Forensic Science International 310 (2020) 110250

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Use of standardized bioinformatics for the analysis of fungal DNA signatures applied to sample provenance

Julia S. Allwood[a,*], Noah Fierer[b,c], Robert R. Dunn[d], Matthew Breen[a], Brian J. Reich[e], Eric B. Laber[e], Jesse Clifton[e], Neal S. Grantham[e], Seth A. Faith[a,f]

[a] Department of Molecular Biomedical Sciences, North Carolina State University, 1060 William Moore Dr., Raleigh, NC, 27607, USA
[b] Department of Ecology and Evolutionary Biology, University of Colorado, 216 UCB, Boulder, CO, 80309-0216, USA
[c] Cooperative Institute for Research in Environmental Sciences, University of Colorado, 216 UCB, Boulder, CO, 80309-0216, USA
[d] Department of Applied Ecology, North Carolina State University, 100 Brooks Ave., David Clark Labs, Raleigh, NC, 27607, USA
[e] Department of Statistics, North Carolina State University, 2311 Stinson Dr., Raleigh, NC, 27607, USA
[f] Battelle Memorial Institute, 505 King Ave., Columbus, OH, 43201, USA

ARTICLE INFO

Article history:
Received 24 February 2020
Accepted 11 March 2020
Available online 12 March 2020

Keywords:
Forensic microbiology
Bioinformatics
Metabarcoding
Sample provenance

ABSTRACT

The use of environmental trace material to aid criminal investigations is an ongoing field of research within forensic science. The application of environmental material thus far has focused upon a variety of different objectives relevant to forensic biology, including sample provenance (also referred to as sample attribution). The capability to predict the provenance or origin of an environmental DNA sample would be an advantageous addition to the suite of investigative tools currently available. A metabarcoding approach is often used to predict sample provenance, through the extraction and comparison of the DNA signatures found within different environmental materials, such as the bacteria within soil or fungi within dust. Such approaches are combined with bioinformatics workflows and statistical modelling, often as part of large-scale study, with less emphasis on the investigation of the adaptation of these methods to a smaller scale method for forensic use. The present work was investigating a small-scale approach as an adaptation of a larger metabarcoding study to develop a model for global sample provenance using fungal DNA signatures collected from dust swabs. This adaptation was to facilitate a standardized method for consistent, reproducible sample treatment, including bioinformatics processing and final application of resulting data to the available prediction model. To investigate this small-scale method, 76 DNA samples were treated as anonymous test samples and analyzed using the standardized process to demonstrate and evaluate processing and customized sequence data analysis. This testing included samples originating from countries previously used to train the model, samples artificially mixed to represent multiple or mixed countries, as well as outgroup samples. Positive controls were also developed to monitor laboratory processing and bioinformatics analysis. Through this evaluation we were able to demonstrate that the samples could be processed and analyzed in a consistent manner, facilitated by a relatively user-friendly bioinformatic pipeline for sequence data analysis. Such investigation into standardized analyses and application of metabarcoding data is of key importance for the future use of applied microbiology in forensic science.

1. Introduction

A fundamental aspect of forensic biology is the investigative capability to detect links, where present, between individuals, objects and locations. While such links are often demonstrated through the use of human DNA, the use of environmental material to contribute information relevant to criminal investigations long antecedes the use of DNA for human identification, featuring in the stories of Sherlock Holmes and in the works of Locard in the early twentieth century [1,2]. Plant and fungal material, as identified morphologically by experts, has been accepted as evidence in the judicial systems of many countries (for example refer to [3,4]). Morphologically identified plant and fungal materials continue to be a valuable, if perhaps often overlooked, source of information for criminal investigations. Examples of such uses include the successful identification of the original mass burial sites of relocated bodies during criminal investigations of the Homeland

* Corresponding author. Present address: Manaaki Whenua Landcare Research, 231 Morrin Road, St Johns, Auckland 1072, New Zealand.
E-mail address: allwoodj@landcareresearch.co.nz (J.S. Allwood).

http://dx.doi.org/10.1016/j.forsciint.2020.110250
0379-0738/

War in Former Yugoslavia [5] and cases of aggravated assault or murder [6,7]. The use of massively paralleled DNA sequence-based approaches (hereafter referred to as 'sequencing') for questions of forensic biology targeting human DNA are being investigated, with some already implemented for use. In addition, sequencing approaches are being evaluated in a forensic context when targeting environmental DNA, such as DNA from plants, fungi, invertebrates and bacteria. With the majority of these studies thus far applying a metabarcoding approach, environmental DNA is being evaluated for its use for sample provenance [8–10], post-mortem interval prediction [11–16] and human identification via skin microbiomes [17–20]. These studies, and many others, operate under the overarching objective of developing applications ultimately capable of contributing investigative leads.

The benefits of these approaches are many. They include the capacity to use a variety of types of material present at crime scenes, the ability to simultaneously generate data from multiple samples and genetic targets, and potentially high sensitivity and minimal sample consumption. Yet, the implementation of such approaches in a forensic casework setting presents some unique challenges. Many DNA-based methods currently employed in forensic biology use purchased commercial kits, complete with instructions for use and often analysis. For example, for human identification, the GlobalFiler™PCR Amplification Kit (Thermo-Fisher Scientific, MA, USA) and the ForenSeq DNA Signature Prep Kit (Verogen Inc., CA, USA) may be used. Each purchased kit is often accompanied by documentation for method use, data processing and results interpretation, with some kits requiring specific processing software (ForenSeq DNA Signature Prep Kit, ForenSeq Universal Analysis Software, Verogen Inc., CA, USA). The implementation of DNA-based methods is directed by the adherence to specific guidelines agreed upon by the forensic community. Such methods include those outlined by the Scientific Working Group on DNA Analysis Methods (SWGDAM) [21] and are subject to requirements of relevant accrediting bodies and/or standards, as well as in-house validation testing tailored to each laboratory. Guidelines such as SWGDAM [21] have been adjusted in recent years to include points for developmental validation specific to sequence-based approaches. As yet, there is little guidance available for the development and implementation of non-human DNA sequence-based approaches, aside from recommendations of what such criteria should be and repeated calls for the need of such guidelines [22–24].

The ability to accurately predict the origin of an environmental sample, such as soil or dust, has clear advantages as an application for forensic science and biosecurity. There are many examples of large-scale environmental biology metabarcoding studies [25–27], with a growing number focused on objectives relevant to forensic science [8,28,29]. Very few studies have focused on the establishment of standard protocols on a smaller scale for forensic applications. This may be, in part, due to several aspects that are particularly problematic when translating such approaches to a forensic context, including but not limited to the need for bioinformatics processing capabilities and application of an appropriate model to address test sample hypotheses. To build on previous work using fungal sequences obtained from dust to perform geolocation across the continental United States [28], and to contribute to this emerging field of forensic biology, work was undertaken to explore geolocation on a global scale [30], and to develop an accompanying standardized operating procedure (SOP) for method application. Previously, a reference dataset was generated by the collection of dust samples from 35 countries, analyzed for fungal sequences and used to train a predictive DeepSpace model [30]. To apply the resulting prediction model to subsequent test samples, a forensic SOP was developed and tested. To facilitate this, a smaller scale laboratory processing method was trialed, as well as the development of a customized bioinformatics pipeline, to curate and prepare resulting sequence data for model implementation. To appropriately demonstrate this method, swab samples retained from the global dust collection [30] were processed to demonstrate the use and investigate the capability of the final protocol and bioinformatics, as would be applied by the end user to prepare samples for geolocation prediction. Here we present the results of this standardized method development and evaluate the application of 76 test samples of varying origins on model performance. This work demonstrates a reproducible methodology that generates consistent results using a sequence-based metabarcoding approach, a key step in the implementation of these research methods for forensic science applications.

2. Materials and methods

2.1. Swab samples and DNA extraction

Dust swabs were collected as part of a larger study from 35 different countries using Bode SecurSwab dual-headed swabs (Bode Technology, VA, USA). Ten to 15 swabs were collected from each country sampled. Sample collection was completed by different individuals depending on the location by following a simple prescribed protocol of swabbing outdoor areas where dust naturally accumulated. Sampling requirements were to find a location outdoors (e.g. window sills, fence posts etc.) and ensure that the area was dry at the time of collection. Sample collection involved briefly swabbing the selected area using each of the dual swab heads. Swabs were stored and shipped at room temperature and frozen at −20 °C upon laboratory receipt prior to processing. The majority of the swabs collected underwent processing for prediction model training and development [30], however several swabs (3–5 depending on the original number submitted) were retained from each country for testing purposes following method and model training as conducted herein. For the purposes of the present study, a pair of the retained swabs from each of 12 different randomly selected countries from within the global dataset were processed for demonstration testing, along with a single swab from each of an additional 7 countries, resulting in swabs being included from 19 different countries (N: 31). All of the swabs processed for this study were retained and not used in method development or model training, but were collected from the same countries and at the same time as swabs that were included in model training.

Evaluation of mixtures was a required component of the larger study to ascertain whether samples could be identified as such. Mixed samples were included in this study and processed in the same way as single source samples. Mixed samples were set up using pooled ratios of extracted DNA from different samples rather than specific quantitation of DNA concentrations for several reasons. Dust samples are low biomass and so were often below detectable limits of standard DNA quantitation measures. Additionally, general dsDNA quantitation would measure DNA from non-target organisms as well as target material, and therefore would mislead target proportions in sample set up. While a target-specific quantitation method could have been used, the target region is a multi-copy gene, therefore resulting quantities would not be representative of the proportion of fungal starting material within each dust sample. Three different groups of mixtures (composed of volume mixtures of DNA extracts per country) were set up for testing. These included two- and three-country mixtures, as well as random mixtures based on shared contextual similarities which may share similarities in biota such as latitude, hemisphere and continent (consisting of two or three countries per mixture depending on context). Two-country mixtures were set up at five different proportion settings (1:20, 1:10, 1:5, 1:2, and 1:1) using two different pairs of randomly selected countries (A and B, Turkey

and Georgia respectively; C and D, Costa Rica and Trinidad and Tobago respectively) with proportions set up in duplicate, allowing samples representing each country of the pair as the major and minor contributor at all proportions (four countries total, A–D, with each pair of countries tested in a variety of proportional samples, N: 18). Three-country mixtures were set up at equal proportion ratios (1:1:1) of randomly selected countries identified as EFG (Cyprus, Pakistan and Macedonia respectively), FGH (Pakistan, Macedonia and Oman respectively), GHI (Macedonia, Oman and Hungary respectively) and HIJ (Oman, Hungary and Nigeria respectively) (N: 4). Loosely themed context mixtures were all set up as equal proportion ratios and included mixtures of countries of similar latitude (Colombia and Malaysia), the Southern Hemisphere (South Africa, New Zealand and Australia), South East Asia (South Korea, Vietnam and Malaysia), Asia (Kazakhstan, Pakistan and Georgia), Africa (Ghana, Nigeria and South Africa), Europe (Czechia, Hungary and Macedonia), the Middle East (Jordan, Qatar and Oman), and America (Mexico, Costa Rica and Colombia) (N: 8).

In addition, swabs and electrical tape samples were collected from the USA (N: 12, 6 tape and 6 swab samples) and Panama (N: 3) to act as outgroup samples (N: 15). Samples from the USA and samples collected from electrical tape were not sample types used in model training or development. Panama was also not included in model training and additionally the Panama swabs were collected from indoor rather than outdoor areas. These samples were included to evaluate model performance on samples from regions not used in model training. They were included and processed as ordinary test samples to replicate a blind-testing scenario, where the applicability of test samples is unknown beforehand (i.e. test samples from truly unknown origins).

Positive control swabs were developed to appropriately monitor swab processing and analysis. Given the low biomass found within dust generally, and therefore subsequent low DNA yields following DNA extraction, the inclusion of a positive control to monitor processing was deemed appropriate. Species of fungi were selected as positive control material development based on several key factors including: environmental ubiquity, non-pathogenic status, and representing genera previously detected in dust samples using a cultivation-independent metabarcoding method [31]. Briefly, strains of *Cladosporium inversicolor* and *Rhodotorula toruloides* were ordered from the American Type Culture Collection (ATCC) which were rehydrated following the manufacturers protocol. Equal volumes of hydrated material was pipetted onto sterile swab heads and left to dry overnight at room temperature. Positive control swabs produced high sequence counts pre- and post-sequence data analysis pipeline, and a positive control swab batch was made as described above. Positive control swabs were stored at −20 °C with the inclusion of one swab per each sample processing batch.

DNA was extracted from the swab heads and cut section of tape (approximately 1 cm^2, which was then cut into smaller pieces using a sterile scalpel) using the DNeasy PowerSoil Kit (Qiagen, Hilden, Germany) following the manufacturers protocol. Samples were processed in five small batches (approximately twelve swabs at a time), with each batch including an extraction negative (reagent blank, ENEG) and a positive control. Resulting DNA extracts were not quantified given low DNA yields recovered from dust and were all submitted for target amplification and downstream processing.

2.2. Target amplification and DNA sequencing

To analyze the fungal component of the dust samples, the internal transcribed spacer (ITS1) region of the rRNA operon was amplified using fungal specific barcoding primers [27,31]. These primers generate sequence-ready amplicons and along with primer sequence, include appropriate adaptor sequences compatible with the Illumina sequencing system. The reverse primer also included a unique 12 bp Golay auto-correcting bar code (in the same format as used elsewhere [27,32]) to allow for individual sample identification without employing additional indexes prior to sequencing. Amplification was performed in triplicate reactions per sample in 96-well plate format using an ABI Veriti thermocycler (Applied Biosystems by ThermoFisher Scientific, MA, USA). Triplicate amplification followed the recommendations of the Earth Microbiome Project [27] for low biomass samples. Thermocycling conditions consisted of an initial denaturation of 94 °C for 3 min, followed by 35 cycles of 94 °C for 45 s, 50 °C for 1 min, 72 °C for 1.5 min and a final extension step at 72 °C for 10 min. Each 25 μL reaction was composed of 12.5 μL Promega GoTaq Hot Start Colorless Master Mix (Promega, WI, USA), 4 μL of primer mix (forward and reverse primers at 2.5 μM), 1 μL of DNA template and 7.5 μL of ddH_20.

In order to include all extraction negatives (ENEGs) relevant to sample batches being tested, reaction input for each ENEG well included an equal volume mixture of template from two extraction negatives, with the exception of ENEG3 which was just template from a single extraction negative. The resulting product of each reaction was visualized using 2.2 % agarose gels on the FlashGel System (Lonza Rockland Inc., ME, USA) to ensure amplification. The remaining amplified product from each replicate reaction was pooled per sample. The amplified product was then purified by applying 25 μL of pooled amplified material to the SequalPrep™ Normalization Plate Kit (Thermo Fisher Scientific, MA, USA) following the manufacturer's instructions, resulting in 20 μL purified product, normalized to approximately 2 nM. Equal volumes of each of the purified normalized amplicons (96 samples in total, including controls) were pooled together for sequencing.

The library pool was quantified in triplicate using the Qubit 3.0 (Thermo Fisher Scientific, MA, USA), with the average result used to represent the sample pool. Molarity was calculated using an estimate of average library fragment length based on the amplicon sizes of the two positive control strains, while accounting for primer lengths. Sample pool was then submitted to the Genomic Sciences Laboratory (NCSU core sequencing facility) for sequencing on an Illumina MiSeq using a MiSeq Reagent Kit v2 (500 cycles) (Illumina, CA, USA) performed as a 2 × 250 bp paired end run. Final library quantity loaded was 7 pmol with a 15 % phiX spike in. Resulting sequence data consisted of a single undetermined R1, R2 and I1 fastq file for downstream processing.

2.3. Sequence data analysis (pipeline)

Sequence data analysis was performed using a bioinformatics pipeline customized for reproducible ITS1 sequence data analysis. This pipeline culminates in the generation of sample sequence data into a suitable format for prediction model application, as well as the more standard metabarcoding output files such as representative sequence and OTU table equivalent. This pipeline was designed to be employed independent of internet or server access on a standalone computer. The pipeline is performed entirely within R studio [33], in two steps. The first, using QIIME2 [34,40] via the terminal window (facilitated via Windows Subsystem for Linux) and the second, using R software [35] operated via an R Markdown document [36]. Sequence files are demultiplexed, trimmed and separated into final fastq files per sample using QIIME2 [34,40].

The first section of the R markdown [36] document instructs the user which script to copy and paste into the terminal window to perform the QIIME2 [34,40] portion of the pipeline. The commands used are also added to the R markdown file in order to

document the commands used to facilitate each step. Following the completion of step one, the user is instructed to make specific changes within the remainder of the R markdown document in regards to file names and other specifics pertinent to each run and then the R markdown document is saved and step two is initiated, processing the R markdown document by selecting "Knit". Processing the R markdown document results in performance of all of the commands embedded within the document using the files prepared through step one. Once complete, several output files are generated including the final data file for model implementation (similar to an OTU table) as well as a text document capturing the entire run (including time stamp and other processing tracking information) automatically generated upon run completion, detailing the R markdown document in its entirety, thereby documenting both the instructions for processing as well as what was completed for step one. Data processing steps, in particular those within step two, follow recommendations and tutorials for generating amplicon sequence variants (ASVs) using the DADA2 R software package [37]. Several data quality assessment and processing steps are performed within the R markdown document including filtering for quality (using settings for maximum expected error rates per forward and reverse strands, and limiting sequences containing 'N' nucleotides), with error learning applied to both forward and reverse strands [37]. Sequences are then dereplicated, and 'denoised' using the learned error rate, and then merged [37]. Chimeric sequences are removed and a table detailing unique sequence strings (ASVs) and counts per sample is generated (similar to an OTU table). These unique sequences are then compared against the ASV fasta comparison file that was compiled from the set of global swabs (35 different countries sampled, totaling 487 samples for model development) as originally used to train the prediction model. This assigns identifiers to unique sequences found in test sample batches that match back to sequences observed in the global comparison file. Once this assignment has been made, a final sample table is generated using the sequence identifiers that correlate with the global comparison set and the counts of those sequences per sample, which is then applied to the model for country of origin prediction.

2.4. Model prediction

The prediction model was developed using the final set of global data, consisting of 487 swabs from 35 different countries and built upon earlier work by Grantham and colleagues [30]. A spatially-aware deep learning model was developed and employed for sample country of origin prediction. Briefly, this model applies a random partitioning system across the earth's surface and employs a deep learning model to predict, via probability assignment, which of these partitions a sample is likely to have originated from. The final predictions are based on the average over many random partitions. The 'best' prediction for the present study is taken as the top prediction, the highest probability across all the partitions for a given sample. The specifics of the model are beyond the scope of this manuscript, for greater detail please refer to [30]. The resulting model was trained on ITS1 sequence data and applied to subsequent samples through a customized graphical user interface software, dubbed ASVtracer, allowing for country of origin prediction and accompanying prediction visualizations. Results generated through ASVtracer include global location graphic files as well as probabilities listed in a text file. Demonstration samples were prepared as described above and applied as anonymous samples by a second user to the ASVtracer for blind model implementation. Samples were required to have generated $\geq$3000 sequence counts assigned to ASV taxa within the comparison global reference file to be eligible for prediction via implementation to the model software. Resulting outputs were assessed by interpreting the top three probabilities generated per sample, which includes coordinates and corresponding country per prediction. A prediction was deemed accurate if the top three predicted countries matched the known country of origin/collection. A prediction was deemed a near-miss if the top three predicted countries were geographically neighboring countries to the known country of origin/collection and deemed incorrect if the known country of origin/collection was not among the top three predictions.

3. Results

3.1. Samples and data processing

Each DNA extraction underwent ITS1 amplification in triplicate as described, with resulting amplicons per reaction visualized using gel electrophoresis. Sample triplicate reactions were pooled and purified regardless of whether visible amplicons were observed during gel electrophoresis given the low biomass of dust samples. Extraction negatives (ENEGs) and no template controls (NTCs) were sequenced with some showing minimal read counts ($\leq$118 and $\leq$68 reads respectively) and zero read counts following entire pipeline processing. It is therefore assumed that the majority of these original reads were primer dimers, chimeric product or other extraneous amplicons. Positive controls were sequenced and showed >100, 000 reads across all three controls sequenced. The two ASV sequences that were present in positive control samples (in greater than 15 read counts) were matched back to the expected species for *R. toruloides*, and matching to several species within the *Cladosporium* genus.

3.2. Single country swab predictions

Previously unprocessed swabs collected from 12 countries that were included in the prediction model training dataset were randomly selected and analyzed to demonstrate SOP performance of the laboratory methods and sequence data analysis pipeline. Additionally, these samples were used to assess the performance of the prediction model and GUI software when applied to subsequent test samples obtained from countries used to train the model. The samples were applied to the prediction model in an anonymous way, therefore the model user did not know which country the resulting data file per sample originated from. Two swabs were used per country resulting in 24 samples processed (Table 1). Two samples did not generate enough final sequence counts to be applied to the model (Country 8, Kazakhstan and Country 12, Cyprus). Of the remaining 22 samples, 19 resulted in a predicted country that was the correct country of origin, while three samples resulted in predictions that did not match their known country of origin. Incorrectly predicted samples were both samples from Qatar (Country 6), predicted as Bahrain, and one sample from New Zealand (Country 1), predicted as Uruguay. In addition to these samples, single swabs were processed from seven additional countries where material was used to make up the artificial mixture, in order to have single contribution comparison data from the same per country material. All of these samples generated sufficient sequence data to be applied to the model, with five samples correctly predicted to their country of origin (Table 1). The two incorrectly predicted samples were from Oman (Country H), predicted as Bahrain, and Nigeria (Country J), predicted as Ghana. Therefore of the entire set of single country swabs tested (N: 31, 12 countries with duplicate swabs and seven countries with a single swab), discarding the two samples that did not generate sufficient sequence data for model application, overall prediction

Table 1
Single country swab sample details and results, including identifier (ID), original collection country, sequence counts (as final output results total and of the total, those that could be assigned to sequences within the comparative global database), prediction and result. This includes single country swabs analyzed for DNA extracts used in mixtures (indicated by letters in ID column). * Indicates DNA extracts that were processed as single country test samples and additionally also used as components in mixed samples.

ID:	True Country:	Sequence Counts		Predicted Country:	Result:
		Total:	Global:		
Country 1	New Zealand	79888	54781	New Zealand	Correct
Country 1	New Zealand	95085	23279	Uruguay	Incorrect
Country 2	Czechia	57433	53749	Czechia	Correct
Country 2	Czechia	55922	23131	Czechia	Correct
Country 3	Mexico	65720	39791	Mexico	Correct
Country 3	Mexico	80293	49714	Mexico	Correct
Country 4	Jordan	64580	52905	Jordan	Correct
Country 4	Jordan	99962	81039	Jordan	Correct
Country 5	Australia	71349	59489	Australia	Correct
Country 5	Australia	98947	30912	Australia	Correct
Country 6	Qatar	54515	33556	Bahrain	Incorrect
Country 6	Qatar	89684	3609	Bahrain	Incorrect
Country 7*	Istanbul, Turkey	64925	28835	Turkey	Correct
Country 7	Istanbul, Turkey	70597	53138	Turkey	Correct
Country 8	Kazakhstan	1409	1409	–	–
Country 8	Kazakhstan	63955	42306	Kazakhstan	Correct
Country 9*	Trinidad & Tobago	169134	41420	Trinidad & Tobago	Correct
Country 9	Trinidad & Tobago	95072	70415	Trinidad & Tobago	Correct
Country 10	South Korea	21016	9486	South Korea	Correct
Country 10	South Korea	12899	6463	South Korea	Correct
Country 11	Ghana	102310	85596	Ghana	Correct
Country 11	Ghana	79098	73763	Ghana	Correct
Country 12*	Cyprus	99436	55767	Cyprus	Correct
Country 12	Cyprus	269	222	–	–
Country B	Georgia	138842	69592	Georgia	Correct
Country C	Costa Rica	68787	18656	Costa Rica	Correct
Country F	Pakistan	101400	98271	Pakistan	Correct
Country G	Macedonia	58446	36464	Macedonia	Correct
Country H	Oman	38716	29234	Bahrain	Incorrect
Country I	Hungary	78330	63183	Hungary	Correct
Country J	Nigeria	103269	40049	Ghana	Incorrect

accuracy was 82.8 % (correct predictions made for 24 of 29 samples).

3.3. Mixed country predictions

Two country mixtures were set up in five different proportion settings for two different pairs of countries, A and B (Turkey and Georgia respectively), and C and D (Costa Rica and Trinidad and Tobago respectively). The top three predicted countries were recorded per reaction per pair. For all mixed reactions for A and B, regardless of proportions or major and minor contributing country, the top three predictions were consistently Georgia (Table 2). For the majority of the mixed reactions for C and D, regardless of proportions or major and minor contributing country, the top three predictions were Trinidad and Tobago and Brazil (both second and third predictions) respectively (Table 2). The only exception to this was for a single reaction (Trinidad and Tobago 1:20 Costa Rica) where the top three predictions were instead Honduras (first and second predictions) and Trinidad and Tobago respectively. This reaction is the extreme proportion set up, allowing the maximum proportion for Costa Rica in this paired assessment. All 18 of these mixed samples of varying proportions were correctly predicted back to one of the countries within the composition. However, one component country of each pair was always predicted as the country of origin, regardless of whether it was the major or minor mixture component.

Three country mixtures were set up with extracts from a single swab from each of six different countries. Mixtures included the following combinations of countries: Cyprus, Pakistan, and Macedonia (denoted as E, F, G respectively); Pakistan, Macedonia, Oman (denoted as F, G, H respectively); Macedonia, Oman and Hungary (denoted as G, H, I respectively); and Oman, Hungary, and Nigeria (denoted as H, I, J respectively). Each combination was tested at equal composition proportions (e.g. mixed together sample at 1:1:1). The first two sets of mixtures (combinations E, F, G and F, G, H) resulted in all three top predictions of country of origin as Pakistan (a country included in the mixture) (Table 2). The second two sets of mixtures (combinations G, H, I and H, I, J) all resulted in Hungary as the top prediction (included in the mixture), with either Hungary or Croatia (not included within the mixtures) as the second and/or third best prediction (Table 2).

The contextually themed mixtures consisted of eight mixtures (a single mixture with two countries, the remainder with three) with composition based on either similar latitude (Colombia and Malaysia), hemisphere (South Africa, Australia and New Zealand) or geographic locality (South East Asia – South Korea, Vietnam, Malaysia; Asia – Kazakhstan, Pakistan, Georgia; Africa – Ghana, Nigeria, South Africa; Europe – Czechia, Hungary, Macedonia; Middle East – Jordan, Qatar, Oman; and America – Mexico, Costa Rica, Colombia). All but one of these mixtures generated accurate predictions of a country of origin to one of the countries included within the mixture as the top three predictions (Table 2). Of these accurately predicted samples, the third best prediction for the 'Europe' sample was an exception to this, with Croatia predicted (not included in the mixture), however the first and second prediction were both Hungary (included in the mixture). The sample consisting of mixed material from Middle Eastern countries was incorrectly predicted as originating from Kuwait (not included in the mixture). With the exception of the 'Europe' sample (which predicted Hungary and Croatia), the predictions for these samples did not represent more than one single country component, with each sample's top three predictions indicating the same single country, rather than top three predictions split across the true components of the mixture (Table 2).

Table 2
Artificial country mixture details and model prediction results. Country of origin predictions are as detailed in 'Best', '2nd' and '3rd' indicating the top three predictions generated when applying the results output data to the model via the ASVtracer interface.

Countries:	Mix:	Best	2nd	3rd
Turkey, Georgia	1:20	Georgia	Georgia	Georgia
Turkey, Georgia	1:10	Georgia	Georgia	Georgia
Turkey, Georgia	1:5	Georgia	Georgia	Georgia
Turkey, Georgia	1:2	Georgia	Georgia	Georgia
Turkey, Georgia	1:1	Georgia	Georgia	Georgia
Georgia, Turkey	1:20	Georgia	Georgia	Georgia
Georgia, Turkey	1:10	Georgia	Georgia	Georgia
Georgia, Turkey	1:5	Georgia	Georgia	Georgia
Georgia, Turkey	1:2	Georgia	Georgia	Georgia
Costa Rica, Trinidad & Tobago	1:20	Trinidad & Tobago	Brazil	Brazil
Costa Rica, Trinidad & Tobago	1:10	Trinidad & Tobago	Brazil	Brazil
Costa Rica, Trinidad & Tobago	1:5	Trinidad & Tobago	Brazil	Brazil
Costa Rica, Trinidad & Tobago	1:2	Trinidad & Tobago	Brazil	Brazil
Costa Rica, Trinidad & Tobago	1:1	Trinidad & Tobago	Brazil	Brazil
Trinidad & Tobago, Costa Rica	1:20	Honduras	Honduras	Trinidad & Tobago
Trinidad & Tobago, Costa Rica	1:10	Trinidad & Tobago	Brazil	Brazil
Trinidad & Tobago, Costa Rica	1:5	Trinidad & Tobago	Brazil	Brazil
Trinidad & Tobago, Costa Rica	1:2	Trinidad & Tobago	Brazil	Brazil
Cyprus, Pakistan, Macedonia	1:1:1	Pakistan	Pakistan	Pakistan
Pakistan, Macedonia, Oman	1:1:1	Pakistan	Pakistan	Pakistan
Macedonia, Oman, Hungary	1:1:1	Hungary	Croatia	Hungary
Oman, Hungary, Nigeria	1:1:1	Hungary	Hungary	Croatia
Colombia, Malaysia	1:1	Malaysia	Malaysia	Malaysia
South Africa, New Zealand, Australia	1:1:1	New Zealand	New Zealand	New Zealand
South Korea, Vietnam, Malaysia	1:1:1	Malaysia	Malaysia	Malaysia
Kazakhstan, Pakistan, Georgia	1:1:1	Pakistan	Pakistan	Pakistan
Ghana, Nigeria, South Africa	1:1:1	Ghana	Ghana	Ghana
Czechia, Hungary, Macedonia	1:1:1	Hungary	Hungary	Croatia
Jordan, Qatar, Oman	1:1:1	Kuwait	Kuwait	Kuwait
Mexico, Costa Rica, Colombia	1:1:1	Mexico	Mexico	Mexico

All of the randomly assigned three country and context themed mixtures were correctly predicted back to one of the component countries as the best country of origin prediction, with input proportion having no discernible effect.

3.4. Outgroup and positive controls

Outgroup samples consisted of tape lifts (using electrical tape) and swabs collected from two outdoor environments within the same city (Raleigh, NC) in the continental United States (US), as well as indoor swabs collected in Panama. Both the US and Panama were not included within the global dataset, nor were tape samples or dust swabs collected from indoor environments. Swabs from the US locations all showed similar top predictions of Argentina or Uruguay, and resulted in final total sequence count numbers ranging from 68, 000 to almost 104, 000. The tape samples from the same locations showed more sporadic country of origin predictions, including South Korea, Oman, Bahrain, Argentina, Djibouti and South Africa. Tape samples resulted in final sequence count numbers ranging from 31, 000 to over 120, 000, similar to that of the swabs, but with one sample not applicable for implementation into the model as a result of an insufficient number of sequences matching the global dataset. The indoor swab samples from Panama generated read counts ranging from 41, 587 to over 80, 000, with all three samples resulting in sufficient sequence count numbers matching the global set to allow for country of origin prediction. These three Panama samples resulted in top predictions of Malaysia, Somalia and Trinidad and Tobago for each swab respectively.

4. Discussion

4.1. Overall performance

This work was undertaken to test the development and feasibility of adopting a standardized operating procedure (SOP) from a large-scale metabarcoding study. The use of fungal DNA and the resulting SOP was evaluated as an investigative tool for forensic science. The resulting SOP included each key step necessary to target, sequence and analyze the fungal DNA within a dust swab. These steps were sample processing from initial swab DNA extraction, sequence generation and bioinformatics to generating sample test data suitable for implementation to a previously developed DeepSpace model [30] for country of origin prediction. Performance of the predictive model was evaluated by testing samples that had been collected and retained from countries including those that had been used to train the model, as well as samples originating from countries the model was naive to. In this way, all swabs analyzed here were collected at the same time and from the same country as swabs used to train the model, but were only used to test the model herein (i.e. one large set of swabs, split so the majority is used for model training, and a smaller subset for testing). In addition, artificial mixtures consisting of combinations of extracted material from different countries used to train the model were also used. The initial components of the SOP are based upon well-established methods of DNA extraction and target amplification of sequence ready products [27,32]. These components performed as expected. The bioinformatics pipeline processed sequence data as described using the R markdown [36] file via R studio [33]. As this file offers directions and performs data analysis, minimal end-user interaction was required and output data files were generated without issue. A small proportion of samples tested failed to generate a high enough number of sequence counts (given the threshold set). As a consequence, those samples did not pass filtering settings during data processing and/or provide enough sequences that matched the global comparative set to warrant application to the prediction model. Given the low biomass of dust samples, and the variation in yield observed sample to sample, it is not surprising that some samples would fail to generate sufficient data for model application, therefore the inability to analyze some samples used to test the SOP here was not

unexpected. When designing similar studies, it is reasonable to anticipate that an appreciable number of samples tested will fail to generate sufficient data for model implementation.

4.2. Prediction accuracy

Single country testing illustrated a high rate of prediction accuracy (greater than 82 %) and performed with similar accuracy to that observed previously using dust samples collected across the continental US [28], and as similarly demonstrated with country of origin prediction during model development [30]. In addition, where errors occurred, they tended to be related to predictions that were within the right region but the wrong country (e.g., Bahrain instead of Oman) or in the right biogeographic region and climate but wrong country (e.g., Brazil instead of Costa Rica). Multiple country mixtures generally resulted in correct predictions to one of the country components within a mixed reaction, however there was no evidence that the country predicted was influenced by the proportion of mixture contents. Whether predicting one of the countries of origin of a mixture is useful in a forensic context will depend on the details of the particular forensic case. Additionally, the top three predictions of the majority of the mixed reactions were all a single country and generally did not indicate predictions of more than one of the correct countries included in the mixture. The outgroup samples showed variable predictions, with the majority of the swabs collected from the US predicted to South American countries, although notably not Mexico, which geographically would be the closest country included in model training. Tape samples exhibited a broad mix of predicted origins with no observable pattern, as did the indoor swabs from Panama.

4.3. Forensic aspects of processing

Translating environmental DNA sequence-based approaches from a research setting to implementation within a forensic science setting presents numerous challenges. A method must be accurate, reproducible and precise, among other requirements, when operating within established performance parameters, in order to be considered suitable for use as an investigative tool. Several key aspects of the present work have been tailored towards addressing these requirements, including the use of sequence data as amplicon sequence variants (ASVs), a constrained bioinformatics pipeline, and the evaluation and establishment of processing controls.

The use of ASVs in metabarcoding studies instead of the often used OTUs is becoming a more frequent choice for sequence data analysis (as discussed elsewhere [38,39]). The generation of each set of OTUs is contextually dependent upon the dataset that is run, and thus can vary in exact data outputs run to run unless all previous data is reanalyzed with new data [38]. With ASVs exact sequence strings of information are used as data units, which can then be detected and recorded in the same manner across different processing runs, as well as matched and counted exactly across different datasets. This also allows for exact comparison back to a curated reference database (as used here), or could also be compared against lists of exact taxa of interest for screening or surveillance projects such as may be applied in a biosecurity context.

Another aspect critical to uniform data generation and ease of use of an investigative tool built upon a metabarcoding approach is a user-friendly bioinformatics pipeline for sequence data analysis. Other key aspects in a forensic setting may also include implementation of bioinformatics processing in an off-network environment or on an independent computer. The bioinformatics pipeline presented herein allows the undertaking of sequence data processing entirely within the freely available R Studio software [35], which many users may already be familiar with, employing R software packages commonly used for metabarcoding analysis (such as DADA2 [37]). The pipeline is alternately directed and facilitated by the use of an R Markdown [36] document that directs the first portion of sequence data processing, while actively processing the second portion of the pipeline. This allows for sequence data processing with minimal end-user intervention, with no adjustment required for any key data curation steps, and culminates in an ASV data file ready for model implementation. Additionally, the use of the R markdown document provides other project tracking advantages such as time stamping data processing, as well as automatic generation of a run report (as a word document) along with the generation of other pertinent data files for record keeping. In this context, each run of sequencing data is processed in the same manner, regardless of different operators or network connectively, and each step of bioinformatics processing is detailed and saved.

The inclusion of processing controls is considered best practice in a laboratory setting and are a useful way to monitor correct undertaking of many different DNA-based methods. The use of a positive control in particular is advantageous for detection-based testing and/or when testing samples of low yield. As well as including and evaluating the results of extraction negatives (or reagent blanks) from DNA extraction through to sequence data processing (discussed above), positive control swabs were developed and included throughout sample testing. The use of positive control swabs is a recommended part of the SOP, proving useful when extracting materials of such low biomass, to ensure that all laboratory processing steps were undertaken correctly prior to sequencing, with resulting sequence data additionally acting as a control for bioinformatics pipeline processing. The use of positive and negative controls has been advised repeatedly when applying metabarcoding approaches to forensic settings [22,23]. The positive control developed herein is simple to set up and effective, and while designed for the present work, should be broadly applicable to other studies targeting fungal material.

4.4. Limitations and future work

The SOP offers a novel way to process environmental DNA samples in the form of dust swabs, from DNA extraction through to bioinformatics processing and data output generation. The use of exact sequence variants, a constrained bioinformatics pipeline to reduce user subjectivity and run to run variation, and culminating in the generation of processing documents detailing methods used, are all beneficial to the application of a metabarcoding method in a forensic context. The design and inclusion of a positive control is also beneficial and aligns with best practice for forensic processing [21]. However, through the course of SOP development and application of additional country samples, artificial country mixtures and outgroup samples has revealed several limitations and aspects requiring further study. The predictive model was accurate when applied to single country swabs originating from countries used in model training and development. However, there is a clear requirement for an interpretation threshold for model prediction results. The model will always generate a suite of predictions (as long as a sample presents sufficient sequence counts), without any measure of confidence for interpretation or acceptable probability range (i.e. accept probabilities above a certain level, while disregarding others below). In this way, a prediction is always made, whether truly likely or not, and as evidenced by the application of outgroup samples the model had not 'seen' before, leads to incorrect country of origin predictions. An interpretation threshold for best prediction probabilities must be established to fully explore the accuracy of the geolocation

model. In the absence of such thresholds for prediction interpretation, the application of artificial mixtures premature. There was nothing obvious in the model results to indicate that these mixed samples were interpreted differently to single contributor samples within the model, or that could allow the identification of such samples by the user from the model output. In short, mixed samples could not be identified as such. This is likely to have been further compounded by the variable yield of biological material across dust swabs as well as the multi-copy nature of ITS1. Additionally, it is worth noting that these aspects of low biomass samples of varying yield combined with a multicopy genetic target is why the positive control sample should only be interpreted qualitatively and further demonstrates why a quantitative application here would not be truly replicative of test samples nor an appropriate control.

Although beyond the SOP evaluation here, other points of future work to improve the model training would be the investigation of intra-sample variability when collected at the same point and time; sampling across different areas, ecosystems and seasons within a county; and investigating the role that rare or unique biota and taxon-specific read counts have on prediction accuracy.

The bioinformatics process is relatively user friendly and reduces sequence data analysis subjectivity, however it still requires some end-user intervention for processing set up and some familiarity with R Studio is required. While the use of an R Markdown document displays each processing step to the user and therefore is not a 'black box' approach to data analysis, the user does not have to be cognizant of all aspects of processing due to the automatic processing of the R Markdown method. Therefore, the onus would be upon the end user to ensure familiarity and understanding with each aspect of processing for defensible reporting. Settings within the pipeline remain unadjusted run to run. This assurance of processing continuity is a conservative approach, rather than a focus upon maximizing data recovery through settings adjusted run to run by the end user.

5. Conclusions

A standard operation procedure was developed to allow the application of environmental dust samples as test samples to a previously developed geolocation model for country of origin prediction. The SOP included implementation of a customized bioinformatics processing pipeline to facilitate the analysis and preparation of metabarcoding sequence data for model implementation. Use of this SOP to process subsequently applied test samples revealed that the majority were predicted to true country of origin with high accuracy, however incorrect predictions were made when applying samples from countries not used for model development. Additionally artificially mixed samples were generally predicted accurately to one of the component countries, but could not be identified as mixtures. Collectively SOP evaluation testing demonstrated key aspects of sample processing and method development relevant to the use of a geolocation method in a forensic context, including aspects that require further development such as threshold establishment and interpretation guidelines.

Declaration of Competing Interest

None.

Acknowledgements

This work was supported by the U.S. Army Research Office and the Defense Forensic Science Center (DFSC) and was accomplished under Cooperative Agreement Number W911NF-16-2-0195. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, DFSC, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.forsciint.2020.110250.

References

[1] B.H. Kaye, Dust chapter, Science and the Detective: Selected Reading in Forensic Science, John Wiley & Sons, New York, 2008.
[2] E. Locard, The analysis of dust traces. Part I, Am. J. Police Sci. 1 (1930) 276–298.
[3] C. Ladd, H.C. Lee, The use of biological and botanical evidence in criminal investigations, The Use of Biological and Botanical Evidence in Criminal Investigations, (2005) .
[4] P.E.J. Wiltshire, Mycology in palaeoecology and forensic science, Fungal Biol. 120 (2016) 1272–1290, doi:http://dx.doi.org/10.1016/j.funbio.2016.07.005.
[5] A.G. Brown, The use of forensic botany and geology in war crimes investigations in NE Bosnia, Forensic Sci. Int. 163 (2006) 204–210, doi:http://dx.doi.org/10.1016/j.forsciint.2006.05.025.
[6] D.L. Hawksworth, P. Wiltshire, Forensic mycology: current perspectives, Res. Rep. Forensic Med. Sci. (2015) 75, doi:http://dx.doi.org/10.2147/RRFMS.S83169.
[7] P.E.J. Wiltshire, D.L. Hawksworth, J.A. Webb, K.J. Edwards, Two sources and two kinds of trace evidence: enhancing the links between clothing, footwear and crime scene, Forensic Sci. Int. 254 (2015) 231–242, doi:http://dx.doi.org/10.1016/j.forsciint.2015.05.033.
[8] C. Fløjgaard, T.G. Frøslev, A.K. Brunbjerg, H.H. Bruun, J. Moeslund, A.J. Hansen, R. Ejrnæs, Predicting provenance of forensic soil samples: linking soil to ecological habitats by metabarcoding and supervised classification, PLoS One 14 (2019)e0202844, doi:http://dx.doi.org/10.1371/journal.pone.0202844.
[9] S. Giampaoli, A. Berti, R.M. Di Maggio, E. Pilli, A. Valentini, F. Valeriani, G. Gianfranceschi, F. Barni, L. Ripani, V. Romano Spica, The environmental biological signature: NGS profiling for forensic comparison of soils, Forensic Sci. Int. 240 (2014) 41–47, doi:http://dx.doi.org/10.1016/j.forsciint.2014.02.028.
[10] E.M. Jesmok, J.M. Hopkins, D.R. Foran, Next-generation sequencing of the bacterial 16S rRNA gene for forensic soil comparison: a feasibility study, J. Forensic Sci. 61 (2016) 607–617, doi:http://dx.doi.org/10.1111/1556-4029.13049.
[11] A. Belk, Z.Z. Xu, D.O. Carter, A. Lynne, S. Bucheli, R. Knight, J. Metcalf, Microbiome data accurately predicts the postmortem interval using random forest regression models, Genes 9 (2018) 104, doi:http://dx.doi.org/10.3390/genes9020104.
[12] K.L. Cobaugh, S.M. Schaeffer, J.M. DeBruyn, Functional and structural succession of soil microbial communities below decomposing human cadavers, PLoS One 10 (2015)e0130201, doi:http://dx.doi.org/10.1371/journal.pone.0130201.
[13] H.R. Johnson, D.D. Trinidad, S. Guzman, Z. Khan, J.V. Parziale, J.M. DeBruyn, N.H. Lents, A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval, PLoS One 11 (2016)e0167370, doi:http://dx.doi.org/10.1371/journal.pone.0167370.
[14] J.L. Metcalf, L. Wegener Parfrey, A. Gonzalez, C.L. Lauber, D. Knights, G. Ackermann, G.C. Humphrey, M.J. Gebert, W. Van Treuren, D. Berg-Lyons, K. Keepers, Y. Guo, J. Bullard, N. Fierer, D.O. Carter, R. Knight, A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system, ELife 2 (2013), doi:http://dx.doi.org/10.7554/eLife.01104.
[15] J.L. Metcalf, Z.Z. Xu, S. Weiss, S. Lax, W. Van Treuren, E.R. Hyde, S.J. Song, A. Amir, P. Larsen, N. Sangwan, D. Haarmann, G.C. Humphrey, G. Ackermann, L.R. Thompson, C. Lauber, A. Bibat, C. Nicholas, M.J. Gebert, J.F. Petrosino, S.C. Reed, J.A. Gilbert, A.M. Lynne, S.R. Bucheli, D.O. Carter, R. Knight, Microbial community assembly and metabolic function during mammalian corpse decomposition, Science 351 (2016) 158–162, doi:http://dx.doi.org/10.1126/science.aad2646.
[16] J.L. Pechal, T.L. Crippen, M.E. Benbow, A.M. Tarone, S. Dowd, J.K. Tomberlin, The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing, Int. J. Legal Med. 128 (2014) 193–205, doi:http://dx.doi.org/10.1007/s00414-013-0872-1.
[17] N. Fierer, C.L. Lauber, N. Zhou, D. McDonald, E.K. Costello, R. Knight, Forensic identification using skin bacterial communities, Proc. Natl. Acad. Sci. 107 (2010) 6477–6481, doi:http://dx.doi.org/10.1073/pnas.1000162107.
[18] S.E. Schmedes, A.E. Woerner, B. Budowle, Forensic human identification using skin microbiomes, Appl. Environ. Microbiol. 83 (2017), doi:http://dx.doi.org/10.1128/AEM.01672-17.
[19] S.E. Schmedes, A.E. Woerner, N.M.M. Novroski, F.R. Wendt, J.L. King, K.M. Stephens, B. Budowle, Targeted sequencing of clade-specific markers from skin

microbiomes for forensic human identification, Forensic Sci. Int. Genet. 32 (2018) 50–61, doi:http://dx.doi.org/10.1016/j.fsigen.2017.10.004.
[20] A.E. Woerner, N.M.M. Novroski, F.R. Wendt, A. Ambers, R. Wiley, S.E. Schmedes, B. Budowle, Forensic human identification with targeted microbiome markers using nearest neighbor classification, Forensic Sci. Int. Genet. 38 (2019) 130–139, doi:http://dx.doi.org/10.1016/j.fsigen.2018.10.003.
[21] S.W.G. on D.A.M. Methods, Validation Guidelines for DNA Analysis Methods, (2016) .
[22] B. Budowle, Quality assurance guidelines for laboratories performing microbial forensic work, Forensic Sci. Commun. 5 (2003).
[23] B. Budowle, N.D. Connell, A. Bielecka-Oder, R.R. Colwell, C.R. Corbett, J. Fletcher, M. Forsman, D.R. Kadavy, A. Markotic, S.A. Morse, R.S. Murch, A. Sajantila, S.E. Schmedes, K.L. Ternus, S.D. Turner, S. Minot, Validation of high throughput sequencing and microbial forensics applications, Invest. Genet. 5 (2014) 9, doi:http://dx.doi.org/10.1186/2041-2223-5-9.
[24] S.A. Hardwick, I.W. Deveson, T.R. Mercer, Reference standards for next-generation sequencing, Nat. Rev. Genet. 18 (2017) 473–484, doi:http://dx.doi.org/10.1038/nrg.2017.44.
[25] A. Barberán, J. Ladau, J.W. Leff, K.S. Pollard, H.L. Menninger, R.R. Dunn, N. Fierer, Continental-scale distributions of dust-associated bacteria and fungi, Proc. Natl. Acad. Sci. 112 (2015) 5756–5761, doi:http://dx.doi.org/10.1073/pnas.1420815112.
[26] M. Delgado-Baquerizo, A.M. Oliverio, T.E. Brewer, A. Benavent-González, D.J. Eldridge, R.D. Bardgett, F.T. Maestre, B.K. Singh, N. Fierer, A global atlas of the dominant bacteria found in soil, Science 359 (2018) 320–325, doi:http://dx.doi.org/10.1126/science.aap9516.
[27] L.R. Thompson, J.G. Sanders, D. McDonald, A. Amir, J. Ladau, K.J. Locey, R.J. Prill, A. Tripathi, S.M. Gibbons, G. Ackermann, J.A. Navas-Molina, S. Janssen, E. Kopylova, Y. Vázquez-Baeza, A. González, J.T. Morton, S. Mirarab, Z. Zech Xu, L. Jiang, M.F. Haroon, J. Kanbar, Q. Zhu, S. Jin Song, T. Kosciolek, N.A. Bokulich, J. Lefler, C.J. Brislawn, G. Humphrey, S.M. Owens, J. Hampton-Marcell, D. Berg-Lyons, V. McKenzie, N. Fierer, J.A. Fuhrman, A. Clauset, R.L. Stevens, A. Shade, K. S. Pollard, K.D. Goodwin, J.K. Jansson, J.A. Gilbert, R. Knight, J.L.A. Rivera, L. Al-Moosawi, J. Alverdy, K.R. Amato, J. Andras, L.T. Angenent, D.A. Antonopoulos, A. Apprill, D. Armitage, K. Ballantine, J. Bárta, J.K. Baum, A. Berry, A. Bhatnagar, M. Bhatnagar, J.F. Biddle, L. Bittner, B. Boldgiv, E. Bottos, D.M. Boyer, J. Braun, W. Brazelton, F.Q. Brearley, A.H. Campbell, J.G. Caporaso, C. Cardona, J. Carroll, S.C. Cary, B.B. Casper, T.C. Charles, H. Chu, D.C. Claar, R.G. Clark, J.B. Clayton, J.C. Clemente, A. Cochran, M.L. Coleman, G. Collins, R.R. Colwell, M. Contreras, B.B. Crary, S. Creer, D.A. Cristol, B.C. Crump, D. Cui, S.E. Daly, L. Davalos, R.D. Dawson, J. Defazio, F. Delsuc, H.M. Dionisi, M.G. Dominguez-Bello, R. Dowell, E. A. Dubinsky, P.O. Dunn, D. Ercolini, R.E. Espinoza, V. Ezenwa, N. Fenner, H.S. Findlay, I.D. Fleming, V. Fogliano, A. Forsman, C. Freeman, E.S. Friedman, G. Galindo, L. Garcia, M.A. Garcia-Amado, D. Garshelis, R.B. Gasser, G. Gerdts, M.K. Gibson, I. Gifford, R.T. Gill, T. Giray, A. Gittel, P. Golyshin, D. Gong, H.-P. Grossart, K. Guyton, S.-J. Haig, V. Hale, R.S. Hall, S.J. Hallam, K.M. Handley, N.A. Hasan, S. R. Haydon, J.E. Hickman, G. Hidalgo, K.S. Hofmockel, J. Hooker, S. Hulth, J. Hultman, E. Hyde, J.D. Ibáñez-Álamo, J.D. Jastrow, A.R. Jex, L.S. Johnson, E.R. Johnston, S. Joseph, S.D. Jurburg, D. Jurelevicius, A. Karlsson, R. Karlsson, S. Kauppinen, C.T.E. Kellogg, S.J. Kennedy, L.J. Kerkhof, G.M. King, G.W. Kling, A.V. Koehler, M. Krezalek, J. Kueneman, R. Lamendella, E.M. Landon, K. Lane-deGraaf, J. LaRoche, P. Larsen, B. Laverock, S. Lax, M. Lentino, I.I. Levin, P. Liancourt, W. Liang, A.M. Linz, D.A. Lipson, Y. Liu, M.E. Lladser, M. Lozada, C.M. Spirito, W.P. MacCormack, A. MacRae-Crerar, M. Magris, A.M. Martín-Platero, M. Martín-Vivaldi, L.M. Martínez, M. Martínez-Bueno, E.M. Marzinelli, O.U. Mason, G.D. Mayer, J.M. McDevitt-Irwin, J.E. McDonald, K.L. McGuire, K.D. McMahon, R. McMinds, M. Medina, J.R. Mendelson, J.L. Metcalf, F. Meyer, F. Michelangeli, K. Miller, D.A. Mills, J. Minich, S. Mocali, L. Moitinho-Silva, A. Moore, R.M. Morgan-Kiss, P. Munroe, D. Myrold, J.D. Neufeld, Y. Ni, G.W. Nicol, S. Nielsen, J.I. Nissimov, K. Niu, M.J. Nolan, K. Noyce, S.L. O'Brien, N. Okamoto, L. Orlando, Y.O. Castellano, O. Osuolale, W. Oswald, J. Parnell, J.M. Peralta-Sánchez, P. Petraitis, C. Pfister, E. Pilon-Smits, P. Piombino, S.B. Pointing, F.J. Pollock, C. Potter, B. Prithiviraj, C. Quince, A. Rani, R. Ranjan, S. Rao, A.P. Rees, M. Richardson, U. Riebesell, C. Robinson, K.J. Rockne, S.M. Rodriguezl, F. Rohwer, W. Roundstone, R.J. Safran, N. Sangwan, V. Sanz, M. Schrenk, M.D. Schrenzel, N. M. Scott, R.L. Seger, A. Seguin-Orlando, L. Seldin, L.M. Seyler, B. Shakhsheer, G. M. Sheets, C. Shen, Y. Shi, H. Shin, B.D. Shogan, D. Shutler, J. Siegel, S. Simmons, S. Sjöling, D.P. Smith, J.J. Soler, M. Sperling, P.D. Steinberg, B. Stephens, M.A. Stevens, S. Taghavi, V. Tai, K. Tait, C.L. Tan, N. Taş, D.L. Taylor, T. Thomas, I. Timling, B.L. Turner, T. Urich, L.K. Ursell, D. van der Lelie, W. Van Treuren, L. van Zwieten, D. Vargas-Robles, R.V. Thurber, P. Vitaglione, D.A. Walker, W.A. Walters, S. Wang, T. Wang, T. Weaver, N.S. Webster, B. Wehrle, P. Weisenhorn, S. Weiss, J.J. Werner, K. West, A. Whitehead, S.R. Whitehead, L.A. Whittingham, E. Willerslev, A.E. Williams, S.A. Wood, D.C. Woodhams, Y. Yang, J. Zaneveld, I. Zarraonaindia, Q. Zhang, H. Zhao, A communal catalogue reveals Earth's multiscale microbial diversity, Nature 551 (2017), doi:http://dx.doi.org/10.1038/nature24621.
[28] N.S. Grantham, B.J. Reich, K. Pacifici, E.B. Laber, H.L. Menninger, J.B. Henley, A. Barberán, J.W. Leff, N. Fierer, R.R. Dunn, Fungi identify the geographic origin of dust samples, PLoS One 10 (2015)e0122605, doi:http://dx.doi.org/10.1371/journal.pone.0122605.
[29] N. Procopio, S. Ghignone, A. Williams, A. Chamberlain, A. Mello, M. Buckley, Metabarcoding to investigate changes in soil microbial communities within forensic burial contexts, Forensic Sci. Int. Genet. 39 (2019) 73–85, doi:http://dx.doi.org/10.1016/j.fsigen.2018.12.002.
[30] N.S. Grantham, B.J. Reich, E.B. Laber, K. Pacifici, R.R. Dunn, N. Fierer, M.J. Gebert, J.S. Allwood, S. Faith, Global forensic geolocation with deep neural networks, ArXiv Preprint (2019). https://arxiv.org/abs/1905.11765.
[31] T.J. White, T. Bruns, S. Lee, J. Taylor, Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, PCR Protocols: A Guide to Methods and Applications, vol. 18(1990) , pp. 315–322.
[32] N. Fierer, M. Hamady, C.L. Lauber, R. Knight, The influence of sex, handedness, and washing on the diversity of hand surface bacteria, Proc. Natl. Acad. Sci. 105 (2008) 17994–17999, doi:http://dx.doi.org/10.1073/pnas.0807920105.
[33] RStudio, RStudio: Integrated Development Environment for R, Boston, MA, (2018) . http://www.rstudio.org/.
[34] J.G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F.D. Bushman, E.K. Costello, N. Fierer, A.G. Peña, 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, R. Knight, QIIME allows analysis of high-throughput community sequencing data, Nat. Methods 7 (2010) 335–336, doi:http://dx.doi.org/10.1038/nmeth.f.303.
[35] R Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.R-project.org/.
[36] J. Allaire, Y. Xie, J. McPherson, J. Luraschi, K. Ushey, A. Atkins, H. Wickham, J. Cheng, W. Chang, rmarkdown: Dynamic Documents for R, (2018) . https://CRAN.R-project.org/package=rmarkdown.
[37] B.J. Callahan, P.J. McMurdie, M.J. Rosen, A.W. Han, A.J.A. Johnson, S.P. Holmes, DADA2: high-resolution sample inference from Illumina amplicon data, Nat. Methods 13 (2016) 581–583, doi:http://dx.doi.org/10.1038/nmeth.3869.
[38] B.J. Callahan, P.J. McMurdie, S.P. Holmes, Exact sequence variants should replace operational taxonomic units in marker-gene data analysis, ISME J. 11 (2017) 2639–2643, doi:http://dx.doi.org/10.1038/ismej.2017.119.
[39] R.C. Edgar, UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing, BioRxiv (2016)081257, doi:http://dx.doi.org/10.1101/081257.
[40] Evan Bolyen, et al., Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2, Nat. Biotechnol. (2019), doi:http://dx.doi.org/10.1038/s41587-019-0209-9 in press.