Document Title:	Development of a Self-Sustaining, Open Access Forensic STR Sequence Diversity Database
Author(s):	Seth Faith Ph.D., Kelly Meiklejohn Ph.D.
Document Number:	254584
Date Received:	March 2020
Award Number:	2015-DN-BX-K062

Final Summary Overview

National Institute of Justice

NCSU/Forensic Sciences Institute – FSI

Award number 2015-DN-BX-K062

“Development of a Self-Sustaining, Open Access Forensic STR Sequence Diversity Database”

Prepared by: Seth Faith PhD (Co-PI)
faiths@battelle.org

Kelly Meiklejohn PhD (Co-PI)
kameikle@ncsu.edu

Submitted by: Sponsored Programs Office

Email: sps@ncsu.edu Telephone: (919) 515-2444

Submitted: March 26th 2019

Recipient Organization:

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

Project Period: January 1, 2016 – December 31, 2018

Stefanie D. Saunder
Associate Director, Operations
NCSU Sponsored Programs

Project Synopsis

Next-generation sequencing (NGS), also referred to as massively parallel sequencing (MPS), is evolving at a rapid pace and opening new doors to more sensitive methods for genetic typing of forensic samples. This technology allows for extremely high sample and genetic marker multiplexing of conventional forensic loci and also investigative data. In the scope of short tandem repeat (STR) typing, recent publications have described NGS methods and discovered sequence diversity in the core repeat and flanking regions of many forensically relevant loci that have not been previously reported or catalogued. Sequencing of STRs using NGS has permitted the identification of STR isoalleles, sequence variants that do not differ by size (length) but do differ in the underlying sequence. These isoalleles have excited the forensic community, who are eager to use this additional sample variation in instances where traditional capillary electrophoresis limited interpretation, such as complex mixtures and degraded samples.

However, no comprehensive resource was previously available that compiles allele frequencies of sequence-based STR loci across multiple populations. Thus, this research study led by North Carolina State University (NC State) aimed to characterize STR alleles from a large U.S. cohort and build an online database, The POPSeq Human STR Sequence Diversity Database (https://popseq.cvm.ncsu.edu).

Purpose: *To provide the forensics community an open-access, sustainable sequence-based population database for genetic markers produced with NGS.* The final product will be a self-sustaining resource that enables NGS technology to be applied to criminal casework, missing persons, paternity, cold and unsolved cases by providing a comprehensive database of STR loci. To achieve this outcome, this project had three goals.

Goal 1: To generate a robust and accurate human population STR sequence dataset using commercially available NGS kits.

Objectives: Human samples from 900 individuals from four population groups will be analyzed in duplicate using two NGS methods and quality reviewed for concordance and sequence accuracy.

Goal 2: To develop a cloud-based, open-access tool for visualizing population data, and performing genotype analysis at the locus sequence level.

Objectives: The information technology team at NC State will build a cloud-hosted database for STR data produced by NGS. The final database will provide advanced analytical capabilities to the forensics community including population statistics and sequence based matching analytics.

Goal 3: To provide a consolidated database for forensic laboratories producing NGS data.

Objectives: Upon completion of proficiency testing, external laboratories will be identified as contributors to the online database. The dynamic web application will allow automated data submission with computer-supported quality and content review to provide a community supported, sustainable and long-lived database resource.

Project Design and Methods

1.1 Population sampling

Anonymized genomic DNA samples were collected from volunteers and external collaborators [Federal Bureau of Investigation DNA Support Unit (FBI DSU) and Pontificia Universidade Catolica do Rio Grande] under North Carolina State IRB protocols # 9213 and 6569, respectively. All samples admitted to the database were amplified with the PowerSeq™ Auto/Y System Prototype kit (Promega Corp., Madison, WI), currently marketed as PowerSeq™ 46GY. Sequencing libraries were constructed in 96-sample pools using automation and were sequenced with the MiSeq System (Illumina Inc., San Diego, California).

1.2 Data analysis using next-generation sequencing

The methods for analyzing STR data with Cloud computing were developed and described in Bailey *et al.* 2017, which allowed for accurate and efficient STR analysis. Genotypes were dissociated and locus-specific tables were built in a separate indexed database schema (SQL) for the website. Thus, no individual genotypes can be accessed in the POPSeq database.

1.3 Data review and quality control

The most recent guidance on NGS data and STR population databasing were followed (ISFG Guidelines, Parson *et al.* FSI:Gen 2016 and Gusmão *et al.* 2017) to include: 1) indexing of the sequences to the forward strand of the human reference GRCh38 (https://www.ncbi.nlm.nih.gov/grc/human), 2) reporting homozygotes as two distinct alleles, 3) maintaining backwards compatibility to fragment size-based STR databases, and 4) the removal of profiles from duplicate and related individuals. Additionally, based on the current guidance, the data admitted to the database also underwent rigorous quality control (QC) review, following at least one of three QC levels.

- *Level One*: Single sequencing analysis conducted using the PowerSeq™ Auto/Y System Prototype kit. Profiles were independently reviewed and interpreted by two different trained analysts, using the previously described set of criteria.
- *Level Two:* Size concordance. Final sequence-based profiles generated with the PowerSeq™ Auto/Y System Prototype kit were compared to CE-generated genotypes and admitted only if concordant.
- *Level Three:* Size and sequence concordance. Final sequence-based profiles generated with the PowerSeq™ Auto/Y System Prototype kit were compared to CE-generated genotypes and confirmed for sequence using a second NGS STR assay (ForenSeq™ DNA Signature Prep kit; Illumina Inc).

Additional details on the quality control may be found on the Documentation page of the website.

1.4 Website design

The environment was hosted by Amazon Web Services (AWS) using EC2 computing resources, Amazon RDS for SQL-like databasing, S3 buckets for data storage and retrieval, Amazon Lambda for serverless functions. All of the resources were contained in one Amazon Virtual Private Cloud (VPC) to include a private subnet for group security policies, data encryption, and secured internet data in-transit protocols and a public subnet for the web hosting. The development team maintained a secured "sandbox" environment to develop and test tools. Once verified, the production level website (https://popseq.cvm.ncsu.edu) was updated from the sandbox code. The NC State domain was used for Internet Protocol (IP) and DNS purposes, but the whole environment was hosted in AWS and did not physically reside on NC State servers. Users of the website could browse the home page summary data and the documentation page, but required a user name and password for login. The population data was explored through the website using the HTML application layer provided by R Shiny (https://shiny.rstudio.com). Data could be visualized with tables and graphs, as well as downloaded in a .csv format to include sequence, size, and population frequency.

Changes in approach/scope

The following details changes to the original proposed approach and the reasons for the changes. All changes were documented in the semi-annual reports to NIJ.

1. Kickoff meeting discussion and advisory input determined that the sample analysis plan should increase the number of samples tested in single, with fewer samples to be analyzed in duplicate. To achieve this the FBI and NCSU planned to split the workload of analyzing DNA samples from the FBI DNA Support Unit (DSU).
2. Delays arose in receiving DNA samples from FBI DSU (450 samples received on May 30, 2017). So, the following corrective actions were taken to test human DNA samples: Pontifical Catholic University of Rio Grande do Sul (Brazil) provided 150 samples for testing, and we submitted and received approval for an IRB protocol at NCSU to collect and test these DNA samples. A no-cost extension was also requested and approved to accommodate the delay, which extended the program end from Dec 31, 2017, to Dec 31, 2018.
3. Dr. Faith resigned from his position at NC State University and accepted a role as Research Leader in Genomics at Battelle Memorial Institute, effective January 2, 2018. To allow for successful completion of this project, Dr. Faith drafted a plan of work to include his continued role as PI/PM while under subcontract to NC State, which included a co-PI and current IT staffing at NC State. The plan was provided to NIJ for review and approved. All approvals and contracts were made to proceed with execution of the grant with Dr. Faith as a co-PI subcontractor at Battelle.

4. Plan for proficiency testing and collaborating laboratory tasks were not fully executed due to lack of available collaborating laboratories. Only one laboratory, The University of North Texas (PI: Bruce Budowle), expressed interest to share/submit NGS data. However, despite multiple attempts to coordinate collaboration, the University of North Texas Laboratory was not been able/willing to commit. There remain too few research laboratories generating NGS data for STRs that could contribute to this project, and no forensics laboratories have gone online with NGS to serve as data providers.

Summary of findings

The PowerSeq™ workflow could be automated for up to 96 samples, producing robust and highly accurate STR sequence data. Bioinformatic tools could be deployed in a Cloud environment and scalable to on-demand resourcing in a secured environment. The fundamental approach is high performance computing that identifies the unique flanks of each STR and reports the STR sequence and fragment-based size for backwards compatibility (Bailey *et al.* 2017).

The NGS approach was successfully used to characterize family trios from Brazil for population level analysis and paternity typing (Silva *et al.* 2017, 2018). Further, the full profiles generated in this study that were concordant with standard capillary electrophoresis analysis (Thermo Globalfiler) have been submitted to STRidER (https://strider.online/) for external quality control (Silva *et al.* 2019, in prep).

For the complete population study, autosomal STR loci were analyzed a for ~900 samples and frequencies were calculated based on sequence composition. Population genetics studies were performed and Hardy–Weinberg equilibrium, polymorphic information content (PIC), and observed and expected heterozygosity were assessed. Overall, sequence-based allelic variants (*i.e.*, isoalleles) were observed in 20 out of 22 different STR loci commonly used in forensic DNA genotyping, with the highest number of isoalleles observed in locus D12S391. The highest increase in allelic diversity and in PIC was observed in D3S1338 and D8S1179. Some unique alleles were observed in specific populations, such as a vWA allele in Brazilians, and PentaD deletions in the flanks (2.2 and 3.2) in African Americans. Complete population details can be reviewed in (Silva *et al.* 2019, in prep).

The website went live on 18 August 2017 with version 1.0 and was updated 18 December 2017 to v1.1. Over 50 people registered for use of the website. The database summary is as follows:
PopSeq Human STR Sequence database version 1.1
Release date: 18 December 2017
Total number of samples: 878
Total number of STR alleles: 49,764
Total number of unique allele sequences: 1028

Figure 1 shows the home page for the website and the self-service analytics for customized searching of the database by selectable metadata filters/tags.

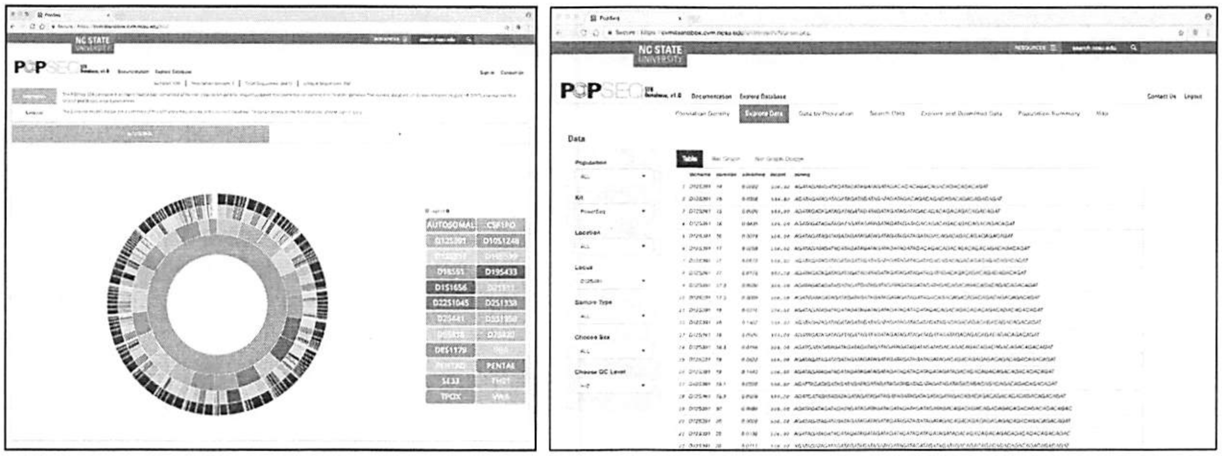

Figure 1 – POPSeq website, Homepage (left panel), self-service analytics (right panel).

Future of the database

While the population data will be published in a forensics journal and deposited with the online STR database, STRidER (https://strider.online/), the POPSeq database cannot continue without funding. A continuation grant was proposed to the FY18 Research and Development in Forensic Science for Criminal Justice Purposes, but was not selected. Thus, the yearly operating costs (~$6,000) for POPSeq are not presently available. Consequently, the website was discontinued on December 31, 2018 at the end of this grant. The development team has imaged the website and database using Amazon AMI and Glacier, as well as Github, to preserve the website and its architecture for later use. Battelle Memorial Institute will preserve the AMIs and Github. The team will continue to search for funding for the website, and in the event of such resources will use the archived material to quickly re-establish the website.

Impacts

The major accomplishments of this research program will help the criminal justice community move forward with adoption of NGS technology and include: 1) the development of an automated laboratory workflow for NGS STR analysis, 2) the development of a Cloud-based bioinformatics tool, 3) the characterization of ~900 human samples from five major population groups for sequence-based STRs, and 4) the development of an online, opensource website for exploring STR data. The version 1.1 database is the first known STR sequence frequency database, as related to NGS analysis of forensic markers.

The gap in an STR sequence diversity database has been recognized by numerous forensic science bodies, such as the FBI Scientific Working Group on DNA Analysis and Methods (SWGDAM) and the International Society for Forensic Genetics (ISFG). Our research and the website product (POPSeq) filled a critical gap for laboratories advancing to NGS based assays. The work reported here will complement and enhance with contemporaneous and future NGS studies required for full adoption of NGS technology in crime laboratories. Further, two members of this study were task members for the SWGDAM NGS working group and sought to draft guidelines for NGS validation and interpretation that will likely be approved and adopted in 2019. The research performed on this study helped to inform the SWGDAM participants on the most pertinent matter for validation and interpretations for STR analysis, as well as mitochondrial DNA analysis. Lastly, knowledge learned on this project also assisted two of the study members who were on a joint task force with the Ohio Bureau of Criminal Investigation to validate STR sequencing with NGS for missing persons cases. Their knowledge was applied to create the study validation plan and write the report to the FBI's NDIS board for approval of the NGS kit.

Products

The work products from this grant program are detailed below, to include Peer-reviewed and published manuscripts in forensic journals, presentations and conference proceedings, inventions, websites, and novel academic instructional courses.

Publications

- Silva D., Scheible M.K., Bailey S.F, Williams C.L., Schutter J., Skomrock N., Minard-Smith A., Allwood J.S., Barker-Scoggins N., Eichman C.F, Meiklejohn K.A., and Faith S.A. Sequence-based autosomal STR characterization in four US populations using PowerSeq™ Auto/Y System. FSI:Genetics.
    - Prepared as of 12/31/2018, but awaiting submission to journal until confirmation of data from STRidER (data submitted to STRidER 01/04/2019).
- Silva D., Sawitzki F., Scheible M., Bailey S., Alho C., and Faith S.A. Paternity testing using Massively Parallel Sequencing and the PowerSeq™ AUTO/Y system for short tandem repeat sequencing. Electrophoresis (2018). 39(21):2669-2673. https://doi.org/10.1002/elps.201800072
- Silva D., Sawitzki F., Scheible M., Bailey S., Alho C., Faith S.A. Genetic analysis of Southern Brazil subjects using the PowerSeq™ AUTO/Y system for short tandem repeat sequencing. FSI:Genetics (2018). 33:129-135. https://doi.org/10.1016/j.fsigen.2017.12.008
- Bailey S.F., Scheible M.K., Williams C., Silva D., and Faith S.A., Secure and Robust Cloud Computing for High-throughput Forensic Microsatellite Sequence Analysis and Databasing. FSI:Genetics (2017). 31: 40-47. https://doi.org/10.1016/j.fsigen.2017.08.008
- Faith, S.A. and Scheible, M. Analyzing Data from Next Generation Sequencers Using the PowerSeq® Auto/Mito/Y System. [Internet] 2016. [cited: 2016, Apr 1]. Available from: http://www.promega.com/resources/profiles-in-dna/2016/analyzing-data-from-next-generation-sequencers-using-the-powerseq-automitoy-system/

Conference posters

- Williams C., Faith S.A. Evaluation of portable nanopore STR sequencing coupled to a customized Cloud-enabled data analysis platform. 28th International Symposium on Human Identification. Seattle, WA (2017).
- Silva D., Sawitzki F., Scheible M., Bailey S., Alho C., Faith S.A. Genetic analysis of Southern Brazil subjects using the PowerSeq™ AUTO/Y system for short tandem repeat sequencing. 28th International Symposium on Human Identification. Seattle, WA (2017).
- Hoggan M., Scheible M.K., and Faith S.A. An assessment of next-generation sequencing on forensic touch DNA samples. National Conference on Undergraduate Research (NCUR), Nashville, TN (2017).
- Scheible, M.K., Bailey S, Silva D., Hoggan M., and Faith S.A. Developmental Validation of A Method for Quantitative High-Throughput Forensic Microsatellite (STR) Sequencing. 26th International Symposium on Human Identification. Minneapolis, MN. (2016).

Presentations

- PopSeq: The Human STR Sequence Diversity Database. NextGen Dx Summit. Washington DC (2017).
- The "Next-generation" of forensic DNA: Education, Casework, and Databasing. Federal Bureau of Investigation. Quantico, VA (2017).
- Educating the "next-generation" of forensic DNA scientists. Virginia DFS' Regional Symposium on DNA Forensics. Norfolk, VA (2017).
- PopSeq: The Human STR Sequence Diversity Database. SWGDAM Round Table (2017).
- Next-generation sequencing overview. Centre for Forensic Sciences (CFS), Toronto, CA (2016).

Websites

- POPSeq, Human STR Sequence Diversity Database https://popseq.cvm.ncsu.edu
- STRGazer – an R based tool for viewing STR profiles https://github.com/sethadam30/STRGazer
- Eppendorf automation scripts for PowerSeq™ https://github.com/sethadam30/Automation_scripts_Eppendorf.git

Inventions, Patent applications, and/or licenses

Invention Report 17192, NC State University, *Secure and Robust Cloud Computing for High-throughput Forensic DNA Sequence Analysis and Databasing.* A complete front-to-end Cloud system was developed to upload, process, and interpret raw Next-generation sequencing (NGS) data using a web browser.

Other products

Dr. Faith and Melissa Scheible designed a new course at NC State called BIT 495/595 Next-gen DNA forensics in Fall 2016. The laboratory workflow and the *Altius* tool developed on this project were used as part of the curriculum. The class was offered for the second time in the Fall of 2017. Fifteen students of undergraduate and graduate level in the Biotechnology program successfully passed this class with high marks.

Trainees, Participants & Other Collaborating Organizations

Training and Development

The following five trainees participated in this research program as part of their undergraduate, graduate, or post-doctoral education and training at NC State University. NIJ grant funding was directly responsible for these training opportunities.

1. Post-doctoral fellow, Deborah Silva PhD – NGS and bioinformatics. Presently a post-doctoral fellow at Battelle Memorial Institute, an NPO (Columbus, OH).
2. Undergraduate student, Marina Hoggan – Sample processing and stutter calculations. Presently a Lab Technician Sorenson Forensics, (Provo, UT).
3. Undergraduate student, Sarah Ermatinger – Population genetics. Presently a PhD candidate at Duke University (Durham, NC).
4. Graduate researcher, Varun Ramachandra Sekar – STR databasing. Presently a programmer at Amazon (Seattle, WA).
5. Graduate researcher, Hannah Seddon - bioinformatics and data analysis. Presently a Biologist at the FBI-TEDAC (Huntsville, AL).

Full roster of individuals participating on this research grant

Staff	Role	Contribution	Funding Support	Location	Collaborated with foreign country	Country(ies) of foreign collaborator
Seth A. Faith	Co-PI	Project Lead, bioinformatics development	This Grant	Columbus, OH	Yes	Brazil, Philippines
Kelly Meiklejohn	Co-PI NC State Lead	Oversight of NC State activities	This Grant	Raleigh, NC	No	
Melissa Scheible	Lab Lead	Optimized lab workflow, provided input to bioinformatic model, tested bioinformatics	This Grant	Raleigh, NC	No	
Sarah Bailey	Bioinform-atics lead	Developed bioinformatics, developed Web tools and database schema	This Grant	Raleigh, NC	No	
Christopher Eichmann	Database/ Web Lead	Constructed AWS environment, organized web team	This Grant	Raleigh, NC	No	
Nikki Scroggins Barker	Web Developer	Contributed to database and web design requirements	This Grant	Raleigh, NC	No	

Marina Hoggan	Undergrad assistant	Quality control of database and bioinformatic tools	This Grant	Raleigh, NC	No	
Hannah Seddon	Graduate Research Assistant	QA/QC review of data	This Grant	Raleigh, NC	No	
Deborah Silva	Post-doc	Laboratory sample processing, data review, and analysis	This Grant	Raleigh, NC	Yes	Brazil
Bill Derosiers	Database Administrator	Database requirements and execution	This Grant	Raleigh, NC	No	
David Luong	Systems Engineer	Cloud design and execution	This Grant	Raleigh, NC	No	
Fred Wright	Statistics Lead	Statistical Analysis	This Grant	Raleigh, NC	No	
Sarah Ermatinger	Undergraduate Researcher	Method Development	This Grant	Raleigh, NC	No	
Jared Schuetter	Statistician	Data Analysis	This Grant	Columbus, OH	No	
Jeremy Sampson	Developer	Develop new Cloud tools	This Grant	Columbus, OH	No	

Other Collaborating Organizations

Organization Name	Organization Type	Location	Contribution	Foreign or Domestic	Point(s) of Contact
FBI DNA Support Unit	Federal Government	Quantico, VA	In-kind support (DNA samples), Collaborative research, technical idea exchange	Domestic	Rebecca Just PhD, Jodi Irwin PhD
University of Central Florida	Academic	Orlando, FL	Personnel exchanges/subject matter expertise	Domestic	Jack Ballantyne PhD
North Carolina State Crime Laboratory	State Government	Raleigh, NC	Personnel exchanges/subject matter expertise	Domestic	David Freeman PhD
Amazon Web Services	Private Industry	Seattle, WA	Personnel exchanges/subject matter expertise	Domestic	Peyton Biggs, Benjamin Snively

Pontificia Universidade Catolica do Rio Grande	Academic	Porto Alegre, Brazil	In-kind support (DNA samples)	Foreign	Clarice S. Alho PhD

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.