# Forensic Resource/Reference On Genetics – knowledge base: FROG-kb User's Manual

**Updated June, 2017**

# Table of Contents

# 1. Introduction

The purpose of the Forensic Resource/Reference On Genetics – knowledge base (FROG-kb) database is to provide users with online tools for comparing user-provided data with underlying allele frequencies in populations. The database serves as a teaching and research web interface to facilitate the use of single-nucleotide polymorphisms (SNPs) in the forensic field.

FROG-kb is a freely available, web-accessible, and curated resource and reference of allele frequency data for SNPs and other genetic polymorphisms. The data used in FROG-kb calculations derive from the ALlele FREquency Database (ALFRED http://alfred.med.yale.edu), a continually updated database of allele frequency data on SNPs and other genetic polymorphisms.

Current SNP panel categories within FROG-kb include Individual Identification SNPs (IISNPs), Ancestry Inference SNPs (AISNPs), and Phenotype Interference SNPs (PISNPs). IISNPs provide the ability to calculate random match probabilities for user-specified genotypes. AISNPs provide the ability to calculate relative likelihoods of ancestry from different reference populations for user-specified genotypes. PISNPs provide the ability to predict a visible trait for user-specified genotypes for the relevant SNPs.

This User's Manual provides detailed information on these SNP panels and instructions for the navigation and searching of each within FROG-kb. Additional information on FROG-kb and ALFRED can be accessed at each database's website and found in the publications listed in the Reference section (Section 9) of this User's Manual.

# 2. Accessing FROG-kb Home Page and Features

FROG-kb can be accessed at **http://frog.med.yale.edu/FrogKB/**. The recommended browsers for accessing FROG-kb are the latest versions of Mozilla Firefox and Google Chrome. Some versions of Internet Explorer are not compatible. If you have any issues please reach out to the database curator using the **Contact Us** page described in **Section 4.3** of this User's Manual.

# 3. Home Page and Navigation Tools

***Exhibit 1*** presents the FROG-kb Home Page, which provides users with information about FROG-kb and a brief description of the types of data sets within the database. The database's Main Menu is shown across the top of the page (see red box). The Main Menu allows easy navigation among menu options and is accessible from any page on the FROG-kb database.

**Exhibit 1. FROG-kb showing the Main Menu (A) options and the Home (B) page description**


**Welcome to the New FROG-kb**

FROG-kb seeks to make allele frequency data for panels of SNPs more useful in a forensic setting. The primary objective of FROG-kb is to provide a web interface that, from a forensic perspective, is useful for teaching and research and can serve as a tool facilitating forensic practice. The underlying data are housed in the already extensively used and referenced ALlele FREquency Database (ALFRED).

The FROG-kb interface makes the information usable for forensic purposes, including computational tools for comparing user-provided data with underlying allele frequencies in populations. The buttons across the top provide top level navigation. This new FROG-kb interface has all of the previous functionalities with supplementary reference material and "space" for additional documentation that will be added over the next few months. FROG-kb functions best using Mozilla Firefox and Google Chrome; some versions of Internet Explorer are not compatible.

FROG-kb focuses on two types of SNP panels, Individual Identification (IISNPs) and Ancestry Inference (AISNPs), but also includes a Phenotype Inference (PISNPs) panel for eye color. IISNP and AISNP panels provide the ability to enter genotypes of an individual at multiple SNPs and calculate probabilities of that multisite genotype in each of several populations. More information on each SNP panel can be found in Section 4.1 of the User Manual (under the Documentation button) along with information on searching each SNP panel within FROG-kb. Also, go to Announcements for updates.

If you come across any issues or have suggestions use the Contact Us button to reach us. "If you see something - say something"

As shown in Exhibit 1, the features of the Main Menu are as follows:

- Home
- IISNP
- AISNP
- PISNP
- Documentation
- Contact Us
- Announcements

# 4. Navigating FROG-kb

The purpose of this section is to provide the user with an overview of the Main Menu's features, as well as a description of the information that can be found in each.

## 4.1 SNP Panel Categories: IISNP, AISNP, PISNP


The IISNP, AISNP, and PISNP features on the FROG-kb Main Menu provide access to multiple published SNP panels reviewed and entered by the curator. Clicking any of these three features opens a page that lists the SNP panels under that category.

As shown in ***Exhibit 2***, each SNP panel shown in the list includes a reference for the panel. In addition, the right side of each panel includes a link (***Navigate to ALFRED***) that allows the user to access ALFRED for more detail about the panel. Each SNP panel also shows a **Go** button above the panel, which users can click to open the panel.

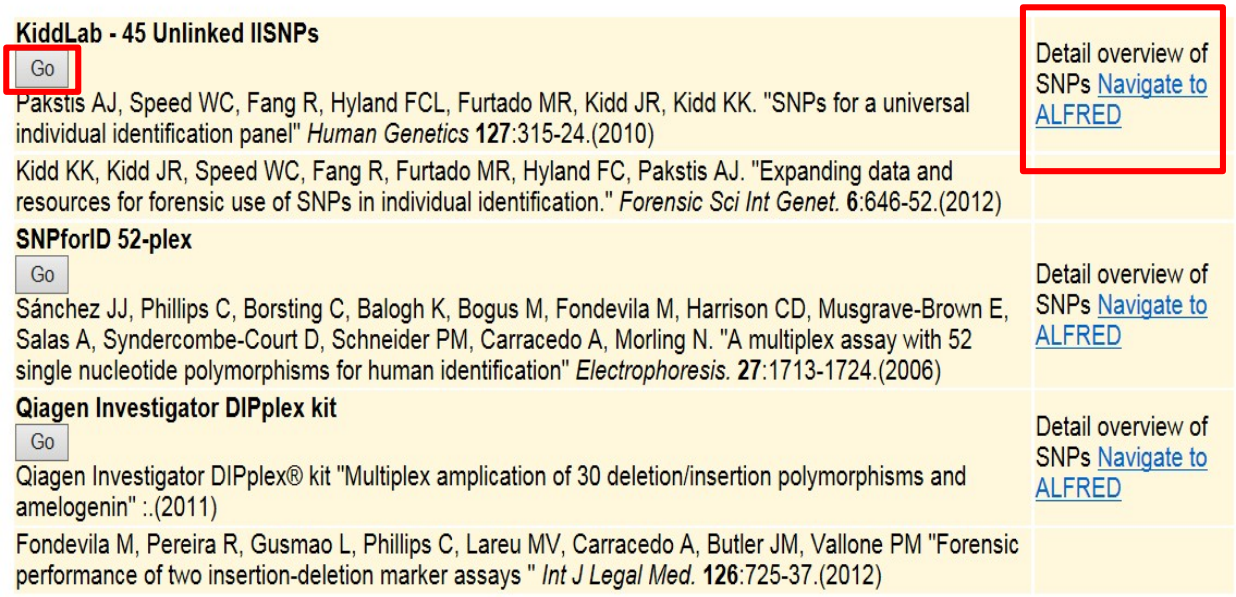
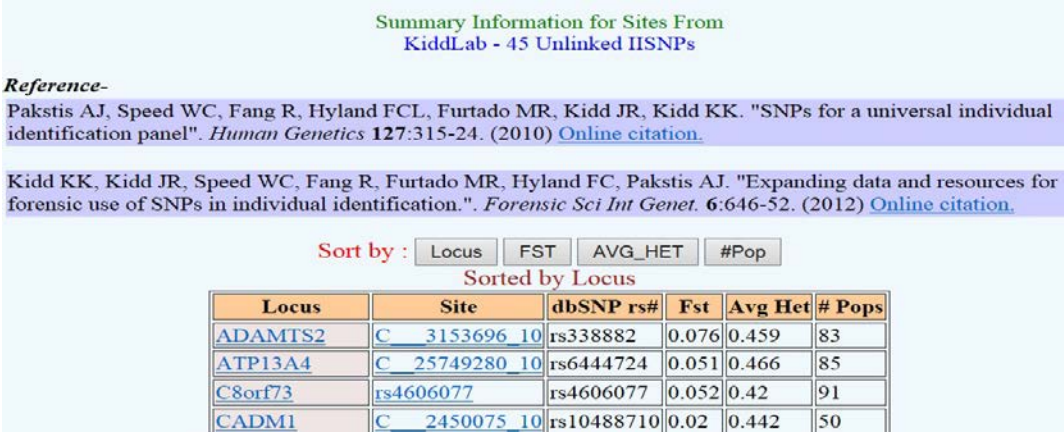



**Exhibit 2. KiddLab -45 Unlinked IISNPs illustrating layout seen in IISNP, AISNP, and PISNP options.**

Clicking on the ***Navigate to ALFRED*** link opens the SNP Set page within ALFRED in a new browser window (***Exhibit 3***). The SNP Set module in ALFRED has multiple options for viewing the presented information, including sorting the SNPs by FST and average heterozygosity.



Summary Information for Sites From
KiddLab - 45 Unlinked IISNPs

*Reference-*

Pakstis AJ, Speed WC, Fang R, Hyland FCL, Furtado MR, Kidd JR, Kidd KK. "SNPs for a universal individual identification panel". *Human Genetics* **127**:315-24. (2010) Online citation.

Kidd KK, Kidd JR, Speed WC, Fang R, Furtado MR, Hyland FC, Pakstis AJ. "Expanding data and resources for forensic use of SNPs in individual identification.". *Forensic Sci Int Genet.* **6**:646-52. (2012) Online citation.

Sort by : Locus FST AVG_HET #Pop

Sorted by Locus

| Locus | Site | dbSNP rs# | Fst | Avg Het | # Pops |
|---|---|---|---|---|---|
| ADAMTS2 | C___3153696_10 | rs338882 | 0.076 | 0.459 | 83 |
| ATP13A4 | C__25749280_10 | rs6444724 | 0.051 | 0.466 | 85 |
| C8orf73 | rs4606077 | rs4606077 | 0.052 | 0.42 | 91 |
| CADM1 | C___2450075_10 | rs10488710 | 0.02 | 0.442 | 50 |

**Exhibit 3. Additional information found on ALFRED database for KiddLab -45 Unlinked IISNPs**

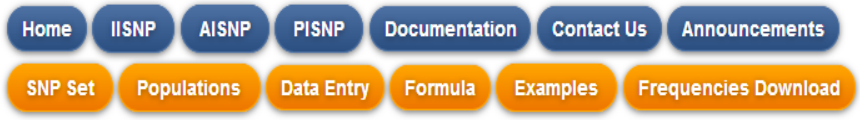
Clicking the **Go** button above an SNP panel opens the panel, as well as a row of option buttons for viewing the panel. As shown in ***Exhibit 4***, these option buttons are as follows:

- SNP Set,
- Populations,
- Data Entry,
- Formula, and
- Examples.

Each SNP panel category (IISNP, AISNP, and PISNP) presents the same row of option buttons, except the PISNP panel, which does not include Populations.

**Exhibit 4. Panel of 45 IISNPs SNP Set information.**

The following subsections provide information on each of the option buttons accessible on a SNP panel.




**KiddLab - 45 Unlinked IISNPs**

(The table is sortable - click on column header to sort) Download SNP List

| Navigate to ALFRED | Navigate to dbSNP ▼ | chromosome | chromosome position |
|---|---|---|---|
| SI001900K | rs10092491 | 8 | 28,411,072 |
| SI001899B | rs10488710 | 11 | 115,207,176 |
| SI001402H | rs1058083 | 13 | 100,038,233 |
| SI015392U | rs10773760 | 12 | 130,761,696 |
| SI015046Q | rs10776839 | 9 | 137,417,308 |
| SI001909T | rs1109037 | 2 | 10,085,722 |

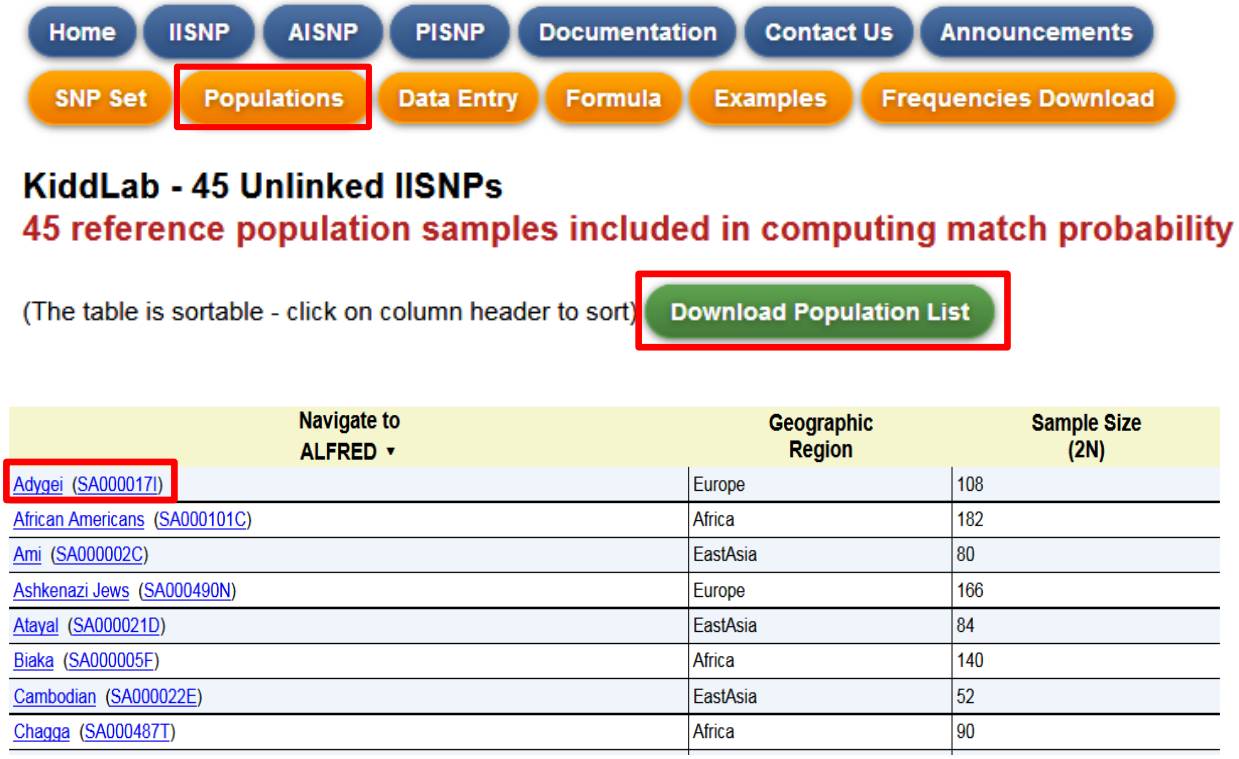
### *SNP Set*

The default view for a SNP panel is a sortable table that presents the **SNP Set** (the first option button shown). As shown previously in Exhibit 4 and to the right, the first column of this table presents an ALFRED UID (e.g., SI001900K) that links to ALFRED for polymorphism information. The second column of the table shows a list of rs numbers (e.g., rs10092491); each number is an active link to the dbSNP record for that SNP. Note that within the SNP Set in ALFRED, additional populations may have data for some, but not all, populations, and that those populations are not included in the calculations in FROG-kb.

| Navigate to ALFRED | Navigate to dbSNP |
|---|---|
| SI001900K | rs10092491 |
| SI001899B | rs10488710 |
| SI001402H | rs1058083 |
| SI015392U | rs10773760 |
| SI015046Q | rs10776839 |

### *Populations*

Clicking the **Populations** option button provides a list of populations for which comparable calculations can be made (***Exhibit 5***). This is the set of populations for which all SNPs in the set have allele frequency data. Each population name and the sample UID shown in the list is an active link. Clicking on  a link opens a new browser window that presents ALFRED information on the population and the sample.




| Navigate to ALFRED | Geographic Region | Sample Size (2N) |
|---|---|---|
| Adygei (SA000017I) | Europe | 108 |
| African Americans (SA000101C) | Africa | 182 |
| Ami (SA000002C) | EastAsia | 80 |
| Ashkenazi Jews (SA000490N) | Europe | 166 |
| Atayal (SA000021D) | EastAsia | 84 |
| Biaka (SA000005F) | Africa | 140 |
| Cambodian (SA000022E) | EastAsia | 52 |
| Chagga (SA000487T) | Africa | 90 |

**Exhibit 5. Panel of 45 IISNPs Population information**

### *Data Entry*

Clicking the **Data Entry** option button allows the user to specify an individual, multi-site genotype and to calculate the probability of that genotype in each of the populations. As shown in ***Exhibit 6***, two options are provided at the bottom of the Data Entry page for the user to enter genotype data for a particular SNP panel: **File Upload** and **Selection by Radio Button**. ***Section 5.1*** of this User's Manual provides more information on entering data in FROG-kb.


**Data entry for KiddLab - 45 Unlinked IISNPs in FROG-kb**

There are two options for users to enter genotype data for a particular SNP panel; 'Selection by Radio Button' and 'File Upload'.

**Selection by Radio Button** opens the ability to specify an individual's multi-site genotype using radio button. The list of SNPs can be sorted by rs-numbers, chromosome, siteUId and chromosomal position for ease of working with the set. For each SNP listed in the data entry table, users can navigate to ALFRED using the ALFRED UID URL link, and navigate to dbSNP using the rs-number URL link. Also displayed in the data entry table is the chromosome and chromosomal position. Radio buttons for the selection of possible genotypes are also displayed in the data entry table. Selecting the Compute button will calculate the probability of that genotype in each of the populations.

**File Upload** function provides users with an option to enter SNP information and corresponding genotype for a panel in a text area. Example downloadable files for each SNP panel is provided. Users can save these files and modify them by replacing the 'NN'(unknown) genotype to observed genotype in the individual. The file starts with the SNP Set tag. The tag provides information to the internal code on the type of SNP Set. For example, 'ai34' for SNPforID 34-plex and 'ii52' for SNPforID 52-plex. Copy and paste this file on to the text area provided in the Input Genotype for a Panel function. Select Upload to upload and run the computation.

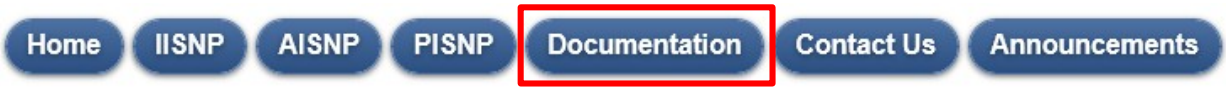
File Upload | Selection by Radio Button

**Exhibit 6. Data Entry for KiddLab – 45 Unlinked IISNPs in FROG-kb.**

### *Formula and Examples*

Clicking the **Formula** option button opens a page that presents an explanation on how the probabilities are derived. The **Examples** option button allows users to access a pre-entered data entry page for one individual from a specific population. Both the Example and Formula options are explained further in **Section 6** and **Section 7** of this Users' Manual, respectively.

## 4.2 Documentation




Clicking the **Documentation** feature in the Main Menu opens an additional set of option buttons for users to learn more about FROG-kb capabilities. Similar to the options that appear on the SNP panel pages, the Documentation options appear as orange buttons under the Main Menu.


As shown above, the option buttons that appear under Documentation are as follows:

- Functionalities,
- Pipeline,
- FAQ (Frequently Asked Questions),
- Manual, and
- Resources.

### *Functionalities*

The **Functionalities** option provide an overview for navigating through the SNP panel information, SNP panel example data, SNP panel data entry, and results of the data entry calculations (***Exhibit 7***). The Functionalities serves as an online, secondary location for information provided in this User's Manual.

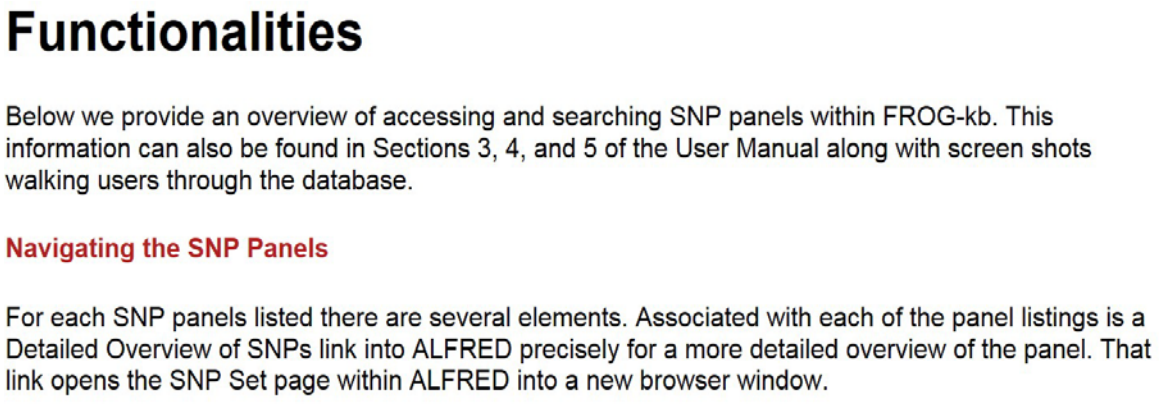




**Exhibit 7. Functionalities within the Main Menu option for Documentation provides navigation overview of FROG-kb.**

### *Pipeline*

Clicking on the **Pipeline** option opens a graphical representation that illustrates the process for inputting, searching, and retrieving data from FROG-kb.

### *FAQ*

The **FAQ** option presents answers to Frequently Asked Questions about FROG-kb (***Exhibit 8***). Users who have additional questions not found on the FAQ page should follow the process outlined in **Section 4.3** of this User's Manual to submit their questions to the database curator.

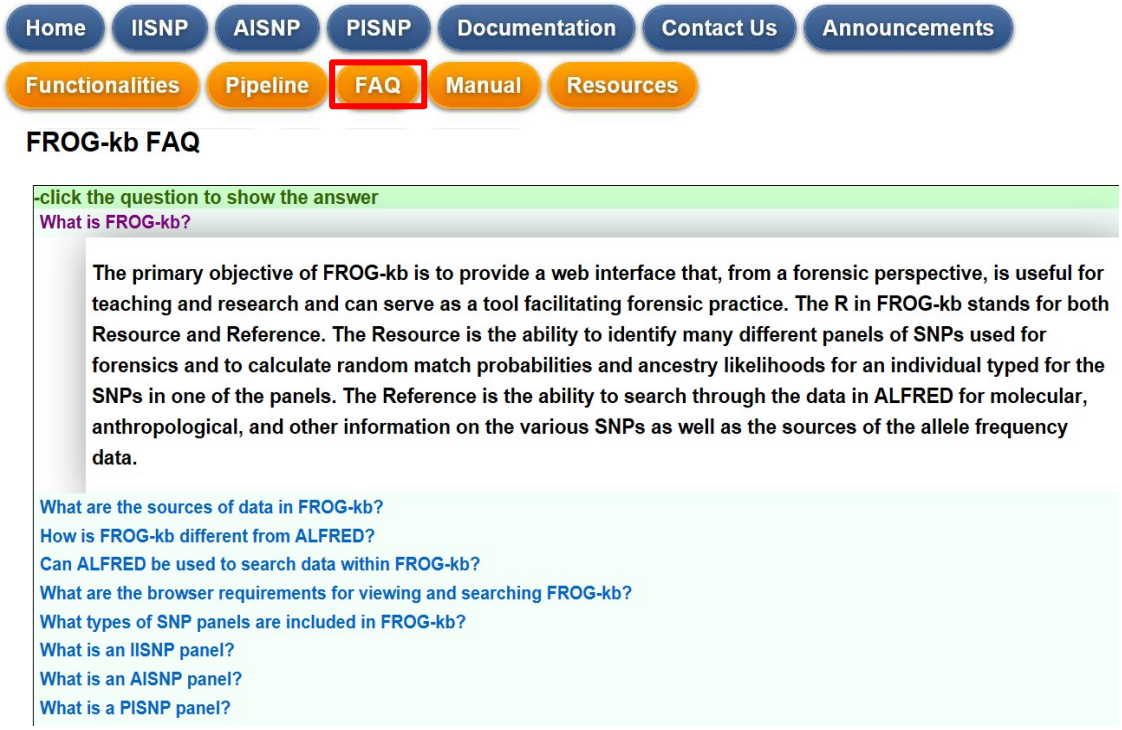



**Exhibit 8. List of Frequently Asked Questions for FROG-kb.**

### *Manual*

Clicking the **Manual** option opens a copy of this User's Manual.

### *Resources*

The **Resources** option button provides users with links (STR- and SNP-based) and references to potentially relevant databases and sites. This page also lists poster presentations pertaining to FROG-kb that have been presented at annual conferences and that are freely accessible to users.

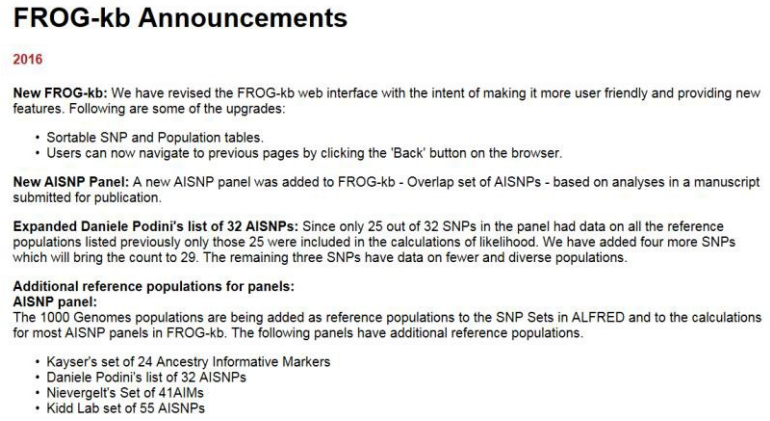
## 4.3 Contact Us


The **Contact Us** feature in the FROG-kb Main Menu serves as a resource for users to provide feedback and questions to the database curator. Users can also inquire about submitting data and providing additional information that will assist the community. As shown in ***Exhibit 9***, the Contact Us page presents a form where users can provide their email address and a subject and message. Clicking the **Submit** button at the bottom of the page sends the message to the database curator.


**Exhibit 9. Contact Us submission form**

## 4.4 Announcements


The **Announcements** feature on the Main Menu provides users with updates pertaining to FROG-kb (***Exhibit 10***). These updates may include additional functionalities or upgrades to the web interface, additional panels, and expansion of existing panels in FROG-kb.



**FROG-kb Announcements**

2016

**New FROG-kb:** We have revised the FROG-kb web interface with the intent of making it more user friendly and providing new features. Following are some of the upgrades:

- Sortable SNP and Population tables.
- Users can now navigate to previous pages by clicking the 'Back' button on the browser.

**New AISNP Panel:** A new AISNP panel was added to FROG-kb - Overlap set of AISNPs - based on analyses in a manuscript submitted for publication.

**Expanded Daniele Podini's list of 32 AISNPs:** Since only 25 out of 32 SNPs in the panel had data on all the reference populations listed previously only those 25 were included in the calculations of likelihood. We have added four more SNPs which will bring the count to 29. The remaining three SNPs have data on fewer and diverse populations.

**Additional reference populations for panels:**
**AISNP panel:**
The 1000 Genomes populations are being added as reference populations to the SNP Sets in ALFRED and to the calculations for most AISNP panels in FROG-kb. The following panels have additional reference populations.

- Kayser's set of 24 Ancestry Informative Markers
- Daniele Podini's list of 32 AISNPs
- Nievergelt's Set of 41AIMs
- Kidd Lab set of 55 AISNPs

**Exhibit 10. FROG-kb Announcements Page**

# 5. Searching IISNP, AISNP, and PISNP Data

**Section 4.1** of this User's Manual, ***SNP Panel Categories: IISNP, PISNP, AISNP***, provides an overview for navigating the SNP panels and the functionalities within each. The following sections expand on that information by walking users through the process for inputting, searching, and retrieving data in FROG-kb.

***Exhibit 11*** shows the Data Entry default page for the 45 Unlinked IISNP panel. The page presents two options for users to enter genotype data for an SNP panel: **File Upload** and **Selection by Radio Button**.

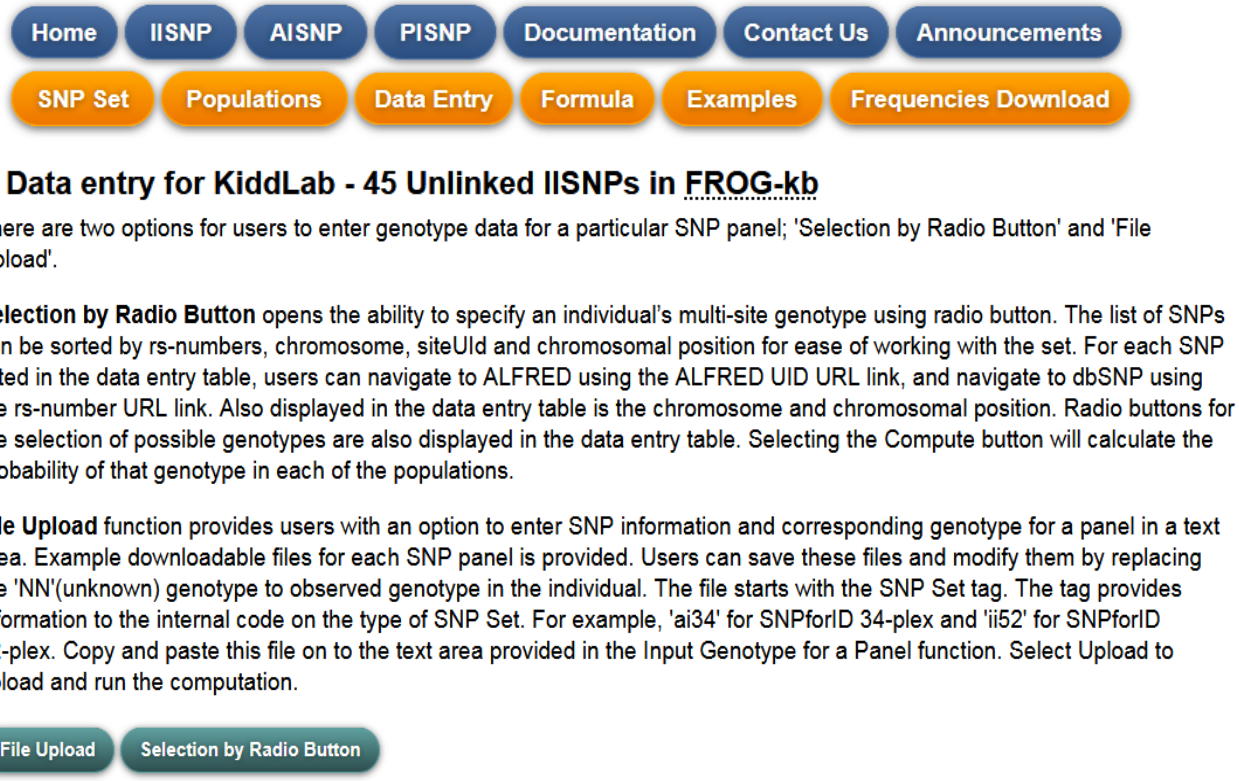



**Exhibit 11. Data Entry Page for IISNP Data Entry.**

## 5.1 Inputting and Searching Data: Selection by Radio Button

***Exhibit 12*** shows the data entry form for the **Selection by Radio Button** option. This form allows the user to specify an individual's multi-site genotype. By default, the list of SNPs is

sorted by rs numbers; however, users can also sort the list by any of the first four columns by clicking on the preferred header. There is an option in this page to enter a label for the analysis providing an automatic identification record of the results generated for that particular input genotype.

As shown in the first two columns of the table, for each SNP on the list, the ALFRED UID and dbSNP rs numbers are displayed as active links to the respective databases. The third column presents the chromosome (chr), and the fourth column provides the chromosome position. As shown in the red box, the next three columns present radio buttons for categorizing the possible genotypes. Users can enter the genotype by simply clicking on the radio button for the genotype. If the genotype is unknown, the last column in the data entry table allows the user to select the Unknown (NN) option.

The two action buttons of this page are **Compute** and **Set all unselected to known and compute**. These action buttons are discussed in the following subsections.

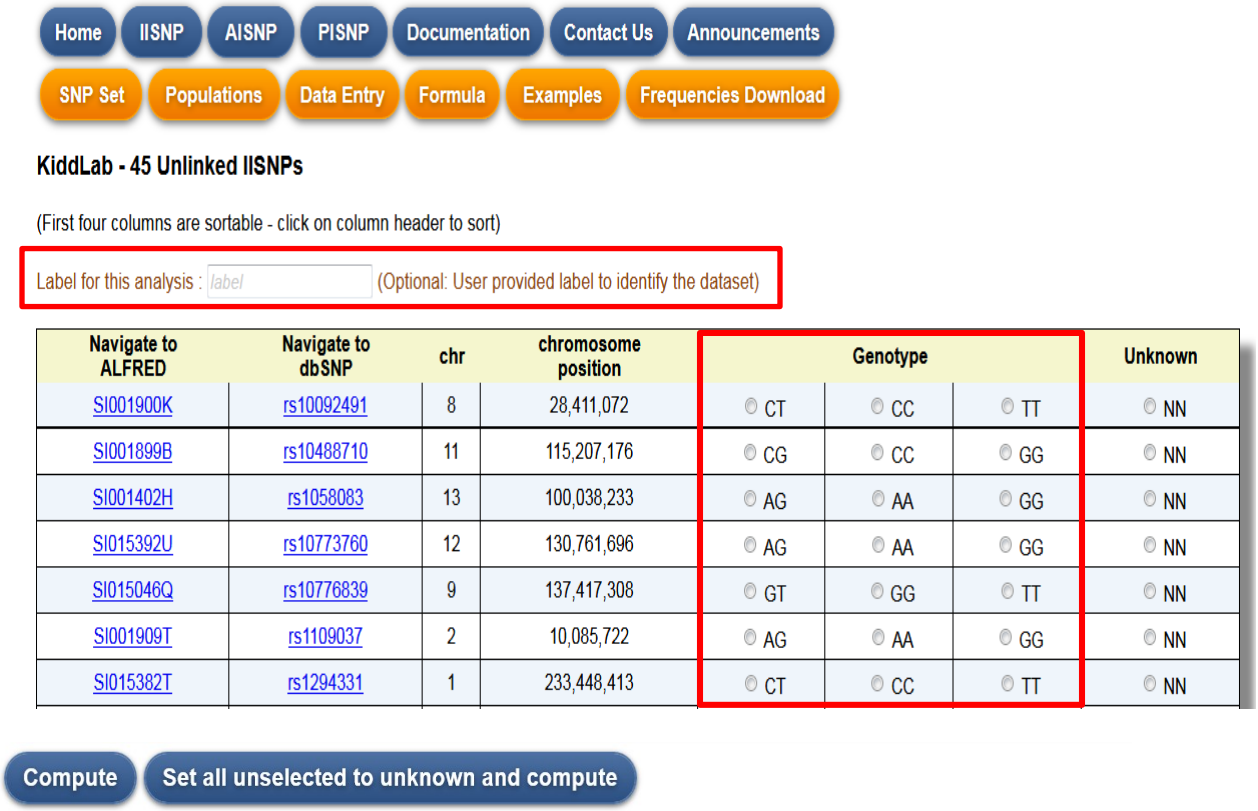



**Exhibit 12. KiddLab – 45 Unlinked IISNPs Data Entry table.**

### *Compute*

Once all the information has been entered, the user can select **Compute** at the bottom of the Data Entry table to initiate the calculation and display the results. If any SNPs do not have a radio

button entry (Genotype or Unknown), the system will flag the missing entry and a warning box will appear indicating that the user must add the missing information. Once a radio button has been selected for each SNP, the user must again click Compute to run the calculation.

### *Set All Unselected to Unknown and Compute*

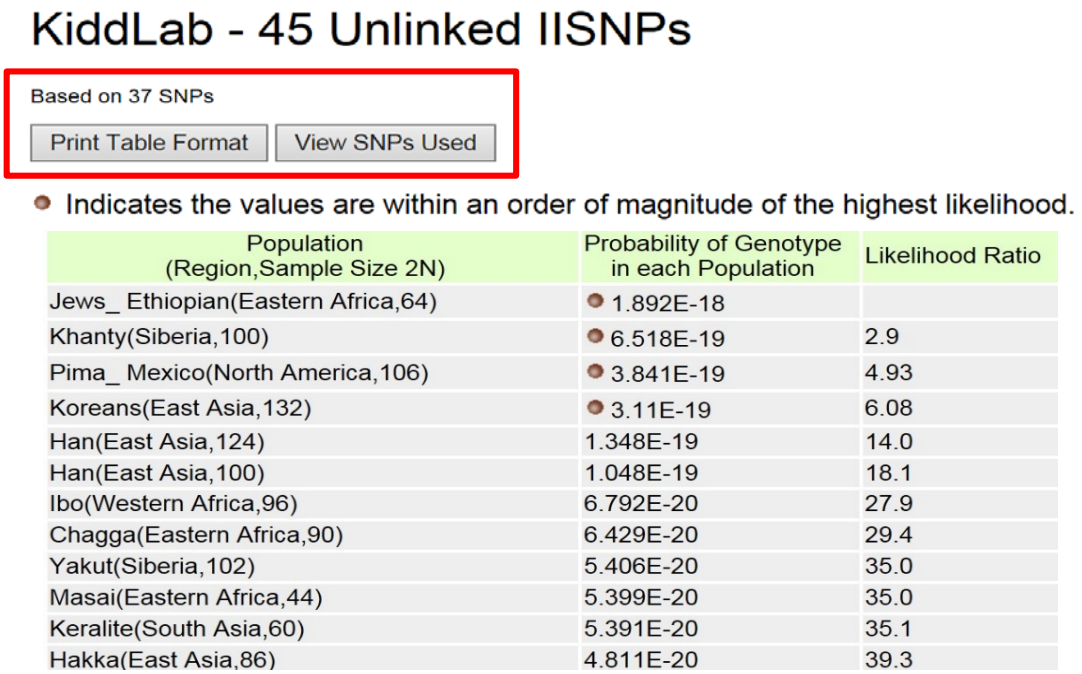
FROG-kb also provides user with a **Set all unselected to unknown and compute** option, which is shown at the bottom of the Data Entry table next to the Compute button. If preferred, a user can click this option instead of Compute, and the system with automatically capture any genotypes for which a radio button was not selected and mark these as Unknown. Note that using this feature assumes that the user has already entered in all known genotypes.

Both **Compute** and **Set all unselected to unknown** initiate the FROG-kb calculation and display the results when all SNPs have an entry.

## 5.2 The Statistical Results

For each of the Individual Identification and Ancestry Inference SNP panels FROG-kb calculates the probability of the user-entered multi-locus genotype in each of the reference populations. As shown in ***Exhibit 13,*** the results of a calculation are displayed as a table with three columns that outline the following: (1) the name of the reference population sampled with its geographic region and the sample size, (2) the probability of the entered genotype occurring in that population, and (3) the likelihood ratio of the highest probability (in the top row) to the probability in the specific reference population. The populations are ordered by their probabilities of generating the entered genotype from highest to lowest.

The text above the table (see red box in exhibit) indicates the number of SNPs used in the calculation, and the **Print Table Format** and **View SNPs Used** buttons allow the user to print the output table and to see the list of SNPs used.




| Population (Region,Sample Size 2N) | Probability of Genotype in each Population | Likelihood Ratio |
|---|---|---|
| Jews_ Ethiopian(Eastern Africa,64) | 1.892E-18 | |
| Khanty(Siberia,100) | 6.518E-19 | 2.9 |
| Pima_ Mexico(North America,106) | 3.841E-19 | 4.93 |
| Koreans(East Asia,132) | 3.11E-19 | 6.08 |
| Han(East Asia,124) | 1.348E-19 | 14.0 |
| Han(East Asia,100) | 1.048E-19 | 18.1 |
| Ibo(Western Africa,96) | 6.792E-20 | 27.9 |
| Chagga(Eastern Africa,90) | 6.429E-20 | 29.4 |
| Yakut(Siberia,102) | 5.406E-20 | 35.0 |
| Masai(Eastern Africa,44) | 5.399E-20 | 35.0 |
| Keralite(South Asia,60) | 5.391E-20 | 35.1 |
| Hakka(East Asia,86) | 4.811E-20 | 39.3 |

**Exhibit 13. Results from KiddLab-45 Unlinked IISNPs data entry.**

Irrespective of the IISNP or AISNP panel used, the probabilities calculated have two interpretations. One interpretation is that the probability of the entered genotype is equivalent to a random match probability (RMP) assuming no deviation from Hardy-Weinberg ratios in the population. The alternative is that the probability can be considered to be the likelihood that the specific population is the origin of the entered genotype. If ancestry inference is the objective, the absolute value has no meaning; only the relative likelihoods are meaningful. The population with the highest probability is the most likely ancestral population among the set of reference populations. (see Section 7 for discussion of ancestry inference). The populations are ordered by the probabilities of the entered genotype from highest to lowest; therefore, the likelihood ratios range from 1 to some larger number for the least likely population, representing how many times more likely the entered genotype is in the highest-ranking population compared to the specific reference population. 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 than the best.

Separately, the log of the probabilities ***(Exhibit 14***) of the entered genotype is plotted from highest to lowest as a mouse-over graph that specifies the population and value at the point indicated.

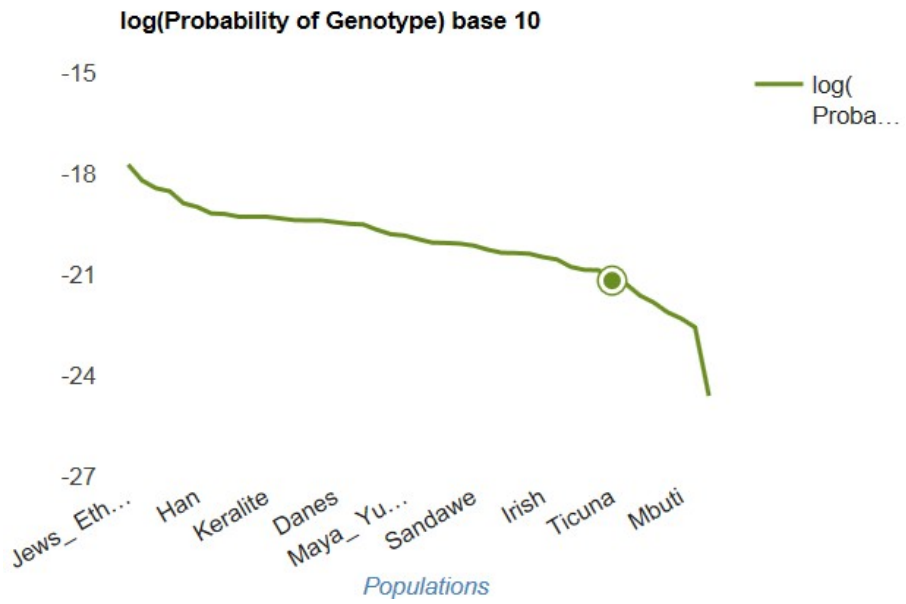



**Exhibit 14. Log of probabilities for data entered in KiddLab-45 Unlinked IISNPs.**

The IISNP panels have less information on actual ancestry than the AISNP panels. The relative likelihoods for the populations do not generally provide useful ancestry information because the SNPs in the IISNP panels were chosen to have similar allele frequency values around the world. Exhibit 13 shows that the four highest likelihoods represent four very different parts of the world and have likelihoods that are not significantly different. This is what is expected for an IISNP panel. It is up to the user to decide whether the relevant breadth of global coverage is sufficient to use the highest probability for the entered genotype as a forensically relevant maximum probability for a random match in a global context.

## 5.3 Inputting and Searching Data: File Upload

The second option to input data is the File Upload function shown previously in Exhibit 12. The File Upload function provides users with an option to enter SNP information and a corresponding genotype for a panel in a text-area. ***Exhibit 15*** shows the default File Upload page. Note that in addition to the input-specific buttons shown at the bottom of this page (**File Function** and **Input Genotype for Panel**), the page provides an **About** option that serves as an online reference for users on inputting genotype data using the File Upload function. The information provided on the **About** page is the same as the instructions outlined in this User's Manual.

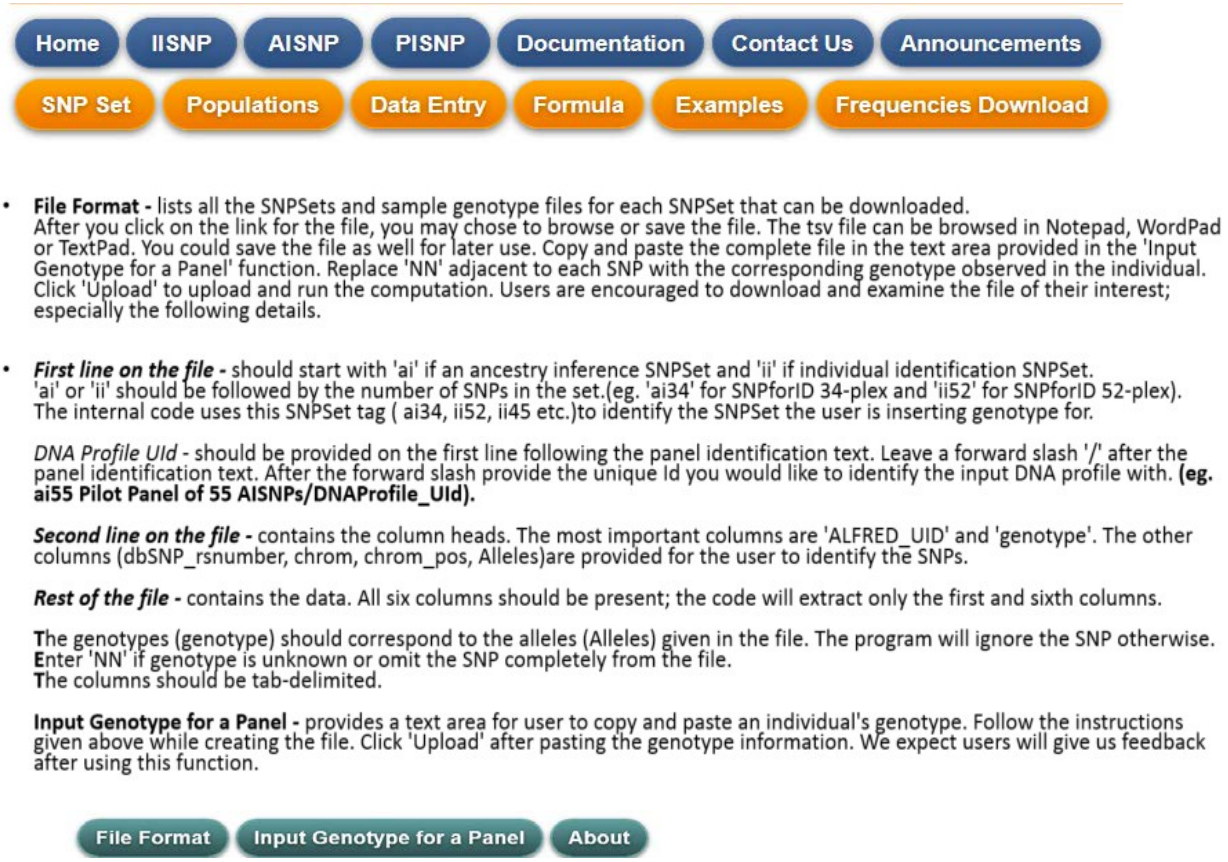
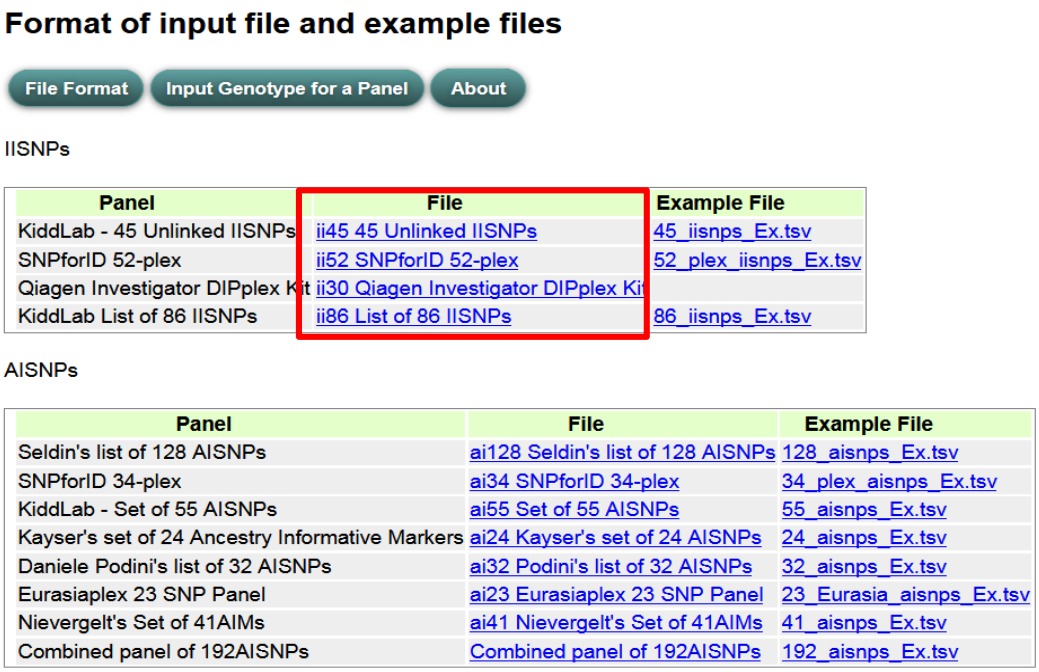



**Exhibit 15. Explanation on using the File Upload option for Data Entry**

Clicking on the **File Format** option at the bottom of the File Upload page lists all the SNP Sets and the sample genotype files for each set can be downloaded, as shown in ***Exhibit 16***.



**Format of input file and example files**

File Format | Input Genotype for a Panel | About

IISNPs

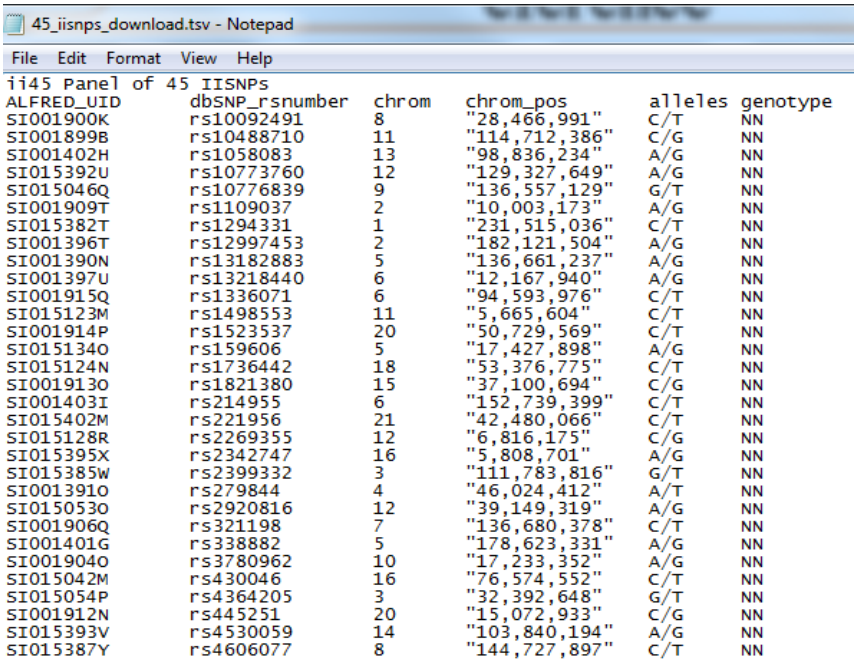
| Panel | File | Example File |
|---|---|---|
| KiddLab - 45 Unlinked IISNPs | ii45 45 Unlinked IISNPs | 45_iisnps_Ex.tsv |
| SNPforID 52-plex | ii52 SNPforID 52-plex | 52_plex_iisnps_Ex.tsv |
| Qiagen Investigator DIPplex Kit | ii30 Qiagen Investigator DIPplex Kit | |
| KiddLab List of 86 IISNPs | ii86 List of 86 IISNPs | 86_iisnps_Ex.tsv |

AISNPs

| Panel | File | Example File |
|---|---|---|
| Seldin's list of 128 AISNPs | ai128 Seldin's list of 128 AISNPs | 128_aisnps_Ex.tsv |
| SNPforID 34-plex | ai34 SNPforID 34-plex | 34_plex_aisnps_Ex.tsv |
| KiddLab - Set of 55 AISNPs | ai55 Set of 55 AISNPs | 55_aisnps_Ex.tsv |
| Kayser's set of 24 Ancestry Informative Markers | ai24 Kayser's set of 24 AISNPs | 24_aisnps_Ex.tsv |
| Daniele Podini's list of 32 AISNPs | ai32 Podini's list of 32 AISNPs | 32_aisnps_Ex.tsv |
| Eurasiaplex 23 SNP Panel | ai23 Eurasiaplex 23 SNP Panel | 23_Eurasia_aisnps_Ex.tsv |
| Nievergelt's Set of 41AIMs | ai41 Nievergelt's Set of 41AIMs | 41_aisnps_Ex.tsv |
| Combined panel of 192AISNPs | Combined panel of 192AISNPs | 192_aisnps_Ex.tsv |

**Exhibit 16. List of all SNP Sets and sample genotypes for download**

After clicking on the link for the preferred file (see red box in Exhibit 16), users can choose to browse or save the file. The .tsv file can be browsed in Notepad, WordPad, or TextPad. ***Exhibit 17*** shows a .tsv file for the 45 IISNPs panel opened in Notepad.



45_iisnps_download.tsv - Notepad

File Edit Format View Help

```
ii45 Panel of 45 IISNPs
ALFRED_UID      dbSNP_rsnumber  chrom   chrom_pos       alleles genotype
SI001900K       rs10092491      8       "28,466,991"    C/T     NN
SI001899B       rs10488710      11      "114,712,386"   C/G     NN
SI001402H       rs1058083       13      "98,836,234"    A/G     NN
SI015392U       rs10773760      12      "129,327,649"   A/G     NN
SI015046Q       rs10776839      9       "136,557,129"   G/T     NN
SI001909T       rs1109037       2       "10,003,173"    A/G     NN
SI015382T       rs1294331       1       "231,515,036"   C/T     NN
SI001396T       rs12997453      2       "182,121,504"   A/G     NN
SI001390N       rs13182883      5       "136,661,237"   A/G     NN
SI001397U       rs13218440      6       "12,167,940"    A/G     NN
SI001915Q       rs1336071       6       "94,593,976"    C/T     NN
SI015123M       rs1498553       11      "5,665,604"     C/T     NN
SI001914P       rs1523537       20      "50,729,569"    C/T     NN
SI015134O       rs159606        5       "17,427,898"    A/G     NN
SI015124N       rs1736442       18      "53,376,775"    C/T     NN
SI001913O       rs1821380       15      "37,100,694"    C/G     NN
SI001403I       rs214955        6       "152,739,399"   C/T     NN
SI015402M       rs221956        21      "42,480,066"    C/T     NN
SI015128R       rs2269355       12      "6,816,175"     C/G     NN
SI015395X       rs2342747       16      "5,808,701"     A/G     NN
SI015385W       rs2399332       3       "111,783,816"   G/T     NN
SI001391O       rs279844        4       "46,024,412"    A/T     NN
SI015053O       rs2920816       12      "39,149,319"    A/G     NN
SI001906Q       rs321198        7       "136,680,378"   C/T     NN
SI001401G       rs338882        5       "178,623,331"   A/G     NN
SI001904O       rs3780962       10      "17,233,352"    A/G     NN
SI015042M       rs430046        16      "76,574,552"    C/T     NN
SI015054P       rs4364205       3       "32,392,648"    G/T     NN
SI001912N       rs445251        20      "15,072,933"    C/G     NN
SI015393V       rs4530059       14      "103,840,194"   A/G     NN
SI015387Y       rs4606077       8       "144,727,897"   C/T     NN
```

**Exhibit 17. 45 IISNPs .tsv file opened in Notepad**

Users are encouraged to download and examine the files of their interest in a .tsv file, especially the following details:

- The first line on the file should start with 'ai' if an ancestry inference SNPSet and 'ii' if individual identification SNPSet.
  - 'ai' or 'ii' should be followed by the number of SNPs in the set.(e.g., 'ai34' for SNPforID 34-plex and 'ii52' for SNPforID 52-plex).
  - The internal code uses this SNPSet tag (e.g., ai34, ii52, ii45) to identify the SNPSet for which the user is inserting the genotype.
- The second line on the file contains the column headers. The most important columns are 'ALFRED_UID' and 'genotype'. The other columns (dbSNP_rsnumber, chrom, chrom_pos, Alleles) are provided for the user to identify the SNPs.
- The rest of the file contains the data. All six columns should be present; the code will extract only the first and sixth columns.
- The genotypes (genotype) should correspond to the alleles (Alleles) given in the file. The program will ignore the SNP otherwise. Enter 'NN' if genotype is unknown, or omit the SNP completely from the file. The columns should be tab-delimited.

Once users have examined the text within a .tsv file, users should copy the complete file in the text area and paste it into the provided **Input Genotype for a Panel** function (***Exhibit 18***). Users should replace NN adjacent to each SNP with the corresponding genotype observed in the individual.

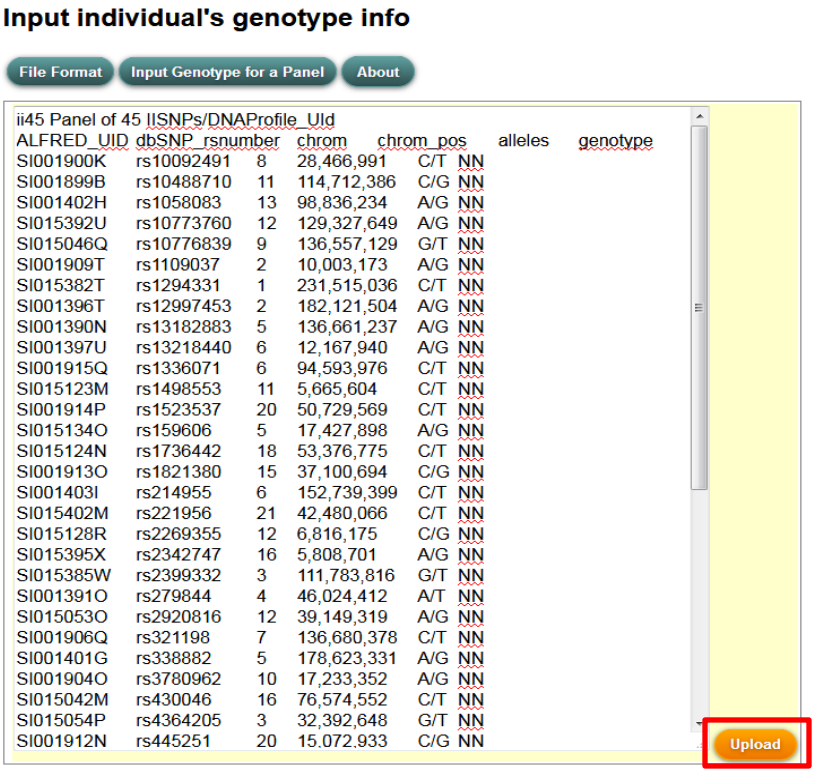



**Exhibit 18. SNP set data pasted into the 'Input Genotype for a Panel' function**

To upload and run the computation, the user should click **Upload** at the bottom of the page. The results from the data entry using the File Upload are provided in the same format as described in **Section 5.2** The Statistical Results

# 6. Search Examples

Users can access a pre-entered data entry page for one individual from a specific population. These pages are dynamic, and users can run the **Compute** function from these pages. The buttons that access these pages are named after different populations (e.g., Korean, Hungarian) ***Exhibit 19*** shows an example of a Korean population for the KiddLab – 45 Unlinked IISNPs in Frog-kb.

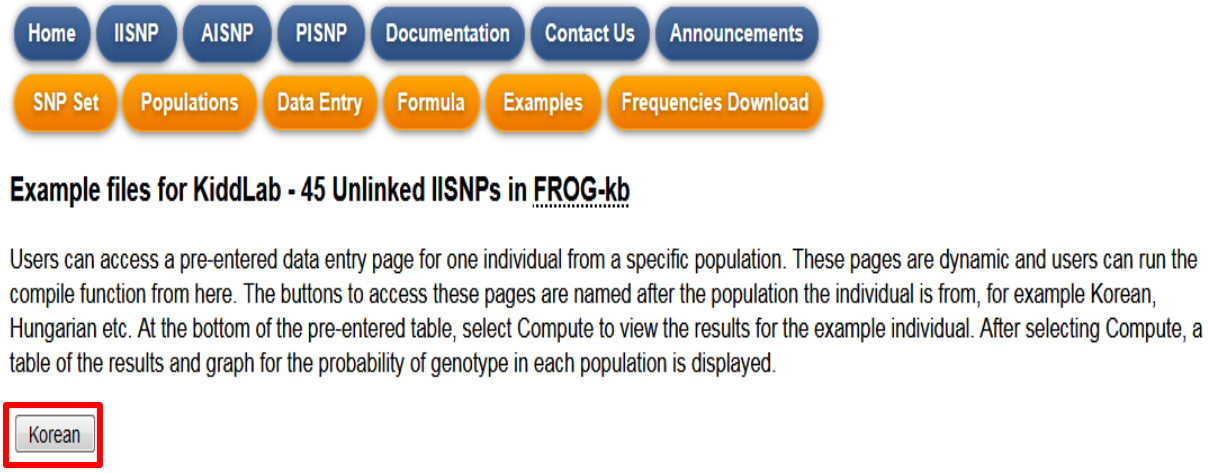



**Exhibit 19. Korean individual example for KiddLab – 45 Unlinked IISNPs in Frog-kb**

Selecting the Korean population in Exhibit 19 provides a data entry table similar to Exhibit 12 but it is pre-entered for the genotypes of one Korean individual. At the bottom of the pre-entered table, select **Compute** to view the results for the example population. After selecting **Compute**, a table of the results and a graph outlining the probability of the genotype in each population is displayed (as described in **Section 5.2** of this User's Manual).

Search examples can be found in each of the SNP panels: IISNP, AISNP, and PISNP. These are good place to start if users have questions on inputting and viewing the resulting data.

# 7. Interpreting the results from FROG-kb

## 7.1 The statistical results

The results table for calculations using any of the Individual Identification (IISNP) panels or Ancestry Inference (AISNP) panels was described above in section 5.2. FROG-kb uses the genotype frequencies for the reference populations and calculates the probability that the target individual's multi-locus genotype could occur in each of the populations. Those reference population genotypes are pre-calculated from the population allele frequencies assuming H-W proportions. For those probabilities to be comparable among the reference populations, the reference populations must have data (allele frequencies) for all of the SNPs in a panel. However, the target individual need not have data for all of SNPs; FROG-kb will use only the reference genotypes for the SNPs present in the target individual. The results of the calculation are displayed as a table with three columns: each line contains 1) the name of the reference population sample with its geographic region and sample size, 2) the probability of the entered genotype occurring in that population, and 3) the likelihood ratio of the highest probability (in the top row) to the probability in specific reference population (Exhibits 13 and 20). The populations are ordered from highest to lowest by their probabilities of generating the entered genotype.

## 7.2 Random Match Probability

The probability of the entered genotype is equivalent to the probability of a random match probability (RMP) assuming no deviation from Hardy-Weinberg ratios in the population. The results for one of the IISNP panels provide an indication of how rare the genotype is globally. (Examples 13 & 14). The largest value, the one listed at the top, provides an upper bound for the RMP among the populations tested. The relative likelihoods for the populations do not generally provide useful ancestry information because the SNPs in the IISNP panels were generally chosen to have similar allele frequency values around the world. The results for an AISNP panel can also be interpreted simply as an indication of the upper bound for the RMP among the reference populations.

## 7.3 Inference of ancestry

The probability calculated for each reference population can be considered to be the likelihood that the specific population is the origin of the entered genotype. If ancestry inference is the objective, the absolute value has no meaning; only the relative likelihoods are meaningful. The population with the highest probability is the most likely ancestral population among the set of reference populations. Note, the probability calculated does not consider multiple ancestries for an "admixed" individual, only ancestry from individual reference populations.

Multiple factors affect how one interprets the results from FROG-kb for one of the AISNP panels. Many are discussed in Kidd KK, 2016 and those issues can be different for different sets of SNPs.

One of the first is that none of the panels contains reference populations truly representative of the human species. The inference of biogeographic ancestry for an unknown DNA sample (individual) can only be as good as the global coverage of the reference populations. If the true population of origin is not among the reference populations, the results cannot identify it. If the unknown comes from one of closely related reference populations, any distinction is questionable *a priori* because the true population of origin may not be the most likely or significantly different from the most likely. Note, in Exhibit 20 the Hungarian population ranks sixth for this Hungarian individual.

Using the likelihood framework makes it clear that the "most likely" may not be meaningfully different from other highly likely populations. A very relevant point is that there is a finite probability of the unknown genotype arising in almost every population in the world. Thus, the "most likely" is simply that, the most likely, and others are less likely to extremely unlikely. If the likelihood ratio among the more likely populations is within a factor of 10 of the most likely, there is no meaningful basis for distinguishing among those potential ancestral populations. Even a ratio of up to 100 includes populations that cannot be meaningfully excluded from possibly being ancestral for the specific genotype. At some value a likelihood ratio of greater than 100 becomes so unlikely as to be excludable. Understanding how these likelihood ratios relate to the number and value of markers is easy if one uses one of the examples for a panel to calculate the most likely population of origin but examines the populations with likelihood ratios less than 100. After changing a significant (say half or more) number of SNPs to not tested (NN) and then recalculating, one usually sees that more populations now have likelihood ratios less than 100 and hence are not highly unlikely or excluded.

With these considerations in mind the interpretation of the results in Exhibit 20 is relatively simple. The Hungarian individual has a multi-locus genotype that is common in Europe, especially in more northern populations. Were the origin of this individual not known, it would be impossible to confidently specify any of the seven most likely populations of origin—the likelihood ratios are all less than one order of magnitude. There are an additional six European populations that are less likely, but cannot be excluded. Finally, there are populations with likelihood ratios of more than two orders of magnitude worse (starting with the Chuvash) and could be considered highly unlikely to be the source of this individual. But note, the 1000 Genomes sample of Finns has a likelihood ratio of 468 compared to the Danes; also, a separate sample of Finns has the third highest likelihood. Exhibit 21 shows the log likelihoods can be as low as -70 (for Biaka Pygmies) as populations more distant from Europe are considered as possible origins of the Hungarian Example individual.

The fact that the "most likely" ancestral population cannot be interpreted as the true ancestral population may be easier to understand when one considers the fact that the SNPs being used are

polymorphic and hence different individuals in the same population will have different genotypes.

**KiddLab - Set of 55 AISNPs**

Population likelihoods based on 55 SNPs and 139 reference populations for : Hungarian Example

Print Table Format | View SNPs Used

● **Indicates the values are within an order of magnitude of the highest likelihood.**

| Population (Region,Sample Size 2N) | Probability of Genotype in each Population | Likelihood Ratio |
|---|---|---|
| Danes (Europe,102) | ● 4.563E-12 | |
| Irish (Europe,232) | ● 2.627E-12 | 1.74 |
| Finns (Europe,72) | ● 1.985E-12 | 2.3 |
| Mixed Europeans(CEU) (Europe,198) | ● 1.922E-12 | 2.37 |
| Russians_Archangel'sk (Europe,68) | ● 1.656E-12 | 2.76 |
| Hungarians (Europe,184) | ● 8.239E-13 | 5.54 |
| Russians (Europe,96) | ● 4.705E-13 | 9.7 |
| Komi-Zyrian (Asia,94) | 2.266E-13 | 20.1 |
| British(GBR) (Europe,182) | 1.799E-13 | 25.4 |
| Mixed Europeans (Europe,190) | 1.25E-13 | 36.5 |
| Toscani(TSI) (Europe,214) | 1.241E-13 | 36.8 |
| Iberian(IBS) (Europe,214) | 6.889E-14 | 66.2 |
| Chuvash (Europe,84) | 3.519E-14 | 130.0 |
| Ashkenazi Jews (Europe,166) | 2.696E-14 | 169.0 |
| Adygei (Europe,108) | 2.048E-14 | 223.0 |
| Turkish (Asia,154) | 1.887E-14 | 242.0 |
| Finns(FIN) (Europe,198) | 9.748E-15 | 468.0 |
| Iranians (Asia,88) | 6.138E-15 | 743.0 |
| Turkish Cypriots (Europe,120) | 5.925E-15 | 770.0 |
| Greeks (Europe,104) | 3.046E-15 | 1500.0 |
| Roman Jews (Europe,54) | 2.126E-16 | 21500.0 |
| Sousse_Tunisia (Africa,98) | 6.1E-17 | 74800.0 |

**Exhibit 20. The top of the table of rank-ordered populations for ancestry inference of the Hungarian Example individual for the KiddLab panel of 55 SNPs.**

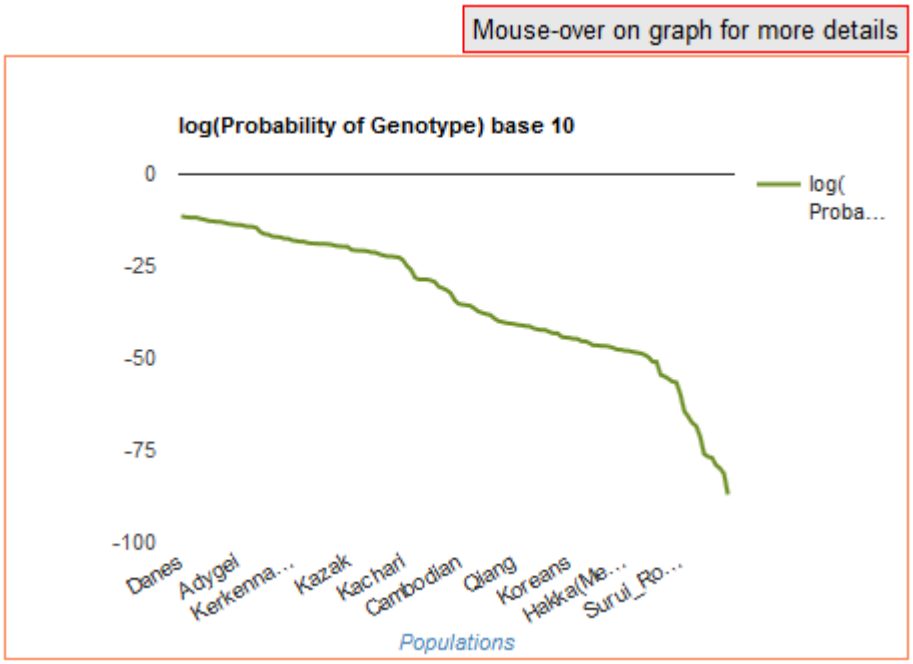



**Exhibit 21. The graph of log likelihoods for the Hungarian Example individual for the KiddLab panel of 55 SNPs.**

# 8. Formulas and Explanations used to Derive Probabilities

## 8.1 Derived Probabilities for IISNP and AISNP Panels

Calculations for IISNP and AISNP panels assume Hardy-Weinberg proportions of the genotypes based on the allele frequencies available for each SNP in each population. This probability of each genotype is stored in FROG-kb after being pre-calculated from the allele frequencies in ALFRED.

Thus, the probability of each genotype at one locus is given as

A.) $P_r$(homozygous allele 1) = (frequency allele 1) $^2$ = $p^2$
B.) $P_r$(heterozygous) = 2*(frequency allele 1)*(frequency allele 2) = 2pq
C.) $P_r$(homozygous allele 2) = (frequency allele 2) $^2$ = $q^2$

where “p” is the frequency of allele 1 and “q” is the frequency of allele 2 in the population.

The probability of a specific multi-locus genotype in a specific population is then simply the product across all loci of the locus-specific genotype probabilities, known as the product rule in forensics. The program implements this basic rule of probability by using the genotype probabilities corresponding to the genotypes entered by the user. The calculation is repeated for each population, and the population-specific probabilities are printed and graphed with the

populations. The probabilities are ordered from highest to lowest. If there are no input data for one locus, that locus is skipped in the calculations for all populations.

For the IISNP panels, these are the population-specific match probabilities. For the AISNP panels, these are the probabilities of the input, multi-locus genotype arising in the different populations and can be considered the relative likelihoods of the populations being the origin of the multi-locus profile.

Special consideration is needed for those situations in which one allele is not observed in a population. The simple estimate of the allele frequency is zero, and two genotypes would strictly be estimated to be zero (i.e., absent in the population). Given the sample sizes involved, an allele at a frequency of less than 1% might never be seen, and one can never prove the absence of an allele. Entering a zero into the multiplication would clearly lead to an incorrect conclusion that the genotype could not exist in that population. A very small allele frequency needs to be assumed to be conservative. Taking sample size into account, one can assume that the next individual sampled would be the first heterozygote for the previously unseen allele. Then the allele frequency estimate would be 1/(2n+2), where n is the original sample size in which the allele was not seen. That assumption has been made in the pre-calculation of the genotype probabilities stored in the FROG-kb database.

## 8.2 Derived Probabilities for PISNP Panel

The eye color of each individual can be probabilistically predicted based on his or her genotypes and the derived alpha and beta. The probabilities of each individual being brown ($\pi_1$), blue ($\pi_2$), and otherwise ($\pi_3$) is

$$\pi_1 = \frac{\exp(\alpha_1 + \sum \beta(\pi_1)_k x_k)}{1 + \exp(\alpha_1 + \sum \beta(\pi_1)_k x_k) + \exp(\alpha_2 + \sum \beta(\pi_2)_k x_k)}$$

$$\pi_2 = \frac{\exp(\alpha_2 + \sum \beta(\pi_2)_k x_k)}{1 + \exp(\alpha_1 + \sum \beta(\pi_1)_k x_k) + \exp(\alpha_2 + \sum \beta(\pi_2)_k x_k)}, \text{ and}$$

$$\pi_3 = 1 - \pi_1 - \pi_2 .$$

where $x_k$ is the number of the alleles used in the calculation of the $k^{th}$ SNP.

The model parameters alpha and beta were derived based on 3,804 Dutch individuals in the model-building set of a previous study (Lui et al., 2009). These probabilities can be calculated using the macro provided in the link ***here*** (Walsh et al., 2011). ***Exhibit 20*** shows the allele information relative to the forward strand and is used in the probability calculations provided in

the macro.

| rsnumber | ancestral/derived | allele used in calculation |
|---|---|---|
| rs12913832 | A/G | A |
| rs1800407 | C/T | T |
| rs12896399 | G/T | G |
| rs16891982 | C/G | C |
| rs1393350 | G/A | A |
| rs12203592 | C/T | T |

**Exhibit 20. Alleles relative to forward strand used in probability calculations**

# 9. References


Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC, Kayser M. Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol. 19*(5):192- 193. (2009)

Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M. IrisPlex: a sensitive DNAtoolfor accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic *Sci Int Genet 5*:170-80. (2011)

Kidd KK. Thoughts on Estimating Ancestry. *Chapter 7 In:* Handbook of Forensic Genetics: Biodiversity and Heredity in Civil and Criminal Investigation. Eds. Antonio Amorim and Bruce Budowle - Vol 2, 131-144. (2016) Series Title - Security Science and Technology Series