Complex DNA Mixture Analysis

Darrell O. Ricke, Martha S. Petrovick, Catherine R. Cabrera, Eric D. Schwoebel, and James C. Comolli

Lincoln Laboratory has developed a suite of technologies that enables rapid analysis of samples that contain DNA from multiple sources. These new techniques can process 100 million DNA sequences in five minutes. Moreover, the Laboratory's system compares a DNA profile against a database of 20 million profiles in five seconds. These capabilities are aimed at helping law enforcement and national security professionals expeditiously identify suspected criminals or terrorists, or their relatives.

» **Of all the leave-behind signatures for** forensics- and biometrics-reliant missions, only DNA can link an individual to a crime with a high degree of certainty. Complex human DNA mixtures from three or more contributors have proven difficult to analyze with current methods. This difficulty applies especially for touch samples, with low DNA input, and the type of DNA mixtures that are often obtained in realistic criminal and intelligence investigations, such as those involving gang-owned guns, drugs, illicit cash, terrorist devices, and terrorist attack staging areas (Figure 1). Current methods are not capable of matching a reference profile to a low-concentration mixture containing DNA from four or more individuals, such as might be collected from a door knob, shared cell phone, or currency. Obtaining usable DNA profiles from a much higher proportion and much wider variety of samples than currently achievable would have a significant positive impact on forensic analyses and intelligence collection. Lincoln Laboratory is developing new tools that enable the analysis of complex DNA mixtures with up to 10 contributors. Researchers at the Laboratory are developing new technologies and analysis capabilities for unmet DNA forensics needs of the Department of Defense and law enforcement. Using a theoretical framework [1], Lincoln Laboratory has demonstrated identification of as many as 10 individuals in touch mixtures from a variety of substrates using 2,311 single nucleotide polymorphisms (SNPs), massively parallel sequencing (MPS), and advanced bioinformatics tools. With this approach, at least one person was identified at a probability of random man not excluded (P(RMNE))

FIGURE 1. Because of limitations with the current DNA analysis methods, forensic scientists have focused their efforts on samples that are likely to be from a single source (a). However, for many realistic settings, such as a terrorist attack staging area (b), in which one needs to match a suspect's DNA, single-source samples are rare or nonexistent.

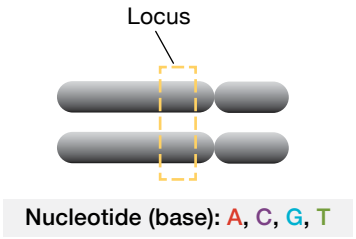

FIGURE 2. A locus is a position on a chromosome. A nucleotide, or base, is the type of monomer that is contained in the polymer that is DNA. The four types of nucleotides in DNA are commonly abbreviated A, C, G, and T.

that is less than 1×10^{-9} in 97 out of 100 touch samples containing 3 to 12 contributors. Very few, if any, of these samples would be amenable to analysis using current short tandem repeat (STR) methods.

All human cells, with the exception of germline cells, contain 23 pairs of chromosomes. A specific location on a chromosome is called a locus, and a particular DNA sequence that might be found at a specific locus is called an allele (Figure 2).

Current DNA forensic techniques rely upon determining the STRs (Figure 3). An alternative approach for complex DNA mixture analysis proposed by Voskoboinik and Darvasi [1] uses a large panel of rare SNPs with low minor allele frequencies (Figure 3). On the basis of this theoretical framework, Lincoln Laboratory implemented and validated MPS SNP panels for complex DNA

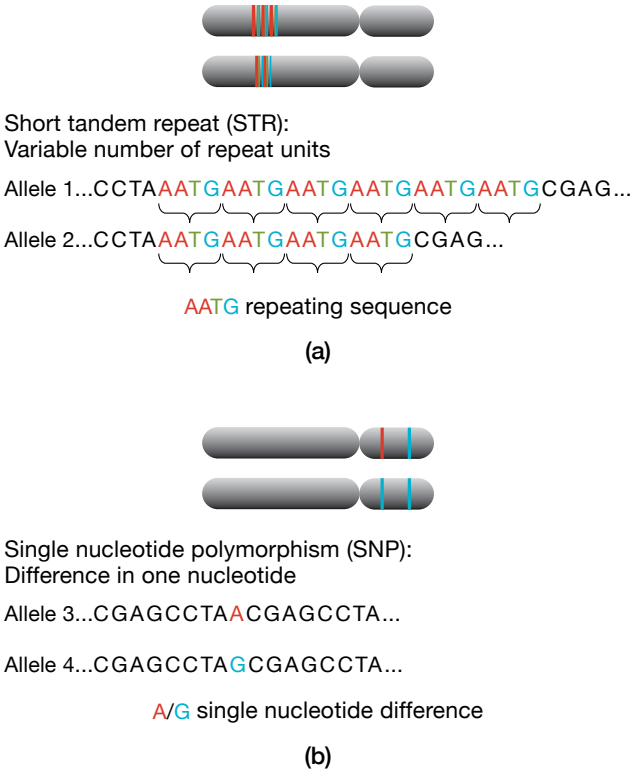

FIGURE 3. Two different classes of DNA variants are used for forensic analysis and discussed in this article: (a) short tandem repeats (STRs) and (b) single nucleotide polymorphisms (SNPs). The STRs are repeats of four to five bases, and the number of repeats differ between people (Allele 1 and Allele 2), while SNPs vary at a single base (Allele 3 and Allele 4).

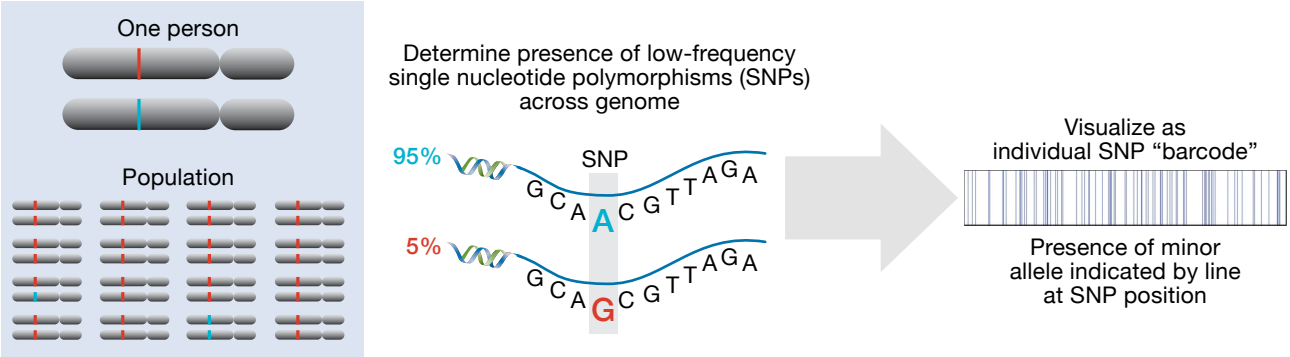

FIGURE 4. Each person within the population has a unique combination of rare SNPs (shown in blue) that can be imagined as a digital barcode.

mixture analysis and kinship analysis. More specifically, we designed a panel of 2,311 SNPs with low minor allele frequencies that tend to be consistent between populations to enable complex mixture analysis across multiple ethnicities. Selection of rare SNPs (~5 percent frequency in the target population) creates unique minor allele signatures (i.e., barcodes) for individuals and enables effective differentiation of multiple barcodes in a mixture (Figure 4). Using high-throughput sequencing to generate these SNP data allows for thousands of sequence reads at each SNP locus, thereby enabling sensitive detection of minor contributors.

Experiments were performed to determine the ratio of DNA at which a contributor can be detected in two-person mixtures with this SNP panel. In contrast to the current lower limit for mixtures analyzed by STR sizing of roughly one in 20, mixture contributors have been detected down to one in 400 [2]. For standard SNP mixture analysis [3], the upper limit for the number of contributors that can be identified in the same mixture is about 9 to 10 [2]. In addition, we developed an investigative method (TranslucentID) that enables the identification of contributors in mixtures of 10 to 20 contributors [4]; see Figure 5.

After establishing that the Lincoln Laboratory MPS SNP approach to complex mixture analysis performs well on laboratory-generated, i.e., controlled, samples, an extensive set of experiments was performed to assess the method's performance with uncontrolled touch samples. Complex touch samples mixtures were generated from individuals handling different objects multiple times over the course of days or weeks. The individuals logged each touch so that the analytical results could be compared to truth data. As shown in Figure 6, Lincoln Laboratory's MPS SNP approach successfully identified significantly more contributors to these complex mixture samples than current forensic approaches can detect, suggesting a potential paradigm shift in forensic DNA analysis that would enable successful analysis and generation of actionable data from samples that would currently go uncollected or unanalyzed.

Sample analysis time is another important consideration. Transitioning from STR-based panels with fewer than 20 loci to SNP panels with thousands to tens of thousands of loci represents computational scaling issues that that have been resolved by Lincoln Laboratory through the development of a suite of analysis tools, as described below. Analysis time includes sequence alignment, allele (DNA variant) calling, identification searches, and mixture analysis. Typically, MPS runs generate as many as 100 million DNA sequences. We developed GrigoraSNP (grígora is Greek for fast) for rapid alignment and allele calling of MPS SNP sequences [5]. GrigoraSNP typically processes 100 million sequences in approximately five minutes using four computational threads [5].

We developed FastID for rapid analysis of large SNP panels for both identification and mixture analysis [6]. FastID bit encodes SNP profiles and implements mixture analysis in three hardware instructions (exclusive OR/XOR, logical AND, and population count). FastID can compare a profile against a database of 20 million profiles in five seconds by using only a single computational thread. In addition to FastID for SNP

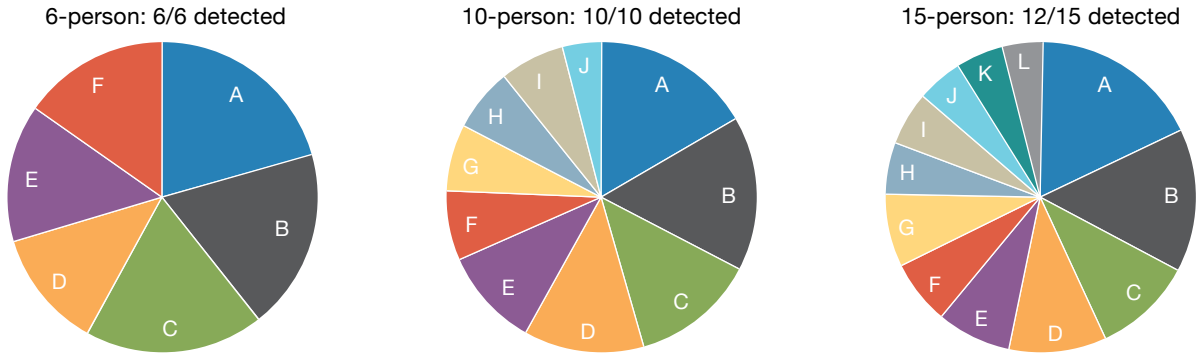

FIGURE 5. Shown are results from lab-generated equimolar mixtures of human DNA. For each pie chart, the wedge size corresponds to each subject's (identified by letter) estimated relative DNA contribution to the mixture. For the 6-person figure, the probability of random man not excluded P(RMNE) is 2.1e-54 for each of the identified individuals; for the 10-person figure, the P(RMNE) values are in the range of 5e-10 to 5e-21; and for the 15-person figure, the P(RMNE) values are in the range of 2.2e-8 to 4.5e-24.

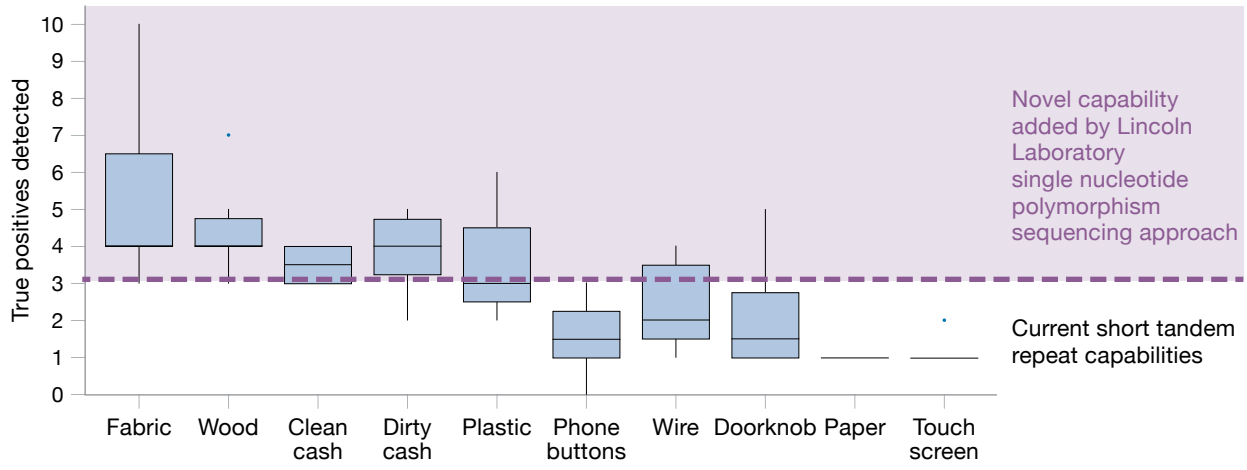

FIGURE 6. This summary of results from multiple experiments analyzing touch samples from different materials shows that the Laboratory's single nucleotide polymorphism (SNP) approach is successful with mixtures that are too complex to analyze with currently used methods.

analysis, we developed TachysSTR (tachýs is Greek for rapid) for working with STR profiles; together, FastID and TachysSTR were a 2018 R&D 100 Award winner [7] for rapid DNA forensic profile analysis. Increasing the number of DNA loci strained the statistical analysis of results in terms of computational time and numerical precision [8]. We developed Fast P(RMNE) (probability of random man not excluded) to reduce the computational time and avoid the numerical precision problems encountered with working with tens of thousands of loci [8].

In addition to developing custom methods for generating optimized SNP data and analytical tools to extract useful information from that data, Lincoln Laboratory developed the IdPrism Advanced DNA Forensics system to integrate the MPS SNP analysis tools into a graphical user interface system coupled to a relational database. As part of this system, a Linux program script runs every five minutes checking for completed sequencing runs. The binary format MPS sequence data files are transferred from the sequencer to the analysis system, converted to text format, processed by GrigoraSNPs, and uploaded to

the IdPrism database. New references are compared to all references and also compared to all mixture profiles with FastID [6] and Fast P(RMNE) [8]. New mixtures are compared to profiles of all known references with FastID and Fast P(RMNE). In addition, close relatives are detected by kinship analysis [9].

Technology advancements are enabling new DNA forensics capabilities. Our SNP mixture panel enables the characterization of complex DNA mixtures of 10 or more individuals. IdPrism has been successfully transitioned to the FBI Research Laboratory for evaluation, and discussions for commercialization have been initiated. Future expansion capabilities include customized MPS panels of tens of thousands of loci for improved prediction biogeographic ancestries for individuals, extensions of kinship for more distant relatives, and more. ■

References

1. L. Voskoboinik and A. Darvasi, "Forensic Identification of an Individual in Complex DNA Mixtures," *Forensic Science International: Genetics*, vol. 5, no. 5, 2011, pp. 428–435.
2. D. Ricke, P. Fremont-Smith, J. Watkins, T. Boettcher, and E. Schwoebel, "Estimating Individual Contributions to Complex DNA SNP Mixtures," *Journal of Forensic Sciences*, vol. 64, no. 5, 2019, pp. 1468–1474.
3. J. Isaacson, E. Schwoebel, A. Shcherbina, D. Ricke, J. Harper, M. Petrovick, et al., "Robust Detection of Individual Forensic Profiles in DNA Mixtures," *Forensic Science International: Genetics*, vol. 14, 2015, pp. 31–37.
4. D.O. Ricke, J. Watkins, P. Fremont-Smith, M.S. Petrovick, T. Boettcher, and E. Schwoebel, "TranslucentID: Analysis of Complex DNA SNP Mixtures with Large Numbers of Donors," *Australian Journal of Forensic Sciences*, published Dec. 2019 online at doi: 10.1080/00450618.2019.1699958.
5. D.O. Ricke, A. Shcherbina, A. Michaleas, and P. Fremont-Smith, "GrigoraSNPs: Optimized Analysis of SNPs for DNA Forensics," *Journal of Forensic Sciences*, vol. 63, no. 6, 2018, pp. 1841–1845.
6. D.O. Ricke, "FastID: Extremely Fast Forensic DNA Comparisons," paper in *Proceedings of the 2017 IEEE High Performance Extreme Computing Conference*, 2017.
7. L. French, "Creating 'DNA Barcodes,' Researchers Work to Speed up Forensic Analysis," *R&D Magazine* online at https://www.rdmag.com/news/2019/02/creating-dna-barcodes-researchers-work-speed-forensic-analysis, posted 14 Feb. 2019.
8. D. Ricke and S. Schwartz, "Fast P(RMNE): Fast Forensic DNA Probability of Random Man Not Excluded Calculation" [version 1; referees: awaiting peer review], 2017, available as DOI:10.12688/f1000research.13349.1.
9. B.S. Helfer, P. Fremont-Smith, and D.O. Ricke, "The Genetic Chain Rule for Probabilistic Kinship Estimation," bioRxiv. 2017, available at https://www.biorxiv.org/content/10.1101/202879v3.

Appendix

Case Study: DNA Sequencing for Forensic Geographic Attribution

Lincoln Laboratory is developing tools that use next-generation DNA sequencing technology to enhance geographic source attribution, that is, the determination of the origin of an item of interest. Knowledge gained from tracking objects back to their source, such as the site of device manufacture or assembly or the launch location of an airborne vehicle, could help characterize and eventually disrupt asymmetric threat networks. Geographic source attribution is most often used in criminal forensics for evidence and investigation, and accomplished through detailed analysis of dust. The particles within dust are derived from environmental sources, such as soil, plants, and animals, or from human activity, which are all particular to a geographic location, so can serve as a rich source of information about the site of origin. However, forensic dust analysis is currently performed by manual microscopy, so is laborious, requires extensive expertise, and takes weeks to complete. This relatively long time-to-answer greatly limits the use of geographic attribution for intelligence or tactical decision-making purposes for which an answer is needed quickly. We sought to address this limitation by developing tools that can rapidly provide attribution information from a large number of samples.

Our team focused on using a technique called metabarcoding to analyze DNA in dust on objects, in particular DNA from traces of plants and fungi. DNA barcodes are small regions of the genome that are shared by groups of organisms (in this case plants or fungi) but whose sequence differs by species. Metabarcoding involves the simultaneous analysis of multiple barcodes. By amplifying multiple plant and fungal barcodes from trace DNA extracted from dust on objects (metabarcoding) and then subjecting the barcodes to next-generation sequencing, we can rapidly obtain the signatures of plant/fungal DNA, corresponding to hundreds of species per sample, in numerous samples in parallel. These signatures were used for attribution in two ways: first for an unguided query to address the question, "Where in the world did this object reside?" and second for a guided query to compare samples from different objects to determine their common origin, in other words to address the question, "Did this sample come from a similar location as that sample?"

For unguided attribution, plant sequencing reads derived from dust metabarcoding were used to identify the originating plant species by matching to sequence records in the National Center for Biotechnology Information sequence reference database. Then, the known geographic distributions of the identified plant species were retrieved from a publicly available biogeography database Biodiversity Information Serving Our Nation (BISON). We overlaid the individual distribution maps in a geographic information system to produce a point-to-grid heat map indicating the regions where the distributions overlapped (Figure A, top). This map gave an estimate of the geographic origin of the dust sample. By comparing the estimate of our dust sample's origin to the known origin of a control dust sample, we were then able to assess the accuracy and resolution of our method.

Roughly one-third of dust samples collected from four different U.S. locations over a year provided accurate regional attribution, defined as accurate geolocation to within 600 kilometers of the site of origin. This was the first demonstration that an estimate of geographic attribution could be achieved rapidly from crude plant DNA found in dust. The accuracy of the technique depended on the number of species identified, which was significantly lower in the winter months, and on the location from which the dust sample was collected. For instance, with dust collected during spring and summer months at sites in New Mexico or South Carolina, greater than 70 percent of the samples provided accurate geolocation information. The accuracy and applicability were limited by the plant species available in both the sequence and biodistribution databases, so expansion of these databases to include more species at better resolution could result in significant improvements to the percentage of samples

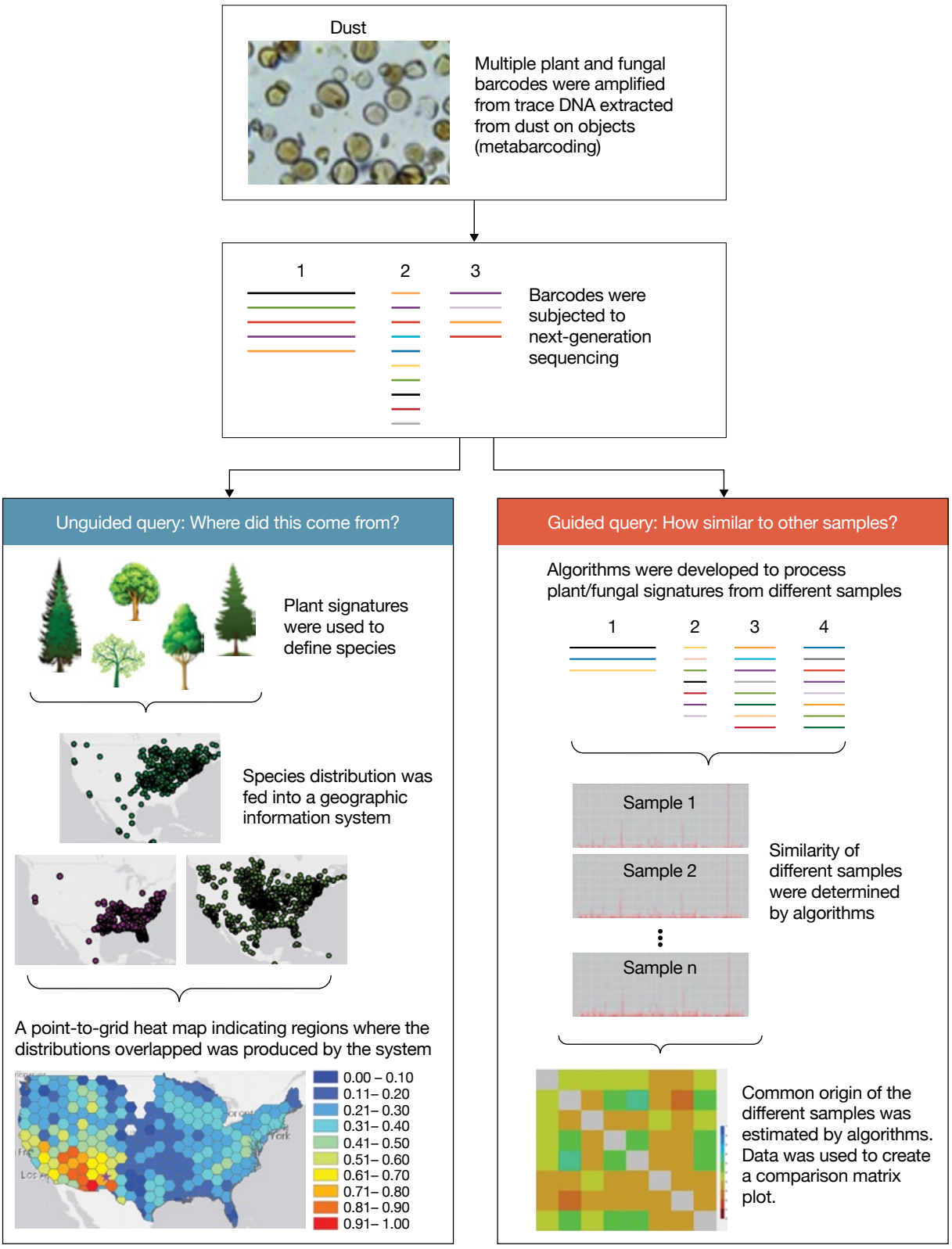

FIGURE A. Shown is a geographic attribution pipeline for an unguided and a guided query from dust samples. For the unguided query (left), plant signatures were used to define species, the distribution of which was fed into a geographic information system to produce a point-to-grid heat map indicating the regions where the distributions overlapped. For the guided query (right), algorithms were developed to process plant/fungal signatures from different samples, determine their similarity, and estimate their common origin.

that yield accurate attribution and to the geolocation resolution that could be achieved.

The comparative guided attribution tool (Figure A, bottom) was designed to rapidly compare the plant and fungal sequence reads obtained from metabarcoding in different samples, determine their similarity, and estimate their common origin. Because each sample produced hundreds to thousands of different plant and fungal barcode reads, we compressed these reads into a metabarcode "signature" by employing a shingling method similar to that used for email spam detection that computationally compares signatures to determine their similarity. To test the method, we compared signatures from dust on samples that were left in Massachusetts for one week then relocated to South Carolina for one week to those from dust on samples that remained in a single location (Massachusetts or South Carolina) for one week. As expected, dust samples from objects at only one site had a significantly higher similarity to reference samples from that site than did dust samples that had been in different locations. However, samples that were in Massachusetts for as little as one day then relocated to South Carolina had similarity to reference samples from both sites, indicating that the metabarcode signature from Massachusetts was detectable after relocation. This finding suggested that the plant and fungal DNA within dust from a relocated sample enabled attribution back to its site of origin.

The two novel geographic attribution pipelines developed at Lincoln Laboratory provide initial proof of concept that environmental DNA contained within dust has utility in tracking samples and estimating their point(s) of origin. Both our work and recent technological advances in next-generation sequencing technology lead us to conclude that metabarcoding has the potential to serve as the basis for an automated system that can rapidly gather geolocation information on numerous samples or objects to provide information in a field-forward, operational setting. Further enhancement of the metabarcoding methodology, along with the expansion of the scope and accuracy of available sequence and biodistribution information, could improve attribution resolution to 50 kilometers or less within a well-referenced geographic area.

About the Authors

Darrell O. Ricke is a member of the technical staff in the Biological and Chemical Technologies Group at Lincoln Laboratory. He has broad experience in programming and software engineering, bioinformatics, molecular biology, genomics, and functional genomics. He has worked extensively on understanding the mechanisms of human diseases, biology discovery, data integration, and data mining. His current biomedical research focuses on applications in metagenomics for identification of infectious disease organisms, advanced DNA forensics, epigenetics, transcriptomics, proteomics, and disease mutation analysis. He is applying high-performance technologies to projects that include the application of current sequencing technologies to forensics and complex DNA mixtures, metagenomics, transcriptomics, epigenetics, development of multiple bioinformatics tools, and integrated data architecture design. He has been granted 10 patents and is the recipient of a 2018 R&D 100 Award for innovative methods for comparing DNA samples. He holds bachelor's degrees in computer science and in genetics and cell biology, and a master's degree in computer science from the University of Minnesota, and a doctorate in molecular biology from the Mayo Graduate School.

Martha S. Petrovick, a member of the technical staff in the Biological and Chemical Technologies Group, joined Lincoln Laboratory in 1998 to develop a cell-based sensor for biological agents and became the cell engineering leader. She also led the forensic signature science task for the Accelerated Nuclear DNA Equipment program, which was sponsored by the U.S. Departments of Defense, Justice, and Homeland Security to develop automated, rapid human DNA profiling capabilities for field biometrics and forensics applications. She currently works on improving methods for analyzing forensic DNA samples and is investigating novel markers for physical and cognitive resilience. She earned bachelor's and master's degrees in pathobiology at the University of Connecticut, and a doctorate in cell and developmental biology at Harvard University.

Catherine R. Cabrera is the leader of the Biological and Chemical Technologies Group at Lincoln Laboratory. She joined the Laboratory in 2002, initially working on hardware and software development for the identification of biowarfare agents. She was part of the team that received an R&D 100 Award for the development of the PANTHER automated cell-based bioaerosol sensor, which has since transitioned to operational use for building protection and plant pathogen detection. She currently

oversees a diverse portfolio of programs that include ones on molecular biomarkers for heath and performance, advanced DNA forensics, and engineered and synthetic biology. Her areas of technical expertise include microfluidics; biodefense technologies, systems, and architectures; red/blue team analysis; microbiome and human health; point-of-need diagnostics; genetic and epigenetic biomarkers of health and activity; and use of physiological status indicators to provide early warning of exposure to chemical warfare agents or pathogens. She holds bachelor's degrees in biochemistry and chemical engineering from Rice University and a doctorate in bioengineering from the University of Washington. Her doctoral research was focused on developing fieldable technologies to detect pathogens in resource-limited environments.

Eric D. Schwoebel is a member of the technical staff in the Biological and Chemical Technologies Group, where he has recently worked on analysis of DNA mixtures, assessment of kinship between individuals, and prediction of physical characteristics from DNA markers. He joined Lincoln Laboratory in 2001 to assist in the development of cell-based sensors for the detection and identification of biological warfare agents. He received a doctorate in cell biology from Baylor College of Medicine, worked in vaccine development at the Institute of Primate Research in Nairobi, Kenya, and returned to Baylor College of Medicine to examine the biochemistry of protein transport between the cytoplasm and nucleus.

James C. Comolli, a member of the technical staff in the Biological and Chemical Technologies Group, joined Lincoln Laboratory in 2015. His research interests include infection therapeutics and diagnostic technologies, biosensors, bacterial communities, pathogen-host interactions, and synthetic biology. Prior to joining the Laboratory, he was the leader of the Biomedical Engineering Group at Draper Laboratory; research director at ECI Biotech, a startup developing microbial diagnostics; and a staff researcher at Johnson & Johnson, working on medical devices. He has a bachelor's degree in biology from Johns Hopkins University and a doctoral degree in cellular and molecular biology from Harvard University. He completed postdoctoral training in bacterial pathogenesis at the University of California, San Francisco, and in microbial physiology at the University of Wisconsin–Madison.