OSAC RESEARCH NEEDS ASSESSMENT FORM

Title of research need: Development of infrastructure to compile and share raw electronic data for training and tool development

Describe the need:

We have entered the age of "big data" and artificial intelligence. The forensic DNA community has yet to take full advantage of current computational power, methods and establish an efficient means of sharing data. Unfortunately, the limited availability of large and diverse DNA profile databases currently hinders the implementation of computationally advanced solutions that can profoundly impact DNA mixture interpretation and impede wide scale assessments of currently used methods of interpretation. The value of "big data" has only been recently realized in the forensic community through development and assessment of new software tools and statistical models that have already made an impact on the quality of forensic DNA analysis. However, the means to develop these tools and assessment methods are reliant on the availability of large-curated datasets.

In September 2016 the *President's Council of Advisors on Science and Technology Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*[1] was published to advise the president and the nation on the state of forensic science in the United States and to identify areas of improvement in the scientific methods and the interpretation of forensic data. The advisory group specifically identified a need for publicly available, large collections of DNA profiles to serve as an educational resource for public and private forensic laboratories as well as independent entities. The report stated, *"Such efforts will be aided by the creation and dissemination (under appropriate data-use and data-privacy restrictions) of large collections of hundreds of DNA profiles created from known mixtures—representing widely varying complexity with respect to (1) the number of contributors, (2) the relationships among contributors, (3) the absolute and relative amounts of materials, and (4) the state of preservation of materials—that can be used by independent groups to evaluate and compare the methods. In addition to scientific studies on common sets of samples for the purpose of evaluating foundational validity, individual forensic laboratories will want to conduct their own internal developmental validation studies to assess the validity of the method in their own hands*"[1].

Although many research, developmental and validation projects have compiled large datasets, many are not available due to confidentiality or other related concerns. In addition, these sets may lack the diverse and comprehensive types of DNA profiles (or other related types of data) that would permit robust software development of comparative studies. This research need calls for the development of large and diverse data sets that are accessible to the public and the infrastructure to support such databases. The infrastructure includes (1) the development of computational resources to support housing of large data sets, (2) means to properly curate the databases—quality assurance/control methods and tools, documentation and monitoring throughout the "life" of the database, and (3) the development of legal, policy and data sharing framework to allow for the use and updating of databases among academic, governmental and private

institutions. The PROVEDIt dataset[2*] is a principal example of the type of database and architecture that can be used effectively to realize the benefits of these types of datasets/databases.

Keyword(s): Database, validation, NGS, DNA mixture, mixture interpretation, artificial intelligence

Submitting subcommittee(s): Human Biology **Date Approved:** 10/05/201

(If SAC review identifies additional subcommittees, add them to the box above.)

Background Information:

1. Does this research need address a gap(s) in a current or planned standard? (ex.: Field identification system for on scene opioid detection and confirmation)

Yes, ANSI/ASB Standard 020, Standard for Validation Studies of DNA Mixtures, and Development and Verification of a Laboratory's Mixture Interpretation Protocol

2. Are you aware of any ongoing research that may address this research need that has not yet been published (e.g., research presented in conference proceedings, studies that you or a colleague have participated in but have yet to be published)?

Yes, several research projects and proposals as well as mixture analysis reviews including the recent NIST Review on DNA Mixture Interpretation.

3. Key bibliographic references relating to this research need:

1) PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY. President's Council of Advisors on Science and Technology. *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. (2016).
2) Alfonse, L. E., Garrett, A. D., Lun, D. S., Duffy, K. R. & Grgicak, C. M. A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt. *Forensic Sci. Int. Genet.* **32,** 62–70 (2018).
3) Buitler, J. M., Iyer, H., Press, R., Taylor, M.K., Vallone, P.M. and Willis, S. DNA Mixture Interpretation: A NIST Scientific Foundation Review NISTIR 8351-DRAFT. (2021). https://doi.org/10.6028/NIST.IR.8351-draft.

4. Review the annual operational/research needs published by the National Institute of Justice (NIJ) at https://nij.ojp.gov/topics/articles/forensic-science-research-and-development-technology-working-group-operational#latest? Is your research need identified by NIJ?

No

5. In what ways would the research results improve current laboratory capabilities?

Generally, the development of these databases can impact the ability of researchers to develop new computational tools to better address the analytical and interpretational needs of the forensic DNA analyst.

[*] PROVEDIt Initiative (Project Research Openness for Validation with Experimental Data) is a publicly available resource of 25,000+ DNA profiles [1].

Specific examples, however, cannot be provided, the value in this is in providing the "raw materials" (data) to be able to explore new computational methods. Additional benefits include ability to use these data sets for training of new analysts and students training to become forensic practitioners.

6. In what ways would the research results improve understanding of the scientific basis for the subcommittee(s)?

New computational tools, specifically using artificial intelligence, are highly dependent on data availability and the quality of the data. If these databases are used for the training and testing of the newly developed computational tools, the subcommittee will have access to these data and thus be able to better understand the response, use and results of these computational tools.

7. In what ways would the research results improve services to the criminal justice system?

Provide the raw data needed to further develop new computational tools that will improve the confidence and quality of results. Additional benefits include ability to use these data sets for training of new analysts and students training to become forensic practitioners. These databases and methods will permit independent assessments of new and historically relevant computational tools, allowing for a direct comparison of different methods using the same data sets.

8. Status assessment (I, II, III, or IV): I

	Major gap in current knowledge	Minor gap in current knowledge
No or limited current research is being conducted	I	III
Existing current research is being conducted	II	IV

This research need has been identified by one or more subcommittees of OSAC and is being provided as an informational resource to the community.