Kongoh version 3.2.2 User Manual

Sho Manabe (manabesh@hirakata.kmu.ac.jp)

2022-10-21

Contents

Information on Kongoh **3**

Version changes **4**

Changes in ver. 3.2.2 . . . 4

Important updates . . . 4

Minor changes . . . 4

Changes in ver. 3.2.1 . . . 4

New features . . . 4

Minor changes . . . 4

Changes in ver. 3.2.0 . . . 5

New features . . . 5

Minor changes . . . 6

Tutorial **7**

Getting started . . . 7

Input files . . . 9

Deconvolution . . . 14

Likelihood ratio . . . 23

Set analysis methods **35**

Create an analysis method . . . 35

Edit an analysis method . . . 43

Estimate parameters **45**

Input files . . . 45

Set conditions . . . 49

Review the results . . . 60

Set information on typing kits **72**

Create information on a typing kit 72

Edit information on a typing kit 83

Edit repeat correction **85**

References **89**

Information on Kongoh

Kongoh (named after the Japanese word "mixture") is open-source software for DNA evidence interpretation based on a quantitative continuous model[1, 2]. The software is a graphical user interface and has been released as the R-package *Kongoh* (>= ver. 3.2.0).

Kongoh can perform mixture deconvolution and the assignment of the likelihood ratio for crime stain profiles typed by the AmpFlSTR® Identifiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA) and GlobalFiler™ PCR Amplification Kit (Thermo Fisher Scientific). Profiles typed by kits other than Identifiler Plus and GlobalFiler can be analyzed if experimental data are prepared from the kit. A peak located at the position of the back stutter, forward stutter, double-back stutter, and minus 2-nt stutter need not be designated as an allele or stutter because the derivation of the peak in the stutter position can be determined probabilistically. Hence, the stutter filters of all loci can be removed. Moreover, Kongoh considers allelic drop-out, which is the event of a peak under the analytical threshold. By contrast, drop-in is not considered; however, spontaneous drop-in peaks can be explained based on additional unknown contributors.

Version changes

Note

Users are recommended to switch to version 3.2.2 because of the following updates.

Changes in ver. 3.2.2

Important updates

- A calculation error was fixed. In the previous version of Kongoh, the error occurred in the case that one of the expected forward stutter ratios (FSRs) or the expected double-back stutter ratios (DSRs) were not greater than zero under the model for these stutters. In this case, the new version of Kongoh does not regard the observed peak at the stutter position as only the stutter product.
- A calculation error was fixed. In the previous version of Kongoh, the error occurred when there were some locus names with spaces (e.g., Penta D and Penta E) in the input files.
- An error about setting analytical thresholds in the "Set analysis methods" was fixed.

Minor changes

- The default file "GF_29cycles_3500xL_24sec.csv" is updated. The update does not affect the result of the mixture deconvolution and the likelihood ratio.

Changes in ver. 3.2.1

New features

- The likelihood ratio is displayed on both normal and logarithmic scales in the "Result" of the "Likelihood ratio" tab.
- Users can export likelihoods and LRs of all assumed hypotheses.

Minor changes

- The term "Likelihood" has been changed to "Log10 (Likelihood)" in the "Result" of the "Deconvolution" tab.
- The following minor issue was fixed. The unexpected error message "Error: The total number of contributors is smaller than the number of known contributors." was displayed when the assumed number of contributors was smaller than the number of reference profiles inputted.

Changes in ver. 3.2.0

New features

- Kongoh can be used as the R-package *Kongoh*.
- UI design has been changed.
- Project data can be saved and loaded.
- A new feature "mixture deconvolution" has been implemented. In this feature, probabilistic genotyping can be performed without reference profiles.
- The expected peak height of drop-out alleles (Q) is regarded as that of the most frequent allele in Q regardless of known and unknown contributors.
- A new feature "Set analysis methods" has been implemented.
  - Users can set analysis methods per typing kit in the software.
  - Analytical thresholds can be set per locus.
  - The "mixture ratio filter" is added to exclude unrealistic genotype combinations.
  - The files for Monte Carlo parameters and the information on allele repeat correction can be specified when setting analysis methods. Therefore, users no longer need to load these files in the "Files" tab.
- The file format for Monte Carlo parameters is changed as follows.
  - Names of parameters in each model are explicitly declared in the column.
  - The scale of "Mean" parameters for AE and Hb is changed from a logarithmic scale to a linear scale.
  - See example files for GlobalFiler (GF_29cycles_3500xL_24sec.csv) and Identifiler Plus (IDP_28cycles_3130xl_10sec.csv) which are located at extdata > parameters in the package *Kongoh*.
- A new feature "Edit repeat correction" has been implemented. Users can edit information on allele repeat correction used for models of stutter ratios.
- The default files of allele repeat correction are updated. See files for GlobalFiler (GF_repeat_correction.csv) and Identifiler Plus (IDP_repeat_correction.csv) which are located at extdata > repeat_correction in the package *Kongoh*.
- When the input allele frequencies are represented as allele counts, the input frequencies can be viewed in both the format of allele counts and the format of allele probabilities estimated by Dirichlet distribution.

- Estimated gamma distributions are replaced with the expected peak heights of gamma distributions in the graph of probabilistic genotyping.
- There are some improvements in the feature "Estimate parameters".
  - Project data can be saved and loaded.
  - The information on "Sample File" and "Dye" in the file of experimental data is no longer needed. See the example file "Experimental-data_GF_Example.csv" which is located at extdata > example in the package *Kongoh*.
  - Parameters of min AE, min Hb, max BSR, max FSR, max DSR, and max M2SR can be determined based on the experimental data.
  - Functions for checking input files have been updated to detect the difference in the locus set and the sample names between each file.
  - The graph for stutter ratios has been improved when the option "Multiple loci together" is selected for the method of modeling.
- A new feature "Set information on typing kits" has been implemented. Users can create or edit information on the locus set, alleles, and sizes of each allele.

Minor changes

- The term "mixture ratio" has been changed to "mixture proportion".
- The term "Weight" has been changed to "Weight (0-1 scale)" in the result of probabilistic genotyping.
- An error message has been improved when there is a lack of information about allele repeat correction for the representative of the unobserved alleles (Q).

Tutorial

Getting started

1. Ensure that R (>= 4.2.0) is installed. It is available from the R Development Core Team website (http://www.R-project.org).

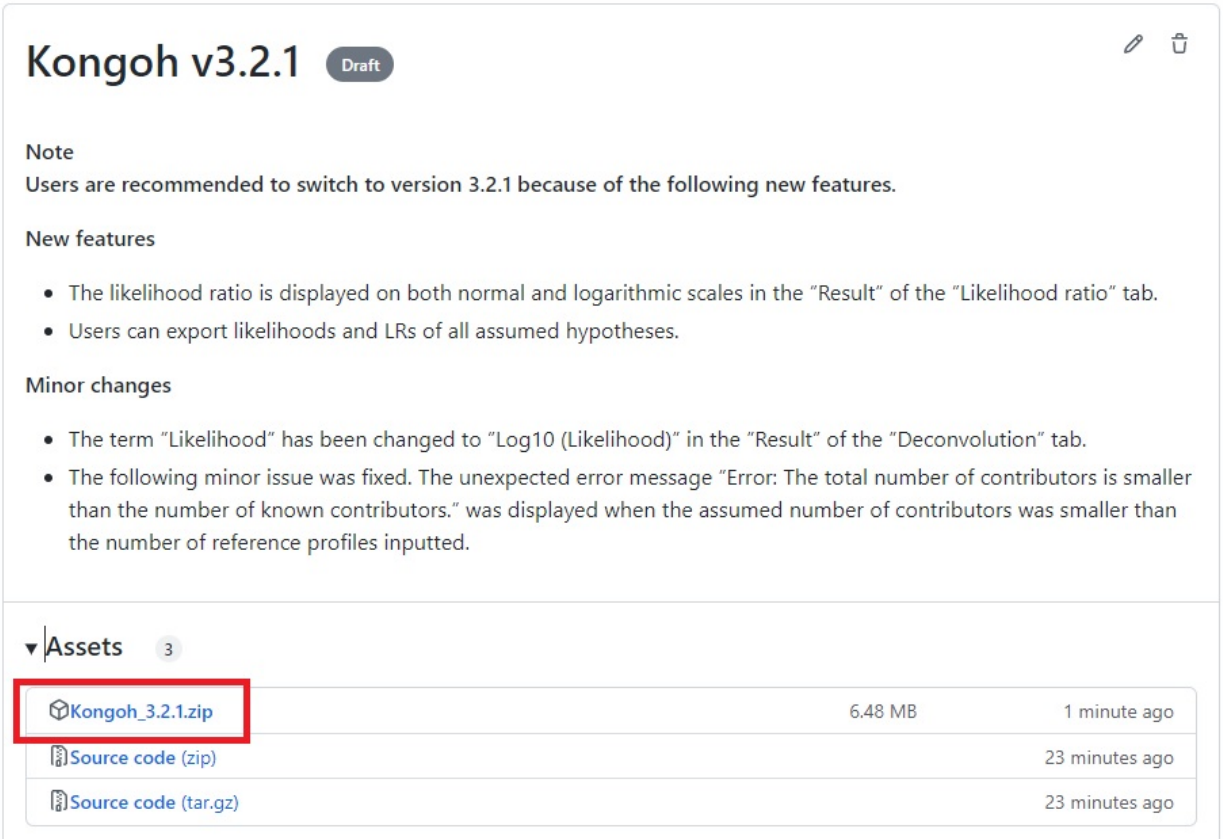
2. Begin an R session.

3. Execute the following command in R to install required packages.

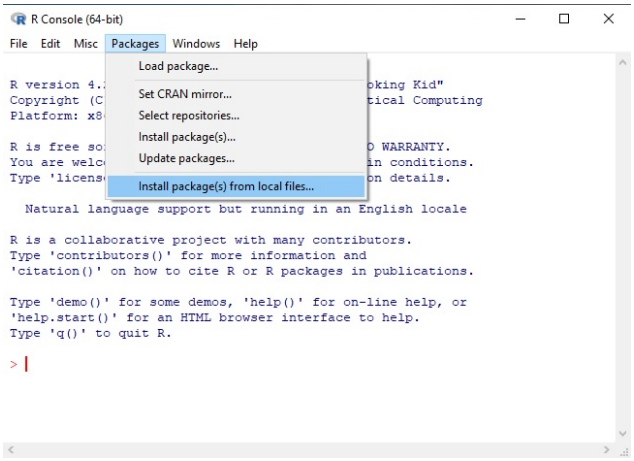
```
install.packages(c("tcltk2", "gtools", "truncnorm", "GenSA"))
```

4. Go to https://github.com/manabe0322/Kongoh/releases.

5. Download "Kongoh_3.2.2.zip".



6. Install ”Kongoh_3.2.2.zip” from ”Install package(s) from local files...”.

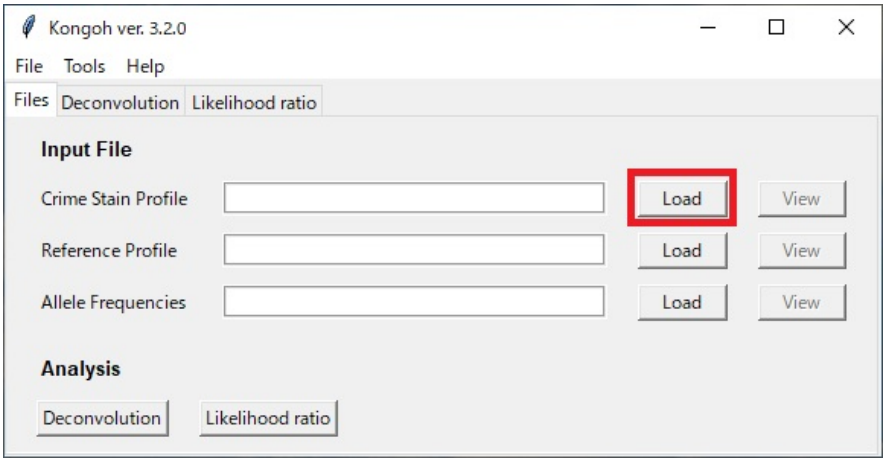


7. Execute the following commands in R to start GUI.

```
library(Kongoh)
Kongoh()
```

Input files

1. Load a file of the crime stain profile.



Note

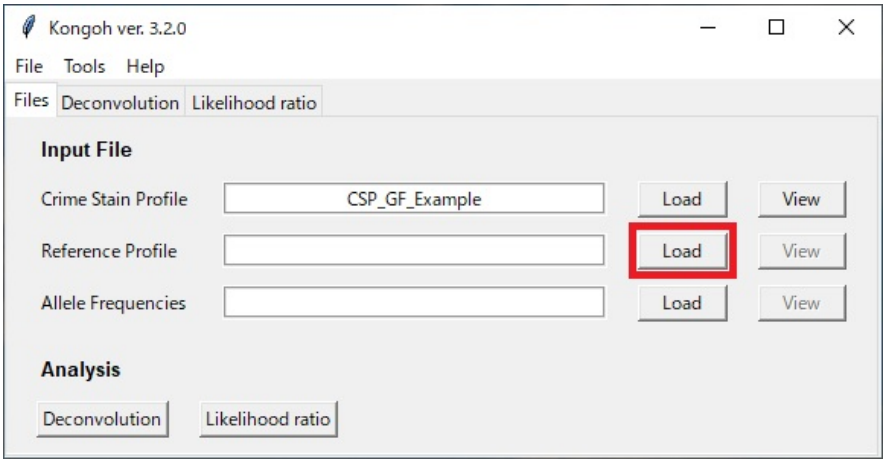
- The example file "CSP_GF_Example.csv" is provided. Go to extdata > example in the package *Kongoh*.
- This file can be exported from the GeneMapper® ID-X software.
- This file must include information regarding the "Sample Name", "Marker", "Allele", "Size", and "Height" as shown in Figure 1.
- Back stutters, forward stutters, double-back stutters, and minus 2-nt stutters are not necessarily removed manually[1].
- Pull-up peaks and noises must be removed manually.

[1] Whether these stutters should be removed depends on the setting of the Monte Carlo parameters.

Sample Name	Marker	Allele 1	Allele 2	Allele 3	...	Size 1	Size 2	Size 3	...	Height 1	Height 2	Height 3	...
Example	D3S1358	15	16	17		121.35	125.35	129.39		1800	262	1494	
Example	vWA	13	14	15		164.95	169.06	172.92		46	1640	93	
Example	D16S539	8	9	10		239.82	243.93	247.93		137	1311	3200	
Example	CSF1PO	8	9	10		291.15	295.15	299.07		59	1831	165	
Example	TPOX	8	10	11		350.77	359.01	362.95		1150	2395	1100	
Example	Yindel												
Example	AMEL	X				98.81				11136			
Example	D8S1179	11	12	13		138.5	142.62	146.78		262	3557	4117	
Example	D21S11	28	29	30		199.64	203.65	207.57		1715	79	2502	
Example	D18S51	13	14	17		285.76	289.76	301.73		400	7163	1908	
Example	DYS391												
Example	D2S441	9	10	11		80.86	84.99	89.15		70	1526	1500	
Example	D19S433	12	12.2	13		141.6	143.55	145.59		134	61	2321	
Example	TH01	6	7	8		187.24	191.31	195.31		922	929	628	
Example	FGA	18	19	21		243.46	247.67	255.6		63	1731	137	
Example	D22S1045	15	16	17		109.42	112.39	115.34		1508	155	3497	
Example	D5S818	8	9	10		142.7	146.78	150.87		55	1657	169	
Example	D13S317	7	8	9		206.87	210.82	214.81		2071	1955	3971	
Example	D7S820	8	9	10		270.6	274.68	278.69		58	1656	3747	
Example	SE33	18	18.2	19		362.31	364.3	366.36		182	47	1940	
Example	D10S1248	13	14	15		106.04	110.06	114.08		231	2959	2107	
Example	D1S1656	14	14.2	15		180.11	182.01	184.23		190	46	4297	
Example	D12S391	17	18	19		228.52	232.48	236.46		161	2919	184	
Example	D2S1338	18	19	20		309.14	313.13	317.35		214	892	2961	

Figure 1: Format of the crime stain profile

2. Load a file of the reference profiles.



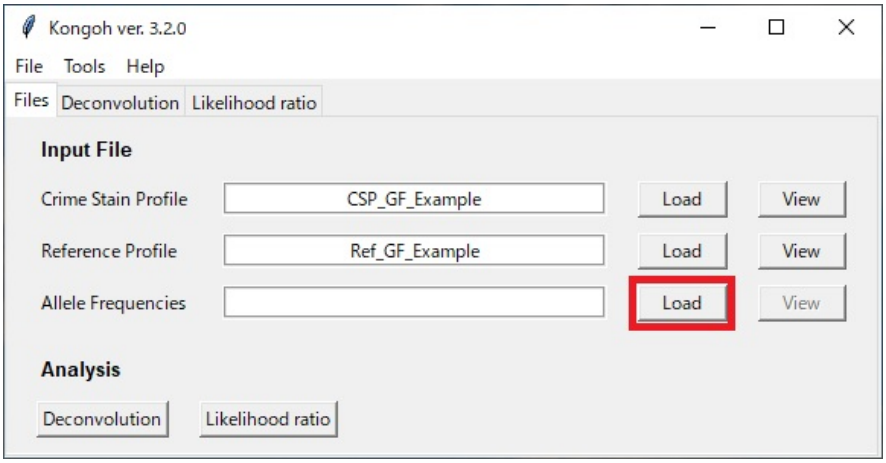
Note

- The example file "Ref_GF_Example.csv" is provided. Go to extdata > example in the package *Kongoh*.
- This file must include information regarding the "Marker" and the name of each profile (e.g., victim and suspect), as shown in Figure 2.
- Two alleles in homozygotes must be entered in each column.

Marker	victim	victim	suspect	suspect
D3S1358	15	18	17	19
vWA	16	17	14	18
D16S539	10	10	9	13
CSF1PO	9	12	11	12
TPOX	10	10	8	11
Yindel				
AMEL	X	X	X	X
D8S1179	13	13	12	14
D21S11	28	30	30	32.2
D18S51	14	14	14	17
DYS391				
D2S441	11	14	10	12

Figure 2: Format of the reference profiles

3. Load a file of the allele frequencies.



Note

- Users can input two formats of allele frequencies: the allele counts (Figure 3) and the allele probabilities (Figure 4).
- If the allele frequencies are represented as the allele counts, the probabilities of each allele are determined by the expected values based on the Dirichlet distribution.
- If the allele frequencies are represented as the probabilities, these values are directly used for the probabilities of each allele. When a crime stain profile or the reference profiles contain alleles that are not listed in the input file, then the minimum allele frequency is used as the frequency of these alleles.
- The allele frequencies for the Identifiler[3] and the GlobalFiler[4] in the Japanese population are provided. Go to extdata > example in the package *Kongoh*. The allele frequencies for the Identifiler are represented as the probabilities. The allele frequencies for the GlobalFiler are represented as the allele counts.

Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA	D22S1045	D5S818	D13S317
5											4				
6											663				
7			1	37		1				1	811			8	6
8			5	3	1357				7		194			18	786
8.1									1						
9			1077	151	354	8					1190			275	392
9.1									110						
9.2										1					
9.3											113				
10			602	648	100	387		6	805		27			616	338
10.1									7						
10.2										3					
10.3															
11			562	620	1072	319		14	1034	10			595	865	675
11.1									1						
11.2										2					
11.3									68						
12	6		517	1267	114	370		140	511	132			4	692	608
12.2										13					
13	3	1	209	208	3	678		598	145	874			5	497	158
13.2										89					
14	79	586	26	52	2	625		643	298	1038			7	28	37
14.1															
14.2										250					
15	1192	78	3	14		410		493	14	160			934	3	2

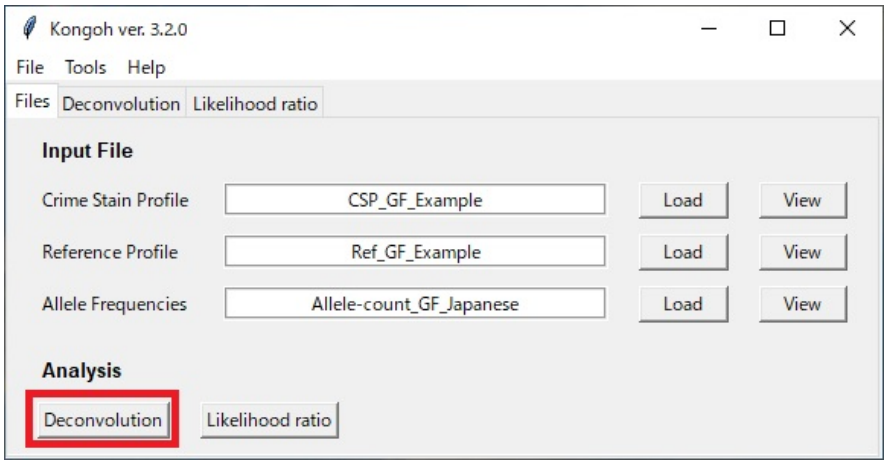
Figure 3: Format of the allele frequencies, which are represented as the allele counts

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
5						0.001852									
6						0.223621									
7	0.001852		0.003332	0.010367		0.266938	0.001852	0.001852				0.001852		0.002962	
8			0.12699	0.001852		0.066642	0.267308	0.002221				0.454646		0.007034	
9	0.001852		0.045909	0.04702		0.398741	0.129211	0.353943				0.114402		0.086264	
9.1			0.001852												
9.2										0.001852					
9.3						0.035172									
10	0.132914		0.219178	0.223621		0.009256	0.115143	0.196964				0.363199	0.002592	0.201037	
10.1			0.001852												
10.2										0.001852					
10.3			0.001852												
11	0.109219		0.328767	0.208071			0.221399	0.187338		0.004073		0.363199	0.004813	0.292484	
11.2										0.001852					
12	0.122917		0.235098	0.418734	0.002221		0.202518	0.178823		0.040726		0.035913	0.04813	0.235468	
12.2										0.005553					
13	0.225102		0.035172	0.069234	0.001852		0.051462	0.072936		0.287671	0.001852	0.001852	0.199556	0.166975	
13.2										0.030359					
14	0.205109		0.006294	0.018141	0.029248		0.013328	0.008515		0.34987	0.194372	0.001852	0.22251	0.009256	
14.2										0.088486					
15	0.134765		0.001852	0.005553	0.39615		0.001852	0.001852		0.051092	0.027027		0.168456	0.001852	
15.2										0.115143					
16	0.064421			0.001852	0.306553				0.008886	0.005553	0.184376		0.125879		
16.2										0.019622					
17	0.006664				0.199926				0.097742		0.282858		0.081822		0.003702

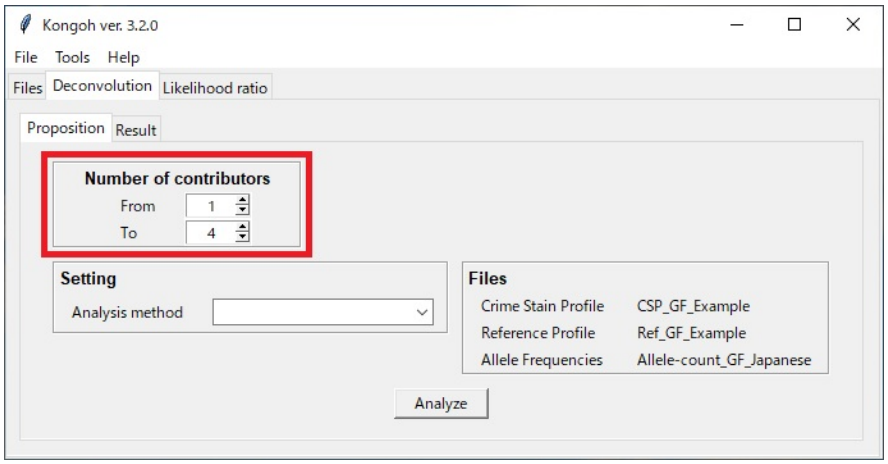
Figure 4: Format of the allele frequencies, which are represented as the allele probabilities

Deconvolution

1. Click the ”Deconvolution” button in the ”Files” tab.



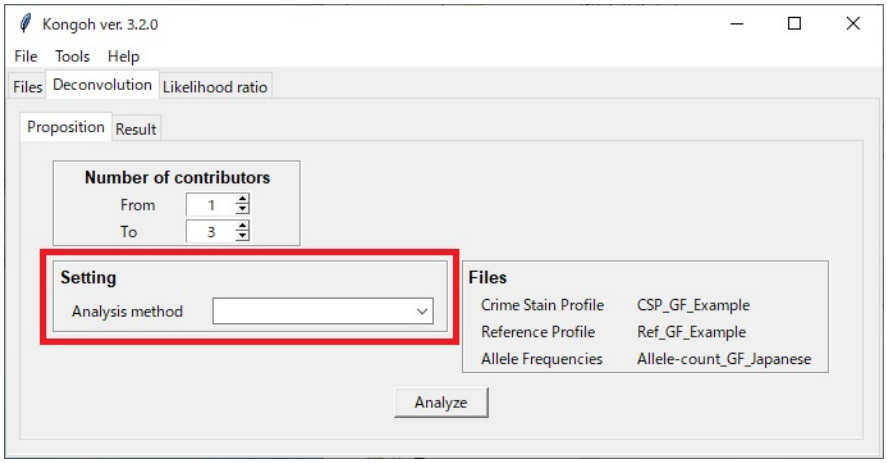
2. Set the range of the assumed number of contributors.



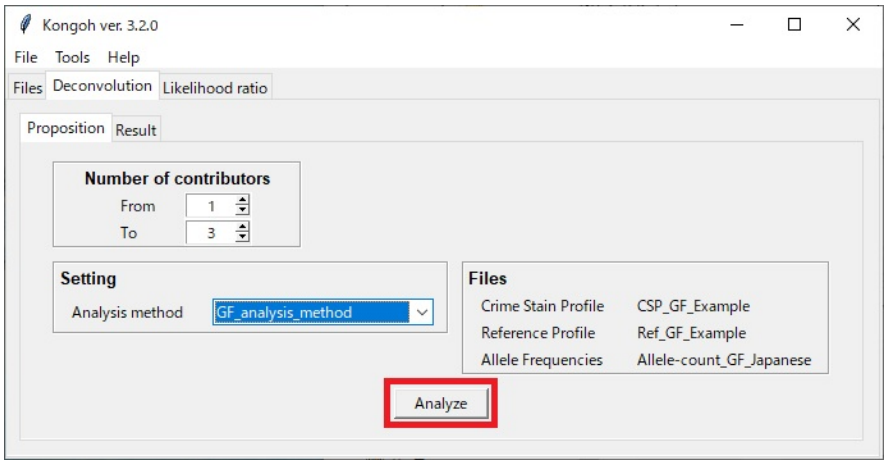
Note

The upper limit of the assumed number of contributors is four.

3. Select an analysis method[2].



4. Click the "Analyze" button.

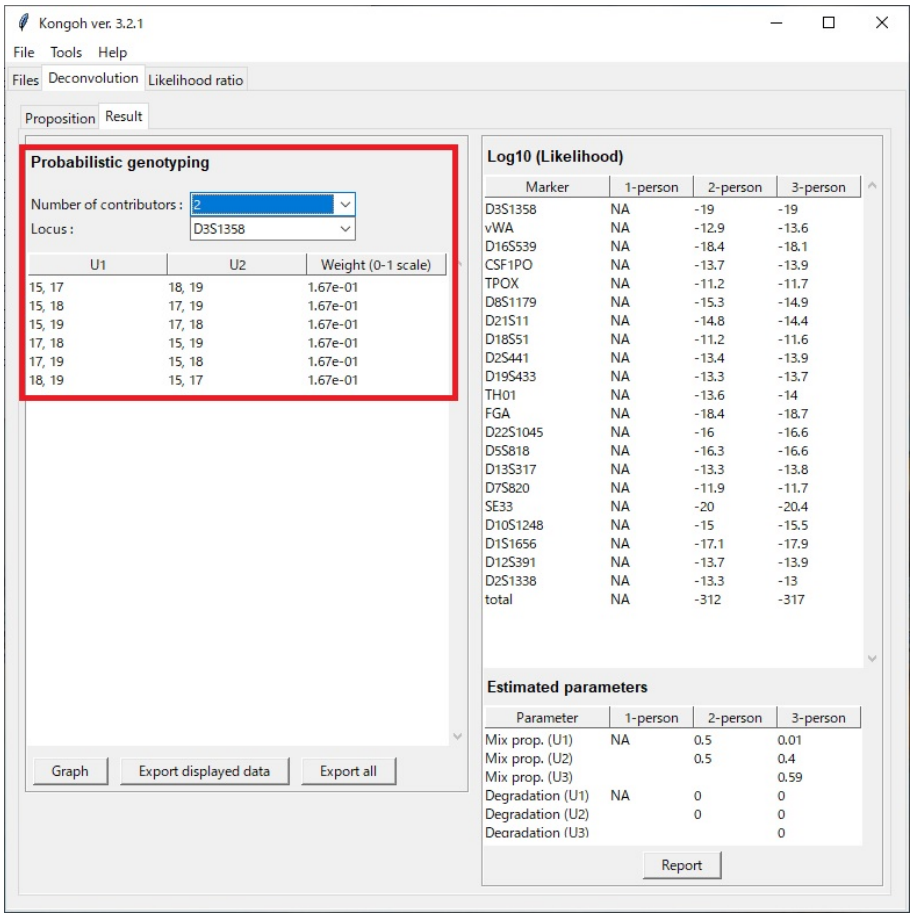


[2]Analysis methods can be set from Tools > Set analysis methods. See the section "Set analysis methods".

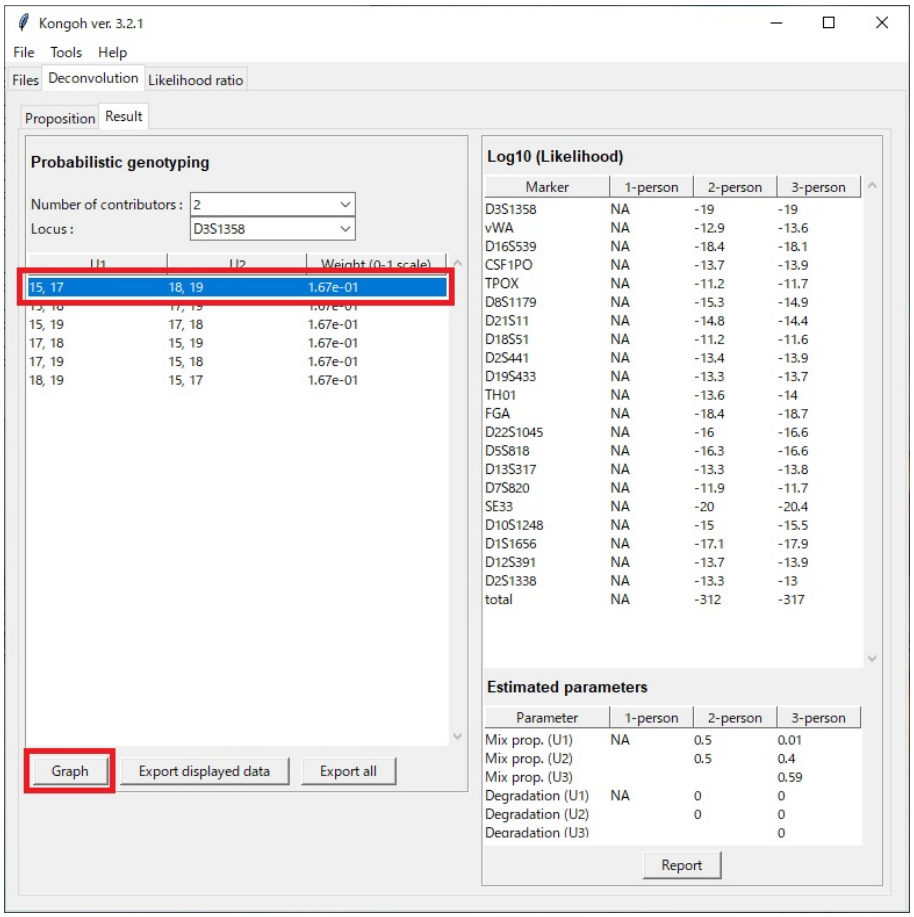
Then, the ”Progress Bar” window will appear. After completing the analysis, the ”Result” tab will appear automatically.

5. Review the result of probabilistic genotyping.

- The weight values of each genotype combination under the selected conditions are displayed.



- The observed peak heights and the expected peak heights of gamma distributions under the selected genotype combination can be compared by clicking the "Graph" button. An example of the graph is shown in Figure 5.



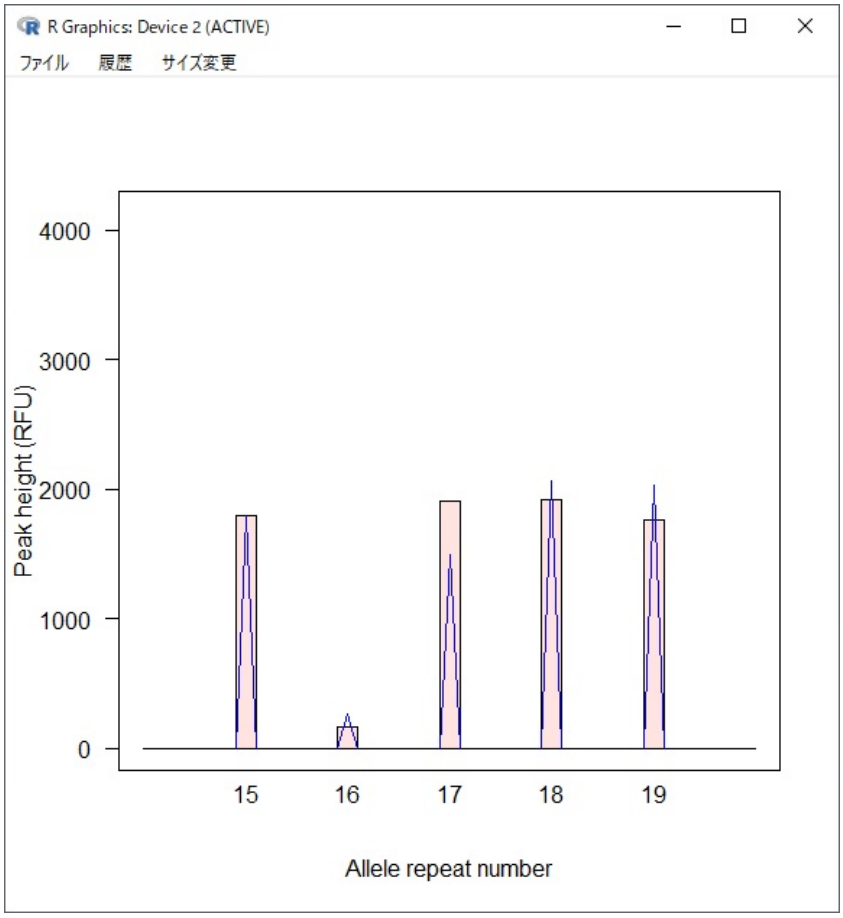
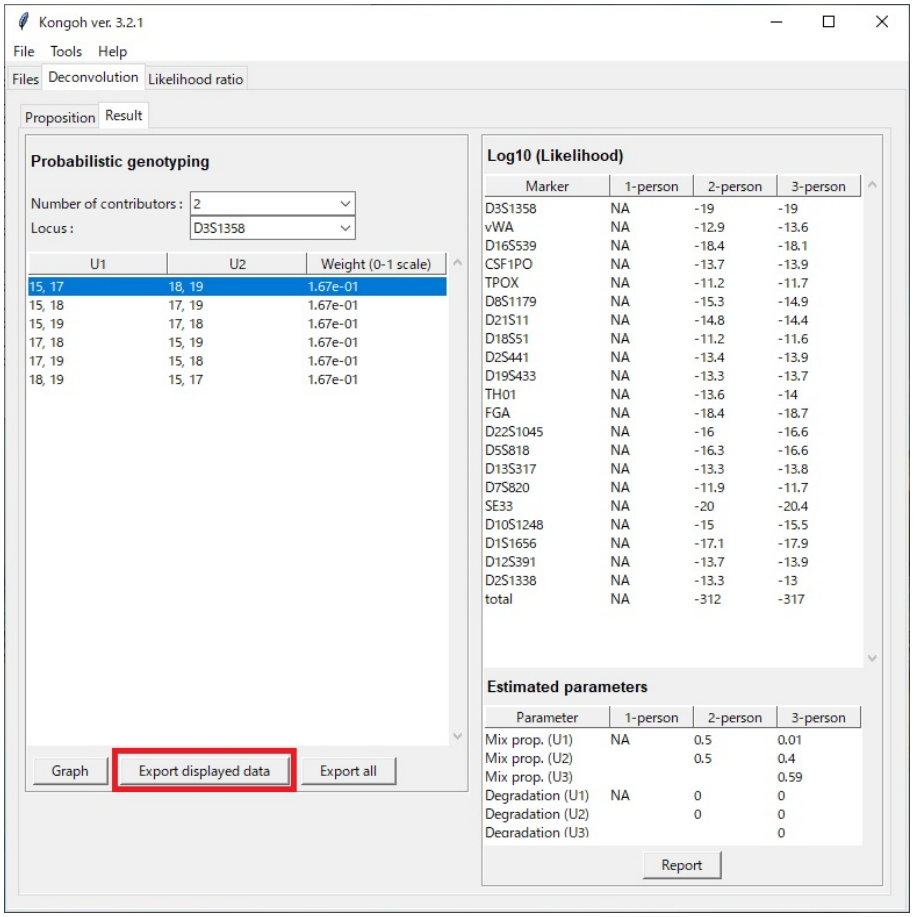
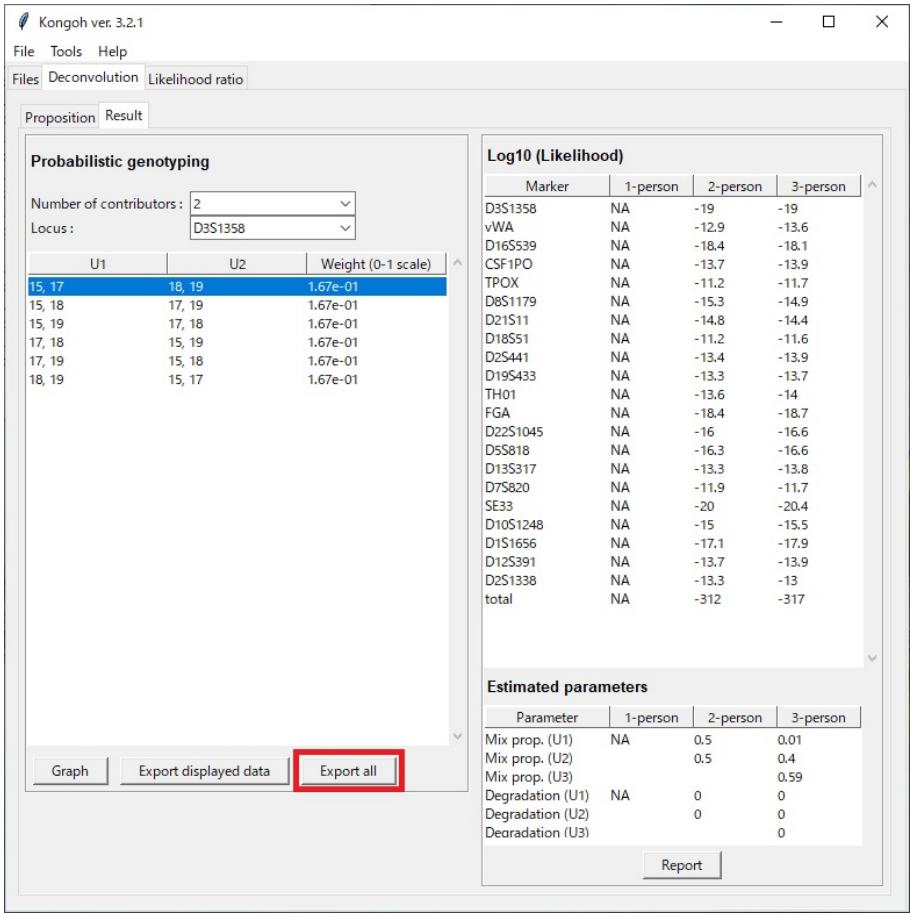


Figure 5: An example of the comparison between the observed peak heights (blue) and the expected peak heights (red)

- Displayed data can be exported by clicking the ”Export displayed data” button.

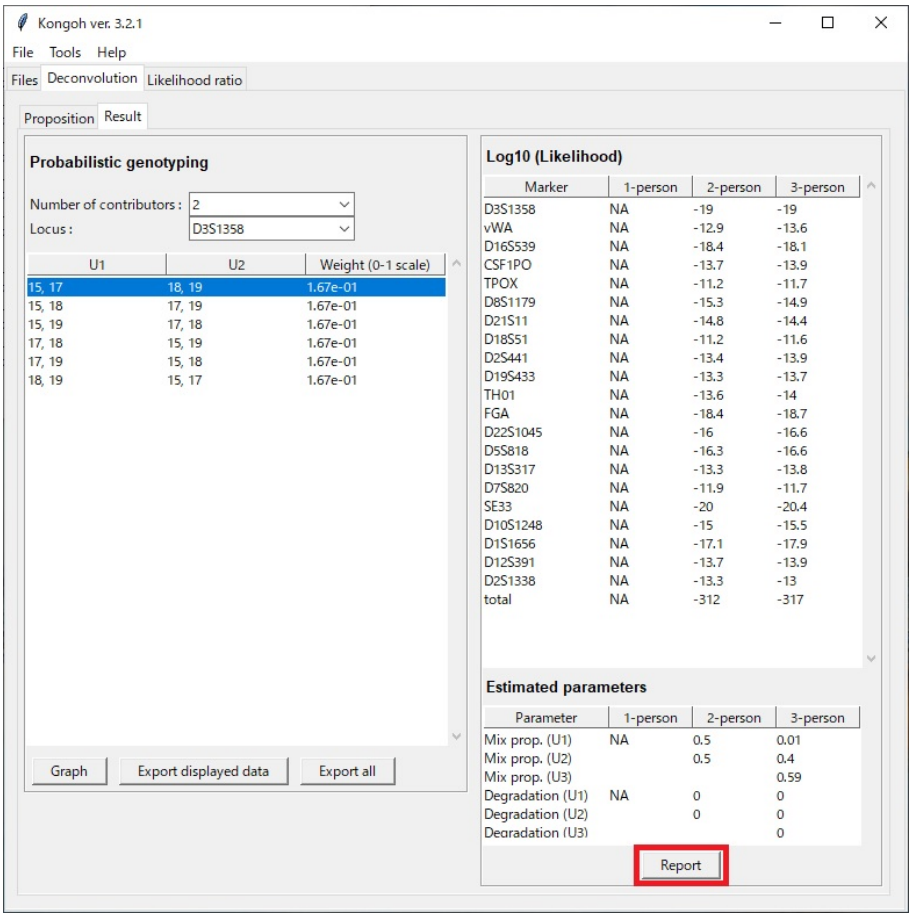


- The results of all conditions can be exported by clicking the ”Export all” button.



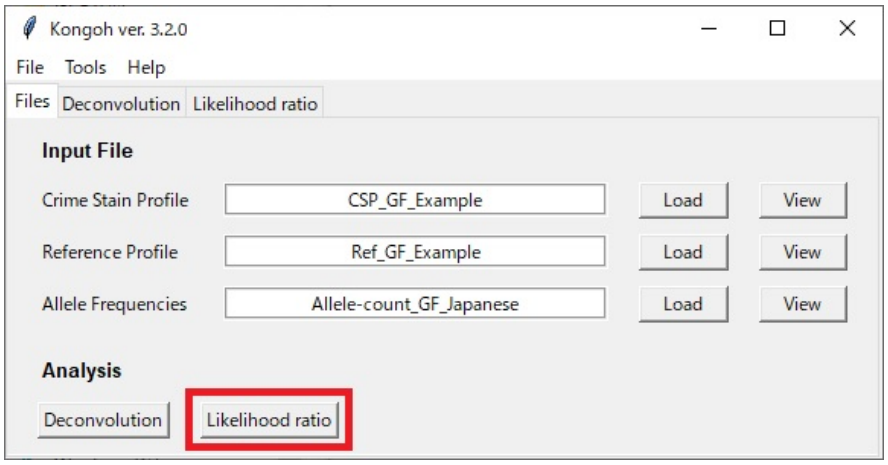
6. Review the result of the likelihoods and the estimated parameters.

   - The report of probabilistic genotyping can be exported by clicking the ”Report” button.



Likelihood ratio

1. Click the "Likelihood ratio" button in the "Files" tab.



2. Set both the prosecutor (Hp) and defense (Hd) hypotheses. Check the individuals to include them as contributors in each hypothesis.

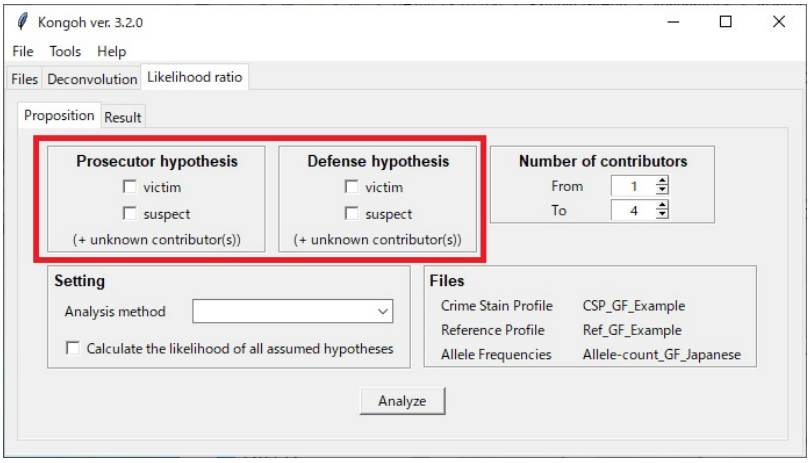


Fig. 6 shows an example of setting the hypotheses:

- Hp: victim + suspect (+ unknown contributor(s))
- Hd: victim (+ unknown contributor(s))

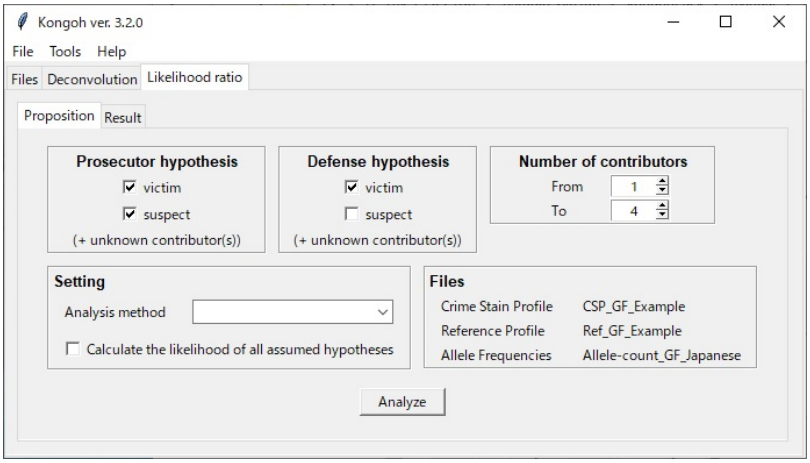
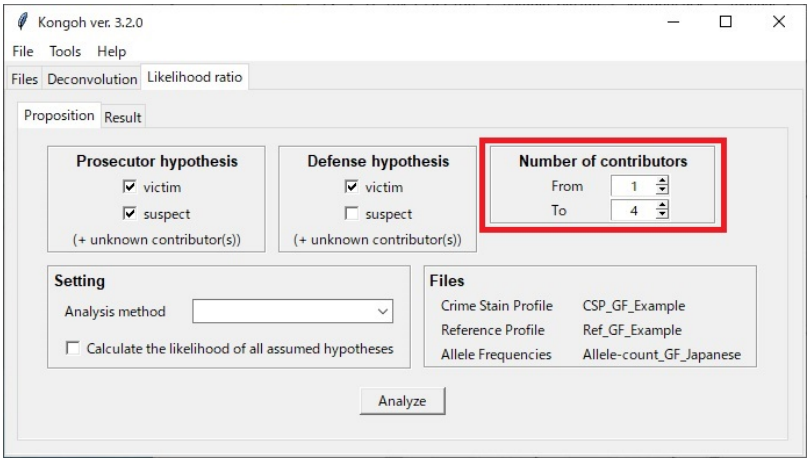


Figure 6: An example of setting the hypotheses

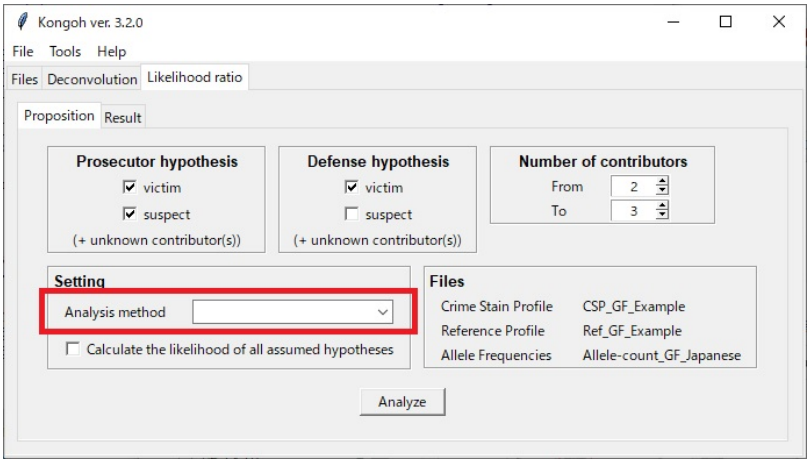
3. Set the range of the assumed number of contributors. The likelihoods of all set numbers are calculated for both Hp and Hd.



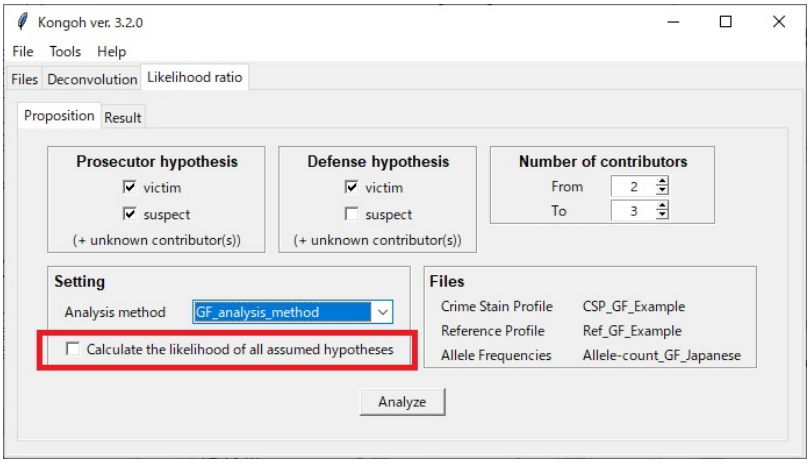
Note

The upper limit of the assumed number of contributors is four.

4. Select an analysis method[3].



5. The likelihoods of all assumed hypotheses can be calculated by checking the "Calculate the likelihood of all assumed hypotheses".

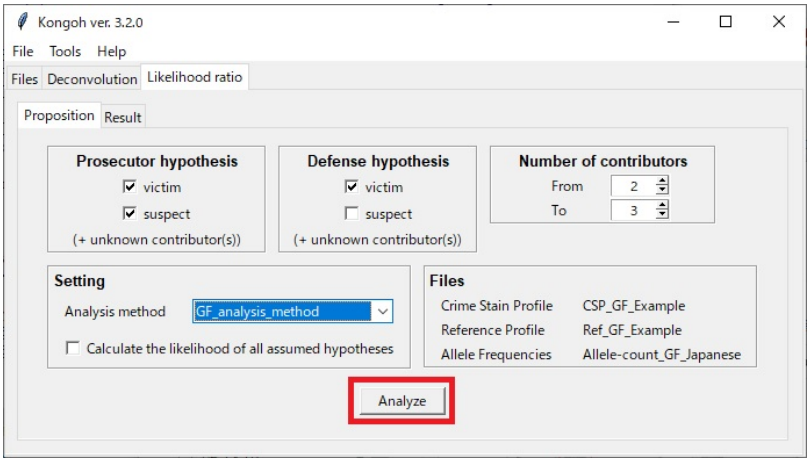


[3]Analysis methods can be set from Tools > Set analysis methods. See the section "Set analysis methods".

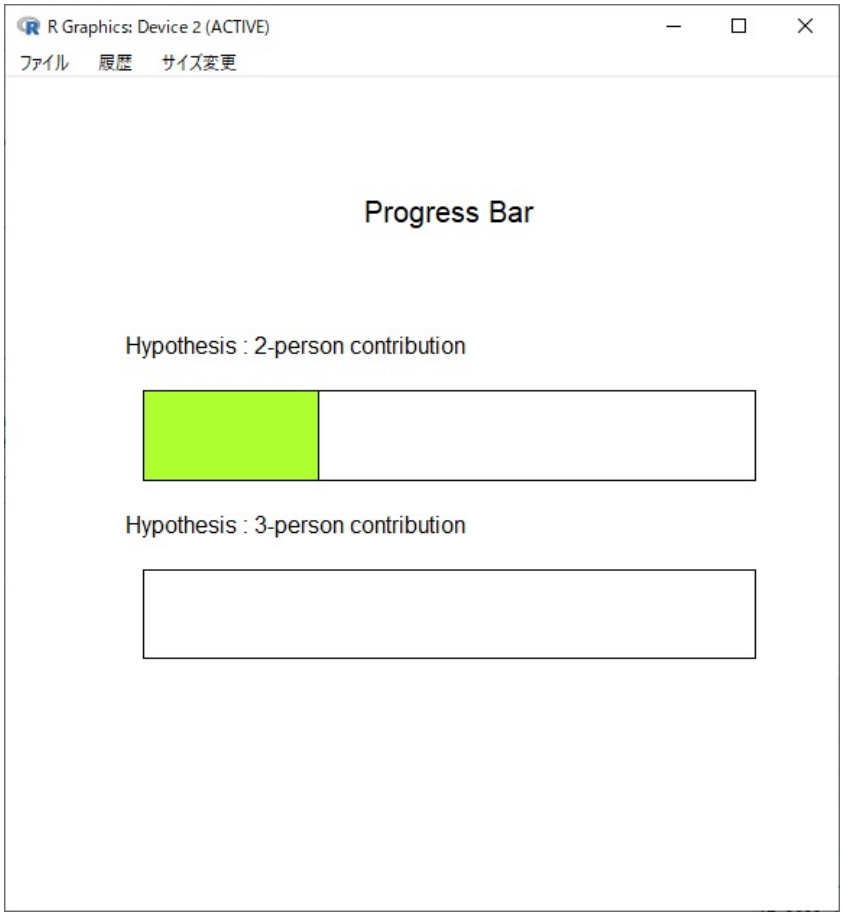
For example, when the reference profiles of the victim and the suspect are inputted and the range of the assumed number of contributors is 2 to 3, all assumed hypotheses are as follows:

- two unknown contributors
- victim + one unknown contributor
- suspect + one unknown contributor
- victim + suspect
- three unknown contributors
- victim + two unknown contributors
- suspect + two unknown contributors
- victim + suspect + one unknown contributor

6. Click the "Analyze" button.

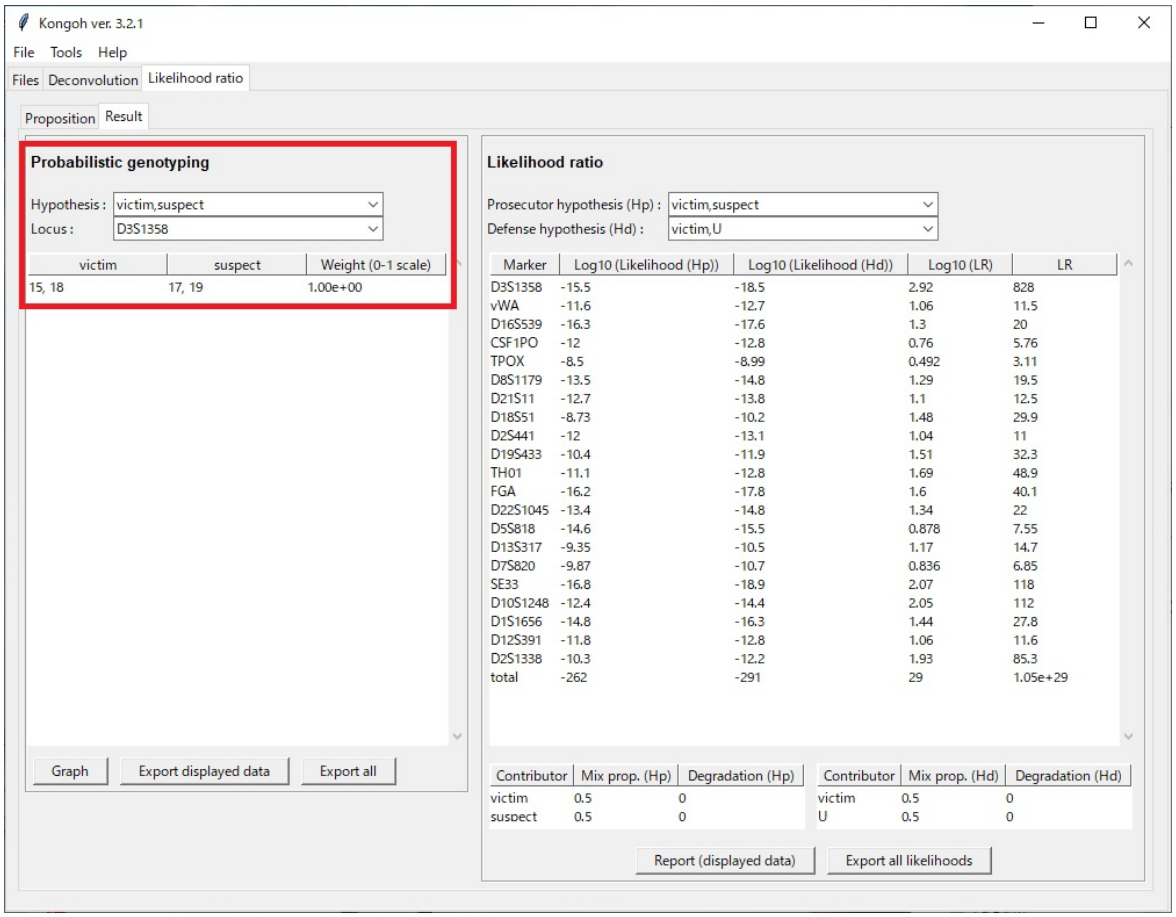


Then, the ”Progress Bar” window will appear. After completing the analysis, the ”Result” tab will appear automatically.

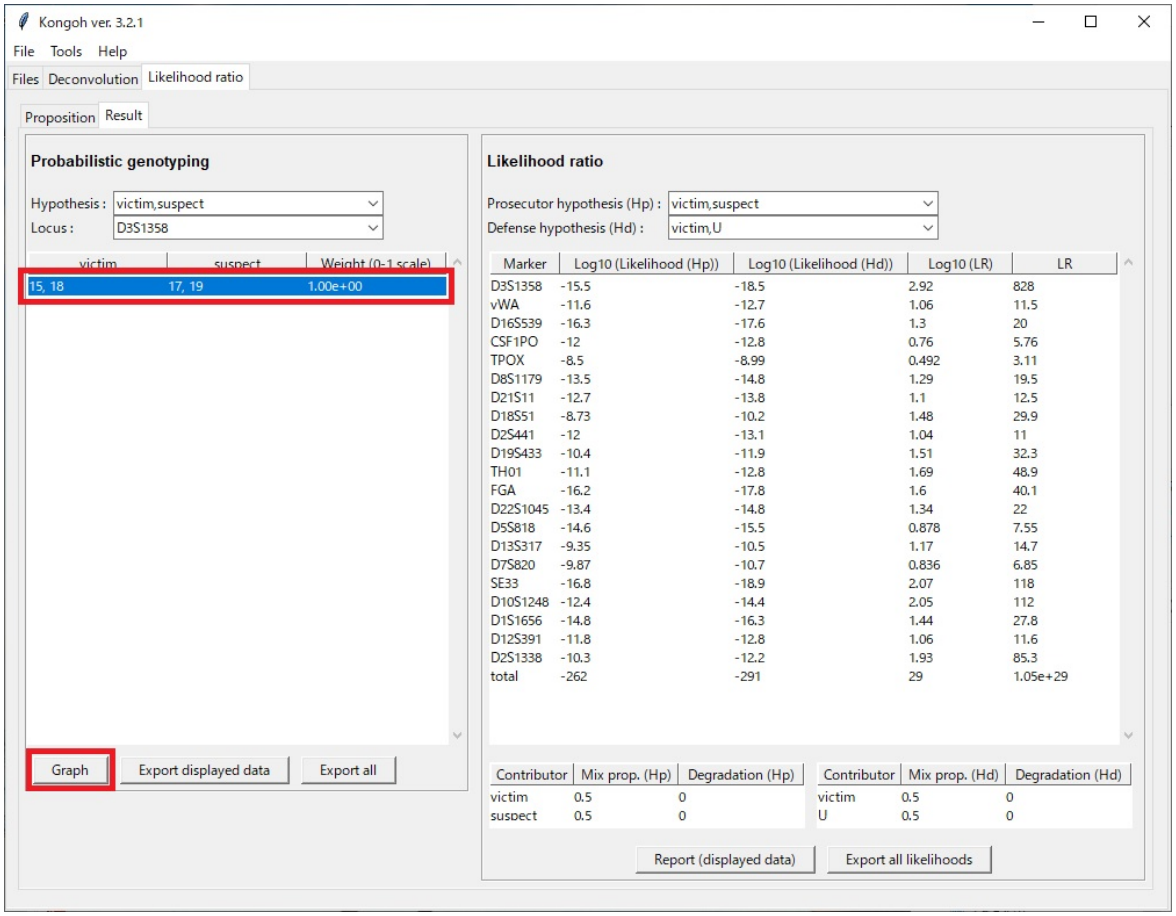


7. Review the result of probabilistic genotyping.

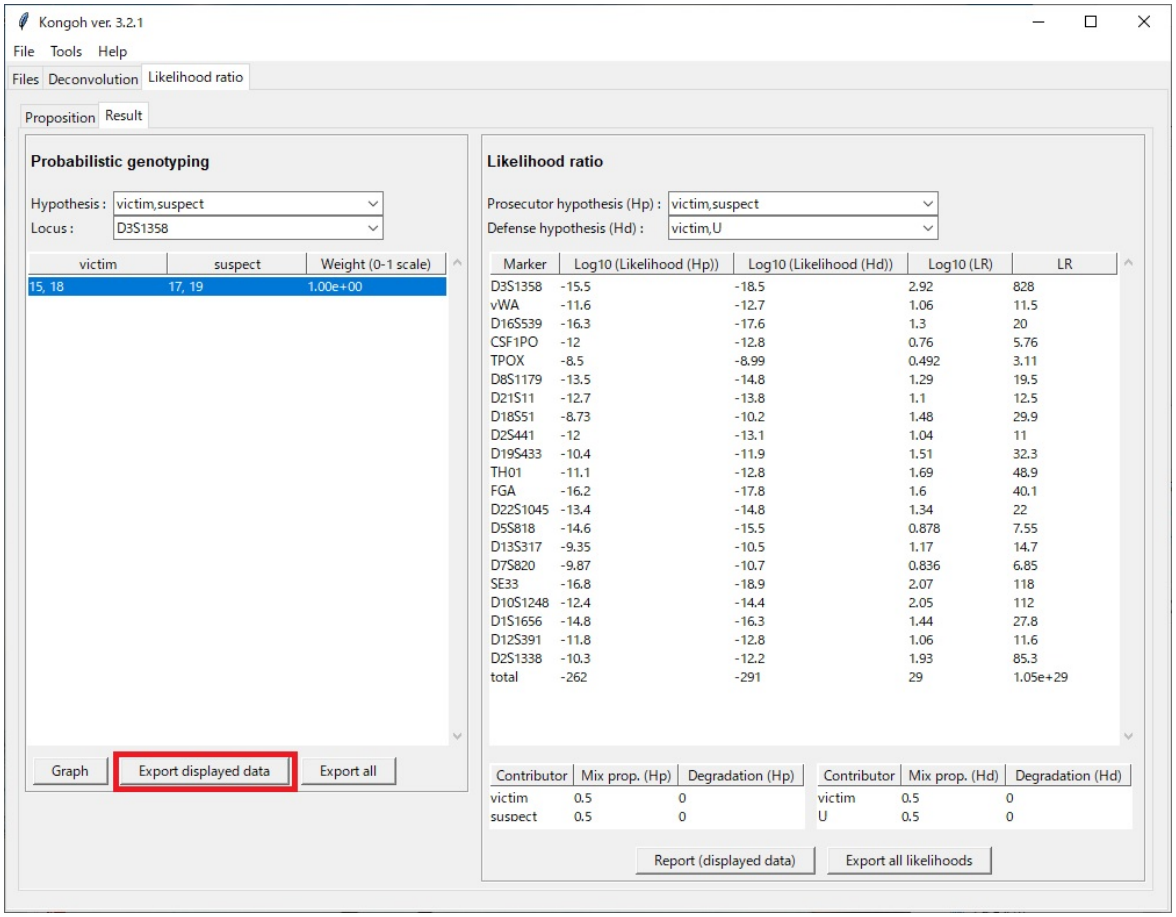
- The weight values of each genotype combination under the selected conditions are displayed.



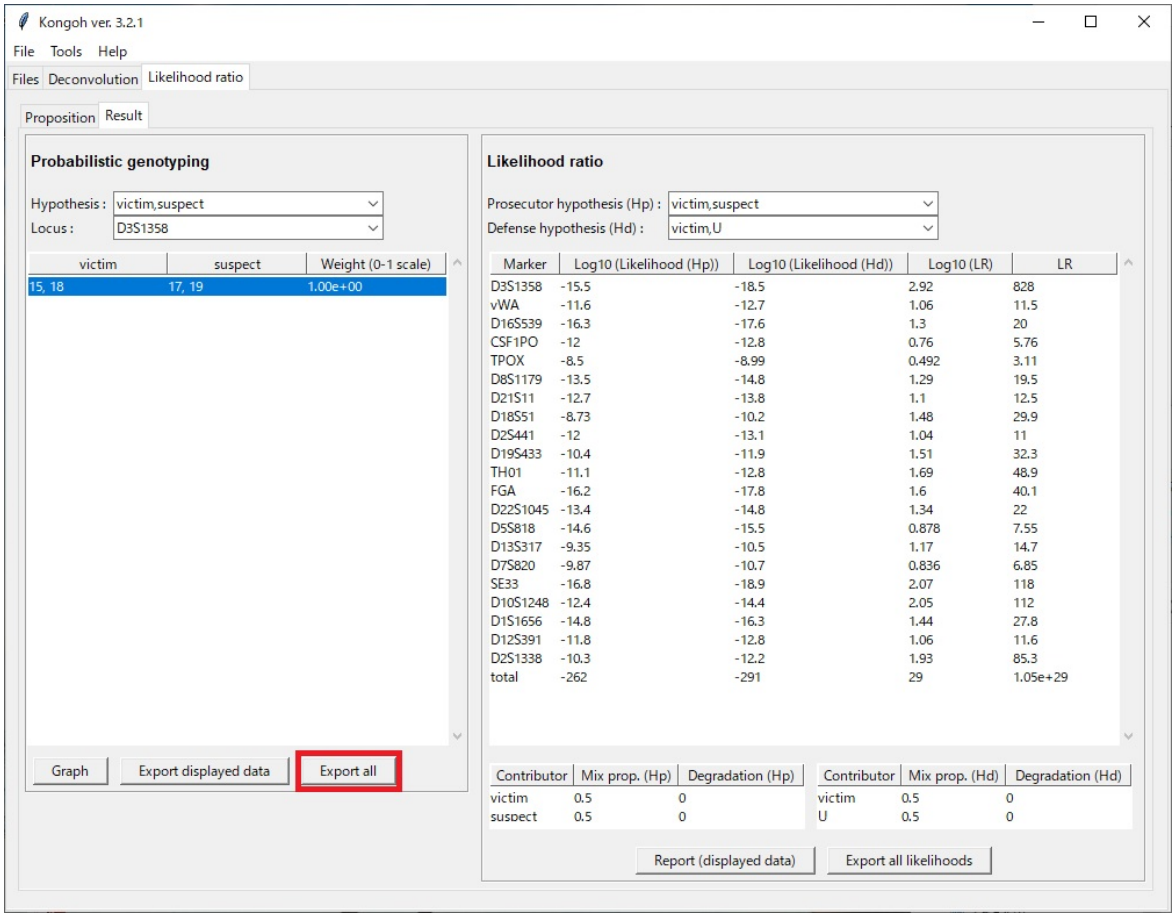
- The observed peak heights and the expected peak heights of gamma distributions under the selected genotype combination can be compared by clicking the ”Graph” button.



- Displayed data can be exported by clicking the ”Export displayed data” button.

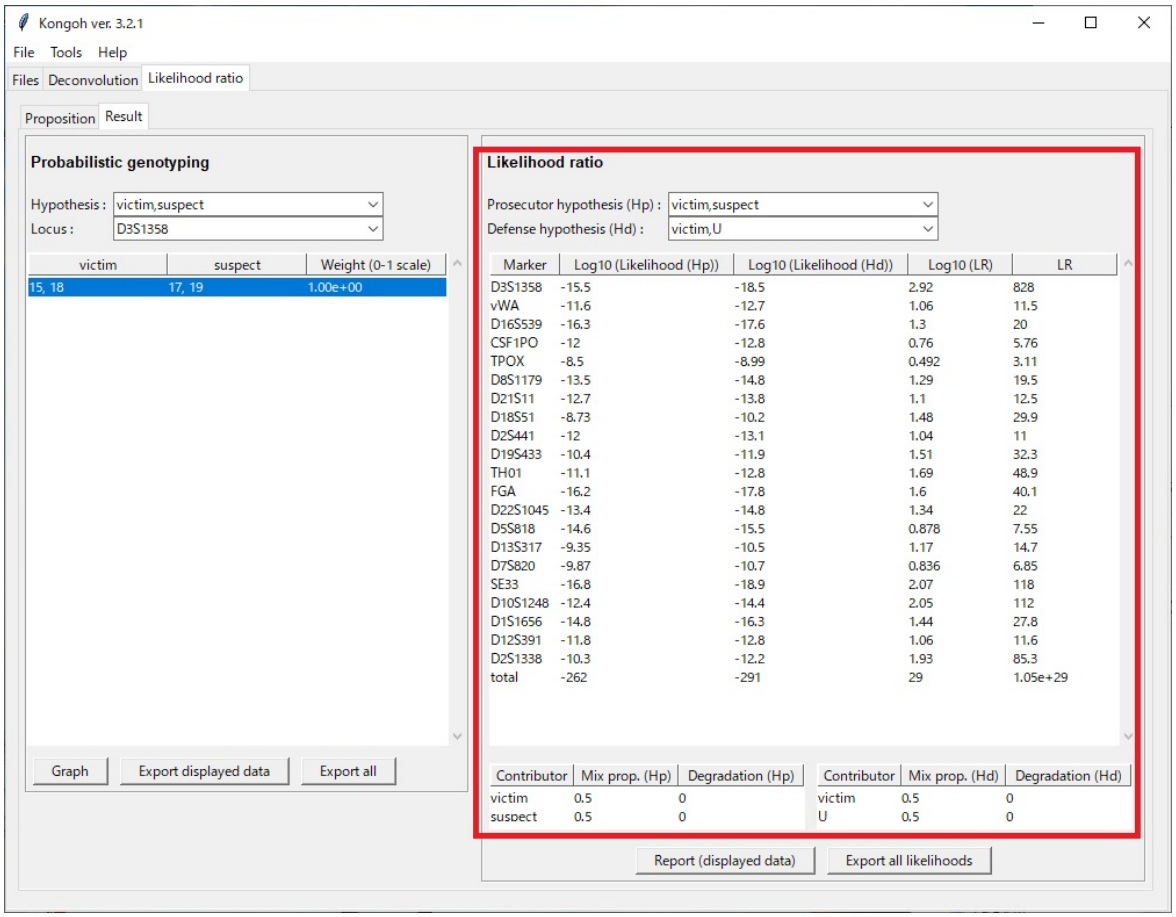


- The results of all conditions can be exported by clicking the ”Export all” button.

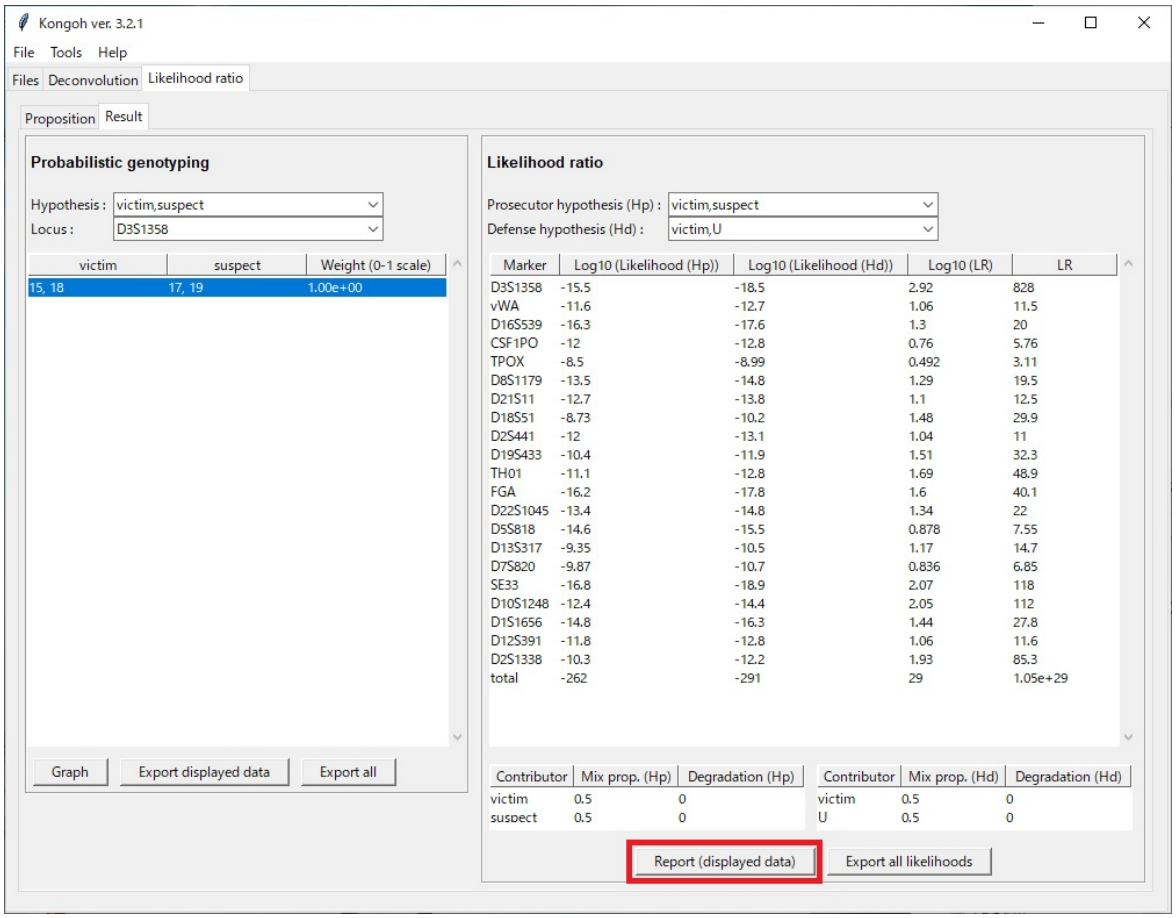


8. Review the result of the likelihood ratio.

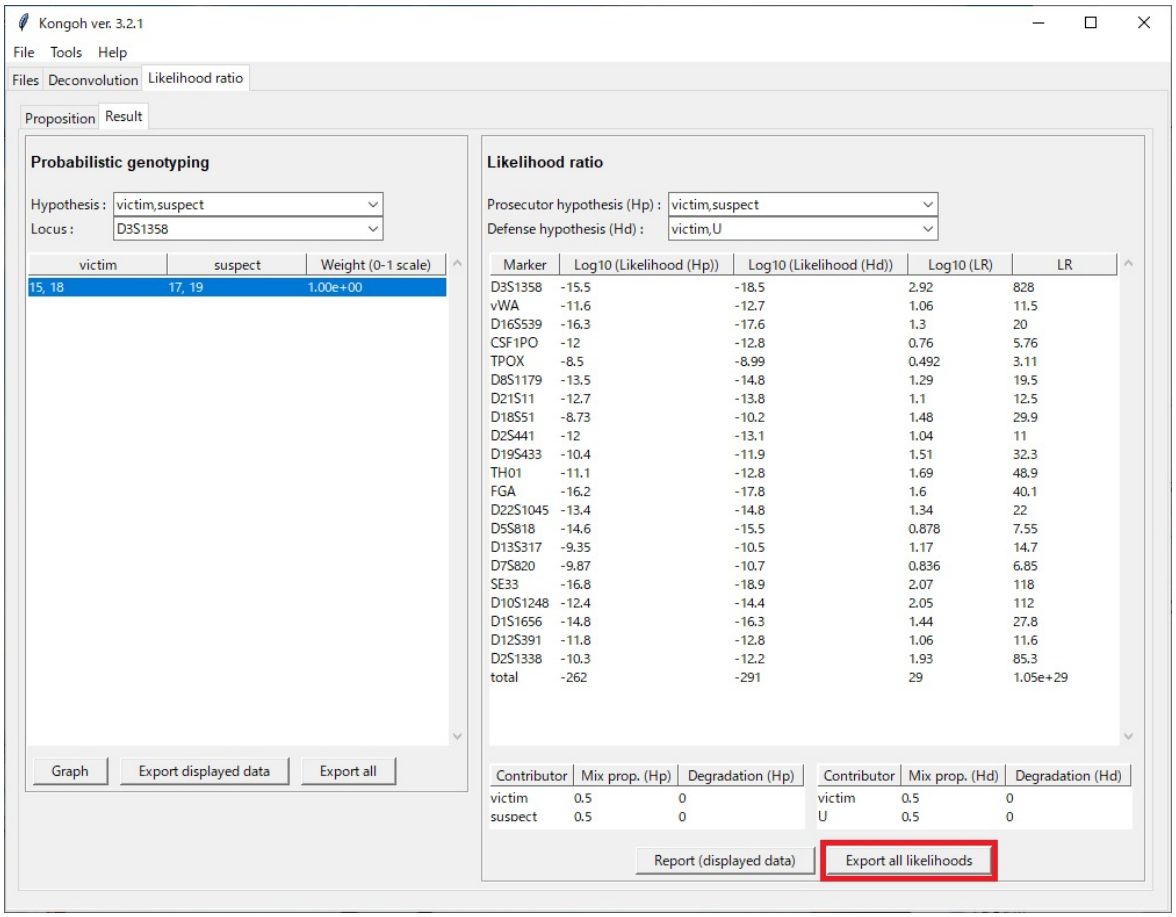
   - The likelihood ratio, the estimated mixture proportion, and the estimated degradation parameter under the selected hypotheses are displayed.



- The report of the likelihood ratio for displayed data can be exported by clicking the ”Report (displayed data)” button.



- Likelihoods and LRs of all assumed hypotheses can be exported by clicking the ”Export all likelihoods” button.

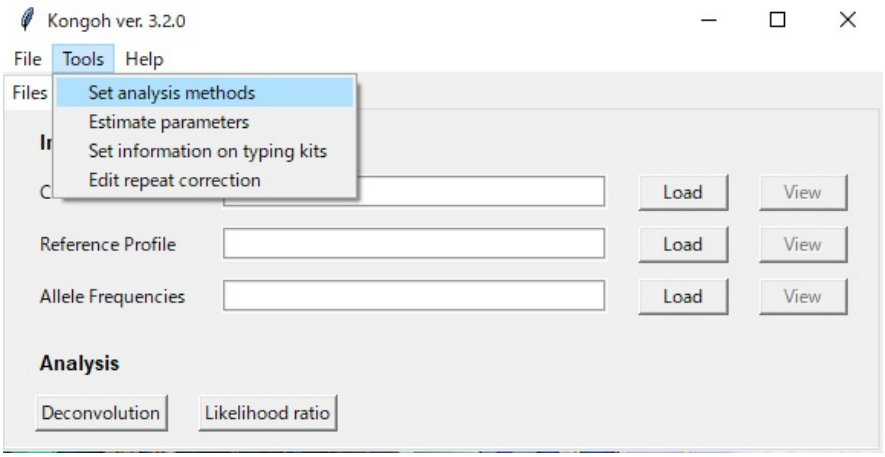


Set analysis methods

Since ver. 3.2.0, users must create an analysis method to determine the conditions for performing deconvolution and assigning likelihood ratios. This section describes how to create and edit an analysis method.

Create an analysis method

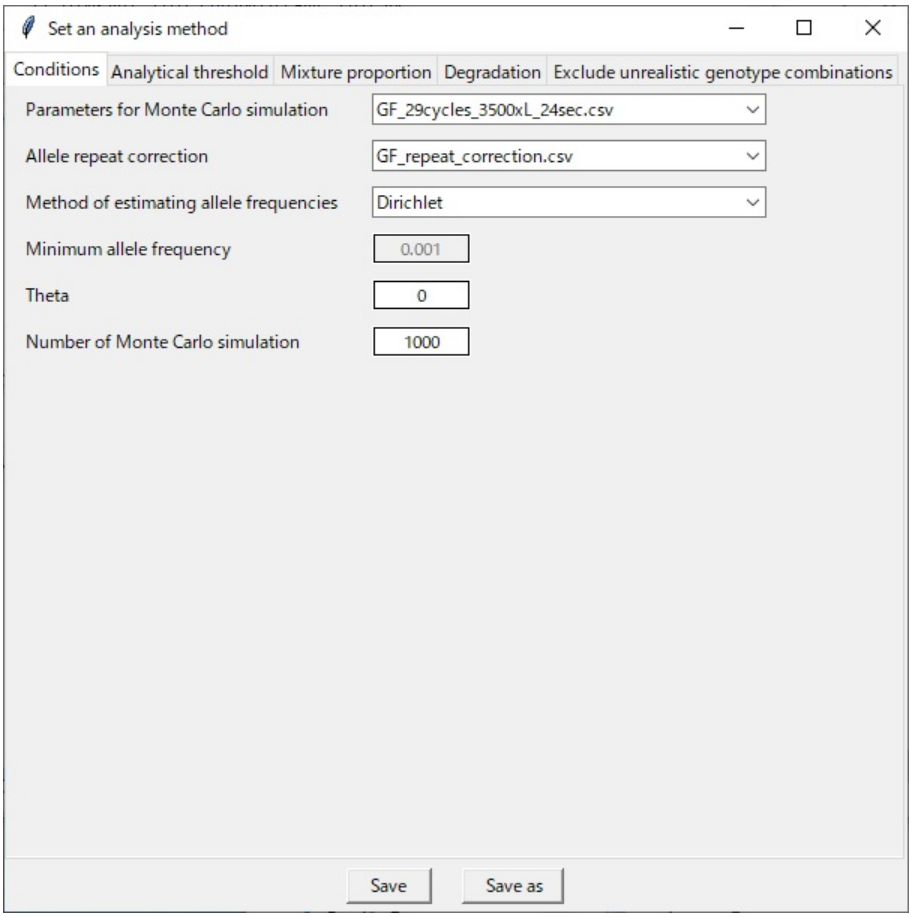
1. Go to Tools > Set analysis methods.



2. Click the "New" button to create an analysis method.

3. Enter or select the settings shown in the figures on the following pages.

"Conditions" tab



Parameters for Monte Carlo simulation

Select a file of the parameters for Monte Carlo simulation. The file can be generated in the function of "Estimate parameters"[4]. The default files for Identifiler Plus (IDP_28cycles_3130xl_10sec.csv)[5] and GlobalFiler (GF_29cycles_3500xL_24sec.csv)[6] are provided in the package *Kongoh*.

[4]See the section "Estimate parameters".

[5]Default parameters of Identifiler Plus were estimated using 392 single-source profiles, which were obtained from publicly available datasets in the Project Research Openness for Validation with Empirical Data (PROVEDIt: https://lftdi.camden.rutgers.edu/provedit/files/). These profiles were derived from non-degraded DNA of 50 individuals. The DNA amount of each sample was 0.0078–0.73 ng. The DNA samples were amplified at 28 cycles, and PCR products were analyzed on an Applied Biosystems^{TM} 3130xl Genetic Analyzer (Thermo Fisher Scientific) with injection parameters set at 3 kV for 10 s.

[6]Default parameters of GlobalFiler were estimated using 300 single-source profiles. These profiles were derived from pristine DNA of 50 individuals. The extracted DNA was diluted to 0.1, 0.05, 0.025, 0.0125, 0.00625, and 0.003125 ng/ L. A total of 300 diluted DNA solutions were amplified in 29 cycles using the GlobalFiler kit. Subsequently, PCR products were analyzed on an Applied Biosystems^{TM} 3500xL Genetic Analyzer (Thermo Fisher Scientific) with injection parameters set at 1.2 kV for 24 s in the Data Collection Software v3.1 (Thermo Fisher Scientific). See Manabe et al., 2021[5].

Allele repeat correction

Select a file of the allele repeat correction. The file can be generated in the function of "Estimate parameters". The default files for Identifiler Plus (IDP_repeat_correction.csv) and GlobalFiler (GF_repeat_correction.csv) are provided in the package *Kongoh*.

Method of estimating allele frequencies

If the allele frequencies of input files are represented as the allele counts, select "Dirichlet". If the allele frequencies of input files are represented as the allele probabilities, select "No correction".

Minimum allele frequency

The minimum allele frequency is applied to the frequency of the unobserved alleles when the "Method of estimating allele frequencies" is set to "No correction".

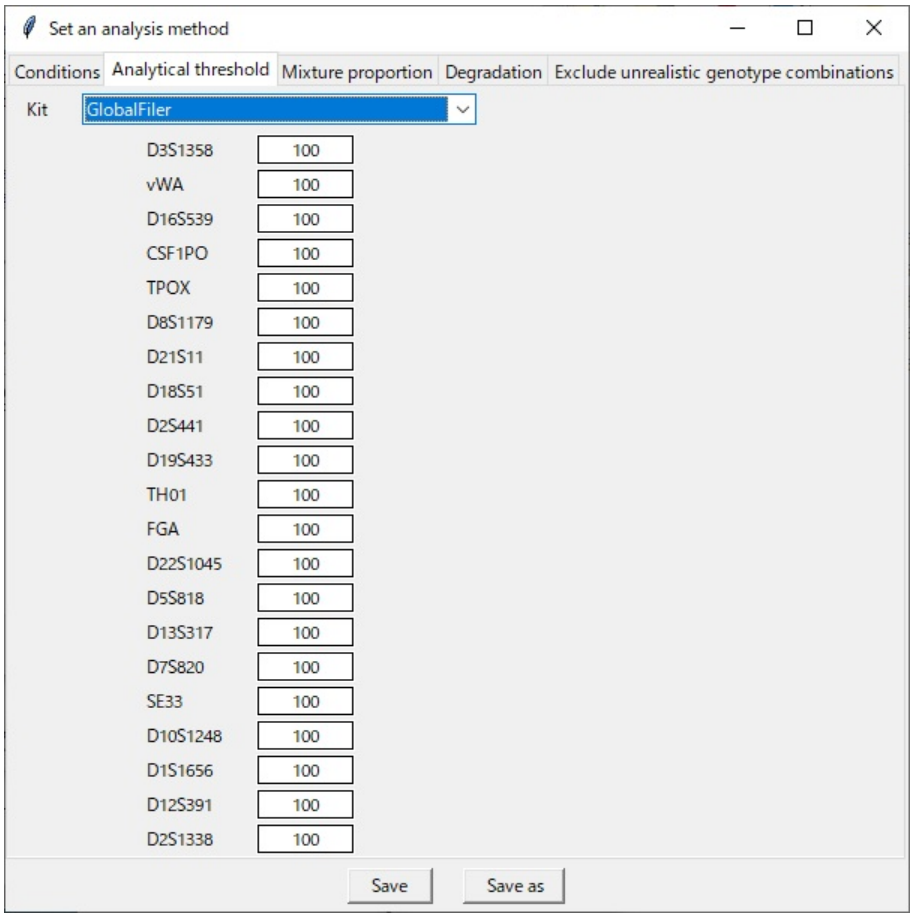
Theta

The theta value is used to consider the subpopulation effect.

Number of Monte Carlo simulation

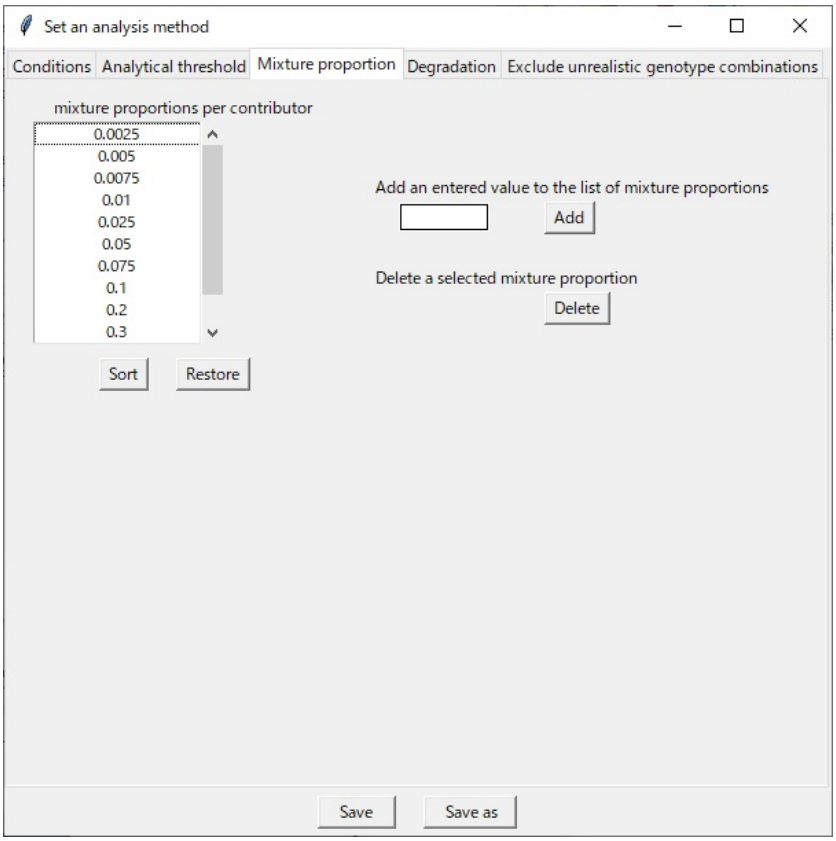
The default value is 1,000, which is considered sufficient for simulations based on developmental validation. If the number of simulations is increased, then the result becomes more robust; however, the runtime increases as well.

”Analytical threshold” tab



Select a kit, then enter the analytical thresholds of each locus.

”Mixture proportion” tab



Users can manually add or delete the mixture proportion per contributor. MP_n is the DNA proportion of the contributor $n(n = 1, 2, ..., N)$ in a crime stain profile, where N is the number of contributors. In Kongoh, MP_n must satisfy the following conditions:

$\sum_{n=1}^{N} MP_n = 1,$

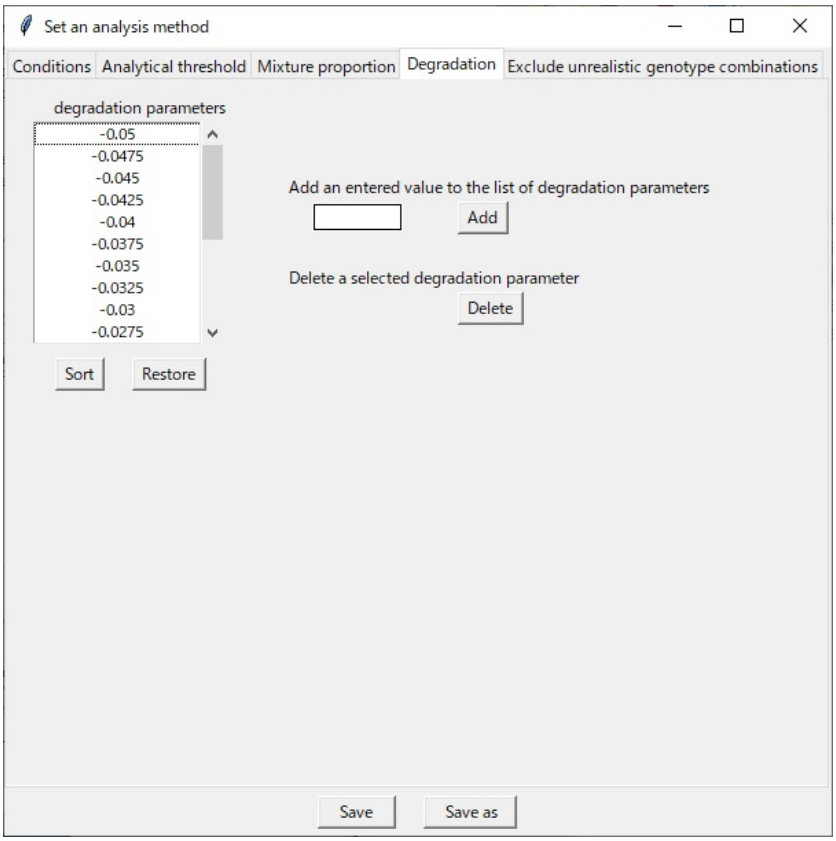
$0 < MP_1 \leq MP_2 \leq ... \leq MP_{N-1} \leq MP_N (N \geq 2).$

In the ”Mixture proportion” tab, candidate values for $MP_n(n = 1, 2, ..., N-1)$ can be set. MP_N is calculated as $1 - \sum_{n=1}^{N-1} MP_n$ in Kongoh.

Note

In addition to the sets of MP_n calculated by the above mentioned method, a set of the same amount of DNA in each contributor is automatically considered in Kongoh (e.g., $MP_1 = 0.33$, $MP_2 = 0.33$, $MP_3 = 0.34$ for $N = 3$).

”Degradation” tab

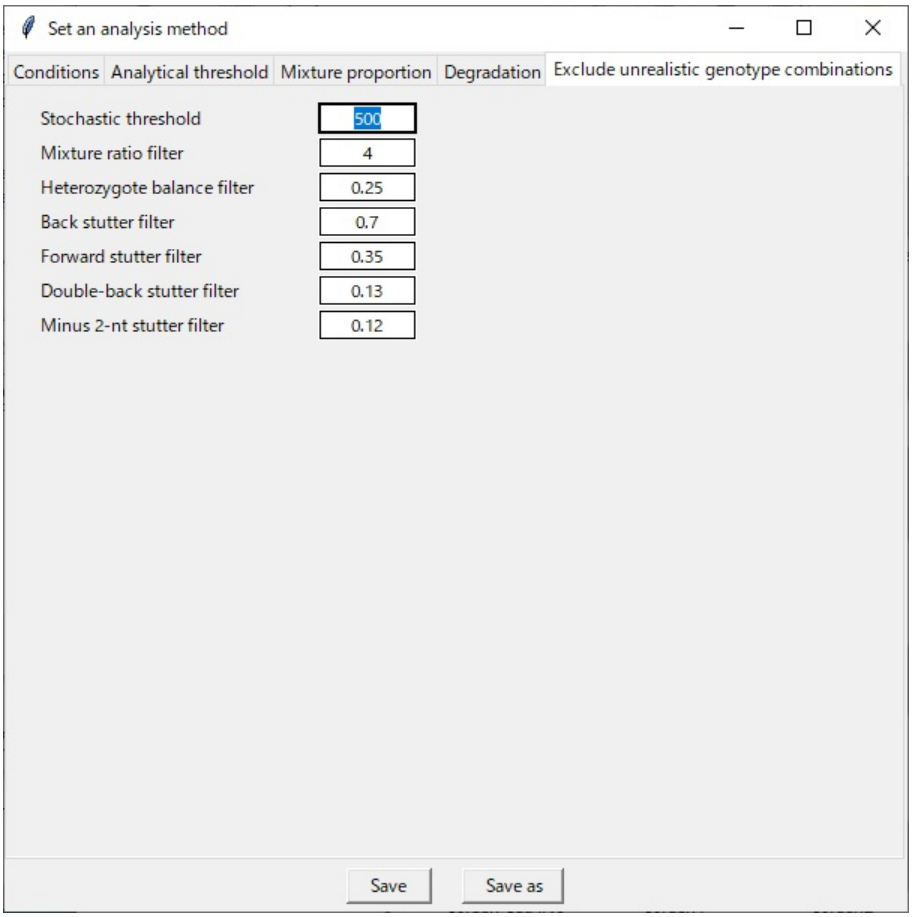


Users can manually add or delete the degradation parameters (d). $d = 0$ means no degradation. The smaller the d value, the greater the degree of degradation.

Note

The default range of d is $[-0.05, 0]$, which is discretized into 20 intervals of width 0.0025. When $d = -0.05$, DNA has significantly degraded. For example, when an allelic peak height located at the 100 base is 4,000 RFU, the peak height of the other allele located at the 225 base (which is approximately the middle detection point in the Identifiler Plus system) is expected to be only 8 RFU (i.e., typically less than the analytical threshold).

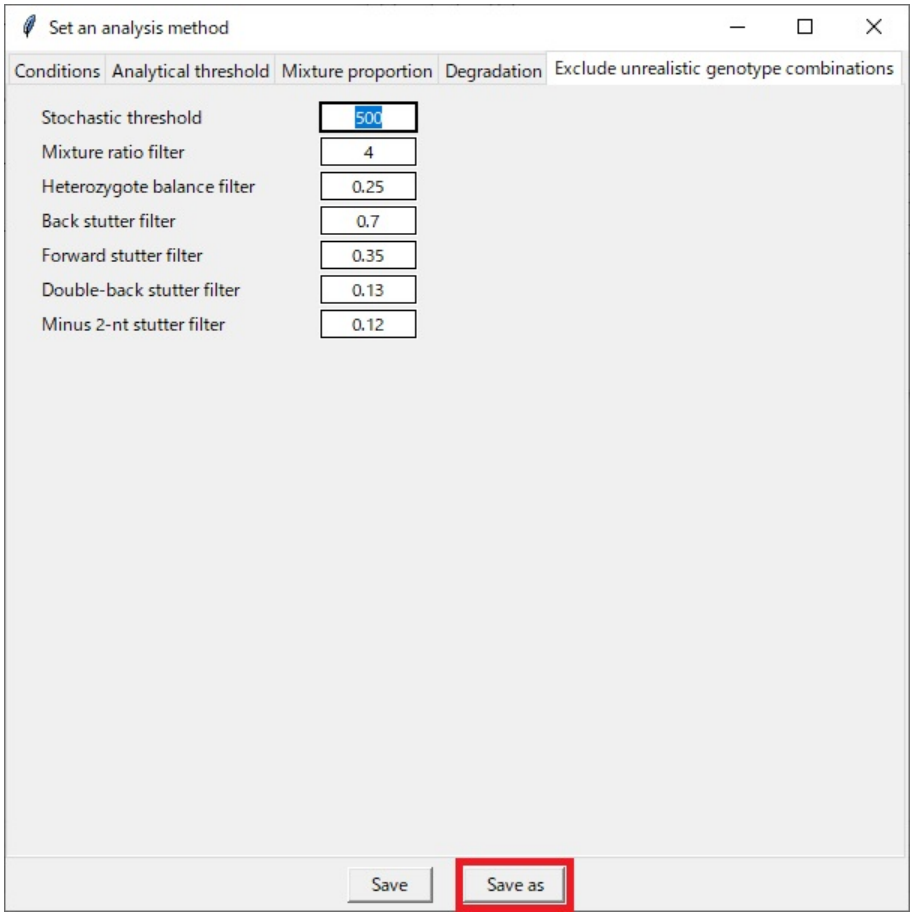
"Exclude unrealistic genotype combinations" tab



- **Stochastic threshold**: The mixture ratio filter and the heterozygote balance filter do not work if one of the peaks is below the threshold.
- **Mixture ratio filter**: A candidate genotype combination of minor and major contributors is inappropriate when the peak height of the minor (h_{minor}) is much larger than that of the major (h_{major}). If h_{minor}/h_{major} is greater than the mixture ratio filter, the genotype combination is excluded. h_{minor} and h_{major} are exactly the mean peak heights of the unique alleles in each.
- **Heterozygote balance filter**: If the peak height ratio of the two arbitrary peaks is smaller than the set value, the two peaks are not regarded as the heterozygote derived only from a single contributor.
- **Back stutter filter**: If the back stutter ratio is greater than the set value, the peak in the position of the back stutter is not regarded as the product derived only from the back stutter.
- **Forward stutter filter**: If the forward stutter ratio is greater than the set value, the peak in the position of the forward stutter is not regarded as the product derived only from the forward stutter.

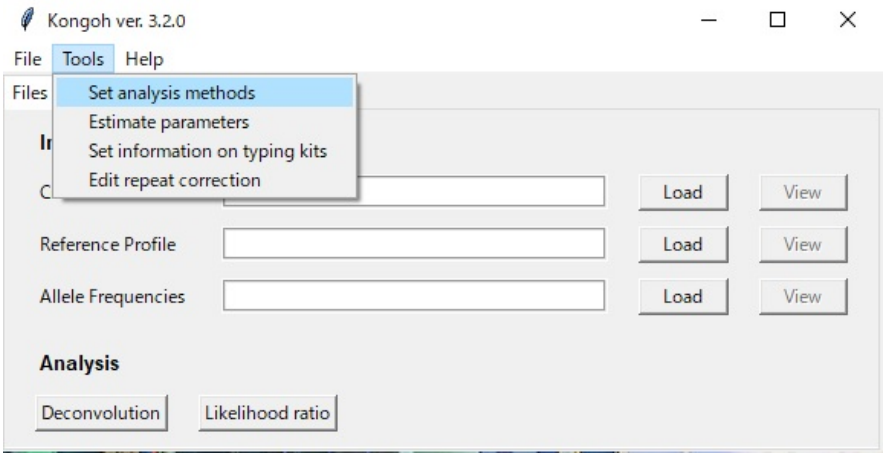
- **Double-back stutter filter**: If the double-back stutter ratio is greater than the set value, the peak in the position of the double-back stutter is not regarded as the product derived only from the double-back stutter.
- **Minus 2-nt stutter filter**: If the minus 2-nt stutter ratio is greater than the set value, the peak in the position of the minus 2-nt stutter is not regarded as the product derived only from the minus 2-nt stutter.

4. Click the ”Save as” button after finishing all the settings.

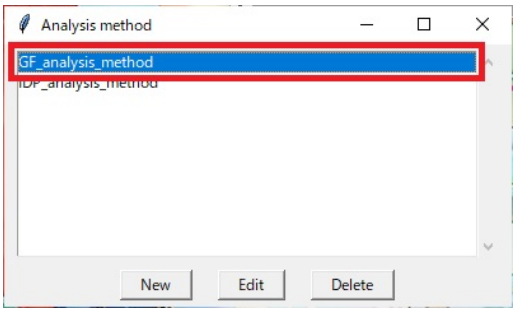


Edit an analysis method

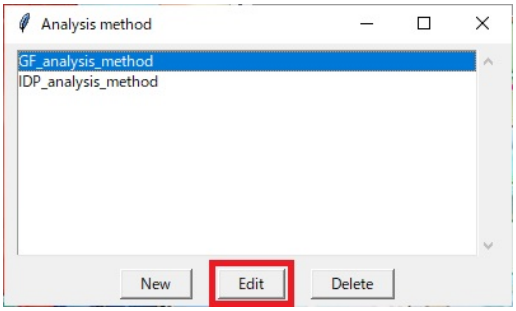
1. Go to Tools > Set analysis methods.



2. Select an analysis method.

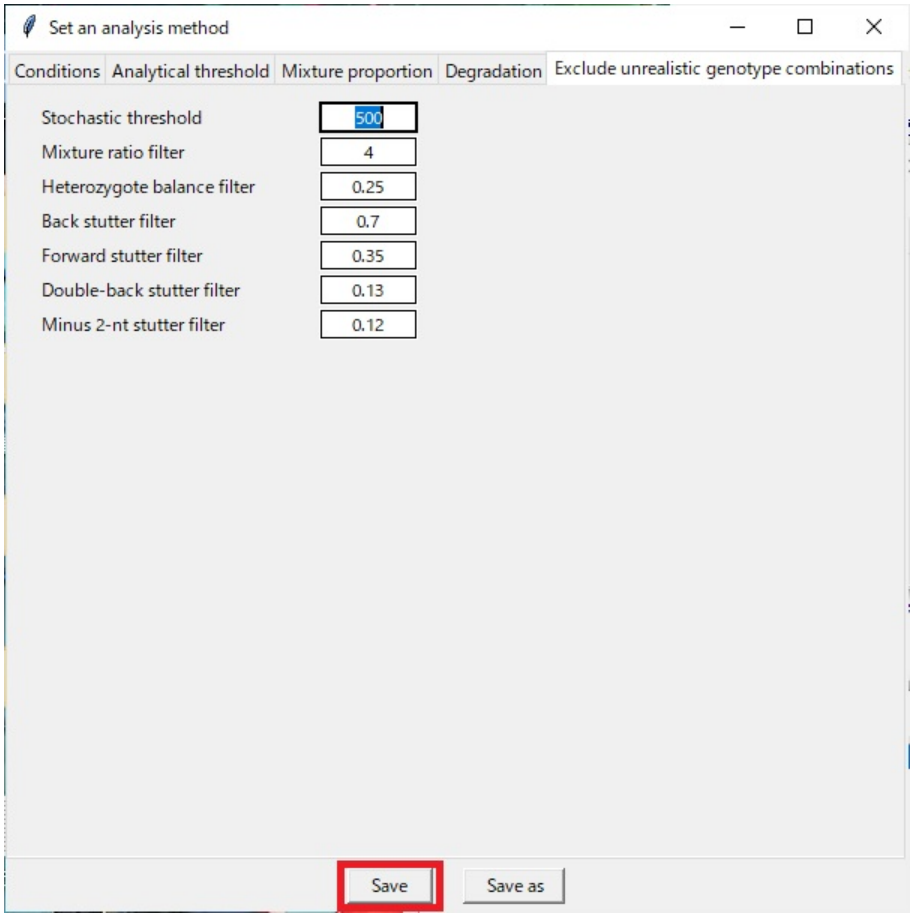


3. Click the "Edit" button to edit the selected analysis method.



4. Enter or select the settings just like creating an analysis method.

5. Click the ”Save” button after finishing all the settings.

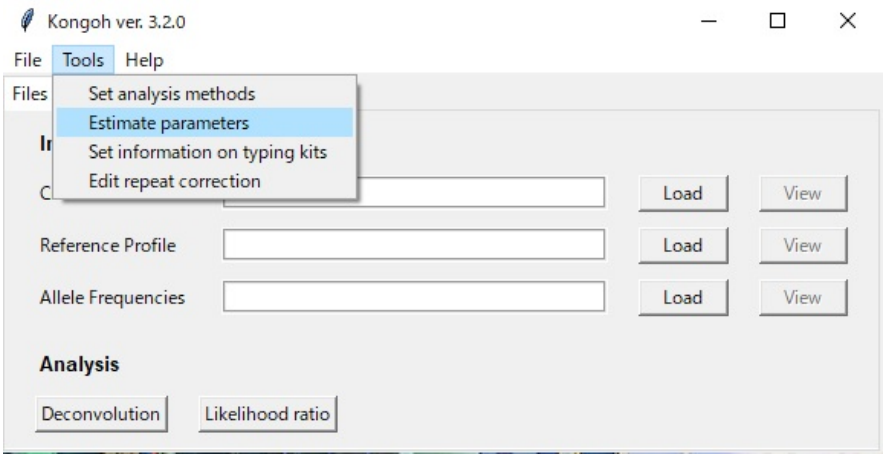


Estimate parameters

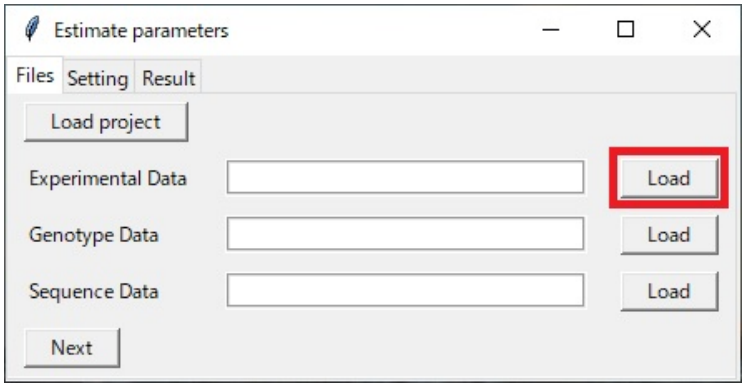
In Kongoh, probability distribution models for locus-specific amplification efficiency (AE), heterozygote balance (Hb), back stutter ratio (BSR), forward stutter ratio (FSR), double-back stutter ratio (DSR), and minus 2-nt stutter ratio (M2SR) can be estimated using experimental data prepared by users. The maximum likelihood estimation (MLE) for each parameter of the probability distribution models (e.g., mean, variance, etc.) is performed by the generalized simulated annealing, which is implemented into the R package "GenSA". In this section, the procedure for the estimation of parameters is described.

Input files

1. Go to Tools > Estimate parameters.



2. Load a file of the experimental data.



Note

The experimental data should satisfy the following conditions:

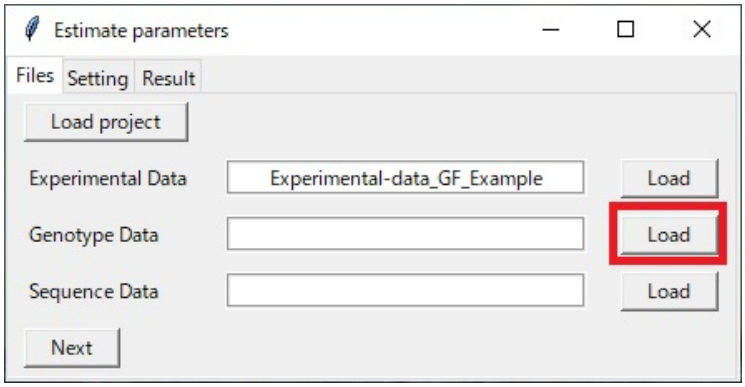
- The data are composed of single-source profiles.
- The profiles are derived from pristine (non-degraded) DNA samples.
- DNA samples should be prepared to the maximum extent, including samples of various amounts of DNA and various individuals.
- All DNA samples are analyzed based on one protocol.
- No filters are to be used to remove stutter peaks when analyzing sample electrophoretic data in software programs such as the GeneMapper® ID-X software.
- It is desirable to remove pull-up peaks and noises manually, but not required[7].

Note

The example file "Experimental-data_GF_Example.csv" is provided. Go to extdata > example in the package *Kongoh*.

- The file of the experimental data can be exported from the GeneMapper® ID-X software.
- The file of the experimental data must include information regarding the "Sample Name", "Marker", "Allele", "Size", and "Height".

3. Load a file of the genotype data.



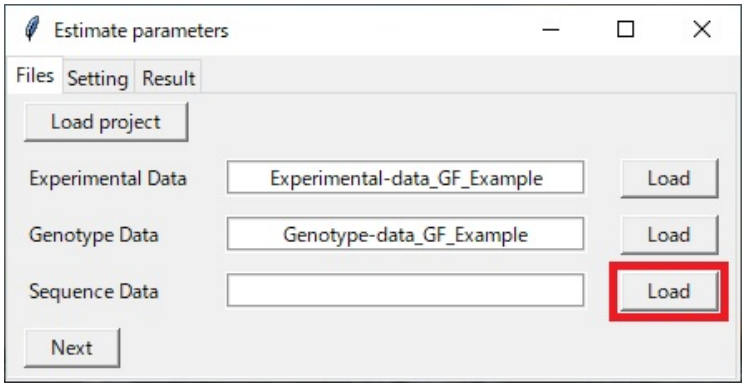
[7]Kongoh can automatically determine allele and stutter peaks based on the genotypes of each experimental data, as well as remove peaks located within ±1 base[6] of an allele peak(s) in different color channels to exclude the effect of the spectral pull-up.

Note

The example file "Genotype-data_GF_Example.csv" is provided. Go to extdata > example in the package *Kongoh*.

- This file must include information regarding the "Sample Name", "Marker", "Allele 1", and "Allele 2".
- Information regarding the "Sample Name" in the "Genotype Data" file must be the same as that in the "Experimental Data" file.
- Two alleles in homozygotes must be entered in each column (except for Yindel and DYS391).

4. Load a file of the sequence data.



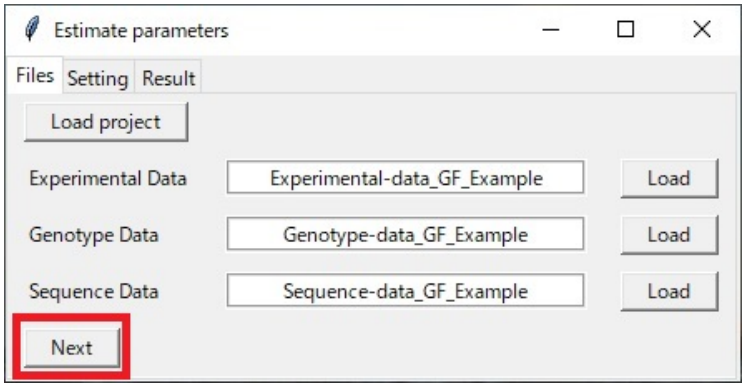
Note

The example file "Sequence-data_GF_Example.csv" is provided[8]. Go to extdata > example in the package *Kongoh*.

- This file must include information regarding the "Marker", "Allele", "Repeat Region", and "Count". Information on the "Flanking Region" is optional.
- "Count" implies the number of observations in sequence-based population data.

[8]The example data is derived from sequence-based U.S. population data for SE33[7] and for loci other than SE33[8].

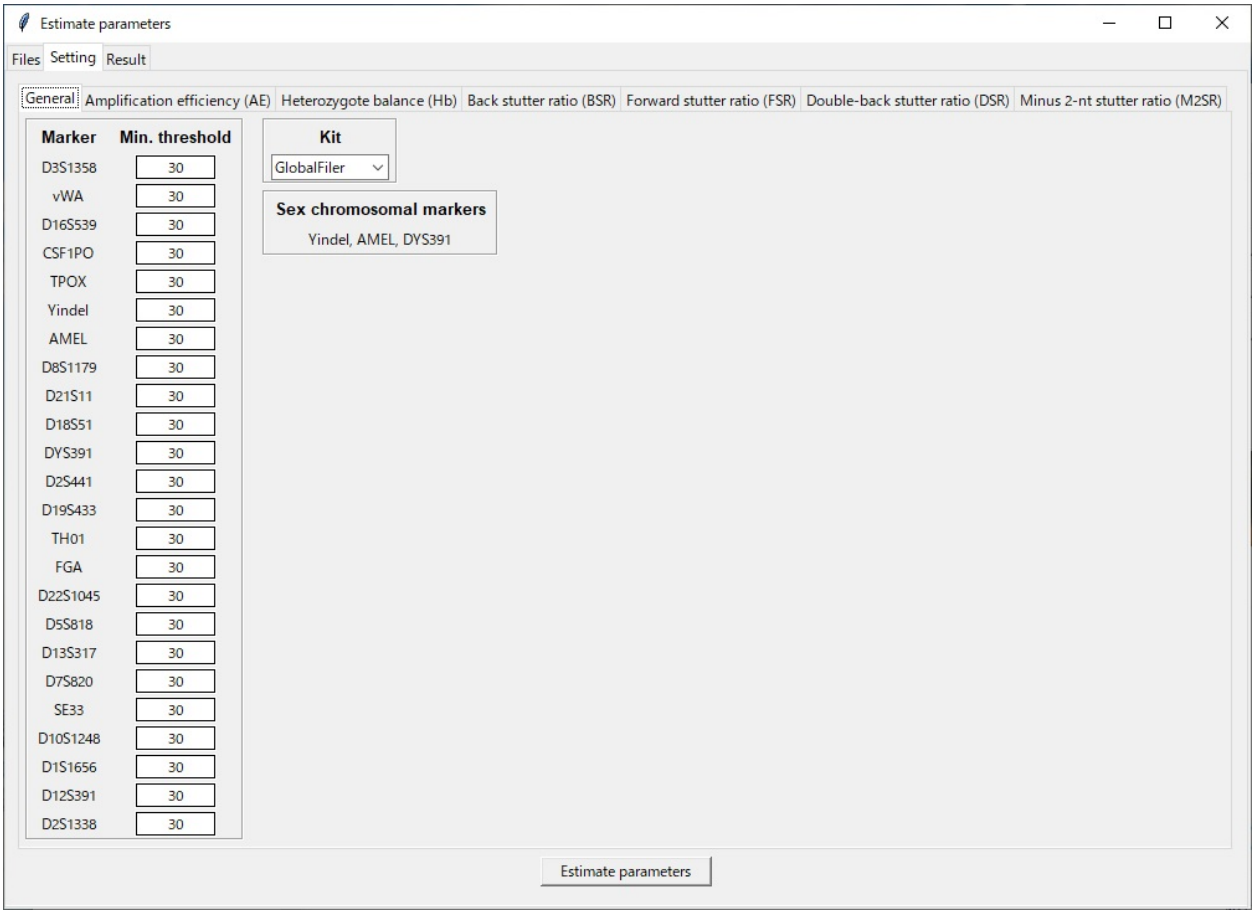
5. Click the ”Next” button. Then the ”Setting” tab will be automatically opened.



Set conditions

1. Set conditions shown in the figures on the following pages.

   "General" tab



Min. threshold

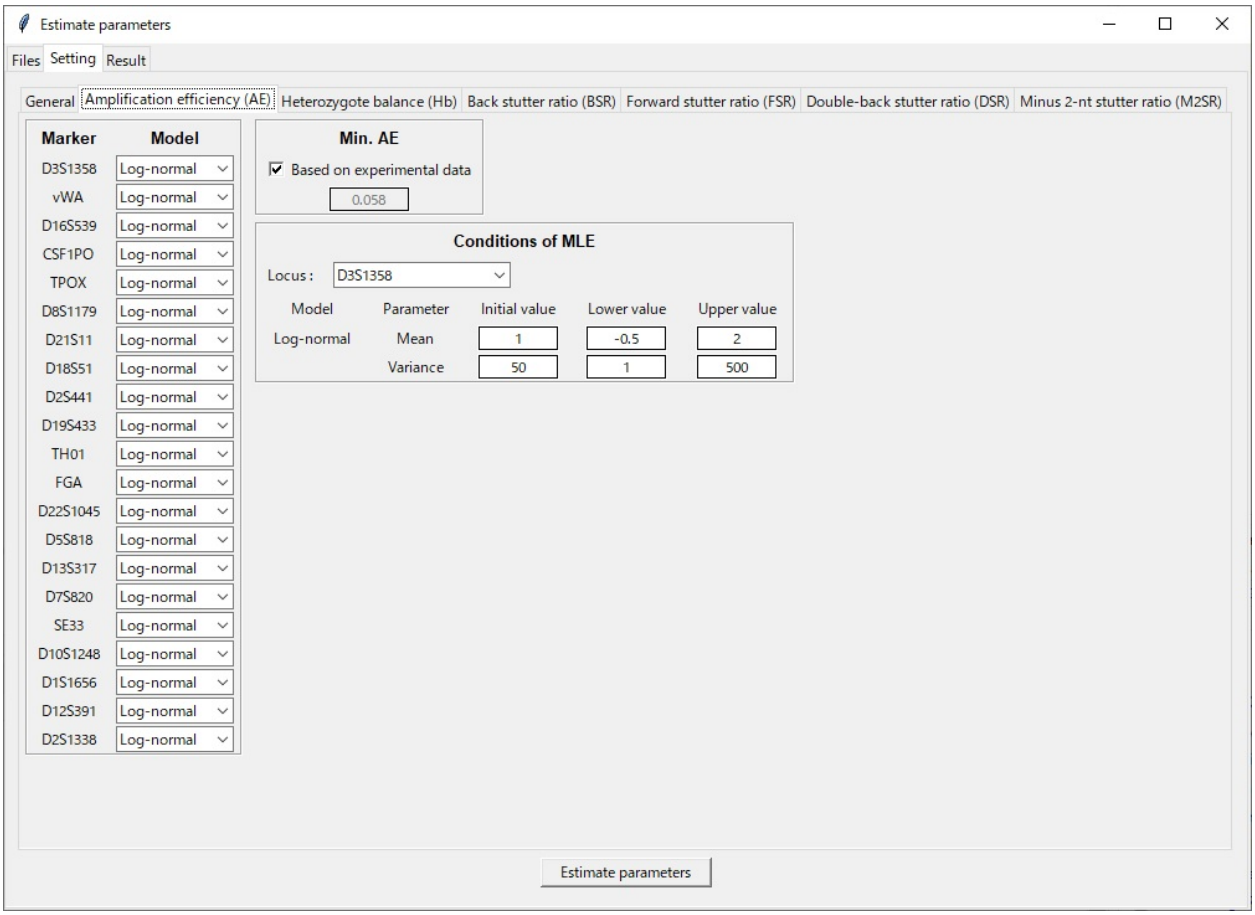
Peaks below the thresholds are not used to estimate the parameters.

Kit

The kit name is automatically selected based on the input files. Users should select the appropriate kit name when more than one kit with the same locus set is registered in the feature "Edit information on typing kits[9]" (e.g., Identifiler and Identifiler Plus).

[9]See the section "Edit information on typing kits" in the user manual.

”Amplification efficiency (AE)” tab



Model

The model for AE is fixed to the “Log-normal” distribution in the current version of Kongoh.

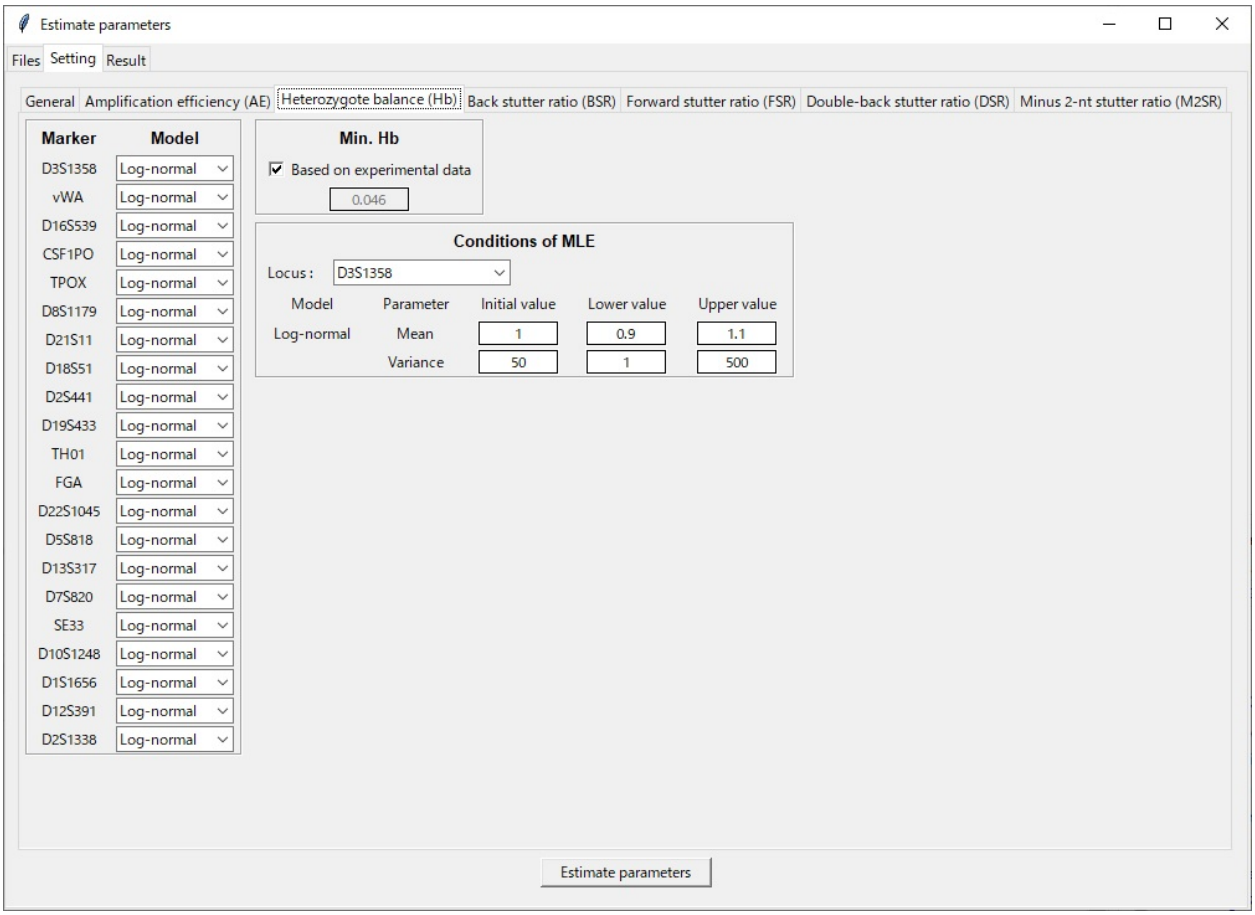
Min. AE

Min. AE is the lower limit of the truncated log-normal distribution for AE. The upper limit is automatically set to 1 / Min. AE. If ”Based on experimental data” is checked, Min. AE is set to the minimum of the experimental AE values. If ”Based on experimental data” is unchecked, users can enter an arbitrary value for Min. AE.

Conditions of MLE

Users can set initial, lower, and upper values of each parameter in the maximum likelihood estimation (MLE). These values do not need to be changed in the initial analysis of MLE. If the estimated probability distribution does not fit to the experimental data, users may try to change these values.

”Heterozygote balance (Hb)” tab



Model

The model for Hb is fixed to the “Log-normal” distribution in the current version of Kongoh.

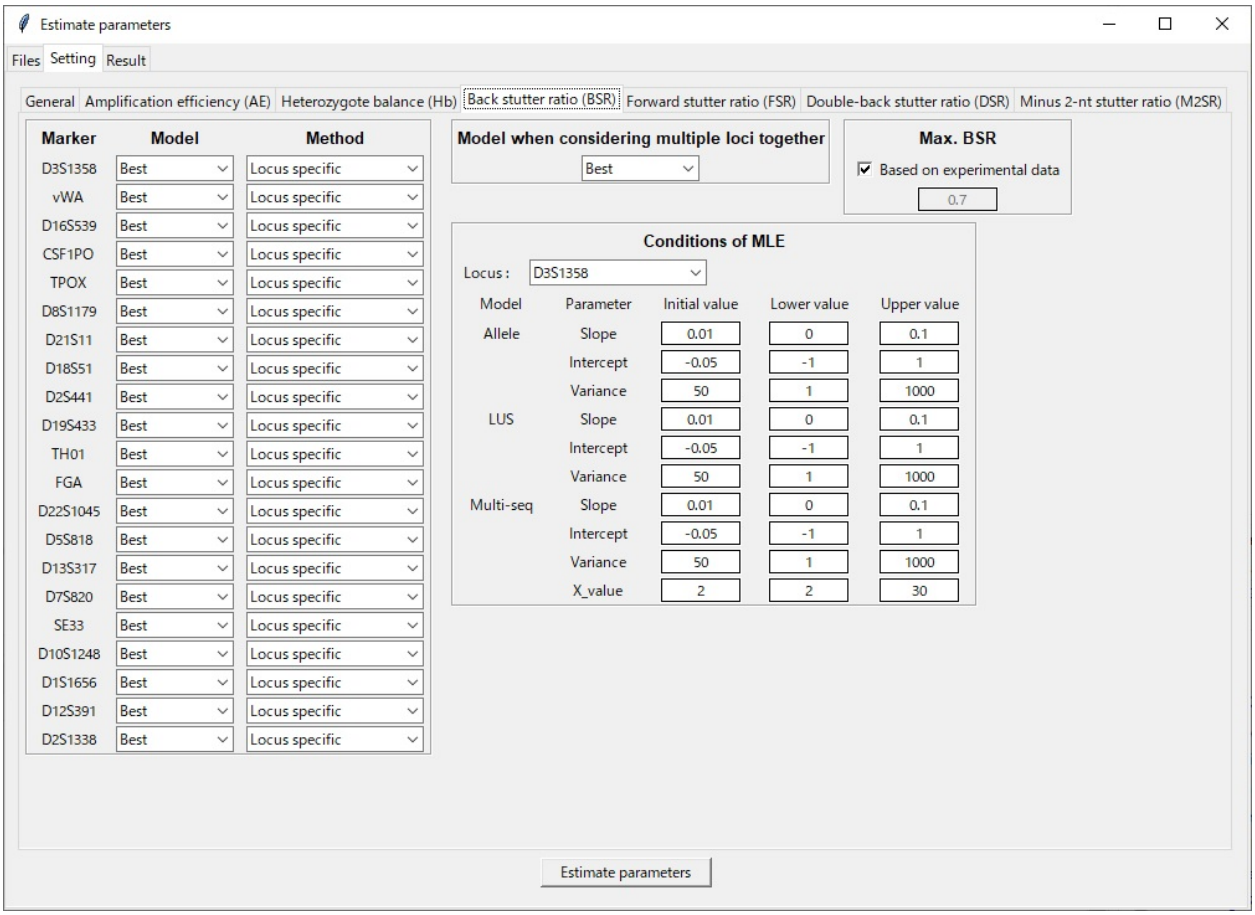
Min. Hb

Min. Hb is the lower limit of the truncated log-normal distribution for Hb. The upper limit is automatically set to 1 / Min. Hb. If ”Based on experimental data” is checked, Min. Hb is set to the minimum of the experimental Hb values. If ”Based on experimental data” is unchecked, users can enter an arbitrary value for Min. Hb.

Conditions of MLE

Users can set initial, lower, and upper values of each parameter in the maximum likelihood estimation (MLE). These values do not need to be changed in the initial analysis of MLE. If the estimated probability distribution does not fit to the experimental data, users may try to change these values.

"Back stutter ratio (BSR)" tab



Model

There are four options for the model for BSR.

- **Best**: The best model is selected from Allele, LUS, and Multi-seq models based on the AIC (Akaike information criterion) values of each model.
- **Allele**: The BSRs are positively correlated with the allele numbers[9].
- **LUS**: The BSRs are positively correlated with the longest uninterrupted stretch (LUS) values[10, 11].
- **Multi-seq**: The BSRs are positively correlated with the corrected allele numbers, which consider not only the LUS, but also other uninterrupted stretches[12].

Method

There are three options for the methods of modeling.

- **Locus specific**: The experimental data of the locus is only used for estimating the probability distribution for this locus.

- **Multiple loci together**: The experimental data of all loci with the option "Multiple loci together" is used for estimating the probability distribution for these loci.
- **Not consider**: The probability distribution model is not considered in the locus.

Model when considering multiple loci together

Users can select models from four options (i.e., Best, Allele, LUS, and Multi-seq when considering multiple loci together.

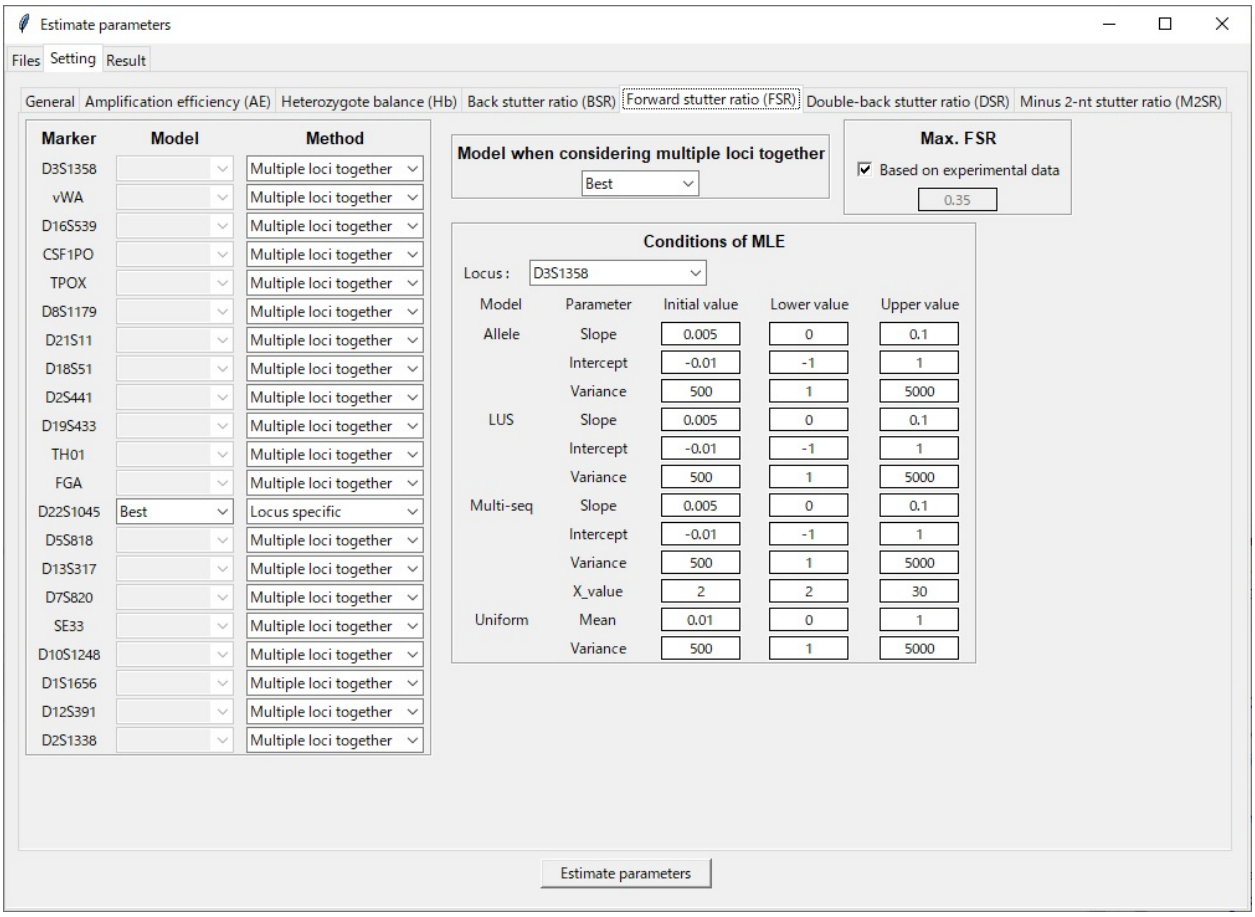
Max. BSR

Max. BSR is the upper limit of the truncated log-normal distribution for BSR. If "Based on experimental data" is checked, Max. BSR is set to the maximum of the experimental BSRs in all loci. If "Based on experimental data" is unchecked, users can enter an arbitrary value for Max. BSR.

Conditions of MLE

Users can set initial, lower, and upper values of each parameter in the maximum likelihood estimation (MLE). These values do not need to be changed in the initial analysis of MLE. If the estimated probability distribution does not fit to the experimental data, users may try to change these values.

"Forward stutter ratio (FSR)" tab



Model

There are five options for the model for FSR.

- **Best**: The best model is selected from Allele, LUS, Multi-seq, and Uniform models based on the AIC values of each model.
- **Allele**: The FSRs are positively correlated with the allele numbers.
- **LUS**: The FSRs are positively correlated with the longest uninterrupted stretch (LUS) values.
- **Multi-seq**: The FSRs are positively correlated with the corrected allele numbers, which consider not only the LUS, but also other uninterrupted stretches.
- **Uniform**: The mean value of the FSRs is uniform regardless of the allele number.

Method

There are three options for the methods of modeling.

- **Locus specific**: The experimental data of the locus is only used for estimating the probability distribution for this locus.

- **Multiple loci together**: The experimental data of all loci with the option "Multiple loci together" is used for estimating the probability distribution for these loci.
- **Not consider**: The probability distribution model is not considered in the locus.

Model when considering multiple loci together

Users can select models from five options (i.e., Best, Allele, LUS, Multi-seq, and Uniform when considering multiple loci together.

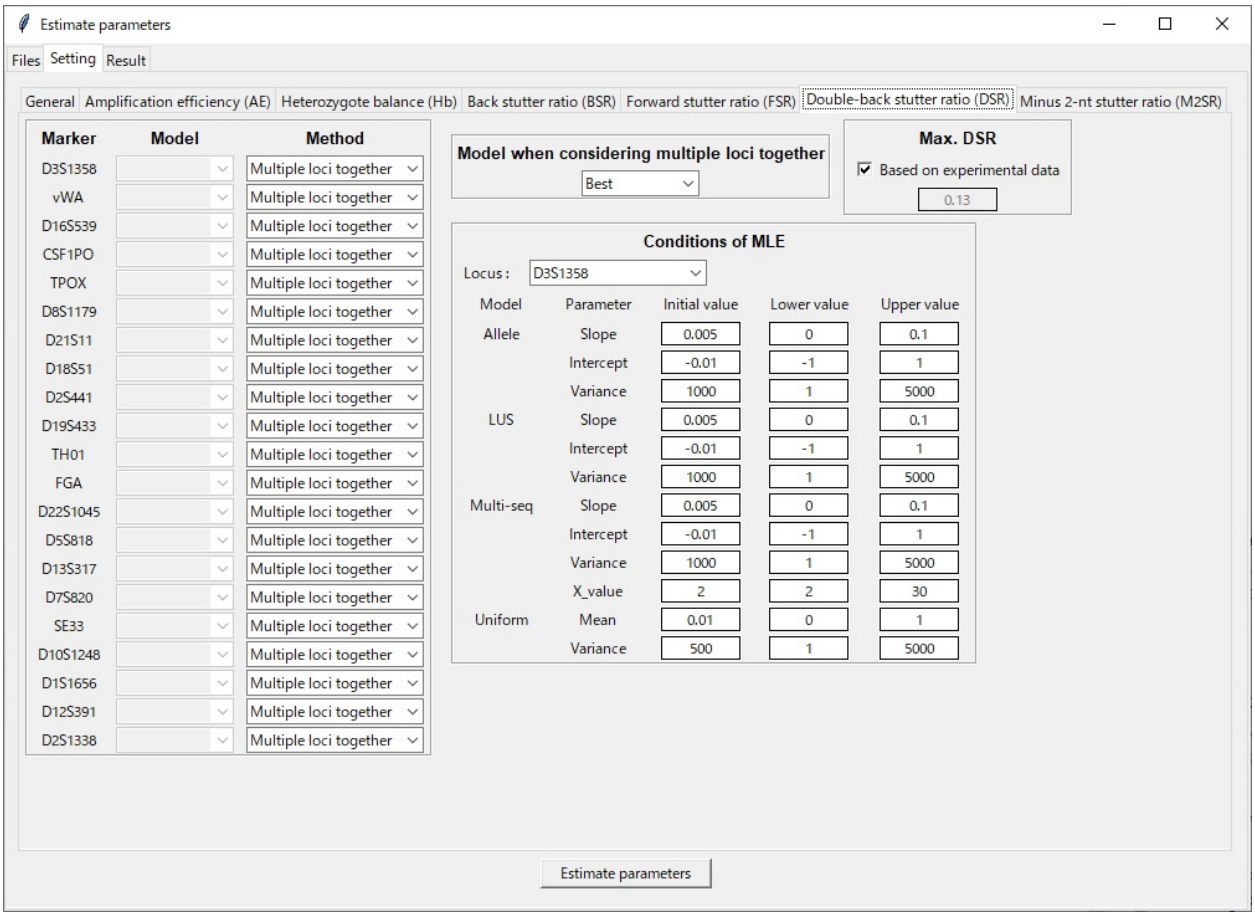
Max. FSR

Max. FSR is the upper limit of the truncated log-normal distribution for FSR. If "Based on experimental data" is checked, Max. FSR is set to the maximum of the experimental FSRs in all loci. If "Based on experimental data" is unchecked, users can enter an arbitrary value for Max. FSR.

Conditions of MLE

Users can set initial, lower, and upper values of each parameter in the maximum likelihood estimation (MLE). These values do not need to be changed in the initial analysis of MLE. If the estimated probability distribution does not fit to the experimental data, users may try to change these values.

"Double-back stutter ratio (DSR)" tab



Model

There are five options for the model for DSR.

- **Best**: The best model is selected from Allele, LUS, Multi-seq, and Uniform models based on the AIC values of each model.
- **Allele**: The DSRs are positively correlated with the allele numbers.
- **LUS**: The DSRs are positively correlated with the longest uninterrupted stretch (LUS) values.
- **Multi-seq**: The DSRs are positively correlated with the corrected allele numbers, which consider not only the LUS, but also other uninterrupted stretches.
- **Uniform**: The mean value of the DSRs is uniform regardless of the allele number.

Method

There are three options for the methods of modeling.

- **Locus specific**: The experimental data of the locus is only used for estimating the probability distribution for this locus.

- **Multiple loci together**: The experimental data of all loci with the option "Multiple loci together" is used for estimating the probability distribution for these loci.
- **Not consider**: The probability distribution model is not considered in the locus.

Model when considering multiple loci together

Users can select models from five options (i.e., Best, Allele, LUS, Multi-seq, and Uniform when considering multiple loci together.

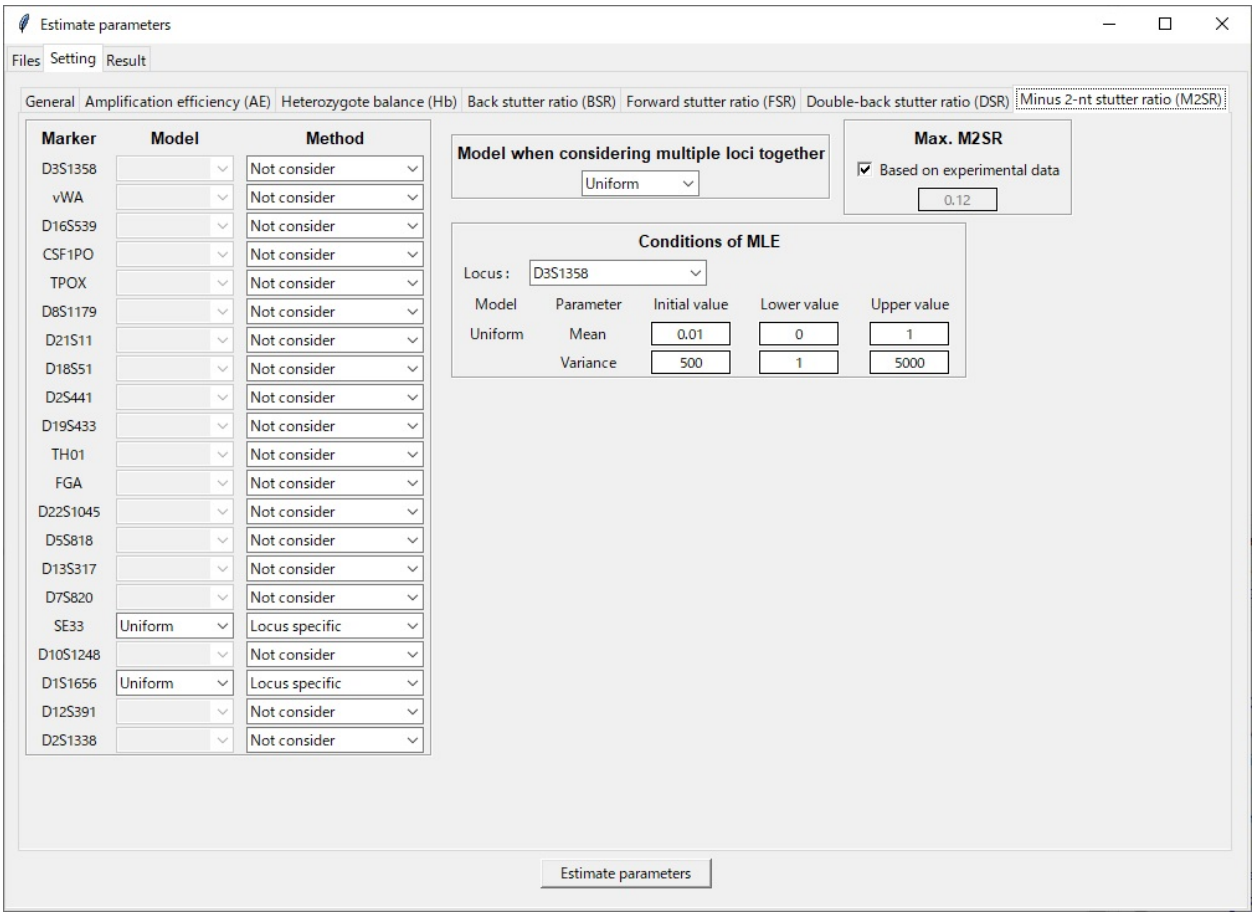
Max. DSR

Max. DSR is the upper limit of the truncated log-normal distribution for DSR. If "Based on experimental data" is checked, Max. DSR is set to the maximum of the experimental DSRs in all loci. If "Based on experimental data" is unchecked, users can enter an arbitrary value for Max. DSR.

Conditions of MLE

Users can set initial, lower, and upper values of each parameter in the maximum likelihood estimation (MLE). These values do not need to be changed in the initial analysis of MLE. If the estimated probability distribution does not fit to the experimental data, users may try to change these values.

”Minus 2-nt stutter ratio (M2SR)” tab



Model

The model for M2SR is fixed to the “Uniform” distribution in the current version of Kongoh.

Method

There are three options for the methods of modeling.

- **Locus specific**: The experimental data of the locus is only used for estimating the probability distribution for this locus.
- **Multiple loci together**: The experimental data of all loci with the option ”Multiple loci together” is used for estimating the probability distribution for these loci.
- **Not consider**: The probability distribution model is not considered in the locus.

Model when considering multiple loci together

The model for M2SR is fixed to the “Uniform” distribution in the current version of Kongoh when considering multiple loci together.

Max. M2SR

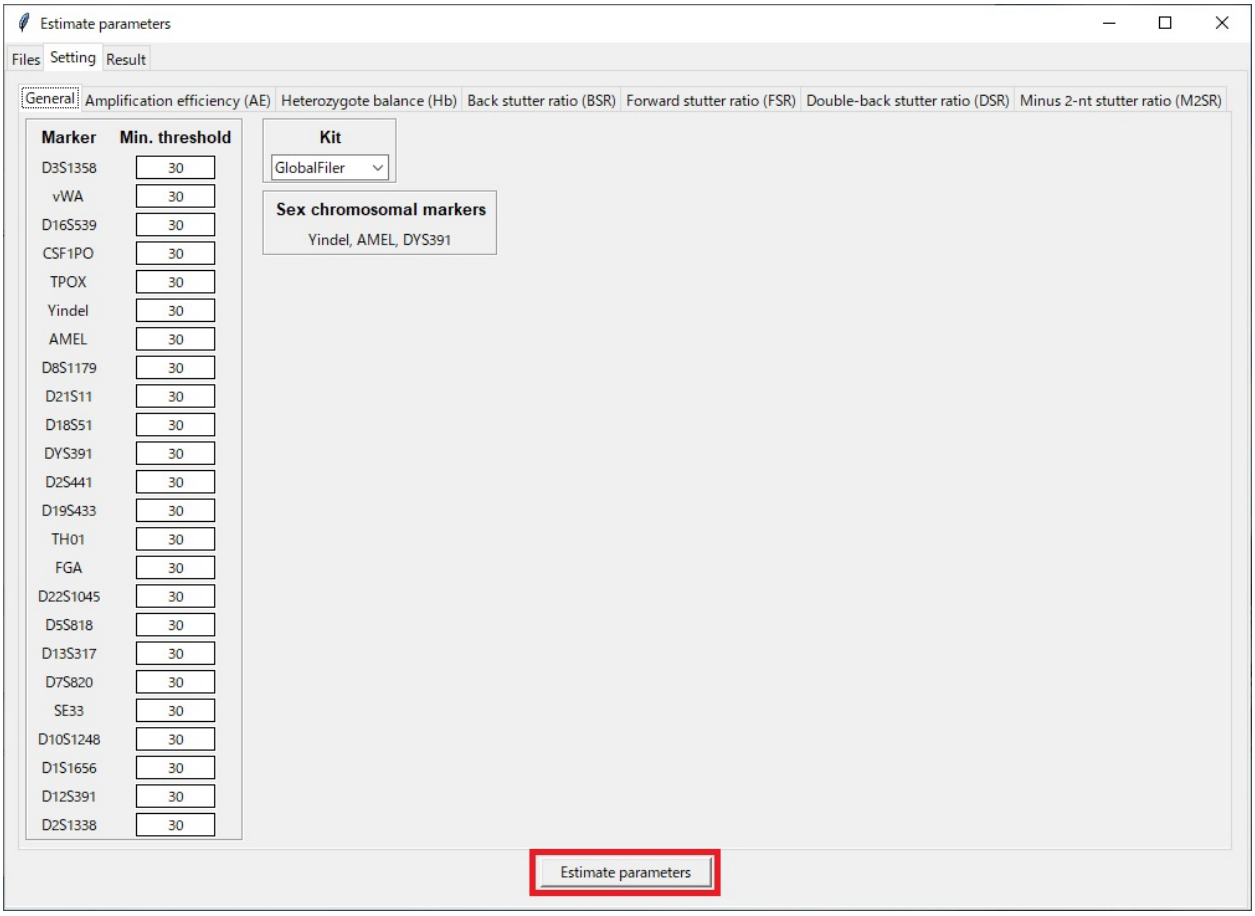
Max. M2SR is the upper limit of the truncated log-normal distribution for M2SR. If ”Based

on experimental data" is checked, Max. M2SR is set to the maximum of the experimental M2SRs in all loci. If "Based on experimental data" is unchecked, users can enter an arbitrary value for Max. M2SR.

Conditions of MLE

Users can set initial, lower, and upper values of each parameter in the maximum likelihood estimation (MLE). These values do not need to be changed in the initial analysis of MLE. If the estimated probability distribution does not fit to the experimental data, users may try to change these values.

2. Click the "Estimate parameters" button after setting all conditions.



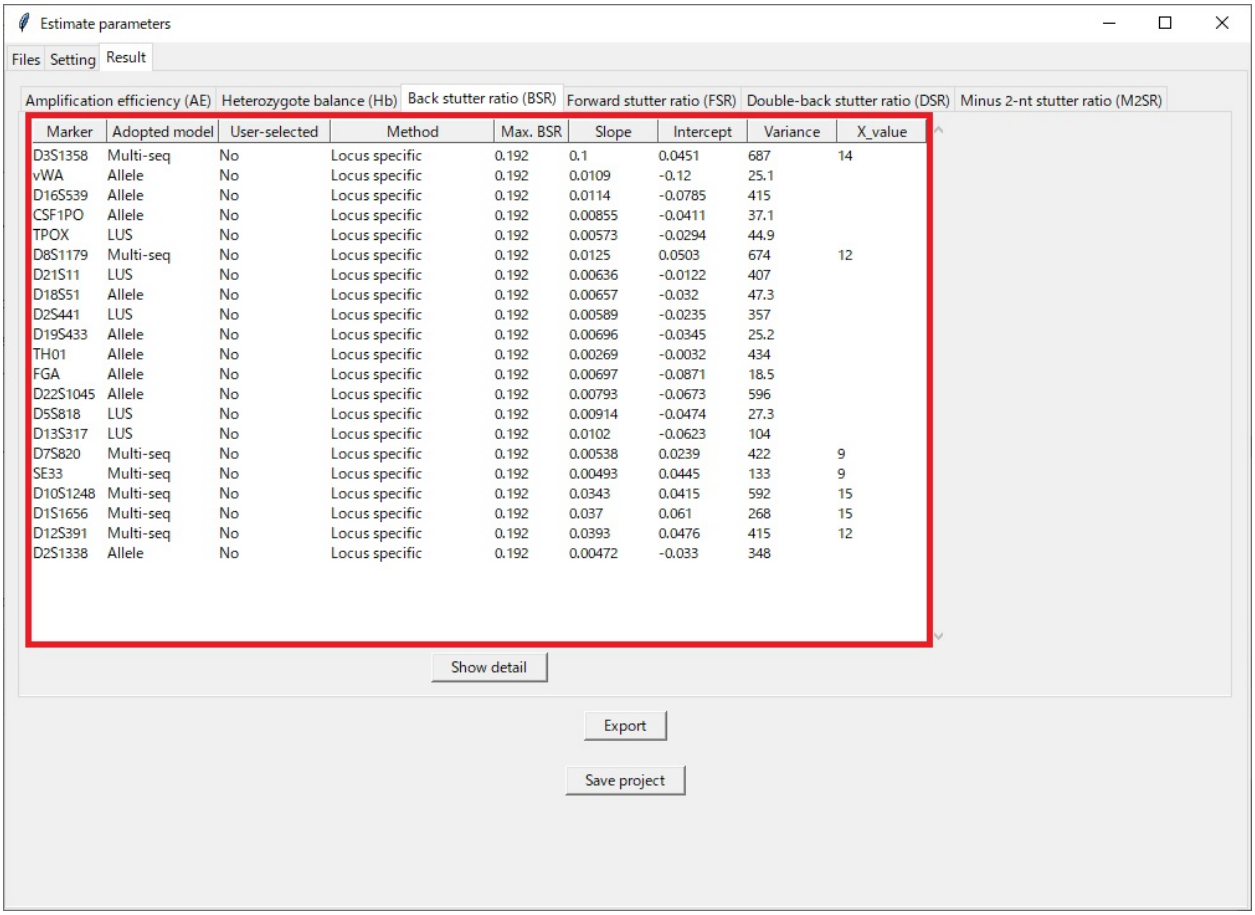
Note

It takes a few hours to estimate parameters. After the parameter estimation is completed, the "Result" tab appears automatically.

Review the results

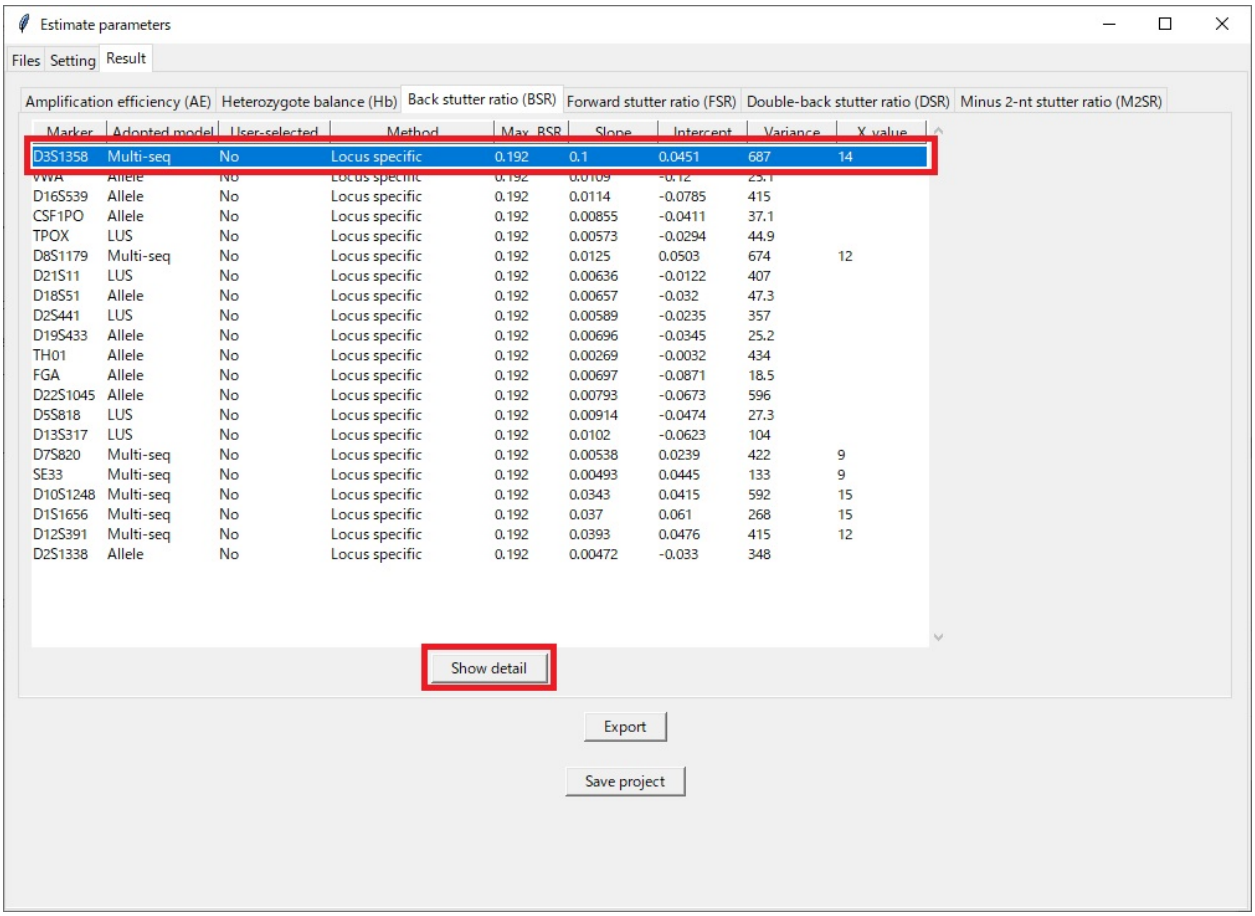
1. Review the results of estimated parameters.

    - Adopted models and estimated parameters in each locus are displayed on the left side of the screen.

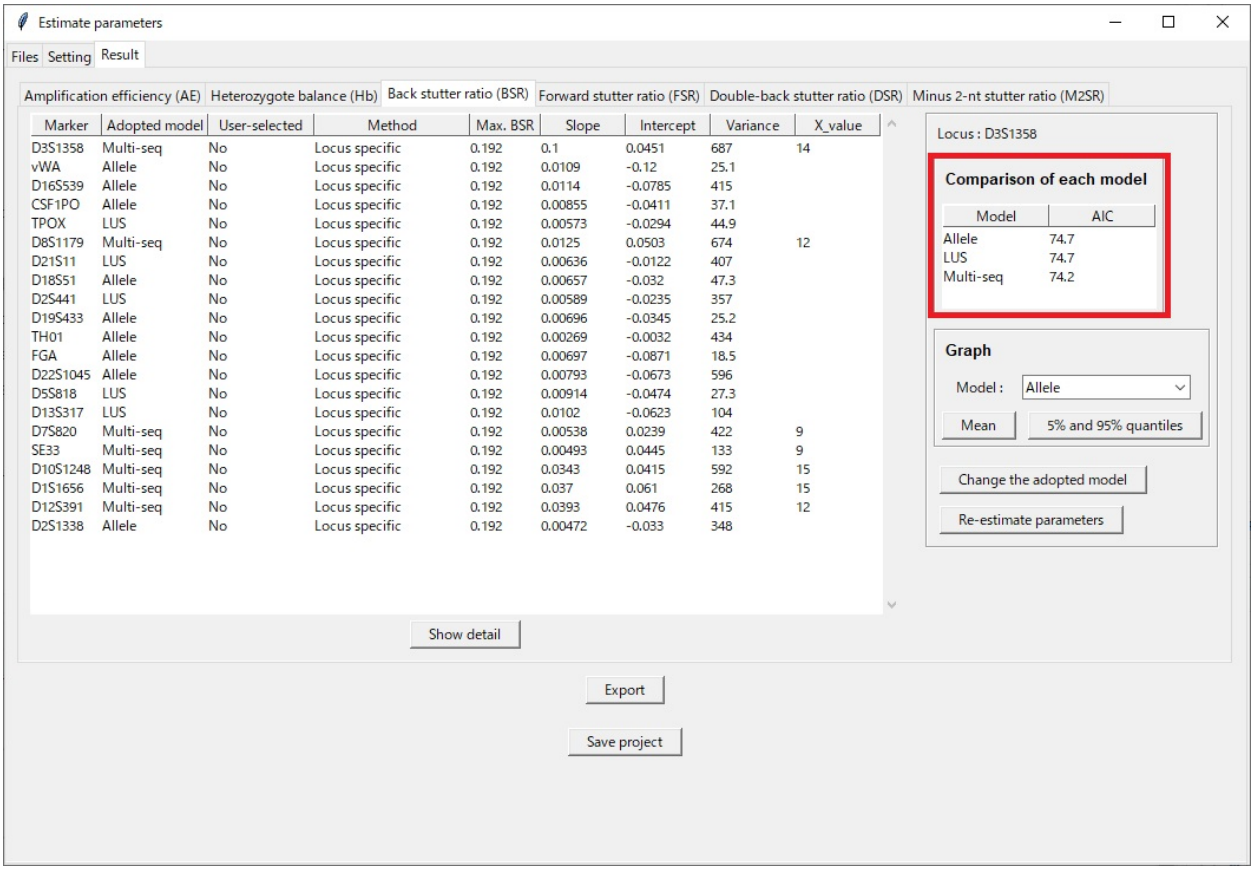


Marker	Adopted model	User-selected	Method	Max. BSR	Slope	Intercept	Variance	X_value
D3S1358	Multi-seq	No	Locus specific	0.192	0.1	0.0451	687	14
vWA	Allele	No	Locus specific	0.192	0.0109	-0.12	25.1	
D16S539	Allele	No	Locus specific	0.192	0.0114	-0.0785	415	
CSF1PO	Allele	No	Locus specific	0.192	0.00855	-0.0411	37.1	
TPOX	LUS	No	Locus specific	0.192	0.00573	-0.0294	44.9	
D8S1179	Multi-seq	No	Locus specific	0.192	0.0125	0.0503	674	12
D21S11	LUS	No	Locus specific	0.192	0.00636	-0.0122	407	
D18S51	Allele	No	Locus specific	0.192	0.00657	-0.032	47.3	
D2S441	LUS	No	Locus specific	0.192	0.00589	-0.0235	357	
D19S433	Allele	No	Locus specific	0.192	0.00696	-0.0345	25.2	
TH01	Allele	No	Locus specific	0.192	0.00269	-0.0032	434	
FGA	Allele	No	Locus specific	0.192	0.00697	-0.0871	18.5	
D22S1045	Allele	No	Locus specific	0.192	0.00793	-0.0673	596	
D5S818	LUS	No	Locus specific	0.192	0.00914	-0.0474	27.3	
D13S317	LUS	No	Locus specific	0.192	0.0102	-0.0623	104	
D7S820	Multi-seq	No	Locus specific	0.192	0.00538	0.0239	422	9
SE33	Multi-seq	No	Locus specific	0.192	0.00493	0.0445	133	9
D10S1248	Multi-seq	No	Locus specific	0.192	0.0343	0.0415	592	15
D1S1656	Multi-seq	No	Locus specific	0.192	0.037	0.061	268	15
D12S391	Multi-seq	No	Locus specific	0.192	0.0393	0.0476	415	12
D2S1338	Allele	No	Locus specific	0.192	0.00472	-0.033	348	

- Detailed results of one locus can be displayed by selecting a locus and clicking the ”Show detail” button.

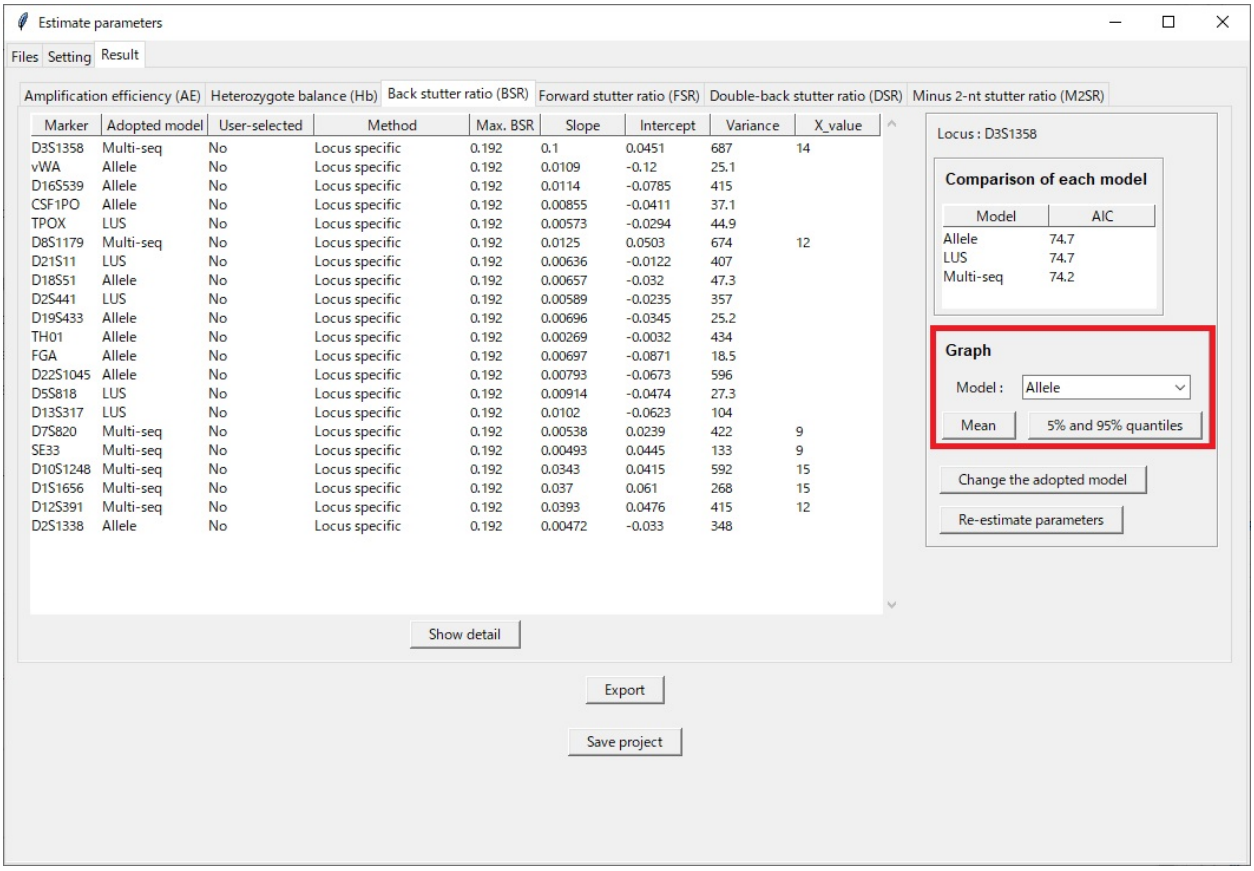


- AIC (Akaike information criterion) values[10] of each model are displayed in the frame of "Comparison of each model".



[10] A model featuring a low AIC is considered a superior fit for the experimental SR data as compared with a model featuring a high AIC.

- Users can verify whether the estimated parameters are fitted to the experimental data. The graph for ”5% and 95% quantiles” can be described in terms of the AE and Hb. The graphs for ”Mean” and ”5% and 95% quantiles” can be described in terms of the BSR, FSR, DSR, and M2SR. For example, the graphs for ”Mean” and ”5% and 95% quantiles” of BSR can be described as shown in Figures 6 and 7, respectively.



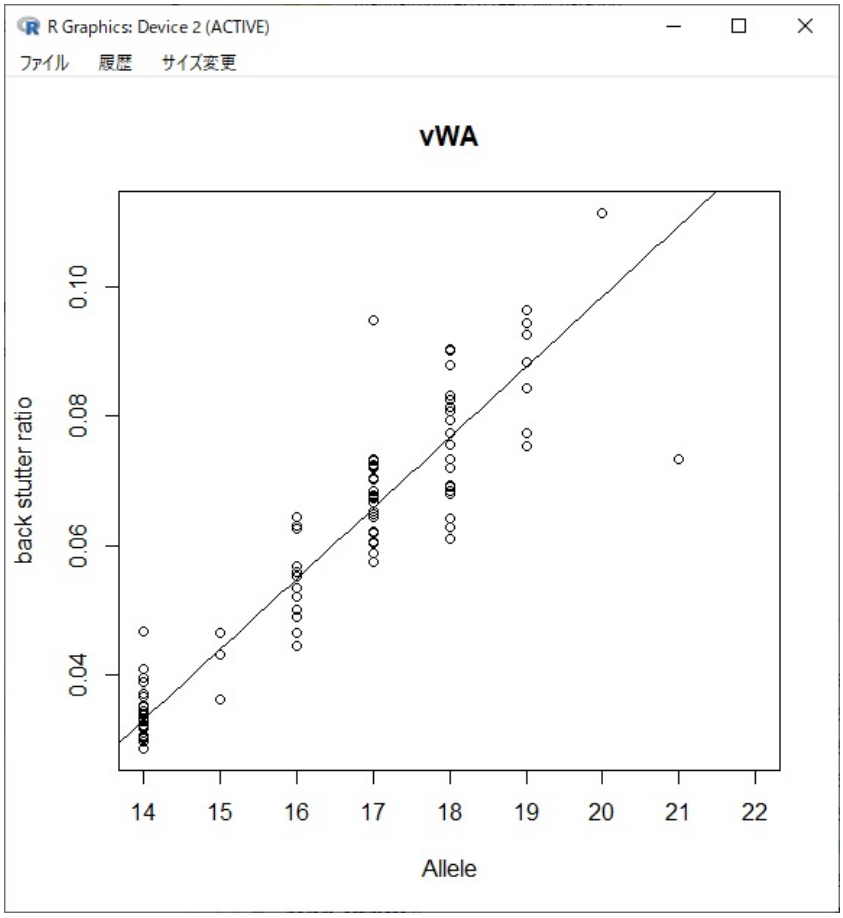


Figure 7: An example of the graph for ”Mean” of back stutter ratios. The solid line indicates a regression line of mean values.

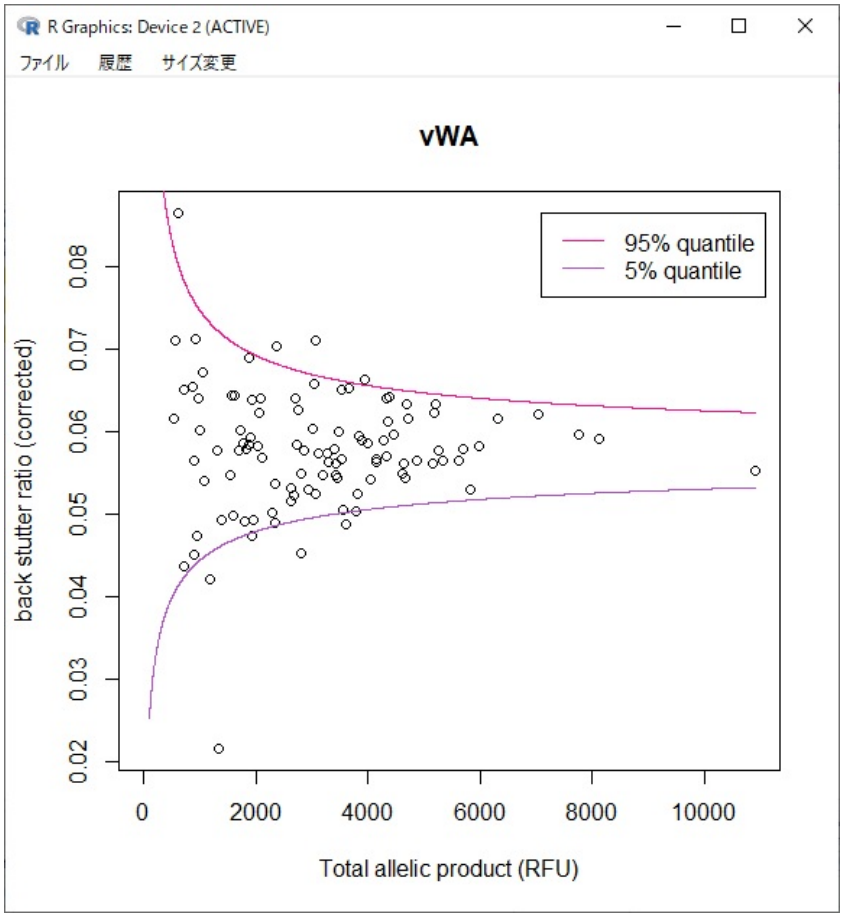
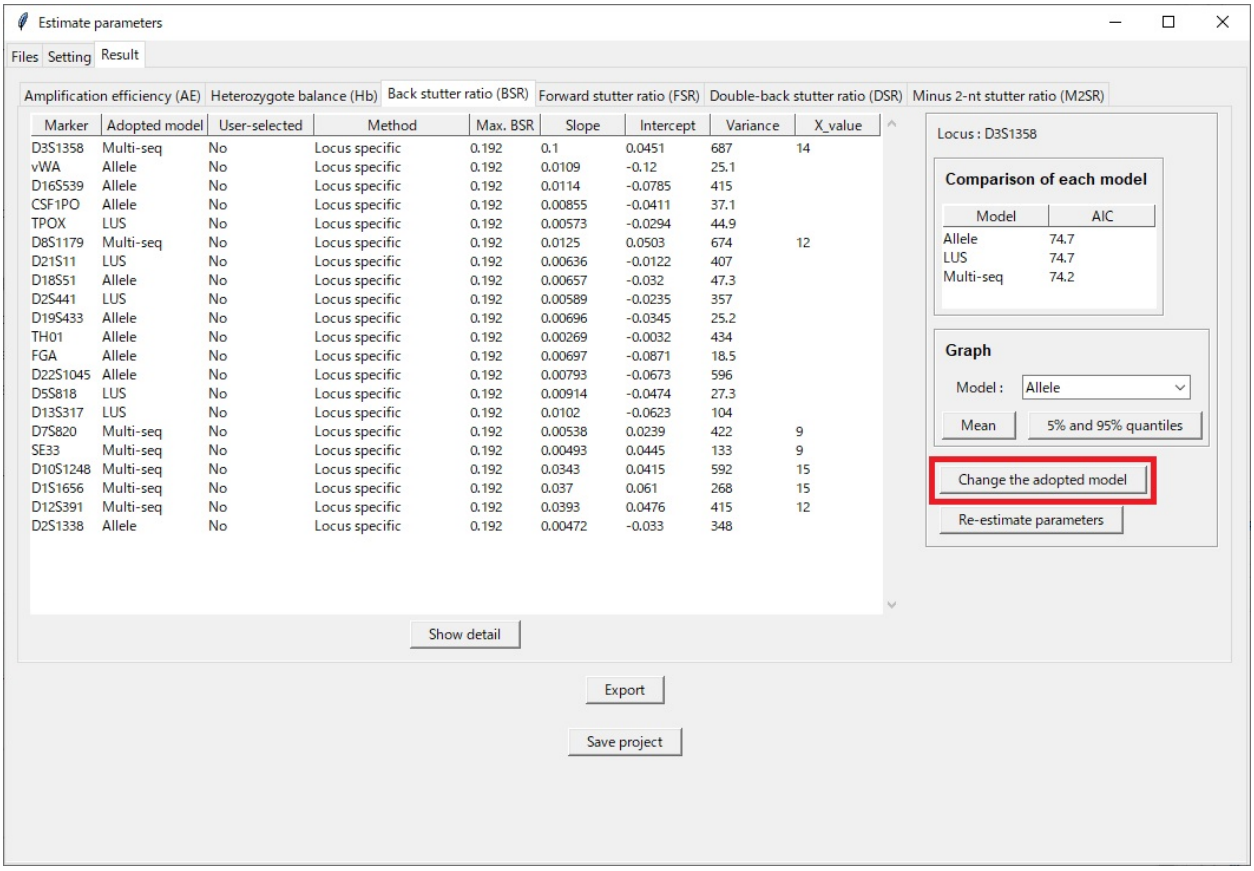
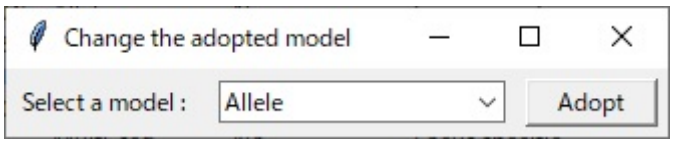


Figure 8: An example of the graph for "5% and 95% quantiles" of back stutter ratios. The relationship between back stutter ratios and total allelic products is shown. The upper and lower solid lines indicate 95% and 5% quantiles of back stutter ratios, respectively. Back stutter ratios are corrected by eliminating the difference in mean values in each allele.

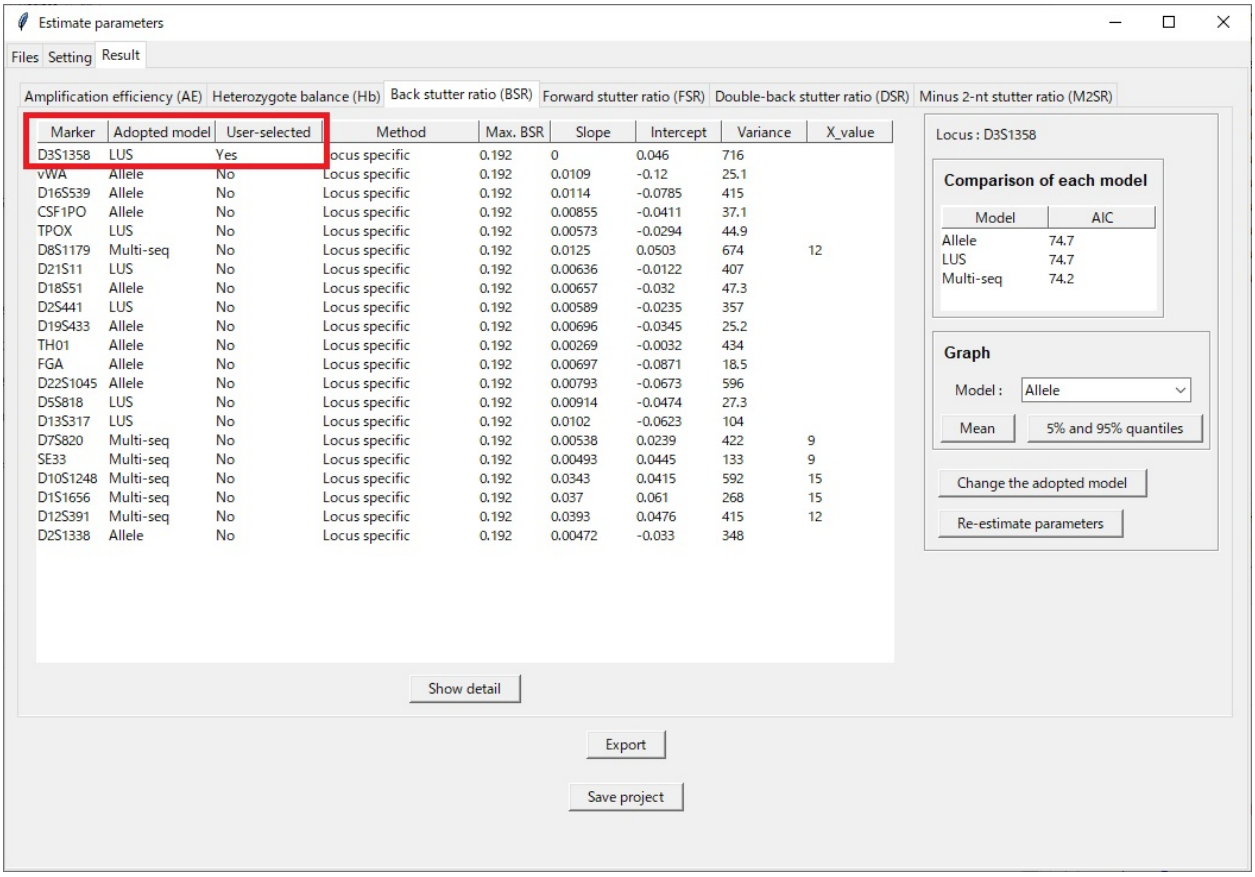
- Users can change the adopted model for BSR, FSR, and DSR by clicking the ”Change the adopted model” button.



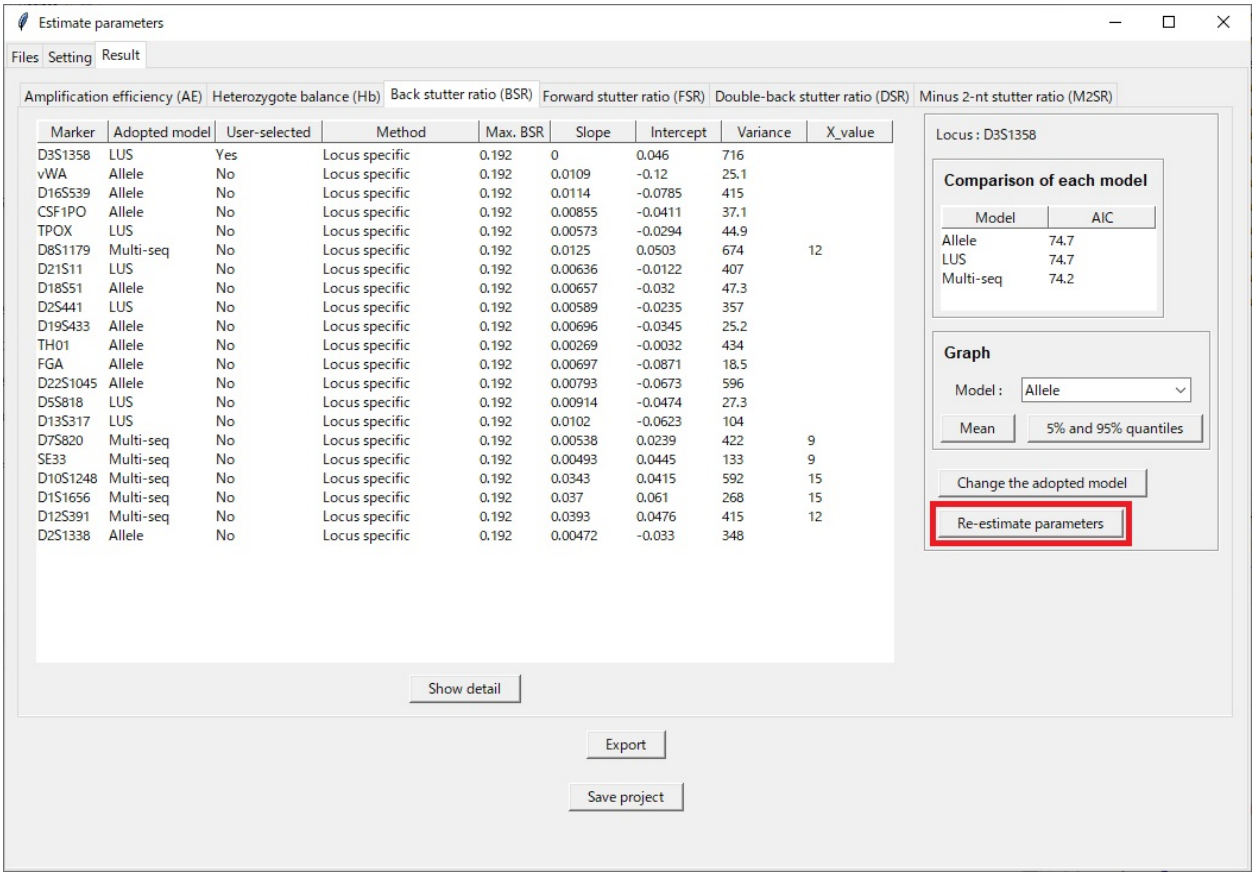
Then, a new window will appear. Select the desired model, and click the “Adopt” button.



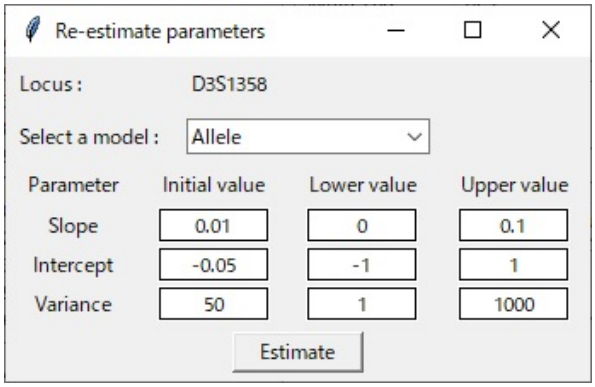
The user-selected model will be reflected and the "User-selected" column of the target locus is changed to "Yes".



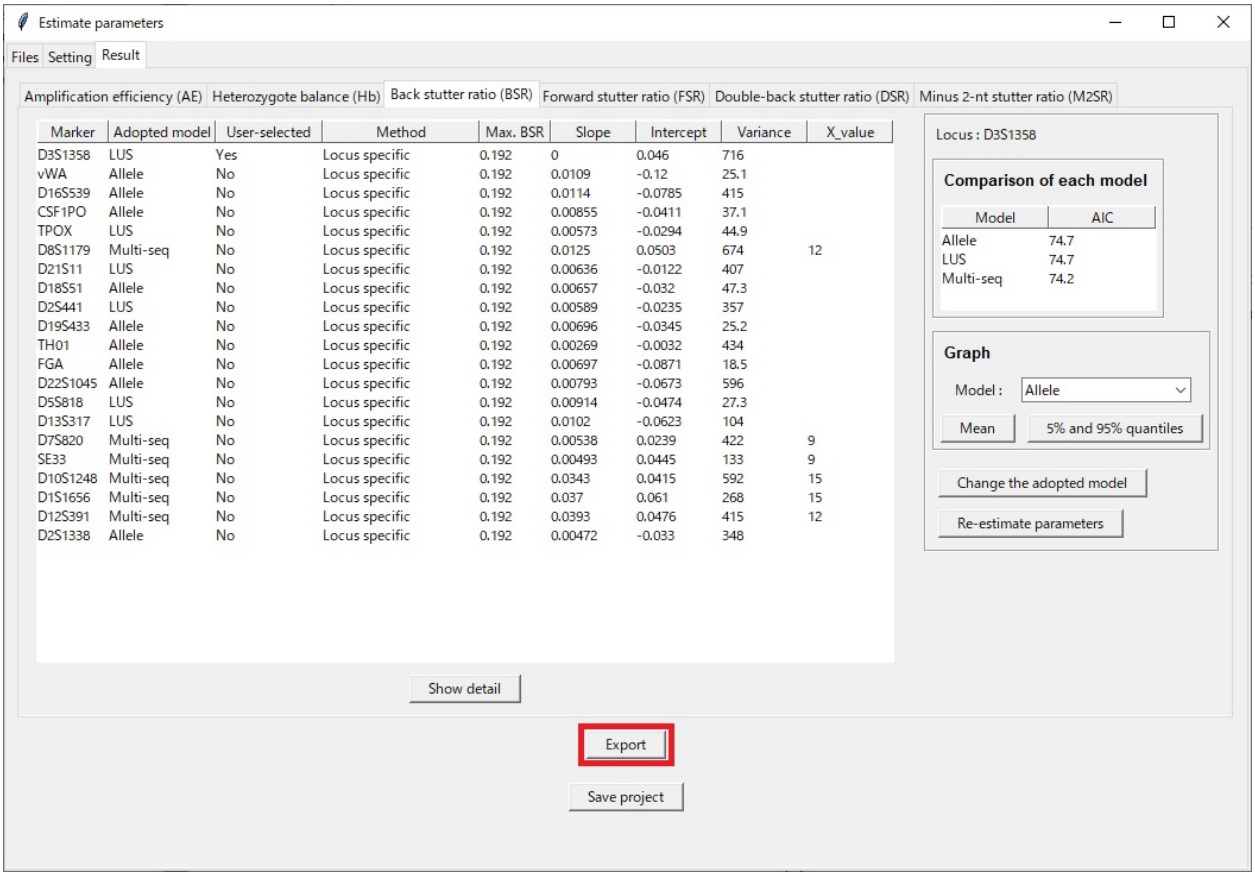
- Users can re-estimate the parameters by clicking the ”Re-estimate parameters” button.



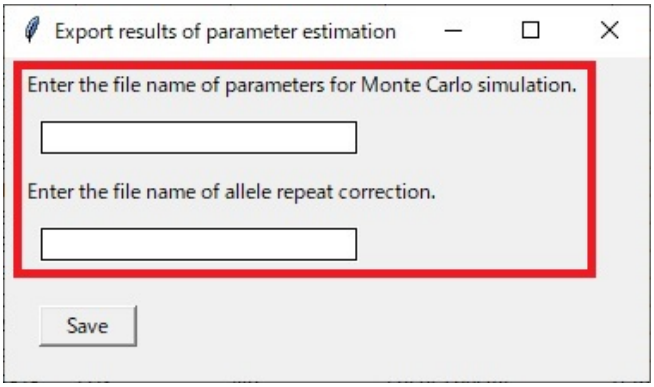
Then, a new window will appear. Select a model and set initial, lower, and upper values of each parameter in the maximum likelihood estimation (MLE). After that, click the ”Estimate” button to start the re-estimation.



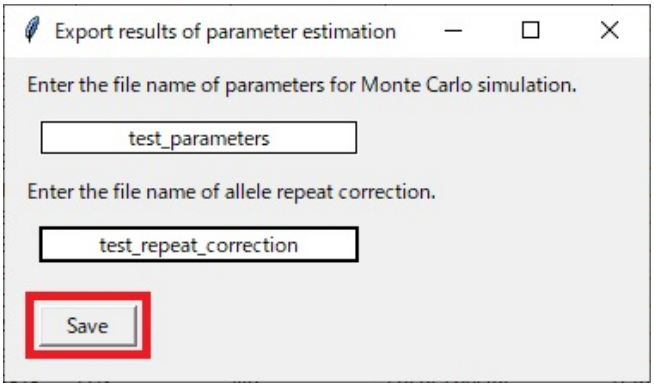
2. After reviewing all the results of estimated parameters, click the ”Export” button.



3. Enter names for ”Parameters for Monte Carlo simulation” and ”Allele repeat correction”.



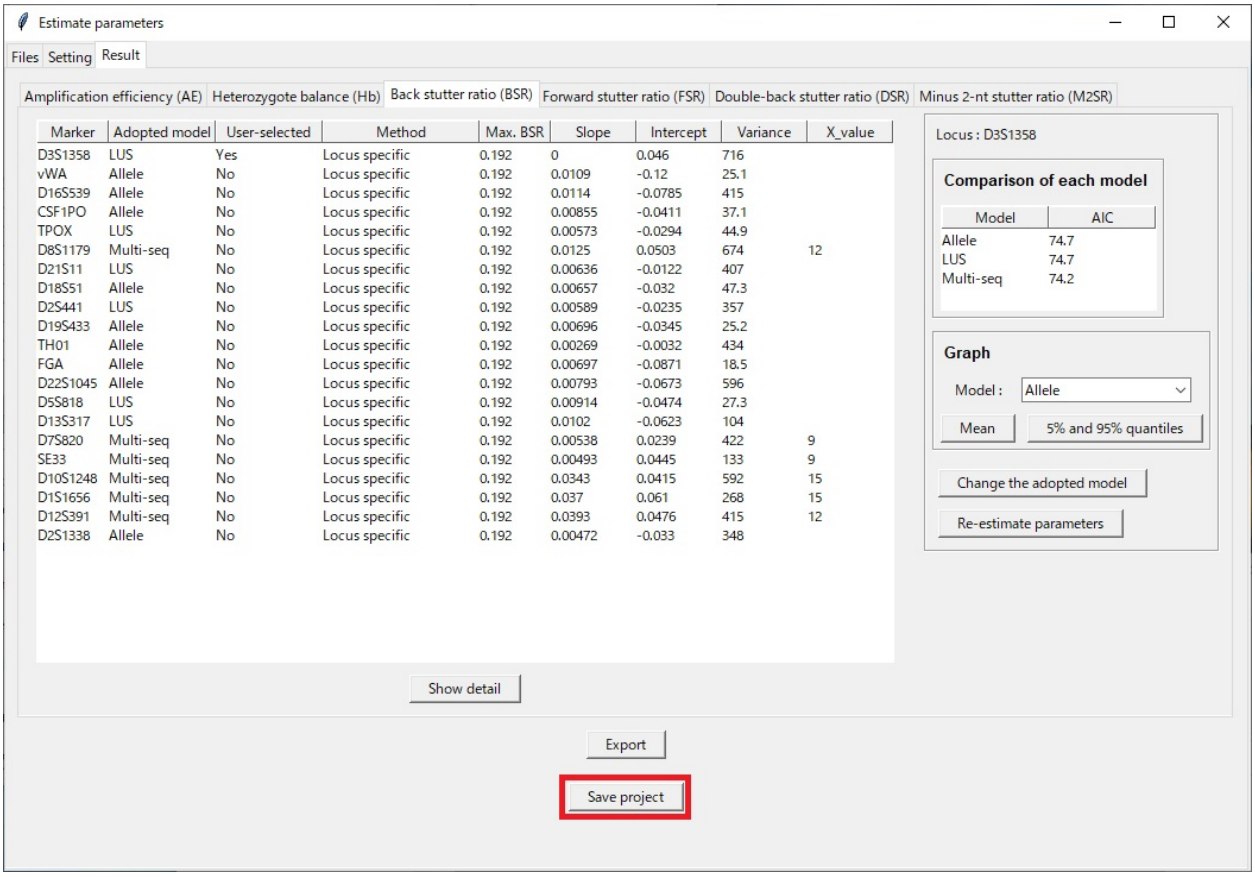
4. Click the ”Save” button.



Note

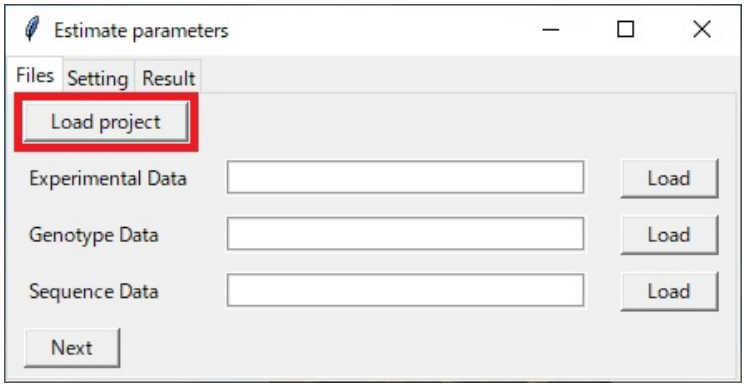
Two exported files can be directly used for setting analysis methods of ”Parameters for Monte Carlo simulation” and ”Allele repeat correction”.

5. Click the ”Save project” button to save the current project.



Note

Users can load the saved project by clicking the ”Load project” button in the ”Files” tab.

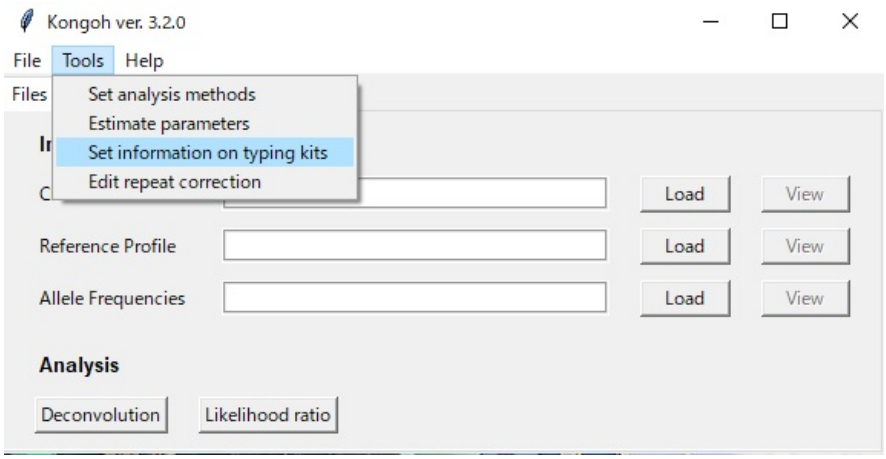


Set information on typing kits

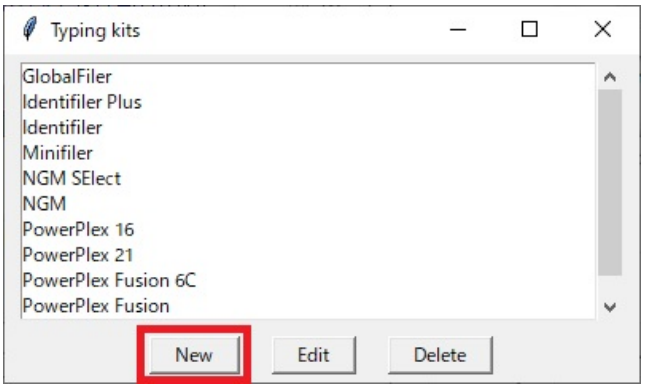
Users can create or edit information on typing kits. The information is essential to perform deconvolution and the assignment of likelihood ratios.

Create information on a typing kit

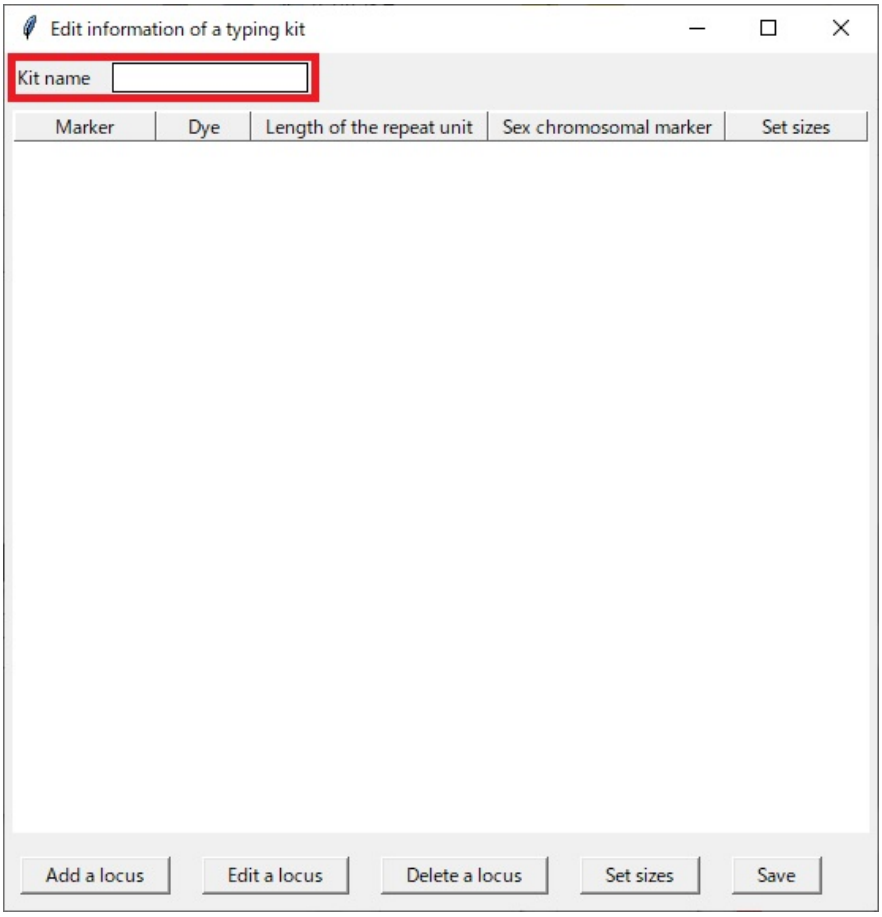
1. Go to Tools > Set information on typing kits.



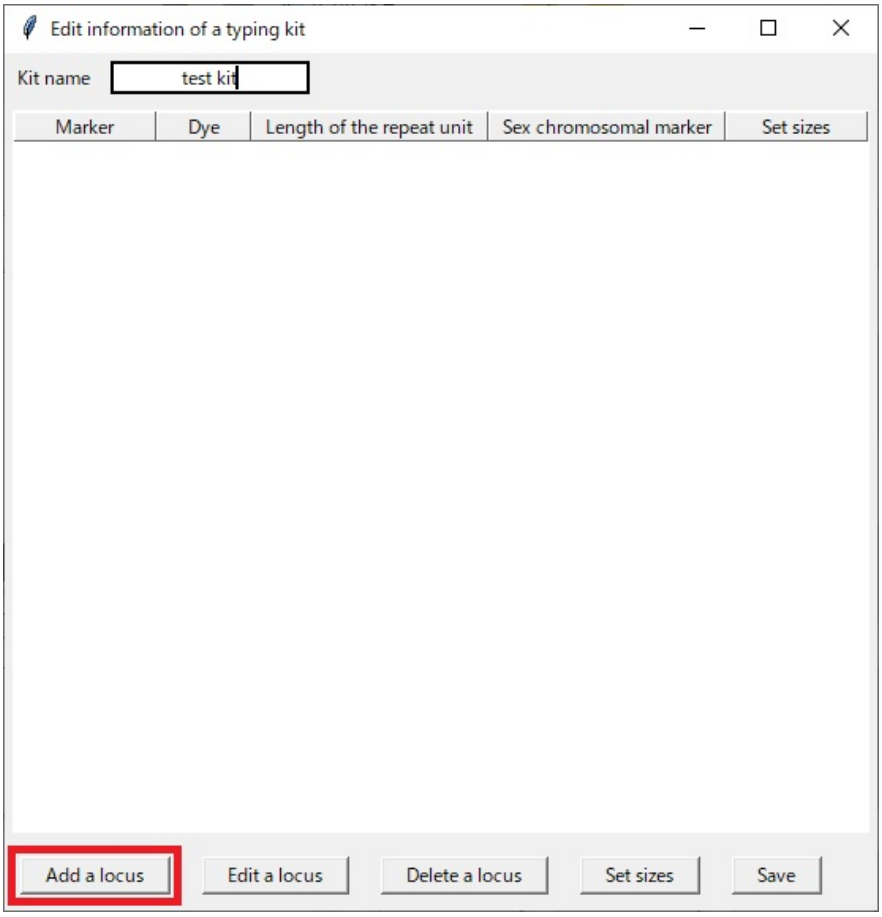
2. Click the "New" button.



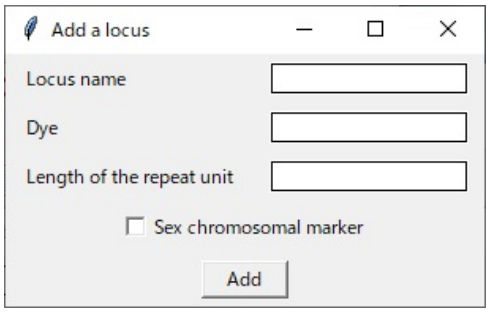
3. Enter a kit name.



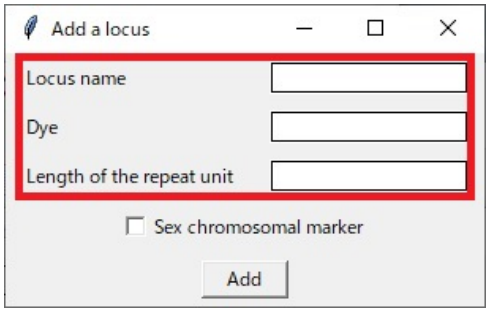
4. Click the ”Add a locus” button.



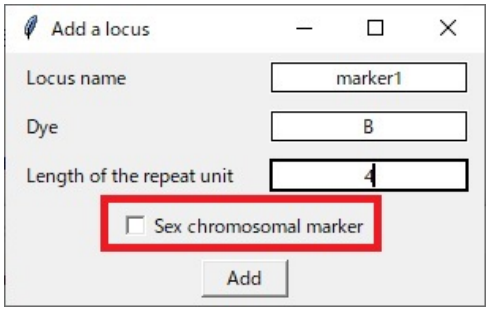
Then, a new window will be automatically opened.



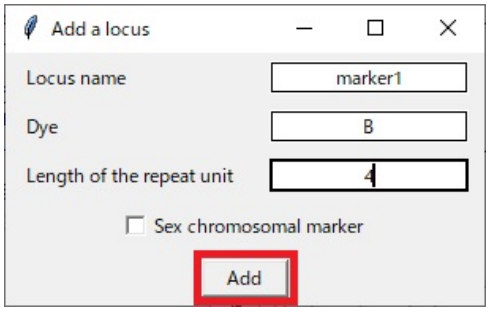
5. Enter a locus name, the dye information, and the length of the repeat unit[11].



6. If the locus is a sex chromosomal marker, Check the "Sex chromosomal marker".

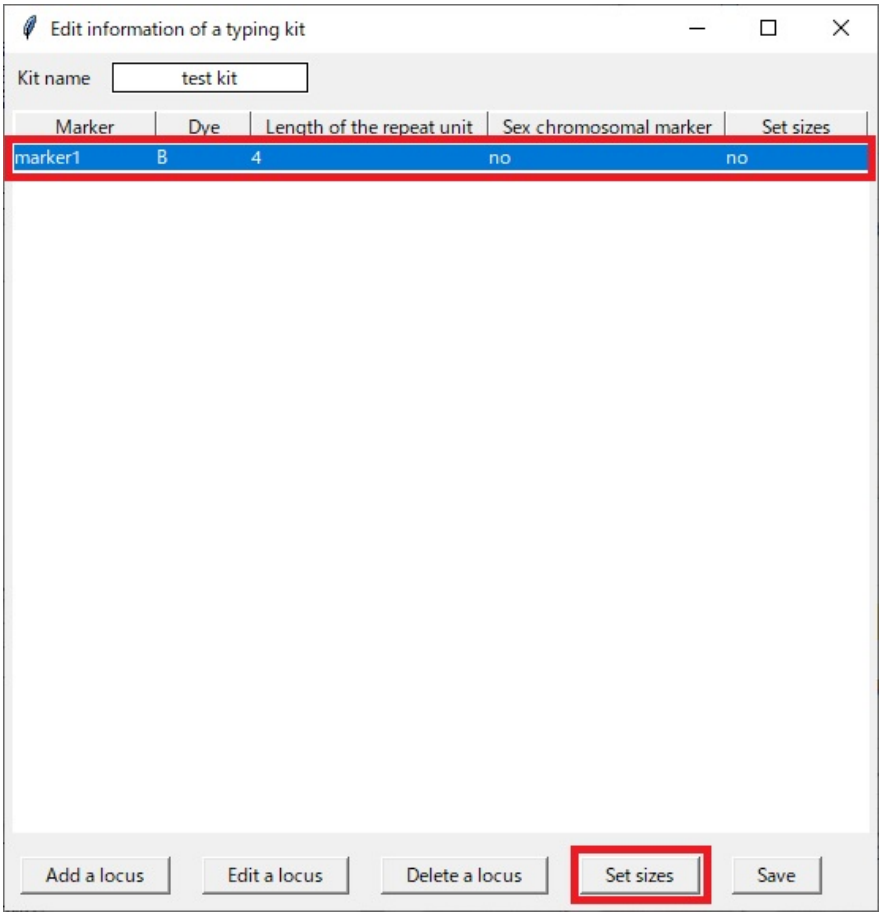


7. Click the "Add" button. Then, the information on the locus will be displayed.



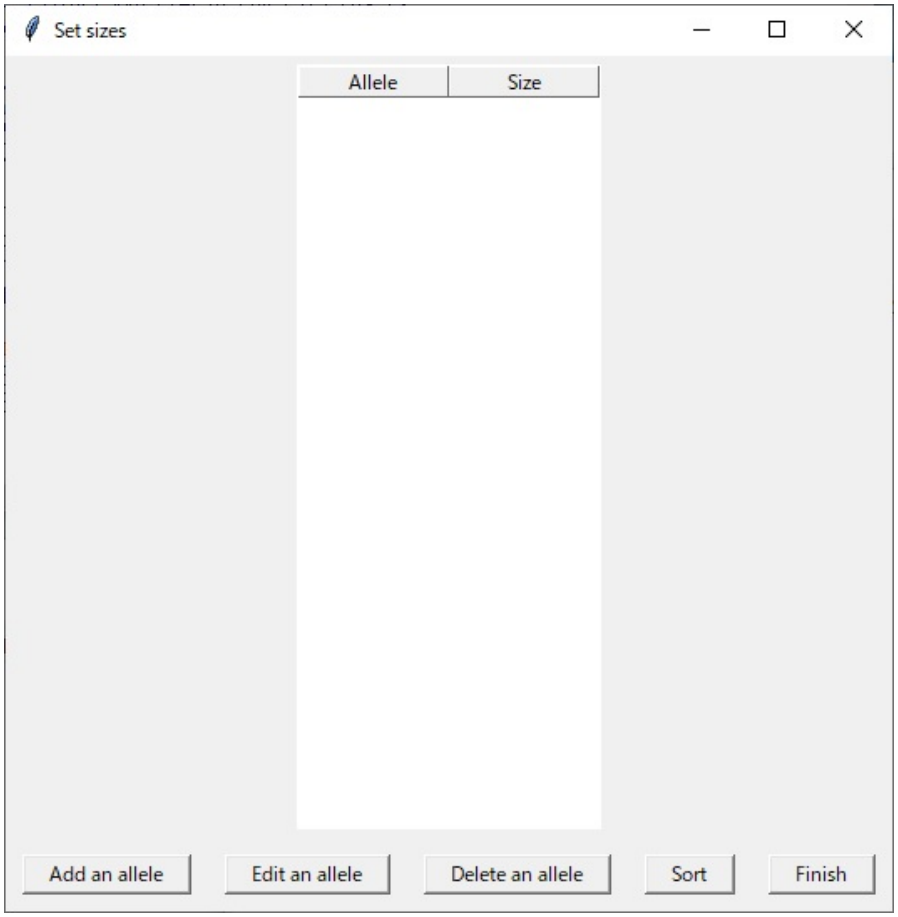
[11] Kongoh is only applicable to the STR typing kit; therefore, the length of the repeat unit must be set in all loci other than sex chromosomal markers.

8. Select a locus and click the ”Set sizes” button to set the sizes of each allele[12].

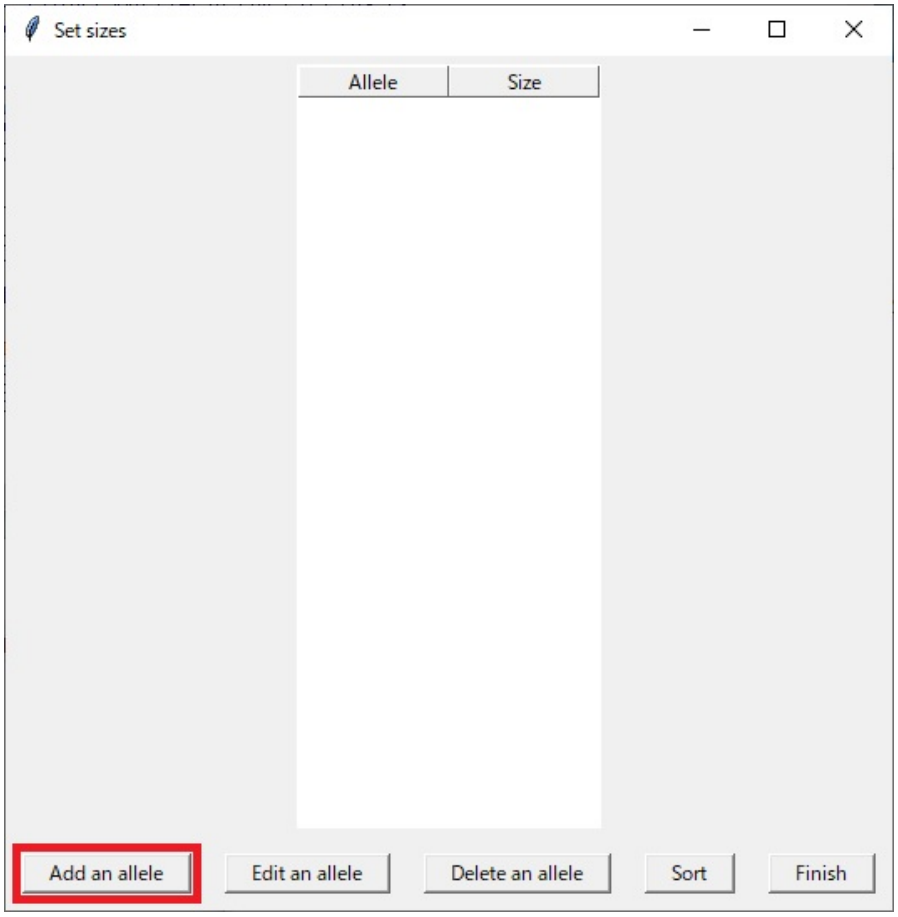


[12]The information on the sizes is used to estimate the size of a drop-out allele in the case of locus drop-out when performing deconvolution and assigning likelihood ratios.

Then, a new window will be automatically opened.

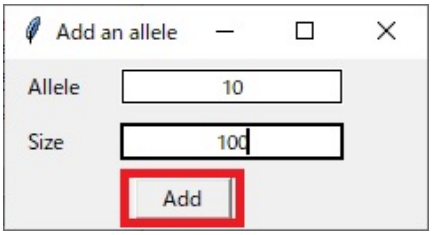


9. Click the ”Add an allele” button.

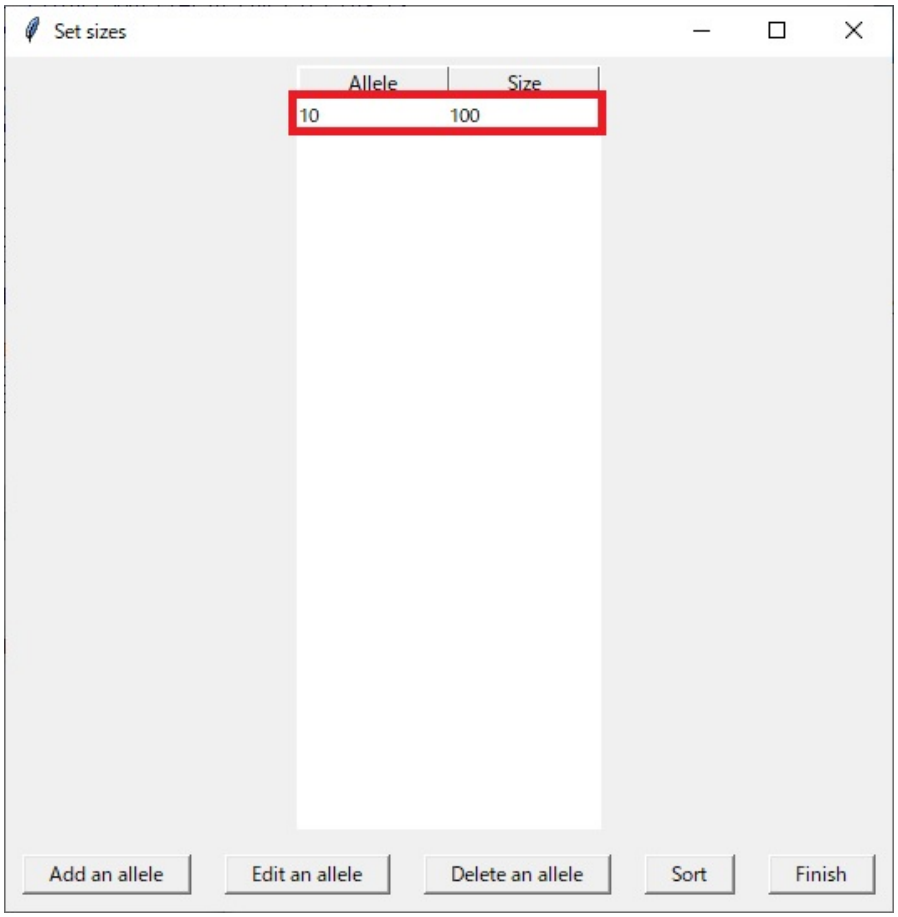


10. Enter an allele name, and the size.

11. Click the ”Add” button.



Then, the information on the allele and the size will be displayed.

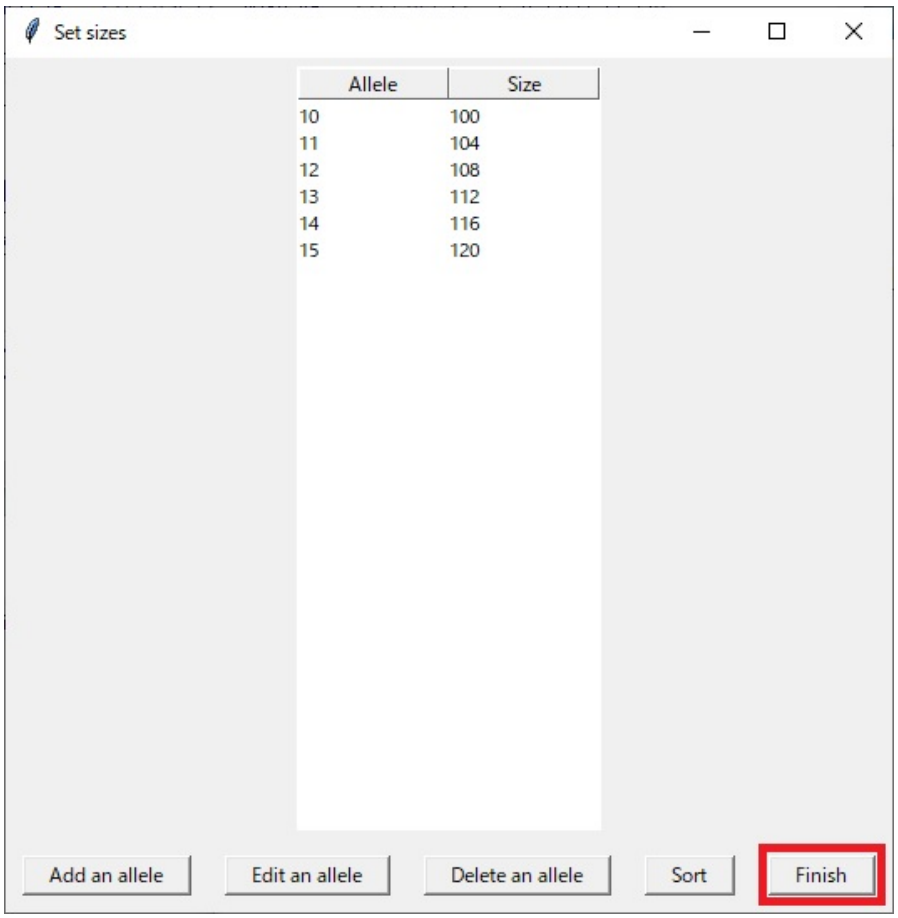


12. Repeat the addition of alleles and the sizes as many as possible.

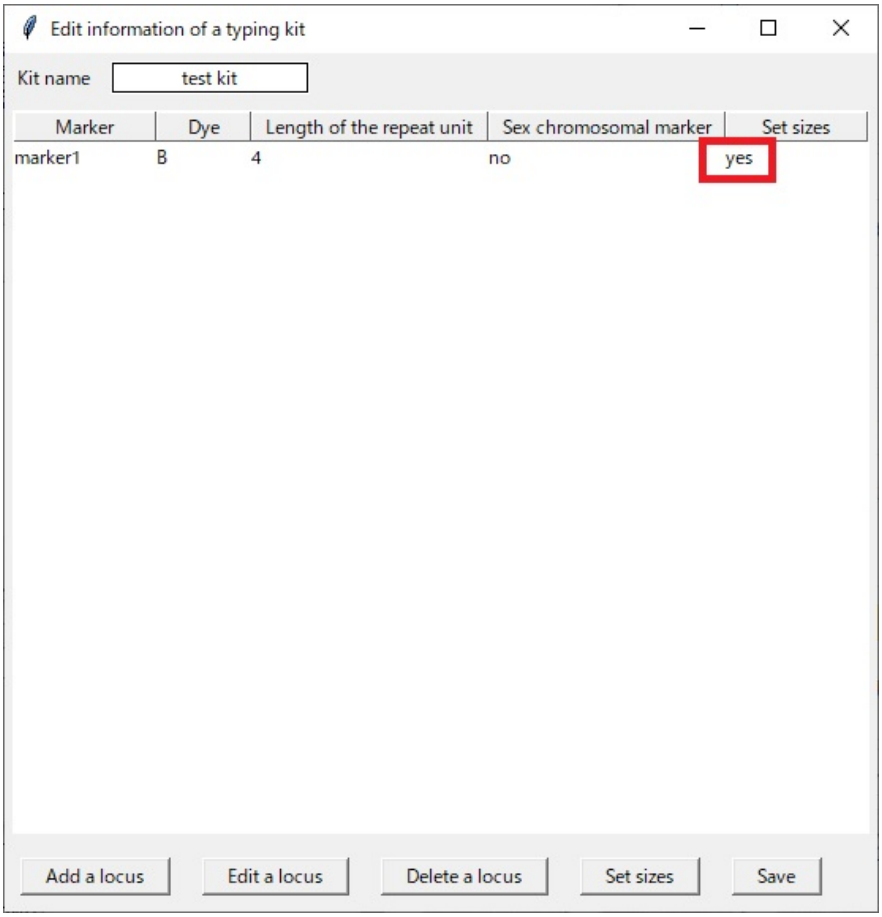
Note

Users can edit an allele and the size from the ”Edit an allele” button. Users can also delete an allele and the size from the ”Delete an allele” button. Alleles can be sorted from the ”Sort” button.

13. Click the ”Finish” button.

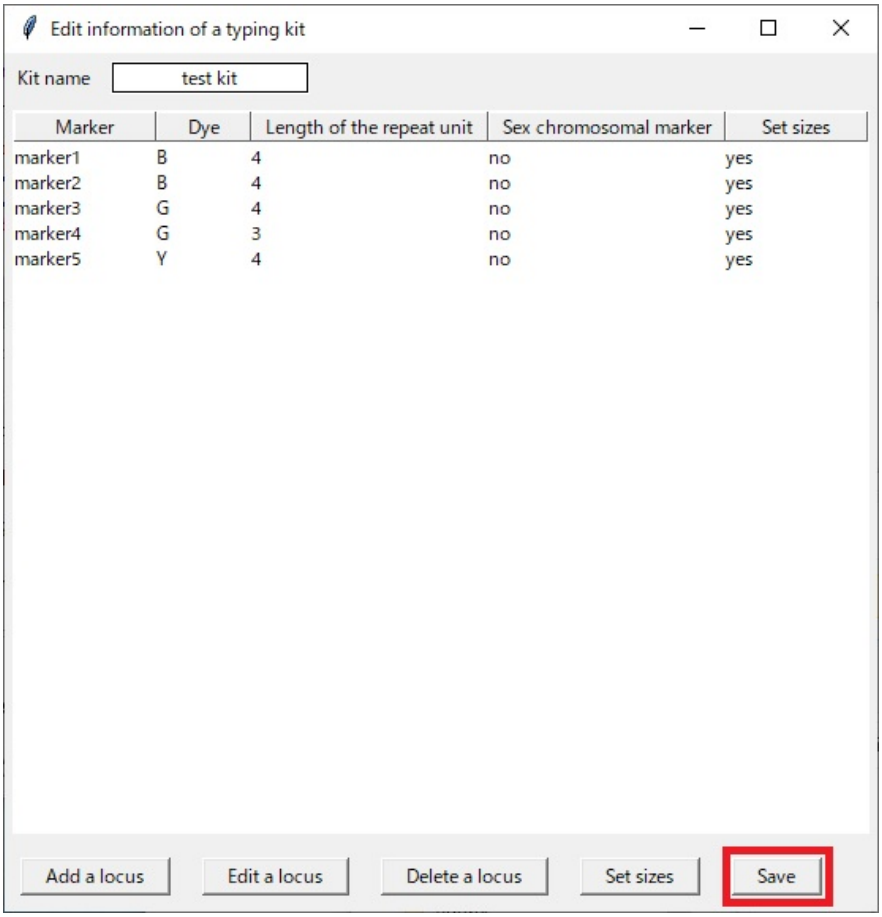


Then, the ”Set sizes” column will be changed to ”yes”[13].



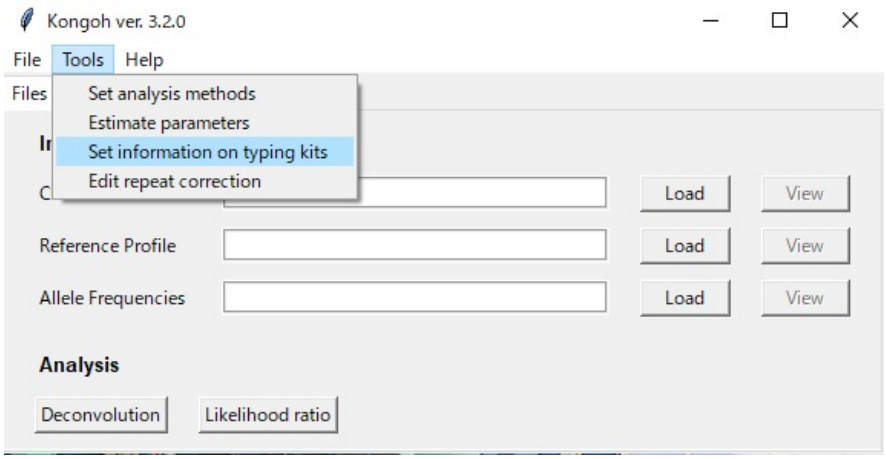
[13]Users must set sizes of all loci.

14. Click the ”Save” button after setting all markers.

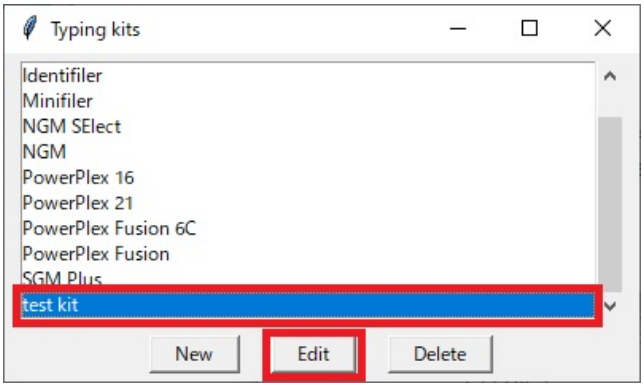


Edit information on a typing kit

1. Go to Tools > Set information on typing kits.

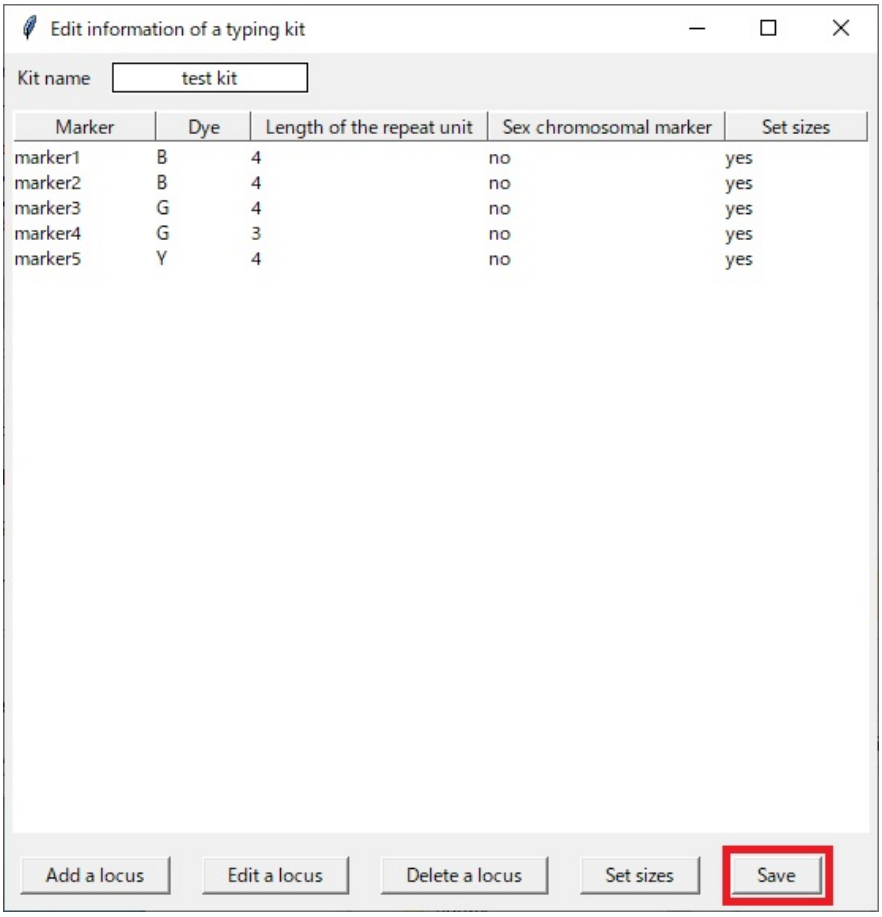


2. Select a kit and click the "Edit" button.



3. Edit the information on the selected kit just like creating information on a typing kit.

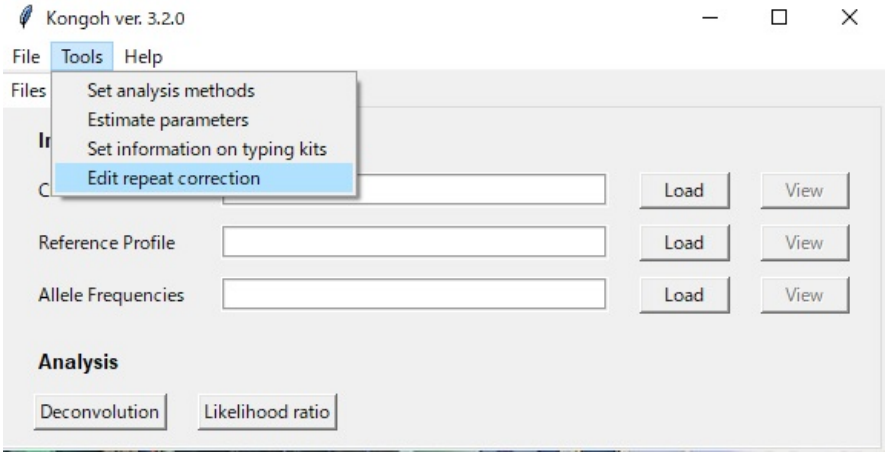
4. Click the ”Save” button.



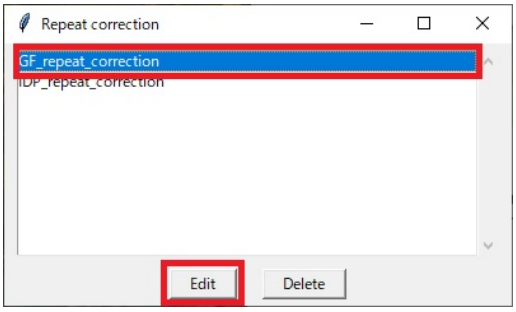
Edit repeat correction

Users can edit LUS, CA_BSR, CA_FSR, and CA_DSR. LUS is the abbreviation of the longest uninterrupted stretch[11], which is used in the LUS model for stutter ratios. CA_BSR, CA_FSR, and CA_DSR are corrected allele numbers used in the multi-sequence model for back stutter ratios, forward stutter ratios, and double-back stutter ratios, respectively.

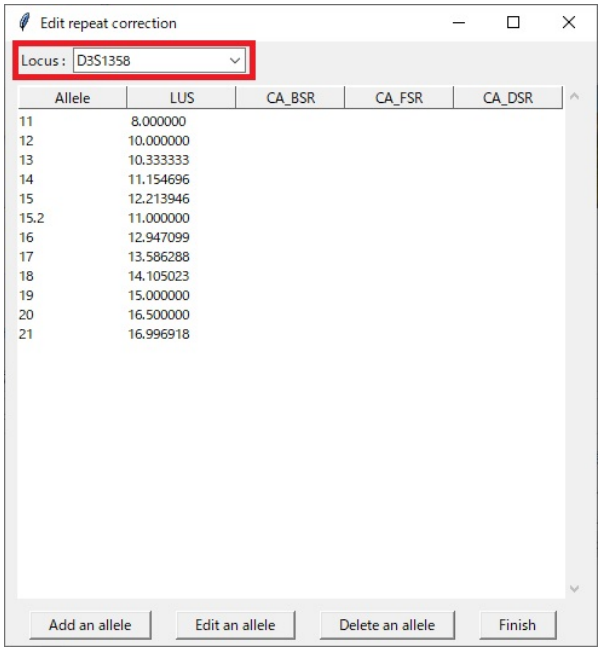
1. Go to Tools > Edit repeat correction.



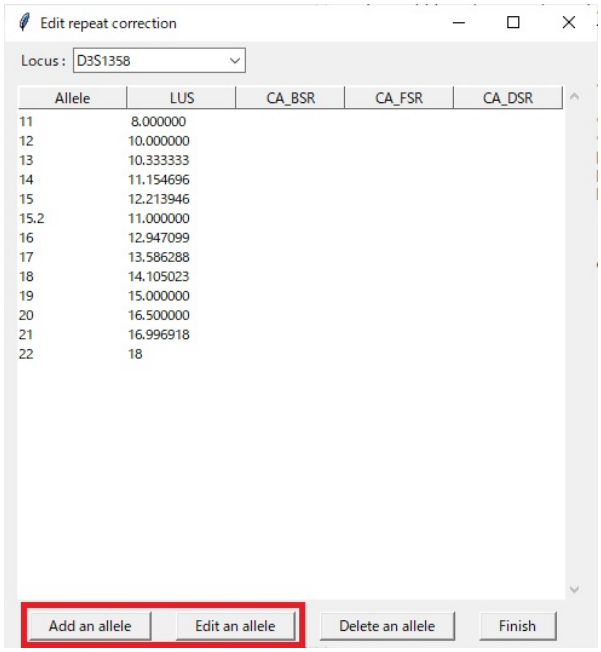
2. Select a dataset of allele repeat correction and click the "Edit" button.



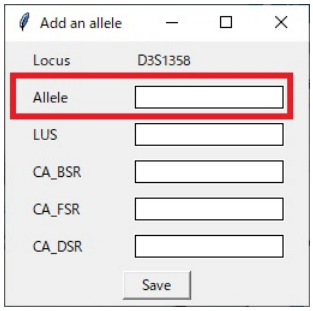
3. Select a locus.



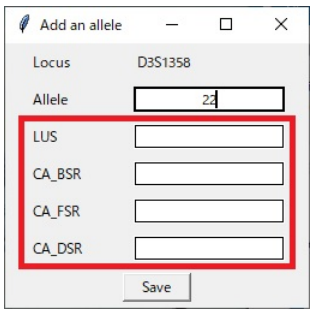
4. Add or edit LUS, CA_BSR, CA_FSR, or CA_DSR from the "Add an allele" button or the "Edit an allele" buttons, respectively.



5. Enter the name of an allele.



6. Enter values of LUS, CA_BSR, CA_FSR, or CA_DSR as needed.



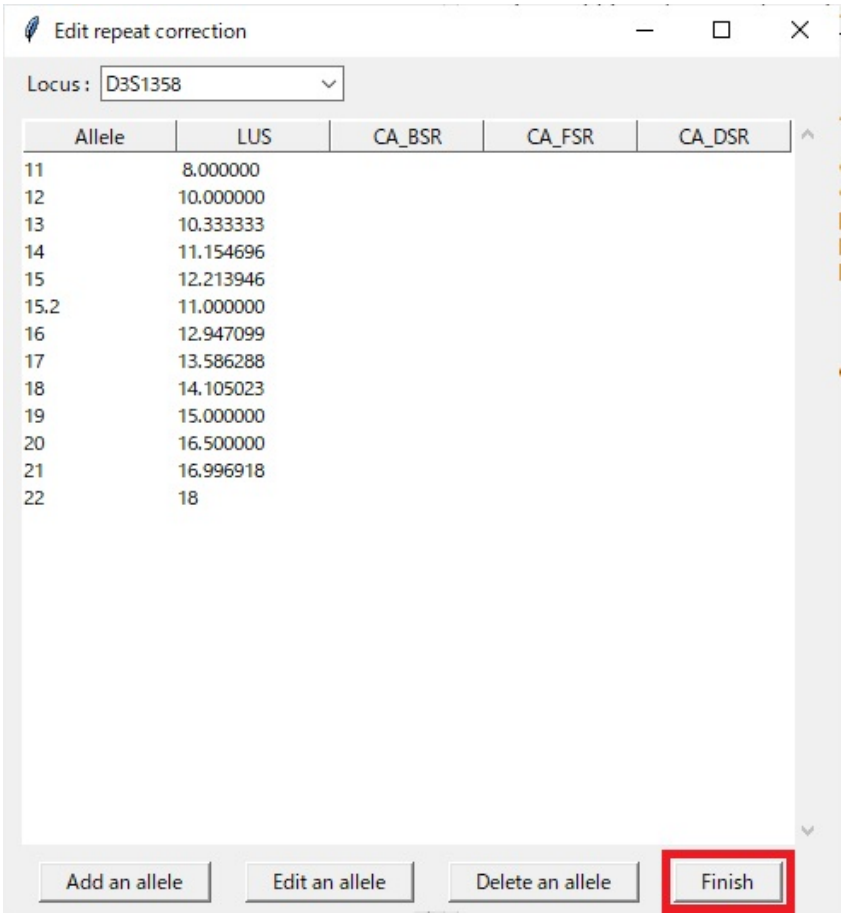
7. Click the ”Save” button.

8. Repeat the edit of LUS, CA_BSR, CA_FSR, and CA_DSR.

Note

Users can delete an allele from the ”Delete an allele” button.

9. Click the ”Finish” button after finishing all the settings.



References

[1] S. Manabe, C. Morimoto, Y. Hamano, S. Fujimoto, K. Tamaki, Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model, PLoS One 12 (11) (2017) e0188183.

[2] S. Manabe, T. Fukagawa, K. Fujii, N. Mizuno, K. Sekiguchi, A. Akane, K. Tamaki, Development and validation of Kongoh ver. 3.0.1: Open-source software for DNA mixture interpretation in the GlobalFiler system based on a quantitative continuous model, Leg Med (Tokyo) 54 (2022) 101972.

[3] K. Yoshida, K. Takahashi, K. Kasai, Allele frequencies of 15 loci using AmpFlSTR Identifiler Kit in Japanese population, J Forensic Sci 50 (3) (2005) 718–719.

[4] K. Fujii, H. Watahiki, Y. Mita, Y. Iwashima, T. Kitayama, H. Nakahara, N. Mizuno, K. Sekiguchi, Allele frequencies for 21 autosomal short tandem repeat loci obtained using GlobalFiler in a sample of 1501 individuals from the Japanese population, Leg Med (Tokyo) 17 (5) (2015) 306–308.

[5] S. Manabe, K. Fujii, T. Fukagawa, N. Mizuno, K. Sekiguchi, K. Inoue, M. Hashiyada, A. Akane, K. Tamaki, Evaluation of probability distribution models for stutter ratios in the typing system of GlobalFiler and 3500xL Genetic Analyzer, Leg Med (Tokyo) 52 (2021) 101906.

[6] K. Fujii, T. Fukagawa, H. Watahiki, Y. Mita, T. Kitayama, N. Mizuno, Ratios and distances of pull-up peaks observed in GlobalFiler kit data, Leg Med (Tokyo) 34 (2018) 58–63.

[7] L. Borsuk, K. Gettings, C. Steffen, K. Kiesler, P. Vallone, Sequence-based US population data for the SE33 locus, Electrophoresis 39 (2018) 2694–2701.

[8] K. Gettings, L. Borsuk, C. Steffen, K. Kiesler, P. Vallone, Sequence-based U.S. population data for 27 autosomal STR loci, Forensic Sci Int Genet 37 (2018) 106–115.

[9] P. Walsh, N. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic Acids Res 24 (1996) 2807–2812.

[10] C. Brookes, J. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Sci Int Genet 6 (2012) 58–63.

[11] J. Bright, D. Taylor, J. Curran, J. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci Int Genet 7 (2013) 296–304.

[12] D. Taylor, J. Bright, C. McGoven, C. Hefford, T. Kalafut, J. Buckleton, Validating multiplexes for use in conjunction with modern interpretation strategies, Forensic Sci Int Genet 20 (2016) 6–19.