REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
27-08-2019	Final Report	16-Aug-2016 - 28-Feb-2019

4. TITLE AND SUBTITLE
Final Report: Forensic Geolocation via Biological Signatures

5a. CONTRACT NUMBER
W911NF-16-2-0195

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

North Carolina State University
2701 Sullivan Drive
Admin Srvcs III, Box 7514
Raleigh, NC 27695 -7514

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
68784-LS-RIF.9

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Seth Faith
UU	UU	UU	UU		19b. TELEPHONE NUMBER 614-424-3368

RPPR Final Report
as of 12-Sep-2019

Agency Code:

Proposal Number: 68784LSRIF **Agreement Number: W911NF-16-2-0195**
INVESTIGATOR(S):

Name: Seth A. Faith
Email: faiths@battelle.org
Phone Number: 6144243368
Principal: Y

Organization: **North Carolina State University**
Address: 2701 Sullivan Drive, Raleigh, NC 276957514
Country: USA
DUNS Number: 042092122 EIN: 566000756

Report Date: 31-May-2019 Date Received: 27-Aug-2019
Final Report for Period Beginning 16-Aug-2016 and Ending 28-Feb-2019
Title: Forensic Geolocation via Biological Signatures
Begin Performance Period: 16-Aug-2016 **End Performance Period:** 28-Feb-2019
Report Term: 0-Other

Submitted By: Kelly Landolfi Email: Landolfi@battelle.org
Phone: (614) 424-4428

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 1 **STEM Participants:** 2

Major Goals: The purpose of this research study was to establish laboratory methods and develop robust software for forensic examination of materials to determine the geographic origin and a probability estimate that the materials originated in the expected geographic region. The following goals were established for the study.

Goal 1 is to establish an easy laboratory workflow for DNA scientists to generate sequencing data from forensic samples.
Goal 2 is to refine a statistical model and database with Defense Forensic Science Center (DFSC) specific data generated from OCONUS samples on this project.
Goal 3 is to convert the model from a developmental tool to a production quality, end-user designed software deliverable.
Goal 4 is to train the end-users at DFSC on the laboratory method and software.

Accomplishments: The overarching objective of this project was to investigate the use of genetic signatures from environmental materials found within dust, such as plants and fungi, in order to determine the point of origin. Study tasks to achieve the objective are as follows:
Task 1. Kick-off meeting and requirements gathering -The team will meet with stakeholders and end-users at DFSC to define the requirements of the system.
Task 2. Study plan - A plan will be prepared that outlines the testing/development and software production tasks. DFSC will review and approve the plan prior to commencement of work.
Task 3. Sample collection - Working through DFSC and other DoD partners, environmental samples will be collected from DFSC high priority regions of the world.
Task 4. Data generation – OCONUS environmental samples will be assessed for fungi (ITS1) and plant (trnL) marker gene sequences using the specialized laboratory to establish the database.
Task 5. Model development - The existing statistical model will be re-trained using the new OCONUS data and evaluated for precision and accuracy using cross-validation.
Task 6. Software production - The final, production-quality, implementation of our software will be built with end-users specified graphical user interface and extensively tested, optimized, and documented prior to delivery to DFSC.
Task 7. SOPs and Manuals - Convert the laboratory methods to a forensic protocol according to DFSC requirements. A demonstration of the method using forensic unknowns will be conducted and complete laboratory SOPs and software manuals will be delivered.
Task 8. Training - Hands-on laboratory and software training will be provided to DFSC using validated SOPs developed on this program.

RPPR Final Report
as of 12-Sep-2019

Task 9. Reporting – Project flyer, monthly reports, quad charts, and a final technical report will be provided. Major accomplishments were made for each task. Overall resulting methodology, including sample testing and model use, were evaluated for proof of concept and standard operating procedure development. The results showed that a fungal genetic marker (ITS1) and not a plant genetic marker (trnL), displayed extremely high potential as a forensic marker in dust samples, providing over 20,000 unique sequence signatures across 487 tested samples. Implementation of a custom designed, spatially aware deep learning model (DeepSpace) yielded ~86% prediction accuracy for single source samples using ITS1. The laboratory methods and prediction software were assessed in a demonstration test on a variety of operationally relevant samples. Testing displayed ideal use cases and also current gaps that may be addressed in future research to enhance this novel and powerful technology for routine use in forensic laboratories. Final deliverables included: laboratory SOPs, custom bioinformatics scripts, the software analysis tool (ASVTracer), raw and analyzed data, and training materials. A scholarly manuscript was prepared that describes the statistics methods in finer detail (Grantham et al. 2019) and is available from the ARO extranet and final electronic data deliverable disk.

To date, this project demonstrates that largest OCONUS sampling of environmental dust signatures, and subsequent genetic analysis. The power of this dataset was reflected in building a cutting-edge model to make predictions of geographic origin for unknown samples. The study team demonstrated that the fungal genetic marker ITS1 is very effective as a highly diverse biomarker with geographic origin information that can be used in deep learning (artificial intelligence) models. Here, the spatially aware deep learning model, DeepSpace, was custom designed to be implemented for the specific use of geographic prediction from NGS analyzed environmental samples. The model was highly accurate for country of origin and was readily developed into end-user designed software for routine analysis. Demonstration testing showed that single source samples analyzed with the ITS1 NGS protocol and the ASVTracer software could be readily predicted to the correct country of origin, while mixed source samples presently do not yield correct mixed source predictions. In conclusion, this study shows the extreme potential of NGS genetic analysis, coupled to artificial intelligence, to provide mission critical information for providence of materials and opens the doors to transition of this technology into routine casework.

Training Opportunities: Scientists at the Defense Forensic Science Center (DFSC) were trained by the research study staff for the specialized laboratory methods in next-generation DNA sequencing and analysis approaches using machine learning.

Results Dissemination: One scientific publication was prepared and submitted for peer-review.

Grantham N.S., Reich B.J., Laber E.B., Pacifici K., Dunn R.R., Fierer N., Gebert M., Allwood J.S., & Faith S.A. Forensic Geolocation with Deep Neural Networks. J. Royal Statistical Society: Series C (2019) - in review.

One doctoral dissertation was produced that contained some research from this study.

Grantham N. “Statistical Methods for High-Dimensional, Spatially-Distributed Microbiome Data from Next-Generation Sequencing.” Thesis submitted to NC State University - Summer 2017.

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: No IP filings were made due to this study, but software was developed as a project deliverable and was transferred to the Defense Forensic Science Center.

PARTICIPANTS:

Participant Type: Co-Investigator
Participant: Seth Adam Faith
Person Months Worked: 6.00
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Funding Support:

Participant Type: Co-Investigator
Participant: Matthew Breen
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Faculty
Participant: Robert Dunn
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Faculty
Participant: Eric Laber
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Faculty
Participant: Brian Reich
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Faculty
Participant: Noah Fierer
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Jesse Clifton
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N

Other Collaborators:

Participant Type: Undergraduate Student
Participant: Saran Ahluwalia
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Julia Allwood
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

DISSERTATIONS:

Publication Type: Thesis or Dissertation
Institution: North Carolina State University
Date Received: 07-Sep-2017 Completion Date: 8/31/17 8:19PM
Title: Statistical Methods for High-Dimensional, Spatially-Distributed Microbiome Data from Next-Generation Sequencing.
Authors: Neal S. Grantham
Acknowledged Federal Support: **Y**

FINAL TECHNICAL REPORT

Army Research Office

Submitted to Defense Forensic Science Center (DFSC)

ARO Agreement W911NF-16-2-0195

"Forensic Geolocation Via Biological Signatures"

PI: Matthew Breen PhD
Email: matthew_breen@ncsu.edu
Co-PI: Seth A. Faith PhD
Phone: (614) 688-1809 Email: faith.3@osu.edu

Submitted by: Seth A. Faith

Submitted: 24 May 2019
Revised and resubmitted: 26 Aug 2019

DUNS 042092122 and EIN: 56-6000756

Recipient Organization:

North Carolina State University
Research Administration/SPARCS
2701 Sullivan Drive
Admin Services III; Box 7514
Raleigh, NC 27695-7514

Project Period: 16 Aug 2016 – 28 Feb 2019

Final Technical Report:
Forensic Geolocation via Biological Signatures

Executive Summary

This report encapsulates the work completed on project *'Forensic geolocation via biological signatures'* in accordance with the cooperative research agreement W911NF-16-2-0195. The overarching objective of this project was to investigate the use of genetic signatures from environmental materials found within dust, such as plants and fungi, in order to determine the point of origin. Point of origin, or geolocation, was predicted through the development and evaluation of various statistical models. Overall resulting methodology, including sample testing and model use, were evaluated for proof of concept and standard operating procedure development. The results showed that a fungal genetic marker (ITS1), and not a plant genetic marker (trnL), displayed extremely high potential as a forensic marker in dust samples, providing over 20,000 unique sequence signatures across 487 tested samples. Implementation of a custom designed, spatially aware deep learning model (DeepSpace) yielded ~86% prediction accuracy for single source samples using ITS1. The laboratory methods and prediction software were assessed in a demonstration test on a variety of operationally relevant samples. Testing displayed ideal use cases and also current gaps that may be addressed in future research to enhance this novel and powerful technology for routine use in forensic laboratories. Final deliverables included: laboratory SOPs, custom bioinformatics scripts, the software analysis tool (ASVTracer), raw and analyzed data, and training materials. A scholarly manuscript was prepared that describes the statistical methods in finer detail (Grantham et al. 2019) and is available from the ARO extranet and final electronic data deliverable disk.

Table of Contents

1. Problem Statement

Develop and evaluate forensic examination and analysis methodology for the use of DNA signatures obtained from environmental materials for prediction estimation of geographic origin in a global context.

2. Introduction

The purpose of the 'Geolocation via Biological Signatures' project was to build on existing research (Barberán et al., 2015; Grantham et al., 2015) to perform advanced technologically analysis of environmental DNA signatures, targeting plant and fungal material. The primary scope of this work was to develop, test and deliver a customized laboratory workflow and method, interpretation software, and reference global database for the forensic examination of such materials. This method was focused on the biological components of outdoor dust, with dusts swabs collected and analyzed from many different countries across the globe. The scope of the present work was to build upon earlier biogeographic prediction methods and capabilities established on samples from the continental United States, and develop the methodology to samples acquired outside the continental United States (OCONUS). Samples collected were analyzed and compiled into a representative global dataset and used to train and evaluate multiple prediction algorithms. A standard operating procedure was developed to process test samples on a smaller scale (as compared to reference dataset construction) using next-generation sequencing (NGS) and evaluated through demonstration test. DNA signature data curation and prediction software input and analysis protocols were developed and presented as the final examination system for global DNA signatures as collected from outdoor dust.

3. Sample Collection and Processing

Dust-associated fungal and plant DNA was collected and analyzed from nearly 35 countries across 6 continents (excluding Antarctica), the largest database of global dust samples known to date (see Appendix A for details on the global dataset). Outdoor dust samples were collected with a dry Bode SecureSwab2 collector or a sterile cotton swab between November 2016 and February 2017 by employees of Bode Cellmark stationed worldwide. The collection SOP was a project deliverable and not presented in this report. Standardized meta-data collected for each sample included a qualitative description of the sampling location, country of origin, and latitude/longitude

coordinates (to 100th of a degree resolution). Dust samples were stored at room temperature and sent to the University of Colorado laboratory for molecular analyses. DNA was extracted using the Qiagen PowerSoil htp-96 well Isolation Kit and a modified method of Barberan et al. (2015). To target fungal taxa, the first internal transcribed space (ITS1) region of the rRNA operon was amplified using ITS1-F/ITS2 barcoded primers (McGuire et al., 2013). To target plant taxa, primers targeting a region of the trnL chloroplast gene (as described in Craine et al., 2017) were used. For both primer pairs, each sample was assigned a unique 12-bp error-correcting barcode (Caporaso et al., 2010) and PCR-amplified in triplicate. After gel visualization to confirm amplicon size/amounts, all amplicons for a given primer pair were pooled together using SequalPrep Normalization Plates and sequenced on an Illumina MiSeq instrument running the 2x250 bp MiSeq kit. Sequences were demultiplexed using a custom Python script; pair-end reads were merged using the usearch7 mergepairs feature (Edgar, 2010); adapter sequences were trimmed from the merged reads using fastx clipper (https://github.com/agordon/fastx_toolkit); and reads were quality filtered using usearch7 (Edgar, 2010). Finally, sequences were clustered into amplicon sequence variants (ASVs at 100% sequence similiarity) using DADA2 (Callahan et al., 2016) and their taxonomic identities were determined using a Bayesian classifier (Wang et al., 2007) against either the UNITE database (Abarenkov et al., 2010) for fungi or an in-house reference database of plant trnL sequences compiled from the NCBI GenBank archive.

4. Data Generation

4.1. Global Dataset

A total of 517 dust samples were analyzed. These samples were collected from each of the following 35 countries representing several geographic regions: the Americas (Mexico, Colombia, Trinidad and Tobago, Uruguay, Argentina, Brazil, Costa Rica, Honduras), Africa (Ghana, Nigeria, South Africa, Djibouti and Somalia), East Europe (Czechia, Croatia, Hungary, Macedonia, Moldova), West Asia (Turkey, Cyprus, Jordan), Middle East (Bahrain, Kuwait, Qatar, Oman, Georgia, Azerbaijan), Central Asia (Afghanistan, Kazakhstan, Pakistan), East Asia (Vietnam, South Korea, Malaysia), and Oceania (Australia, New Zealand). Together, 15-20 samples per country were processed using the plant (trnL) and fungal (ITS1)-targeted sequencing approaches described above. After applying a minimum of 3,000 read dataset for each sample, 487 total samples were analyzed from 34 countries (Djibouti and Somalia were treated as one country). A total of >20,000 and ~1,000 fungal and plant taxa (unique amplicon sequence variants), respectively, were detected across the sample set.

4.2. Test Samples and SOP

To make sample processing as standardized as possible, a semi-automated data analysis pipeline was assembled to circumvent or minimize potential points of end-user

subjectivity or chances of human error. To facilitate this, a pipeline was built based on that used to generate the global dataset, and implemented within R Studio (standard operating procedures and R scripts delivered to the sponsor with final data delivery package). R Studio is a free, easily accessible user interface that facilitates the operation of R software as well as other user-friendly attributes. One such attribute is the capacity to use R Markdown (RMD) documents. These files allow for the customization of analyses to be performed as well as addition of user notation and guidance. These documents contain analysis-driving script, that allow for a 'push play' approach when performing analyses using R software that otherwise would be largely command line-based. This functionality allows for reproducible analysis and was adopted to facilitate and document sample processing and direct consistent output file generation, suitable for model implementation. To process test samples and to perform the SOP, the RMD is used simultaneously as a processing end-user guide as well as an analysis tool. The first part instructs the user how to perform the required functions in the terminal, while the second part instructs the R software to process the curated files. Initial steps require the use of QIIME2, which requires a Linux-based system or internet connectivity via a virtual machine. To circumvent this, the Windows Subsystem for Linux (WSL) for Windows 10 was used, at the recommendation of the sponsor. Within the R Studio platform and using WSL, a command line terminal can be run, which allows for the initial sample processing of run data using QIIME2. The second part of the pipeline using DADA2 can be run by the 'push play' approach using the RMD template. Use of the RMD culminates in the generation of a document detailing the processes that were applied, along with run, operator and time-stamp details. The data file required for model implementation is also generated in the format required for immediate use.

Basic quality control is performed when the sequence data comes off the instrument in accordance with the platform used, which typically includes removing nonsensical "junk" reads and phiX sequences. In order to begin analyzing these files a combination of python functions are utilized through QIIME2 to curate and demultiplex the sequence data. In brief, demultiplexing is performed using the parameters set forth within QIIME2 for the Earth Microbiome Project, which is the methodology for sequencing that has been used for the present project. This assumes that sequence data is paired-end (sequenced in both directions) and multiplexed (with each sample having a unique 12-bp barcode for differentiation). Sequence data is then demultiplexed using QIIME2 which removes adapter and primer sequences, and produces individual read one and two files per sample, whilst retaining quality information. These steps prepare the sequence data for analysis which is done primarily using R software and the DADA2 package in R Studio. This pipeline identifies and counts the unique amplicon sequence variants (ASVs) observed within the samples processed and matches them back to those discovered within the global dataset. A key part of accurately determining sequence error from true ASVs, is facilitated by DADA2 which 'denoises' (removes errors) from forward and reverse reads separately. Therefore, contrary to other comparable pipelines, forward and reverse reads are not merged until data processing (which is

beneficial given that error rates are not necessarily the same for the forward and reverse reads).

Prepared individual sample files were then filtered and trimmed. During this step, no sequences with 'N' ambiguity codes are tolerated and any remaining phiX sequences are removed. Quality parameters are set as maximum expected error values for forward and reverse reads. In order to 'denoise' the sequences, several steps are performed including establishing the error rate and dereplication to reveal unique final sequences. The error rates are then applied to the dereplicated sequences to produce the final ASVs. Sample read files are then merged, with a minimum overlap of 20 base pairs and a maximum mismatch of zero, using the denoised and dereplicated data, resulting in a final sequence table detailing samples and ASV observance counts. Chimeric sequences (PCR-mediated recombination sequences) are removed at this point before the sequences are assigned an identifier against the global sequence database. Output files are generated including final ASV table with global database identification and observances, as well as a file containing representative sequence data for each ASV determined within the run being processed. Interpretation caveats include assessing any sequence information produced within the reagent blanks and negative controls, and only submitting samples for model prediction that consist of equal or greater than 3000 sequence observances in the final file. For detailed step by step data generation for test samples and as employed in the SOP (along with the data generation summary that is described here above), please see the laboratory SOP deliverable document.

5. Model Development

5.1. Overview of the DeepSpace Model

The basic idea of the spatially aware deep learning model, DeepSpace, is to (1) break up the Earth's surface into many random partitions; (2) for each of these random partitions, train a deep learning model that predicts which cell of the partition a sample came from; and (3) average the predictions over all of the random partition to get a more fine-grained prediction. The model uses a mixture of 50 fine and coarse partitions of the globe (see Figure 1 for an example of coarse and fine partitions of the United States).

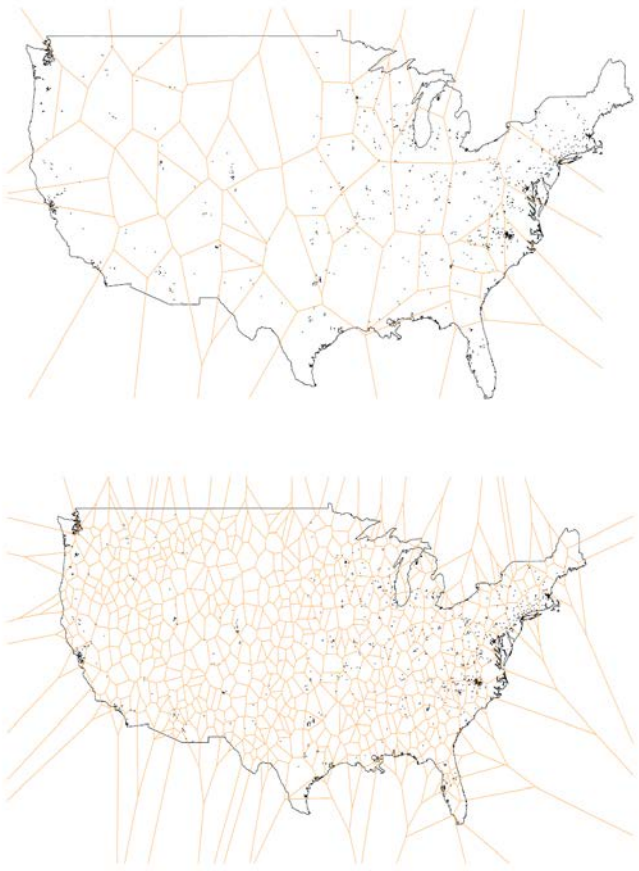

Figure 1: A coarse (top) and fine (bottom) random partition of the continental United States.

To obtain point-predictions, we divide each of the countries in the training set into a collection of small regions. The final output of a DeepSpace model prediction is a probability at each of these small regions, where the probability at a given region is the average of the probabilities of the cells in each random partition which contain that region.

5.2.Training Data

The final dataset consists of 487 samples prepared by the method described in Sections 3 and 4, after excluding: 1) samples with a total count of less than 3000 across all taxa, and 2) taxa which did not have any counts in any sample. The latter filter left 19842 out of 20446 ITS1 taxa in the final training data set.

Table 1 displays the countries included in the final dataset alongside the number of samples from each country.

Afghanistan (5) Argentina (15) Australia (16) Azerbaijan (15) Bahrain (16) Brazil (24) Colombia (15) Costa Rica (15) Croatia (14) Cyprus (13) Czechia (13) Djibouti-Somalia (13)	Georgia (13) Ghana (16) Honduras (15) Hungary (15) Jordan (15) Kazakhstan (13) Kuwait (14) Macedonia (FYROM) (14) Malaysia (15) Mexico (15) Moldova (16) New Zealand (10)	Nigeria (17) Oman (10) Pakistan (15) Qatar (15) South Africa (13) South Korea (11) Trinidad and Tobago (15) Turkey (16) Uruguay (15) Vietnam (15)

Table 1. Countries in the DeepSpace dataset (number of samples from country in parentheses). Sampling and analysis strategy attempted to maximize the geographic breadth of sampling within each country, given the diversity of actual samples collected.

5.3. DeepSpace Model Technical Details

Let *D* be a spatial domain; in our case, *D* is a collection of regions within each of the 34 countries in the final dataset on which the DeepSpace model was fit. Geolocation may be formalized in the language of spatial point process theory, where the response is a spatial location *s* in *D*, which is regressed onto nonspatial covariate evidence x in X. In our application, x is a vector of microbial presence/absence data such that X = { 0, 1}^{p}, where *p* is in the tens of thousands. We assume that the spatial point process follows a nonhomogenous Poisson process with intensity surface $\lambda(s \mid x)$for all *s* in *D*. The challenge lies in choosing an appropriate model for the intensity surface $\lambda(s \mid x)$.

Let $P = \{P_k\}_{k=1,}^{K}$ denote a partition of *D.* Given *P*, we assume the Poisson intensity is constant in each cell of this partition, and denote $\lambda_k(x)$ as the intensity for all *s* in P_k. Let $|P_k|$ denote the area of region P_k and take $f_k(x) = log\ [|\ P_k|\lambda_k(x)]$ and the probability on region *k* is

$$P(s\ \in P_k \mid x)\ =\ \frac{exp\ [\ f_k(x)\ ]}{\Sigma_{l=1}^{K}\ exp\ [f_l(x)]}$$

for *k=1, …, K.*

We estimate $f_1, \ldots, f_K$ by training a supervised classifier on available data points as follows. Define $h(s) = k$ for s in P_k, which serves as the label for the cell to which *s* belongs. Then the classifier is trained on $\{(x_i, h_j(s_i))\}_{i=1}^{n}$. The trained classifier yields estimators $\widehat{f_1}, \ldots, \widehat{f_K}$ and subsequently the intensity surface estimator $\widehat{\lambda_k}(x)\ =\ |\ P_k|^{-1}\ exp\ [\hat{f}_k(x)]$.

The spatial prediction is the location that maximizes the fitted intensity surface. A piecewise constant intensity surface does not have a unique maximum, and thus the precision of the geolocator is limited by the resolution of partition *P*. To obtain finer accuracy, we construct multiple random partitions of *D*, train a separate classifier on each partition, and estimate the intensity surface as the average over the random partitions. To generate partition *j*=1, …, *N*, we draw K_j seeds from the domain, $v_{jk} \sim^{iid}$ Uniform(*D*), and define a Voronoi partition $P^j = \{ P^j_k \}_{k=1}^K$ where $P^j_k = \{s \in D: (\forall l \neq k) \, || s - v_{jk}|| < ||s - v_{jl}||\}$.We train a supervised classifier on $\{(x_i, h_j(s_i))\}_{i=1}^n$ to yield $\hat{F}_j = \{\hat{f}_{j1}, \ldots, \hat{f}_{jK_j}\}$and subsequently $\hat{\lambda}_{jk}(x) = |P^j{}_k|^{-1} \, exp\,[\hat{f}_{jk}(x)]$.

We obtain geolocation predictions by averaging pointwise over our collection of models $M = \{ \hat{F}_j \}_{j=1}^N$. The estimated intensity $\hat{\lambda}(s \mid x) = \frac{1}{N} \Sigma_{j=1}^N \hat{\lambda}_{j h_j(s)}(x)$,and so the estimated location for a sample with microbiome composition *x* given by

$$\hat{s} = arg \max_{s \in D} \hat{\lambda}(s \mid x, M).$$

5.4.Model Fitting

The probability estimator used by DeepSpace is a feed-forward neural network with these specifications:

- 3 fully connected hidden layers with 2048, 2048, and 1024 neurons;
- Dropout rate of 0.3 at each layer;
- Adam optimizer;
- Implemented in Keras package in Python 3.

The DeepSpace model was fit on the dataset as described in Section5.4, and the resulting 50 deep learning models (corresponding to 50 Voronoi partitions) saved as locally stored .hdf5 files. These files are accessed by the ASVTracer software in order to make predictions on user-uploaded samples (see Software Production).

The first studies in model fitting deployed routine machine learning and dep-learning models to assess both the ideal model and the performance of TIS1, trnL or combined ITS1 and trnL. The larger study can be review in Grantham et al. (2019), a manuscript in review, but provided to the sponsor. Table 2 provides a summary of the testing. Various models were assesed for the genetic data including, assessment of the partition sizes (coarse, mixed and fine), statistical models (KNN – k-nearest-neighbor, RF – Random Forest, NN – Neural Network, DNN – Deep Neural Network, DS – DeepSpace), and DNA sequence sets (ITS1 – fungi alone, trnL – plant alone, ITS1+trnL – fungi and plant combined). ITS1 alone consistently provided the highest accuracy for unknown sample testing (determination of correct country), while trnL alone and ITS1+trnl combined provided weaker accuracies for predication. The DeepSpace (fine) model with ITS1 was

the showed the highest predictive accuracy of all tests and was determined to be the model for implementation into the final software.

Some key take-ways are provided through this testing. First, the ITS1 gene, that yielded ~20,000 unique sequences, was powerful as a genetic marker to predict unknown sample origin (86.7% accuracy for country of origin. Second, the plant marker (trnL) was minimally accurate (< 46% across all cases), and also reduced accuracy when combined in the analysis with ITS1. This may be due to the lack of high polymorphism (uniqueness) in the length of DNA analyzed for trnL, as well as confounding and variable distributions of plant signatures globally. Thirdly, the DeepSpace model showed much higher accuracy than classical machine learning models (e.g., KNN, NN), likely attributed to the spatial awareness of the model that classifies on space (geo-coordinates), rather the geopolitical boundaries (i.e., country borders). Thus, ITS1 in the DeepSpace model is a major advancement for trace analysis and prediction of point of origin analysis.

Seeds	Model	ITS1	trnL	ITS1-trnL
		Accuracy (%) country	Accuracy (%) country	Accuracy (%) country
Coarse	Spatial KNN	29.3	30.1	36.0
Coarse	Spatial RF	36.9	39.8	36.8
Coarse	Spatial NN	45.5	44.2	57.3
Coarse	DeepSpace	46.6	43.1	62.2
Mixed	Spatial KNN	29.0	29.3	33.3
Mixed	Spatial RF	38.8	36.9	42.5
Mixed	Spatial NN	45.8	45.8	56.3
Mixed	DeepSpace	45.3	46.9	64.9
None	DNN	44.7	38.8	79.2
Fine	Spatial KNN	60.7	46.1	34.6
Fine	Spatial RF	72.9	45.3	43.5
Fine	Spatial NN	80.7	20.3	55.1
Fine	**DeepSpace**	**86.7**	44.7	63.2

Table 2. Model testing for parameters of partition size, genetic marker, and Machine learning/deep learning model.

Given that the DeepSpace model assigns probabilities and point estimates on space rather than country, the summary above (Table 2), may not provide the full value of the model. In figure 2, the accuracy is presented as predicted country versus true country of origin, but the countries are aligned along the axes by the geographic proximity. In a perfect situation the samples would show as one-to-one matches are only the boxes along the diagonal would contain information. In figure 2, the samples that are mis-assigned (outside of the diagonal), are in many case cases the diagonal indicating a neighboring country. So, this presentation of data now demonstrates that many of the

inaccurate country calls were to neighboring countries, which in some cases may be a very real or close to target prediction.

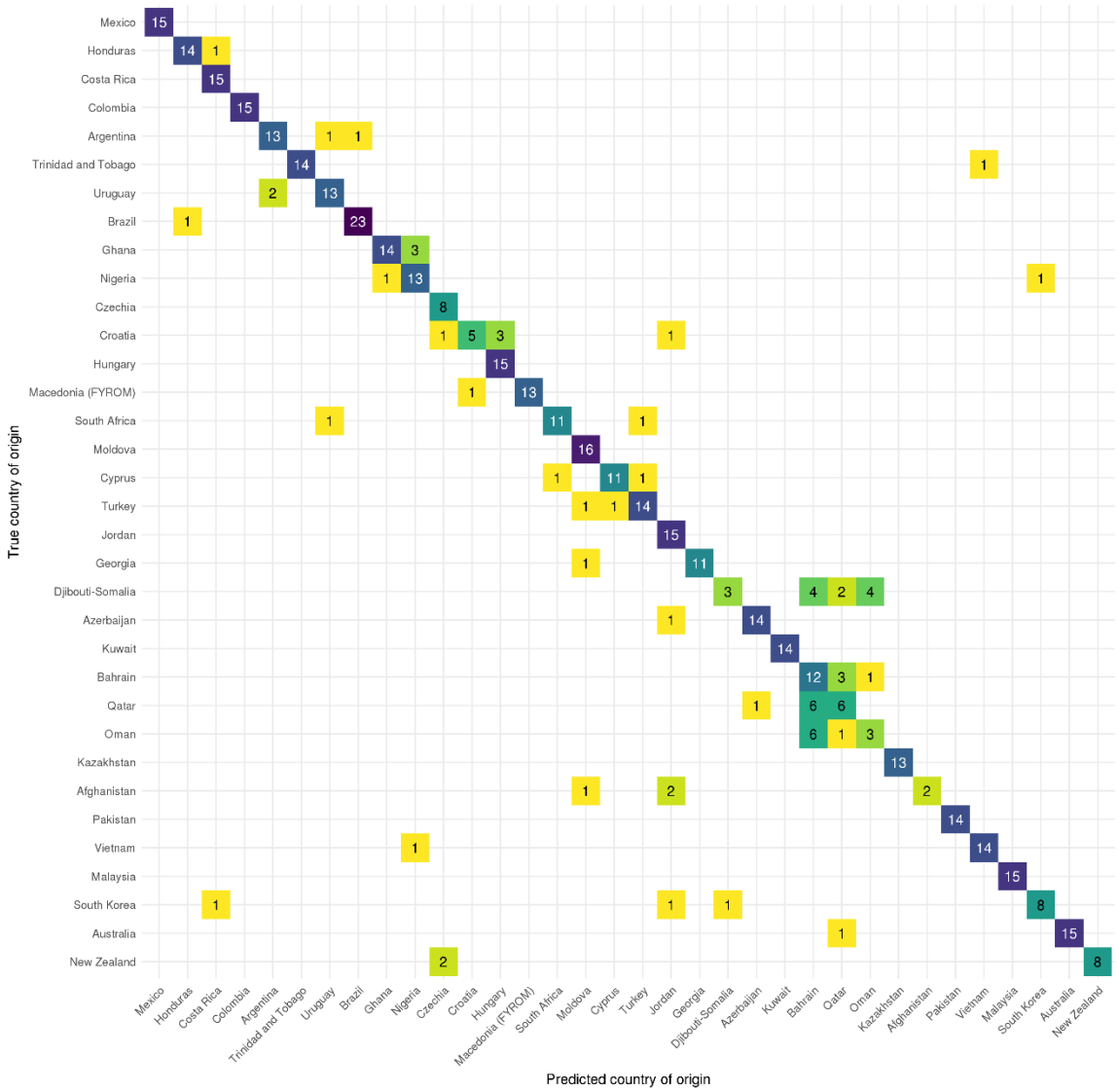

Figure 2 - Country Classification matrix for DeepSpace (mixed) model. The confusion matrix for sample prediction with ITS1 analysis is presented to show the model determination of samples from predicted (x-axis) versus their true (y-axis) origins. The Amplicon Sequence Variant (ASV) protocol was applied to identify ITS1 taxa (unique sequences) in each sample that were analyzed using Illumina MiSeq sequencer (see project test plan). Numbers inside the boxes reflect the total samples classified. Thus, samples along the diagonal are correct assignments. The axes are oriented by country proximity. Thus, misclassification close to the diagonal are in neighboring/adjacent countries.

6. Software Production

The ASVTracer software was developed in order to facilitate the use of the DeepSpace model to make probabilistic predictions about the geospatial location associated with data described in Section 5.2.

ASVTracer provides a user interface which supports the following workflow for making predictions on sample data:

1. Upload a CSV or Excel file containing microbial presence/absence data (and display error message if this file is incorrectly formatted).
2. Compute predictions for each sample in the uploaded file using the saved DeepSpace model; when these predictions are made, the model output (i.e., probabilities, country names, latitude, and longitude for each spatial location in the prediction region) is automatically saved to a local time-stamped text file.
3. Select one of the analyzed samples to display confidence regions associated with the predicted probabilities from the DeepSpace model, along with a marker indicating the highest-probability location, on a world map in the user interface.
4. Define a circular region of Earth's surface by choosing its radius in kilometers and the coordinates of its center for which to compute a probability; that is, the predicted probabilities of each location within the user-defined region are added up to yield the probability of the sample having come from somewhere in that region. Display the associated region and the probability in the user interface.
5. Generate a time-stamped PDF containing a copy of the model output, the map display, and the results of any hypothesis tests.

The initial version of the software was presented at the in-person training (Section 8.3), which resulted in feedback on what features the final version of the software should include, including user-friendly error messages for incorrectly formatted data; display of confidence regions associated with a sample, rather than probabilities at each spatial location; display of the highest-predicted-probability spatial location; display of the user-defined region for hypothesis testing; and automatic generation of time-stamped PDF reports corresponding to the complete model output and all user-conducted hypothesis tests. These features were included in the final version of the ASVTracer software.

ASVTracer is implemented in Python 3. The user interface was implemented using the Tkinter package, with Matplotlib's Basemap library used to implement the display of spatial predictions on a world map. The ASVTracer software was a final deliverable to the sponsor.

7. Demonstration Testing

7.1. Overview

The demonstration testing was included in the geolocation project study plan to illustrate the forensic version of the research method in conjunction with the final prediction model. Demonstration testing culminated in the production of the SOP and the assessment of model prediction accuracy and capability. To appropriately demonstrate this method, a variety of samples and mixtures were set up to investigate the capability of the entire final method, as would be applied by the end-user to subsequent test samples. Samples included those collected during the global database generation phase, outgroup samples (i.e. indoor and country excluded from model development) and artificial mixtures, in accordance with the test criteria in the study plan, as well as positive control reference materials. All samples were treated in the same manner throughout processing and applied to the model as would any subsequent test samples submitted by an end-user. For a comprehensive report on the demonstration testing undertaken please refer to the standalone demonstration testing report. For details regarding reference materials and all method specifications please refer to the stand-alone SOP document.

7.2. Approach

Dust samples were collected following the protocol described above. Of the swabs retained from the global collection, two swabs were used from each of 23 different countries for demonstration testing. In addition, outgroup samples were used, which included swabs from Panama (collected from indoor environments), swabs and tape lifts collected locally (Raleigh, NC) and control samples (positive and negative). DNA was extracted using the method described above using lower throughput options (i.e. single-tube extractions instead of plate extractions). A total of 24 samples were processed from 12 countries (two swabs per country), along with a single swab from seven additional countries to make up artificial mixed samples. Swabs used for most mixtures were also processed as single country samples. A total of 18 outgroup samples were processed, which included locally collected samples and Panama (indoor) samples. Artificial mixtures (composed of volume mixtures of DNA extracts per country) were set up in a range of different proportions to represent two- and three-country mixtures. A total of 18 different two-country mixtures were set up and analyzed, composed of two pairs of countries (four countries total), set up at different proportions. A total of 12 different three-country mixtures were set up and analyzed, composed of six different countries in variable proportions. Additionally, eight context-themed mixtures (e.g. similar latitude, same continent etc.) were also set up, consisting of equal proportions. The total of samples analyzed during demonstration testing therefore included: 31 single country samples, 18 two-country mixtures, 12 three-country mixtures, 8 themed mixtures, 18 outgroup samples, and 9 controls.

For demonstration testing, only fungal signatures were generated, as per sponsor guidance during the critical design review. The ITS1 region of the fungal rRNA gene was targeted, amplified and sequenced following the PCR and sequencing method described above. This method used the unique primer pair combination per reaction to both amplify the ITS1 target, and simultaneously prepare resulting amplicons for sequencing on an Illumina platform. All amplification was performed in triplicate reactions, with pooled product used for subsequent processing and library preparation (as described in 'Sample Processing' above). Demonstration testing samples were sequenced using an Illumina MiSeq with a MiSeq Reagent Kit v2 (500 cycles) (Illumina, San Diego, CA), performed as a 2x250bp 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. Sequence data was processed as described above in 'Data Generation – Test Samples and SOP'. All resulting data were applied to the model in an anonymous way, as if no prior information was available regarding sample origins.

7.3. Results and Discussion

Of the 31 single country samples analyzed, two did not generate sufficient data for implementation to the model for country of origin prediction. A root cause analysis was not performed to determine why the two samples failed. Of the remaining 29 samples, 24 were correctly predicted to their country of origin (Table 3). Five samples were incorrectly predicted to another country of origin, these included both samples from Qatar (predicted as Bahrain), and one sample each from New Zealand (predicted as Uruguay), Oman (predicted as Bahrain) and Nigeria (predicted as Ghana). This results in an overall prediction accuracy of 82.8% for these single country samples.

ID:	**Country:**	**Sequence Counts**		**Prediction:**	**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

Table 3: 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.

Two-country mixtures were made up of 9 different proportion settings using two different country pairs (Turkey and Georgia, Costa Rica and Trinidad and Tobago), resulting in 18 reactions. Regardless of proportion set up, one country of each pair was consistently predicted as the country of origin, being Georgia for the first pair, and Trinidad and Tobago for the second pair (Table 4). For the latter pair, Brazil was also commonly predicted after Trinidad and Tobago as the second and/or third best prediction. The only exception to this pattern was at the maximum proportion input for Costa Rica, where the top predictions were then Honduras, followed by Trinidad and Tobago.

Three-country mixtures were set up in various combinations and at three different proportion settings. Mixtures included the following combinations of countries: Cyprus, Pakistan, and Macedonia; Pakistan, Macedonia, Oman; Macedonia, Oman and Hungary; and Oman, Hungary, and Nigeria. The first two sets of mixtures resulted in all three top predictions of country of origin as Pakistan, regardless of the proportion of the mixture that the Pakistan sample represented (Table 4). The latter two sets of mixtures all resulted in Hungary as the top prediction, regardless of the proportion of the mixture that the sample from Hungary represented. Hungary was also either the second or third best prediction for these samples, with Croatia also being predicted as second or third.

ID:	Countries:	Mix:	Best	2nd	3rd
AB-Mix1	Turkey, Georgia	1:20	Georgia		
AB-Mix2	Turkey, Georgia	1:10			
AB-Mix3	Turkey, Georgia	1:5			
AB-Mix4	Turkey, Georgia	1:2			
AB-Mix5	Turkey, Georgia	1:1			
BA-Mix1	Georgia, Turkey	1:20			
BA-Mix2	Georgia, Turkey	1:10			
BA-Mix3	Georgia, Turkey	1:5			
BA-Mix4	Georgia, Turkey	1:2			
CD-Mix1	Costa Rica, Trinidad & Tobago	1:20	Trinidad & Tobago	Brazil	
CD-Mix2	Costa Rica, Trinidad & Tobago	1:10			
CD-Mix3	Costa Rica, Trinidad & Tobago	1:5			
CD-Mix4	Costa Rica, Trinidad & Tobago	1:2			
CD-Mix5	Costa Rica, Trinidad & Tobago	1:1			
DC-Mix1	Trinidad & Tobago, Costa Rica	1:20	Honduras		Trinidad & Tobago
DC-Mix2	Trinidad & Tobago, Costa Rica	1:10	Trinidad & Tobago	Brazil	
DC-Mix3	Trinidad & Tobago, Costa Rica	1:5			
DC-Mix4	Trinidad & Tobago, Costa Rica	1:2			
EFG	Cyprus, Pakistan, Macedonia	1:1:1	Pakistan		
FGH	Pakistan, Macedonia, Oman	1:1:1	Pakistan		
GHI	Macedonia, Oman, Hungary	1:1:1	Hungary	Croatia	Hungary
HIJ	Oman, Hungary, Nigeria	1:1:1	Hungary		Croatia
SL1	Colombia, Malaysia	1:1:1	Malaysia		
SH1	South Africa, New Zealand, Australia	1:1:1	New Zealand		
SEAS1	South Korea, Vietnam, Malaysia	1:1:1	Malaysia		
AS1	Kazakhstan, Pakistan, Georgia	1:1:1	Pakistan		
AF1	Ghana, Nigeria, South Africa	1:1:1	Ghana		
EU1	Czechia, Hungary, Macedonia	1:1:1	Hungary		Croatia
ME1	Jordan, Qatar, Oman	1:1:1	Kuwait		
AM1	Mexico, Costa Rica, Colombia	1:1:1	Mexico		

Table 4: 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.

The contextually themed mixtures consisted of eight samples 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). The majority of these samples generated accurate predictions of a country of origin to one of the countries included within the mixture as the top three predictions (Table 4). 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 one sample predicted incorrectly to a country not included within the sample was from the Middle Eastern sample, which 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 all of these samples did not represent more than one single country component.

None of the outgroup samples were correctly predicted to country of origin (Panama or USA). Locally collected swab samples were predicted as originating from Argentina or Uruguay, with tape samples from the same collection points predicted more sporadically to include South Korea, Oman, Bahrain, Argentina, Djibouti and South Africa (Table 5). Indoor swabs originating from Panama were predicted as originating from Malaysia, Somalia and Trinidad and Tobago. Therefore, none of the outgroup samples were correctly predicted to the true country of origin.

Demonstration testing illustrated the use of the method on small-scale sample processing and revealed that the model is highly accurate when applied to independent single-source swab samples obtained from countries that were used in model training and development. When applied to artificial mixtures composed from countries that the model has 'seen' before in training and development, sample origin was predicted correctly to one of the composite countries for the majority of samples applied. Beyond this, no pattern was observed in regards to proportions or country combinations used, demonstrating that this method requires further evaluation for the application to mixed samples. Outgroup testing demonstrated that while the model performs well when analyzing material from countries used in training and development, swabs from countries that have not been analyzed before cannot be predicted correctly (as would be expected, highlighting the importance of compiling a comprehensive database). For a more thorough report on the demonstration testing please refer to the standalone demonstration testing report.

ID:	Country:	Details:	Total:	Global:	Best	2nd	3rd
Env tape 1	USA	Tape lift, Site 1	86958	29213	South Korea		
Env tape 2		Tape lift, Site 1	16489	13215	Oman	Bahrain	
Env tape 3		Tape lift, Site 1	123864	88836	Argentina		
Env tape 4		Tape lift, Site 2	63746	44696	South Korea		
Env tape 5		Tape lift, Site 2	31722	3014	-		
Env tape 6		Tape lift, Site 2	72802	33632	Djibouti	South Africa	
Env swab 1		Outdoor swab, Site 1	78406	64424	Argentina		Uruguay
Env swab 2		Outdoor swab, Site 1	103975	26718	Argentina		
Env swab 3		Outdoor swab, Site 1	100171	92310	Argentina		Uruguay
Env swab 4		Outdoor swab, Site 2	84535	55951	Argentina		
Env swab 5		Outdoor swab, Site 2	68253	24245	Argentina	Uruguay	
Env swab 6		Outdoor swab, Site 2	82285	51038	Argentina		
Pan1	Panama	Indoor swab	88420	67579	Malaysia		
Pan2		Indoor swab	41587	3914	Somalia	Djibouti	
Pan3		Indoor swab	95445	7772	Trinidad & Tobago	Brazil	Guyana

Table 5: Outgroup sample details, sequence counts (as final output results total and of the total, those that match to sequences in the comparative global database), 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 interface.

8. Documentation and Training

8.1. Laboratory SOP

Documentation detailing the standard operating procedure (SOP) for sample processing was developed as a required project deliverable. The expected scope of this document was to describe the small-scale forensic adaptation of sample processing based on the large-scale sample processing as used in the research and model development phase of this project. Additionally, the final SOP was to be used to direct the processing of samples to be used in the demonstration testing. The final SOP documentation includes small-scale laboratory processing instruction as intended, as well as sequence data analysis guidance to culminate test sample processing in data preparation for model implementation. The SOP document was drafted and used during the on-site training, with adjustments and edits made following training completion. The final SOP document is 27 pages in length and includes the following main sections:

- Purpose statement, materials and equipment, table of contents

- Introduction (Background, Sequence Data Treatment)
- Methods I (Sampling and DNA Extraction)
- Methods II (Sample Amplification and Library Preparation)
- Methods III (Library MiSeq Loading)
- Methods IV (Sequence Data Pipeline)
- Positive Controls
- Data Output Results
- References, Terminology and Acronym's, Manufacturer Instruction Documents, Appendices

8.2.Software Manual

An initial draft of a manual describing in detail how to use the ASVTracer software and interpret the model output was prepared and shared with the end user. This includes all of the information a user needs to follow the workflow outlined in Section 6 and to access the resulting text and PDF files; guidelines for the interpretation of the model output and the associated map display; and a sample PDF file generated by the software.

Feedback from the end user as to the content of the manual was subsequently incorporated into the finalized software manual, which was delivered along with the final version of the software.

8.3.Training

On-site training was conducted at Fort Gillem, Forest Park, Georgia from the 23 – 25 October 2018. The objective of this training was to detail and familiarize users with the laboratory SOP for test sample processing, the model prediction software, and to present the findings of the demonstration testing. Training began with an overview and background of the geolocation project, followed by a review of the laboratory processing and data analysis methods detailed in the SOP. Demonstration results were presented and discussed, followed by presentation of statistical background as relevant to the final developed model, and user guidelines for model software use. This training included live loading of a prepared library onto the Illumina MiSeq, as well as live demonstrations of the SOP data processing pipeline (using a small set of samples), and live usage of the software for geolocation prediction.

On-site training resulted in the generation of feedback for constructive edits to all final documents, including the SOP and software manual. Laboratory SOP included updates regarding explicit step by step loading of the MiSeq as a result of on-site participation. Software demonstration resulted in suggestions for adjustments of various options within software, output file edits in keeping with end-user requirements, and requests for further elaboration of data output interpretation within the software manual.

9. Summary

This study represents major accomplishments and discoveries that help advance capabilities in trace evidence analysis. To date, this project demonstrates that largest OCONUS sampling of environmental dust signatures, and subsequent genetic analysis. The power of this dataset was reflected in building a cutting-edge model to make predictions of geographic origin for unknown samples. The study team demonstrated that the fungal genetic marker ITS1 is very effective as a highly diverse biomarker with geographic origin information that can be used in deep learning (artificial intelligence) models. Here, the spatially aware deep learning model, DeepSpace, was custom designed to be implemented for the specific use of geographic prediction from NGS analyzed environmental samples. The model was highly accurate for country of origin (~86%) and was readily developed into end-user designed software for routine analysis. Demonstration testing showed that single source samples analyzed with the ITS1 NGS protocol and the ASVTracer software could be readily predicted to the correct country of origin, while mixed source samples presently do not yield correct mixed source predictions. In conclusion, this study shows the extreme potential of NGS genetic analysis, coupled to artificial intelligence, to provide mission critical information for providence of materials and opens the doors to transition of this technology into routine casework.

Bibliography

Software packages and consortia

R Markdown - Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., … Chang, W. (2018). rmarkdown: Dynamic Documents for R. (Version R package version 1.10.). Retrieved from https://CRAN.R-project.org/package=rmarkdown

DADA2 - Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13(7), 581–583. https://doi.org/10.1038/nmeth.3869

QIIME2 - Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7(5), 335–336. https://doi.org/10.1038/nmeth.f.303

R and R Studio - R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/; RStudio. (2018). RStudio: Integrated development environment for R. Boston, MA. Retrieved from http://www.rstudio.org/

Earth Microbiome Project - Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., … The Earth Microbiome Project Consortium. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. Nature, 551, 457.

Scientific publications

Abarenkov, K., Nilsson, R. H., Larsson, K. H., Alexander, I. J., Eberhardt, U., Erland, S., ... & Sen, R. (2010). The UNITE database for molecular identification of fungi–recent updates and future perspectives. New Phytologist, 186(2), 281-285.

Barberán, A., Ladau, J., Leff, J. W., Pollard, K. S., Menninger, H. L., **Dunn, R. R.**, & **Fierer, N.** (2015). Continental-scale distributions of dust-associated bacteria and fungi. *Proceedings of the National Academy of Sciences*, *112*(18), 5756-5761.

Craine, J. M., Barberán, A., Lynch, R. C., Menninger, H. L., **Dunn, R. R.**, & **Fierer, N.** (2017). Molecular analysis of environmental plant DNA in house dust across the United States. Aerobiologia, 33(1), 71-86.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460-2461.

Grantham, N. S., **Reich, B. J.**, Pacifici, K., **Laber, E. B.**, Menninger, H. L., Henley, J. B., ... & **Dunn, R. R.** (2015). Fungi identify the geographic origin of dust samples. PLoS One, 10(4), e0122605.

Grantham N.S., Reich B.J., Laber E.B., Pacifici K., **Dunn R.R., Fierer N., Gebert M., Allwood J.S., & Faith S.A.** Forensic Geolocation with Deep Neural Networks. J. Royal Statistical Society: Series C (2019) - submitted.

McGuire, K. L., Payne, S. G., Palmer, M. I., Gillikin, C. M., Keefe, D., Kim, S. J., ... & Massmann, A. L. (2013). Digging the New York City skyline: soil fungal communities in green roofs and city parks. PloS one, 8(3), e58020.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol., 73(16), 5261-5267.

Appendix A

Table A1: Countries sampled for the final ITS1 dataset.

	Count of Swabs Received	**Count of Swabs Sequence Analyzed**
Afghanistan	5	5
Antalya, Turkey	10	8
Argentina	20	15
Australia	21	16
Azerbaijan	20	15
Bahrain	18	17
Brazil	30	30
Colombia	20	15
Costa Rica	20	15
Croatia	20	15
Cyprus	19	14
Czechia	20	15
Djibouti-Somalia	20	18
Georgia	20	15
Ghana	20	15
Honduras	21	18
Hungary	20	15
Istanbul, Turkey	10	8
Jordan	20	15
Kazakhstan	19	14
Kuwait	20	15
Macedonia	20	15
Malaysia	20	15
Mexico	19	14
Moldova	20	17
New Zealand	14	9
Nigeria	20	16
Oman	20	15
Pakistan	20	15
Qatar	20	15
South Africa	18	13
South Korea	20	15
Trinidad and Tobago	20	15
Uruguay	20	15
Vietnam	20	15
Grand Total	**664**	**517**