KinBN v2.0.0 user manual

Contents

1. What is *KinBN*?

KinBN is a free software (GNU General Public License v3.0) for kinship analysis based on the Bayesian network[1]. It can be applied to short tandem repeat (STR) markers commonly used in forensic genetics to calculate the likelihood ratio (LR) considering the linkage between loci, mutation, and drop-out. The software is a graphical user interface written in R language. The software has been validated under various conditions.

Characteristics of the current version of *KinBN* are:
● It is based on the Bayesian network and calculates the LR for kinship analysis.
● It considers the effects of linkage, mutation, and drop-out on kinship determination.
● It facilitates the assessment of complex relationships, such as incest.
● It can be used to simulate the LR distribution of computationally generated DNA profiles based on assumed relationships.
● It generates a report of calculation results that can be saved and checked as required.

KinBN has been developed by Morimoto C et al. of the Department of Forensic Medicine, Kyoto University Graduate School of Medicine. Questions regarding the software should be sent to the following e-mail address: morimoto@fp.med.kyoto-u.ac.jp.

2. Changes in v2.0.0
Added functionalities:
● The software can be used as the R-package *KinBN*.
● The software can calculate the LR value while accounting for allele drop-out.
● Project data can be saved and loaded as required.
Minor changes:
● Layout of the software has changed.
● The user can set the calculation conditions, such as the estimation method of allele frequency, in the Setting tab.

---

[1] Morimoto C et al. Forensic Sci Int Genet., 47, 102279, 2020.

3. Tutorial

A) Starting software

   KinBN is compatible with R language (v.4.3 or v.4.2) available on the R Development Core Team website (http://www.R project.org). The current version of *KinBN* (v2.0.0) is freely available on GitHub (https://github.com/ChieMorimoto/KinBN/releases). The user needs to install the R language prior to using this software.

   The user can initiate an R session and enter the code below the R console to install *KinBN* and other necessary packages.

```
>
install.packages('https://github.com/ChieMorimoto/KinBN/releases/download
/v2.0.0/KinBN_2.0.0.zip',repos=NULL,type='win.binary')
> install.packages(c("tcltk2", "bnlearn", "gRbase", "gRain", "tkrplot",
"kinship2"))
```

Then, the user can enter the codes below to start the graphical user interface (GUI).

```
> library(KinBN)
> KinBN()
```
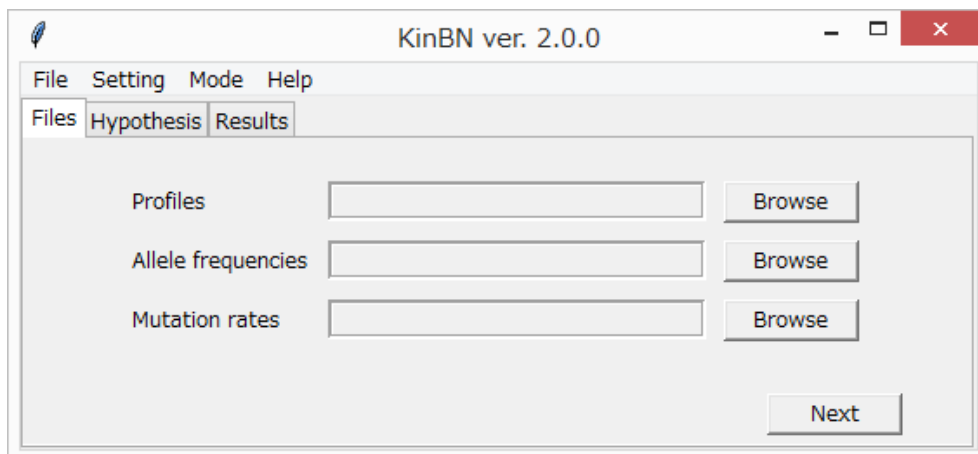
This will open the top screen of *KinBN* (Fig. 1).

Fig. 1 Top screen of *KinBN* available on the Files tab in "case analysis" mode.

B) Preparation of input files

KinBN can be used to perform two modes: case analysis and simulation. For case analysis, the user requires three CSV files: STR profiles of the case (Fig. 2), observed numbers of alleles at each locus in the population (Fig. 3), and mutation rates at each locus (Fig. 4). Any locus set can be accommodated in this software. Locus names should be unified in the files. For simulation, the user requires two CSV files: observed numbers of alleles at each locus in the population (Fig. 3) and mutation rates at each locus (Fig. 4).

The user can import the three input files by clicking on the *Browse* button. Example files are provided in the GitHub repository (https://github.com/ChieMorimoto/KinBN/releases).

	A	B	C	D	E
1	Marker	ID1	ID1	ID2	ID2
2	D3S1358	16	17	16	16
3	vWA	17	17	17	18
4	D16S539	9	9	9	9
5	CSF1PO	12	12	12	13
6	TPOX	8		8	10
7	D8S1179	15	15	10	14
8	D21S11	30	30	30	32.2
9	D18S51	13		13	14
10	D2S441	10	14	10	14
11	D19S433	13.2	14	14	14
12	TH01	7	8	7	8
13	FGA	23	24	22	24
14	D22S1045	11	15	15	16
15	D5S818	11	12	11	12
16	D13S317	12	13	12	13
17	D7S820	11	11	11	11
18	SE33			18	26.2
19	D10S1248	14	15	14	15
20	D1S1656	12	18	17	17.3
21	D12S391	18	19	18	19
22	D2S1338	23	23	23	25

Fig. 2 STR genotypes of two people (ID1 and ID2). The user should enter two alleles for homozygotes in each column (e.g., D16S539 of ID1 and ID2) and leave a blank in loci with possible allele drop-out (e.g., TPOX and D18S51 of ID1). Two blanks indicate a locus drop-out (e.g., SE33 of ID1).

Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA	D22S1045	D5S818	D13S317	D7S820	SE33
2.2																	
3.2																	
4																	
4.2																	
5					1						1						
6					1						170						
6.3																	
7											140			2		20	
8			13	4	379	10					69			4	87	104	
8.1																1	
9			77	10	92	4			1		86			30	56	121	
9.1									1								
9.3											249						
10			41	159	36	74		6	152	1	6			40	34	185	
10.1																	
10.2																	
10.3																	
11	1		227	223	182	55		7	248	4	1		101	257	235	148	1
11.2																	
11.3									44								
12		1	227	260	30	121		82	34	51			9	280	194	115	5
12.2										1							

Fig. 3 Observed numbers of alleles at each locus in the population. This can be used to estimate the allele frequencies. First column and row indicate the allele number and locus, respectively.

Marker	P (-2)	P (-1)	P (0)	P (+1)	P (+2)	M (-2)	M (-1)	M (0)	M (+1)	M (+2)
D3S1358	3.70E-05	0.001424	0.997353	0.001148	3.77E-05	0	0.000193	0.999697	7.56E-05	3.44E-05
vWA	0.000146	0.002718	0.994906	0.002201	2.92E-05	0	0.000125	0.999378	0.000445	5.25E-05
D16S539	8.41E-05	0.000934	0.997629	0.001353	0	0	0.000484	0.999412	0.000104	0
CSF1PO	0	0.002043	0.996111	0.001846	0	0	0.000213	0.999408	0.00038	0
TPOX	0	0.000116	0.999697	0.000187	0	0	7.51E-05	0.999849	7.61E-05	0
D8S1179	4.82E-05	0.001301	0.99743	0.001196	2.39E-05	0	0.000134	0.999702	0.000142	2.22E-05
D21S11	3.70E-05	0.000868	0.998013	0.001082	0	3.60E-05	0.000108	0.999727	0.000129	0
D18S51	0.000147	0.000943	0.997459	0.001421	2.98E-05	2.13E-05	0.000117	0.999364	0.000476	2.15E-05
D2S441	5.44E-05	0.001167	0.997464	0.001251	6.42E-05	1.49E-05	0.00027	0.999494	0.000212	8.71E-06
D19S433	0	0.000958	0.998425	0.000452	0.000164	0	0.000505	0.999435	5.97E-05	0
TH01	0	4.48E-05	0.999895	5.97E-05	0	1.70E-05	5.46E-05	0.99992	8.29E-06	0
FGA	0.000278	0.001497	0.995438	0.002708	7.94E-05	0	0.000348	0.999358	0.000293	0
D22S1045	5.44E-05	0.001167	0.997464	0.001251	6.42E-05	1.49E-05	0.000494	0.999494	0.000212	8.71E-06
D5S818	0	0.001421	0.996841	0.001198	0.000541	0	0.000259	0.999462	0.000279	0
D13S317	0	0.001369	0.997194	0.001397	3.96E-05	3.33E-05	7.08E-05	0.999657	0.000239	0
D7S820	0	0.001303	0.997827	0.00087	0	0	6.22E-05	0.999813	0.000125	0
SE33	5.44E-05	0.001167	0.997464	0.001251	6.42E-05	1.49E-05	0.00027	0.999494	0.000212	8.71E-06
D10S1248	5.44E-05	0.001167	0.997464	0.001251	6.42E-05	1.49E-05	0.00027	0.999494	0.000212	8.71E-06
D1S1656	5.44E-05	0.001167	0.997464	0.001251	6.42E-05	1.49E-05	0.00027	0.999494	0.000212	8.71E-06
D12S391	5.44E-05	0.001167	0.997464	0.001251	6.42E-05	1.49E-05	0.00027	0.999494	0.000212	8.71E-06
D2S1338	3.70E-05	0.000868	0.998013	0.001082	0	3.60E-05	0.000108	0.999727	0.000129	0

Fig. 4 Locus- and sex-specific mutation rates at each locus. First column indicates the locus. P (x) and M (x) indicate the paternal and maternal mutation rates, respectively. x indicates the distance of mutation.

C) Case analysis mode

After importing three files, the user can click *Next* to open the Hypothesis tab and set the hypotheses for LR (Fig. 5).

Fig. 5 Hypothesis tab. After importing the files, names of known profiles are automatically displayed.

In this tab, Hypotheses 1 (H1) and 2 (H2) can be set by manually creating a pedigree. Sex, father, and mother were recorded for each person. *Add* and *Delete* buttons refer to adding and deleting extra persons in the assumed pedigree. If a person is the founder, the user checks *founder*. For example, when the user tests a pairwise sibship between ID1 and ID2, the setting is as shown in Fig. 6. The user can confirm the pedigree using the *View pedigree tree* button. This software restricts the number of family members (up to 25).

Fig. 6 Example of hypotheses settings for pairwise sibship analysis. In hypothesis 1, UK1 and UK2 indicate the extra individuals.

Before the calculation, the calculation conditions can be set from the Setting. Please check section 4.B) for details.

Once the user clicks *Calculate*, the LR calculation begins. The results are shown in the Results tab (Fig. 7).
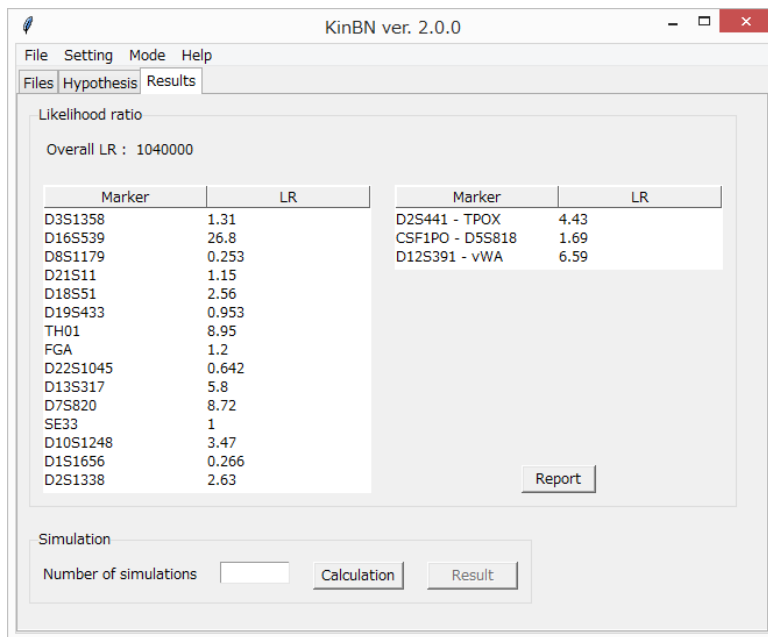
Fig. 7 Calculation result in the case analysis mode. Overall LR indicates the combined value of all loci. LR value of two linked loci is calculated as one value and displayed in the right table.

The user can save a case report (CSV file) from *Report* button. Moreover, by entering the number of simulations, the simulation of the LR values in the assumed relationship can be performed in a manner similar to that of the simulation mode (Fig. 8). If the LR values are 0 and the distribution cannot be drawn, the graph will not be displayed.
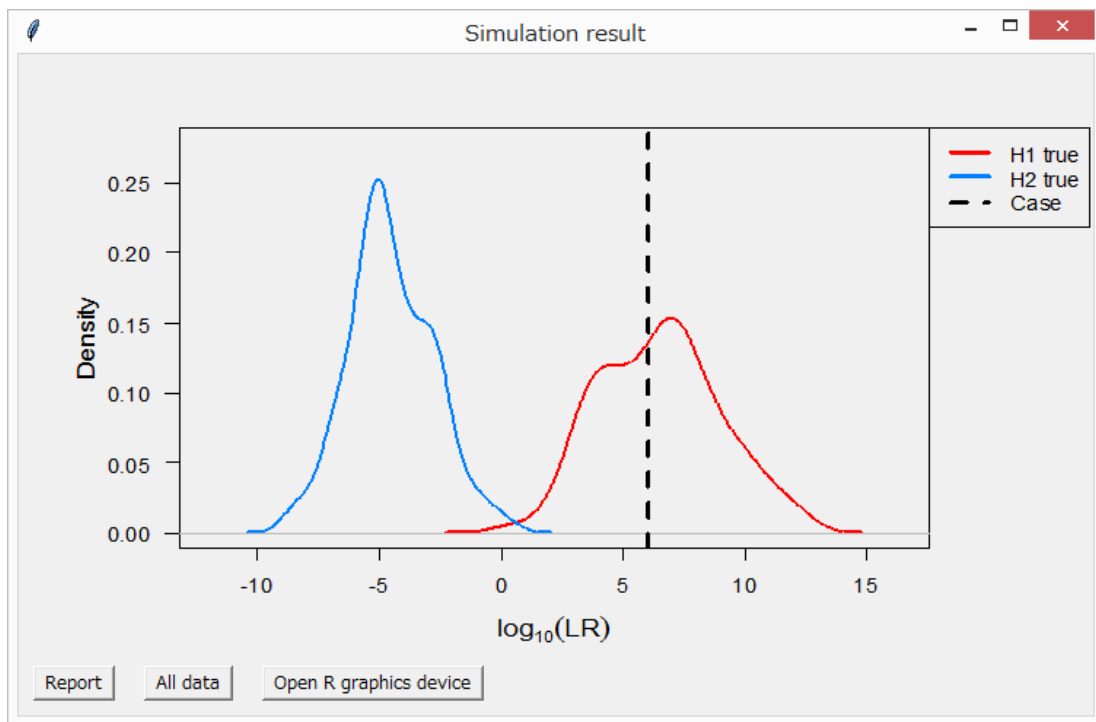
Fig. 8 Simulation result in the case analysis mode. LR distributions of H1 true and H2 true are displayed. Dashed line indicates the LR value in case analysis. The user can save a report of the simulation summary (CSV file) by clicking on the *Report* button. All LR values (CSV file) can also be obtained by clicking on the *All data* button. The entire graph can be saved by clicking on the *Open R graphics device* button.

D) Simulation mode

After the user selects *Simulation* in Mode tool bar, the Files tab is opened (Fig. 8).

Fig. 8 Files tab in simulation mode.

The user imports the two input files from *Browse* buttons. Once the user selects *Next*, Hypothesis tab opens (Fig. 9).
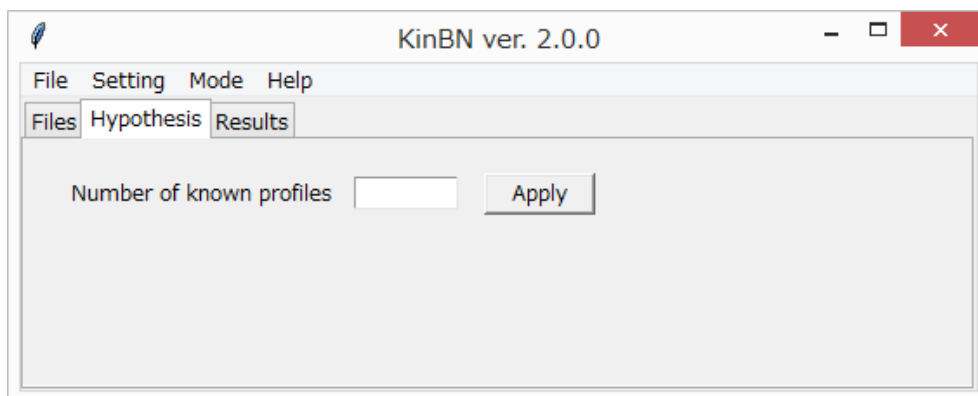
Fig. 9 Hypothesis tab in simulation mode.

In the Hypothesis tab, the user enters the number of known profiles. Once the user clicks *Apply*, the screen for the hypothesis setting is displayed (Fig. 10). As with case analysis, each hypothesis can be set manually by creating a pedigree.

Before the calculation, the calculation conditions can be set from the Setting. Please check section 4.B) for details. However, drop-out is not considered in the simulation mode.

To start the simulation, the user enters the number of simulations and clicks *Calculate* button. Many familial genotypes are computationally generated by considering linkages and mutations according to the user setting. LR values were calculated based on genotypes and provided the distribution under the targeted family tree. The progress bar indicates the state of simulation.

Fig. 10 Hypothesis tab in simulation mode. Here, the number of known profiles shown is two (ID1 and ID2).

The resulting tab was opened after completing the simulation (Fig. 11). Graph of the LR distribution is shown. The user can save a simulation summary (CSV file) from *Report* button. All LR value data can also be obtained from *All data* buttons.
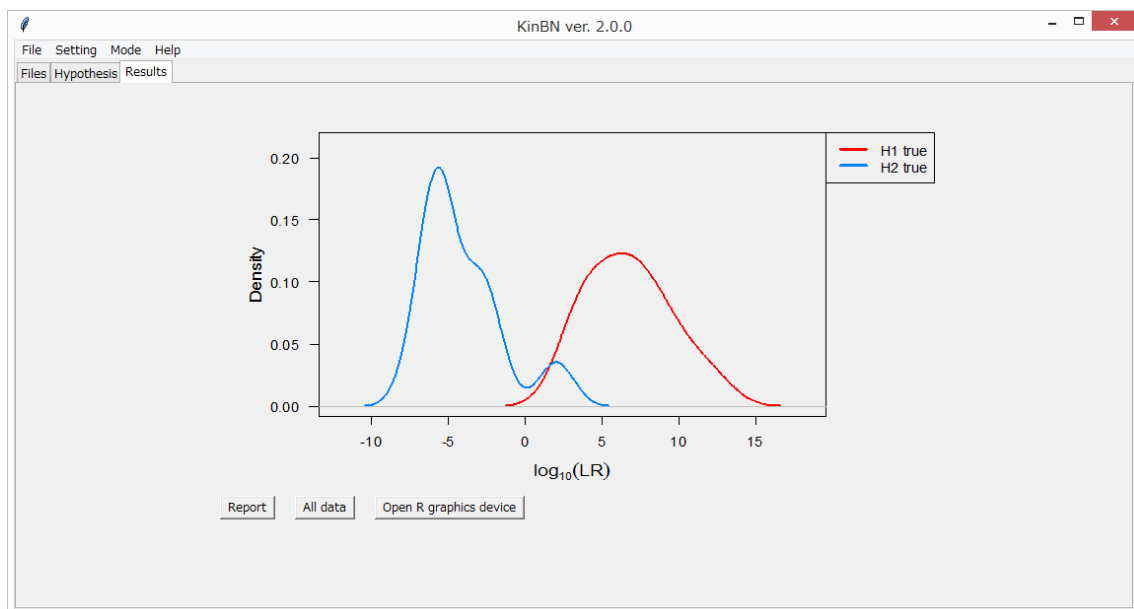

Fig. 11 Calculation result in simulation mode.

4. Tool bar

A) File

KinBN v2.0.0 has four options about project (Fig. 12).

New project: Start new project.

Load project: Select previously saved .RData and load the project.

Save project: Save the project information (input files, hypotheses, and calculation settings).

Quit: Quit R session.

Fig. 12 File options on the tool bar.

B) Setting

KinBN v2.0.0 can be set to three calculation conditions: linkage, drop-out, and allele frequency (Fig. 13).
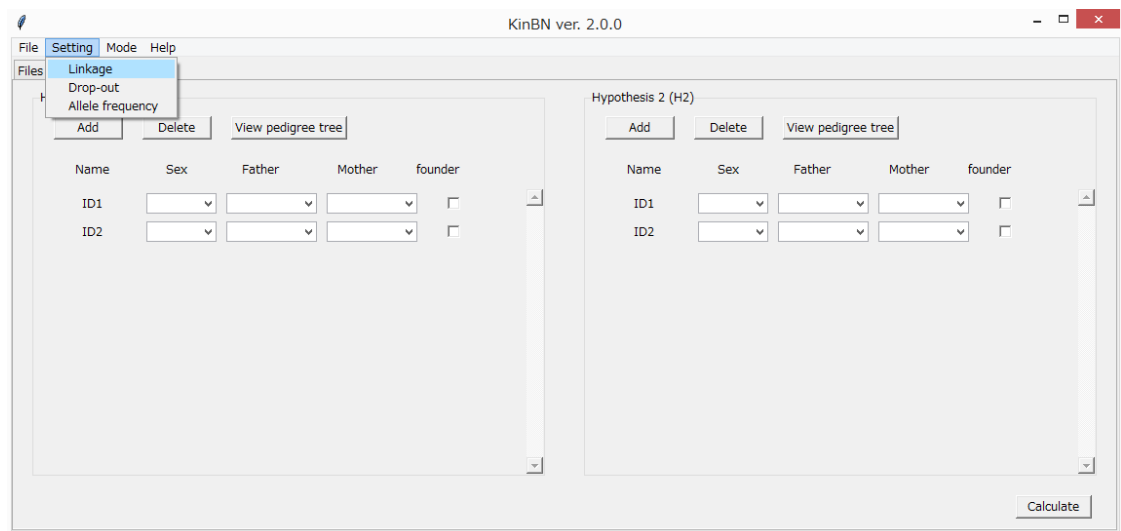
Fig. 13 Setting options on the tool bar.

i. Linkage

The user can select the linkage settings from three modes (Fig. 14).

Default setting:

If the user uses multiplex kits following GlobalFiler (Thermo Fisher Scientific, Waltham, MA, USA), Identifiler (Thermo Fisher Scientific), PowerPlex Fusion (Promega, Madison, WI, USA), PowerPlex 16 (Promega), or NGM SElect (Thermo Fisher Scientific), *Default settings* can be selected.

Manual setting:

The user can freely set the linked loci and recombination rate. The user adds and deletes pairs of linked loci using the *Add* and × buttons, respectively.

No linkage:

If the user assumes independence between loci, *No linkage* mode can be used.

Fig. 14 Linkage setting screen.

ii. Drop-out

The user can select the drop-out settings from three modes if CSV file of STR profiles has the blanks (Fig. 15). Appendix describes the details of the calculation method. In the case of locus drop-out, *KinBN* outputs 1 as LR of the locus whichever mode is selected.

LR considering all possible genotypes equally (Method 1):

For possible drop-out loci, both homozygous and heterozygous genotypes were considered equally when calculating LR.

LR calculated with Pr(D) (Method 2):

Pr(D) represents the probability of allele drop-out. For possible drop-out loci, LR was calculated by accounting for Pr(D). The user should enter Pr(D) value.

User-defined LR:

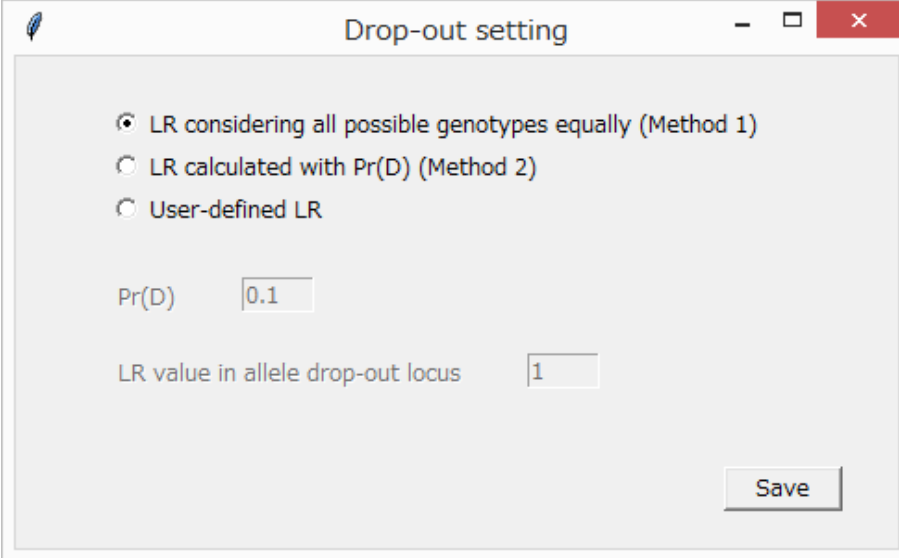For possible allele drop-out loci, user-defined LR values were outputted. The user should enter a fixed value.

Fig. 15 Drop-out setting screen.

iii.   Allele Frequency

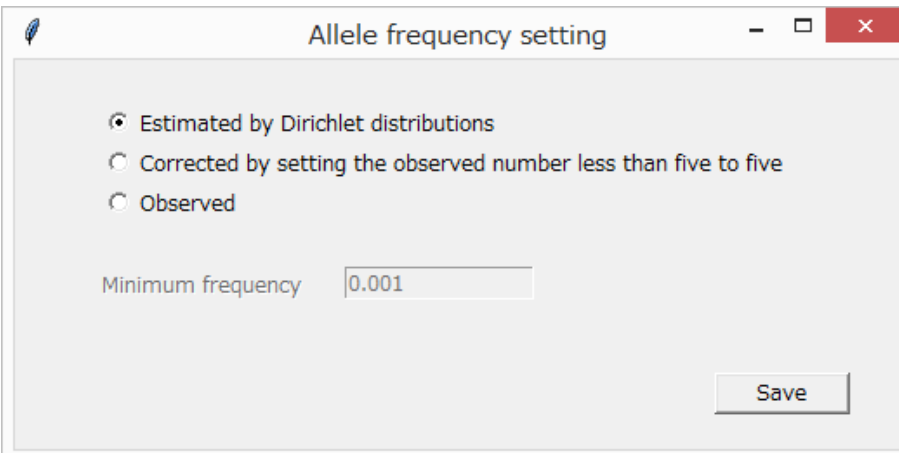The user can select the allele frequency settings from three modes (Fig. 16).

Fig. 16 Allele frequency setting screen. The user can select the allele frequency setting from three modes: *Estimated by Dirichlet distributions*, *Corrected by setting the observed number less than five to five*[2], and *Observed.* In *Observed*, the user can enter the minimum allele

---

[2] This setting is based on that specified by the National Research Council II (1996).

frequency for unobserved alleles.

C) Mode

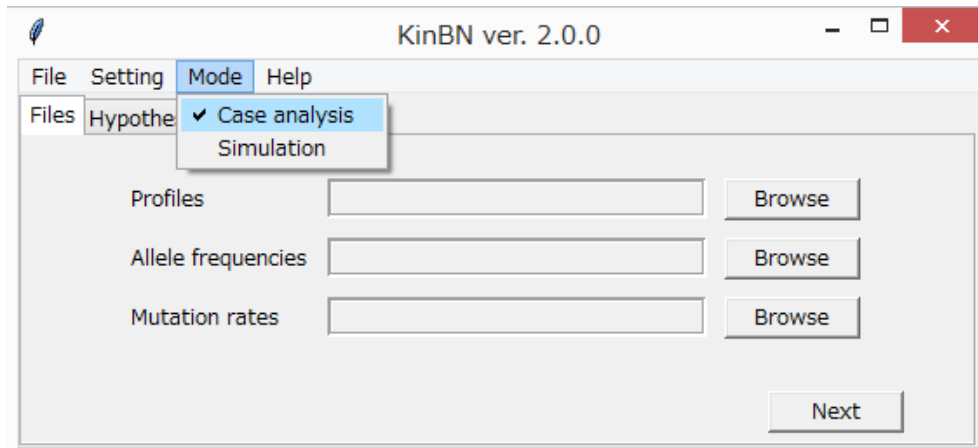The user can switch between two modes: Case analysis and Simulation (Fig. 17).

Fig. 17 Mode options on the tool bar.

D) Help

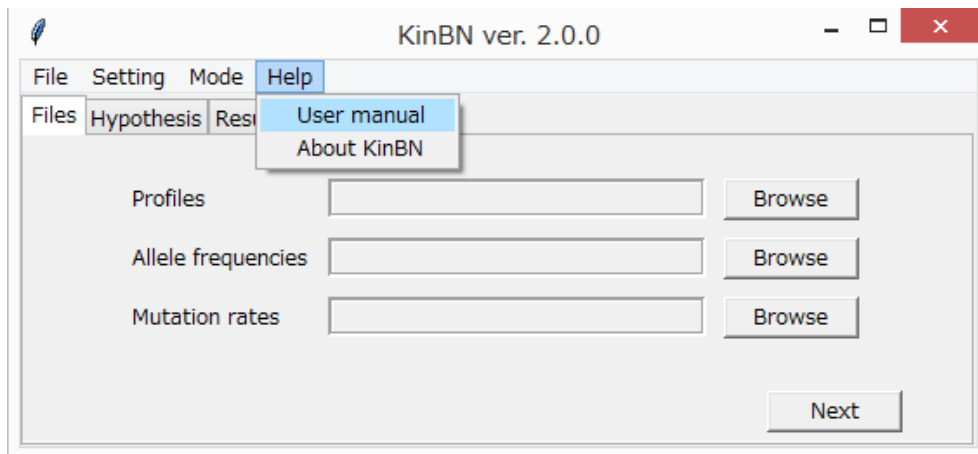The user manual can be checked using *User manual* (Fig. 18).

Fig. 18 Help options on the tool bar.

5. Appendix: Calculation of allele drop-out

KinBN is based on a Bayesian network, which is useful for evaluating multiple probabilistic events, such as mutations and linkages. In v2.0.0, the true STR type (genotype: *gt*) before drop-out occurs and the STR type (drop-out genotype: *dgt*) after drop-out occurs are used as nodes to consider allele drop-out in *KinBN*. A link was drawn from *gt* to *dgt* to represent the occurrence of allele drop-outs. For example, if *gt* is heterozygous for (*a, b*) but allele *b* is not detected (allele drop-out), *dgt* becomes (*a,*).

For example, assume that in paternity duo analysis between the alleged father (AF) and child (C), only allele *a* is detected in AF at the locus where three types of alleles (*a*, *b*, and *c*) can be observed and two alleles, a and b, are detected in C (heterozygote). However, if the peak height of allele *a* of AF is low and the possibility of allele drop-out is assumed, we must consider three possible true STR types (node *gt*) for AF: homozygous for (*a, a*), heterozygous for (*a, b*), and heterozygous for (*a, c*). The three patterns of probabilities considered in the link from *gt* to *dgt* are as follows:

(1) $\Pr\big(AFdgt = (a,)\big|AFgt = (a,a)\big)$: probability that the apparent STR type is (*a,*) when the true STR type is (*a, a*)

(2) $\Pr\big(AFdgt = (a,)\big|AFgt = (a,b)\big)$: probability that the apparent STR type is (*a,*) when the true STR type is (*a, b*)

(3) $\Pr\big(AFdgt = (a,)\big|AFgt = (a,c)\big)$: probability that the apparent STR type is (*a,*) when the true STR type is (*a, c*)

(*a,*) shows the STR results when only allele *a* was detected.

The probabilities of (1), (2), and (3) are related to the incidence of allele drop-out. *KinBN* provides two methods, Method 1 and Method 2, to set the probabilities of (1), (2), and (3).

Method 1 assumes that the probabilities of (1), (2), and (3) are equal, i.e., the probability that the apparent STR type is (*a,*) is equal, regardless of the true STR type of AF.

On the other hand, Method 2 is considering the probability of dropping out an allele for heterozygote as $\Pr(D)$. In case (1), since allele *a* in the true STR type has been observed even apparently, it means that none of the two alleles in *AFgt* has dropped out, and the probability of (1) is $1 - \Pr(D)^2$. In case (2), allele *a* in the true STR type is apparently observed, but allele *b* is not observed. Therefore, one of the two alleles of *AFgt* is dropped out and the other is not dropped out. Therefore, the probability of (2) is $\Pr(D)(1 - \Pr(D))$. The probability of (3) is $\Pr(D)(1 - \Pr(D))$ in the same way as (2).

By incorporating the probabilities set in Method 1 and Method 2 into the calculation of the likelihood ratio, the likelihood ratio assuming allele drop-out can be calculated. If no peak is

observed in a certain locus (locus drop-out), the likelihood ratio for that locus is set to 1, regardless of the difference between Method 1 and Method 2.