Document Title:	Continued Development of FROG-kb: A Forensic Resource/Reference on Genetics Knowledge Base
Author(s):	Kenneth K. Kidd
Document Number:	251821
Date Received:	July 2018
Award Number:	2014-DN-BX-K030

Final Summary Report

National Institute of Justice grant NIJ 2014-DN-BX-K030

Covering December 1, 2014 to December 31, 2016

Project Title:“Continued development of FROG-kb: a forensic resource/reference on genetics knowledge base”

Kenneth K. Kidd (PI) Professor Emeritus of Genetics

Email: Kenneth.Kidd@yale.edu

Telephone: 203-785-2654

Department of Genetics

Yale University School of Medicine

Submitted by K.K. Kidd

This project has been jointly funded by NIJ (grant 2014-DN-BX-K030) and NSF (grant BCS-1444279; "Ongoing Development of a Human Population Genetics Resource with Forensic Application"). No attempt is made to distinguish parts of the project by funding source; this is one project and the report covers the entire project.

Major goals and specific aims

The overall goal of the project has been to continue expanding the contents of our large existing database, ALlele FREquency Database (ALFRED, https://alfred.med.yale.edu/alfred/), to make necessary modifications to it, and to develop the FROG-kb database interface (Forensic Resource/Reference of Genetics knowledge base, http://frog.med.yale.edu/FrogKB/) to provide new data and functionality designed specifically as a prototype for forensic teaching and research, to facilitate forensic practice, and as a forensic reference tool. The specific goals were (1) **To enhance the contents of the underlying database** by including population frequency data on SNP sets associated with Individual Identification SNP (IISNP), Ancestry Inference SNP (AISNP), and Phenotype Informative SNP (PISNP) panels as well as SNPs in general typed on multiple populations; **(2) To add new functions and enhance the user interfaces** by improving computations and implementing new functions associated with specific SNP sets and ultimately provide the users with a more useful, user-friendly, and intuitive interface to FROG-kb; **(3) To develop FROG-kb as a knowledge base** by developing functions in FROG-kb to facilitate input of ideas and suggestions from users from different forensic backgrounds and also by adding didactic explanatory materials to the available functions, thus enhancing the teaching/educational aspects of the knowledge base components.

Specific Objectives

The specific objectives of the project were to identify and then make publically available high quality allele frequency data relevant to studies of human genetic variation with all allele frequency tables linked to the source (publication), to the molecular definition (usually dbSNP), and the anthropologic description of the population (various web sites or summary descriptions). The main objective for ALFRED and FROG-kb was to expand the frequency data available for SNPs studied in many populations. For the SNPs included in the forensic panels that are available in the ALFRED SNPset page and in FROG-kb we emphasized adding additional population samples even if the entire SNP set is not available for a population. New data comes from the published literature, the Kidd laboratory, and from collaborating researchers. Another goal was to increase the frequency data available on more reference populations from around the world for a specific forensic SNP panel so that more meaningful interpretations of calculations would be available in FROG-kb. An additional objective was to create a more user-friendly interface and to publicize FROG-kb through the collaboration with the Forensic Technology Center of Excellence (FTCoE) section of Research Triangle Institute International (RTI).

Accomplishments

A key ongoing component of this project involves searching the literature for relevant material, curating the information for entry into ALFRED, and then implementing the relevant data into FROG-kb. The contents of ALFRED increased from about 37,064,000 frequency tables when the funding began in December 2014 to about 40,119,488 as of Oct. 2016. Through our curatorial efforts various new SNPsets and additional reference population samples for the existing SNPsets have been added. In the current FROG-kb interface we have3 SNPsets under the IISNP component, 10 panels under the AISNP component and one SNP panel under the PI SNP component. Each of these panels has supporting population data. URL links exist to pages in ALFRED for more details and allele frequency data tables for specific populations. Data exist and are accessible through ALFRED not only for the populations used for calculation but also for each SNP-population combination for which data are available. Additional SNPsets not included yet in FROG-kb database are available through the ALFRED SNPset page. We have been working with the view that the discriminatory power for individual identification will be population specific, and ancestry inference will only be as good as the set of reference populations. The recent data additions to the database have worked toward overcoming this limitation by systematically adding additional population data for a panel as data become available. For example, the ‘KiddLab set of 55 AISNPs’ now has 139 reference populations (Pakstis et al., 2017) and the recently included ‘Combined panel of Kiddlab-55, Seldin's-128, and SNPforID34-plex AISNPs’ with data on 75 population samples.

Major areas of accomplishment related to FROG-kb

(1) With the collaboration of the Forensic Technology Center of Excellence a new user-friendly interface with greater flexibility was developed and put online in January of 2015 aimed at increasing awareness of and traffic to FROG-kb. An elaborate 20 page “user manual” with detailed textual, graphical elements and flow charts designed to make FROG-kb navigation easier for the user was created and added to the interface. The ‘pipeline’ link was implemented giving a diagrammatic representation of the data entry and computation details for the panels. Other links added to the interface include an ‘announcements’ link to keep the user up to date on the recent additions and changes to the database and a ‘contact us’ link.

(2) Special efforts were made to fill in the otherwise largely empty matrix of populations with polymorphism frequency data. This involved routinely scanning the literature, identifying and curating the information for quality and adding population-specific allele frequency data from the anthropology, human genetics, and forensic literature. We also made direct efforts to encourage other researchers to study new population samples on the existing panels of forensic SNPs. An emphasis was madeto add datasets relevant to forensic applications involving individual identification (II SNP) and Ancestry Information (AI SNP) panels.

(3) Various new functions and enhancements to already existing functionalities were made to make the database more user-friendly and easier to navigate. A very important improvement consisted of systematically changing the nucleotide codings of all SNPs in forensic SNP sets so that they represent the forward strand. In the earlier version of FROG-kb the allele designations for the genotypes listed on the data entry forms were not consistent with any single standard orientation because of the heterogeneity of their representations in the literature and public databases like dbSNP. As a consequence, the results of the likelihood computations implemented in FROG-kb were unpredictable because of the variability in genotype coding. The current updated database (ALFRED and FROG-kb) has all the SNPs in a SNP

panel represented on the forward strand orientation. This major modification standardizes data entry. Since this update, all further data entries in FROG-kb are consistent to represent the alleles on the forward (5’->3’) orientation. Some other new functionalities updates include sortable column fields for population and SNPs in a SNPset, two ‘data entry’ options for ‘Genotype entry using selection by radio button’ and ‘File upload’, and the availability of ‘Example files’ for each panel. We also added to the current interface a downloadable file option of the ‘Genotype frequencies’ used for the computations for all the panels and also an option to enter the ***‘DNA profile id’*** providing an easy identification record of the results generated for a particular input genotype profile.

(4) Significant efforts towards publicizing the FROG-kb database were made including various posters and talk presentations and also through the feedback function available in the public interface. This has already been used to request easy access to the exact frequencies used for likelihood calculations; a new tab was added to download those data directly. To develop the Knowledgebase aspect of the database we have added links from relevant forensic websites and also links to the posters and talks presented by Dr. Kenneth Kidd at the various meetings highlighting the features of the FROG-kb database (the list of the presentations and talks is included in the presentations/papers section of the report).

(5) We are publishing in peer reviewed journals. From our study of SNPs from 21 different AI (ancestry informative) SNP panels we published a research paper (Soundararajan et al., 2016) highlighting the “empty matrix” issue which is a serious problem for the forensic field. The SNPset from this overlap study have also been included in FROG-kb as an AI SNP panel. We have a manuscript in preparation highlighting the new FROG-kb interface, the various new SNP panels incorporated, and also the progress in increasing the reference populations for existing panels.

In summary, a completely new FROG-kb interface with detailed user manual has been put online. Data on many population samples data have been added to ALFRED/FROG-kb increasing the reference populations available for various SNP panels. Our overlap study paper involving 21 different AISNP panels and the outcome of the analysis from the study was published (*Soundararajan et al., 2016)* highlighting the ‘empty matrix’ issue which is a serious problem for the forensic field. FROG-kb already accesses a much larger set of accumulated information and cites other additional panels that could be added readily. The FROG-kb and ALFRED interfaces have new functions for the convenience of users. Our ultimate goal was to provide a ‘one-stop shop’ pertaining to the forensic SNPs panels for the forensic and human population genetics communities and the key outcomes contributing to this goal have been implemented in the current FROG-kb interface with supporting and supplemental data in ALFRED.

Adverse Events

In March of 2016 ALFRED and FROG-kb servers were hacked. After a short period of time the database's public interfaces were brought up with no compromise to data. The new public interface has better security to minimize the likelihood of a repeat. However, the curator interface (an interface used by the curator to enter allele frequency data and other details related to the data curation and entry) was completely down and is in the process of being rebuilt from the ground up. This has slowed the addition of new data.

Databases

ALlele Frequency Database, ALFRED (https://alfred.med.yale.edu)

ALFRED is a database that has been designed to make allele frequency data on anthropologically defined human population samples readily available to the scientific community and to link these polymorphism data to the molecular genetics-human genome databases. ALFRED contains data on a large number of populations with clear and extensive anthropological, ethnographic, and linguistic descriptions and definitions of the specific samples linked to citations on which allele frequencies are based. The detailed curation involved in locating, assembling, and recording the data from various sources is the key feature of ALFRED and such assembled relevant data in ALFRED is a necessary precursor to implementation of the relevant data into FROG-kb calculations.

Forensic Resource\Reference on Genetics-knowledgebase, FROG-kb (http://frog.med.yale.edu)

FROG-kb seeks to make allele frequency data for SNPs and other genetic polymorphisms more useful in a forensic setting. The primary objective of FROG-kb is to provide a web interface that, from a forensic perspective, is useful for teaching and research and can serve as a tool facilitating forensic practice. The underlying data are housed in the already extensively used and referenced ALlele FREquency Database (ALFRED). The FROG-kb interface makes the information usable for forensic purposes, including computational tools for comparing user-provided data with underlying allele frequencies in populations.

These tools are organized by the methodology followed and the different published SNP/marker panels available.

Opportunities for training and professional development from this project

Much effort has been put into the training and development of the staff. All individuals involved in the project benefit from 'cross-training' created by the frequent, and necessary, interaction with other staff members. Programmers are constantly interacting with other programmers at the Yale Medical Informatics Team. The curator and student involved in the project interact with the programmer and other members of the project team to understand the informatics and population genomics involved in ALFRED and FROG-kb. Collaboration and interaction with RTI (Research Triangle Institute International) team with expertise in website design had been tremendously beneficial.

Dissemination/publicity of products

The broad objective of the project was to assemble and make available otherwise widely dispersed data in an integrated fashion. The large number of users attests to the successful accomplishment of that objective. Recently in 2016, the average number of users each month was 322 for FROG-kb and 10,379 for ALFRED. About forty percent of the users have internet addresses in the United States and Canada. Elsewhere around the world high ranking user countries include: China, Japan, the United Kingdom, France, Germany, Italy, Brazil, Singapore, Russia, India, Australia, and Spain.

There have been two peer reviewed publications from this project. A short communication (Pakstis et al., 2015) is published in *Forensic Science International: Genetics* reporting the large increase in population frequency data and to publicize the existence of the database/web site. Another publication (Soundararajan et al., 2016) from the overlap study of 21 published AI SNP panels was published emphasizing the 'empty matrix' problem.

Our collaboration with the FTCoE of Research Triangle Institute International (RTI) has increased the visibility of and traffic to FROG-kb. As part of the publicity initiative we added links from relevant forensic websites. FTCoE website is one of them: https://www.forensiccoe.org/Our-Impact/Advancing-Technology/Databases. Northeastern Association of Forensic Scientists (NEAFS) and Southwestern

Association of Forensic Scientists (SWAFS) are other websites RTI is working on getting to link to FROG-kb.

Various aspects of the work supported by this project were also presented at numerous slide talks and poster sessions at international meetings and visits to University laboratories. These meetings of forensic researchers, anthropologists, and population geneticists were held at locations in the United States, Europe, and China. Examples of such meetings include but were not limited to: Bode Technology meetings, the International Symposium on Human Identification (ISHI), the International Society of Forensic Genetics (ISFG), Green Mountain DNA Conferences, Gordon Research Conferences, the American Academy of Forensic Sciences, and the International Conference on Genomics.

Impact of this project on forensics

The objective of the project is facilitation of access to and use of data in studies of human genetic variation broadly with an emphasis on use in a forensic setting. The FROG-kb interface provides a unique forensic resource and the underlying database ALFRED is a unique resource of population genetic data. The forensic web front-end of the database with extensive SNP data will facilitate understanding of and use of SNPs in forensic settings. The educational aspects of FROG-kb should be of great value in helping the forensic technicians understand and use SNPs when eventually match and ancestry calculations are accepted in the courts in the future. If the calculations are accepted in the courts, the results produced by FROG-kb in a printed report (as already exists) could be directly submitted as evidence.

Journal publications specifically related to this project

Soundararajan, U., L. Yun, M. Shi, K.K. Kidd, **2016**. Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. *Forensic Science International: Genetics* 23:25-32. DOI: 10.1016/j.fsigen.2016.01.013.

Pakstis, A.J., E Haigh, L Cherni, A Ben Ammar ElGaaied, A Barton, B Evsanaa, A Togtokh, J Brissenden, J Roscoe, O Bulbul, G Filoglu, C Gurkan, KA.Meiklejohn, JM Robertson, C-X Li, Y-L Wei, H Li, U Soundararajan, H Rajeevan, JR Kidd, KK Kidd, **2015**. 52 additional reference population samples for the 55 AISNP panel. *Forensic Science International: Genetics* 19:269-271.

Andrew J. Pakstis, Longli Kang, Lijun Liu, Zhiying Zhang, Tianbo Jin, Elena L. Grigorenko, Frank R.Wendt, Bruce Budowle, Sibte Hadi, Mariam Salam Al Qahtani, Niels Morling, Helle Smidt Mogensen, Goncalo E. Themudo, Usha Soundararajan, Haseena Rajeevan, Judith R. Kidd, Kenneth K. Kidd, **2017**. Increasing the reference populations for the 55 AISNP panel: the need and benefits. *Int J Legal Med in press*

Kidd, K.K., 2016. Chapter 7: "Thoughts on estimating ancestry" In: A. Amorim and B. Budowle (Eds), Handbook of Forensic Genetics--Biodiversity and heredity in civil and criminal investigation. London: Imperial College Press. (Describes the methods in FROG-kb and their interpretation.)

Other publications, selected conference papers and presentations

Kidd, K.K., H. Rajeevan, K. Moore, U. Soundararajan, R.Satcher, P. Melton, A.J. Pakstis, J. Ropero-Miller. (2016). *"Population genetics and the interpretation of forensic DNA data—FROG-kb" Poster presentation*. Gordon Research Conference--Forensic Analysis of Human DNA. Waterville Valley, New Hampshire, amd at ISHI, Minneapolis, Minnesota.

Kenneth Kidd, Haseena Rajeevan, Katherine N. Moore, Richard Satcher, Patricia A. Foley-Melton, Jeri D. Ropero-Miller (2016). *"Updates to the Forensic Research/Reference on Genetics Knowledge Base (FROG-kb) Database". Poster presentation*. American Academy of Forensic Sciences meeting. Las Vegas, Nevada

K.K. Kidd, W.C. Speed, Sharon Wootton, Robert Lagace, Reina Langit, E. Haigh, Joseph Chang, A.J. Pakstis (2015). *Advances in identifying and characterizing microhaplotypes for forensics". Meeting dates: June 22-26 for the Conference on Forensics and Anthropologic Genetics*. 9th Annual meeting of the International Society of Applied Biological Sciences, Bol, Croatia.

K.K. Kidd (2015). *Now Generation Sequencing (NGS) for Forensics*. Bode 11th Annual DNA Technical Workshop. San Diego, CA.

K.K. Kidd (2015). *Future uses of DNA in forensics*. Connecticut State Forensics Laboratory. Meriden, CT

K.K. Kidd (2014). *Presentation on Next generation sequencing role for forensic applications.* Annual DNA & Investigators Workshop. Bode Mid-Atlantic, Arlington VA.

Manuscripts in preparation

U. Soundararajan, H. Rajeevan, K.K. Kidd, 2016. New enhancements for FROG-kb.

K.K. Kidd, W.C. Speed, A.J. Pakstis, other co-authors including those from collaboration with ThermoFisher Scientific, order to be determined depending on participation in analyses and writing; 2016. Working title: “Evaluating 132 Microhaplotypes across a Global Set of 83 Populations”.

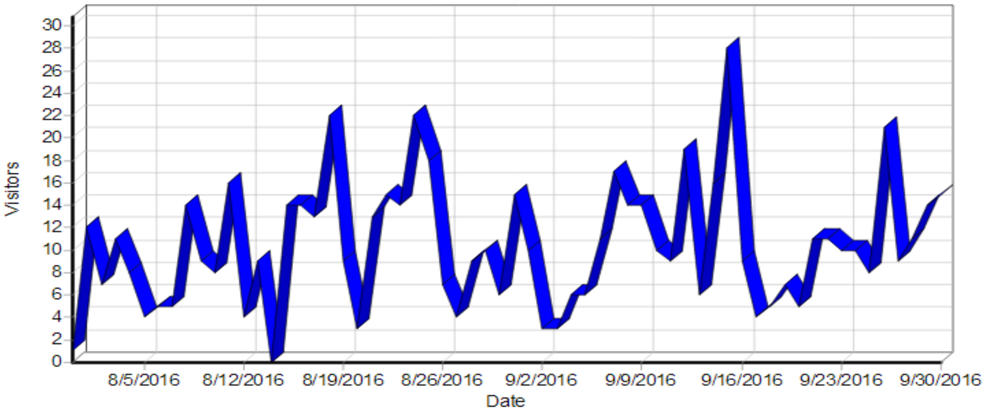
FROGkb usage summary--Aug 2016 and Sep 2016

Hit - A request for any file (page, image, etc).

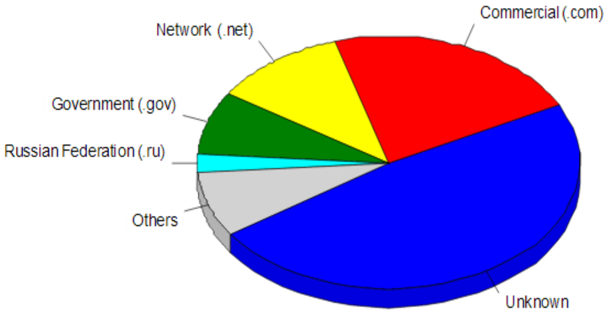
We have tried our best to exclude hits from web crawlers/spiders.

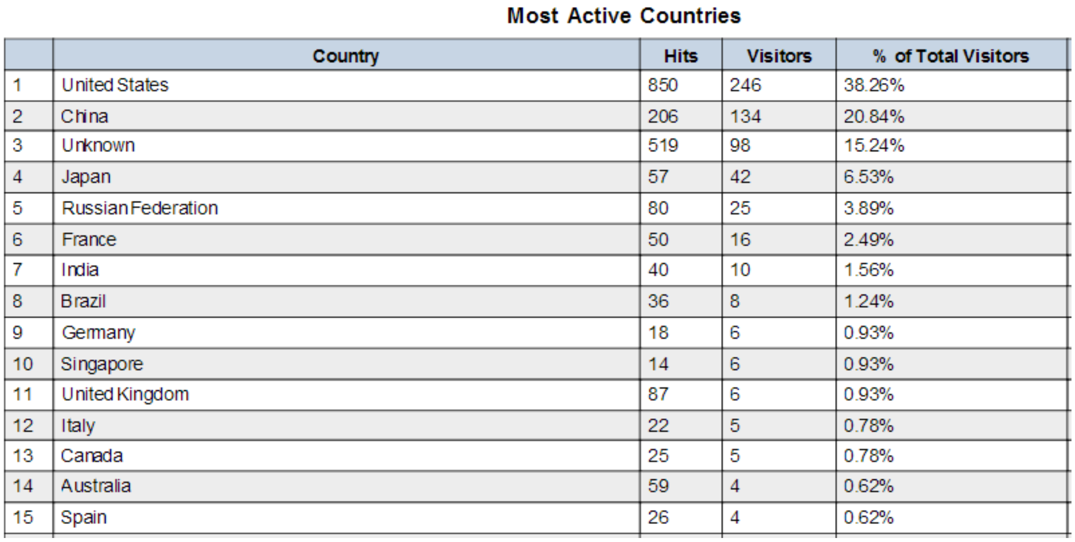
General Statistics

Total Hits	2,280
Visitor Hits	2,280
Average Hits per visitor	3.5
Total Visitors	643
Average Visitors per Day	10
Total Unique IPs	328



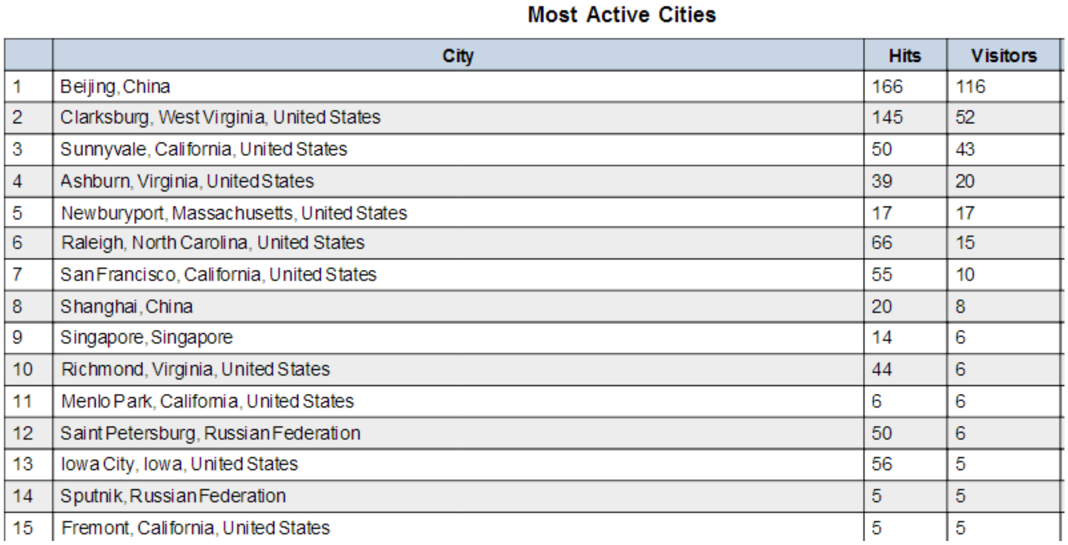
Top-Level Domains





Most Active Countries

	Country	Hits	Visitors	% of Total Visitors
1	United States	850	246	38.26%
2	China	206	134	20.84%
3	Unknown	519	98	15.24%
4	Japan	57	42	6.53%
5	Russian Federation	80	25	3.89%
6	France	50	16	2.49%
7	India	40	10	1.56%
8	Brazil	36	8	1.24%
9	Germany	18	6	0.93%
10	Singapore	14	6	0.93%
11	United Kingdom	87	6	0.93%
12	Italy	22	5	0.78%
13	Canada	25	5	0.78%
14	Australia	59	4	0.62%
15	Spain	26	4	0.62%



Most Active Cities

	City	Hits	Visitors
1	Beijing, China	166	116
2	Clarksburg, West Virginia, United States	145	52
3	Sunnyvale, California, United States	50	43
4	Ashburn, Virginia, United States	39	20
5	Newburyport, Massachusetts, United States	17	17
6	Raleigh, North Carolina, United States	66	15
7	San Francisco, California, United States	55	10
8	Shanghai, China	20	8
9	Singapore, Singapore	14	6
10	Richmond, Virginia, United States	44	6
11	Menlo Park, California, United States	6	6
12	Saint Petersburg, Russian Federation	50	6
13	Iowa City, Iowa, United States	56	5
14	Sputnik, Russian Federation	5	5
15	Fremont, California, United States	5	5