December 2015 Update from the author:

This Report was written in mid-2013 covering the development of a pilot version of FROG-kb and describing the version that existed in June, 2013, when funding ended. After a period of stasis, funding for FROG-kb resumed in late 2014 (co-funded by NIJ grant 2014-DN-BX-K030 and NSF grant BCS-1444279) and extensive changes to the web site <frog.med.yale.edui>, functionality, and the data resources have been undertaken. By early 2016 the web interface will be considerably different from that described in this 2013 Report.

National Institute of Justice
Final Technical Report
Grant # 2010-DN-BX-K226
Project Title:
Forensic Resource/Reference on Genetics Knowledge Base (FROG-kb)

Kenneth K. Kidd (PI)
Professor of Genetics
Email: kenneth.kidd@yale.edu
Telephone: 203-785-2654
Department of Genetics
Yale University School of Medicine

Abstract

A significant amount of research is being focused on highly informative single nucleotide polymorphisms (SNPs) and on development of panels of these SNPs that can potentially be considered for human identification and description in forensic, biomedical, association, as well as epidemiological studies. Our specific goals included: (1) Design, develop and publish a web application that would be useful from a forensic standpoint, (2) Modify the structure of the existing database ALFRED (ALlele FREquency Database, http://alfred.med.yale.edu), the data source for the application to store forensically relevant information, and (3) Enter various forensic SNP panels that are developed in our lab, obtained from collaborators, and published in the literature. During the grant period we developed an open access web application FROG-kb (Forensic Research/Reference on Genetics-knowledge base, http://frog.med.yale.edu) that can be used as a tool for forensic practices.

Using our previous knowledge in database design the schema of ALFRED was modified to accommodate the tables and columns that were essential for FROG-kb. The logic supporting the design of the additional tables and relationships is as follows. There can be one or many sites associated with a defined forensic panel. Every panel is linked to at least one publication. The marker phenotype frequency for each 'site - population sample' combination is pre-calculated and saved. None of these modifications called for changes to the ALFRED application. The database underlying FROG-kb is implemented using Oracle version 10 on one of Yale's institutional database servers where it is maintained.

The user interface layout of FROG-kb is designed to reflect the organization of the contents and functionality, as well as ease of use of the web interface. Every set of pages relative to a function on FROG-kb originates from a tab on the 'Main Menu'. The SNP panels are organized by the type of panel: 'IISNP' for Individual Identification SNPs, 'AISNP' for Ancestry Inference SNPs and 'PISNP' for Phenotype Informative SNPs. Functions and data related to each panel can be navigated to by first selecting the type of panel and then a specific panel. The most important function 'calculation of maximum likelihood' is augmented with example images of input and output screen, preselected data-entry pages for two individuals from different populations, and two options of data-entry facilities. The two data-entry options provide a flexible user-interface for data input while the results pages for both are uniform. Users have the option of printing the results page, and the data-input page. The lists of populations used in the maximum likelihood calculation are given as well. The web front end is built using web developing technologies such as Java, Java Servlet, JSP, JQuery,

and GoogleCharts. Almost all of the client-code utilizes JQuery, and the server implementation is in Java.

The different functions on FROG-kb are implemented on three different IISNPs and seven different AISNPs panels. Each panel provides examples and the ability to calculate match probabilities for user-specified genotypes in each of many populations that have allele frequencies available for all SNPs in the panel. For PISNPs panel we provide a panel of six SNPs for eye color prediction (IrisPlex) along with ability to specify an individual's genotype and predict eye color from that. A paper on this knowledge base and web site has been published (Rajeevan et al., 2012). Usage has been reasonably steady since early 2013 when the basic production version went public. During that period nearly two dozen individuals per day have used the database.

Table of Contents

Executive Summary

Original objectives

Single nucleotide polymorphisms (SNPs) have great potential for forensics. However, two obstacles to their implementation have existed: lack of commercial kits for a well documented set of SNPs and lack of databases and functionality to document and interpret the SNPs used. Both obstacles need to be addressed simultaneously. This project was proposed to undertake in a forward-looking and proactive way the database and functionality issue. Therefore, the primary objective of the project was to provide a web interface 'Forensic Research/Reference On Genetics knowledge base' (FROG-kb) on DNA polymorphisms conducive for teaching and research and as a referential tool from a forensic standpoint. The underlying data would be provided by the already extensively used and referenced ALlele FREquency Database, ALFRED (http://alfred.med.yale.edu). Many of the markers already in ALFRED are those used in forensics or published in the forensic literature for possible use. The project would allow even more specifically forensic SNP data and information to be curated and entered. The new web based interface for FROG-kb would be designed to make these data available to forensic students, researchers, and practitioners in a relevant user-friendly manner. Modified versions of search tools already available from ALFRED would be implemented on FROG-kb to make it more suitable for forensic purposes. In addition to displaying data in an organized manner, multiple computational tools that operate utilizing the underlying allele frequency and user provided data would be available from FROG-kb. These tools would be organized by the calculation methodology to be used and the different published SNP/marker panels. Computations would involve both determining the match probability of an input multi-SNP profile using the population data already stored as well as calculations estimating the relative likelihoods of the multi-SNP profile having ancestry from any population in the database. Interfaces for these calculations would be specifically designed for defined panels of SNPs. Functions that can facilitate user defined SNP panels would be available for research and teaching purposes. A FROG-kb wiki application would be established to encourage forensic expert involvement in the building of the interface and functionality. As described in this final technical report, we have exceeded our expectations in all except the final of those proposed developments: a wiki application. Advice from the colleagues we consulted was that other aspects of the development were more important.

Though the grant was technically awarded in 2010, funding did not start until early 2011. Thus, this final report covers the 30 months from January 2011

through June 2013 when funding in the extension period for personnel was exhausted.

Progress during the grant period

At the time of the grant application for this project an illustrative prototype of FROG-kb with very limited functionality was available on the web. During the first 6 months of the project (1/1/11 to 6/30/11), development of FROG-kb progressed significantly in three areas: (1) Amendments were made to the backend database (ALFRED) by creating new tables and new fields were added to already existing tables to accommodate information required for FROG-kb, (2) Various forensically relevant datasets were uploaded into the database, and (3) A functional pilot implementation of a web interface (http://frog.med.yale.edu) was designed, implemented , and deployed.

By the end of the first year of funding on the project (1/1/11 to 12/31/11) we had accomplished many of the overall objectives. (1) A solid database structure was in place. (2) A web interface was online and available for comment (the web interface had already implemented some of the proposed statistical functions.) (3) We had begun to work on a manuscript to introduce the web site to the forensic community. (4) We had also assembled the data for several more forensic panels and had already put five into FROG-kb accessible on the web. (For all the data sets we first needed to curate and upload the data into ALFRED and then make the frequencies available for the calculations in FROG.) At the end of this period, the FROG-kb version accessible on the web had data on 5 different SNP panels.

By half way through the second year 1) A manuscript to introduce the web site to the forensic community had been submitted and accepted by 'Investigative Genetics'; 2) Multiple additions and enhancements had been made to the web interface; updates had been made to the database structure to support the amended functions; 3) New forensically relevant datasets were uploaded into the database. 4) Research was undertaken toward development of better statistics for evaluating the significance of ancestry inference.

By the end of the second year development of FROG-kb had advanced in the following areas. (1) Several functional amendments were made on the web interface. (2) New forensically relevant datasets were uploaded into the database. (3) Programming bugs were identified and fixed during this period. The work during this period resulted in the deliverable of an improved version of FROG-kb put online in January, 2013. An email announcement of the improved

version was sent to nearly 100 forensic scientists for whom we had email addresses.

During the final six months of the project the focus was on contents and ease of use. Five new forensic panels were added. Several seemingly small changes were added to make the interface more user-friendly. We developed a likelihood approach for eye color that gives the same results as Snipper (http://mathgene.usc.es/snipper/) but did not have time to implement it. At the end of the project (June 30, 2013, when funds for personnel ended) we had a fully functional web application for eleven pre-defined forensic panels of SNPs.

Current status of FROG-kb

The current version of FROG-kb will remain online for at least a year (courtesy of a colleague who will pay the costs of maintaining the database on Yale's Oracle server) or until new funding is available. During that time it will be essentially static as will ALFRED since its funding also ended at the end of June, 2013. The following paragraphs summarize the current functionalities available in the current version of FROG-kb and the current contents.

Structure of the database

The underlying database for both ALFRED and FROG-kb functionalities is implemented using a traditional relational structure. The database is implemented using Oracle version 10 on one of Yale's institutional database servers where it is maintained. FROG-kb is built as a separate web based interface entirely different from the ALFRED interface. The intent and design issues supporting the initial database structure and subsequent modifications of ALFRED can be found in Cheung et al. (2000), Osier et al. (2002), and Rajeevan et al. (2005, 2011). The user interface layout and functionalities of FROG-kb are elaborated in the Investigative Genetics paper (Rajeevan et al. 2012). The web front end of FROG-kb is built using web developing technologies such as Java, Java Servlet, JSP, JQuery, and GoogleCharts. Almost all of the client-code utilizes JQuery, and the server implementation is in Java.

Nature of the interface

The user interface layout of FROG-kb is designed to reflect the organization of the contents and functionality, as well as ease of use. Every set of pages relative to a function on FROG-kb originates from a tab on the 'Main Menu' that appears on the left-hand side of every page (appendix Figure 2). The 'Home Page' gives a brief summary of the functions available in FROG-kb. Explanatory information about FROG-kb can be found under 'About'. The core functions on FROG-kb can

be reached by selecting the tabs ‘File Upload’, ‘IISNP’, ‘AISNP’, or ‘PISNP’. Under each of the SNP types is the list of panels available and the primary reference for each. Under each specific panel are several functions. ‘SNP Set’ gives the list of SNPs in the panel by rs number and an active link to dbSNP for molecular specifics. The ‘Populations’ button gives the list of populations for which there are data on all SNPs and for which likelihoods can be calculated. The ‘Functionalities’ button under each type of panel gives a brief description of the functions available. The ‘Examples’ button leads to two options, (1) a static picture of input and output and (2) input datasets the user can run and modify. The ‘Data Entry’ button provides two different options to enter an individual’s genotype data for a particular SNP panel-- 'Selection by Radio Button' and 'File Upload'-- both of which are explained. The ‘Formula’ button explains the calculation used for the specific panel.

After the genotype data for an individual are entered, the ‘Calculate’ button needs to be clicked to calculate and display (1) the random match probabilities for the IISNP panel, (2) the likelihoods of originating from diverse populations for the AISNP panels, or (3) the eye color probabilities for the PISNP panel.

Contents and functionality

FROG-kb has three different sets of IISNPs: (1) the Kidd Lab set of 45 unlinked SNPs (Kidd et al., 2012 & Pakstis et al., 2010), (2) the SNPforID set of 52 SNPs (Sanchez et al., 2006), and (3) the Qiagen DIPplex set of insertion/deletion markers (Fondevila et al., 2012). There are seven different panels of AISNPs: (1) the Seldin group set of 128 SNPs (Kosoy et al., 2009 & Kidd et al., 2011), (2) the SNPforID set of 34 SNPs (Phillips et al., 2007), (3) the Kidd Lab set of 55 SNPs (Kidd et al., 2013 submitted), (4) Kayser’s set of 24 SNPs (Lao et al., 2010), (5) Podini’s set of 32 SNPs (Gettings et al., 2014), (6) the SNPforID Eurasiaplex of 23 SNPs (Phillips et al., 2013 & Bulbul et al., 2011), and (7) Nievergelt’s set of 41 SNPs (Nievergelt et al., 2013). There is one panel available under PISNPs: Irisplex for eye color prediction (Kayser et al., 2011). The reference for each panel is in FROG-kb associated with the index of the panels of each type. Each panel has a link to the SNP Set in ALFRED from which all information available on each SNP can be accessed.

Several functions exist as a series of buttons for each panel. The list of the specific SNPs in a panel exists with active url links to dbSNP. Each panel has a link to the ‘SNP Set’ page in ALFRED from which all information available on each SNP can be accessed. Different numbers of populations have data for the different sets and the populations available are listed for each set under the ‘Populations’ button. Some of the SNPs are found in more than one of the

AISNP panels because of several well known SNPs with large allele frequency differences. “Functionality” and “Formula” buttons link to explanatory text. “Data entry” leads to the options for data input on an individual. “Examples” leads to two different views of how data can be entered and results that can occur, depending on the specific panel.

Results for IISNPs and AISNPs are essentially identical. Populations are ranked by the probability of the input data arising from that population. The probability is also given as is the likelihood ratio of the highest probability to each other result. Also displayed is a graph of the entire set of probabilities with mouse-over functionality to give the population and value at any point on the graph. In the case of the single set of six SNPs in the Irisplex phenotype SNPs the results are the relative probabilities of the three possible eye colors: light, intermediate, dark.

Project outcome

During this grant period, in keeping with the original objectives of the project, we have developed a ‘one-stop shop’ database prototype for the forensic use of SNP panels. It was designed to be useful for the forensic community. The current version of the FROG-kb web interface is versatile in its functionalities and comprehensive in the population data available for many of the SNP sets. The interface allows viewing and retrieval of data, as well as calculation of statistics on several forensically relevant SNP sets. Data on seven Ancestry, three Individual, and one Phenotype Informative panels (AISNP, IISNP, and PISNP) with all the functionalities (detailed in the previous section ‘Contents and functionality’) are currently available in FROG-kb. Good explanatory and didactic material to the existing functions in FROG-kb are also included.

However, FROG-kb is a work in progress and the current implementation is not without limitations (discussed in detail in the 4.1 ‘Discussions’ section under ‘Conclusions’ of this report). Our initial efforts in developing this resource for the forensic community has necessarily focused on the database structure and the website interface with various different panels implemented, as was the objective of the project. Funding was not sufficient, nor was it intended to be, for comprehensive inclusion of all forensically relevant SNP panels. Hence, though this is a pilot effort, the current available interface is highly a functional one and serves as a prototype for one approach to a database that can be a reference and resource on genetics for the forensic community. The direction of future development, if undertaken, will be determined in large part by feedback from the forensic community.

Usage of FROG-kb

Since no practicing forensic lab, at least that we know of in the United States, is using any of these panels, we expect current usage to be related to understanding what these panels can do as well as to exploratory or educational usage. Since the new version was put in place in January 2013, usage has been moderate at an average of 21 visitors per day and a total of over a thousand unique Internet Protocol (IP) addresses. The largest % of those visitors have unresolved IP addresses, but for the IP addresses that we can resolve, there have been 300 commercial visitors, over 250 ",net" visitors, 18 educational (.edu) visitors, 16 US government (.gov), and 10 ".org" visitors. Foreign countries among those with multiple visitors were China (65), Brazil (19), Russian Federation (18), Germany (15), Singapore (13), and between 12 and 9 visitors each from Spain, Italy, Australia, and The Netherlands.

Implications of policy and practice

Policy: FROG-kb was designed to allow the adoption of SNPs in various aspects of forensic practice by proactively providing the resources needed to evaluate the data collected on SNPs. Commercial kits are becoming available now for use in forensic practice; they implement many of the panels already available in FROG-kb. It is our expectation that the match and ancestry calculations will eventually be accepted in the courts since the underlying databases are and will be public and sufficiently large to match the existing STRP databases for population frequencies.

Practice: The didactic aspects of FROG-kb should be of great value in helping the forensic technicians understand and use SNPs. When commercial kits allow implementation of SNP typing in forensic labs, the database structure and basic web interface allow other SNP panels to be added should they be different from the currently implemented panels. If the kits and calculations are accepted in the courts, the results produced by FROG-kb in a printed report (as already exists) might be directly submitted as evidence.

I) Introduction

1.1 Statement of the Problem:

Single nucleotide polymorphisms (SNPs) have great potential for forensics. However, two obstacles to their implementation have existed: (1) lack of commercial kits for a well documented set of SNPs and (2) lack of databases and functionality to document and interpret the SNPs used. Both obstacles need to be addressed simultaneously. This project has undertaken, in a forward-looking and proactive way, the database and functionality issue.

1.2 Background and Rationale:

More than a decade ago the forensic community settled on a set of short tandem repeat polymorphisms (STRPs) for human identity. These markers are multi-allelic and are excellent for individual matching of suspect and crime scene DNA. While 13 STRs form the core of the FBI Laboratory's CODIS (Combined DNA Index System), the 10 core loci used in the UK and much of Europe consist of eight loci that overlap with CODIS plus seven additional markers that include the five new European Standard Set (ESS). Discussions on the best options on expanding the core sets of loci are underway (Ge et al., 2012 and Hares, 2012). Online tools and databases have followed to allow users to reference and predict population affiliations using these multi-allelic markers including STRBase (http://www.cstl.nist.gov/strbase/), PopAffiliator (http://cracs.fc.up.pt/popaffiliator/) and pop.STR (http://spsmart.cesga.es/popstr.php). The extensive allele frequency data that have been accumulated over the years in large public databases also allow population-specific estimates of the probability of a random match of two unrelated individuals. However, it is exactly the high level of polymorphism in almost all populations that limits the ability of these markers to determine ancestry of an individual.

Considering the ease, accuracy, and efficiency in typing single nucleotide polymorphisms (SNPs) and their essentially zero rate of recurrent mutation when compared with STRPs, SNPs have the potential to be considered for human identification and description in forensic, biomedical, association, as well as epidemiological studies (Amorim et al, 2005, Gill et al,2004). Also SNP markers serve an important role in analyzing challenging forensic samples, such as those that are degraded, for augmenting the power of kinship analyses and family reconstructions for missing persons and unidentified human remains, as well as for providing investigative lead value in some cases without a suspect (Budowle & van Daal, 2008). With all these advantages of using SNPs over STRPs, more and more SNP panels are being developed for various forensic purposes. These

include panels of identity-testing SNPs, ancestry informative SNPs, lineage informative SNPs and phenotype informative SNPs for various forensic applications. Another class of markers, Insertion-deletion polymorphisms (InDels) also has the desirable characteristics of SNPs and has begun to be used in forensics both for individual identification and for ancestry inference (Watkins et al., 2003 &Pereira et al., 2012). However, considerable research is required to establish a reliable set of panels (SNPs or InDels) containing sufficient numbers of markers to provide excellent discriminatory power comparable to or exceeding that of STR markers. Not only do multiple populations need to be studied to identify the best markers but interpretation of results in any application needs the reference allele frequencies in multiple populations. The discriminatory power for individual identification will be population specific and ancestry inference will only be as good as the set of reference populations. As more data on these SNP and Indel panels accumulate there is a need for assembling the data in a forensic setting to make it available for forensic research and practices.

Online tools and databases based on multi-allelic STRPs (CODIS database) are actively used in forensic teaching, research, and investigations. Though considerable research is being conducted on SNPs for forensic purposes, databases that can be used for forensic research and investigation are limited. Online web tools that demonstrate classification algorithms are being developed for SNP sets like 'The Snipper' app suite (http://mathgene.usc.es/snipper/) for three ancestry informative AISNP sets (34, 32, and 77 markers), Phillips et al., 2007 and SNP*for*ID browser(http://www.snpforid.org/) for the forensic panels developed from SNP*for*ID project. Currently, the main limitations of these tools are the number of SNP sets and the range of population data available for computation. To overcome this limitation we developed an open access web application, FROG-kb (Forensic Research/Reference on Genetics-knowledge base) <http://frog.med.yale.edu> that allows viewing and retrieval of data as well as calculation of statistics on several forensically relevant SNP sets to be useful for teaching and research relevant to forensics.

1.3 Purpose and specific Goals:

The original purpose of the work undertaken under NIJ funding was to provide a web interface to forensically relevant allele frequency data with emphasis on SNPs and their allele frequencies in multiple populations. This web site and underlying database would be useful for teaching and research from a forensic perspective and proactively provide the resource needed to implement SNPs in forensic practice. Such a web site would also focus discussion on the specifics that the forensic community would ultimately want such a resource to provide.

Our FROG-kb seeks to make allele frequency data for SNPs and other genetic polymorphisms more useful in a forensic setting and to serve as a tool facilitating forensic practice. Though FROG-kb is built as a separate interface, it accesses data from the already established and popularly used ALlele FREquency Database, ALFRED (http://alfred.med.yale.edu) described in several publications (e.g., Cheung et al., 2000; Osier et al., 2002, Rajeevan et al., 2012).

The project had multiple specific goals: (1) to develop a web interface to allow searching, viewing, and retrieval of the data in a manner more straight-forward and useful to a forensic investigator and also to allow calculation of elementary statistics for the forensic scientist based on retrieved SNP data. (2) to modify slightly the structure of the existing underlying database ALFRED (ALlele FREquency Database) to store forensically relevant information and to allow access to defined forensic datasets. (3) to enter various forensic SNP panels that were developed in our lab and/or data from our collaborators as well as panels that were published in the literature with the objective of making them retrievable and useful to the forensic scientist, investigator, or student. (4) to develop functions in FROG-kb to facilitate input of ideas and suggestions from individuals from different forensic backgrounds. (5) to develop FROG-kb as a knowledge base. (It is the access to the extensive data in and through ALFRED that helps make FROG-kb a “knowledge base”.) Thus, the ultimate goal is that FROG-kb would allow the adoption of SNPs in various aspects of forensic practice by proactively providing the resources needed to evaluate the data collected and provide a specific view of what the forensic community needed as a basis for future discussion.

The pilot implementation of the FROG-kb web application (http://frog.med.yale.edu/FrogKB) was developed and put online in June, 2011, and was “announced” in a poster at the 2011 NIJ meeting. A full paper introducing the web site and the knowledge base to the forensic community was published in 2012 (Rajeevan et al. 2012), marking the effective public debut of FROG-kb. An updated FROG-kb version went online in January, 2013, and included additional data, functionality, and didactic and explanatory text. Based on the feedbacks and suggestions (mainly from the advisory panel established for an application to the NIJ to continue support and development of the database) some additional improvements and datasets were implemented before the personnel funding ended in June, 2013. The database will remain essentially static but online until new funding is obtained.

Three types of SNP panels are currently implemented in FROG-kb: Ancestry Inference (AISNP), Individual Identification (IISNP), and Phenotype Informative

(PISNP). For the IISNP and AISNP panels, the interface allows a user to input the genotype of an individual for multiple SNPs and have the likelihoods of that multisite genotype for each of several populations calculated and displayed. Various functions have been added including genotype input by file upload and a color code system to flag subsets of SNPs when relevant. The reference functions make use of ALFRED by links to documentation on each SNP, population, and sample for gene frequency estimation. The user interface, SNP sets, and the functionalities provided for the SNP sets were revised continuously as we added new data to ALFRED.

1.4 Review of Literature

Two types of literature review are presented here: one reviewing the available forensic SNP databases and another section citing the papers publishing SNP data that are useful in forensic investigations. The forensic SNP databases include:

The Snipper app suite (http://mathgene.usc.es/snipper/) is an online web site that computes SNP classification of individuals. This open access web portal demonstrates the classification algorithms and error estimation systems documented in Phillips et al. (2007a, 2009) and Pereira et al., (2012). The data input interface and instructions leading to the analysis are simple and straightforward. Many different calculation options are available. Computations can be based on the population data already stored or can utilize custom datasets. The results returned are well documented. Several other options of population genetics relevance are offered.

SPSmart (http://spsmart.cesga.es/): The database is designed for accessing and combining large-scale genomic databases of SNPs for use in population genetics. The available datasets includes the Hapmap data, 1000 genomes phase I data, Perlegen data and the HGDP data from Stanford & University of Michigan. There are separate browsers available for each of these datasets. The browser in the database that is associated with forensic marker data includes the SNP*for*ID (http://www.snpforid.org/) browser that is based on the forensic SNP panels from the SNP*for*ID project. The forensic marker sets included in the browser are the 52-plex individual identification (Sanchez et al., 2006) and the 34-plex ancestry informative marker sets (Phillips et al., 2007b, Fondevila et al., 2012). Functionality includes selecting populations and calculating various statistics (Fst, In, Hobs, Hexp , etc.) and downloading data.

A more general database linking to others and providing methodologic and other materials is the 'SNP section' page in the National Institute of Standards and Technology (NIST) web site http://www.cstl.nist.gov/biotech/strbase/SNP.htm which gives an aggregation of information on forensically informative SNPs.. Another one is the 'Ancestry SNPminer' (https://research.cchmc.org/mershalab/AncestrySNPminer/login.php) which is a data-querying tool and there are many good features for identifying ancestry informative SNPs in a research context but no direct application to forensic practice.

FROG-kb does not supplant any of the above mentioned databases. However, one of the major limitations with the reviewed databases is that they include a restricted number of datasets/ forensic SNP panels with data on a limited number of populations. The current version of FROG-kb already accesses a much larger set of accumulated information through ALFRED and also cites other additional panels that could be added readily. FROG-kb also has functions not otherwise available and the recent enhancements provide even greater depth and enhanced functionality. Our ultimate goal is to provide a 'one-stop shop' pertaining to the forensic SNPs panels for the forensic community and the researchers as well.

The review of forensic literature involves description of several forensic panels--the ones developed from our lab for various purposes, others from the literature and panels that were made available to us through our collaborators. All these are included in FROG-kb and the description of all the relevant forensic panels along with the citations are detailed in the Methods and Results section of this report.

II) Research design and methods

As stated before, though FROG-kb is built as a separate web based interface entirely different from the ALFRED interface, we access the same data as the ALFRED interface. Because many of the relational structures and connections already existed in ALFRED, prototyping the new functionality was made easier by building on existing database structures. Hence, amendments were made to ALFRED database to accommodate new FROG-kb functions. Here, we provide background on ALFRED's basic data structure and the amendments that were made to accommodate FROG-kb functions. The intent and design issues supporting the initial database structure and subsequent modifications of ALFRED can be found in Cheung et al. (2000), Osier et al. (2002), and Rajeevan et al. (2005, 2011). In the following sections only the data structure and system

design of ALFRED that are relevant to FROG-kb are included. The user interface layout and functionalities of FROG-kb are elaborated in the Investigative Genetics paper (Rajeevan et al, 2012).

2.1 ALFRED and FROG-kb Database Design

ALFRED is implemented using a traditional relational structure which is illustrated in appendix Figure 1a. An individual polymorphism (or **Site**) is contained within a locus on the genome. Ethnic populations are organized by their geographic location (**Geographic_Region**). Multiple samples may be drawn from a particular population. For such highly heterogeneous populations as African American or European American, special care is taken to delineate the specific geographic region of the population. Population samples are typed to determine the frequency of alleles at a site. The **Typed_Sample** table bridges samples and polymorphisms and also associates the typing method, which is detailed in the **Typing_Method** table. The allele frequency values for a **Typed_Sample** are stored in the **Frequencies** table. Information about the contributor of particular allele frequency data is kept in the **Contributors** Table.

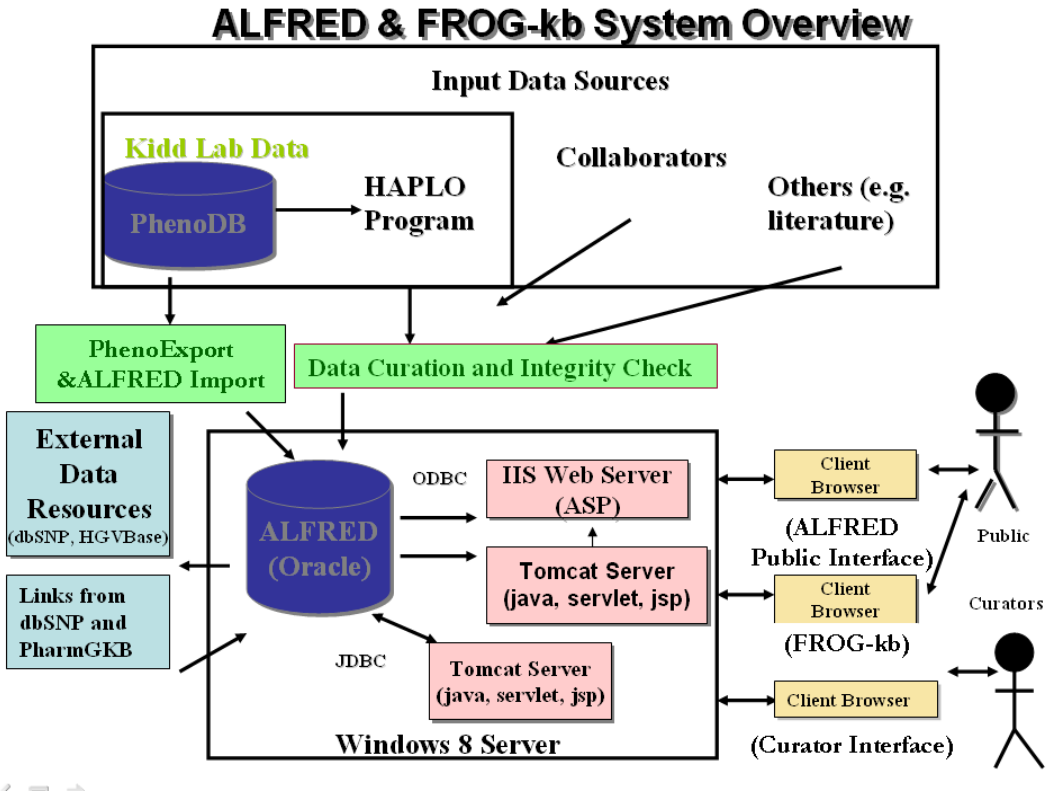
Additional tables were added to the underlying relational database of ALFRED to accommodate information essential for the human identity testing applications in FROG-kb. Figure 1b in the appendix gives the database tables and relationships incorporating only the supplementary tables relevant to FROG-kb. The logic supporting the design of the additional tables and relationships follows. There can be one or many sites (**Sites_Forensic_Panel**) associated with a defined forensic panel (**Forensic_Panel**). Every panel is linked to at least one publication (**Pub_Forensic_Panel**). Such links are clearly identified when the underlying data are unpublished to document the source of the data. The marker phenotype (equivalent to 'genotype' based on multiple unavoidable assumptions) frequency for each 'site - population sample' combination is pre-calculated and saved in '**Forensicphenofreq**'. Since these population-marker frequencies do not change for existing data, pre-calculation of the phenotype summary is reasonable and expedites the involved case-specific computations. While all allele frequency data required for running the computations in FROG-kb were already in the ALFRED database, the new tables provide the framework for displaying information related to the different forensically relevant SNP sets in an efficient and user-friendly manner.

2.2 ALFRED and FROG-kb System Design

The architectural components of our system are depicted in Figure A below. Multiple data sources are used and referenced but a large majority of the gene

frequency data has been obtained from the Kidd Laboratory, from public repositories, and from collaborators. The data stored in the database are made accessible to the public via the web interfaces for ALFRED and FROG-kb. To broaden the utility of the data, links to other relevant web sites are added. For example, we have established links to NCBI's PubMed to connect site and sample descriptions and published frequencies with the original literature. Unseen by the user, there are curatorial software tools that provide integrity checks and allow the curators to more efficiently annotate entries and add web links to appropriate entries in other databases. The tools used to upload allele frequency tables of SNPs are semi-automated to ensure data quality.

Figure A. Basic system overview of ALFRED and FROG



2.3 Informatics Development

The database for FROG-kb is implemented using Oracle version 10 on one of Yale's institutional database servers where it is maintained. The web front end of FROG-kb is built using web developing technologies such as Java, Java Servlet, JSP, JQuery, and GoogleCharts. Almost all of the client-code utilizes JQuery, and the server implementation is in Java. The currently deployed version of FROG-kb has been tested on both PC and Mac, using many different browsers: Mozilla Firefox11.0, Internet Explorer 8.0, Safari 5.1.4, and Google Chrome 17.0

on PC, and Firefox11.0, Internet Explorer 5.0, Safari 5.0 on Mac. We are using Tomcat as our web server which runs on a Windows XP machine.

The user interface layout of FROG-kb is designed to reflect the organization of the contents and functionality, as well as ease of use. Every set of pages relative to a function on FROG-kb originates from a tab on the 'Main Menu' (appendix Figure 2) that appears on the left-hand side of every page. The 'Home Page' gives a brief summary of the functions available in FROG-kb and what can be expected soon. The menu-items are as follows; Home, About, File Upload, IISNP, AISNP, PISNP, Pipeline, Search and Contact Us. All the core functions on FROG-kb can be reached by selecting the tabs 'File Upload', 'IISNP', 'AISNP', or 'PISNP'.

The 'Back' button on the browser does not work on the FROG-kb interface. Therefore, the web pages are designed to allow easy movement among the functionalities. Every set of pages relative to a panel on FROG-kb originates from a labeled link on the 'Main Menu' that appears on the left-hand side of every page. The buttons to all functions relative to a specific SNP set appear on the top of that SNP-set page.

Selection of SNP panels: Users can navigate into the functions relevant to each type of SNP panel by selecting the appropriate tab: 'IISNP' for Individual Identification SNPs, 'AISNP' for Ancestry Inference SNPs and 'PISNP' for Phenotype Informative SNPs. Following selection of a particular SNP panel category (IISNP, AISNP, or PISNP), there are multiple published panels listed. The IISNPs tab has implementation of three different SNP panels, AISNPs tab has implementation of seven panels, and PISNPs tab leads to one panel. Details on each panel are given in a separate section of this report. The citation information for each of the panels, a 'Go' button to navigate into the selected panel, and a 'Detailed Overview of SNPs' link to navigate into ALFRED are provided for each (appendix Figure 3). The link to ALFRED opens the 'SNP Sets' page within ALFRED into a new browser window. The SNP Set module in ALFRED has multiple functions, including the ability to see for each SNP a pie chart on Google Maps of frequencies for all populations with data.

Several options are possible after entering a SNP panel page. The functions related to the selected panel are available by selecting the appropriate buttons at the top of the page (shown in Figure 4 in the appendix). The 'SNP Set' option provides the list of SNPs in the panel. The list includes the dbSNP rs-numbers with active links to the corresponding dbSNP record for molecular characterization of the SNP. The 'Populations' button provides the list of

populations for which comparable calculations can be made. This is the set of populations for which all SNPs in the set have allele frequency data. Conversely, many populations have data on additional SNPs; those SNPs are not included for the calculations. Within the 'SNP Set' functionality in ALFRED additional populations may have data for some, but not all SNPs; those populations are not included in the calculations. Each population name within FROG-kb is an active link to information on the population stored within ALFRED; that page opens in a new browser window. The geographic region of each population is included. A world map in which regions are divided on an arbitrary but convenient basis is available from the link 'Geographic Region Map'.

Functionalities link: There is a 'Functionalities' button under each type of panel which gives a brief description of different functions available for that panel including the detailed overview of SNPs and link to the SNPset page in ALFRED etc.

Examples: In the current version of FROG-kb, all the example files from the main page of each SNP set are consolidated under the 'Examples' button under each panel. There are two distinct example options for users to explore in FROG-kb. Screen shots of sample runs are saved as images with data input and output screen on adjacent panels. These are static screens. The second example option is where a pre-entered data entry page for one individual from a specific population is displayed. This page is dynamic and users can run the compile function from here. The buttons to access these pages are named after the population the individual is from, for example Korean, Hungarian, etc.

Data Entry options: The most significant interactive function is accessed via 'Data Entry' (figure 5 in the appendix) that provides two different options to enter genotype data for a particular SNP panel: **'Selection by Radio Button'** and **'File Upload'.**

a.'Selection by Radio Button' opens the ability to specify an individual's multi-site genotype using radio buttons. The list of SNPs in a set is sorted by rs-numbers for ease of working. For each SNP on the list the ALFRED UID, dbSNP rs-number, chromosome number, and chromosomal position are displayed (appendix Figure 5a). The ALFRED UID and rs-number are URL links to ALFRED and to dbSNP SNP information pages, respectively. This is followed by radio buttons for the possible genotypes. The genotype is entered by simply clicking on the radio button for the genotype at each SNP. An obvious assumption is that there is no allele drop out, that is, that a typing result (phenotype) with only one allele detected is really a homozygote. A radio button

labeled ‘NN’ is provided for missing data for each SNP, but it is not necessary to click on the ‘NN’ for missing data. For large SNP sets, if the user’s SNP set is also ordered by rs-number in a spreadsheet, selecting the appropriate genotype radio-button should be relatively effortless. At the bottom of the list are three buttons (appendix Figure 5a): Set all unselected to unknown, Print Format, and Compile. The Print Format will generate a condensed version of the input data that can be printed as a permanent record of the input data. The information in the pop-up window can also be copied and pasted into a text editor which in turn can be opened in an Excel spreadsheet. The Compile will initiate calculation and display the results. If there are SNPs with no genotype selected, a warning will be sent with the missing data rows highlighted for easy detection and the option exists to examine which SNPs have no entry and to either enter a genotype or use the ‘Set all unselected to unknown’ option to fill those with ‘NN’. It is necessary to click on ‘Compile’ again to generate results.

b. *‘File Upload’* function (appendix figure 5b) provides users with an option to enter SNP information and corresponding genotype for a panel in a text area. Data input using the radio button option can be tedious for SNPsets that include many markers. Therefore, the file upload option is provided on the main-menu on the left-hand-side of the web interface. This function enables users to input an individual’s genotype for a specific SNPSet in a much more user friendly manner. The ‘File Format’ page within the function provides a sample file for each SNPSet in FROG-kb. The ‘Input Genotype’ page provides a text area for users to insert the required information to compute likelihoods on the inserted data. This input should have information regarding the panel the data will be checked against. The ALFRED_UID and the genotype are the only two fields that are important. The genotype should correspond to the alleles given in the file; otherwise the program will ignore the SNP. The columns are expected to be tab-delimited. Example downloadable files for each SNP panel are provided to provide users with a template. Users can save these files and modify them by replacing the 'NN' (unknown) genotype with the observed genotype of the individual. Each file starts with the SNPSet tag. The tag provides information to the internal code on the type of SNPSet, e.g., 'ai34' for SNP*for*ID 34-plex and 'ii52' for SNP*for*ID 52-plex, etc. The file upload option was originally included only under the ‘Main Menu’. We then realized that someone using the ‘Data Entry’ function from the individual SNP panel page might never notice the ‘File Upload’ option from the main page. Therefore, we have also placed the ‘File Upload’ option with appropriate text under the ‘Data Entry’ for each panel so a user entering the ‘Data entry’ page of each panel gets 2 options to choose from: ‘File upload’ and ‘Selection by radio button’. Detailed text with example files explaining the format that is required for the data input is provided in the page that comes

up when the 'File upload' option is clicked. We believe these options give the user a greater flexibility for data inputs.

It was brought up by one of our users that after computing the likelihood for an individual's available genotypes it was impossible to go back to that particular 'Data Entry' page with genotypes still filled in for that individual since the 'back button" does not work in FROG-kb. Changes were made to the code so that clicking the 'Data Entry' button will take the user back to the user's selections. The user is notified about this with a note displayed right above the likelihood graph.

The '*Formula'* button gives an explanation on how the probabilities are derived.

Results of the calculation are displayed as a table with three columns: the name of the population sampled with its geographic region and the sample size, the probability of the entered multilocus genotype in that population, and the likelihood ratio of the most probable population to each specific population (appendix Figure 6). The new addition to the results page is the third column giving the likelihood ratio of the best relative to each subsequent population. This makes it easier to see that the best is, for example, twice as likely as the second best, or 200 times as likely as the 12th best population. The probability of the entered genotype is equivalent to the likelihood of that population being the origin of the genotype assuming no deviation from Hardy-Weinberg ratios in the population. The populations are ordered by their likelihoods as the origin of the entered genotype from highest to lowest; therefore, the likelihood ratios range from 1 to some larger number for the least likely population of origin, representing how many times more likely the most likely population is compared to the specific population as origin of the entered genotype. As a rough measure of significance a flag indicates those populations with probability values that are within an order of magnitude of the best and are therefore considered not significantly less likely to be the most likely. This "duplicates" information in the likelihood ratio column but adds emphasis since it is important to recognize uncertainty.

Color code system to flag SNPs: This is a functionality that was developed to accommodate addition of SNP panels that are derived from multiple SNP panels that are already in FROG-kb. Amendments were made on the web interface to accomplish this feature. Most specifically, the panel of 55 AISNPs is derived from two unpublished AI panels (Set of 39 & Set of 44) from our lab. These two panels are published in ALFRED (http://alfred.med.yale.edu/alfred/selectedSnpSet.asp?setId=141 and

http://alfred.med.yale.edu/alfred/selectedSnpSet.asp?setId=241).The panel of 55 is published in FROG-kb. It was important for the authors to convey to the users the original SNP panels and which SNPs differed. This was achieved by adopting a color code system wherein yellow signified a SNP from “Set of 39” and blue a SNP from “Set of 44”. These codes are displayed on the ‘Data Entry’ page and ‘SNP List’ page (appendix Figure 4 and 5a). The URL to the corresponding complete set in ALFRED is also provided. The same system could be utilized to represent addition of a new SNP to a panel already in FROG-kb.

The ‘Pipeline’ link in the Main menu gives the graphical summarization of the procession through the interface. A ‘Contact Us’ button is available under the Main menus for users to contact the ‘FROG-kb Team’.

2.4 Data curation and entry of Forensic SNPsets:

The usefulness of any database is dependent on the quality of its contents. The detailed curation involved in locating, assembling, and recording the data from the literature is the key feature of ALFRED and such assembled forensically relevant data in ALFRED ‘SNPsets’ page are accessed by FROG-kb. Besides the panels that are developed in our lab for forensic purposes, we systematically screen the literature in order to include population allele frequency data for any new panel that is published. Also population allele frequency data for any SNP in any of the existing panels are added. When a new population has data for all SNPs in a panel it is included in the calculations in FROG-kb.

The manual curation of the data involves thorough scanning of the literature using search engines like Pubmed, Scopus, Google Scholar, etc. using various combinations of key words (eg. forensic snps, forensic data, etc). All the journals that publish the majority of the forensic SNP-related data have been individually scanned regularly for early access ‘Article in press’ sections. These include the International Journal of Legal Medicine, Forensic Science International: Genetics supplement series, Journal of Forensic Sciences, etc. In addition various journals that routinely publish genetics data including Human Genetics, European Journal of Human Genetics, Investigative Genetics, Genetics and Molecular Research, Mutation Research, Annals of Human Genetics, Human Biology, etc. are scanned to locate any forensically relevant SNP data for FROG-kb. Apart from these, all the forensically relevant journals that are available with full access to articles through the Yale Library website are also checked to make sure we do not miss any forensic SNP data on hand in the literature.

Once the publication having relevant information is identified, the data, if available in the publication, are assembled and formatted for entry into ALFRED

and included under the relevant SNPset. The citation is also entered and linked to the data. For those papers that do not provide the frequency data we send requests to the authors for submission of data. In some cases, curation also involves matching of the markers from a set if identified by different names.

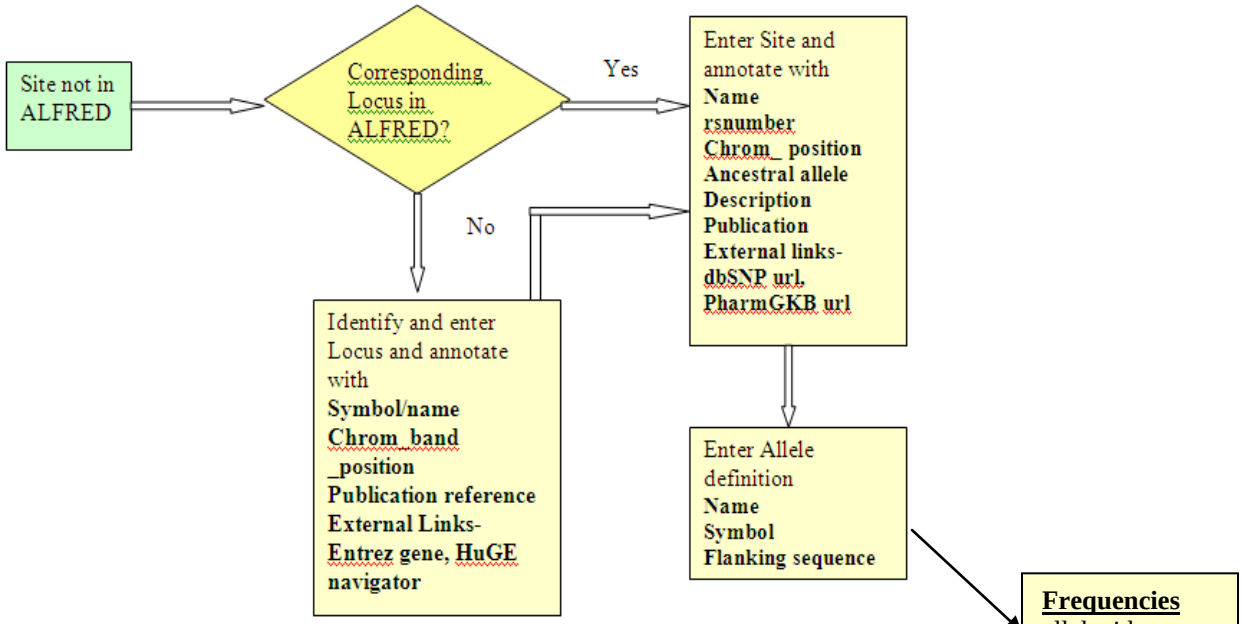
For the panels that are developed in our lab, the dataset is assembled and entered in ALFRED and then defined as a SNPset. Besides the curatatorial steps in defining the systems and population/sample (described below) entry of data involves an additional step of mapping of the ALFRED system ids to our lab database Phenodb ids and extracting the data for uploading into ALFRED using a program 'Phenoexport' in place for this purpose.

Curation also involves identifying and entering explicit links to dbSNP and additional databases to define the molecular nature of the polymorphism in genomic context. Also, explicit links are inserted to define the population and population sample underlying any frequency data. The general curatorial steps involved in entering the data in ALFRED are illustrated in the flowchart below. For the forensic SNP panels, once the data are entered in ALFRED, the SNPs are defined as a SNP set in the specially created "SNP Sets" function page under the 'Search' tab on the ALFRED homepage. These data are eventually accessed by FROG-kb and defined as a panel in FROG-kb.

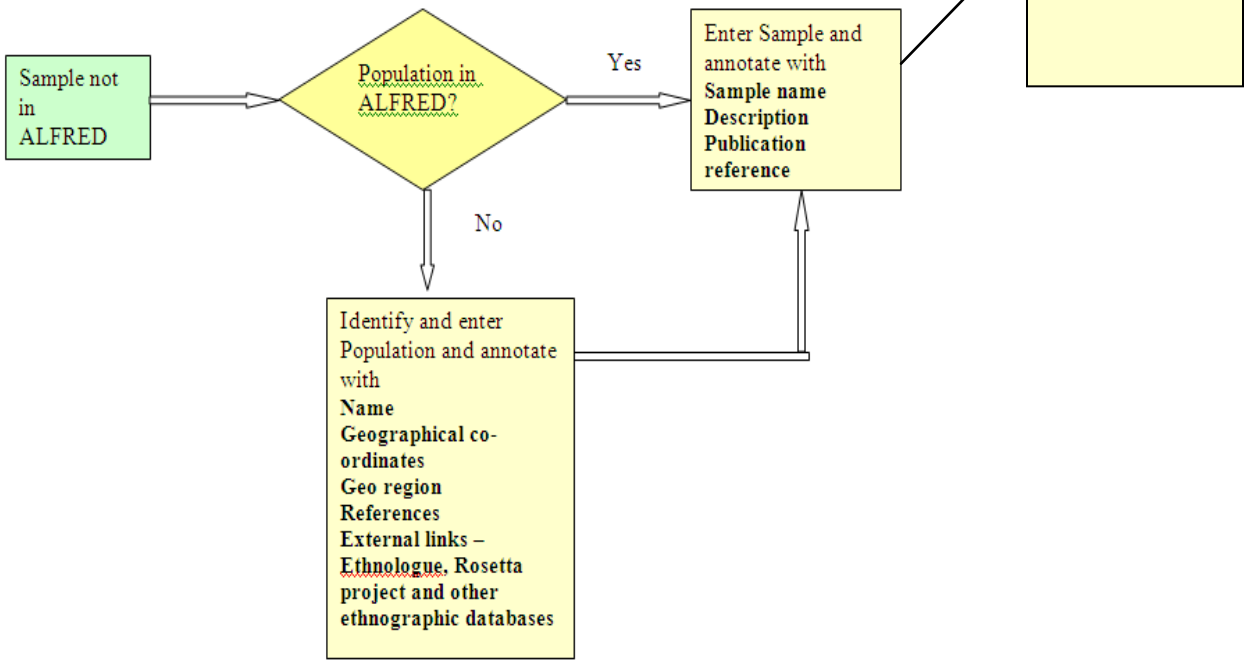
All these features of the data curation make FROG-kb a 'unique resource' of forensic data. These efforts are time consuming and are not innovative, but high quality of the data is critical and the collation into one central place is important. While much more data could be incorporated into FROG-kb, the budget did not provide sufficient personnel time to do more than currently exists in the static version of FROG-kb. When more funding becomes available, additional data already curated in ALFRED can be incorporated into FROG-kb.

This flow chart gives a quickview of the required details of curation for new data to be entered before the frequencies are uploaded into the ALFRED database.

Entering Locus, Site and Allele



Entering Population and Sample



Based on the above mentioned curatorial steps various panels were defined in ALFRED as SNPsets and have been included in the FROG-kb interface. These include panels from our lab, literature curation, and those that were made available to us through requested data submissions and from our collaborators. The current version of FROG-kb includes the following eleven panels.

II (Individual Identification) SNP panels:

45 unlinked II SNPs from KiddLab**:** The IISNP markers studied in the process of developing our IISNP panel are entered into ALFRED (Kidd et al., 2006; Pakstis et al., 2007; Pakstis et al., 2010; and Kidd et al., 2012). Thus ALFRED includes data on those SNPs ultimately deemed inappropriate. This 45-SNP panel had been tested on 44 worldwide population samples from our lab and on an independent set of Chinese (Lou et al., 2011). In addition, many of the markers have additional population frequencies included from various sources. All these data are available in ALFRED under the 'SNPset -> II SNPsets' link. The calculations available in FROG are so far restricted to the 44 populations with allele frequency data for all 45 IISNPs.

The 52-plex (SNP*for*ID) panel (Sanchez et al., 2006) is an individual identification panel initially tested on 9 populations; a subset of 26 of these markers have data on at least an additional 44 of our populations in ALFRED. Additional population data have been included for these 52 markers or subsets of 49 SNPs from all these publications (Farzad et al., 2013; Tomas et al., 2013; Barbaro et al., 2012; Ruiz et al., 2012; Poulsen et al., 2011; Drobnic et al., 2010; Montelius et al., 2009; Porras et al., 2009; Barbaro et al., 2009; Tomas et al., 2008; Pereira et al., 2008a, Pereira et al., 2008b). This panel is available in FROG-kb under the "IISNP" tab.

Qiagen Investigator DIPplex kit: This is an IISNP set which we prioritized entering as a panel in FROG-kb because there is a commercial kit available now. This set includes 30 di-allelic deletion/insertion polymorphisms (DIPs aka InDels) developed for multiplex amplification and Amelogenin (http://www.qiagen.com/products/investigatordipplexkit.aspx). We accumulated data from the literature on various population samples typed using the kit for a total of 17 population samples for this marker set (Fondevila et al., 2012; Neuvonen et al., 2012; Turrina et al., 2011; Kis et al., 2012; Zidkova et al., 2013; Friis et al., 2012; Larue et al., 2012; Martín et al., 2013 , Carvalho et al., 2013). Only 13 of these samples have the complete set of data on 30 indels and hence can be included in FROG-kb for probability calculation purposes. However, the ALFRED SNPset provides access to all the available data for this set.

AI (Ancestry informative) SNP panels in FROG-kb:

Kidd Lab set of 55 AISNPs: This panel of AISNPs (manuscript submitted) is the union of two subpanels- KiddLab - Pilot Panel of 39 AISNPs and KiddLab - Set of 44 AISNPs. All SNPs have been tested on 63 population samples in our lab. All the markers that were studied during the process of identifying this subset are entered in ALFRED but FROG-kb includes only the final 'set of 55 AI SNPs'. The full 55-SNP panel has been implemented with color coding (described in the informatics development section of this report) to show which marker belongs to which of the two subpanels. Also available for most of the SNPs from this panel are population sample data from the 1000 Genomes project, the eleven Hapmap Phase 3 populations, and the HGDP populations.

Seldin's set of 128 SNPs (Kosoy et al., 2009): For this set we have included data on all the populations that were studied in Kosoy et al. (2009), 11 HapMap population samples, the HGDP population panel, and data on 73 population samples tested in Kidd Lab (Kidd JR et al., 2011). These additional data from various sources provide a set of 119 populations for calculation of the likelihoods of ancestry based on this panel.

The 34-plex (SNP*for*ID) AIMs panel (Phillips et al., 2007) is an ancestry informative panel initially developed to distinguish between sub-Saharan Africans, Europeans, and East Asians. The original publication includes the data on 51 HGDP population samples for this set. In addition FROG-kb contains the data for Hapmap population samples, data from two other publications which typed this panel on 3 population samples (Phillips et al., 2009; Khodjet-el-khil et al., 2010) and also the 1000 Genomes population samples. Also available in ALFRED for this set are data from various other sources bringing the total number of population samples to above 100 for most of the markers. Fondevila et al. (2013) reported the replacement of one of the underperforming SNPs from the panel by a more informative SNP, rs3827760.

Kayser's set of 24 Ancestry Informative Markers (Lao et al., 2010)*:* Manfred Kayser gave us the genotype data for these AIMs markers typed on the 52 HGDP populations. We calculated the allele frequencies and have included the data as a SNPset under the AISNPs link and also in FROG-kb.

Podini's panel of 32 AISNPs (Gettings et al., 2014): This is a panel of 32 AISNPs from Daniele Podini, George Washington University who submitted the data to us to make it available in FROG-kb. Though this panel has 32 markers, only 28 markers have been typed on a cohesive set of populations (4 of the markers from the set do not have data on HGDP populations). Only 25 markers from this set

are typed on our lab sample collection. In both cases all the population allele frequency data are in ALFRED. So, for reference the ALFRED SNP set have all the 32 markers and available data on each, but in the FROG-kb panel we have only included 25 HGDP SNPs for which comprehensive data exist on more populations allowing for meaningful likelihood ratio calculations. In the 'SNPset' page of this panel in FROG-kb we have included a clickable button saying 'Why only 25' which explains the reason for including only 25 SNPs. When other markers in the panel are typed on the whole list of populations, they will be added to the calculation.

Eurasiaplex 23-SNP panel: We had data submission from Bulbul and Phillips which included the analysis from 85 SNPs related to bio geographic ancestry and pigmentation type from Bulbul et al. (2011). This set of SNPs includes the SNP*for*ID's 34-plex AIMs panel, 32-plex (22 AIMS markers they have identified and 10 pigmentation markers) and the Eurasiaplex panel. They submitted the genotype data for these 85 markers typed on 20 population samples (1000 Genomes phase 1 reference samples and some additional samples from their collection). We have included all this data in ALFRED but not defined as a SNP set. A recent publication by Phillips et al. (2013) highlighting the 23-SNP Eurasiaplex panel was identified through literature search and hence we have included this panel as a SNP set in ALFRED and also in FROG-kb as an AISNP set. The panel is linked to both the earlier publication of Bulbul and the recent one by Phillips. The recent publication provided additional population data for this panel including the 1000 Genomes, Hapmap, HGDP population collection samples as well as some European population samples. We have included all these additional data for this panel thus allowing meaningful likelihood calculation on 76 population samples.

Carolyn Nievergelt's set of 41 AIMs (Nievergelt et al., 2013): This panel includes 41 markers that were selected from the HGDP data. We have typed the KiddLab 47 standard populations for 40 of these markers. All the data are included in FROG-kb thus allowing calculations for 40 of the SNPs on 99 population samples.

Phenotype Informative (PI) marker panel:

Irisplex: The six-SNP Irisplex for eye color prediction from Walsh et al. (2011). Data on these six markers for the 52 HGDP population samples and data on European populations from Walsh et al. (2012) are available for reference in ALFRED. The data on the additional European populations were obtained in response to a data submission request. The implementation in FROG-kb for this

panel differs from that for IISNP and AISNP calculations in that calculations are not dependent on population but only on the genotypes entered. The eye color prediction computation uses the formula in Walsh et al. (2011).

Forensic datasets in ALFRED but not in FROG-kb:

Besides the ones in FROG there are various SNP sets on different populations for forensic use that are entered into ALFRED and defined as SNP sets.

The IISNP sets include Dixon's 21-plex assay (Foren-SNP multiplex kit), the panel of 25 IISNPs from Pietrangeli et al. (2010), Vallone's set of 70 SNPs(Orchid SNPs) and the CODIS set. The forensic related data in ALFRED but not defined as a SNPset for the IISNP category include the 30 SNPs selected from the Affymetrix 500K SNP dataset tested on a sample of 960 Korean individuals (Kim et al., 2010), a 24 SNP set studied on a Korean sample (Lee et al., 2005), a 16 SNP set studied on Japanese (Hiratsuka et al., 2005) , data on two multiplex AIMs panels studied on 16 Latin American population samples (Silva et al., 2010), and a set of 30 Indels studied on 5 Chinese populations (Li et al., 2011). These have been and remain of very low priority for inclusion in FROG-kb because the few populations studied make the likelihood calculations of little value. Were that to change, they could be considered for implementation in FROG-kb.

The AISNP sets include the two multiplex AIMs panels studied on 16 Latin American population samples from Silva et al. (2010) and the Visigen (Visible Trait Genetic consortium) data (Keating et al., 2013). This consortium has developed the Identitas Version 1 (v1) Forensic Chip based on Illumina Infinium technology. Genotype files obtained from the Visigen group contained data for 197,353 SNPs on 29 populations (including limited numbers from 27 Kidd Lab populations and two populations from others in the consortium). We have calculated and uploaded the allele frequency data for all these markers into ALFRED.

The PISNP sets we have entered in ALFRED include both (1) Kayser's 43 SNPs for Hair Color Prediction, which is a set of 43 SNPs defined for hair color prediction based on Branicki et al. (2011) and (2) the Ruiz et al. (2013) 23 SNPs for eye color prediction based on studies on six European populations. We obtained these data by requesting the submission of the data and have included them in ALFRED. The data for using the Ruiz et al. (2013) panel for estimation of eye color likelihoods are not yet included.

Another class of markers for which data have been included in ALFRED includes the Lineage informative (LI) SNPs. These are sets of tightly linked SNPs that function as multi-allelic haplotype markers on the autosomes that collectively may serve for matching individuals to family or clan. Such markers could be used to identify missing persons through kinship analyses. These LISNPs may also serve as Ancestry Inference (AISNPs) markers. ALFRED contains data from Pakstis et al. (2012) "Mini-haplotypes as Lineage Informative SNPs (LISNPs) and Ancestry Inference SNPs (AISNPs)" and the expanded panel of 25 unlinked minihaps presented in a paper being drafted.

The recent work in our lab had been focused on moving to smaller genomic regions in the search for **microhaps** (regions that can be covered in one sequence read) for various forensic purposes. From this initial screening we have identified 30 micro-hap sets that are of forensic interest. The data for 20 of these micro-haps on 57 KiddLab population samples have been added to ALFRED with adequate definitions. These set of markers can be retrieved by typing “Microhap” in the keyword search function in ALFRED. New interfaces will need to be designed and programmed before these can be implemented in FROG-kb.

III) Results

As per the goals of the project, we have made significant progress and have a functional website online accessing a specifically modified and enhanced database. We have published a manuscript in ‘Investigative Genetics’ to introduce the web site and the knowledge base to the forensic community (Rajeevan et al., 2012).

The FROG-kb interface is designed to reflect the organization of the contents and functionality, as well as ease of use. During the first year of the grant amendments were made to the backend ALFRED database and new linking tables were created to accommodate information required for the FROG-kb web application. The pilot implementation was put online in June of 2011which included the IISNPs and AISNP from the menu bar with some information under the ‘About’ and ‘Pipeline’ items to provide information about the underlying calculations. Two published panels of IISNPs (the Kidd Lab 45 IISNPs and the SNP*for*ID consortium 52 IISNPs) and two different AISNP panels (The Seldin Lab’s set of 128 AISNPs and Kidd Lab pilot panel of 39 SNPs) were included with this implementation. The IISNP feature provides examples and the ability to calculate match probabilities for user-specified genotypes and the AISNP feature this pilot implementation provides examples and the ability to calculate relative likelihoods of ancestry from different populations for user-specified genotypes.

Under each panel various buttons provide the list of SNPs, two examples, the populations for which there are data, a blank genotype entry form, a clear option, and two pre-loaded entry forms for real individuals were available. For both of the panels pre-loaded data entry pages are provided as a training set for new users. Explanatory materials explaining the calculations were included later.

During the early period of the second year of the grant various amendments and additions were made to the database. The changes/additions include the modification of likelihood calculations for the SNP sets in response to a feedback. This modification was implemented since special consideration was needed for those situations in which one allele is not observed in a population and hence the observed allele frequency is zero and two genotypes are strictly estimated to be zero. This is especially important for AISNPs since many of the loci are fixed in some populations, but even a SNP in an IISNP panel may not be seen in an isolated population. Given the sample sizes involved, very low frequencies of the 'missing' allele cannot be excluded and using a value of zero in the calculations would be incorrect. The approach used is to simply use a very small allele frequency instead of the zero that is the allele frequency present in ALFRED. The modification was implemented as follows; if one assumes a heterozygote might be seen in the next individual sampled from the population, the allele frequency would then be1/ (2n+2) where n is the original sample size in which the allele was not seen. Various interface developments were added including two new buttons 'Functionalities' and 'Formula' that could be accessed from each SNP set page, A world map in which regions are divided on an arbitrary but convenient basis was made available from the link 'Geographic Region Map' and Populations table which provides the list of populations the calculations are based on has the geographic region and the sample size were added. During the later part of the second year various other functionality amendments were implemented including the 'Color code system to flag SNPs' that are derived from multiple SNP panels, a new option for data entry 'File upload' was created in addition to the data entry method of selection using radio button. The population likelihood table was complemented with flagging of the populations that are within an order of magnitude of the highest likelihood. Various relevant databases and resources were aggregated for the reference aspect of FROG-kb. These were added as 'Links' from the about page. And also all the presentations related to FROG –kb project were added as 'Links' under the 'About ' page in the main menu . A 'Contact Us' link was added that opens up a form where users can enter their email address, a subject and a message.

A new deliverable improved version of FROG-kb was put online in the beginning of January of 2013 with several newly implemented functions to the web interface

described above. We sent the FROG-kb link to several members of forensic community to solicit feedbacks and suggestions. Based on their feedbacks and also suggestions from our own FROG team, functional amendments were made to the web interface and additional datasets were prioritized and added during the final period before the grant funding ended and the staff were laid off at the end of June, 2013. The user interface layout and all the functionalities available in the current version are detailed in the above section (see 2.3 Informatics development) and also detailed in the FROG manuscript which is attached in the Appendix.

The accumulation of all the forensic panels datasets have been an ongoing process throughout the period of the grant. Based on the curational and data entry method described we assembled data for several forensic panels in FROG-kb and in ALFRED (listed in the Research and Methods section under 2.4 Data curation and entry).

The discriminatory power for individual identification will be population specific and ancestry inference will only be as good as the set of reference populations. We believe we have progressed in the right track to accomplish our main goal of prioritizing and incorporating all these additional datasets from various sources into FROG-kb and also systematically adding additional population data for a panel as they become available which provide in the meaningful interpretation of the results. In summary, currently in FROG-kb there are three IISNP panels, seven AISNP panels and one PISNP panel. Each of these panels exists with supporting population data. URL links exist to ALFRED for more details and allele frequency data tables for specific populations. Data exist in ALFRED not only for the populations used for calculation in FROG-kb but also for each SNP-population combination for which data are available.

As a knowledgebase we have added explanatory text and examples for the functionalities in FROG-kb and notes on interpretation of results. Links to other relevant databases have also been added. Several of our colleagues who are forensic practitioners and researchers have agreed to be part of a formal advisory panel (See Appendix). We have been consulting with them and getting feedback in the design and amendments of FROG-kb, including those for future enhancements.

IV) Conclusions

4.1 Discussion

With the original objectives of the project in perspective during the period of this grant we have provided a 'one-stop shop' web site and underlying database on forensic SNP panels that is intended to be useful for the forensic community and researchers. FROG-kb's user interface is versatile in its functionality and comprehensive in the population data available for many SNP sets. The current interface allows viewing and retrieval of data as well as calculation of statistics on several forensically relevant SNP sets. The recent enhancements to the interface including the color coding system to flag subsets of SNPs, a new option of data entry using file upload, and the likelihood ratio column in the results table have made the database more user-friendly. At the same time they allow more meaningful interpretation of the results. The print option allows the user to have a condensed version of the input data as well as the table format of the results as a permanent record. The results page also contains a graphical representation of results with the log of the probabilities of the entered genotype plotted from highest to lowest as a mouse-over graph specifying the population and value at the point indicated. Ancestry, Individual, and Phenotype Informative (AISNP, IISNP, and PISNP) panels studied and published from the Kidd Lab and elsewhere are currently available in FROG-kb. Each of these panels exists with supporting population data. URL links exist to ALFRED pages for more details including allele frequency data tables for all available populations for each SNP in each panel. As a knowledge base we have also added good explanatory and didactic material to the existing functions in FROG-kb. 'Examples' under each SNP set provides the users an option to explore functionality before experimenting by entering a new genotype profile of an unknown or a forensic case. Links to other relevant databases and poster presentations on FROG have been added. The 'contact us' link helps users to give us feedback. Several of our colleagues who are forensic practitioners and researchers have agreed to be part of a formal advisory panel for future development. After the improved version of FROG-kb was put online at the beginning of this year we have been consulting with them and getting feedback in the design and amendments of FROG-kb. While they have raised issues and made suggestions for future development, they helped us set priorities to complete before the end of the project and have agreed that what has been done so far was the most important. Thus, we believe we are moving in the right direction towards our ultimate objective of providing the resources and functionality that will facilitate the implementation of SNPs in forensic labs.

FROG-kb is a work in progress. Hence, the current implementation is not without limitations. FROG-kb can only provide information for the populations comprehensively tested for the entire set of SNPs in a panel. For ancestry inference FROG-kb is currently designed for individuals whose ancestry is overwhelmingly from one population or set of closely related populations. Admixed individuals, in the sense of recent ancestors from geographically and genetically different populations, will not necessarily provide meaningful results. And currently the allele designations for the genotypes listed on the data entry forms are not consistent with any one of several existing standards. Since most SNPs are unambiguous with respect to the strand being called, the user should have no problem making the necessary conversion from the typing data to the genotype codes on the input screen. However, G/C and A/T SNPs need to be specified as to the strand being called; the future standard for FROG-kb will be that the positive strand (pter to qter, 5' to 3') will be the reference even for SNPs in genes that are coded on the reverse strand but that has not been fully implemented.

Our initial efforts in developing this resource for the forensic community has necessarily focused on the database structure and the website interface with various different panels implemented, as was the objective of the project. We have focused on what we believe were the most important features. However, we consider this a pilot effort, albeit a highly functional one. Whether or not the current version of FROG-kb continues to be supported, it serves as a prototype for one approach to a database that can be a reference and resource on genetics for the forensic community. Moreover, any other database that may supplant it will have access to the data that have been accumulated--they are all freely available in the public domain and can be downloaded from ALFRED.

We recognize various limitations to the current implementation and also that adding various functionalities would enhance FROG-kb. Some of the potential enhancements include: (1) Improve the quality and completeness of the printable output to make it presentable in court (with date, time and case no). (2) Develop a standalone version of the likelihood calculator for selected SNP panels. This would be useful when professionals in the field are working on a certain case, most of the information is classified, and running the computation over the internet on a public server would not be appropriate. Also, there could be circumstances in which an internet connection is not available to access FROG-kb to run the module on a set of SNPs. (3) Implement new improved panels, such as panels of Lineage Informative markers and phenotype inference SNPs. (4) Add a search function to identify the existence and affiliations of a set of SNPs in FROG-kb. (5) Generate tools to draw a global density map of

population likelihoods. (6) Allow sorting of SNP lists in each panel by genomic position. (7) Add various SNP panels and add more population data for the existing panels. and (8) Provide a more comprehensive user manual and more didactic information.

4.2 Implications of policy and practice

Policy: In a forward-looking manner FROG-kb was designed to allow the adoption of SNPs in various aspects of forensic practice by proactively providing the resources needed to evaluate the data collected on SNPs. Commercial kits are becoming available now for use in forensic practice; they implement many of the panels already available in FROG-kb. We believe that having a specific forensic web front-end to a large database with extensive SNP data will facilitate understanding of and use of SNPs in forensic settings. It is our expectation that the match and ancestry calculations will eventually be accepted in the courts since the underlying databases are and will be public and sufficiently large to match the existing STRP databases for population frequencies.

Practice: The didactic aspects of FROG-kb should be of great value in helping the forensic technicians understand and use SNPs. The commercial kits allow implementation of SNP typing in forensic labs. If the kits and calculations are accepted in the courts, the results produced by FROG-kb in a printed report (as already exists) might be directly submitted as evidence.

Usage: Because the vast majority of forensic laboratories are not using SNPs, the only “usage” currently is related to education and awareness. But the ultimate objective is to provide the resource and functionality that will facilitate the implementation of SNPs in forensic practice. However, the community is already aware of FROG-kb and with the publicity we hope to generate even more awareness in the community. The companies are also beginning marketing of their kits and are in many cases mentioning the published panels available in FROG.

4.3 Implications for further research

Many ways in which FROG will need to be enhanced are already evident (see above). As another example, one kit being marketed by LifeTech combines in one multiplex assay two different autosomal panels. Those panels have many populations in common, but differ for coverage of other populations. How does FROG handle such datasets? Similarly, FROG does not currently have any lineage inference panels but Kidd lab has just begun making one public through ALFRED. A new interface for radio buttons and for a file upload needs to be

designed for such multi-allelic markers; it will need to be user friendly. Given the large number of SNPs in ALFRED, how can a user with a "custom" or "haphazard" set of SNP data manage to use those data through FROG? Should that even be an objective?

The field is rapidly evolving and there will undoubtedly be new ways that ALFRED and FROG-kb can serve the forensic community. One particular interest is in enhancing the slides used in talks to make the verbal aspects less necessary by incorporating that material into text. Such powerpoint files can then be made readily available almost as a course on SNPs in forensics accessible through FROG-kb.

V) References:

Amorim, A., & Pereira, L. (2005). Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Sci Int, 150*(1), 17-21. doi: 10.1016/j.forsciint.2004.06.018

Barbaro, A., Phillips, C., Fondevila, M., Lareu, M., & Carracedo, A. (2012). Genetic variability of the SNPforID 52-plex identification SNP panel in Italian population samples. *Forensic Sci Int Genet, 6*(6), e185-186. doi: 10.1016/j.fsigen.2012.07.002

Barbaro, A., Phillips, C., Fondevila, M., Lareu, M., & Carracedo, A. (2012). Genetic variability of the SNPforID 52-plex identification SNP panel in Italian population samples. *Forensic Sci Int Genet, 6*(6), e185-186. doi: 10.1016/j.fsigen.2012.07.002

Barbaro AA, Phillips C, Fondevila M, Carracedo A, Lareu MV,. (2009). Population data of 52 autosomal SNPs in Italian population. *Forensic Sci Int: Genetics Supplement, 1*, 351-352.

Budowle, B., & van Daal, A. (2008). Forensically relevant SNP classes. *Biotechniques, 44*(5), 603-608, 610. doi: 10.2144/000112806

BulBul O, Filoglu G, Altuncul H, Aradas AF, Ruiz Y, Fondevila M, Phillips C, Carracedo A, Kreigel AK, Schneider PM. (2011). A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type. *Forensic Science International: Genetics, 3*, e500-e501.

Carvalho, A., & Pinheiro, M. F. (2013). Population data of 30 insertion/delection polymorphisms from a sample taken in the North of Portugal. *Int J Legal*

Med, 127(1), 65-67. doi: 10.1007/s00414-012-0693-7

Cheung, K. H., Miller, P. L., Kidd, J. R., Kidd, K. K., Osier, M. V., & Pakstis, A. J. (2000). ALFRED: a Web-accessible allele frequency database. *Pac Symp Biocomput*, 639-650.

Cheung, K. H., Osier, M. V., Kidd, J. R., Pakstis, A. J., Miller, P. L., & Kidd, K. K. (2000). ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. *Nucleic Acids Res, 28*(1), 361-363.

da Silva, C. V., Matos, S., Costa, H. A., Morais, P., Dos Santos, R. M., Espinheira, R., Santos, J.C., Amorim, A. (2013). Genetic portrait of south Portugal population with InDel markers. *Forensic Sci Int Genet, 7*(4), e101-103. doi: 10.1016/j.fsigen.2013.03.009

Dixon, L. A., Murray, C. M., Archer, E. J., Dobbins, A. E., Koumi, P., & Gill, P. (2005). Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. *Forensic Sci Int, 154*(1), 62-77. doi: 10.1016/j.forsciint.2004.12.011

Fondevila, M., Phillips, C., Santos, C., Freire Aradas, A., Vallone, P. M., Butler, J. M., Lareu, M.V., Carracedo, A. (2013). Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. *Forensic Sci Int Genet, 7*(1), 63-74.

Fondevila, M., Phillips, C., Santos, C., Pereira, R., Gusmao, L., Carracedo, A., Butler, J.M., Lareu, M.V., Vallone, P. M. (2012). Forensic performance of two insertion-deletion marker assays. *Int J Legal Med, 126*(5), 725-737. doi: 10.1007/s00414-012-0721-7

Ge, J., Eisenberg, A., & Budowle, B. (2012). Developing criteria and data to determine best options for expanding the core CODIS loci. *Investig Genet, 3*, 1. doi: 10.1186/2041-2223-3-1

Gettings, K. B., Lai, R., Johnson, J. L., Peck, M. A., Hart, J.A., Dressman, G. H., Schanfield, M. s., Podini, D. S. (2014). A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population. *Forensic Science International: Genetics*, 8,101-108.

Gill, P., Werrett, D. J., Budowle, B., & Guerrieri, R. (2004). An assessment of whether SNPs will replace STRs in national DNA databases--joint considerations of the DNA working group of the European Network of

Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDAM). *Sci Justice, 44*(1), 51-53.

Hares, D. R. (2012). Expanding the CODIS core loci in the United States. *Forensic Sci Int Genet, 6*(1), e52-54. doi: 10.1016/j.fsigen.2011.04.012

Hiratsuka, M., Tsukamoto, N., Konno, Y., Nata, M., Hashiyada, M., Funayama, M., & Mizugaki, M. (2005). Forensic assessment of 16 single nucleotide polymorphisms analyzed by hybridization probe assay. *Tohoku J Exp Med, 207*(4), 255-261.

Keating, B., Bansal, A. T., Walsh, S., Millman, J., Newman, J., Kidd, K., Budowle, B., Eisenberg, A., Donfack, J., Gasparini, P., Budimlija, Z., Henders, A. K., Chandrupatla, H., Duffy, D. L., Gordon, S. D., Hysi, P., Liu, F., Medland, S. E., Rubin, L., Martin, N. G., Spector ,T. D., Kayser, M., International Visible Trait Genetics (VisiGen) Consortium. (2012). First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip. *Int J Legal Med,* 127(3):559-72. doi: 10.1007/s00414-012-0788-1. Epub 2012 Nov 13

Khodjet-el-Khil, H., Fadhlaoui-Zid, K., Cherni, L., Phillips, C., Fondevila, M., Carracedo, A., & Ben Ammar-Elgaaied, A. (2011). Genetic analysis of the SNPforID 34-plex ancestry informative SNP panel in Tunisian and Libyan populations. *Forensic Sci Int Genet, 5*(3), e45-47. doi: 10.1016/j.fsigen.2010.07.007

Kidd, J. R., Friedlaender, F. R., Speed, W. C., Pakstis, A. J., De La Vega, F. M., & Kidd, K. K. (2011). Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investig Genet, 2*(1), 1. doi: 10.1186/2041-2223-2-1

Kidd, K. K., Kidd, J. R., Speed, W. C., Fang, R., Furtado, M. R., Hyland, F. C., & Pakstis, A. J. (2012). Expanding data and resources for forensic use of SNPs in individual identification. *Forensic Sci Int Genet, 6*(5), 646-652. doi: 10.1016/j.fsigen.2012.02.012

Kidd, K. K., Pakstis, A. J., Speed, W. C., Grigorenko, E. L., Kajuna, S. L., Karoma, N. J., Kungulilo, S., Kim, J.J., Lu, R.B., Odunsi, A., Okonofua, F., Parnas, J., Schulz, L.O., Zhukova, O.V., Kidd, J. R. (2006). Developing a SNP panel for forensic identification of individuals. *Forensic Sci Int, 164*(1), 20-32. doi: 10.1016/j.forsciint.2005.11.017

Kim, J. J., Han, B. G., Lee, H. I., Yoo, H. W., & Lee, J. K. (2010). Development of SNP-based human identification system. *Int J Legal Med, 124*(2), 125-131. doi: 10.1007/s00414-009-0389-9

Kis, Z., Zalan, A., Volgyi, A., Kozma, Z., Domjan, L., & Pamjav, H. (2012). Genome deletion and insertion polymorphisms (DIPs) in the Hungarian population. *Forensic Sci Int Genet, 6*(5), e125-126. doi: 10.1016/j.fsigen.2011.09.004

Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., Kittles, R., Alarcon-Riquelme, M. E., Gregersen, P. K., Belmont, J. W., De La Vega, F. M., Seldin, M. F. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat, 30*(1), 69-78. doi: 10.1002/humu.20822

LaRue, B. L., Ge, J., King, J. L., & Budowle, B. (2012). A validation study of the Qiagen Investigator DIPplex(R) kit; an INDEL-based assay for human identification. *Int J Legal Med, 126*(4), 533-540. doi: 10.1007/s00414-012-0667-9

Lee, H. Y., Park, M. J., Yoo, J. E., Chung, U., Han, G. R., & Shin, K. J. (2005). Selection of twenty-four highly informative SNP markers for human identification and paternity analysis in Koreans. *Forensic Sci Int, 148*(2-3), 107-112. doi: 10.1016/j.forsciint.2004.04.073

Li, C. T., Zhang, S. H., & Zhao, S. M. (2011). Genetic analysis of 30 InDel markers for forensic use in five different Chinese populations. *Genet Mol Res, 10*(2), 964-979. doi: 10.4238/vol10-2gmr1082

Lou, C., Cong, B., Li, S., Fu, L., Zhang, X., Feng, T., Su, S., Ma, C., Yu, F., Ye, J., Pei, L. (2011). A SNaPshot assay for genotyping 44 individual identification single nucleotide polymorphisms. *Electrophoresis, 32*(3-4), 368-378. doi: 10.1002/elps.201000426

Martin, P., Garcia, O., Heinrichs, B., Yurrebaso, I., Aguirre, A., & Alonso, A. (2013). Population genetic data of 30 autosomal indels in Central Spain and the Basque Country populations. *Forensic Sci Int Genet, 7*(2), e27-30. doi: 10.1016/j.fsigen.2012.10.003

Neuvonen, A. M., Palo, J. U., Hedman, M., & Sajantila, A. (2012). Discrimination power of Investigator DIPplex loci in Finnish and Somali populations.

Forensic Sci Int Genet, 6(4), e99-102. doi: 10.1016/j.fsigen.2011.09.005

Nievergelt, C. M., Maihofer, A. X., Shekhtman T., Libiger O., Wang X., Kidd K. K. & Kidd J. R. (2013). Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investigative Genetics* 4:13

Pakstis, A. J., Fang, R., Furtado, M. R., Kidd, J. R., & Kidd, K. K. (2012). Mini-haplotypes as lineage informative SNPs and ancestry inference SNPs. *Eur J Hum Genet, 20*(11), 1148-1154. doi: 10.1038/ejhg.2012.69

Pakstis, A. J., Speed, W. C., Fang, R., Hyland, F. C., Furtado, M. R., Kidd, J. R., & Kidd, K. K. (2010). SNPs for a universal individual identification panel. *Hum Genet, 127*(3), 315-324. doi: 10.1007/s00439-009-0771-1

Pakstis, A. J., Speed, W. C., Kidd, J. R., & Kidd, K. K. (2007). Candidate SNPs for a universal individual identification panel. *Hum Genet, 121*(3-4), 305-317. doi: 10.1007/s00439-007-0342-2

Pakstis AJ, Fang R, Furtado MR, Haigh E, Kidd JR, Kidd KK. (2013 (Manuscript submitted)). Validation of mini-haplotypes as valuable markers for familial identification and ancestry inference.

Pereira, R., Phillips, C., Pinto, N., Santos, C., dos Santos, S. E., Amorim, A., Carracedo, A., Gusmao, L. (2012). Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS One, 7*(1), e29684. doi: 10.1371/journal.pone.0029684

Phillips, C., Freire Aradas, A., Kriegel, A. K., Fondevila, M., Bulbul, O., Santos, C., Serrulla Rech, F., Perez Carceles, M .D., Carracedo, Á., Schneider, P.M., Lareu, M. V. (2013). Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Sci Int Genet, 7*(3), 359-366. doi: 10.1016/j.fsigen.2013.02.010

Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gomez-Tato, A., Alvarez-Dios, J., Alonso, A., Balnco-verea, A., Brion, M., Montesino, M., Carracedo, A.,Lareu, M. V. (2009). Ancestry analysis in the 11-M Madrid bomb attack investigation. *PLoS One, 4*(8), e6583. doi: 10.1371/journal.pone.0006583

Phillips, C., Salas, A., Sanchez, J. J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M.C., Ballard, D., Lareu, M.V., Carracedo, A; SNPforID Consortium. (2007). Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int*

Genet, 1(3-4), 273-280. doi: 10.1016/j.fsigen.2007.06.008

Pietrangeli, I., Ottaviani, E., Martone, C., Gabriele, L., Arcudi, G., Potenza, S., Spinella, A., Giardina, E., Novelli, G. (2010). Frequency assessment of 25 SNPs in five different populations. *Forensic Sci Int Genet, 4*(5), e131-133. doi: 10.1016/j.fsigen.2010.01.017

Porras, L., Phillips, C., Fondevila, M., Beltran, L., Ortiz, T., Rondon, F., Barreto, G., Lareu, M.V., Henao, J., Carracedo, A. (2009). Genetic variability of the SNPforID 52-plex identification-SNP panel in Central West Colombia. *Forensic Sci Int Genet, 4*(1), e9-10. doi: 10.1016/j.fsigen.2008.12.003

Poulsen, L., Borsting, C., Tomas, C., Gonzalez-Andrade, F., Lopez-Pulles, R., Gonzalez-Solorzano, J., & Morling, N. (2011). Typing of Amerindian Kichwas and Mestizos from Ecuador with the SNPforID multiplex. *Forensic Sci Int Genet, 5*(4), e105-107. doi: 10.1016/j.fsigen.2011.03.006

Rajeevan, H., Cheung, K. H., Gadagkar, R., Stein, S., Soundararajan, U., Kidd, J. R., Pakstis, A. J., Miller, P.L., Kidd, K. K. (2005). ALFRED: an allele frequency database for microevolutionary studies. *Evol Bioinform Online, 1*, 1-10.

Rajeevan, H., Osier, M. V., Cheung, K. H., Deng, H., Druskin, L., Heinzen, R., Kidd, J. R., Stein, S., Pakstis, A. J., Tosches ,N. P., Yeh, C. C., Miller, P. L., Kidd, K. K. (2003). ALFRED: the ALelle FREquency Database. Update. *Nucleic Acids Res, 31*(1), 270-271.

Rajeevan, H., Soundararajan, U., Kidd, J. R., Pakstis, A. J., & Kidd, K. K. (2012). ALFRED: an allele frequency resource for research and teaching. *Nucleic Acids Res, 40*(Database issue), D1010-1015. doi: 10.1093/nar/gkr924

Rajeevan, H., Soundararajan, U., Pakstis, A. J., & Kidd, K. K. (2012). Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb. *Investig Genet, 3*(1), 18. doi: 10.1186/2041-2223-3-18

Ruiz, Y., Chiurillo, M. A., Borjas, L., Phillips, C., Lareu, M. V., & Carracedo, A. (2012). Analysis of the SNPforID 52-plex markers in four Native American populations from Venezuela. *Forensic Sci Int Genet, 6*(5), e142-145. doi: 10.1016/j.fsigen.2012.02.007

Ruiz, Y., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., Casares de Cal, M., Cruz, R., Maroñas, O., Söchtig, J., Fondevila, M., Rodriguez-Cid, M. J., Carracedo, A., Lareu, M. V. (2013). Further development of forensic eye

color predictive tests. *Forensic Sci Int Genet, 7*(1), 28-40. doi: 10.1016/j.fsigen.2012.05.009

Sanchez, J. J., Phillips, C., Borsting, C., Balogh, K., Bogus, M., Fondevila, M., Harrison, C.D., Musgrave-Brown, E., Salas, A., Syndercombe-Court, D., Schneider, P.M., Carracedo, A., Morling, N. (2006). A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis, 27*(9), 1713-1724. doi: 10.1002/elps.200500671

Sharafi Farzad, M., Tomas, C., Borsting, C., Zeinali, Z., Malekdoost, M., Zeinali, S., & Morling, N. (2013) Analysis of 49 autosomal SNPs in three ethnic groups from Iran: Persians, Lurs and Kurds. *Forensic Sci Int Genet, 7*(4), 471-473. doi: 10.1016/j.fsigen.2013.04.001

Silva, M. C., Zuccherato, L. W., Soares-Souza, G. B., Vieira, Z. M., Cabrera, L., Herrera, P., Tarazona-Santos, E. (2010). Development of two multiplex mini-sequencing panels of ancestry informative SNPs for studies in Latin Americans: an application to populations of the State of Minas Gerais (Brazil). *Genet Mol Res, 9*(4), 2069-2085. doi: 10.4238/vol9-4gmr911

Tomas, C., Diez, I. E., Moncada, E., Borsting, C., & Morling, N. (2013). Analysis of 49 autosomal SNPs in an Iraqi population. *Forensic Sci Int Genet, 7*(1), 198-199. doi: 10.1016/j.fsigen.2012.05.004

Tomas, C., Stangegaard, M., Borsting, C., Hansen, A. J., & Morling, N. (2008). Typing of 48 autosomal SNPs and amelogenin with GenPlex SNP genotyping system in forensic genetics. *Forensic Sci Int Genet, 3*(1), 1-6. doi: 10.1016/j.fsigen.2008.06.007

Turrina S, Filippini G, Leo DD. (2011). Forensic evaluation of the Investigator DIPplex typing system. Forensic Science International: Genetics Supplement Series. *Forensic Science International: Genetics Supplement Series, 3*(1), e331-e332.

Vallone, P. M., Decker, A. E., & Butler, J. M. (2005). Allele frequencies for 70 autosomal SNP loci with U.S. Caucasian, African-American, and Hispanic samples. *Forensic Sci Int, 149*(2-3), 279-286. doi: 10.1016/j.forsciint.2004.07.014

Walsh, S., Liu, F., Ballantyne, K. N., van Oven, M., Lao, O., & Kayser, M. (2011). IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci Int Genet,*

5(3), 170-180. doi: 10.1016/j.fsigen.2010.02.004

Walsh, S., Wollstein, A., Liu, F., Chakravarthy, U., Rahu, M., Seland, J. H., Soubrane, G., Tomazzoli, L., Topouzis, F., Vingerling, J. R., Vioque, J., Fletcher, A. E., Ballantyne, K. N., Kayser, M. (2012). DNA-based eye colour prediction across Europe with the IrisPlex system. *Forensic Sci Int Genet, 6*(3), 330-340. doi: 10.1016/j.fsigen.2011.07.009

Watkins, W. S., Rogers, A. R., Ostler, C. T., Wooding, S., Bamshad, M. J., Brassington, A. M., Carroll, M.L., Nguyen, S. V., Walker, J. A., Prasad, B. V., Reddy, P. G., Das, P. K., Batzer, M. A., Jorde, L. B. (2003). Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res, 13*(7), 1607-1618. doi: 10.1101/gr.894603

Zidkova, A., Horinek, A., Kebrdlova, V., & Korabecna, M. (2013). Application of the new insertion-deletion polymorphism kit for forensic identification and parentage testing on the Czech population. *Int J Legal Med*, 127(1), 7-10. doi: 10.1007/s00414-011-0649-3

VI) Dissemination of Results

The web site itself is the primary result and method of dissemination. A publication has been one method of calling attention to it:

Rajeevan, H., Soundararajan, U., Pakstis, A. J., & Kidd, K. K. (2012). Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb. *Investig Genet, 3*(1), 18. doi: 10.1186/2041-2223-3-18.

Scientific meetings where this work was presented:

1. Poster presentation at NIJ annual meeting, Arlington, Virginia , 2011(announcement of the database to the forensic community with the pilot implementation)

Title: Developing SNP panels for ancestry identification useful in forensic investigations

Authors: Kidd KK, Kidd JR, Pakstis AJ, Speed WC, Donnelly M

2. Poster presentation at the NIJ annual meeting, Arlington, Virginia, 2012

Title: FROG-kb: Forensic Resource/Reference on Genetics- knowledgebase

Authors: Rajeevan H, Soundararajan U, Kidd KK

Invited seminars variously entitled, including “Better SNPs for Better Forensics”, were presented at five forensic venues in China on three separate invited trips in 2011, 2012, and 2013.

K. Kidd is scheduled to give several invited talks taking place in July through September, 2013, in forensic settings. He will be mentioning ALFRED and FROG-kb in addition to his other research.

VII) Appendices

7.1. Figures

Figure 1a: ALFRED database structure

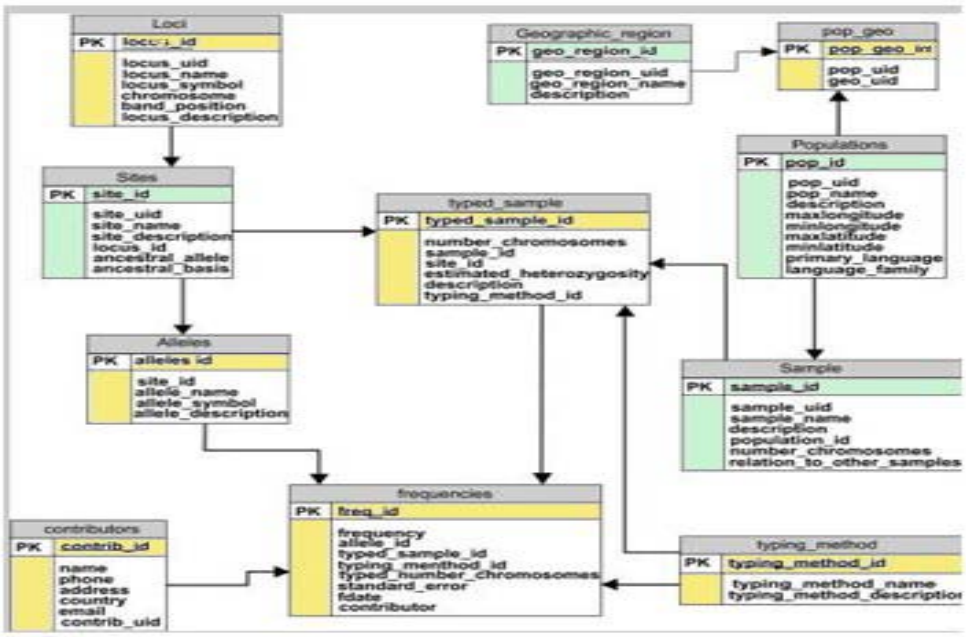


Figure 1b: Tables and relationships from ALFRED database that are used in FROG-kb

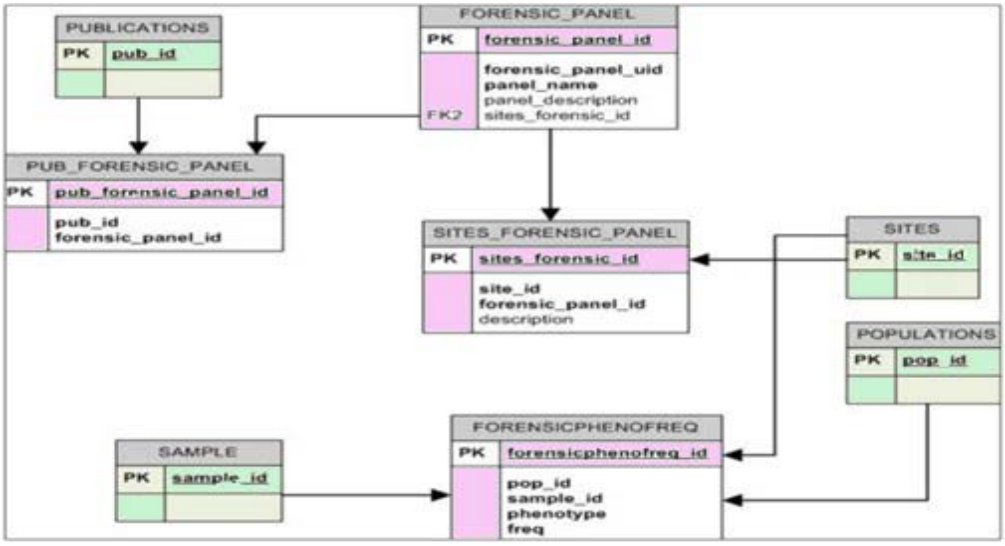


Figure 1a. & 1b: Yellow shaded tables are used only for the ALFRED application, pink shaded tables only for the FROG-kb application, while green shaded tables are shared by ALFRED and FROG-kb.

Figure 2: FROG-kb home page

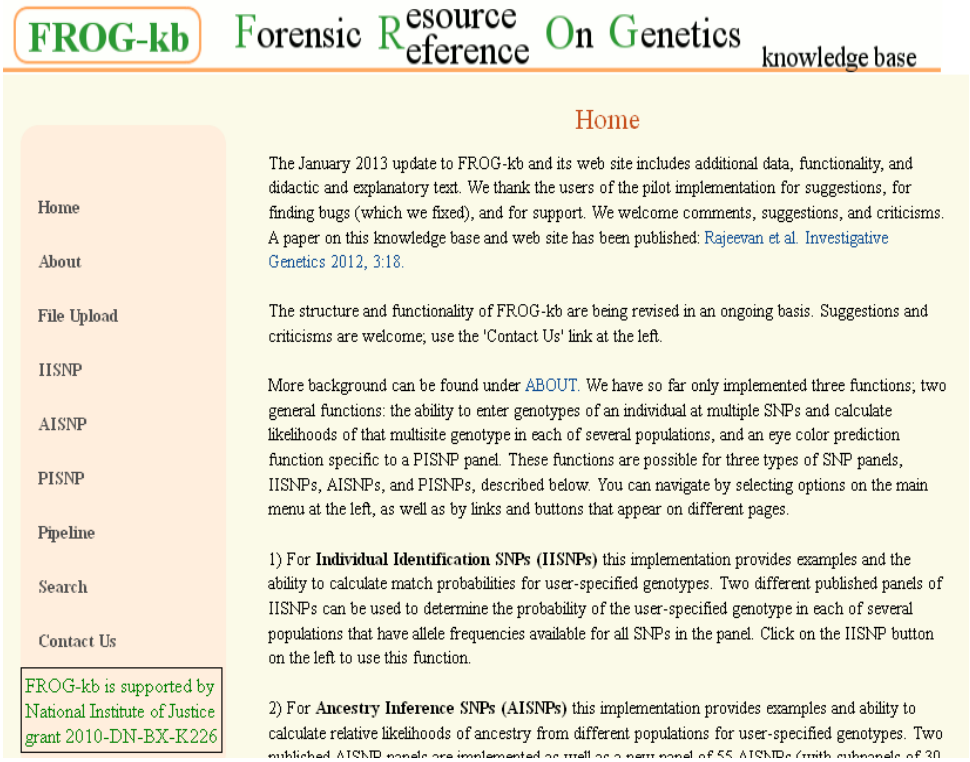


Figure 3: Selection of SNP panel (shown for AISNP category)

Figure 4: Functionalities in FROG-kb (shown with the selection of one AI SNP panel)

Figure 5: Data entry options

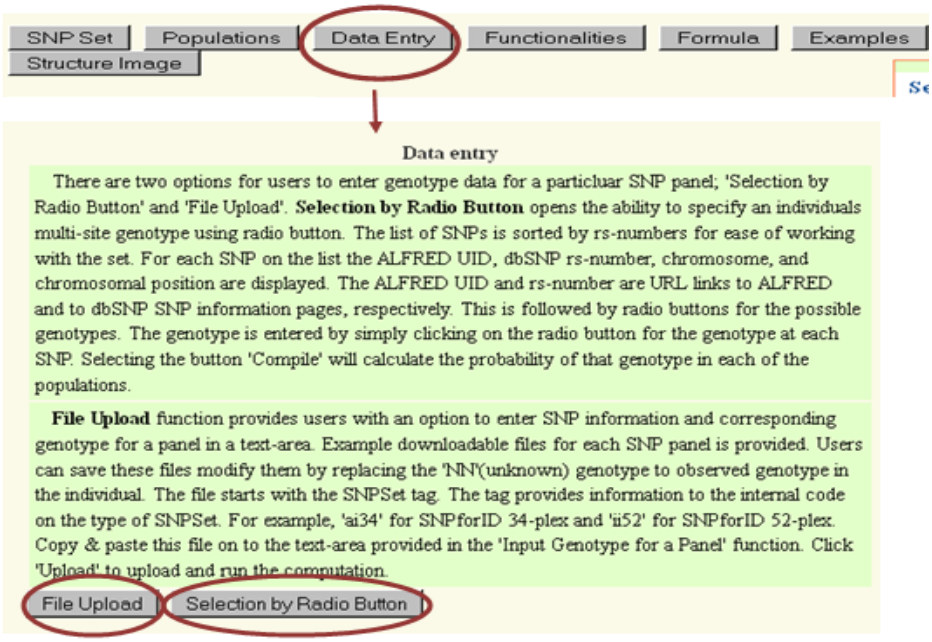


Figure 5a. Data entry selection by radio button

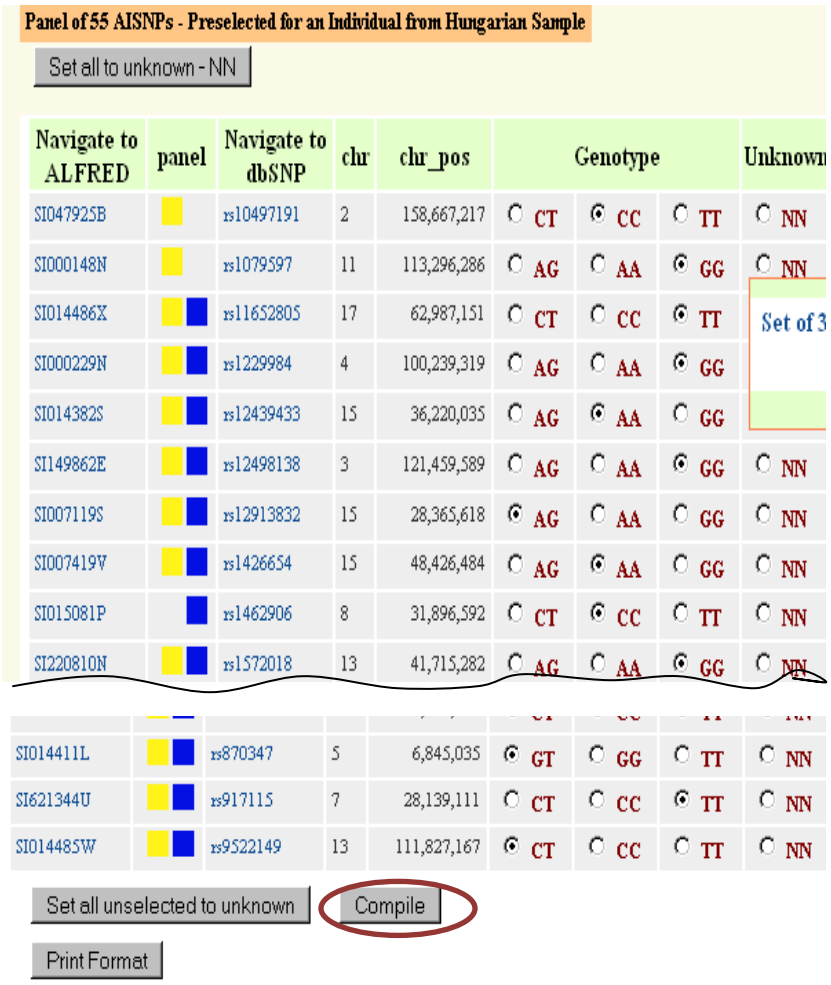


Figure 5b: Data entry option by file upload

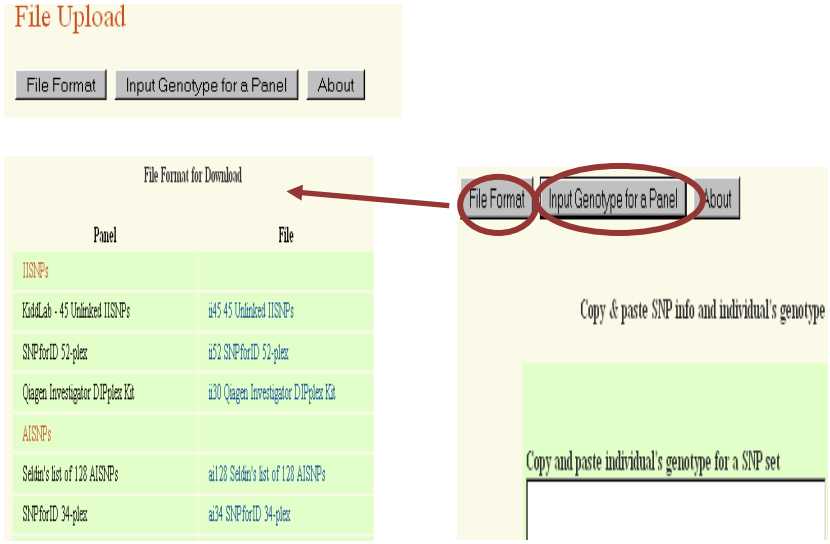
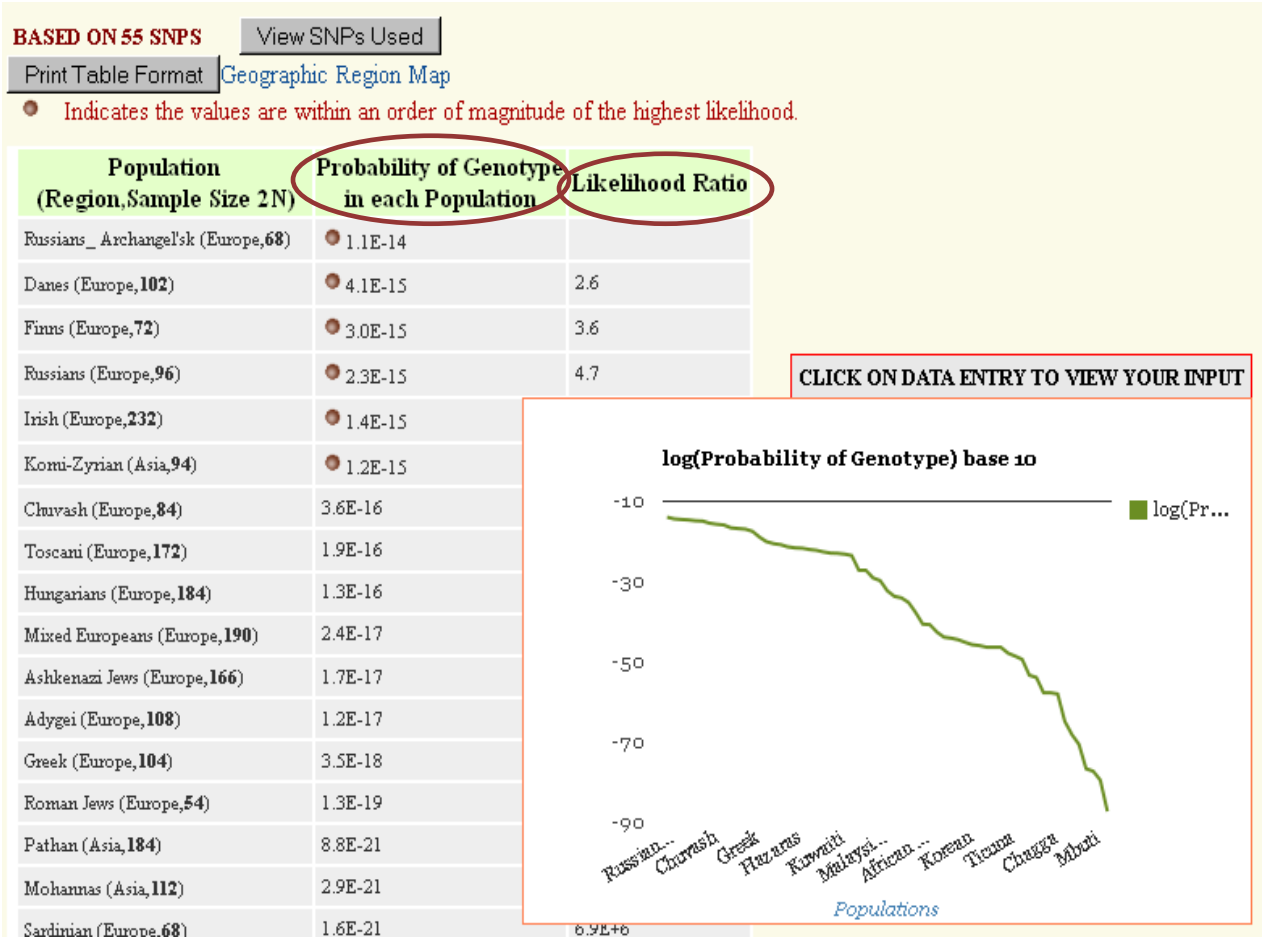


Figure 6. Output results page



7.2. Usage Summary January 2013 to June 2013:

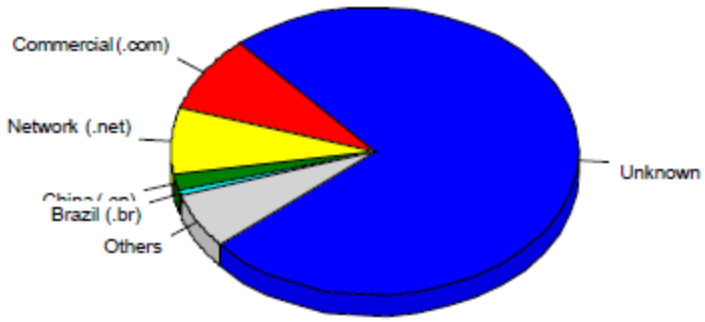
Visitor - The number of visitors is determined by the IP addresses. If a request from an IP address comes after some time (timeout) since the last request from this IP, it is considered to belong to a different visitor. The timeout is set to 30 minutes by default.

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

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

Total Hits	56,053
Average Hits per Day	352
Average Hits per visitor	16.11
Total Visitors	3,479
Average Visitors per Day	21
Total Unique IPs	1,110

Top-Level Domains



Top-Level Domains

	Domain	Hits	Visitors	% of Total Visitors	Bandwidth (KB)
1	Unknown	28,393	2,616	75.19%	294,899
2	Commercial (.com)	17,860	297	8.54%	93,444
3	Network (.net)	4,692	262	7.53%	633,112
4	China (.cn)	456	65	1.87%	19,298
5	Brazil (.br)	451	19	0.55%	4,461
6	Russian Federation (.ru)	55	18	0.52%	3,751
7	Educational (.edu)	398	18	0.52%	16,907
8	Government (.gov)	882	16	0.46%	35,539
9	Germany (.de)	62	15	0.43%	2,245
10	Singapore (.sg)	26	13	0.37%	1,774
11	Spain (.es)	194	12	0.34%	15,740
12	Italy (.it)	793	11	0.32%	5,824
13	Non-profit Organization (.org)	74	10	0.29%	5,720
14	Australia (.au)	193	9	0.26%	13,334
15	Netherlands (.nl)	209	9	0.26%	13,862

The ‘Unknown’ here are unresolved IP addresses. This could include educational and research institutions.

7.3. Advisory Panel Members for FROG-kb Project in 2013

Bruce Budowle, Ph.D.
Department of Forensic and Investigative Genetics
University of North Texas Health Science Center
Fort Worth, Texas

Arthur J. Eisenberg, Ph.D.
Department of Forensic and Investigative Genetics
University of North Texas Health Science Center
Fort Worth, Texas

Joseph Donfack, Ph.D.
Counterterrorism and Forensic Science Research Unit
FBI Laboratory Division
2501 Investigation Parkway
Quantico, Virginia 22135 USA

Daniele S. Podini, Ph.D.
Department of Forensic Sciences
The George Washington University
Washington, D.C.

Investigative Genetics

RESEARCH Open Access

Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb

Haseena Rajeevan[1,2], Usha Soundararajan[1], Andrew J Pakstis[1] and Kenneth K Kidd[1*]

Abstract

Background: Online tools and databases based on multi-allelic short tandem repeat polymorphisms (STRPs) are actively used in forensic teaching, research, and investigations. The Fst value of each CODIS marker tends to be low across the populations of the world and most populations typically have all the common STRP alleles present diminishing the ability of these systems to discriminate ethnicity. Recently, considerable research is being conducted on single nucleotide polymorphisms (SNPs) to be considered for human identification and description. However, online tools and databases that can be used for forensic research and investigation are limited.

Methods: The back end DBMS (Database Management System) for FROG-kb is Oracle version 10. The front end is implemented with specific code using technologies such as Java, Java Servlet, JSP, JQuery, and GoogleCharts.

Results: We present an open access web application, FROG-kb (Forensic Research/Reference on Genetics-knowledge base, http://frog.med.yale.edu), that is useful for teaching and research relevant to forensics and can serve as a tool facilitating forensic practice. The underlying data for FROG-kb are provided by the already extensively used and referenced ALlele FREquency Database, ALFRED (http://alfred.med.yale.edu). In addition to displaying data in an organized manner, computational tools that use the underlying allele frequencies with user-provided data are implemented in FROG-kb. These tools are organized by the different published SNP/marker panels available. This web tool currently has implemented general functions possible for two types of SNP panels, individual identification and ancestry inference, and a prediction function specific to a phenotype informative panel for eye color.

Conclusion: The current online version of FROG-kb already provides new and useful functionality. We expect FROG-kb to grow and expand in capabilities and welcome input from the forensic community in identifying datasets and functionalities that will be most helpful and useful. Thus, the structure and functionality of FROG-kb will be revised in an ongoing process of improvement. This paper describes the state as of early June 2012.

Keywords: Knowledge base, SNP, InDels, Forensics, Individual identification, Ancestry inference markers, Lineage informative markers

Background

It is considerably more than a decade since the forensic community settled on a set of short tandem repeat (STR) polymorphisms (hence also STRP) for human identity testing [1]. These markers are multi-allelic and are excellent for individual matching of suspect and crime scene DNA. While 13 STRs form the core of the FBI Laboratory's CODIS (Combined DNA Index System), the 10 core loci used in the UK and much of Europe consist of eight loci that overlap with CODIS plus seven additional markers that include the five new European Standard Set (ESS) [2]. Discussions on the best options on expanding the core sets of loci are underway [3-5]. Development of reliable commercial multiplex kits tailored specifically for these sets of markers has led to large offender databases and large amounts of allele frequency data accumulated for these markers on a wide range of populations around the world. Online tools and databases have followed to allow users to reference and predict population affiliations: Canadian Random Match Calculator [6] (http://www.csfs.ca/pplus/profiler.htm),

* Correspondence: Kenneth.Kidd@yale.edu
[1]Department of Genetics, Yale University School of Medicine, 333 Cedar Street, P.O.Box 208005, New Haven, CT 06520-8005, USA
Full list of author information is available at the end of the article

European Network of Forensic Science Institute's DNA WG STR Population Database (http://www.str-base.org/index.php), STRBase (http://www.cstl.nist.gov/strbase/) [7], PopAffiliator (http://cracs.fc.up.pt/popaffiliator/) [8], and pop.STR (http://spsmart.cesga.es/popstr.php) [9]. One argument for continuing use of these tools and marker panels is the large number of individual offender DNA profiles in databases allowing 'cold hits', that is, identification of the criminal based on a database match to crime scene DNA. The extensive allele frequency data that have been accumulated over the years in large public databases also allow population-specific estimates of the probability of a random match of two unrelated individuals. However, it is exactly the high level of polymorphism in almost all populations that limits the ability of these markers to determine ancestry of an individual. The large numbers of alleles and high heterozygosity relate to the high mutation rates of these loci; this also means that matching of STRP alleles is matching by state and not of alleles that are identical by descent.

Considering the ease, accuracy, and efficiency in typing single nucleotide polymorphisms (SNPs) and their essentially zero rate of recurrent mutation when compared with STRPs [10-12], SNPs have the potential to be considered for human identification and description in forensic, biomedical, association, as well as epidemiological studies [13-15]. Insertion-deletion polymorphisms (InDels) have most of the desirable characteristics of SNPs and panels have been proposed and have begun to be used in forensics both for individual identification and for ancestry inference [16-18]. However, considerable research is required to establish a reliable set of markers (SNPs or InDels) containing sufficient numbers of markers to provide excellent discriminatory power comparable to or exceeding that of STR markers. Not only do multiple populations need to be studied to identify the best markers but interpretation of results in any application needs the reference allele frequencies in multiple populations. The discriminatory power for individual identification will be population specific and ancestry inference will only be as good as the set of reference populations. Online web tools that demonstrate classification algorithms are being developed for SNP sets as well, 'The Snipper' app suite (http://mathgene.usc.es/snipper/) for three ancestry informative AISNP sets (34, 32, and 77 markers) being one of them [19]. Currently, the main limitations of these tools are the number of SNP sets and the range of population data available for computation.

In this paper we introduce FROG-kb (Forensic Resource/Reference on Genetics knowledge base), an open access web tool that allows viewing and retrieval of data as well as calculation of statistics on several forensically relevant SNP sets. FROG-kb's user interface is versatile in its functionality and comprehensive in the population data available for many SNP sets. The overall goal of FROG-kb is to make allele frequency data for SNPs and other genetic polymorphisms more accessible and useful in a forensic setting. Ancestry, Individual, and Phenotype Informative (AISNP, IISNP, and PISNP) panels [19-24] studied and published from the host lab and elsewhere are currently available in FROG-kb. Each of these panels exists with supporting population data. URL links exist to ALFRED for more details and allele frequency data tables for specific populations not only for the panels themselves but also for each SNP in each panel. Additional information in the underlying ALFRED database and the curated links into other databases make FROG-kb a reference source as well. As frequency data on a population not in the original publication become comprehensively available for the full SNP panel, that population is included in the computations. As new forensically relevant panels of SNPs with meaningful population data are published, they will be systematically added to FROG-kb. The structure and types of contents in FROG-kb are described below followed by descriptions of current functionality with examples.

Methods

Database structure

The underlying data for the FROG-kb implementation are from the allele frequency database ALFRED (http://alfred.med.yale.edu) [25,26]. ALFRED is a relational database and aspects of the structure and relationships of the tables exist in the above mentioned publications. Additional tables accommodate information essential for a human identity testing application. Figure 1 gives the database tables and relationships incorporating only the supplementary tables relevant to FROG-kb. The logic supporting the design of the additional tables and relationships follows. (FORENSIC_PANEL). Every panel is linked to at least one publication (PUB_FORENSIC_PANEL). Such links are clearly identified when the underlying data are unpublished to document the source of the data. The marker phenotype (equivalent to 'genotype' based on multiple unavoidable assumptions) frequency for each 'site - population sample' combination is pre-calculated and saved in 'FORENSICPHENOFREQ'. Since these population-marker frequencies do not change for existing data, pre-calculation of the phenotype summary is reasonable and expedites the involved case-specific computations. While all allele frequency data required for running the computations in FROG-kb were already in the ALFRED database, the new tables provide the framework for displaying information related to the different forensically relevant SNP sets in an efficient and user-friendly manner.

Implementation

The database for FROG-kb is implemented using Oracle version 10 on one of Yale's institutional database servers

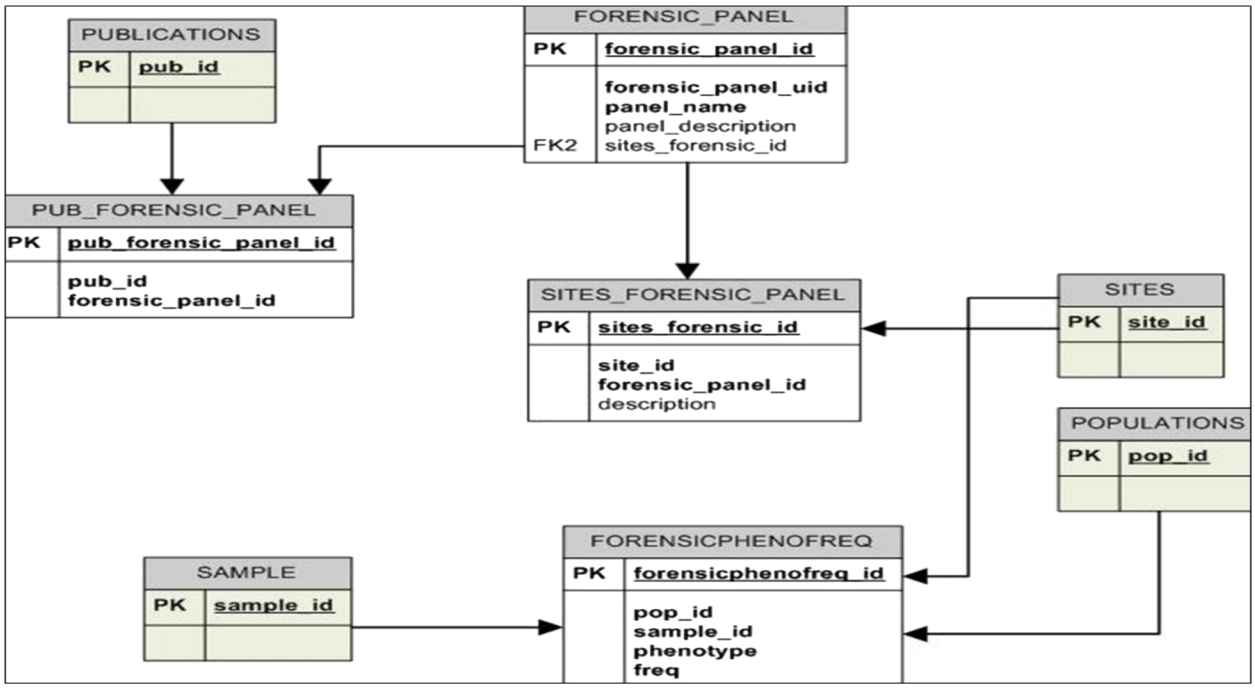


Figure 1 Additional tables added to ALFRED for FROG-kb. The full relational schema of ALFRED is described in [26]; these tables and their interconnections illustrate the additions needed for the functionality of FROG-kb.

where it is maintained. The web front end is built using web developing technologies such as Java, Java Servlet, JSP, JQuery, and GoogleCharts. Almost all of the client-code utilizes JQuery, and the server implementation is in Java. The currently deployed version of FROG-kb has been tested on both PC and Mac, using many different browsers: Mozilla Firefox11.0, Internet Explorer 8.0, Safari 5.1.4, and Google Chrome 17.0 on PC, and Firefox11.0, Internet Explorer 5.0, Safari 5.0 on Mac. We are using Tomcat as our web server which runs on a Windows XP machine.

Functionalities

The user interface layout of FROG-kb is designed to reflect the organization of the contents and functionality, as well as ease of use. Every set of pages relative to a function on FROG-kb originates from a tab on the 'Main Menu' (Figure 2A) that appears on the left-hand side of every page. The 'Home Page' gives a brief summary of the functions available in FROG-kb and what can be expected soon. The procession through the interface, explained below, is summarized graphically under 'Pipeline'.

Users can navigate into the functions relevant to each type of SNP panel by selecting the appropriate tab: 'IISNP' for Individual Identification SNPs, 'AISNP' for Ancestry Inference SNPs and 'PISNP' for Phenotype Informative SNPs. Following selection of a particular SNP panel category (IISNP, AISNP, or PISNP), there are multiple published panels listed. The citation information for each of the panels, a 'Go' button to navigate into the selected panel, and a 'Detailed Overview of SNPs' link to navigate into ALFRED are provided for each (Figure 2B). The link to ALFRED opens the 'SNP Sets' page within ALFRED into a new browser window. The SNP Set module in ALFRED has multiple functions, including the ability to see for each SNP a pie chart on Google Maps of frequencies for all populations with data.

Several options are possible after entering a SNP panel page. The functions related to the selected panel are available by selecting the appropriate buttons at the top of the page (Figure 2C). The option SNP Set provides the list of SNPs in the panel. The list includes the dbSNP rs-number of each SNP with an active link to the corresponding dbSNP record for molecular characterization of the SNP. The Populations button provides the list of populations for which comparable calculations can be made. This is the set of populations for which all SNPs in the set have allele frequency data. Conversely, many populations have data on additional SNPs; those SNPs are not included for the calculations. Within the SNP Set functionality in ALFRED additional populations may have data for some, but not all SNPs; those populations are not included in the calculations. Each population name within FROG-kb is an active link to information on the population stored within ALFRED; that page will open in a new browser window.

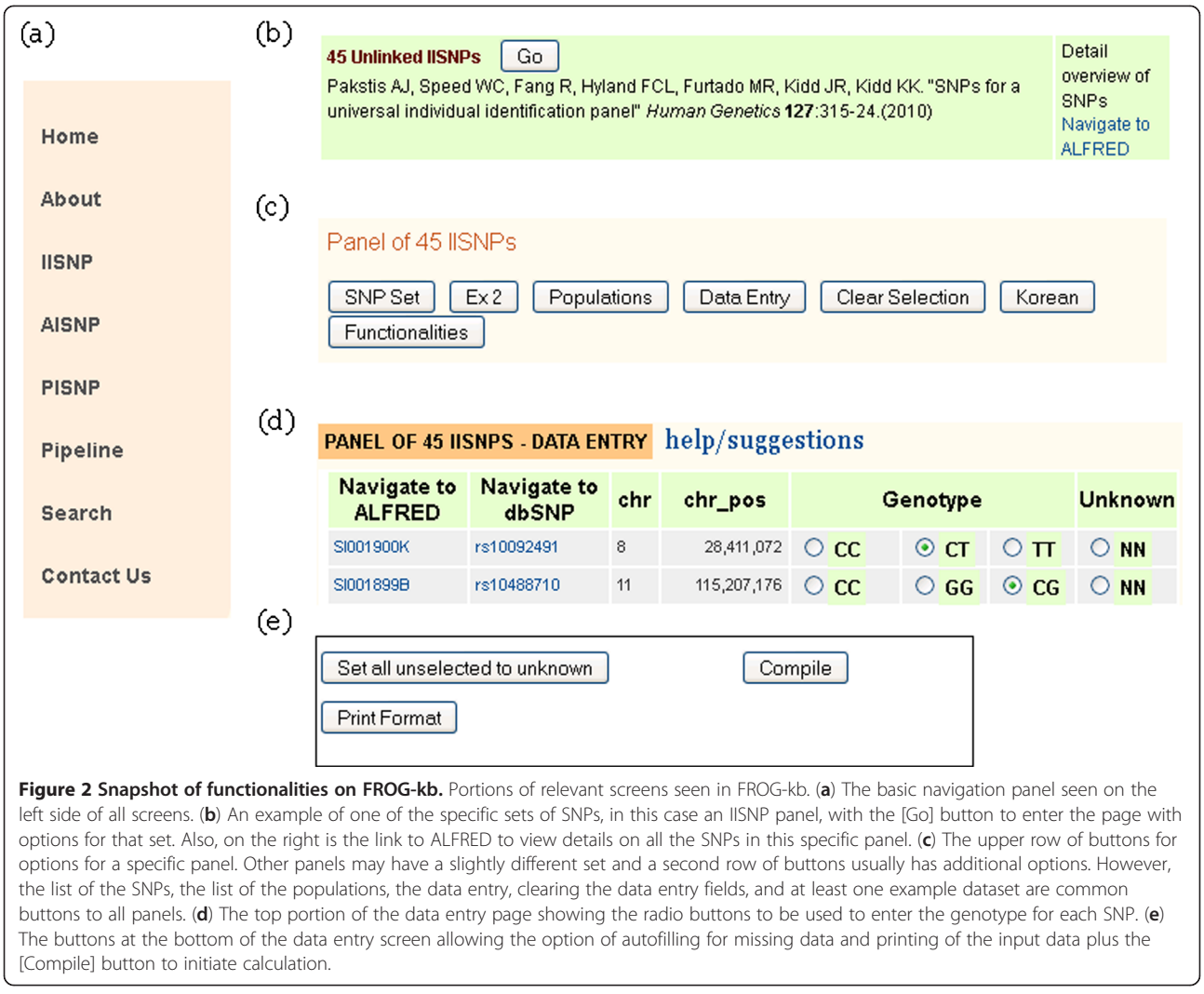


Figure 2 Snapshot of functionalities on FROG-kb. Portions of relevant screens seen in FROG-kb. (**a**) The basic navigation panel seen on the left side of all screens. (**b**) An example of one of the specific sets of SNPs, in this case an IISNP panel, with the [Go] button to enter the page with options for that set. Also, on the right is the link to ALFRED to view details on all the SNPs in this specific panel. (**c**) The upper row of buttons for options for a specific panel. Other panels may have a slightly different set and a second row of buttons usually has additional options. However, the list of the SNPs, the list of the populations, the data entry, clearing the data entry fields, and at least one example dataset are common buttons to all panels. (**d**) The top portion of the data entry page showing the radio buttons to be used to enter the genotype for each SNP. (**e**) The buttons at the bottom of the data entry screen allowing the option of autofilling for missing data and printing of the input data plus the [Compile] button to initiate calculation.

The geographic region of each population is included. A world map in which regions are divided on an arbitrary but convenient basis is available from the link 'Geographic Region Map'. Example options are also accessible using Ex 1 or similar buttons. These are static screen shots to provide examples.

The most significant interactive function is accessed via Data Entry (Figure 2D) that opens the ability to specify an individual's multi-site genotype (strictly, phenotype) and then calculate the probability of that genotype in each of the populations. The list of SNPs is sorted by rs-numbers for ease of working with the set. For each SNP on the list the ALFRED UID, dbSNP rs-number, chromosome, and chromosomal position are displayed. The ALFRED UID and rs-number are URL links to ALFRED and to dbSNP SNP information pages, respectively. This is followed by radio buttons for the possible genotypes. The genotype is entered by simply clicking on the radio button for the genotype at each SNP. An obvious assumption is that there is no allele drop out, that is, that a typing result (phenotype) with only one allele detected is really a homozygote. A radio button labeled 'NN' is provided for missing data for each SNP, but it is not necessary to click on the 'NN' for missing data. For large SNP sets, if the user's SNP set is also ordered by rs-number in a spreadsheet, selecting the appropriate genotype radio-button should be relatively effortless. (We are aware this input can be tedious; more user friendly options are under development.) At the bottom of the list are three buttons (Figure 2E): Set all unselected to unknown, Print Format, and Compile. The Print Format will generate a condensed version of the input data that can be printed as a permanent record of the input data. The information in the pop-up window can also be copied and pasted into a text editor which in turn can be opened in an Excel spreadsheet. The Compile will initiate calculation and display the results. If there are SNPs with no selection, a warning will be sent with the missing data rows highlighted for easy detection and the option exists to examine

which SNPs have no entry and to either enter a genotype or use the Set all unselected to unknown to fill those with 'NN'. Afterward, it is necessary to click on Compile again.

The calculations are essentially identical for all the IISNP and AISNP panels, just different loci (SNPs) and different interpretations are involved. All SNPs are considered statistically independent at the population level and so the probability of the input multisite genotype is simply the product of the probabilities of the genotypes of the individual loci--the 'product rule' in forensics. This calculation is done separately for each population using the allele frequencies estimated for that population and assuming Hardy-Weinberg ratios to estimate the population-specific genotype probabilities. The calculation uses all loci in the panel for which a genotype was entered. A missing genotype, NN in the input, means that locus is skipped for all populations. Only populations with data for all SNPs in the panel are considered in the calculation. Thus, the resulting probabilities/likelihoods that are displayed are based on the same set of loci for all populations.

Special consideration is needed for those situations in which one allele is not observed in a population and hence the observed allele frequency is zero and two genotypes are strictly estimated to be zero. This is especially important for AISNPs since many of the loci are fixed in some populations, but even a SNP in an IISNP panel may not be seen in an isolated population. Given the sample sizes involved, very low frequencies of the 'missing' allele cannot be excluded and using a value of zero in the calculations would be incorrect. The approach used is to simply use a very small allele frequency instead of the zero that is the allele frequency present in ALFRED. If one assumes a heterozygote might be seen in the next individual sampled from the population, the allele frequency would then be $1/(2n+2)$ where n is the original sample size in which the allele was not seen. Other approaches to the problem can exist. Because of different finite sample sizes and the unavoidable possibility of typing error, an absolute small frequency could be used. Another alternative is to simply not include that locus in any calculations involving a population in which one allele is not seen in the reference sample. This last completely circumvents a fixed locus having an undue influence on population rankings for AISNPs, but also overcompensates when those alleles are seen in the focus genotype. Those and other options can be considered by the community for future implementation as alternatives.

The time required to compute the probability of the genotype in each population depends on the number of SNPs in the panel and the number of populations for which the calculations are performed. The results page has a table with the probability values against each population name and a graph of $\log_{10}$ (Probability of Genotype) displayed side-by-side. The number of SNPs used in the computation is given. The button View SNPs Used gives the list of SNPs the computation was based on. Print Table Format generates a printable format of the result. The line graph of $\log_{10}$ (Probability of Genotype) is drawn utilizing the Google-Chart tools. The population name and the corresponding value are displayed when the mouse is hovered over a point (Figure 3). The geographic region displayed adjacent to a population name can be verified using the image at Geographic Region Map.

There are also buttons with population names, for example Hungarian or Korean that will open a pre-entered data entry page for one individual from the specified population. These can be used in an education mode to explore the dependence of the results on particular SNPs and on the total number of SNPs with data.

The above mentioned functionalities are common to all of the IISNP and AISNP panels. The interpretations of the values calculated differ between the IISNP and AISNP panels (see below). We note that, since the calculations are based on only the SNPs with specified genotypes, it is not necessary that an entire panel be genotyped. Indeed, if by happenstance a user has genotyped SNPs that are in a panel, those SNPs will yield a valid result though there may be little discrimination if only a few SNPs are used.

The single PISNP panel differs in that calculations are not dependent on population but only on the genotype entered. The Irisplex panel for eye color prediction under PISNP has the SNP Set, Data Entry and four pre-entered data entry pages. The eye color prediction computation uses the formula provided in the publication [21] and is given on the web site.

Results and Discussion

Interpretation of the calculations

For both the IISNP and AISNP input genotypes the programs currently calculate for each specific population the probability of that multi-locus genotype as the simple product across loci of the frequencies of each specific genotype in the specific population. That simple product is the one displayed for the population in the rank-ordered list and in the graph. The interpretation is different for the two types of SNP panels.

For IISNP panels the 'Probability of Genotype' can be interpreted as the match probability, that is, the probability of finding another unrelated individual in the population with the same multilocus genotype. One assumption is infinite population size with no adjustment for finite size or the statistical uncertainty of the allele frequency estimates for each population other than the 'zero allele frequency' correction noted above. Another assumption is of statistically independent loci, that is, no linkage disequilibrium. This second assumption has been tested, and satisfied, by analyses in the studies that

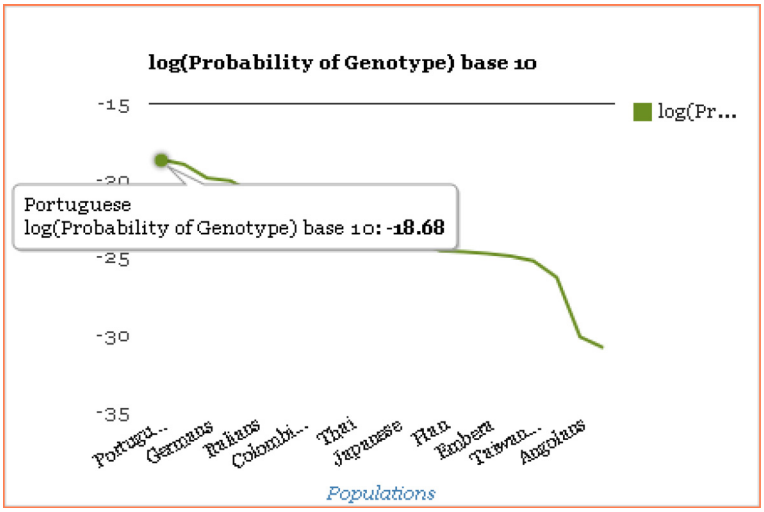


Figure 3 Graph of $\log_{10}$ (Probability of Genotype). An example of the plot of results for one individual input genotype with the populations ordered by the log10 probabilities from largest to smallest. This is part of the results display common to all IISNP and AISNP panel calculations. One can mouse over the graph to see the specific population and its value, as illustrated for Portuguese in this example. Not shown here is the list of those populations with the values also displayed on the same page. The option exists to print the list.

developed and published the panels. The population-specific values are given in the table ordered from highest to lowest probability, but is easier to visualize in the figure as the orders of magnitude showing how sensitive the match probability estimates are to the population from which the DNA sample was obtained. Of particular note is that the top population gives the highest match probability globally, to the extent the populations represent the global variation. Note that these values are acceptable estimates for whatever data were used. If there were many loci with missing data, the probabilities will be less definitive but will still be the best estimates for the loci used.

For AISNP panels the same calculation is made, simple multiplication of the probability of the population-specific genotypes at the several loci. For AISNPs, however, the interpretation of the probability of the genotype given a population can be considered as proportional to the likelihood of the population given the genotype. As likelihoods, the absolute values have no strict interpretation, only the relative values are interpretable. Thus, the rank orders are meaningful with populations having larger probabilities being more likely origins of the individual DNA profile than populations with smaller probabilities. The important question with respect to ancestry inference is how different the likelihoods are. One rough rule of thumb that could be applied is that likelihoods different by less than an order of magnitude are not significantly different. Thus, while the population with the highest likelihood is the most likely origin of the input genotype, it is not necessarily the correct origin and those with similar, albeit lower likelihoods, cannot be excluded. Indeed, even those more than an order of magnitude less likely are not 'excluded', just much less likely.

The likelihoods calculated are the maximum likelihood estimates because they are based on the maximum likelihood estimates for allele frequencies for independent loci with the necessary assumption of Hardy-Weinberg proportions for the genotypes, albeit with a correction for sites with a zero allele frequency. However, those allele frequency estimates have associated standard errors that vary inversely as the sample sizes. Thus, there are greater uncertainties for likelihoods calculated for populations with small sample sizes. Hence, rankings among populations with very similar likelihoods could be different when one or more is based on a very small sample size. The statistical issues with determining significance of an estimated rank order of populations, taking into account all the individual components of uncertainty, are not simple and we currently have no rigorous statistical method identified and implemented. We are not aware of any of the existing AISNP estimation procedures that has solved this problem. We emphasize that users must exercise judgment by recognizing the inherent uncertainty and considering the differences in sample sizes (displayed with each sample name).

Current panels and examples

Two different Individual Identification SNP panels (IISNPs) provide examples and the ability to calculate match probabilities for user-specified genotypes in each of many populations that have allele frequencies available for all SNPs in the panel. The two panels are (A) 45

unlinked IISNPs from the host lab [20] and (B) the SNPforID 52-plex [21]. Similarly, for Ancestry Inference SNPs (AISNPs), this pilot implementation provides examples and ability to calculate relative likelihoods of ancestry from different populations for user-specified genotypes. FROG-kb has already implemented three different AISNP panels. One is a provisional unpublished panel of 39 SNPs assembled specifically to test functionality of this web application. We have included an illustration of the STRUCTURE output for these AISNPs to document their validity. A set of 128 AISNPs [22,23] and the SNPforID panel of 34 SNPs [19] are two additional panels for which examples and calculations are available. For Phenotype Informative SNPs (PISNPs) we provide a panel of six SNPs for eye color prediction (IrisPlex) along with ability to specify an individual's genotype and predict eye color from that [24].

Limitations

FROG-kb can only provide information for the populations comprehensively tested for the entire set of SNPs. For ancestry inference FROG-kb is currently designed for individuals whose ancestry is overwhelmingly from one population or set of closely related populations. Admixed individuals, in the sense of recent ancestors from geographically and genetically different populations, will not necessarily provide meaningful results. For example, one African American genotype gave Ethiopian as the most likely ancestry, understandable because the Ethiopians have allele frequencies at many loci that are intermediate between those of West African and European populations. Other African American genotypes have given West African populations as most likely. Obviously, different African Americans have different combinations of the African and European alleles for the particular AISNPs in a panel. Thus, results for individuals of recent admixed ancestry will be specific to the individual and the AISNP panel.

Currently the allele designations for the genotypes listed on the data entry forms are not consistent with any single standard. Since most SNPs are unambiguous with respect to the strand being called, the user should have no problem making the necessary conversion from the typing data to the genotype codes on the input screen. However, G/C and A/T SNPs need to be specified as to the strand being called; the future standard for FROG-kb will be that the positive strand (pter to qter, 5' to 3') will be the reference even for SNPs in genes that are coded on the reverse strand.

Only a small selection of the SNP and InDel panels that have been published is currently available for use in FROG-kb. The initial effort in developing this resource for the forensic community has necessarily focused on the database structure and the website interface. Several other panels of markers have already been identified and work has begun to curate the data and make the material accessible and useful via FROG-kb. We also expect the forensic and research communities to help us identify data that should be included but is not.

Conclusion

While FROG-kb is a work in progress, the current version of FROG-kb already provides new and useful functionality. We expect FROG-kb to grow and expand in capabilities over the next several months. Indeed, by the time this initial announcement reaches official publication there will likely be changes. We hope that input from the forensic community will help identify those functionalities that are most helpful and useful. We expect FROG-kb to be a useful reference and resource on use of SNP panels in forensics.

Abbreviations

AISNP: Ancestry informative SNP; App: Application; CODIS: Combined Data Index System; FBI: Federal Bureau of Investigation; IISNP: Individual identification SNP; InDels: Insertion-deletions; KB: Knowledge base; PISNP: Phenotype informative SNP; SNPs: Single nucleotide polymorphisms; STR: Short tandem repeat; STRP: Short tandem repeat polymorphism.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HR and KKK are primarily responsible for writing the manuscript. All authors have contributed with suggestions and revisions and have approved the final version.

Electronic sites referenced

FROG-kb[http://frog.med.yale.edu]
ALFRED[http://alfred.med.yale.edu]
Canadian Random Match Calculator[http://www.csfs.ca/pplus/profiler.htm]
DNA WG STR Population Database[http://www.str-base.org/index.php]
PopAffiliator[http://cracs.fc.up.pt/popaffiliator/]
pop.STR[http://spsmart.cesga.es/popstr.php]
STRBase[http://www.cstl.nist.gov/strbase/]
The Snipper app suite [http://mathgene.usc.es/snipper/]

Acknowledgements

This work is supported in part by grant 2010-DN-BX-K226 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice and also in part by NSF grant BCS0938633. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice.

Author details

[1]Department of Genetics, Yale University School of Medicine, 333 Cedar Street, P.O.Box 208005, New Haven, CT 06520-8005, USA. [2]Center for Medical Informatics, Yale University School of Medicine, New Haven, CT 06520, USA.

Received: 27 March 2012 Accepted: 22 June 2012
Published: 1 September 2012

References

1. Butler JM: **Genetics and genomics of core STR loci used in human identity testing.** *J Forensic Sci* 2006, **51**:253–265.
2. Schneider PM: **Expansion of the European Standard Set of DNA Database Loci-the Current Situation.** *Profiles in DNA* 2009, **12**:6–7.
3. Ge J, Eisenberg A, Budowle B: **Developing criteria and data to determine best options for expanding the core CODIS loci.** *Investigative Genet* 2012, **3**:1.

4. Hares DR: **Expanding the CODIS core loci in the United States.** *Forensic Sci Int Genet* 2012, **6:**e52–e54.
5. Hares DR: **Addendum to expanding the CODIS core loci in the United States.** *Forensic Sci Int Genet* 2012, doi:10.1016/j.fsigen.2012.01.003.
6. *The Evaluation of Forensic DNA Evidence*. Washington, DC: NRC, National Academy Press; 1996.
7. Ruitberg CM, Reeder DJ, Butler JM: **STRBase: a short tandem repeat DNA database for the human identity testing community.** *Nucleic Acids Res* 2001, **29:**320–322.
8. Pereira L, Alshamali F, Andreassen R, Ballard R, Chantratita W, Cho NS, Coudray C, Dugoujon JM, Espinoza M, González-Andrade F, Hadi S, Immel UD, Marian C, Gonzalez-Martin A, Mertens G, Parson W, Perone C, Prieto L, Takeshita H, Villalobos H, Zeng Z, Zhivotovsky L, Camacho R, Fonseca NA: **PopAffiliator: online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile.** *Int J Legal Med* 2011, **125:**629–636.
9. Amigo J, Phillips C, Salas T, Formoso FL, Carracedo A, Lareu M: **pop.STR- An online population frequency browser for established and new forensic STRs.** *Forensic Sci Int: Genet Supplem* 2009, doi:10.1016/j.fsigss.2009.08.178.
10. Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D: **Human genome sequence variation and the influence of gene history, mutation and recombination.** *Nat Genet* 2002, **32:**135–140.
11. Huang QY, Xu FH, Shen H, Deng HY, Liu YJ, Liu YZ, Li JL, Recker RR, Deng HW: **Mutation patterns at dinucleotide microsatellite loci in humans.** *Am J Hum Genet* 2002, **70:**625–634.
12. Dupuy BM, Stenersen M, Egeland T, Olaisen B: **Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci.** *Hum Mutat* 2004, **23:**117–124.
13. Amorim A, Pereira L: **Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs.** *Foren Sci Int* 2005, **150:**17–21.
14. Gill P, Werrett DJ, Budowle B, Guerrieri R: **An assessment of whether SNPs will replace STRs in national DNA databases--joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDAM).** *Sci Justice* 2004, **44:**51–53.
15. Sanchez JJ, Borsting C, Hallenberg C, Buchard A, Hernandez A, Morling N: **Multiplex PCR and minisequencing of SNPs--a model with 35 Y chromosome SNPs.** *Foren Sci Int* 2003, **137:**74–84.
16. Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AM, Carroll ML, Nguyen SV, Walker JA, Prasad BV, Reddy PG, Das PK, Batzer MA, Jorde LB: **Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms.** *Genome Res* 2003, **13:**1607–1618.
17. Fondevila M, Pereira R, Gusmao L, Phillips C, Lareu MV, Carracedo A, Butler JM, Vallone PM: **Forensic performance of insertion-deletion marker systems.** *Forensic Sci Int: Genet Suppl Series* 2011, **3:**e443–e444. Qiagen Investigator DIPplex kit.
18. Pereira R, Phillips C, Pinto N, Santos C, dos Santos SE, Amorim A, Carracedo Á, Gusmão L: **Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing.** *PLoS One* 2012, **7:**e29684.
19. Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, AÁlvarez-Dios J, Calaza M, Casaresde Cal M, Ballard D, Lareu MV, Carracedo A, and The SNPforID Consortium: **Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs.** *Forensic Sci Int Genet* 2007, **1:**273–280.
20. Pakstis AJ, Speed WC, Fang R, Hyland FCL, Furtado MR, Kidd JR, Kidd KK: **SNPs for a universal individual identification panel.** *Hum Genet* 2010, **127:**315–324.
21. Sánchez JJ, Phillips C, Borsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N: **A multiplex assay with 52 single nucleotide polymorphisms for human identification.** *Electrophoresis* 2006, **27:**1713–1724.
22. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF: **Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America.** *Hum Mutat* 2009, **30:**69–78.
23. Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, De La Vega FM, Kidd KK: **Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples.** *Investigative Genet* 2011, **2:**1.
24. Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M: **IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information.** *Forensic Sci Int Genet* 2011, **5:**170–80.
25. Cheung KH, Osier MV, Kidd JR, Pakstis AJ, Miller PL, Kidd KK: **ALFRED: an allele frequency database for diverse populations and DNA polymorphisms.** *Nucleic Acids Res* 2000, **28**(1):361–363.
26. Rajeevan H, Soundararajan U, Kidd JR, Pakstis AJ, Kidd KK: **ALFRED: an allele frequency resource for research and teaching.** *Nucleic Acids Res* 2012, **40:**D1010–D1015.

doi:10.1186/2041-2223-3-18
Cite this article as: Rajeevan *et al.*: Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb. *Investigative Genetics* 2012 **3**:18.