RelSearch version 1.0.0 user manual

Contents

1 Initial setup

1. Ensure that R (>= 4.4.0) is installed. It is available from the R Development Core Team website.
2. Begin an R session.
3. Execute the following command in R to install required packages.

```
install.packages(c("shiny", "data.table", "dplyr", "DT", "magrittr",
    "pandoc", "pedtools", "Rcpp", "ribd", "rmarkdown", "shinyFeedback",
    "shinyjs", "shinythemes", "waiter"))
```

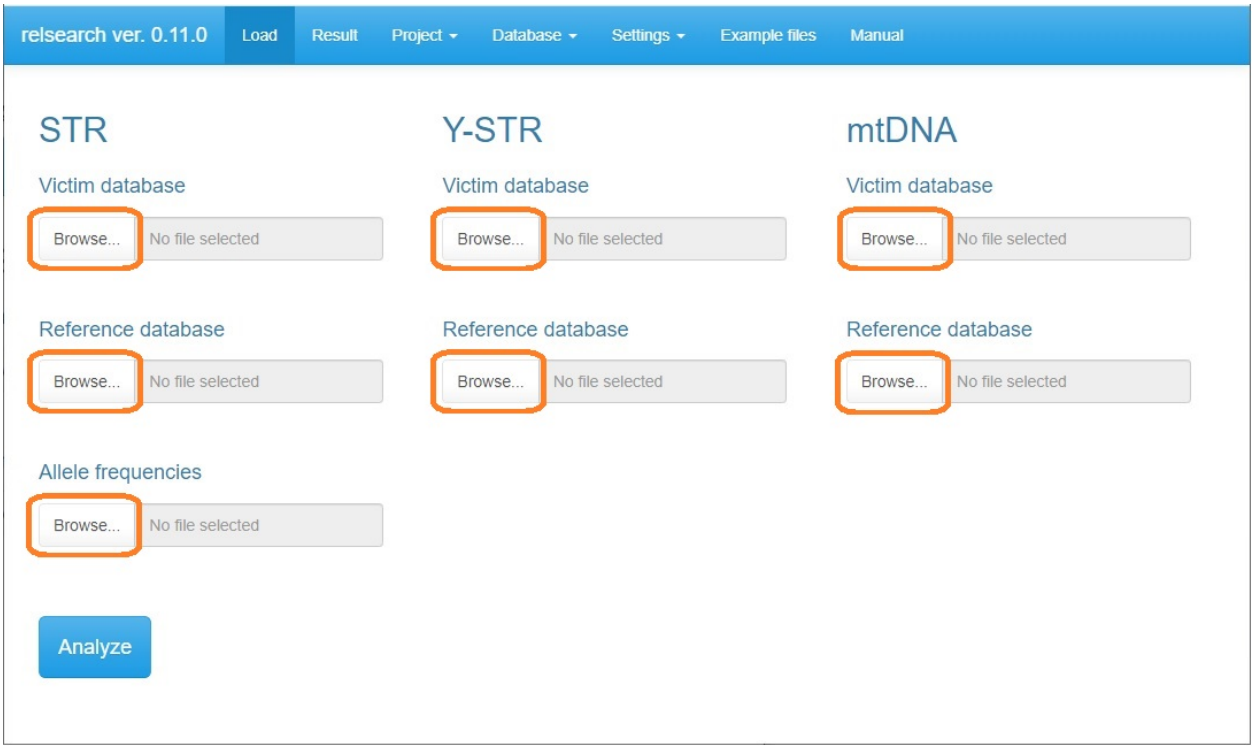
4. Download "RelSearch_1.0.0.zip" from the GitHub repository page.
5. Install "RelSearch_1.0.0.zip" from "Install package(s) from local files…".

2 Quick guide

1. Execute the following commands in R to start GUI.

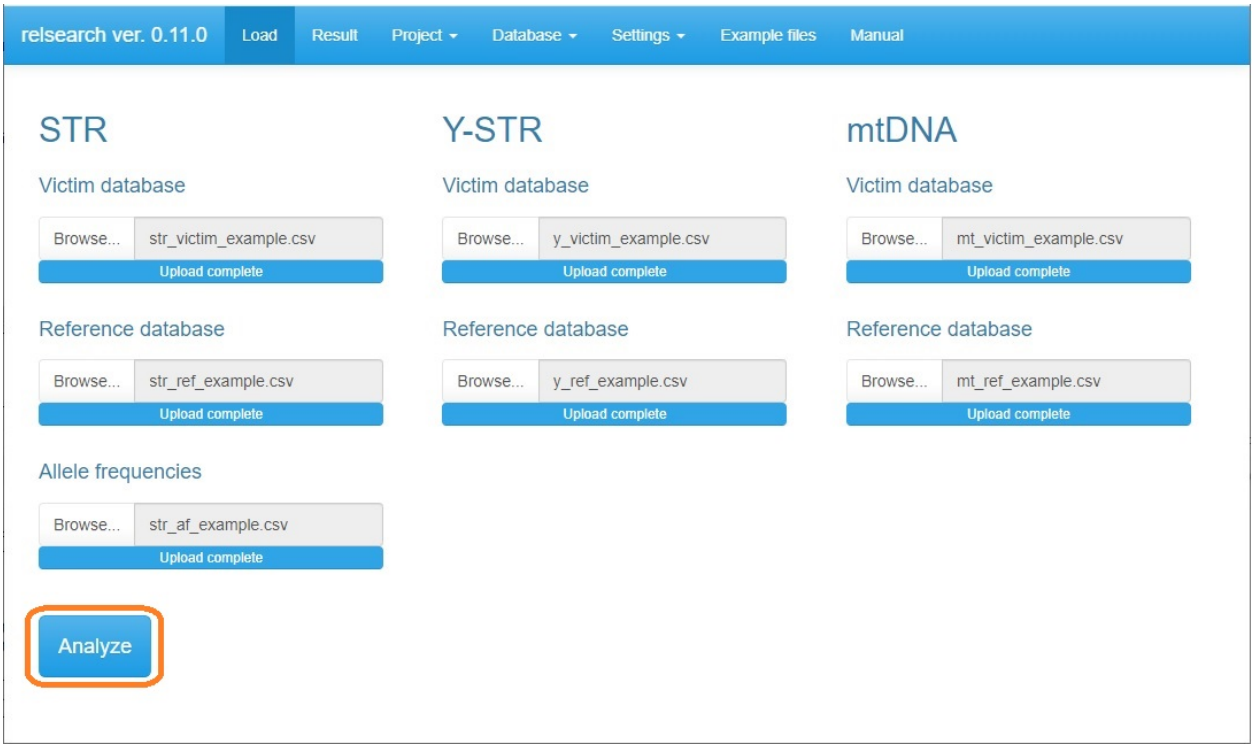
```
library(RelSearch)
RelSearch()
```

2. Load files from each **"Browse..."** button.

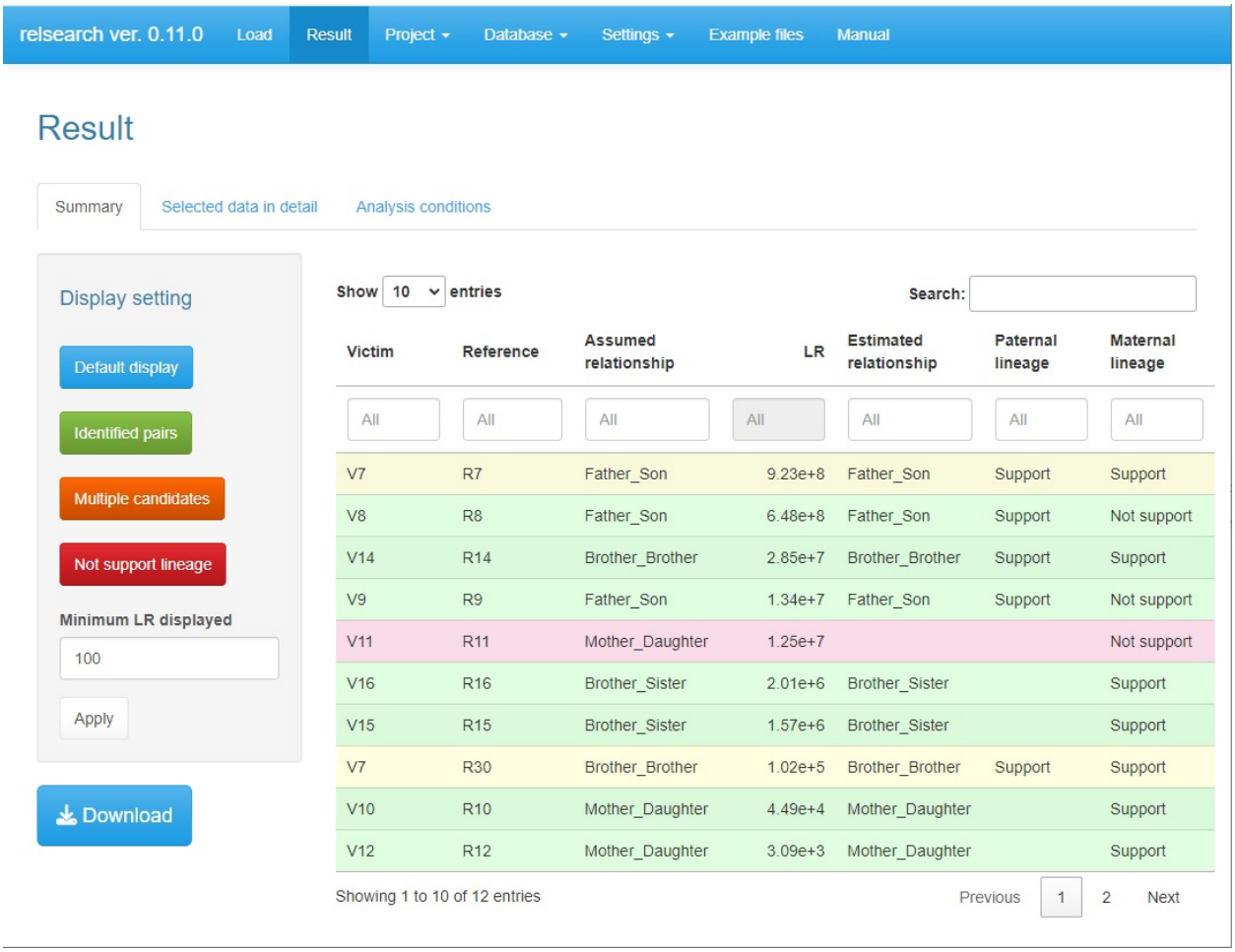


Note See section Files for information on each file.

3. Click the **"Analysis"** button.

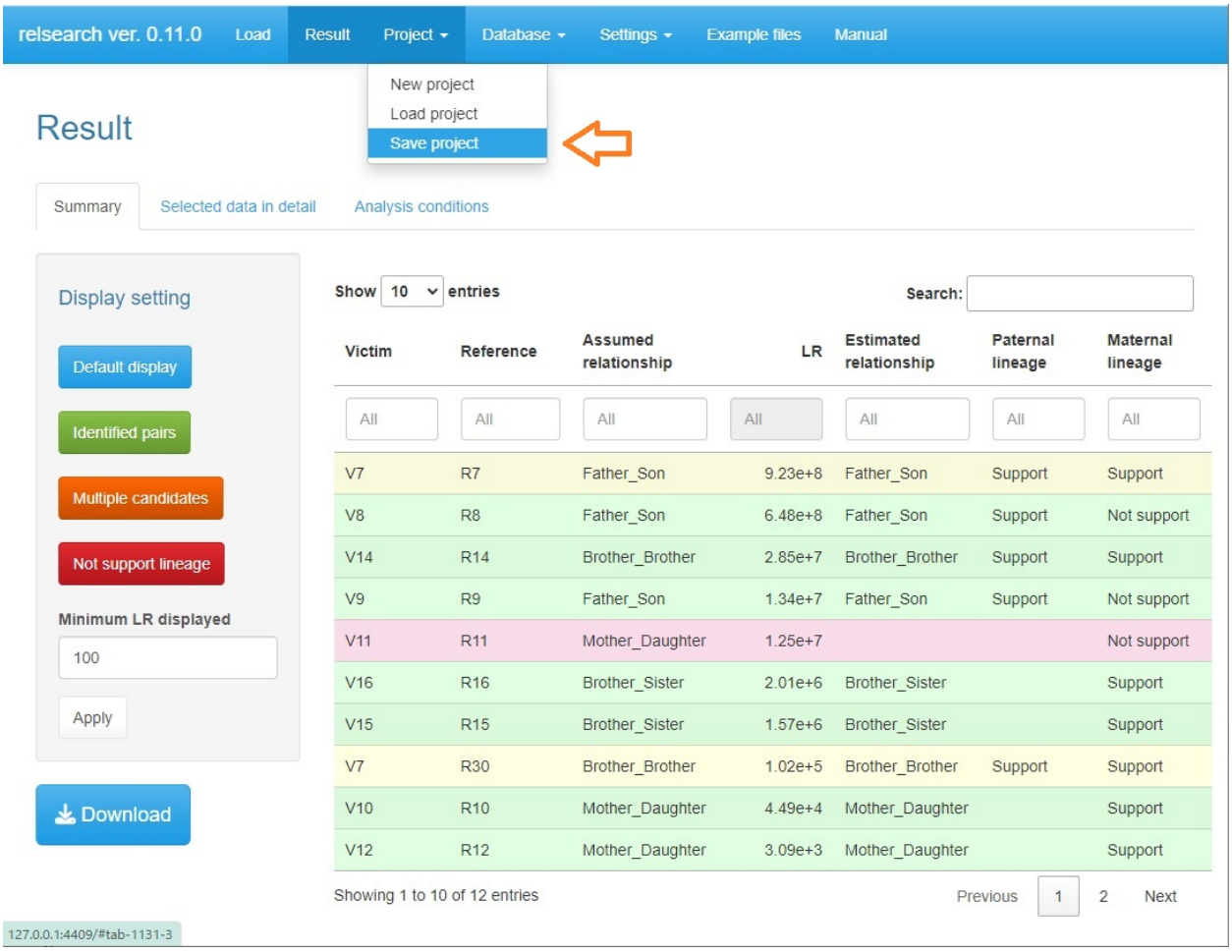


4. After finishing the analysis, the result window is automatically opened.

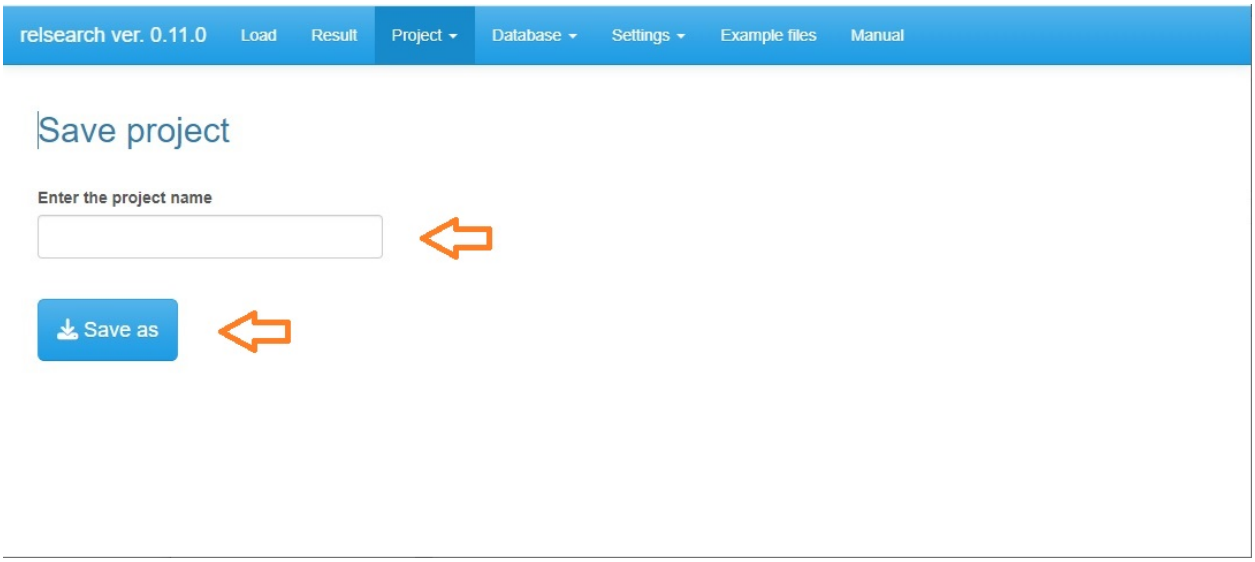


Note See section Result for navigation of the result window.

5. Select **Project ▶ Save project**.



6. Enter the project name and click the **“Save as”** button.



Note The saved project can be loaded from **Project ▶ Load project**. How to handle projects is described in section Other functions.

3 Files

3.1 STR: Victim database

SampleName	D3S1358	D3S1358	vWA	vWA	D16S539	D16S539	CSF1PO	CSF1PO
Victim1	17	17	17		11		12	
Victim2	17		18	18			10	11
Victim3	15	17	17	18	10	10	10	12
Victim4	15	18	15	18	9	12	11	12
Victim5	15	15	18		9	9	10	

Note

- File type: .csv
- This file requires the column "SampleName" and columns for each marker (two columns in each).
- The marker with two empty cells (e.g., D16S539 of the sample "Victim2" in the above table) is ignored when calculating the likelihood ratio.
- The marker with one empty cell (e.g., vWA of the sample "Victim1" in the above table) can be regarded as both homozygote (i.e., no drop-out) and heterozygote with drop-out of one allele when calculating the likelihood ratio.

3.2 STR: Reference database

SampleName	Relationship	D3S1358	D3S1358	vWA	vWA	D16S539	D16S539	CSF1PO	CSF1PO
Reference1	Father_Son	17	17	14	17	10	10	12	12
Reference2	Brother_Brother	16	17	17	17	10	12	13	13
Reference3	Brother_Brother	16	17	14	18	9	9	12	15
Reference3	Father_Son	16	17	14	18	9	9	12	15
Reference4	Brother_Brother	15	16	17		9	12	11	12
Reference5	Uncle_Nephew	15	15	14	16	10	11		

Note

- File type: .csv
- This file requires the columns "SampleName", "Relationship", and columns for each marker (two columns in each).
- The names of the relationship need to be pre-defined in **Settings ▶ Relationships** (see section Setting).

- If a reference has multiple missing family members, add rows for each relationship of the members (e.g., Brother_Brother and Father_Son of the sample "Reference3" in the above table).
- The marker with two empty cells (e.g., CSF1PO of the sample "Reference5" in the above table) is ignored when calculating the likelihood ratio.
- The marker with one empty cell (e.g., vWA of the sample "Reference4" in the above table) can be regarded as both homozygote (i.e., no drop-out) and heterozygote with drop-out of one allele when calculating the likelihood ratio.

3.3 STR: Allele frequencies

Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX
10			602	648	100
11			562	620	1072
12	6		517	1267	114
13	3	1	209	208	3
14	79	586	26	52	2
15	1192	78	3	14	

Note

- File type: .csv
- This file requires the columns "Allele" and columns for each marker (one column in each).
- Allele counts in the population database are entered in columns for each marker.

3.4 Y-STR: Victim database

SampleName	DYS392	DYS518	DYS570	DYS437	DYS385
Victim1	11	37	17	14	13,17
Victim2	13	38	19	14	10,20
Victim3	11	37	17	14	14
Victim4	11	37	16	14	13
Victim5			18		13,17

Note

- File type: .csv
- This file requires the column "SampleName" and columns for each marker (one column in each).

- In the marker with more than one allele, each allele must be separated by a comma without any spaces (e.g., DYS385).
- The marker with an empty cell (e.g., DYS392 of the sample 'Victim5') is ignored for analysis.

3.5 Y-STR: Reference database

SampleName	Relationship	DYS392	DYS518	DYS570	DYS437	DYS385
Reference1	Father_Son	13	38	19	14	10,20
Reference2	Brother_Brother	11	37	17	14	14,17
Reference3	Brother_Brother	11	38	16	14	11
Reference3	Father_Son	11	38	16	14	11
Reference4	Brother_Brother	11	38	19	14	13,15
Reference5	Uncle_Nephew	14		18	14	13,18

Note

- File type: .csv
- This file requires the column "SampleName", "Relationship", and columns for each marker (one column in each).
- The names of the relationship need to be pre-defined in **Settings ▶ Relationships** (see section Setting).
- If a reference has multiple missing family members, add rows for each relationship of the members (e.g., Brother_Brother and Father_Son of the sample "Reference3" in the above table).
- In the marker with more than one allele, each allele must be separated by a comma without any spaces (e.g., DYS385).
- The marker with an empty cell (e.g., DYS518 of the sample 'Reference5') is ignored for analysis.

3.6 mtDNA: Victim database

SampleName	Range	Haplotype
Victim1	73-340 16024-16365	16183C 16189C 16217C 16311C 73G 263G 309.1C 315.1C
Victim2	73-340 16024-16365	16093C 16114A 16223T 16362C 73G 191.1A 194T 263G 309.1C 315.1C
Victim3	73-167 240-340 16024-16365	16095T 16189C 16223T 16265C 16274A 16362C 73G 143A 152C 263G
Victim4	73-265 16024-16284	16223T 73G 152C 263G
Victim5		

Note

- File type: .csv
- This file requires the column "SampleName", "Range", and "Haplotype".
- In the 'Range' column, the first and the last positions in each readable sequence must be written with a hyphen between these positions (e.g, 73-340). Readable sequences must be separated by a single blank space.
- In the 'Haplotype' column, the mtDNA types are expressed as the differences to a reference sequence such as the revised Cambridge Reference Sequence [1] (e.g., 73G, 315.1C, and so on). Each mtDNA type must be separated by a single blank space.
- Empty cells in the 'Range' and 'Haplotype' columns (e.g., the sample 'Victim5') mean that there is no readable sequence.

3.7 mtDNA: Reference database

SampleName	Relationship	Range	Haplotype
Reference1	Mother_Daughter	73-340 16024-16365	16223T 16319A 16362C 73G 152C 263G 309.1C 315.1C
Reference2	Mother_Daughter	73-340 16024-16365	16172C 16223T 16250T 16257A 16261T 73G 150T 263G
Reference3	Brother_Brother	73-340 16024-16365	16172C 16189C 16223T 16355T 16362C 73G 150T 263G
Reference3	Father_Son	73-340 16024-16365	16172C 16189C 16223T 16355T 16362C 73G 150T 263G
Reference4	Sister_Sister	73-340 16024-16365	16223T 16362C 73G 263G 315.1C
Reference5	Aunt_Niece		

Note

- File type: .csv
- This file requires the column "SampleName", "Relationship", "Range", and "Haplotype".
- The names of the relationship need to be pre-defined in **Settings ▶ Relationships** (see section Setting).
- If a reference has multiple missing family members, add rows for each relationship of the members (e.g., Brother_Brother and Father_Son of the sample "Reference3" in the above table).
- In the 'Range' column, the first and the last positions in each readable sequence must be written with a hyphen between these positions (e.g, 73-340). Readable sequences must be separated by a single blank space.
- In the 'Haplotype' column, the mtDNA types are expressed as the differences to a reference sequence such as the revised Cambridge Reference Sequence [1] (e.g., 73G, 315.1C, and so on). Each mtDNA type must be separated by a single blank space.

- Empty cells in the ‘Range’ and ‘Haplotype’ columns (e.g., the sample ‘Reference5’) mean that there is no readable sequence.

4 Setting

4.1 Criteria

You can set the following criteria to support the assumed relationship. Press the **Save** button to reflect the changes. Press the **Reset** button if you want to return to default values.

4.1.1 STR

- Minimum LR: The minimum likelihood ratio that supports the assumed relationship.

4.1.2 Y-STR

- Maximum number of mismatched loci: The upper limit of the number of mismatched loci between victim and reference profiles to support paternal lineage.
- Maximum total mutational steps: The upper limit of total mutational steps (i.e., sum of mutational steps of all loci) between victim and reference profiles to support paternal lineage.

4.1.3 mtDNA

- Maximum number of inconsistency: The upper limit of inconcistency between victim and reference profiles to support maternal lineage.

4.2 Relationship

The names of the relationship in each reference profile need to be pre-defined. There are 44 pre-defined victim-reference relationships as the default.

Information on the table is as follows.

- **Pr(IBD = 2)**: The probability that two alleles are identity by descent (IBD).
- **Pr(IBD = 1)**: The probability that one allele is identity by descent (IBD).
- **Pr(IBD = 0)**: The probability that zero allele is identity by descent (IBD).
- **Paternal lineage**: Information on whether each relationship is paternal lineage or not.
- **Maternal lineage**: Information on whether each relationship is maternal lineage or not.

You can edit, add, and delete the relationships from the left sidebar.

4.2.1 Edit

You can edit the defined relationship name from the **Edit** button.

In the pop-up window, select a relationship from the defined relationships, and enter a new name of the selected relationship. Press the **Save** button to reflect the new name.

4.2.2 Add

You can define a new relationship from the **Add** button.

In the pop-up window, you have to set the relationship name and make a family tree to define the relationship between a victim and a reference.

The following is a list of functions to make a family tree.

- **Add a person**: The function to add an unknown person (UK) to the list of persons.
- **Delete a person**: The function to delete an unknown person (UK) from the list of person.
- **Sex**: The function to select a biological sex (male: M or female: F).
- **Father**: The function to select a person who is the father.
- **Mother**: The function to select a person who is the mother.
- **Founder**: The function to select whether the person is a founder of the family tree or not.

After defining a family tree, press the **View family tree** button to check whether the family tree is correctly set up or not.

If there is no problem in the family tree, press the **Save** button.

Note Information on the IBD probabilities, paternal lineage, and maternal lineage in the new relationship is automatically determined based on your-defined family tree.

4.2.3 Delete

You can delete a defined relationship name from the **Delete** button.

In the pop-up window, select a relationship from the defined relationships. Press the **Save** button to delete the selected relationship.

4.2.4 Reset

If you want to return to default settings, press the **Reset** button. In the pop-up window, press the **Restore default** button.

4.2.5 Family tree

You can check the family tree of each defined relationship. Select a row from the relationship table and press the **Family tree** button.

4.3 Mutation rate

Mutation rates of each autosomal STR marker need to be set to calculate the likelihood ratio (LR). The default mutation rates considering the difference between paternal and maternal and mutational steps (-1, +1, -2, and +2) are reported by Morimoto et al [2]. You can edit, add, and delete these mutation rates from the left sidebar.

4.3.1 Edit

You can edit the mutation rates from the **Edit** button.

In the pop-up window, select a locus and change the mutation rates. Press the **Save** button to reflect the changes.

4.3.2 Add

You can add a locus from the **Add** button.

In the pop-up window, enter a locus name and enter the mutation rates. Press the **Save** button to add the locus.

4.3.3 Delete

You can delete a locus from the **Delete** button.

In the pop-up window, select a locus. Press the **Save** button to delete the selected locus.

4.3.4 Reset

If you want to return to default settings, press the **Reset** button. In the pop-up window, press the **Restore default** button.

4.4 Other parameters

You can set the minimum allele frequency which is used for the unobserved alleles. Press the **Save** button to reflect the changes. Press the **Reset** button if you want to return to the default value.

5 Result

5.1 Summary

After finishing the analysis, summary data is automatically displayed. You can change the displayed data from the left sidebar.

5.1.1 Default display

Default display depends on whether the input data includes autosomal STR data or not.

If the input data includes autosomal STR data, the data where the LR exceeds the minimum LR (see section Setting) is displayed.

If the input data is composed of only Y-STR and/or mtDNA data, the data that satisfies the criteria to support paternal/maternal lineage (see section Setting) is displayed.

If the number of data that satisfies the criteria is more than 10,000, the top 10,000 data is displayed.

5.1.2 Identified pairs

If the data supports that a victim (or a reference) has the assumed relationship with only one reference (or victim), the victim-reference pair is categorized as the identified pair.

5.1.3 Multiple candidates

If the data supports that a victim (or a reference) has the assumed relationship with multiple references (or victims), the victim-reference pair is categorized as the multiple candidates.

5.1.4 Not support lineage

If the LR exceeds the minimum LR but the data does not satisfy the criteria to support paternal/maternal lineage when assuming the paternal/maternal lineage relative, the victim-reference pair is categorized as not support lineage.

5.1.5 Minimum LR displayed

You can change the minimum LR displayed. The lower limit is LR = 1. If the number of data where the LR exceeds the minimum LR is more than 10,000, the top 10,000 data is displayed.

5.1.6 Download summary data

You can download summary data as a .csv file. Press the **Downlaod** button in Summary tab.

5.2 Selected data in detail

The data of a selected victim-reference pair is displayed in **Selected data in detail** tab.

1. Select a row in the summary table.
2. Go to **Selected data in detail** tab.

In **Selected data in detail** tab, you can check and download the analyzed data of STR, Y-STR, and mtDNA in the selected victim-reference pair.

5.3 Analysis conditions

You can check the following analysis conditions in **Analysis conditions** tab.

- **Database**: Input .csv file names of each database are displayed.
- **Allele probability**: Unobserved alleles in the population database are displayed, if any. The frequency of these alleles is set to the user-defined minimum allele frequency and the sum of the frequencies in each locus is corrected to 1.
- **Criteria**: The user-defined criteria to support the assumed relationship (see section Setting) is displayed.
- **Assumed relationship**: Information on the user-defined relationships (see section Setting) is displayed.
- **Mutation rate**: The user-defined mutation rates (see section Setting) are displayed.
- **Parameter**: The user-defined minimum allele frequency (see section Setting) is displayed.

6 Other functions

6.1 Project

6.1.1 New project

To start a new project, go to **Project ▶ New project** and press the **New project** button.

Important: The current project should be saved before starting a new project because the current project will be deleted.

6.1.2 Load project

To load a previous project, go to **Project ▶ Load project**. Select a project file from the **Browse...** button and press the **Load project** button.

Important: The current project should be saved before loading another project because the current project will be deleted.

6.1.3 Save project

To save the current project, go to **Project ▶ Save project**. Enter the project name and press the **Save as** button.

6.2 View database

Each loaded database can be checked in **Database** tab.

6.3 Example files

Example files for each database can be downloaded from the **Example files** tab.

7 Computational principle

7.1 STR

7.1.1 Likelihood ratio

The likelihood ratio (LR) is calculated for each victim-reference pair by assuming the following two hypotheses:

H_1: The victim and the reference are a certain relationship,

H_2: The victim and the reference are unrelated.

The equation of the LR is as follows:

$$LR = \frac{Pr(G_v, G_r|H_1)}{Pr(G_v, G_r|H_2)}$$

where G_v and G_r denote the genotypes of the victim and the reference, respectively.

In RelSearch, users can set the minimum threshold of the LR to support H_1. The default threshold is 100, which is determined based on the verbal scale "Moderate Support" written in the Scientific Working Group on DNA Analysis Methods (SWGDAM) guideline [3].

If full profiles are obtained for a victim-reference pair under the assumption of the relationship other than the parent-child (i.e., without considering drop-out and mutations in calculation), the numerator and the denominator of the LR are calculated using the probabilities that two, one and zero alleles are identity by descent (IBD).

$$\begin{aligned} Pr(G_v, G_r|H) = & Pr(G_v, G_r|IBD = 2)Pr(IBD = 2|H) \\ & + Pr(G_v, G_r|IBD = 1)Pr(IBD = 1|H) \\ & + Pr(G_v, G_r|IBD = 0)Pr(IBD = 0|H) \end{aligned}$$

The specific method for these calculations is based on the method of Wenk et al [4].

7.1.2 Allele probabilities

The allele probabilities are assumed to be based on Dirichlet distributions in each locus [2]. The probability of allele x (p_x) can be estimated using the following formula:

$$p_a = \frac{\alpha_x + 1}{\sum_{x=1}^{X}(\alpha_x + 1)}$$

where α_x is the number of observations of the xth allele ($x = 1, 2, ..., X$) and X is the number of observed allele types.

If some alleles of victims or references are unobserved in the population database, these alleles are added to the observed allele types as zero observation (i.e., $\alpha_x = 0$), and p_x of all alleles are re-estimated [2].

7.1.3 Allelic drop-out

When only one allele is observed and another allele may have dropped out in a locus, RelSearch assumes that the genotype is both homozygous (i.e., no drop-out allele) and heterozygous (i.e., one drop-out allele). The probability of observing the profile is the same in all candidate genotypes. The method corresponds to the "Method 1" proposed by Dørum et al [5].

Here, G_v^* and G_r^* denote the observed profiles of the victim and the reference, respectively. G_{v_i} and G_{r_j} denote the ith candidate genotype of the victim and the jth candidate genotype of the reference, respectively.

Suppose that G_v^* is 15/- (i.e., partial profile) and G_r^* is 16/17. In this case, RelSearch considers the following candidate genotypes of the victim:

G_{v_1}: 15/15,

G_{v_2}: 15/16,

G_{v_3}: 15/17, and

G_{v_4}: 15/Q.

Q denotes any undetected alleles other than 15, 16, and 17.

The candidate genotype of the reference is only 16, 17 because two heterozygote alleles are detected.

In this case, the likelihood function considering allelic drop-out is as follows:

$$
\begin{aligned}
&Pr(G_v^* = 15/-, G_r^* = 16/17|H) \\
&= Pr(G_v^* = 15/-, G_r^* = 16/17|G_{v_i} = 15/15, G_r = 16/17)Pr(G_{v_i} = 15/15, G_r = 16/17|H) \\
&+ Pr(G_v^* = 15/-, G_r^* = 16/17|G_{v_i} = 15/16, G_r = 16/17)Pr(G_{v_i} = 15/16, G_r = 16/17|H) \\
&+ Pr(G_v^* = 15/-, G_r^* = 16/17|G_{v_i} = 15/17, G_r = 16/17)Pr(G_{v_i} = 15/17, G_r = 16/17|H) \\
&+ Pr(G_v^* = 15/-, G_r^* = 16/17|G_{v_i} = 15/Q, G_r = 16/17)Pr(G_{v_i} = 15/Q, G_r = 16/17|H)
\end{aligned}
$$

This equation is generalized as follows:

$$Pr(G_v^*, G_r^*|H) = \sum_i \sum_j Pr(G_v^*, G_r^*|G_{v_i}, G_{r_j})Pr(G_{v_i}, G_{r_j}|H)$$

$Pr(G_v^*, G_r^*|G_{v_i}, G_{r_j})$ can be ignored because the probability of observing G_v^* and G_r^* is the same irrespective of the candidate genotypes. Therefore, the likelihood function can be rewritten as follows:

$$Pr(G_v^*, G_r^*|H) = \sum_i \sum_j Pr(G_{v_i}, G_{r_j}|H)$$

$Pr(G_{v_i}, G_{r_j}|H)$ values are calculated according to equation (1) under the assumption of the relationship other than the parent-child. When assuming the parent-child relationship, mutational events are considered for the calculation as explained in the next session.

7.1.4 Mutation

When assuming the parent-child relationship, mutational events are considered for the calculation of $Pr(G_{v_i}, G_{r_j}|H)$ in equation (2). Suppose that a victim-reference pair is assumed to be parent-child, and G_{v_i} is a/b and G_{r_j} is c/d. RelSearch assumes all inheritance patterns from the parent to the child including mutational events. The LR is calculated according to [6,7].

$$LR = \frac{(\mu_{a\to c} + \mu_{b\to c})p_d + (\mu_{a\to d} + \mu_{b\to d})p_c}{4p_c p_d}$$

where p_a, p_b, p_c, and p_d denote the probabilities of the allele a, b, c, and d, respectively. $\mu_{a\to c}$ denotes the probability that the parent allele a is inherited as the allele c to the child. $\mu_{a\to c}$, $\mu_{b\to c}$, $\mu_{a\to d}$, and $\mu_{b\to d}$ depend on the sex of the parent (i.e., paternal or maternal), and the mutational steps are assumed to be -2, -1, +1, and +2 steps. The mutation rates reported by Morimoto et al [2] are used as the default values.

7.2 Y-STR

RelSearch analyzes whether the Y-STR profiles of each victim-reference pair support paternal lineage. There are two criteria to support paternal lineage as follows:

- Maximum number of mismatched loci: The upper limit of the number of mismatched loci between victim and reference profiles to support paternal lineage.
- Maximum total mutational steps: The upper limit of total mutational steps (i.e., sum of mutational steps of all loci) between victim and reference profiles to support paternal lineage.

If a locus has no alleles in at least one of the victim and the reference, the locus is ignored for the analysis. In addition, if one of the profiles of victim and the reference is partial in a duplicated marker (e.g., DYS385), and the profiles can be explained as paternal lineage (e.g., victim profile: 15, reference profile: 15/16), the locus is also ignored for the analysis.

7.3 mtDNA

RelSearch analyzes whether the mtDNA profiles of each victim-reference pair support maternal lineage. There is one criterion to support maternal lineage as follow:

- Maximum number of inconsistency: The upper limit of the number of inconcistency between victim and reference profiles to support maternal lineage.

Information on the ranges of the mtDNA sequences in each profile is needed because RelSearch targets the ranges read in both victim and reference profiles. The mtDNA profiles are expressed as the differences to a reference sequence such as the revised Cambridge Reference Sequence [1] (e.g., 73G, 315.1C, and so on). RelSearch only recongizes differences in strings of each mtDNA profile; therefore, the rule of the notation for mtDNA profiles should be aligned in each profile by software users.

Reference

[1] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nature Genetics 23 (1999) 147.

[2] C. Morimoto, H. Tsujii, S. Manabe, S. Fujimoto, E. Hirai, Y. Hamano, K. Tamaki, Development of a software for kinship analysis considering linkage and mutation based on a bayesian network, Forensic Science International: Genetics 47 (2020) 102279.

[3] Scientific Working Group on DNA Analysis Methods, Recommendations of the SWGDAM ad hoc working group on genotyping results reported as likelihood ratios., (2018).

[4] R.E. Wenk, M. Traver, F.A. Chiafari, Determination of sibship in any two persons, Transfusion 36 (1996) 259–262.

[5] G. Dørum, D. Kling, C. Baeza-Richer, M. García-Magariños, S. Sæbø, S. Desmyter, T. Egeland, Models and implementation for relationship problems with dropout, International Journal of Legal Medicine 129 (2015) 411–423.

[6] F. Ricciardi, K. Slooten, Mutation models for DVI analysis, Forensic Science International: Genetics 11 (2014) 85–95.

[7] T. Egeland, N. Pinto, A. Amorim, Exact likelihood ratio calculations for pairwise cases, Forensic Science International: Genetics 29 (2017) 218–224.