Document Title:	Advanced Statistical Population Genetics Methods for Forensic DNA Identification
Author(s):	Noah A. Rosenberg, Ph.D.
Document Number:	253932
Date Received:	October 2019
Award Number:	2014-DN-BX-K015

Basic Research and Development in Forensic Science for Criminal Justice Purposes
Department of Justice, Office of Justice Programs
National Institute of Justice
NIJ SL # SL001082
NIJ-2014-3744

Award # 2014-DN-BX-K015

ADVANCED STATISTICAL POPULATION GENETICS METHODS FOR FORENSIC DNA IDENTIFICATION

Prepared by:

Noah A. Rosenberg, PhD
Principal Investigator
Department of Biology
Stanford University
371 Gilbert Building, Room 109
Stanford, CA 94305-5020
Tel: 650 721 2599
Email: noahr@stanford.edu

Prepared on: January 10, 2019

Recipient Organization:
Board of Trustees of the Leland Stanford Junior University
Stanford University
3160 Porter Drive, Suite 100
Palo Alto, CA 94304-8445

Final Progress Report
Project Period: 01/01/2015 – 12/31/2018

Signature of Submitting Official: Robert Loredo, Contract and Grant Officer

Robert Loredo 01/24/2019

Major project goals and objectives

The project focused on application of population-genetic methods of population structure analysis and genotype imputation to problems in forensic genetic analysis. The objectives of the work were (1) to determine the population structure information present in the CODIS forensic markers, (2) to examine cross-compatibility of microsatellite databases with new single-nucleotide polymorphism (SNP) markers by use of techniques of genotype imputation, and (3) to improve the population-genetic basis of advanced forensic genetics techniques such as relatedness profiling.

Research accomplishments

(1) Population structure and CODIS. We performed an extensive analysis of CODIS microsatellite genotypes in a worldwide panel of individuals from the Human Genome Diversity Panel. We focused on comparing the inference of population structure from CODIS microsatellites to similar inferences made on non-CODIS microsatellite sets of similar size, relying on the data from past microsatellite studies from our group (NA Rosenberg et al. 2002 *Science* 298:2381-2385; NA Rosenberg et al. 2005 *PLoS Genetics* 1:660-671; NA Rosenberg 2011 *Human Biology* 83:659-684; TJ Pemberton et al. 2013 *G3: Genes, Genomes, Genetics* 3:891-907).

Our comparative microsatellite analysis demonstrated that the CODIS microsatellites contain a nontrivial level of ancestry information, similar to that of random microsatellite sets. More generally, we found that the level of information about individual identity in a marker set is correlated with the amount of ancestry information. These findings were published in a leading high-impact journal (BFB Algee-Hewitt et al. 2016 *Current Biology* 26:935-942). They received news coverage in a story in *Forensic Magazine*.

(2) Record matching between CODIS profiles and SNP profiles. For our second major study of the worldwide panel of CODIS microsatellites, we showed that a CODIS genotype profile can be linked to a corresponding profile of SNP genotypes, even though the profiles share no markers in common. This "record matching" proceeds from the fact that correlations between microsatellites and SNPs enable partial imputations of the microsatellites from the SNPs. The imputed information available on microsatellite genotypes from SNP genotypes then accumulates across the microsatellites to permit profile matching between microsatellite and SNP databases. Thus, two databases, one on the CODIS microsatellites and one on SNP genotypes, could potentially be used to identify a SNP profile as belonging to the same individual as the contributor of a microsatellite profile.

This work, which enhances potential for design of backward-compatible new forensic marker systems and raises privacy issues in cross-database matching problems, was also published in a high-impact journal (MD Edge et al. 2017 *Proceedings of the National Academy of Sciences USA* 114:5671-5676). The study received news coverage, in *Forensic Magazine*, *Pacific Standard*, *Stanford Report*, and the *University of Michigan Health Lab Report*. It shed light on a previously unknown connection between forensic genetic data and biomedical and genealogical genetic data, a topic that subsequently came to widespread attention in 2018 when a similar type of connection was exploited for use in identifying crime suspects on the basis of publicly available genetic data of their distant relatives.

(3) Record matching of relatives between CODIS profiles and SNP profiles. Our third major study extended our analysis of record matching to close relatives. In particular, we demonstrated that a CODIS genotype profile can be connected to a SNP profile of a close relative, even though CODIS and SNP databases share no markers. The study capitalized both on correlations between microsatellites and SNPs and on classic theory of the effect of relatedness on the identity or non-identity of genotypes in distinct individuals. The study has as a consequence that SNP profiles can potentially be used to perform relatedness matching calculations against a database of STR profiles, and vice versa, with potential for success in a nontrivial fraction of cases. It also expands the privacy concerns revealed by our initial study of record matching between SNP and STR profiles.

We reported this paper in *Cell*, one of the leading journals in the biological sciences (J Kim et al. 2018 *Cell* 175: 848-858). We are pleased that the paper received widespread media coverage, both for its potential to improve forensic searches and for its attention to privacy concerns. News stories appeared in the scientific press, in *Nature*, *New Scientist*, *Science*, *Scientific American*, *Wired*, as well as in mainstream news outlets such as *CNN*, *PBS NewsHour*, and the *Wall Street Journal*.

Summary of research contributions. We are pleased to have completed major research papers concerning all three of our objectives. Broad attention to the work, as evidenced by its appearance in high-profile journals (*Cell*, *Current Biology*, *Proceedings of the National Academy of Sciences USA*), and news coverage of all three main papers, illustrates the interest in topics at the intersection of population genetics and forensic genetics and the potential of population-genetic approaches for advancing forensic genetics.

Training and professional development

The project afforded an opportunity to train postdoctoral researchers and PhD students at the intersection of population genetics and forensic genetics. Dr. Bridget Algee-Hewitt, Dr. Jaehee Kim, and Dr. Michael Edge all conducted research on the project.

Graduate student Michael Edge has graduated from Stanford with a Ph.D. in Biology, entitled “Pick up the pieces: combining information from multiple genetic loci,” and receiving the Department of

Biology's Samuel Karlin Prize for an outstanding Ph.D. thesis in mathematical biology. He has advanced to perform postdoctoral work at the University of California, Davis.

Postdoctoral researcher Bridget Algee-Hewitt has completed her postdoctoral training. She has advanced to a position as a Senior Research Scientist in the Center for Comparative Studies in Race and Ethnicity at Stanford University. Postdoctoral researcher Jaehee Kim is continuing in her studies.

Publications

We have reported three major publications, on population structure inference in the CODIS markers (Algee-Hewitt et al. 2016), record matching between CODIS and SNP databases (Edge et al. 2017), and record matching of relatives between databases (Kim et al. 2018). A fourth study extends the principles uncovered in our population structure analysis to produce general results concerning the behavior of the population structure statistic F_{ST} (Alcala & Rosenberg 2017).

BFB Algee-Hewitt, MD Edge, J Kim, JZ Li, NA Rosenberg (2016). Individual identifiability predicts population identifiability in forensic genetic markers. *Current Biology* 26:935-942.

MD Edge, BFB Algee-Hewitt, TJ Pemberton, JZ Li, NA Rosenberg (2017). Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *Proceedings of the National Academy of Sciences USA* 114:5671-5676.

N Alcala, NA Rosenberg (2017) Mathematical constraints on F_{ST}: biallelic markers in arbitrarily many populations. *Genetics* 206:1581-1600.

J Kim, MD Edge, BFB Algee-Hewitt, JZ Li, NA Rosenberg (2018) Statistical detection of relatives typed with disjoint forensic and biomedical loci. *Cell* 175: 848-858.

Dissemination in news articles (selected)

Beyond its dissemination in the peer-reviewed literature, our work has been described in a number of articles in the popular press. These include:

S Augenstein, *Forensic Magazine* "Does CODIS contain untapped ancestry information?" http://www.forensicmag.com/article/2016/04/does-codis-contain-untapped-ancestry-information, April 13, 2016.

S Augenstein, *Forensic Magazine*, "CODIS has more ID information than believed" https://www.laboratoryequipment.com/2017/05/codis-has-more-id-information-believed-scientists-find, May 15, 2017

E Callaway, *Nature,* "Supercharged crime-scene DNA analysis sparks privacy concerns," https://www.nature.com/articles/d41586-018-06997-8, October 11, 2018

R Lewis, *Genetic Literacy Project*, "Genetic privacy and the case of the Golden-State-Killer: diving into the science" https://geneticliteracyproject.org/2018/05/01/genetic-privacy-and-the-case-of-the-golden-state-killer-diving-into-the-science/, May 1, 2018

M Molteni, *Wired*, "Genome hackers show no one's DNA is anonymous anymore" https://www.wired.com/story/genome-hackers-show-no-ones-dna-is-anonymous-anymore/, October 11, 2018

P Raeburn, *Scientific American*, "How to identify almost anyone in a consumer gene database," https://www.scientificamerican.com/article/how-to-identify-almost-anyone-in-a-consumer-gene-database/, October 11, 2018

S Scutti, *CNN*, "You might not be anonymous, thanks to genealogy databases" https://www.cnn.com/2018/10/11/health/genetic-privacy-study/index.html, October 11, 2018

C Whyte, *New Scientist*, “Police can now use millions more people’s DNA to find criminals,” https://www.newscientist.com/article/2182348-police-can-now-use-millions-more-peoples-dna-to-find-criminals/, October 11, 2018

Presentations

We are also pleased to have been invited to share the results of our research in a variety of conference and university presentations.

Department of Biomolecular Engineering, University of California, Santa Cruz, departmental seminar presentation by Noah Rosenberg, March 2, 2017

Institute for Human Genetics, University of California, San Francisco, departmental seminar presentation by Noah Rosenberg, June 2, 2017

RECOMB-Genetics, conference keynote presentation by Noah Rosenberg, July 27, 2017

Forensic Technology Center of Excellence program of the National Institute of Justice, webinar presentation by Michael Edge, September 21, 2017

Institute for Pure and Applied Mathematics Conference on Algorithmic Challenges in Protecting Privacy for Biomedical Data, University of California, Los Angeles, conference presentation by Noah Rosenberg, January 11, 2018

American Academy of Forensic Science, conference presentation by Michael Edge, February 20, 2018

American Association of Physical Anthropologists, conference presentation by Michael Edge, April 12, 2018

Stanford Biostatistics Workshop, Stanford University, April 26, 2018, departmental seminar presentation by Noah Rosenberg, April 26, 2018

Impact

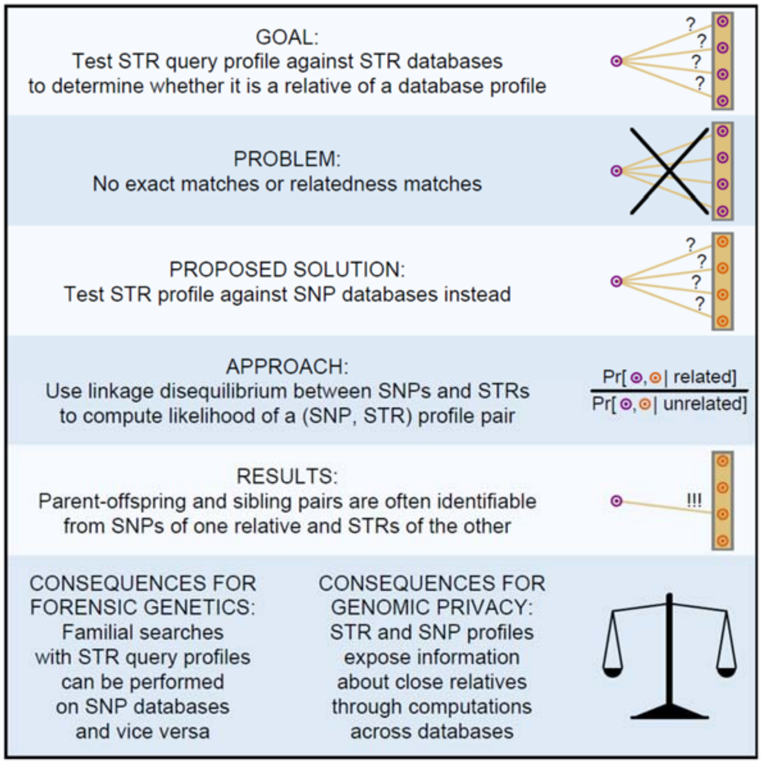
Dissemination of the results. Our work has achieved publication in high-impact journals and recognition in news stories, and it has been the focus of numerous conference presentations and invited lectures. The studies conducted on the project have also now received dozens of citations in the scientific literature, including in major journals in forensic science and forensic anthropology such as *Forensic Science International: Genetics* and the *American Journal of Physical Anthropology*.

The study of Edge et al. (2017) described new ways of connecting disparate genetic databases to test CODIS marker profiles for matches against biomedical or genealogical profiles, commenting on privacy concerns inherent in such methods. With the widely reported description of new forensic genetic techniques used in the "Golden State Killer" case—linking distinct genetic databases and uncovering new privacy concerns—we have seen much interest in our work. In April 2018, when the forensic genetics of the case came to public attention, usage of our article Edge et al. (2017), as tracked by the publishing journal, reached levels comparable to those seen soon after its initial publication. A news story in the Genetic Literacy Project, referring to our study, commented: "*A remarkable study published exactly a year ago revealed the strong correlations among different types of genetic markers that are closely linked on their chromosome*" (R Lewis, May 1, 2018). The news story noted our remark *"The potential for record matching of SNP and CODIS STR profiles, especially with augmented CODIS profiles, uncovers new risks to privacy…"* (Edge et al. 2017).

With the increased attention to relatedness and forensic genetics after the Golden State Killer case, our article of Kim et al. (2018) on record matching of relatives across databases received significant news coverage in dozens of scientific and mainstream publications. An article written by P. Raeburn in *Scientific American* quoted Prof. Bruce Weir of the University of Washington describing the work as "*Practically, it's an enormous advance*."

Development of human resources. The project has contributed to training two postdoctoral researchers and a graduate student in topics of forensic genetics. One postdoctoral researcher, Dr. Bridget Algee-Hewitt, is continuing as a Senior Research Scientist in the Center for Comparative Studies of Race and Ethnicity at Stanford University. A graduate student, Dr. Michael Edge, has graduated in 2016 and performed a brief postdoctoral stint on the project. He has now transitioned to a longer postdoctoral position in statistical population genetics at the University of California, Davis, starting during 2017, where he is further exploring the intersection of forensic genetics and human population genetics. Postdoctoral researcher Dr. Jaehee Kim is continuing her training on the project.

Societal impact. The studies from the project contribute to public discourse on cross-referencing forensic and non-forensic databases for forensic identification, and on privacy implications of such activities. These studies argue that the potential to link databases of forensic markers with databases used in genealogical and biomedical studies extends far beyond what had previously been claimed.



GOAL:
Test STR query profile against STR databases
to determine whether it is a relative of a database profile

PROBLEM:
No exact matches or relatedness matches

PROPOSED SOLUTION:
Test STR profile against SNP databases instead

APPROACH:
Use linkage disequilibrium between SNPs and STRs
to compute likelihood of a (SNP, STR) profile pair

RESULTS:
Parent-offspring and sibling pairs are often identifiable
from SNPs of one relative and STRs of the other

CONSEQUENCES FOR
FORENSIC GENETICS:
Familial searches
with STR query profiles
can be performed
on SNP databases
and vice versa

CONSEQUENCES FOR
GENOMIC PRIVACY:
STR and SNP profiles
expose information
about close relatives
through computations
across databases

As illustrated in the diagram above, the work provides a new method to search for relatedness matches in forensic genetic studies, especially in the scenario in which the crime scene DNA has been lost and the only DNA evidence that is available is a recorded STR profile. The work also highlights new ways in which privacy can be exposed for individuals whose relatives appear in an STR or SNP database.

The project has generated interest in communities working in crime laboratories, including through presentations with significant attendance of forensic DNA practitioners. Results have been cited by researchers in forensic centers, including the National Centre for Forensic Studies of Australia and the Australian Federal Police (Scudder et al. 2018 *Science & Justice* 58: 153-158; Scudder et al. 2018 *Forensic Science International: Genetics* 34: 222-230), and the Forensic DNA Laboratory of the University of the Western Cape, South Africa (Ristow et al. 2018 *Forensic Science International: Genetics* 24: 194-201).