KinBN v1.1.0 user manual

Contents

1. What is *KinBN*?

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

The current version of *KinBN* has the following capabilities:

- Based on a Bayesian network, it calculates LR value for kinship analysis.
- It considers the effects of linkage and mutation that influence kinship determination.
- The user can assume complex relationships such as incest.
- Users can also simulate LR distribution of computationally generated DNA profiles according to the assumed relationship.
- *KinBN* generates a report of calculation results. The user can save it and check the results.

KinBN has been developed by Morimoto et al in Department of Forensic Medicine, Kyoto University Graduate School of Medicine. Questions regarding the software should be addressed to morimoto@fp.med.kyoto-u.ac.jp.

2. Changes in v1.1.0

- **Important update: A calculation error in the case of incest was corrected. The user is recommended to switch to version 1.1.0.**
- The user can delete pairs of linked loci freely in linkage tab.
- A bug of simulation mode was fixed.

3. Tutorial

A) Starting software

KinBN is compatible with R language that is available from the R Development Core Team website (http://www.R project.org). The current version of *KinBN* (v1.1.0) is freely available at GitHub repository (https://github.com/ChieMorimoto/KinBN/releases). In the repository, there are R Workspace data including source code (KinBN v1.1.0.Rdata), R source code (KinBN v1.1.0 source code.R), example files (Allele data in Caucasian.csv, Mutation rates.csv, and Example of STR genotypes.csv), and this user manual (KinBN v1.1.0 user manual.pdf).The user needs to install R language and the current version of *KinBN* prior to use the software.

The user opens KinBN v1.1.0.Rdata with R language and enter the below code at R console.

```
>  KinBN()
```

Then the required packages (*bnlearn*, *RGBL*, *BiocManager*, *BiocGenerics*, *graph*, *gRbase*, *gRain*, *tcltk*, *tcltk2*, *tkRplotR*, and *kinship2*) will be installed automatically and the top screen of *KinBN* will be opened (Fig. 1).

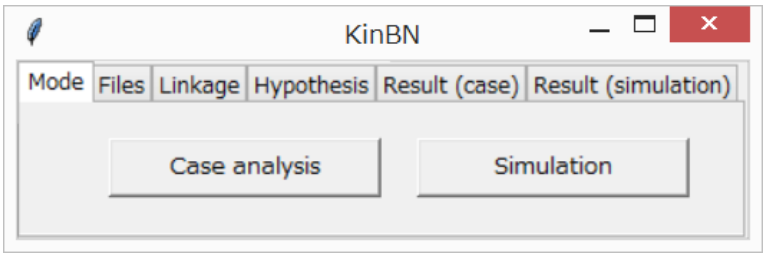


Fig. 1　Top screen of *KinBN*.

B) Preparation of input files

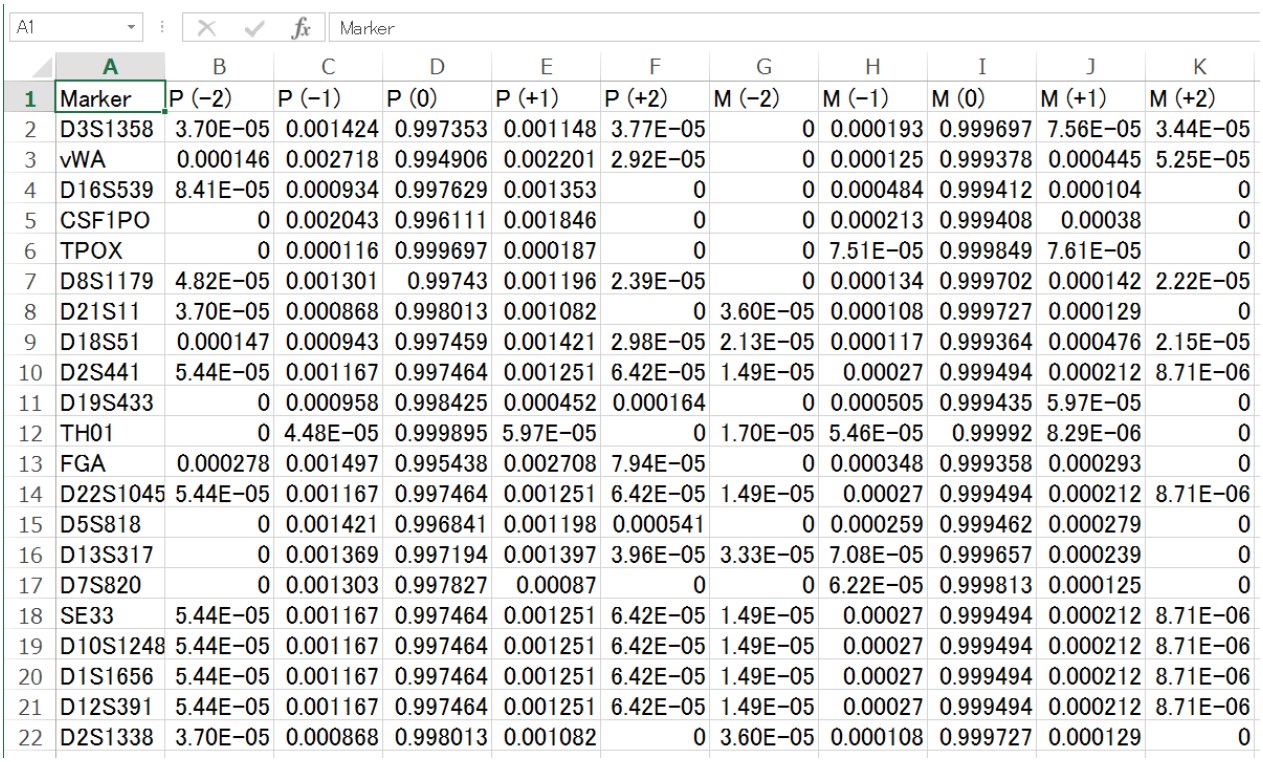
KinBN has two functions; case analysis and simulation. For case analysis, the user prepares three csv files; observed numbers of alleles in each locus at the population (Fig. 2), mutation rates in each locus (Fig. 3), and STR profiles of the case (Fig. 4). Any locus set can be accommodated by the software. Locus name should be unified between these files. For simulation, the user prepares two csv files; observed numbers of alleles in each locus at the population (Fig. 2) and mutation rates in each locus (Fig. 3).

Example files are in the GitHub repository (https://github.com/ChieMorimoto/KinBN/releases).

A1 Allele

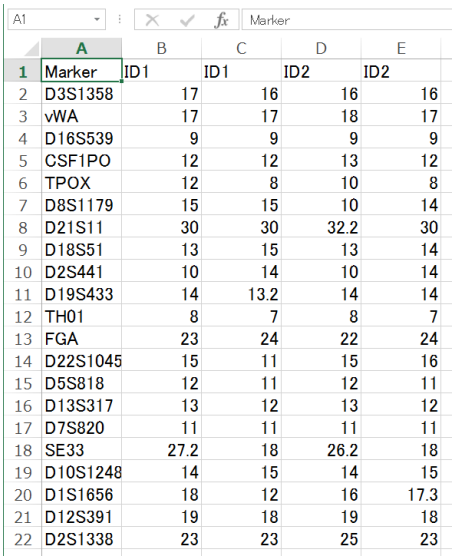
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA
2	2.2												
3	3.2												
4	4												
5	4.2												
6	5					1						1	
7	6					1						170	
8	6.3												
9	7											140	
10	8			13	4	379	10					69	
11	8.1												
12	9			77	10	92	4			1		86	
13	9.1									1			
14	9.3											249	
15	10			41	159	36	74		6	152	1	6	
16	10.1												
17	10.2												
18	10.3												
19	11	1		227	223	182	55		7	248	4	1	

Fig. 2 Example of observed numbers of alleles in each locus at the population. It is used for estimation of allele frequencies. The first column and row indicate allele number and locus names, 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	0.00027	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. 3　Example of locus- and sex-specific mutation rates in each locus. The first column indicates locus names. P (x) and M (x) indicate paternal and maternal mutation rates, respectively. X means the distances of mutation.



Marker	ID1	ID1	ID2	ID2
D3S1358	17	16	16	16
vWA	17	17	18	17
D16S539	9	9	9	9
CSF1PO	12	12	13	12
TPOX	12	8	10	8
D8S1179	15	15	10	14
D21S11	30	30	32.2	30
D18S51	13	15	13	14
D2S441	10	14	10	14
D19S433	14	13.2	14	14
TH01	8	7	8	7
FGA	23	24	22	24
D22S1045	15	11	15	16
D5S818	12	11	12	11
D13S317	13	12	13	12
D7S820	11	11	11	11
SE33	27.2	18	26.2	18
D10S1248	14	15	14	15
D1S1656	18	12	16	17.3
D12S391	19	18	19	18
D2S1338	23	23	25	23

Fig. 4　Example of STR genotypes of two persons (ID1 and ID2).

C) Case analysis

After the user selects *Case analysis* mode in the top screen, Files tab open (Fig. 5).

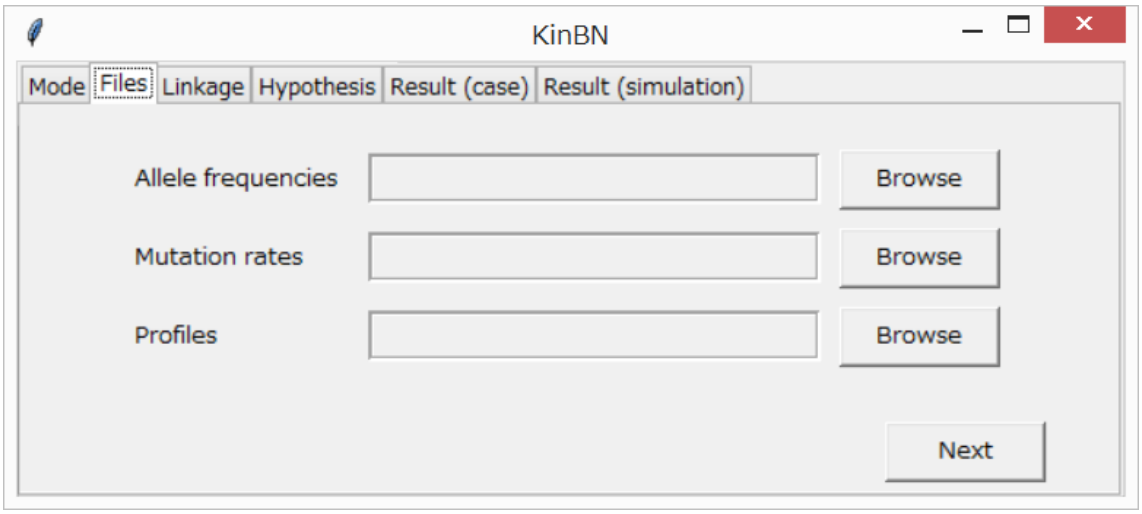


Fig. 5 Files tab for importing three input files.

The user imports three input files from *Browse* buttons. Once the user selects *Next*, Linkage tab for setting information about linked loci will open (Fig. 6).

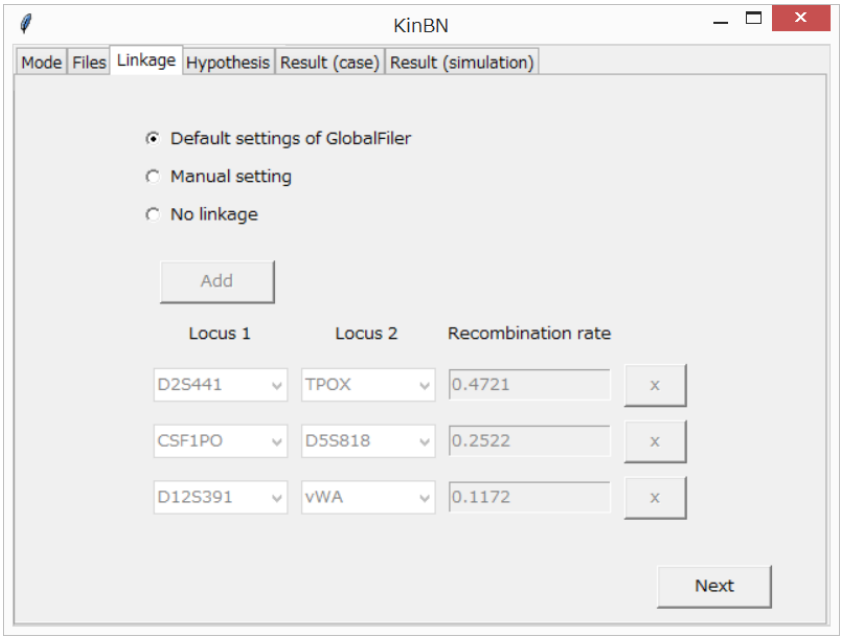


Fig. 6 Linkage tab for setting information about linked loci. Default settings of GlobalFiler are shown as an example.

The user can select linkage setting from three modes; *Default setting*, *Manual setting*, and *No linkage*. 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* allows the user to set linked loci and recombination rate freely. The user add and delete pairs of linked loci by *Add* and × button, respectively. If the user assumes independence between loci, *No linkage* mode can be used.

After the user selects *Next*, Hypothesis tab for setting hypotheses of LR will open (Fig. 7).

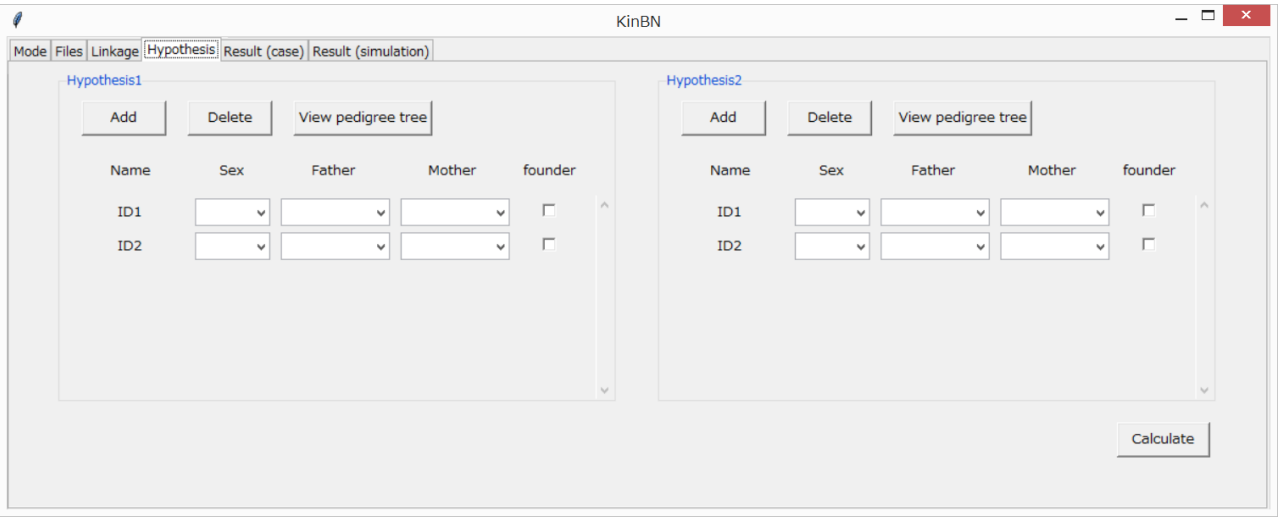


Fig. 7 Hypothesis tab. Following imported file, the names of known profiles are displayed automatically.

In this tab, Hyposesis 1 (H1) and Hypothesis 2 (H2) can be set by making pedigrees manually. In regard to each person, the sex, father, and mother are required. *Add* and *Delete* buttons mean adding and deleting extra persons in the assumed pedigree. If the person is founder, the user checks *founder*. For example, when the user tests pairwise sibship between ID1 and ID2, the setting is shown as Fig. 8. User can confirm pedigrees from *View pedigree tree* button (Fig. 9). This software has a restriction on the number of family members (up to 25).

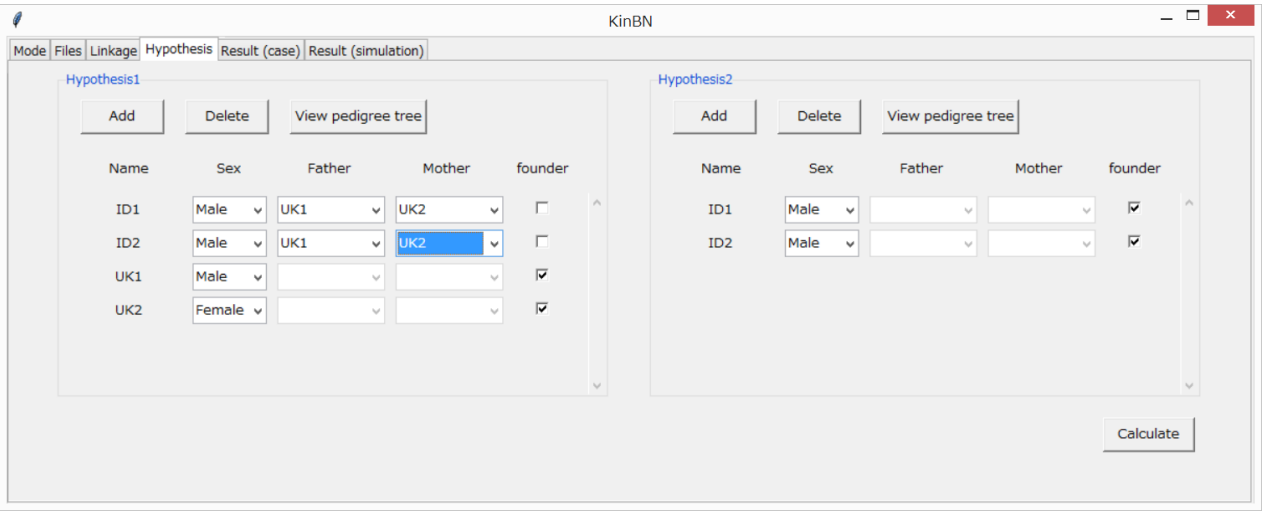


Fig. 8 Example of hyposeses setting at pairwise sibship analysis. In hypothesis 1, UK1 and UK2 are extra persons.

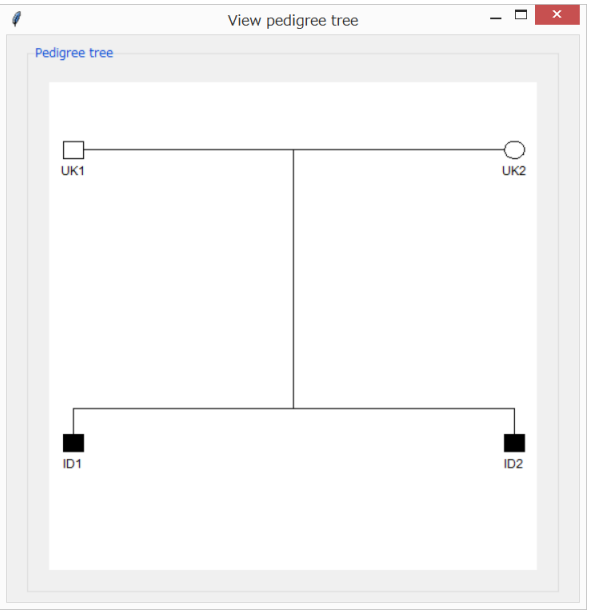


Fig. 9 Pedigree tree of H1 in Fig. 8.

Once the user click *Calculate*, LR calculation will be started. The result will be displayed in Result (case) tab (Fig. 10).

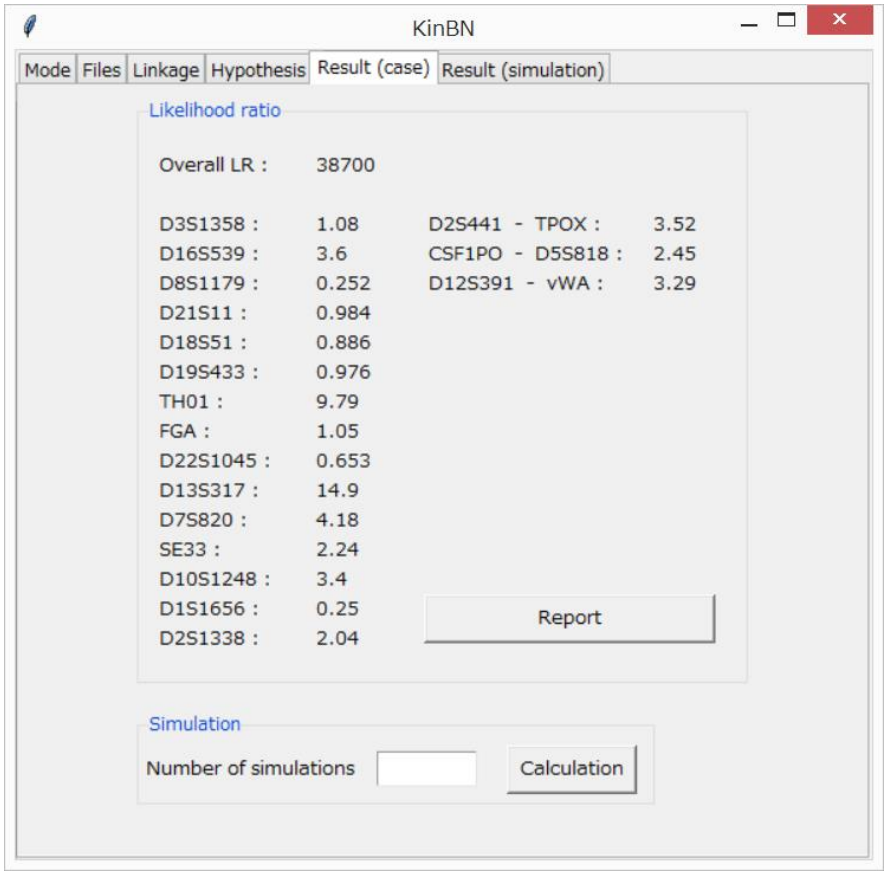


Fig. 10　Example of calculation result of case analysis mode. Overall LR is combined value of all loci. LR value of two linked loci is calculated as one value.

The user can save a report of the case (csv file) from *Report* button. Moreover, by entering *number of simulations*, simulation of LR values in the assumed relationship can be done similar to simulation mode.

D) Simulation

After the user selects *Simulation* in the top screen (Fig. 1), Files tab will open (Fig. 11).

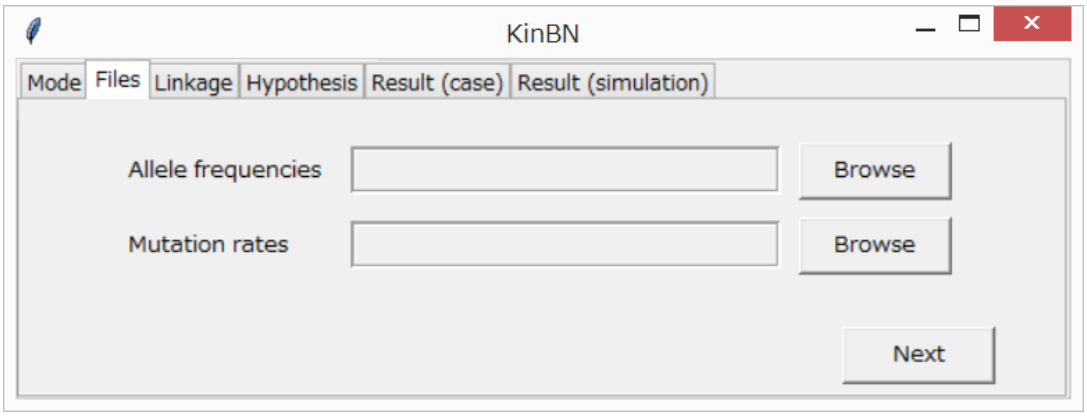


Fig.11 Files tab at simulation mode.

The user imports two input files from *Browse* buttons. Once the user selects *Next*, Linkage tab for setting information about linked loci will open (refer to Fig. 6). The user can select linkage setting similar to case analysis. After the user selects *Next*, Hypothesis tab will open (Fig. 12).

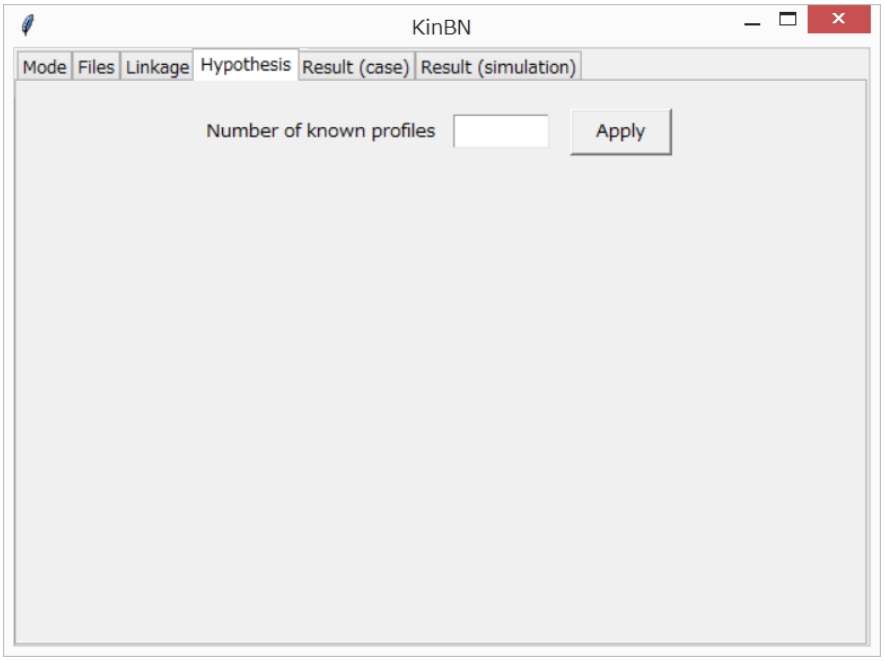


Fig. 12 Hypothesis tab at simulation mode.

In Hypothesis tab, the user enters the number of known profiles. Once the user clicks *Apply*, the screen of hypothesis setting will be displayed (Fig. 13). As with case analysis, each hypothesis can be set by making pedigrees manually. For starting simulation, the user enters the number of simulation and clicks *Calculate* button. A large number of familial genotypes are computationally generated with considering linkage and mutation according to the user's setting. LR values are calculated based on the genotypes and give the distribution under the targeted family tree.

Computation time mainly depends on the assumed pedigree and the number of linked loci. It takes about 4 hours to simulate 10,000 pairwise sibship analyses of 21 STR including three pairs of linked loci by laptop (CPU: Intel Core i7, RAM: 8 GB, storage: 618 GB, OS: Windows 8.1 64 bit). In the case of large family consisted by 21 persons, it takes about 1.5 hours to calculate single LR value of 21 STR including three pairs of linked loci by the same laptop. The calculation time can be reduced to 7 sec by ignoring linkage. If it takes a lot of time to complete the simulation, please consider reducing family members and ignoring linkage. When we performed developmental validation of *KinBN*, considering linkage does not have a general effect on LR distributions in the GlobalFiler 21 loci. In addition, we don't recommend large number of simulations (e.g., > 1,000 times) especially when the user assume a lot of family members and consider linkage.

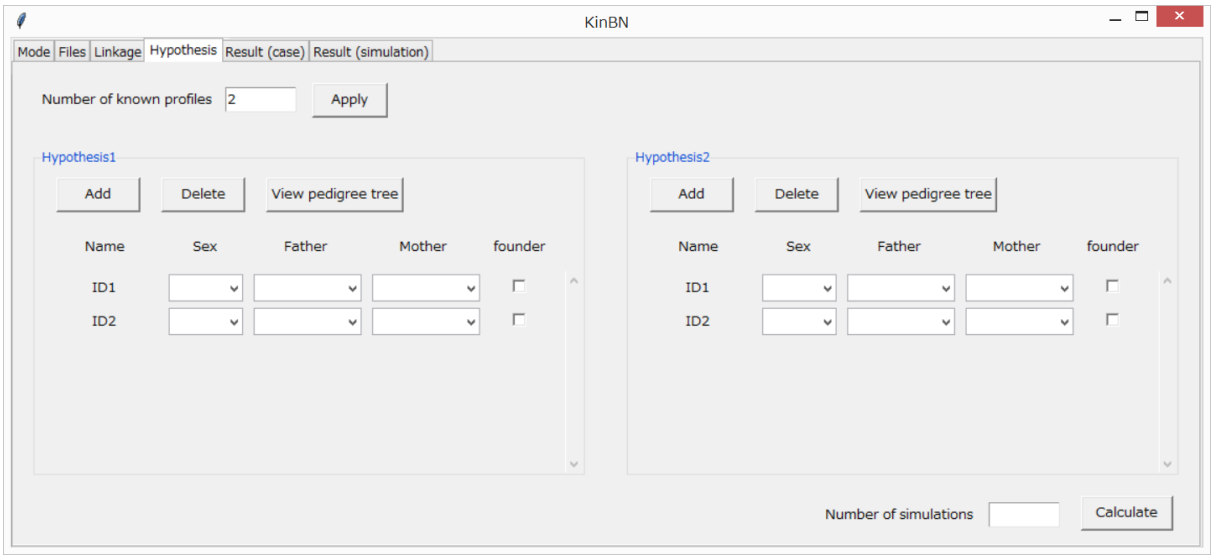


Fig.13 Example of hypothesis tab in simulation mode. In this example, the number of known profiles is two (ID1 and ID2).

Result (simulation) tab will be opened by finishing simulation (Fig. 14). Quantiles (minimum, 5%, 25%, 50%, 75%, 95%, and maximum) of LR values are shown for both H1 true testing and H2 true testing.

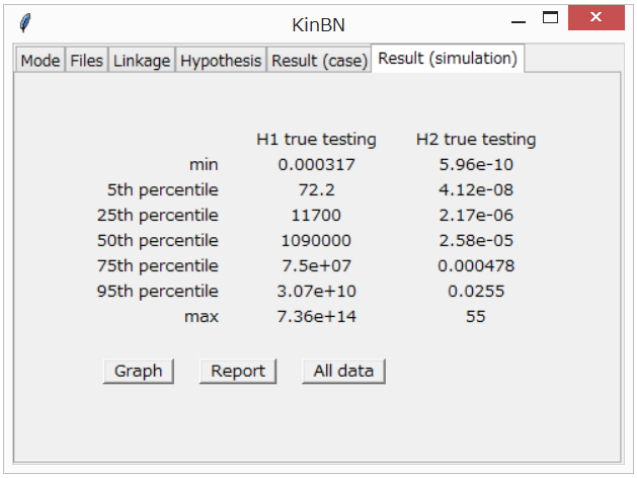


Fig. 14 Example of calculation result of simulation mode.

Graph of LR distributions can be shown from *Graph* button (Fig. 15). The user can save a report of the simulation summary (csv file) from *Report* button. All data of LR values can also be obtained from *All data* button.

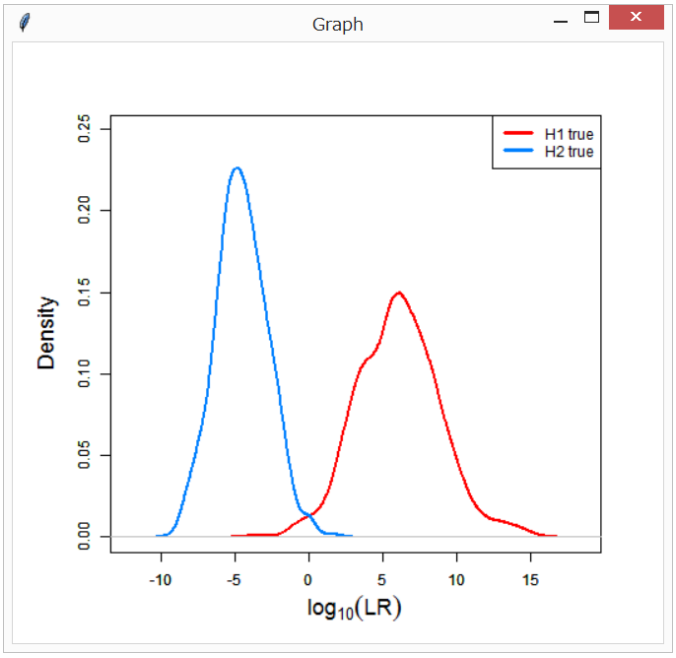


Fig. 15 Example of LR distributions.