysis compared to MDS analysis. However, evidential samples still showed some degree of variation, possibly due to transfer effects and storage during the six weeks of the study (Fig. 2D).

Effect of air-drying reference samples prior to DNA extraction

During the six weeks between the two sampling events, the evidential samples had dried out, and the reference samples were subject to temporal variation during the course of the investigation (Fig. 1B). Therefore, we examined the effect of air-drying reference samples on the discriminatory power of the DNA tests. Air-drying the samples prior to DNA extraction had no significant effect on the number of OTUs detected, either without drying, 309 ± 84 OTUs (mean ± SD) or with drying 274.5 ± 81 OTUs (mean ± SD) (One-way ANOVA, $F_{(1,14)}$=0.710, p= 0.414). Bray-Curtis cluster dendrograms generated with and without air-drying before DNA extraction (Fig. S5) show differences in the relative positioning of two samples collected at the crime scene (5 m A and B); however, in both cases the evidence samples were consistently more similar to the reference samples collected at the location of the body (0 m) than all other reference samples. Furthermore, the BC distance between the evidence samples and the references from the location of the body (0 m) was consistent with and without air-drying (Table 3). This indicates that temporal changes over the course of the investigation did not alter the DNA profile sufficiently to prevent discrimination, and although air-drying reference samples more accurately reflects the desiccation of the evidence samples, it does not improve the similarity of the DNA profiles.

Table 3: The effect of air-drying prior to DNA extraction on similarity between the evidence samples and each reference site. Similarity was measured using the inverse of the Bray-Curtis distance (1-BC).

Reference Location	Without air-drying	With air-drying	Both with and without air-drying
0 m	0.38	0.38	0.38
5 m	0.28	0.26	0.27
3.5 km	0.34	0.29	0.31
11 km	0.20	0.26	0.23

Discussion

This experimental case scenario demonstrates the robustness of this novel forensic technique in comparison to a traditional soil analysis method. The work presented in this study analysed realistic samples recovered from evidential items and incorporated several practical issues, including dessication, seasonal and temporal varation and transfer effects, which would all be likely to be encountered during forensic case work. The comparison of soil DNA profiles to a traditional soil analysis is also an improtant validation step for implimenting new methods into forensic science. Although such a comparison has previously been reported using T-RFLP and ICP-MS by Concheri *et al.* (2), the current study is the first to examine the potential of HTS in a forensic context. In addition, this study utilizes novel analytical methodologies to interpret results, expanding upon previous studies that only analyzed significant differences in Bray-Curtis distance values (35-37). Overall, this study demonstrates the feasible application of HTS in forensic soil analysis and provides prediction statistics as a means to convey the effectiveness and robustness of these methodologies in the court room.

This study demonstrates that DNA profiling using HTS is robust to at least some of the potential limitations commonly associated with casework. Although previous studies describe seasonal fluctuations in soil communities (38, 39), our results demonstrate that the origin of ‘unknown’ evidential samples was correctly predicted with reasonable confidence (i.e. 9.4% error statistic), despite a time lapse of six weeks that included variations in temperature and rainfall. This suggests that variation in eukaryotic diversity as a result of environmental conditions in this scenario was less than the variation between sites, which is a promising result for forensic applications. We also demonstrate that removal of soil from the environment at the time of the crime, and desiccation of soil in a car boot prior to DNA extraction, did not obscure the genetic signal or prevent identification to the crime scene samples. This result is not surprising, as fungal T-RFLP profiles are robust to drying and air-drying samples is a method commonly applied to stabilise soil properties (18). Furthermore, soil particles that had adhered to the shoes and the shovel prior to the crime did not interfere with the biological signal in this scenario. This could be attributed to the sampling, the loss of particles adhered prior to the crime, as well as sequential transfer effects, suggesting that the last soil transferred to the items dominated the genetic signal. However, DNA profiles generated from the evidential samples were more variable than those from the reference sites, which could have been due to subtle temporal, transfer or mixing effects associated with evidential samples. Regardless, soil DNA profiling was successfully used in this study to predict the origin of the soil for three different types of evidential samples. However, it should be noted that the presence of human decomposition can alter soil biota (40-43), and the presence of decay-associated eukaryotic organisms might influence the profile obtained in evidential samples, hindering or assisting in their identification. HTS of reference samples collected from real crime scenes may need to include examination of material several meters from the body to obtain a DNA profile representative of the area prior to human decomposition.

Although both MIR and DNA analysis linked the evidence samples to the crime scene, MIR showed a lower misclassification error rate. Using MIR analysis, all crime scene samples formed a distinct cluster separate from all other reference samples due to the presence of carbonates in all evidential and crime scene samples (0 m and 5 m); this characteristic was absent in the spectra of all other reference samples. However, using this statistical approach, MIR is likely to have been problematic if the crime scene had been located at one of the four organic rich sites, as this method failed to distinguish between these reference locations. The organic soil samples from roadside verges at increased distances from the body (3.5 km, 11 km and 90 km) and the organic soil collected from the

Parkland sample (74 km) all clustered together in both MDS and CAP analysis of MIR data. This lack of resolution between the reference samples increased the misclassification error associated with MIR analysis, potentially inflating the apparent power of the tests. Although statistical analysis of MIR data could be used to complement DNA based approaches, the visual interpretation of MIR spectra by experts with years of training and knowledge could improve resolution based on soil mineralogy.

In contrast, 18S rRNA analysis resulted in clear separation of all reference samples and generated local scale resolution. The misclassification error was reduced when DNA was analysed, as the three organic reference soils from roadside verge locations (3.5 km, 11 km, 90 km) and the organic rich parkland samples (74 km) could all be distinguished using CAP analysis. This result demonstrates the potential of soil DNA to increase resolution between samples where traditional methods, such as MIR analysis, may have failed to distinguish sites. In addition, the predicted origin of all evidential samples was the location of the body, i.e. 0 m sample, rather than the general locality of the crime scene (5 m). This indicates that although the crime scene samples (0 m and 5 m) showed similar mineralogy, each could be specifically differentiated based on eukaryote diversity, demonstrating that eukaryotic soil diversity contains unique signals that enable fine scale resolution. Nevertheless, both MIR and DNA analysis could be used to successfully predict the correct origin of example evidential soil samples in this study, despite a lag-time of six weeks between the ‘criminal activity’ and the collection of reference samples.

In forensic science, the probability that a sample originated from one source, rather than another selected at random, must be evaluated. Human DNA profiling statistics are provided as the Random Match Probability or as a Likelihood Ratio (LR). However, soil analysis significantly differs from human identification, as soil is not a discrete entity and the soil community is vulnerable to influences of both temporal and spatial variation. As Murray and Tedrow (28) state, no two physical objects can ever be the same in a theoretical sense; similarly two soil samples cannot be said in an absolute sense to have originated from a single source. As a result, it is only possible to establish a degree of probability regarding whether or not a sample was derived from a given location. This study demonstrates the use of two different multivariate analytical methods, and highlights the use of each in forensic case work.

In this study, we applied ANOSIM statistical analysis from nMDS, and prediction probabilities from CAP analysis as a means to convey weight of soil evidence depending on the information available to investigators i.e. investigative or evidential phase of a case. Non-constrained nMDS analysis treats each data point independently based on distance measures between samples, indicating within-site variation. Assuming that samples from similar geographical locations or habitat types have similar signals, this type of analysis could be useful for investigative purposes to indicate the likely origin or soil type of an unknown sample. However, if no distinct feature separates the different groupings, resolution between different reference sites can be difficult using nMDS, as observed for 18S rRNA. Such limitation would become problematic for cases requiring evidential value, when a link between samples must be statistically supported. In addition, statistical analysis using ANOSIM is particularly problematic for an evidential sample when no replication is possible. However, samples that show very little similarity in the nMDS can be eliminated from further consideration, assisting in investigative forensic approaches. For example, it was clear that the coastal samples (180 km) had very little similarity to the crime scene samples using both MIR and 18S rRNA analysis. We would suggest projection of an unknown DNA profile onto an nMDS plot obtained from other general localities, and evaluate the similarities between locations using ANOSIM statistics when no conditional statement is known to a case; this approach describes an investigative stage of a case

We show that constrained CAP analysis can provide a more powerful, prediction of the possible origin of an unknown sample given the characteristics of a range of known soil sources; this describes an evidential stage of a case where there is a conditional statement and all references are known to investigators, e.g. alibi sites; this approach describes an evidential stage of a case. CAP analysis maximizes differences among *a priori* groups, whilst minimising differences within the groups (45, 46), therefore increasing the overall resolution between sites. The use of *a priori* groups implies that all the relevant reference groups must be known to the investigation, but the analysis can also be easily updated if new locations or information becomes relevant later in an investigation. CAP analysis also enables a misclassification error to be calculated (47), thus providing a measure of confidence. This is particularly useful for cases with no replicate evidential samples available for standard statistical analysis, as illustrated by the car boot sample in this study. In casework with solid background information, CAP analysis can be used to either include or exclude a suspect from an investigation. Although this experimental case study was successful in predicting the origin of evidential samples, the strength of this result is heavily reliant on the density of reference samples used.

In this study, two independent analytical methods predicted that the evidential soil samples were more similar to those collected at the Quarry Road crime scene in Tooperang, rather than any other location sampled tested. Given that the suspect claimed to have been with the victim on the evening of her disappearance, yet claimed never to have been to this location, soil analysis suggests the suspect was present at the crime scene. The positive DNA profile match linking the suspect to the victim, and now soil DNA and MIR evidence that strongly suggest his presence at the crime scene would strengthen the evidence against the accused and assist in the prosecution of this individual for the murder of the young woman.

Acknowledgements

This work was supported by the Australian Research Council, and the Australian Federal Police. We thank Natasha Mitchell from Forensic Science South Australia (FSSA) for advice on experimental design and members of the Australian Centre for Ancient DNA for helpful comments on the manuscript.

References

1. **Fitzpatrick, R. W., M. D. Raven, and S. T. Forrester.** 2009. A systematic approach to soil forensics: criminal case studies involving transference from crime scene to forensic evidence. Criminal and Environmental Soil Forensics**:**105.
2. **Concheri, G.** 2011. Chemical elemental distribution and soil DNA fingerprints provide the critical evidence in murder case investigation. PloS ONE **6:**e20222.
3. **Muccio, Z., and G. P. Jackson.** 2009. Isotope ratio mass spectrometry. Analyst **134:**213-222.
4. **Walker, G. S.** 2009. Analysis of Soils in a Forensic Context: Comparison of Some Current and Future Options, p. 397-409, Criminal and Environmental Soil Forensics. Springer.
5. **Fitzpatrick, R. W.** 2009. Soil: Forensic Analysis. Wiley Encyclopedia of Forensic Science**:**2377-2388.
6. **Dawson, L. A., and S. Hillier.** 2010. Measurement of soil characteristics for forensic applications. Surface and Interface Analysis **42:**363-377.
7. **Fitzpatrick, R. W.** 2003. Demands on soil classification in Australia. Soil Classification: a Global Desk Reference**:**77-100.
8. **Sensabaugh, G. F.** 2009. Microbial Community Profiling for the Characterisation of Soil Evidence: Forensic Considerations. Springer, Netherlands, 2009. 49-60.
9. **Macdonald, C. A.** 2011. Discrimination of soils at regional and local levels using bacterial and fungal t-RFLP profiling. Journal of Forensic Sciences **56:**61-69.
10. **Young, J. M., L. S. Weyrich, and A. Cooper.** 2014. Forensic soil DNA analysis using high-throughput sequencing: a comparison of four molecular markers. Forensic Science International: Genetics **In Press**.
11. **Fierer, N., C. L. Lauber, N. Zhou, D. McDonald, E. K. Costello, and R. Knight.** 2010. Forensic identification using skin bacterial communities. Proceedings of the National Academy of Sciences **107:**6477-6481.
12. **Kennedy, D. M., J.-A. L. Stanton, J. A. García, C. Mason, C. J. Rand, J. A. Kieser, and G. R. Tompkins.** 2012. Microbial Analysis of Bite Marks by Sequence Comparison of Streptococcal DNA. PLoS ONE **7:**e51757.
13. **Khodakova, A. S., R. J. Smith, L. Burgoyne, D. Abarno, and A. Linacre.** 2014. Random Whole Metagenomic Sequencing for Forensic Discrimination of Soils. PLoS ONE **9:**e104996.

14. **Giampaoli, S., A. Berti, R. Di Maggio, E. Pilli, A. Valentini, F. Valeriani, G. Gianfranceschi, F. Barni, L. Ripani, and V. Romano Spica.** 2014. The environmental biological signature: NGS profiling for forensic comparison of soils. Forensic science international **240:**41-47.

15. **Young, J. M., N. J. Rawlence, L. S. Weyrich, and A. Cooper.** 2014. Limitations and recommendations for successful DNA extraction from forensic soil samples: A review. Science & Justice **54:**238-244.

16. **Lerner, A., Y. Shor, A. Vinokurov, Y. Okon, and E. Jurkevitch.** 2006. Can denaturing gradient gel electrophoresis (DGGE) analysis of amplified 16S rDNA of soil bacterial populations be used in forensic investigations? Soil Biology & Biochemistry **38:**118-1192.

17. **Dequiedt, S., N. P. A. Saby, M. Lelievre, C. Jolivet, J. Thioulouse, B. Toutain, D. Arrouays, A. Bispo, P. Lemanceau, and L. Ranjard.** 2011. Biogeographical patterns of soil molecular microbial biomass as influenced by soil characteristics and management. Glob. Ecol. Biogeogr. **20:**641-652.

18. **Gelsomino, A., and A. Azzellino.** 2011. Multivariate analysis of soils: microbial biomass, metabolic activity, and bacterial-community structure and their relationships with soil depth and type. J. Plant Nutr. Soil Sci. **174:**381-394.

19. **Johnson, M. J., K. Y. Lee, and K. M. Scow.** 2003. DNA fingerprinting reveals links among agricultural crops, soil properties, and the composition of soil microbial communities. Geoderma **114:**279-303.

20. **Macdonald, C. A., R. Ang, S. J. Cordiner, and J. Horswell.** 2011. Discrimination of Soils at Regional and Local Levels Using Bacterial and Fungal T-RFLP Profiling. Journal of Forensic Sciences **56:**61-69.

21. **Macdonald, L. M., B. K. Singh, N. Thomas, M. J. Brewer, C. D. Campbell, and L. A. Dawson.** 2008. Microbial DNA profiling by multiplex terminal restriction fragment length polymorphism for forensic comparison of soil and the influence of sample condition. J. Appl Microbiol **105:**813-821.

22. **Kasel, S., L. T. Bennett, and J. Tibbits.** 2008. Land use influences soil fungal community composition across central Victoria, south-eastern Australia. Soil Biol. Biochem. **40:**1724-1732.

23. **Lenz, E. J., and D. R. Foran.** 2010. Bacterial profiling of soil using genus-specific markers and multidimensional scaling. Journal of Forensic Sciences **55:**1437-1442.

24. **Darby, B. J., D. A. Neher, D. C. Housman, and J. Belnap.** 2011. Few apparent short-term effects of elevated soil temperature and increased frequency of summer

precipitation on the abundance and taxonomic diversity of desert soil micro- and meso-fauna. Soil Biology & Biochemistry **43:**1474-1481.

25. **Bull, P. A., A. Parker, and R. M. Morgan.** 2006. The forensic analysis of soils and sediment taken from the cast of a footprint. Forensic Science International **162:**6-12.

26. **Pye, K., and S. J. Blott.** 2004. Particle size analysis of sediments, soils and related particulate materials for forensic purposes using laser granulometry. Forensic Science International **144:**19-27.

27. **Chazottes, V., C. Brocard, and B. Peyrot.** 2004. Particle size analysis of soils under simulated scene of crime conditions: the interest of multivariate analyses. Forensic Science International **140:**159-166.

28. **Murray, R. C., and J. C. Tedrow.** 1991. Forensic geology.

Supplementary Material

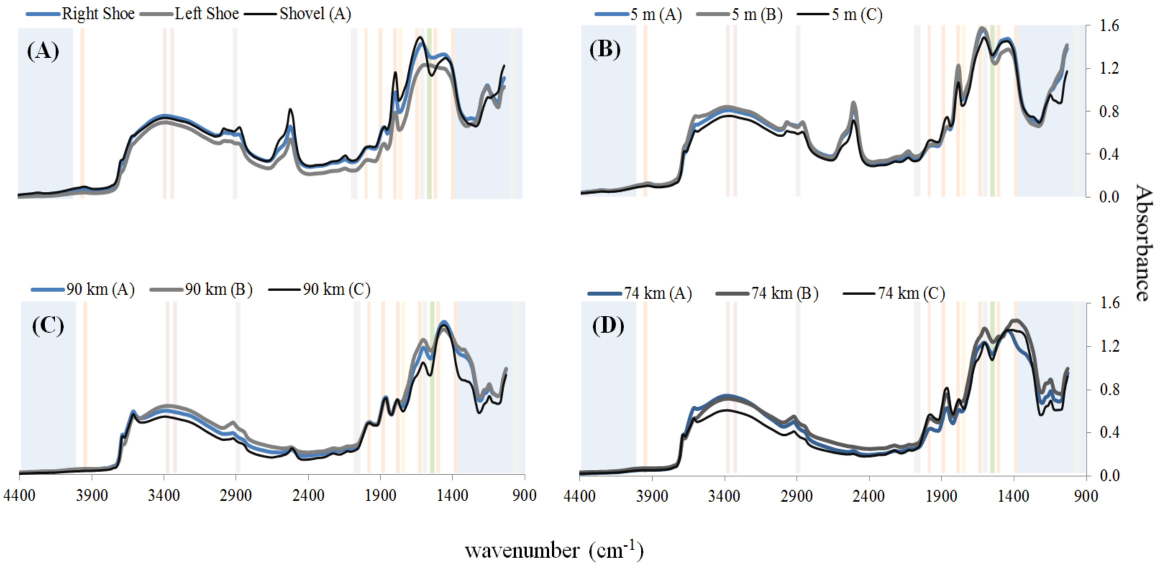


Fig. S1: MIR spectra showing the presence or absence of carbonate peaks in (A) the evidence samples, (B) Quarry road samples (5 m). (C) Barker Road samples (90 km), and (D) Pioneer Woman's Memorial Garden samples (74 km). The reference samples at 90 km and 74 km were used in this comparison as the three replicates within each site showed little variation.

Table S1: 18S sample processing. Raw de-multiplexed sequence reads from Illumina MiSeq sequencing were processed to remove adapters, poor quality and small read lengths that effect downstream processing. Values in parenthesis indicate the percentage of cleaned/raw reads.

Sample	18S index	De-multiplex reads	Adapter trimming	Primer, barcode and length trimming	After Quality Trimming	After closed ref OTU clustering at 97%	Number of OTUs detected	% OTUs retained following EBC filtering
EB1	TCGTGATGTGAC	85576	85576	60653 (71%)	45291 (53%)	35366 (41%)	51	n/a
Right shoe	TAGCCGGCATAG	119187	119187	84296 (71%)	64092 (54%)	16465 (14%)	431	73.8
Left shoe	GTGGAACCACGT	159078	159078	110066 (69%)	89398 (56%)	26683 (17%)	475	74.9
Shovel A	TATTACCGGCAT	109895	109895	82495 (75%)	63065 (57%)	20724 (19%)	404	74.5
Shovel B	GATCCGACACTA	160556	160556	127878 (80%)	101329 (63%)	28177 (18%)	413	75.2
Shovel C	GTGCCATAACCA	181796	181796	135546 (75%)	108188 (60%)	24933 (14%)	491	77.8
Car boot	GTGTTTGGTCGA	159156	159156	101703 (64%)	75743 (48%)	19838 (12%)	329	71.7
EB2	GAAGGAAGCAGG	92161	92161	79116 (86%)	59271 (64%)	40159 (44%)	43	n/a
XA	AGAGAAATGTCG	136347	136347	101922 (75%)	83303 (61%)	33185 (24%)	348	71.8
XB	CTCCCATACCAC	236340	236340	169966 (72%)	138664 (59%)	35877 (15%)	404	73
XC	TACAAACCCTGT	278905	278905	204990 (73%)	166989 (60%)	41829 (15%)	363	73.9
R1A	GTAGAGCTGTTC	291111	291111	210698 (72%)	159003 (55%)	107200 (37%)	191	62.2
R1B	GGAAAGTCGAAG	236586	236586	113638 (48%)	92783 (39%)	28991 (12%)	319	66.7
R1C	TCCACAGGAGTT	214452	214452	171462 (80%)	139341 (65%)	44779 (21%)	345	70.5
R2A	TCATCGCGATAT	149404	149404	119830 (80%)	96476 (65%)	32376 (22%)	338	71.7
R2B	GACTTATAGGCC	169574	169574	141193 (83%)	114564 (68%)	36642 (22%)	320	65.6
R2C	ACGCAACTGCTA	179842	179842	151265 (84%)	119707 (67%)	49126 (27%)	227	62.1
R3A	ACCCTGTACCCT	121994	121994	101847 (83%)	79294 (65%)	48320 (40%)	167	57.9
R3B	GCCATAGGTTTG	212860	212860	174016 (83%)	143548 (67%)	48154 (23%)	232	63.5
R3C	CAGTAACGGCCA	145920	145920	78272 (54%)	63640 (44%)	19938 (14%)	343	72.7
EB3	GGTTCTTATGAC	106616	106616	75946 (71%)	60842 (57%)	25981 (24%)	50	n/a
EB4	TCGTGATGTGAC	297194	297194	224543 (76%)	162159 (55%)	87633 (29%)	73	n/a

Table S1 (continued)

Sample	18S index	De-multiplex reads	Adapter trimming	Primer, barcode and length trimming	After Quality Trimming	After closed ref OTU clustering at 97%	Number of OTUs detected	% OTUs retained following EBC filtering
R4A	TAGCCGGCATAG	328099	328099	298604 (91%)	218131 (66%)	150369 (46%)	127	52.9
R4B	GTGGAACCACGT	333198	333198	282715 (85%)	205577 (62%)	67782 (20%)	260	63.9
R4C	TATTACCGGCAT	398737	398737	357072 (90%)	254437 (64%)	89742 (23%)	193	64.3
R5A	GATCCGACACTA	444636	444636	357208 (80%)	267277 (60%)	94197 (21%)	260	65.3
R5B	GTGCCATAACCA	399116	399116	331809 (83%)	223012 (56%)	77194 (19%)	208	64
R5C	GTGTTTGGTCGA	461300	461300	395527 (86%)	291153 (63%)	146507 (32%)	240	65.8
R6A	GAAGGAAGCAGG	494971	494971	328898 (66%)	240268 (49%)	90239 (18%)	162	63.5
R6B	AGAGAAATGTCG	300923	300923	248592 (83%)	180805 (60%)	77909 (26%)	85	51.4
R6C	CTCCCATACCAC	579982	579982	557391 (96%)	409619 (71%)	346285 (60%)	92	56.4
R7A	TACAAACCCTGT	575094	575094	502588 (87%)	363333 (63%)	271666 (47%)	111	48.1
R7B	GTAGAGCTGTTC	676807	676807	426361 (63%)	322697 (48%)	211030 (31%)	109	53.2
R7C	GGAAAGTCGAAG	352133	352133	306955 (87%)	226371 (64%)	115358 (33%)	177	58.2
EB5	TCCACAGGAGTT	345230	345230	234112 (68%)	160138 (46%)	123466 (36%)	59	n/a
XA_D	TCATCGCGATAT	401784	401784	355305 (88%)	253809 (63%)	130767 (33%)	230	66.7
XB_D	GACTTATAGGCC	424440	424440	317020 (75%)	235815	62650 (15%)	391	73.4
XC_D	ACGCAACTGCTA	447443	447443	336446 (75%)	241943	87238 (19%)	361	73.4
R1A_D	ACCCTGTACCCT	745168	745168	584095 (78%)	400859	270925 (36%)	187	59.2
R1B_D	GCCATAGGTTTG	437970	437970	348639 (80%)	228533	165066 (38%)	155	55.5
R1C_D	CAGTAACGGCCA	421033	421033	367109 (87%)	269623	119134 (28%)	278	65.6
R2A_D	GGTTCTTATGAC	466248	466248	400877 (86%)	289778	88987 (19%)	304	66.5
R3A_D	CTACGACCATTA	641739	641739	405142 (63%)	292964	97394 (15%)	290	68.4
EB6	AGATTACCGGCG	357349	357349	233587 (65%)	166927	51780 (14%)	40	n/a

Table S2: Pair-wise ANOSIM statistics between sample groups.

MIR statistics	Evidence	0 m	5 m	3.5 km	11 km	90 km	74 km	76 km
0 m	0.074							
5 m	0.148	0.037						
3.5 km	1	1	1					
11 km	1	1	1	-0.037				
90 km	1	1	1	0.556	1			
74 km	1	1	1	0.148	0.704	0.444		
76 km	1	1	1	0.778	1	1	1	
180 km	1	1	1	0.407	0.889	1	0.926	1

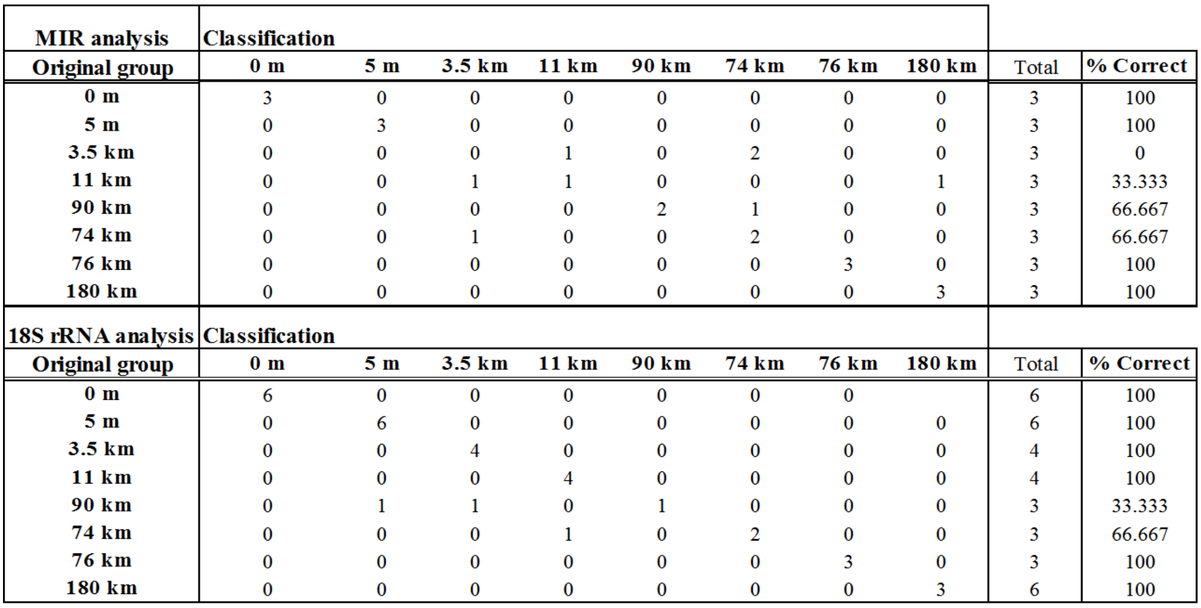
18S rRNA analysis	Evidence	0 m	5 m	3.5 km	11 km	90 km	74 km	76 km
0 m	0.924							
5 m	0.665	0.600						
3.5 km	1	1	0.234					
11 km	1	0.992	0.603	0.802				
90 km	0.963	0.981	0.481	0.407	0.741			
74 km	0.932	0.938	0.735	0.704	0.796	0.704		
76 km	1	1	1	1	1	1	1	
180 km	1	1	0.765	0.963	0.833	0.667	0.852	1

Table S3: The statistics comparing the similarity of the 18S rRNA profile of each reference site to (A) the shoe samples, and (B) the shovel samples *One-way ANOVA and Poc-Hos test results (LSD).*

Source of evidence samples	(I) Origin	(J) Origin	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Lower Bound
Shoes	0 m	5 m	.12780*	.01163	.000	.1010	.1546
		3.5 km	.10020*	.01163	.000	.0734	.1270
		11 km	.15555*	.01163	.000	.1287	.1824
		90 km	.17425*	.01163	.000	.1474	.2011
		74 km	.12125*	.01163	.000	.0944	.1481
		76 km	.24900*	.01163	.000	.2222	.2758
		180 km	.24620*	.01163	.000	.2194	.2730
Shovel	0 m	5 m	.10347*	.00709	.000	.0884	.1185
		3.5 km	.06777*	.00709	.000	.0527	.0828
		11 km	.10640*	.00709	.000	.0914	.1214
		90 km	.11830*	.00709	.000	.1033	.1333
		74 km	.08633*	.00709	.000	.0713	.1014
		76 km	.22683*	.00709	.000	.2118	.2419
		180 km	.19740*	.00709	.000	.1824	.2124

*The mean difference is significant at the 0.05 level.

Table S4: Allocation of observations to groups using CAP analysis and leave-one-out cross validation.



MIR analysis	Classification									
Original group	0 m	5 m	3.5 km	11 km	90 km	74 km	76 km	180 km	Total	% Correct
0 m	3	0	0	0	0	0	0	0	3	100
5 m	0	3	0	0	0	0	0	0	3	100
3.5 km	0	0	0	1	0	2	0	0	3	0
11 km	0	0	1	1	0	0	0	1	3	33.333
90 km	0	0	0	0	2	1	0	0	3	66.667
74 km	0	0	1	0	0	2	0	0	3	66.667
76 km	0	0	0	0	0	0	3	0	3	100
180 km	0	0	0	0	0	0	0	3	3	100
18S rRNA analysis	Classification									
Original group	0 m	5 m	3.5 km	11 km	90 km	74 km	76 km	180 km	Total	% Correct
0 m	6	0	0	0	0	0	0		6	100
5 m	0	6	0	0	0	0	0	0	6	100
3.5 km	0	0	4	0	0	0	0	0	4	100
11 km	0	0	0	4	0	0	0	0	4	100
90 km	0	1	1	0	1	0	0	0	3	33.333
74 km	0	0	0	1	0	2	0	0	3	66.667
76 km	0	0	0	0	0	0	3	0	3	100
180 km	0	0	0	0	0	0	0	3	6	100

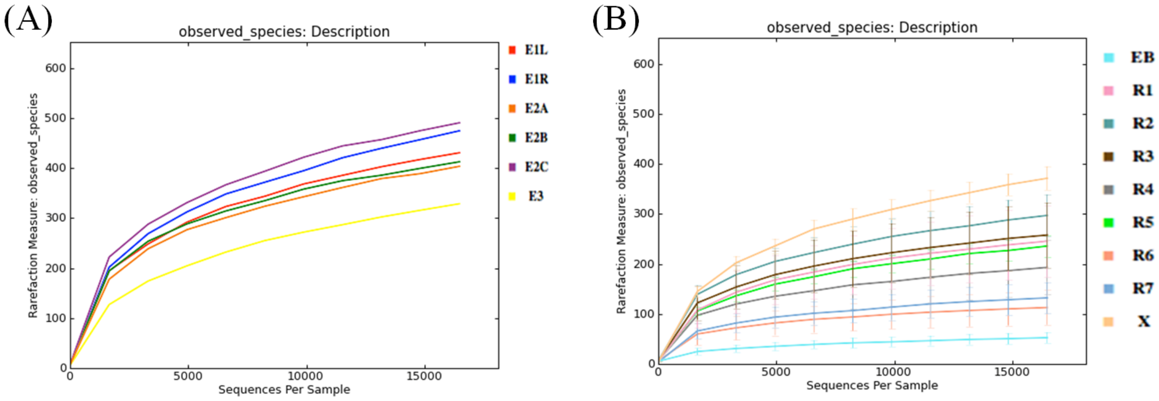


Fig. S2: Rarefaction curves prior to OTU filtering of (A) evidence samples and (B) reference samples.

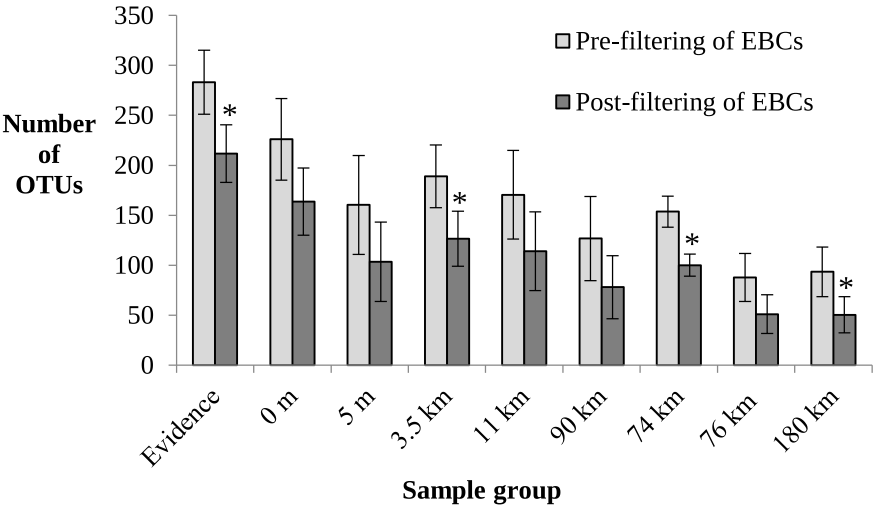


Fig. S3: The mean number of OTUs in the evidence samples (*n*=6) and at each reference site (*n*=3) pre and post filtering of extraction blank controls (EBCs). Error bars represent the standard deviation. * indicates a significant decrease in OTU count post-filtering of EBCs.

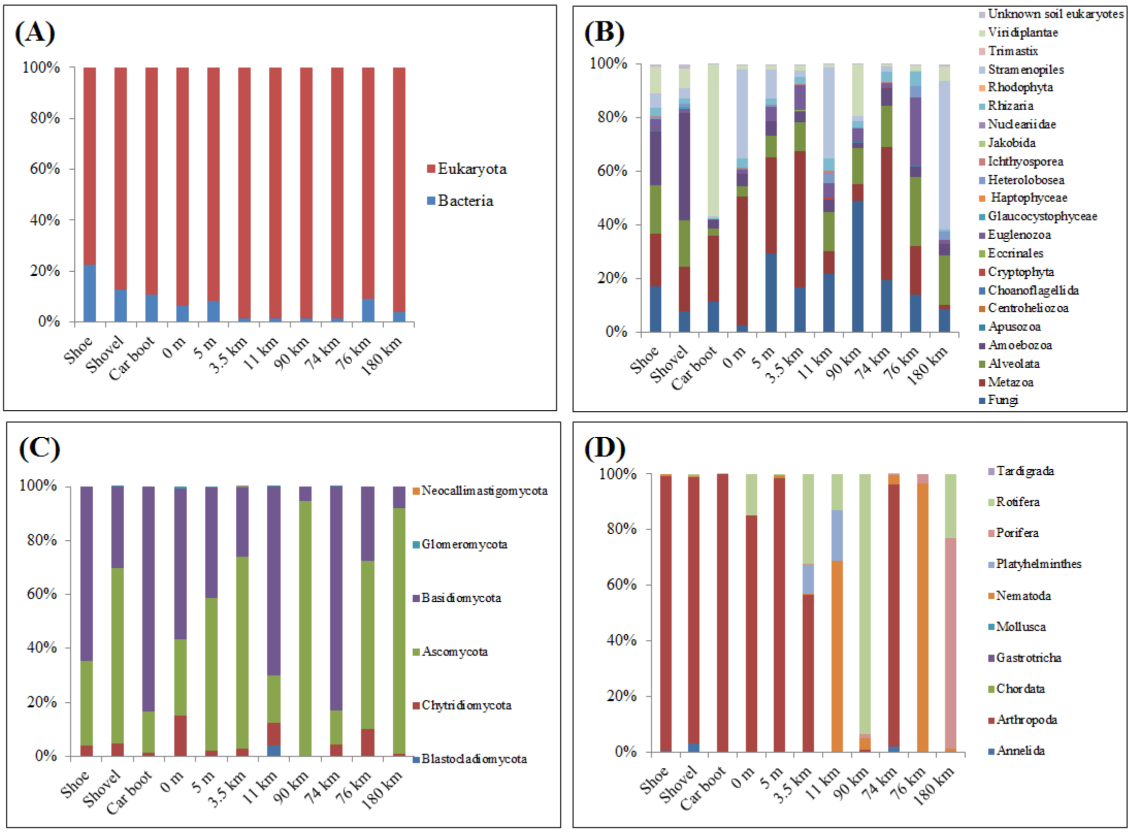
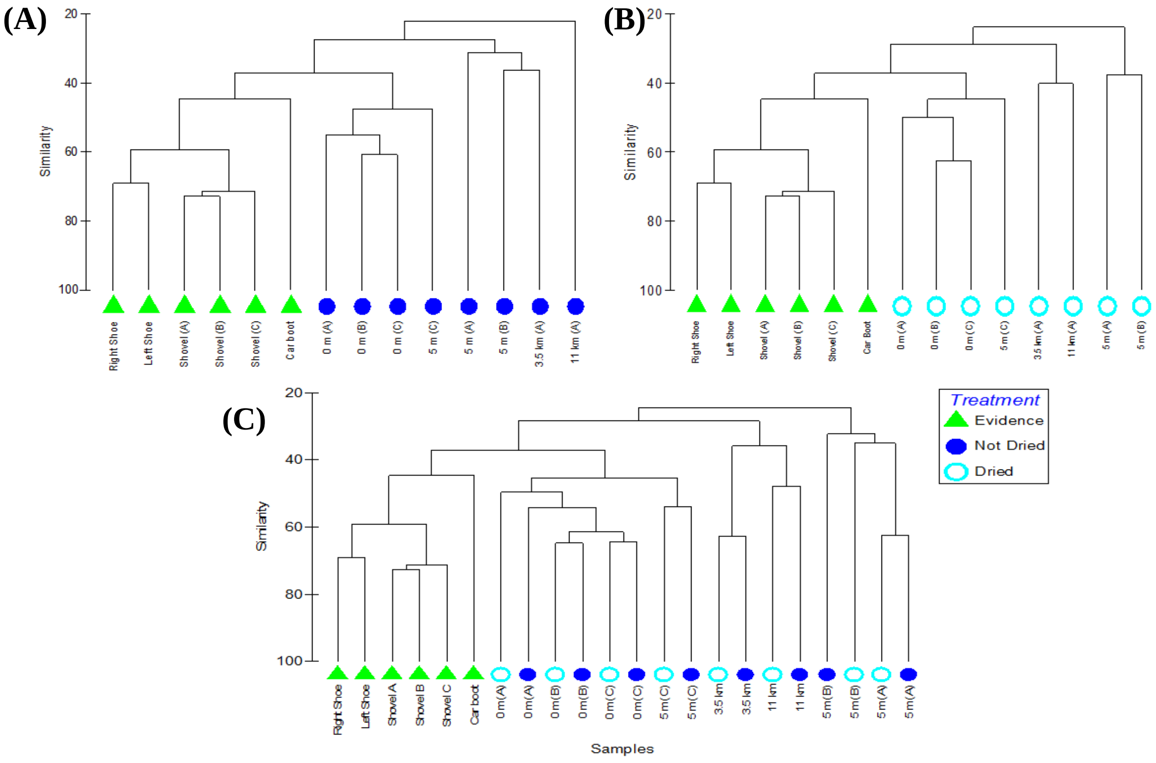


Fig. S4: Taxonomy detected using the 18S rRNA analysis: (A) Domain, (B) eukaryotes only, (C) fungi only, and (D) Metazoa only.



. **Fig. S5: Bray-Curtis Cluster Dendrograms of evidence samples and reference samples** (A) without air-drying prior to DNA extraction, (B) with air-drying prior to DNA extraction and (C) both with and without air-drying prior to DNA extraction

CHAPTER 7

General discussion and concluding remarks

General discussion and concluding remarks

DNA analysis of soil communities has been successfully presented in forensic court cases to establish a link between a suspect and a site, victim or object. However, currently these methodologies have limited resolution, and focus mainly on bacterial biota within soils. This thesis demonstrates the clear potential of HTS technology to discriminate between forensic soil samples by identifying individual species present in a sample. Throughout this thesis, I examine the potential pitfalls associated with this method and demonstrate the robustness of this technique. Chapter 2 features a review article published in *Science and Justice* that discusses potential issues with soil DNA extraction and possible modifications for improving DNA yield. In Chapter 3, (under review in *Forensic Science International: Genetics),* the most appropriate target taxa for use in forensic soil analysis is identified, comparing discriminatory power, reproducibility and contamination risk. This study also highlights the importance of removing low level background DNA in HTS analysis and demonstrates the improvement in discriminatory power between samples when such signal is excluded. Chapters 4 and 5 examine the effect of DNA extraction bias and sample size on soil discrimination using different soil types. This has revealed that DNA extraction modifications can increase successful discrimination and that trace sample sizes can generate reproducible and discriminative DNA profiles. As the optimal DNA extraction method and reproducibility of DNA profiles varied with soil type, I provide recommendations based around soil pH and clay content. By designing a mock case scenario, in Chapter 6 I addressed the environmental factors likely to be encountered during an investigation. Specifically, the impact of temporal variation, and the relocation and storage of evidence samples were of concern as such factors could alter the DNA profile obtained and thus prevent a positive result. However, I demonstrate the variation in the soil community as a result of these factors did not prevent a link between evidence and crime scene samples. This concluding chapter summarises the most significant outcomes and highlights the contribution of my thesis to the field of forensic soil analysis. As this thesis could not address all the potential pitfalls of this methodology, I highlight additional factors that must be tested to further validate this technique, and provide recommendations for optimisation based on soil properties and bioinformatics development. Furthermore, I stress the need for standardisation, robust reference databases and the education of investigation personnel to establish best practice and increase the reliability and strength of soil DNA evidence in court.

Developmental validation of method

The research presented in this thesis provides encouraging insight into the potential of HTS technology for forensic soil science. Before this technique could be fully utilised in casework, research was required to validate the sensitivity, reproducibility, appropriate use of controls and associated biases. To achieve this, I examined multiple molecular markers to identify the most robust target for soil discrimination in a forensic context. As subtle differences in the soil DNA profile may be introduced during laboratory analysis and following criminal activity, the reliability of this highly sensitive technology was of concern. Throughout this thesis, I specifically addressed issues surrounding DNA extraction bias, sample size bias and environmental variation. However, I also repeatedly demonstrate that the DNA metabarcoding and HTS method developed in this thesis is robust, and that samples can be successfully discriminated despite such limitations. Although, this thesis aimed to resolve the potential pitfalls of this method, further validation is required to address transfer effects following a crime.

Examination of different molecular markers

In forensic science, different analysis methods are applied depending on the question and information available. Similarly, for forensic soil DNA analysis different biota can be targeted to address a specific question. To validate the method for soil discrimination, identifying the most appropriate molecular marker with regards to discriminatory power, contamination risk, spatial variability and reproducibility was essential. In Chapter 3, I examined four potential molecular markers, each targeting a different group of soil taxa: plants (*trn*L), bacteria (16S rRNA), fungi (ITS) and eukaryotes (18S rRNA). From these, the ITS gene region (specifically targeting fungi), showed the highest discriminatory power of the markers tested and was also associated with low background DNA levels, a desirable attribute for forensic application. Soil eukaryotes, both 18S and ITS, were also less variable within a single sample than bacterial 16S. This indicates that prokaryotic and eukaryotic markers could be analysed depending on the specific details of the case. For example, to establish a link between soil from a shoe and a particular footprint, perhaps bacterial DNA would be most appropriate for identifying a specific geographic origin, i.e. for evidential analysis, whereas soil eukaryotes may be more appropriate than bacteria for linking the general locality of a forensic sample i.e. for investigative analysis or evidential analysis where no footprint is evident. Nevertheless, I have identified three complementary targets that could be applied to a single sample set to strengthen the analysis. In addition, I found that soil eukaryotes were also less ubiquitous across multiple sites than bacteria, therefore further examination of individual phyla could determine if discrimination between locations can be increased by analysing specific eukaryote groups independently. This could be achieved bioinformatically from the 18S rRNA data, or alternatively by targeting specific phylum using specially designed PCR amplification primers.

In Chapter 3, I showed that plant-specific DNA profiles were the most reproducible within a site. A homogenous distribution of plant DNA could be useful for identifying the likely source of unknown sample in the absence of reference locations, particularly if rare plant taxa can be identified. Such information could direct the focus of an investigation based on the habitat requirements or known distribution of such taxa. However, PCR amplification of *trn*L from soil was problematic due to inhibition by humic acids, therefore visible plant roots and seeds could be isolated from the soil matrix and processed separately. This approach could be particularly useful for identifying a plant species known to be present at the scene, or associated with the victim, that could not be identified morphologically. Furthermore, I found that plant DNA within soil may be problematic for soil discrimination due to the low diversity present. However, the study in Chapter 3 only included two locations and a single barcode region, therefore plant DNA should not be excluded for this application so readily. Additional molecular markers, such as *rbc*L and *trnL_gh*, should also be examined across many location types and geographical regions to determine the most informative plant DNA barcode and the level of between-site resolution possible with plant markers. Matk is recommended by Barcode of Life consortium (BOLD) (http://www.barcodeoflife.org); however state that this region is too long (~600 bp) to successfully amplify from soil. Furthermore, pollen spores are also commonly found at crime scenes (1, 2), and can be present in soils (3). DNA metabarcoding and HTS of pollen in soil could provide another marker for discrimination; however, the dispersal area of pollen may limit the use in a forensic context.

Use of negative controls to monitor background levels of DNA

Contamination is a major issue when validating results for forensic science. For human identification, a PCR product in a negative control is detrimental, and the experiment must be repeated. In DNA metabarcoding, universal primers are designed to amplify a wide range of targets and as a result, negative controls can show sporadic low levels of PCR product (4-6). Laboratory protocols and cleaning procedures should be rigorously designed to minimise the level of such background DNA; however, DNA can also be introduced via the DNA extraction chemistry or PCR reagents (7, 8). Champlot *et al*. (9) examined the efficiency of common de-contamination protocols, including UV-radiation, DNA away, bleach and the use of dUTP in PCR amplifications. Although these protocols reduced contamination levels, a large number of control amplifications were recommended even when maximal caution is exerted during the analysis (9). To minimise the levels of background DNA levels, extreme caution should be implemented at all stages of the analysis. Samples should be collected in a sterile manner using gloves, facemasks, and DNA free tubes. The outside of the tubes should be cleaned with bleach following collection and the tube should be placed into a sealed bag to prevent sample-to-sample contamination. In the laboratory, the weighing equipment (i.e. balance) should be thoroughly cleaned, and the weigh boat, spatula and gloves should be replaced between each sample. DNA extractions and PCRs should also be carried out in dedicated hoods within clean pre-PCR laboratories. Many of these steps are standard practice in forensic science and will prevent sample-to-sample contamination and contamination from human DNA.

Chapter 3 addressed contamination by sequencing the PCR product from extraction blank controls (EBCs) that were visualised upon gel electrophoresis. From this, I determined that background DNA was more of an issue with 16S and 18S rRNA amplification, compared to ITS or *trn*L. By bioinformatically removing reads found in extraction blank control (EBCs) from all samples, the discriminatory power between sites increased using both 16S and 18S as might be expected. Similarly, Porter *et al.* (10) identified contaminant OTUs (0.5%) in 16S bacterial analysis from soil cores, and Schmieder *et al.* (11), reported human contamination in 72% of 202 published microbial and viral metagenomes. However, my study is the first to demonstrate the impact of EBCs on discriminatory power between soil samples, especially in a forensic context, highlighting the control measures required when performing DNA metabarcoding analysis, particularly when bacteria are targeted. I performed both EBCs and PCR negatives; however, PCR negatives were not sequenced as no PCR product was visible on the gel. However, agarose gels were stained using ethidium bromide staining with a detection limit of 0.5 to 5.0 ng/band and so very low quantities of DNA present in blanks may not be visible. Therefore future analyses using this methodology should always sequence both PCR negatives and EBCs as an additional measure to ensure that only OTUs originating from an individual sample are included in downstream analysis.

For validation of forensic soil DNA analysis, demonstrating rigorous monitoring of potential contamination issues is crucial for maintaining credible laboratory standards. Therefore, sequencing of negative controls (EBCs and no template controls) would provide a database of the OTUs commonly detected in the extraction and PCR reagents and could be used to remove background DNA from the sample data so DNA profiles can be presented with confidence. Such experimental controls could be used to compare the levels of contamination afforded by different commercial kits (and reagents) to identify the lowest contamination risk protocol and enable within laboratory validation. Documentation of negative controls could also be useful for monitoring the increase in background DNA levels in reagents over time, as reagents are susceptible to contamination upon opening. This could also provide guidelines on expiration time of a particular reagent and the best storage methods to minimise background DNA levels.

Assessment of DNA extraction bias

Maximising the genetic information recovered from trace quantities of soil is vital to produce robust and reproducible comparisons between forensic samples. Different DNA extraction protocols have reported inefficient DNA extraction and thus variations in abundance of taxa and absence of some taxa altogether (12-15). Such biases could be problematic in achieving an unambiguous result. In forensic science, commercial kits are favoured for ease of use and standardisation. Therefore in Chapter 4, I subjected five soil samples to a commonly used commercial DNA extraction kit and three modified methods of this protocol that could be easily implemented into current laboratory practice. I showed that the standard kit protocol failed to extract DNA from a range of taxonomic groups and demonstrated that different subsets of OTUs were detected upon different treatments.

Detection of additional diversity by different extraction methods could be advantageous for the investigative stage of a case that requires a detailed picture of the diversity to identify indicator taxa and direct the focus of a search. Discrepancies in the DNA profile may have been problematic for evidential stages of an investigation as subtle variations in genetic signature could prevent discrimination. However, my study in Chapter 4 indicates that variation within individual samples due to DNA extraction protocol was less than the variation between samples. This study involved five soils based on varying soil properties; therefore, the biota within these soils should have been very different.

Further validation of DNA extraction bias should perhaps concentrate on soils from similar location types, as subtle variations within a sample may be detrimental when comparing similar soil types from similar ecosystems with different geography. As described in Chapter 6, within site variation limited the differential resolution between soils from similar habitat types (i.e. organic rich soils from roadside verges). Discrepancies due to DNA extraction bias may increase within-site variation, and thus limit the power of the statistical analysis for evidential stages of an investigation. Therefore, comparative forensic soil DNA analysis between samples, and between laboratories, would benefit from standardised DNA extraction protocol.

From the experiment presented in Chapter 4, I cannot exclude that within-sample variation did not contribute to the differences observed between the standard protocol and the two protocols involving incubation at different temperatures; these three extracts were generated from a different sub-sample of each bulk soil and replicates were not performed for each treatment. However, a consistent increase in both DNA yield and OTU count across different soil types using 24 hour room temperature incubation encourages further research; a broad range of soil types from different locations and geographical regions should be examined to thoroughly assess the effect of this treatment in a forensic context. In contrast, the re-extraction protocol was performed directly from the initial soil material, therefore differences in DNA profiles cannot be a result of sampling bias. This is in agreement with previous studies which also show that not all DNA present within a sample is detected by a single extraction and suggest pooling of three successive extractions to minimise extraction bias (14, 16).

The re-extraction protocol offers a novel approach to maximise the use of trace samples. I showed that re-extraction was effective at increasing fungal diversity, particularly from low clay content soils, and should be applied for investigative purposes involving trace quantities. The re-extraction also generated DNA profiles representative of the specific samples, albeit with some subtle variations, therefore could prove useful in cases requiring future analyses as long as that the soil pellet is retained and stored at -20 ^{o}C. The discriminatory power afforded by the OTUs detected exclusively upon re-extraction offers an interesting study. The hypothesis would be that the initial extraction may remove DNA of the most abundant taxa, likely common to multiple locations. In contrast, the re-extraction might detect the rare and ubiquitous OTUs, as the lysis buffer in the second extraction may not be saturated by the high frequency taxa. As a result, the re-extraction protocol may generate DNA extracts with more pronounced genetic variation between sites and thus improve the resolution between sites compared to the initial extraction. This could be particularly useful for comparisons involving soils with similar location types, expected to have similar genetic signals.

Validation of trace quantities of soil

Trace quantities of soil in forensic analysis may not reliably reflect the full diversity of a source area (17, 18). Therefore, the sample size used in DNA extraction is a potential limitation that could introduce discrepancies in the DNA profile (19-23). To address this, Chapter 5 examined the effect of sample size on the ability to discriminate between samples, and determined that the discriminatory power between samples was not hindered by the use of sample sizes as little as 50 mg. This result demonstrates that trace soil samples can generate reliable DNA profiles, suggesting that small sample sizes can be compared to standard quantities (250 mg) of reference samples. Potentially, this would enable limited quantities of soil to be sub-sampled and used for analysis by multiple techniques, such as MIR. However, by reducing the sample size to conserve material in casework, DNA profile reproducibility may be reduced, particularly for fine or coarse soils due to particle size sampling bias. Therefore, larger sample sizes should be used for both fine clay soils and coarse textured soils, where possible.

With this said, DNA profile reproducibility will be dependent on the forensic context. For evidential stages of an investigation, the reproducibility of the DNA profile is less of a concern in terms of the exact dissimilarity value; two extracts from the same sample or location, should be more similar to each other, than to any extract obtained from a different sample, or location. As shown in Chapter 6, within-site variation reduces between site resolution using MDS; however, this is accounted for in CAP analysis. In contrast, investigative stages rely on capturing the complete sample diversity in order to identify individual taxa with specific habitat requirements or restricted geographical distribution. For a reliable result in this context, the DNA profile should be highly reproducible, so that particular taxa are identified with confidence, therefore for this purpose, large sample sizes (250 mg) are recommended, especially for fine clay soils and coarse soil samples.

Robustness to environmental variation

In practice, the soil DNA profile of both reference and unknown samples can be influenced by environmental factors. Most often, a time lapse will be experienced between the crime and the collection of reference soil samples. During this time, fluctuation in rainfall and temperature could introduce differences in the soil biota (24-26). To test such variation in a forensic context, I incorporated a six week lag time into the mock case scenario (Chapter 6) to mimic the likely time scale for an investigation, as advised by scientists at Forensic Science South Australia (FSSA). In addition, removal of the unknown soil from the environment presents a potential limitation of this method (27, 28). To subject soil to storage conditions often encountered in a case, I placed exhibits in the car boot at the time of the crime. After six weeks, soil adhered to shoes and a shovel had dried out, which further introduced concern given high levels of rainfall upon collection of the reference samples. To reflect the condition of the unknown samples, I air-dried a subset of the reference samples prior to extraction and the effect on discriminatory power was examined. Despite variable weather conditions, drying of unknown samples in the car boot and air-drying of reference samples, I demonstrated successful 18S rRNA analysis to predict the origin of soil adhered to the suspect’s belongings. Therefore, Chapter 6 highlights the potential of HTS in practice and suggests that the soil eukaryote DNA profile is robust to environmental variations. Similar evaluation is required to assess bacterial 16S rRNA and fungal ITS in practice. Furthermore, cold cases would involve collection of reference soils years after the crime has occurred. Long term studies are required to assess the variation in soil communities from different locations and habitats, and seasonal variations observed could potentially provide information on the likely transfer time of an evidential soil in a modern case. In an extreme case, the crime scene may have been subjected to a change in land use, for example, previously derelict grasslands may now consist of residential housing. As a result, soil DNA analysis may not be appropriate.

Transfer effects

Unknown samples are commonly subjected to primary and secondary transfer and often layers of soil are acquired over time, for example on car tyres. My study in Chapter 6 tested the effects of primary transfer of soil onto the shovel and shoes and showed that soil present prior to the crime did not interfere with the DNA signal. However, secondary transfer of soil onto the shoes following the crime was not incorporated in this experiment. If the suspect continued to wear the shoes there may have been soil material from multiple locations, thus obscuring the signal from the crime scene. As DNA within soil would be mixed by sampling multiple layers simultaneously, the separation and analysis of different layers might shed light on the validity of DNA analysis in these circumstances.

Investigating the effects of secondary transfer on soil discrimination is an important area of research that should be pursued further. Transfer of soil to materials or objects can alter the characteristics of the soil. Larger soil particles are often lost and only fine particles remain ingrained within the fibres of clothing. The loss of larger particles could impact the soil DNA profile as a fraction of the diversity may be lost. Experimental trials assessing the effect of soil fractionation on DNA profiles and the most efficient means to extract DNA from soil adhered to different materials would be of particular interest. Different molecular markers may produce more reliable and reproducible profiles for individual soil fractions, and the influence of materials on subsequent PCR amplification efficiency may differ between targets. Further research assessing the robustness of the method given transfer effects would provide guidelines on the potential of soil evidence to a particular investigation.

HTS and bioinformatics development

The main advantage of HTS is the ability to generate a vast amount of data relatively quickly; however, analysis of such data can be daunting. Many bioinformatics tools are required and each step requires careful selection of specified threshold parameters that could influence the soil DNA profile obtained. As the field of metabarcoding improves, optimisation of the data analysis methods will massively improve the reliability of HTS in forensic science**.** For forensic soil discrimination, a balance of sensitivity is required; a method should be sensitive to detect differences but avoid false positive results. Although, the approach applied in this thesis consistently showed successful soil discrimination between samples and sites, DNA profile reproducibility from a single sample appeared low (~50% similarity). Such variation between DNA profiles could be due to PCR bias (29, 30), although PCRs were performed in triplicate to minimise this effect. Optimisation of the PCR, inclusion of qPCR assays (31, 32) or a nested PCR approach (33) could potentially increase DNA profile reproducibility by generating a more accurate representation of taxonomic abundance in a sample. However, I propose that the variation observed within a sample or site is largely an artefact of sequencing error and HTS data analysis, which can artificially over-inflate the number of taxa observed (34, 35). Throughout this thesis, the specific data analysis tools were not previously discussed in detail as each Chapter was prepared for publication, therefore the following section describes the steps applied and discusses options for further improvement.

Choice of HTS platform

This thesis used two different HTS platforms, the Ion Torrent PGM and the Illumina MiSeq. The choice of platform for DNA metabarcoding analysis will vary depending on the target being analysed and the level of taxonomic resolution required. The Ion Torrent PGM offers sequencing of long fragment lengths (450 bp) potentially increasing taxonomic resolution. However, sequencing quality is reduced towards the end of the amplicon. In contrast, Illumina MiSeq platform offers a paired end approach (2 x 250 bp) offering bi-directional sequencing. Such approach reduces sequencing errors by only accepting reads that have two successfully merged sequences. In addition, the MiSeq is based on pyrosequencing and sequencing errors can be modelled (36). In contrast, the Ion Torrent PGM relies on a new pH based technology, has a higher error rate in comparison to the Illumina technology and sequencing errors cannot be easily modelled (37). However, the Illumina MiSeq is hindered by the need for customised sequencing primers for DNA metabarcoding analysis. The MiSeq requires an Illumina sequencing primer to be incorporated into the amplicon for successful data generation. However, the combined Illumina adapter/Illumina sequencing primer/index/locus specific sequence is too long for efficient single step fusion primer PCR amplification. As a result, the Illumina sequencing primer is replaced with a customised sequencing primer sequence to shorten the PCR primers; however, at present customised sequencing primers have only been designed for the 16S bacterial rRNA and 18S rRNA gene regions only. This complicates the potential to pool different targets onto a single sequencing run and limits the metagenomic markers that can be analysed on the Illumina platform. To overcome this limitation, a two-step PCR could be trialled using truncated primers; however, this would introduce another step for potential contamination so a single step PCR would be preferred for forensic analysis. Due to this, comparison of the four molecular markers in Chapter 3, and fungal ITS analysis in Chapters 4 and 5 were sequenced using the Ion Torrent PGM, which offers the potential to sequence any loci relatively easily and enables sequencing of different fragment lengths in parallel. Identification of individual taxa would be beneficial for investigative purposes. Therefore, I would recommend the MiSeq for analysis of 16S and 18S, because higher sequence quality can be obtained and thus more reliable taxonomic identification can be achieved. However, forensic soil discrimination does not require identification of individual OTUs to compare samples.

Sequencing depth

The choice of NGS platform will also influence the number of sequences obtained for individual samples; this is referred to as sequence coverage. Low coverage may not capture the full diversity within a sample, resulting in DNA profile variation between replicates. Diversity can be visualised from rarefaction curves, which plateau when the diversity has been saturated. The MiSeq offers greater sequencing coverage (8 GB data) on a single run compared the Ion Torrent PGM 318 chip (1 GB). Increased sequencing depth offered by the MiSeq Illumina would benefit investigative stages of a case by generating a more complete picture of soil diversity. In Chapters 3, 4 and 5, samples were analysed at relatively low sequencing depth (~2,000 per sample) using the Ion Torrent PGM, suggesting sufficient coverage for this soil discrimination purposes. However, in Chapter 6 the MiSeq was chosen to achieve higher coverage and better sequencing quality. In comparison to the Ion Torrent, the increased coverage from the MiSeq (~16,000 per sample) did not increase the similarity between samples within a site; e.g. for 18S 48 ± 14% similarity in Chapter 3 (3,000 sequences per sample) compared to 41 ± 10 % in chapter 6 (16,465 sequences per sample). This suggests that increased sequencing depth did not improve the reproducibility of samples within a site, and therefore the sequencing depth of the Ion Torrent was sufficient. I believe the increased coverage from the MiSeq proportionally increased the diversity and OTU abundance of all samples, compared to the Ion Torrent, and as a result within-site variation was remained the same regardless of sequencing depth. As there is a trade-off between coverage and the number of samples sequenced on a single run, the number of samples run in parallel on the MiSeq could be increased (e.g. *n=100*) without adversely affecting the ability to discriminate. In contrast, the Ion Torrent would be useful for analysing small sample sets (e.g. *n=30*), allowing sequencing runs on a case-per-case basis.

Sequence quality filtering

As sequencing error is an ongoing issue with HTS platforms, particularly with the Ion Torrent PGM (37), quality filtering is an important step applied to ensure that only the most reliable sequences are include in downstream analysis. Chimeras formed during PCR amplification (creating a hybrid DNA molecule that registers as a novel sequence) and pyrosequencing errors (generating base insertion or deletion errors, known as indels) can contribute to inflated diversity estimates. The impact of sequencing errors on forensic soil analysis may differ depending on the application. For investigative stages, taxonomic identification of OTUs may be important, therefore quality filtering is required to ensure only reliable sequences are retained and presence of taxonomy are determined with certainty. Therefore, specialised bioinformatics methods that have been developed to account for different sequencing artefacts will be important (38-40). For examples, Amplicon Pyrosequence Denoising Program (APDP) offers a conservative sequence validation pipeline based on abundance distribution of similar sequences across independent samples, as well as within individual samples (41). In contrast, stringent removal of sequencing errors may be not be as crucial for soil discrimination as individual taxa are not identified. This thesis applied Phred scores (Q>20) to filter low quality sequences (42), and strict zero mismatch thresholds for both the primer and MIDs/index sequences were applied. In addition, I removed sequences of <100 bp to include only sequences of sufficient length for reliable taxonomic identification. Despite applying a less conservative approach to quality filtering successful sample discrimination was achieved. This would be due to a consistent level of 'noise' in all samples. Nevertheless, the impact of more stringent data analysis on resolution between different locations using different molecular markers would benefit forensic soil DNA analysis.

OTU picking

Following quality filtering steps, the remaining sequences are clustered into operational taxonomic units (OTUs) based on a specified sequence similarity. This step also accounts for sequencing errors, and reduces the computational power required for taxonomic identification against a reference database. DNA metabarcoding relies on the formation of operational taxonomic units (OTUs) based on sequence similarity to estimate the diversity within a sample (43); the similarity percentage threshold must be specified and will alter the number of OTUs observed. I used a threshold of 97% similarity; however relaxing this parameter (to 90%) would reduce OTUs and further minimise sequencing error artefacts and allow for better discrimination between samples. Currently, three main OTU picking approaches exist: 1) *de novo* OTU picking, 2) closed reference OTU picking, and 3) open reference OTU picking, which comprises of reference based clustering followed by *de novo* clustering of sequences that do not match database sequences. As this thesis focussed on soil discrimination, *de novo* OTU clustering was used so as not to discard unidentified taxa that could be unique to specific sites. However, by applying a more conservative reference based OTU picking, the noise observed between duplicate extracts from a single sample/site may have been reduced (44). For investigative analysis, a closed reference OTU picking approach could be more appropriate so that all OTUs can be identified against a curated database. It is also common practice to establish a minimum read threshold for which an OTU can be accepted or rejected, i.e. singletons/doubletons are commonly removed (34). This threshold could be important for investigative analysis where the presence of a particular OTU could provide information on the likely origin. For forensic soil discrimination, significantly increasing this boundary could provide a more realistic threshold for extracts originating from the same sample/site. To ensure robust and reliable HTS analysis for forensic casework, each of these parameters requires optimisation and standardisation so that a positive result can be concluded with a high degree of certainty.

Establishing a link between soils in a forensic context

For forensic DNA analysis, three possible outcomes are possible: no match, inconclusive or a positive match. The latter result requires statistical analysis to support the conclusion and provide meaning to the match. Following data filtering and OTU picking, samples can be compared using software programs previously designed for the analysis of ecological data, such as PRIMER6, or alternatively newly developed bioinformatics tools can be applied, e.g. QIIME. In addition, statistical analysis or probability values have to be reported to support the evidence in a forensic context. Therefore, before soil DNA analysis using HTS can be fully utilised in casework, assessment of different sample comparison methods will be required and an extensive reference DNA profile database must be developed to determine the significance of a match. The following section provides recommendations for achieving such goal.

Comparison of soil DNA profiles

The application of robust statistical analysis to predict the likely origin of an unknown forensic soil sample was not performed prior to my thesis. T-RFLP studies used the mean Bray-Curtis distance between samples from the same location as a threshold for determining a match (45-47), or have applied nMDS and ANOVA only (24, 27, 48). In Chapter 6, I demonstrated poor resolution between sites using nMDS and the ability of CAP analysis to improve soil discrimination and generate prediction statistics. Throughout this thesis, Multidimensional Scaling (MDS) was used to illustrate the similarity between samples and within-site variation. ANOSIM statistics were subsequently applied to determine significant differences between samples. In a forensic context, this approach could indicate the most likely habitat type/environment of an unknown sample in investigations where no case specific reference sites are available. However, as demonstrated in Chapter 6, resolution between sites can be poor using MDS and therefore, I introduced Canonical Analysis of Principle coordinates (CAP), as a means to maximise the differences between *a priori* groups (i.e. reference sites), whilst minimising differences within a group. In addition, CAP analysis enabled prediction statistics in a forensic context by utilising the 'leave one out' cross-validation procedure as a classification analysis. Therefore, CAP analysis provides a powerful analysis method for establishing a link between an evidential sample and a specific source when solid background information is available in a case. CAP analysis and prediction statistics would assist evidential stages, whereas MDS and ANOSIM would provide intelligence at investigative stages of a case.

Although CAP offers an effective means to determine the weight of evidence, newly developed Bayesian analysis methods should also be applied to further advance soil analysis in a forensic context. In particular, Source Tracker (49) estimates the proportion of a community that comes from a set of source environments and generates a comprehensive visual representation that can be easily interpreted. Originally Source Tracker was developed to track the background signal associated with extraction blank controls; however, source tracker could have a number of applications in a forensic soil analysis context. First, such analysis could be applied to visualise the proportion of the evidential sample community that originated from the crime scene samples. Given that the crime scene soil represents the highest proportion of the evidential sample such analysis could support the CAP analysis and prediction statistics for evidential stages of an investigation. Source Tracker could also prove useful for investigative stages of an investigation by indicating the likely source of an unknown sample given DNA profiles from different habitat types and geographical regions. Source Tracker also enables prediction of the proportion of each reference sample that could be attributed to the other reference samples to examine potential sample-to-sample contamination. This would be an important tool in forensic science to address an opposing attorneys concern that cross contaminated may have occurred between the evidential sample and crime scene samples. Furthermore, Source Tracker could be applied as a quality assurance measure to demonstrate low background DNA levels attributed to each sample. This would increase confidence in the results by illustrating that rigorous cleaning protocols were followed during sample processing and any OTUs that were detected in negative controls have been removed.

Reference databases

In forensic science, the probability that a sample originated from one source, rather than another selected at random, must be reported by evaluating the weight of evidence. Human DNA profiling statistics are provided as the Random Match Probability or as a Likelihood Ratio (LR). Random Match Probability reflects the frequency of a particular DNA profile in a population, whereas LR involves comparison of the probabilities from two opposing hypotheses and can incorporate prior odds (50). For example, if a genotype is common within the population, the contribution of DNA from the suspect cannot be determined with confidence; however, when a genotype is rare the hypothesis that the suspect contributed to the crime scene sample is stronger. As a result, the key to these statistics is to sample enough individuals to reliably estimate the frequency of the major alleles at a particular locus; this sample size has been published as 100-200 individuals for human population genetics (51). Similar reference databases will be important for soil comparison purposes to understand the frequency of soil DNA profile.

Soil analysis significantly differs from human identification, as soil is not a discrete entity and the soil community is vulnerable to influences of both temporal and spatial variation. Therefore, forensic casework would benefit from generation of a comprehensive and well-documented reference database, comprising of HTS data from forensically relevant locations, e.g. gardens, parks, farmlands, from different geographic locations. Such a database would enable comparison of unknown samples to reference samples from a specific case, as well as state-wide and potentially nationwide reference datasets thus strengthening statistical analysis. Reference sample data would require detailed metadata, such as date of collection, precise location, habitat type, and weather conditions. For example, the Earth Microbiome Project (EMP) is an example of such a database that could be implemented into forensic analyses (52). Since HTS data analysis of large datasets can be computationally intense, the database could be subdivided into geographical regions and/or location types, so that the analysis is directed by prior knowledge of a specific case. For example, if the soil is believed to have originated from a coastal region, comparison to all reference coastal data may be sufficient. HTS offers a means to generate a reference database efficiently and inexpensively, and such a database would enable long-term storage of HTS data for re-analysis upon development of new bioinformatics tools. Ideally, for routine analysis of soils using this technology, the ultimate goal would be to develop a user-friendly software program that enables a scientist to upload HTS sequencing data from a specific case, and analyse the samples against reference data using a standardised bioinformatics pipeline as discussed previously.

Future directions

Development of a reference database would further advance forensic soil DNA analysis and enable such evidence to be applied to many different forensic questions encountered in casework. However, for such a reference database to be effective all stages of the methodology, from sample collection to data analysis, must be standardised to ensure comparable results between laboratories. In addition, investigative teams should be informed of best practice, and the value of soil trace evidence, to ensure reliable results are delivered in court.

Standardisation of new methods for forensics

Standardised methods are required in forensic laboratories to ensure reliable and consistent results. To achieve this, forensic laboratories are required to validate methods and protocols both within the laboratory, and between different laboratories. Within laboratories, scientists can use standard reference materials (SRMs) to check the performance of protocols following a significant modification e.g. RFLP SRMs or PCR-based SRMs to ensure consistent results. However, as DNA metabarcoding is a relatively new discipline, biases associated with each analysis step should be minimised to reduce opportunity for the opposing attorney to contest evidence in court. Therefore, a single standardised protocol should be developed to compare soil samples for forensic purposes and ensure all samples recorded in the reference database were processed identically. Standardisation should include sample collection, sample storage, contamination control, DNA extraction, PCR amplification, PCR purification, HTS sequencing platform and bioinformatics analysis. Where possible, commercial products and kits should be used to implemented to standardise materials across the field (50). However, in-house validation of new commercial kits (e.g. 53) and inter-laboratory validation is also required to demonstrate consistency (e.g. 54). Validation of soil DNA metabarcoding and HTS should be as extensive as possible so that a single protocol can be applied to many laboratories. Collaborative validation would test the limits of this technique and thus strengthen the value of HTS soil evidence in casework. Throughout this thesis, I addressed many of these factors and so the work presented provides the initial steps of this standardisation procedure.

Educating personnel

The investigation team must be advised of best practice to preserve valuable soil DNA evidence. As many forensic investigators are not biologists, they may be unaware that soil biota, particularly bacteria, are vulnerable to changes in the environment. In addition, soil DNA is not as well recognised in casework as other forms of DNA evidence, which include blood, semen, and hair. Therefore, policing bodies and investigation teams may lack training in areas regarding soil contamination, soil collection, and sample preservation. Raising awareness of the potential of this trace evidence, as well as stressing the factors which may jeopardise its value in court, is absolutely vital. For example, shoes are commonly retrieved at a scene and stored as evidence in a crime lab. Therefore, soil adhered to the shoes may remain in storage for long periods of time and be subjected to environmental contamination, both of which can alter the biological signal of that sample (28, 55). Although this thesis indicated that such factors did not prevent a match, eliminating this possibility will prevent opposing attorneys contesting this aspect of evidence. Soil should be collected and extracted during the initial recovery of shoes, in case such evidence is required at a later date. This would also enable the use of soil DNA evidence in cold cases, where such evidence would otherwise be lost. Informing people involved at the early stages of investigations and educating them on the best practices of sample collection and preservation would increase the long-term potential to utilize soil DNA evidence in forensic work. Furthermore, implementing soil DNA evidence as a module into undergraduate training programs would also help raise awareness to young forensic scientists and increase interest for this rapidly developing and highly relevant area of forensic science.

Additional applications of the methodology

This thesis develops and validates the use of DNA metabarcoding and HTS for soil discrimination in forensic science. However, soil DNA metabarcoding and HTS applications are not limited to soil discrimination, and could be extended to address other forensically relevant questions. This technology could provide forensic intelligence in cases where no specific case reference locations are evident. DNA profiles from unknown samples could be compared to the reference database described previously to provide information on the likely location type, or geographical region of an unknown sample. Alternatively, identification of specific taxa with known distributions or specialised habitat requirements could be used to narrow the search area. Soil DNA metabarcoding could also assist geospatial surveys, which often use remote sensing techniques such as aerial photography, topographic mapping and satellite imagery to locate human remains (56, 57). Detection of taxa associated with human decomposition could assist traditional geospatial surveys in identifying burial plots in investigations where no physical remains are evident. By collecting topsoils across such a particular area, HTS DNA profiles which contain taxa indicative of decomposition could identify the specific area where a victim may have been buried. Although this research focuses primarily on soil discrimination, the methodology developed provides the basis for answering a magnitude of forensic related questions, and consequently could be employed as an additional tool in forensic geoscience.

Concluding Remarks

Forensic geoscience is currently considered to be an emerging discipline that can offer significant benefits to forensic investigations. Soil analysis can either provide valuable information on the likely location of an unknown sample to direct the investigation, or alternatively, establish a link between evidence samples and a crime scene. Ultimately, my thesis highlights the potential value of HTS to forensic soil discrimination and informs researchers of the necessary precautions associated with soil DNA evidence. This research demonstrates the first practical application of DNA metabarcoding and high-throughput sequencing (HTS) for forensic soil analyses. I compare the small-scale reproducibility and between-site resolution of four molecular markers (targeting different taxonomic groups), and identify for the first time that soil eukaryotes (particularly fungi) provide greater discriminatory power than bacteria or plants. The approach developed throughout this thesis consistently distinguished between soil samples taken from different localities and was robust to numerous environmental factors and potential limitations commonly encountered in casework. Based on the findings of my thesis, I provide recommendations on important practical and analytical steps required to obtain a robust DNA profile given the wide spatial variability of taxa and sensitivity of HTS to contaminant DNA. The information provided in this body of research will facilitate informed decisions about the most appropriate marker for soil DNA analysis given a specific case. Consequently, my work is aimed at encouraging the use of soil HTS analysis as an additional line of evidence or investigation in forensic casework. I stress the magnitude of possible questions that could be answered using this technique and strongly encourage further research into this area of forensic science.

References

1. **Mildenhall, D. C., P. E. J. Wiltshire, and V. M. Bryant.** 2006. Forensic palynology: Why do it and how it works. Forensic Science International **163:**163-172.
2. **Wiltshire, P. J.** 2009. Forensic Ecology, Botany, and Palynology: Some Aspects of Their Role in Criminal Investigation, p. 129-149. *In* K. Ritz, L. Dawson, and D. Miller (ed.), Criminal and Environmental Soil Forensics. Springer Netherlands.
3. **Bruce, R. G., and M. E. Dettmann.** 1996. Palynological analysis of Australian surface soild and their potential in forensic science. Forensic Science International **81:**77-94.
4. **Epp, L. S., S. Boessenkool, E. P. Bellemain, J. Haile, A. Esposito, T. Riaz, C. Erseus, V. I. Gusarov, M. E. Edwards, A. Johnsen, H. K. Stenoien, K. Hassel, H. Kauserud, N. G. Yoccoz, K. Brathen, E. Willerslev, P. Taberlet, E. Coissac, and C. Brochmann.** 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Mol. Ecol. **21:**1821-1833.
5. **Taberlet, P., E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev.** 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. **21:**2045-2050.
6. **Taberlet, P., E. Coissac, M. Hajibabaei, and L. H. Rieseberg.** 2012. Environmental DNA. Mol. Ecol. **21:**1789-1793.
7. **Tanner, M. A., B. M. Goebel, M. A. Dojka, and N. R. Pace.** 1998. Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. Appl. Environ. Microbiol. **64:**3110-3113.
8. **Corless, C. E., M. Guiver, R. Borrow, V. Edwards-Jones, E. B. Kaczmarski, and A. J. Fox.** 2000. Contamination and sensitivity issues with a real-time universal 16S rRNA PCR. Journal of Clinical Microbiology **38:**1747-1752.
9. **Champlot, S., C. Berthelot, M. Pruvost, E. A. Bennett, T. Grange, and E.-M. Geigl.** 2010. An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. PLoS One **5:**e13042.
10. **Porter, T. M., G. B. Golding, C. King, D. Froese, G. Zazula, and H. N. Poinar.** 2013. Amplicon pyrosequencing late Pleistocene permafrost: the removal of

putative contaminant sequences and small-scale reproducibility. Molecular Ecology Resources **13:**798-810.

11. **Schmieder, R., and R. Edwards.** 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PloS ONE **6:**e17288.

12. **Martin-Laurent, F., L. Philippot, S. Hallet, R. Chaussod, J. C. Germon, G. Soulas, and G. Catroux.** 2001. DNA extraction from soils: Old bias for new microbial diversity analysis methods (vol 67, pg 2354, 2001). Appl. Environ. Microbiol. **67:**4397-4397.

13. **Knauth, S., H. Schmidt, and R. Tippkötter.** 2012. Comparison of commercial kits for the extraction of DNA from paddy soils. Lett. Appl. Microbiol. **56:**222-228.

14. **Feinstein, L. M., W. J. Sul, and C. B. Blackwood.** 2009. Assessment of bias associated with incomplete extraction of microbial DNA from soil. Appl. Environ. Microbiol. **75:**5428-5433.

15. **Zhang, D., W. Li, S. Zhang, M. Liu, and H. Gong.** 2011. Evaluation of the impact of DNA extraction methods on BAC bacterial community composition measured by denaturing gradient gel electrophoresis. Lett. Appl. Microbiol. **53:**44-49.

16. **Jones, M. D., D. R. Singleton, W. Sun, and M. D. Aitken.** 2011. Multiple DNA Extractions Coupled with Stable-Isotope Probing of Anthracene-Degrading Bacteria in Contaminated Soil. Appl. Environ. Microbiol. **77:**2984-2991.

17. **Pye, K., and D. J. Croft.** 2004. Forensic geoscience: introduction and overview. Geological Society, London, Special Publications **232:**1-5.

18. **Ruffell, A.** 2010. Forensic pedology, forensic geology, forensic geoscience, geoforensics and soil forensics. Forensic Science International **202:**9-12.

19. **Ranjard, L., D. P. H. Lejon, C. Mougel, L. Schehrer, D. Merdinoglu, and R. Chaussod.** 2003. Sampling strategy in molecular microbial ecology: influence of soil sample size on DNA fingerprinting analysis of fungal and bacterial communities. Environ. Microbiol. **5:**1111-1120.

20. **Grundmann, L. G., and F. Gourbiere.** 1999. A micro-sampling approach to improve the inventory of bacterial diversity in soil. Applied Soil Ecology **13:**123-126.

21. **Ellingsøe, P., and K. Johnsen.** 2002. Influence of soil sample sizes on the assessment of bacterial community structure. Soil Biol. Biochem. **34:**1701-1707.

22. **Taberlet, P., S. M. Prud'homme, E. Campione, J. Roy, C. Miquel, W. Shehzad, L. Gielly, D. Rioux, P. Choler, and J. C. Clement.** 2012. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. Mol. Ecol. **21:**1816-1820.

23. **Kang, S., and A. L. Mills.** 2006. The effect of sample size in studies of soil microbial community structure. J. Microbiol. Methods **66:**242-250.

24. **Meyers, M. S., and D. R. Foran.** 2008. Spatial and temporal influences on bacterial profiling of forensic soil samples. Journal of Forensic Sciences **53:**652-660.

25. **Baker, K. L., S. Langenheder, G. W. Nicol, D. Ricketts, K. Killham, C. D. Campbell, and J. I. Prosser.** 2009. Environmental and spatial characterisation of bacterial community composition in soil to inform sampling strategies. Soil Biol. Biochem. **41:**2292-2298.

26. **Darby, B. J., D. A. Neher, D. C. Housman, and J. Belnap.** 2011. Few apparent short-term effects of elevated soil temperature and increased frequency of summer precipitation on the abundance and taxonomic diversity of desert soil micro- and meso-fauna. Soil Biology & Biochemistry **43:**1474-1481.

27. **Macdonald, L. M., B. K. Singh, N. Thomas, M. J. Brewer, C. D. Campbell, and L. A. Dawson.** 2008. Microbial DNA profiling by multiplex terminal restriction fragment length polymorphism for forensic comparison of soil and the influence of sample condition. Journal of Applied Microbiology **105:**813-821.

28. **Bainard, L. D., J. N. Klironomos, and M. M. Hart.** 2010. Differential effect of sample preservation methods on plant and arbuscular mycorrhizal fungal DNA. J. Microbiol. Methods **82:**124-130.

29. **Pinto, A. J., and L. Raskin.** 2012. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. PLoS ONE **7:**e43093.

30. **Sipos, R., A. J. Szekely, M. Palatinszky, S. Revesz, K. Marialigeti, and M. Nikolausz.** 2007. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol. Ecol. **60:**341-50.

31. **Murray, D. C., M. Bunce, B. L. Cannell, R. Oliver, J. Houston, N. E. White, R. A. Barrero, M. I. Bellgard, and J. Haile.** 2011. DNA-Based Faecal Dietary Analysis: A Comparison of qPCR and High Throughput Sequencing Approaches. PloS One **6:**e25776

10.1371/journal.pone.0025776.

32. **Prevost-Boure, N. C., R. Christen, S. Dequiedt, C. Mougel, M. Lelievre, C. Jolivet, H. R. Shahbazkia, L. Guillou, D. Arrouays, and L. Ranjard.** 2011. Validation and Application of a PCR Primer Set to Quantify Fungal Communities in the Soil Environment by Real-Time Quantitative PCR. PloS one **6.9**: e24166.

33. **Berry, D., K. B. Mahfoudh, M. Wagner, and A. Loy.** 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl. Environ. Microbiol. **77:**7846-7849.

34. **Unterseher, M., A. Jumpponen, M. Opik, L. Tedersoo, M. Moora, C. F. Dormann, and M. Schnittler.** 2011. Species abundance distributions and richness estimations in fungal metagenomics - lessons learned from community ecology. Mol. Ecol. **20:**275-285.

35. **Kunin, V., A. Engelbrektson, H. Ochman, and P. Hugenholtz.** 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. **12:**118-123.

36. **Logares, R., T. H. Haverkamp, S. Kumar, A. Lanzén, A. J. Nederbragt, C. Quince, and H. Kauserud.** 2012. Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. J. Microbiol. Methods **91:**106-113.

37. **Bragg, L. M., G. Stone, M. K. Butler, P. Hugenholtz, and G. W. Tyson.** 2013. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. PloS Comput. Biol. **9:**e1003031.

38. **Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince, and R. Knight.** 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics **27:**2194-2200.

39. **Quince, C., A. Lanzen, R. J. Davenport, and P. J. Turnbaugh.** 2011. Removing noise from Pyrosequenced Amplicons. BMC Bioinformatics **12**.

40. **Quince, C., A. Lanzen, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan.** 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Meth **6:**639-641.

41. **Morgan, M., A. Chariton, D. Hartley, and C. Quince.** 2013. Improved inference of taxonomic richness from environmental DNA. Plos One **8**.

42. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. Genome Res. **8:**175-185.

43. **Huse, S. M., D. M. Welch, H. G. Morrison, and M. L. Sogin.** 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. **12:**1889-1898.

44. **Kuczynski, J., C. L. Lauber, W. A. Walters, L. W. Parfrey, J. C. Clemente, D. Gevers, and R. Knight.** 2012. Experimental and analytical tools for studying the human microbiome. Nat. Rev. Genet. **13:**47-58.

45. **Heath, L. E., and V. A. Saunders.** 2006. Assessing the Potential of Bacterial DNA Profiling for Forensic Soil Comparisons*. Journal of Forensic Sciences **51:**1062-1068.

46. **Quaak, F. C. A., and I. Kuiper.** 2011. Statistical data analysis of bacterial t-RFLP profiles in forensic soil comparisons. Forensic Science International **210:**96-101.

47. **Pasternak, Z., A. Al-Ashhab, J. Gatica, R. Gafni, and S. Avraham.** 2012. Optimization of molecular methods and statistical procedures for forensic fingerprinting of microbial soil communities. Int Res J Microbiol **3:**363-372.

48. **Macdonald, C. A.** 2011. Discrimination of soils at regional and local levels using bacterial and fungal t-RFLP profiling. Journal of Forensic Sciences **56:**61-69.

49. **Knights, D., J. Kuczynski, E. S. Charlson, J. Zaneveld, M. C. Mozer, R. G. Collman, F. D. Bushman, R. Knight, and S. T. Kelley.** 2011. Bayesian community-wide culture-independent microbial source tracking. Nat Meth **8:**761-763.

50. **Butler, J. M.** 2005. Forensic DNA typing: biology, technology, and genetics of STR markers. Academic Press.

51. **Foreman, L., and I. Evett.** 2001. Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system. International Journal of Legal Medicine **114:**147-155.

52. **Gilbert, J., F. Meyer, D. A. Antonopoulos, P. Balaji, C. T. Brown, C. T. Brown, N. Desai, J. A. Eisen, D. Evers, D. Field, W. Feng, D. Huson, J. Jansson, R. Knight, J. Knight, E. Kolker, K. Konstantindis, J. Kostka, N. C. Kyrpides, R. Mackelprang, A. McHardy, C. Quince, J. Raes, A. Sczyrba, A. Shade, and R. Stevens.** 2010. Meeting Report: The Terabase Metagenomics Workshop and the Vision of an Earth Microbiome Project, vol. 3.

53. **Frank, W. E., B. E. Llewellyn, P. A. Fish, A. K. Riech, T. L. Marcacci, D. W. Gandor, D. Parker, R. Carter, and S. M. Thibault.** 2001. Validation of the AmpFeSTR™ Profiler Plus PCR Amplification Kit for Use in Forensic Casework. Journal of Forensic Sciences **46:**642-646.

54. **Musgrave-Brown, E., D. Ballard, K. Balogh, K. Bender, B. Berger, M. Bogus, C. Børsting, M. Brion, M. Fondevila, and C. Harrison.** 2007. Forensic validation of the SNP< i> for</i> ID 52-plex assay. Forensic Science International: Genetics **1:**186-190.

55. **Tzeneva, V. A., J. F. Salles, N. Naumova, W. M. de Vos, P. J. Kuikman, J. Dolfing, and H. Smidt.** 2009. Effect of soil sample preservation, compared to the effect of other environmental variables, on bacterial and eukaryotic diversity. Research in Microbiology **160:**89-98.

56. **Pringle, J. K., A. Ruffell, J. R. Jervis, L. Donnelly, J. McKinley, J. Hansen, R. Morgan, D. Pirrie, and M. Harrison.** 2012. The use of geoscience methods for terrestrial forensic searches. Earth-Science Reviews **114:**108-123.

57. **Pringle, J. K., C. Holland, K. Szkornik, and M. Harrison.** 2012. Establishing forensic search methodologies and geophysical surveying for the detection of clandestine graves in coastal beach environments. Forensic Science International **219:**e29-e36.