Kongoh version 3.0.1 User Manual

6th July 2021

Sho Manabe

Department of Legal Medicine, Kansai Medical University, Hirakata, Japan

manabesh@hirakata.kmu.ac.jp

Contents

1. What is *Kongoh*?

Kongoh (named after the Japanese word “mixture”) is an open-source software for DNA evidence interpretation based on a quantitative continuous model [1]. The software is a graphical user interface written in R language, and the source code is freely available at GitHub (https://github.com/manabe0322/Kongoh/releases).

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) can be analyzed using *Kongoh*. 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.

Kongoh can calculate the likelihood of one to four contributors. The likelihood ratios are calculated by the ratio of the maximum likelihood in prosecution and defense hypotheses.

2. Changes in ver. 3.0.1

- The crime stain profile typed by GlobalFiler kit can be analyzed.
- The “estimation of the parameter” function was added. The function can estimate parameters based on experimental data prepared by users.
- Crime stain profiles typed by kits other than Identifiler Plus and GlobalFiler can be analyzed if experimental data are prepared from the kit[1].
- Analytical thresholds can be set based on the dye color.
- Allele frequencies can be estimated using the expected values based on the Dirichlet distribution for each locus.
- Candidate mixture ratios per contributor and degradation parameters can be determined arbitrarily.
- Forward stutter ratios, double-back stutter ratios, and minus 2-nt stutter ratios can be considered.
- Uniform, longest uninterrupted stretch (LUS), and multi-seq models for stutter ratios were implemented.
- The upper and lower limits were added to the log-normal distributions for locus-specific amplification efficiency and heterozygote balance.
- The upper limits were added to the log-normal distributions for each type of stutter ratio.

[1] In some typing kits (i.e., other than Identifiler Plus, GlobalFiler, NGM, NGMSElect, Minifiler, SGM Plus, PowerPlex Fusion 6C, PowerPlex Fusion, PowerPlex 21, and PowerPlex 16), crime stain profiles with locus drop-out cannot be analyzed in *Kongoh* because the sizes of drop-out alleles cannot be determined.

- The method for excluding unrealistic genotype combinations was improved by adding a stochastic threshold. If one of the allele peaks in a heterozygous genotype is above the threshold and the other allele peak is below the analytical threshold, then the genotype is regarded as unrealistic.
- The "heterozygote balance filter" is applicable when both of the two heterozygote allele peaks are above the stochastic threshold.
- The "forward stutter filter", "double-back stutter filter", and "minus 2-nt stutter filter" were added to exclude unrealistic genotype combinations.
- The maximum stutter peak height was removed from the thresholds to exclude unrealistic genotype combinations.
- The likelihood of all assumed hypotheses for explaining a crime stain profile, including the number of contributors and the combination of reference individuals, are automatically calculated in the first calculation[2]. When recalculating the likelihood ratio by changing the prosecutor and defense hypotheses in the same crime stain profile, the calculation results are displayed immediately.
- The results of probabilistic genotyping can be verified using the software.

[2] For example, when reference profiles of a victim (V) and a suspect (S) are inputted in *Kongoh* and 1–3 contributors are assumed, likelihoods of the following hypotheses are automatically calculated: V, S, an unknown contributor (U), V+S, V+U, S+U, U+U, V+S+U, V+U+U, S+U+U, and U+U+U.

3. Tutorial

3.1. Getting started

1. Ensure that R software is installed. It is available from the R Development Core Team website (http://www.R-project.org).
2. Download the source code of *Kongoh*, which is freely available at GitHub (https://github.com/manabe0322/Kongoh/releases).
3. Decompress the downloaded file.
4. Begin an R session.
5. Load the "Kongoh v3.0.1.RData" file from "Load Workspace" in the "File" tab (Fig. 1).

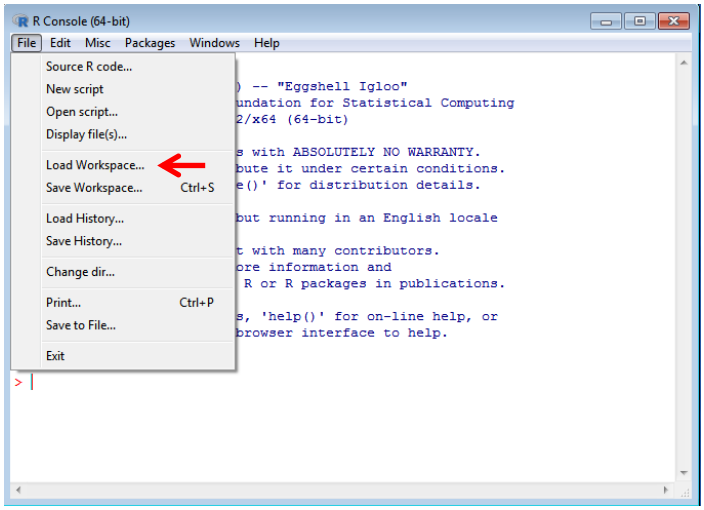

Fig. 1. "Load Workspace" in "File" tab.

6. Execute the following command in R:

```
Kongoh()
```

7. If all required packages used in *Kongoh* (tcltk, tcltk2, gtools, parallel, truncnorm, and GenSA) have not been installed, then these packages will be automatically installed.

8. After all packages are loaded, the "Files" tab opens automatically (Fig. 2).

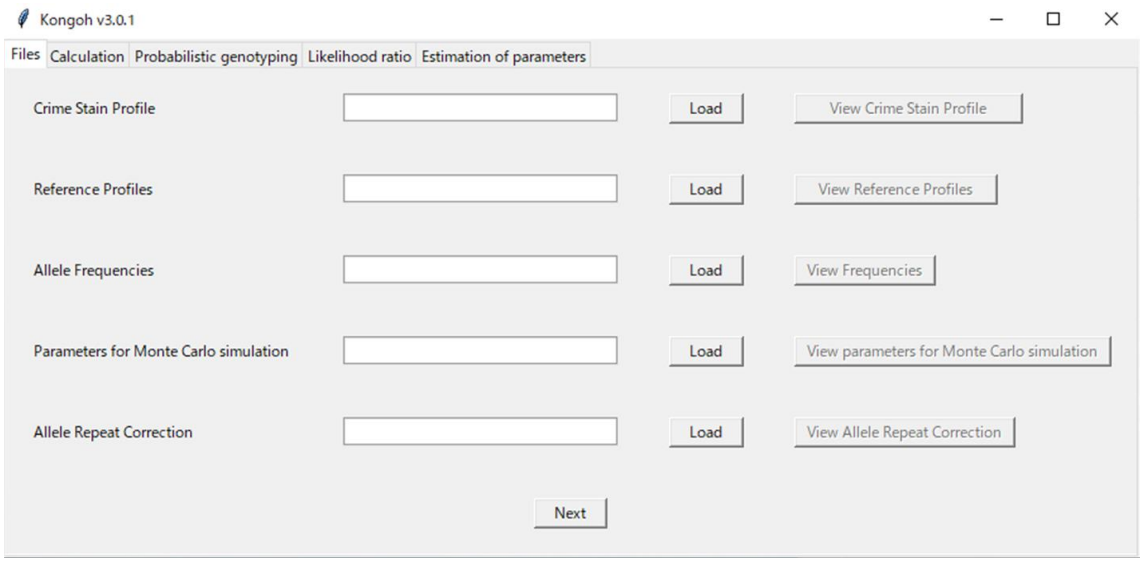

Fig. 2. "Files" tab, where user can import a crime stain profile, reference profiles, allele frequencies, parameters for Monte Carlo simulation, and information of allele repeat correction for models of each stutter ratio.

3.2. Input required files

1. Prepare required files (.csv format) based on the "**Notes**" below.
2. Load five required files by clicking the "Load" button for each file.
3. Click the "Next" button.

Notes:

Crime stain profile

- This file can be exported from the GeneMapper® ID-X software.
- This file must include information regarding the "Sample File", "Marker", "Allele", "Size", and "Height" as shown in Fig. 3.
- Back stutters, forward stutters, double-back stutters, and minus 2-nt stutters are not necessarily removed manually in the software program using sample electrophoretic data (e.g., GeneMapper® ID-X software)[3].
- Pull-up peaks and noises must be removed manually.

[3] Whether these stutters should be removed depends on the setting of the "Monte Carlo parameter" file.

Sample File	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	

Fig. 3. Format of crime stain profile.

Reference profiles

- This file must include information regarding the “Marker” and the name of each profile (e.g., victim and suspect), as shown in Fig. 4.
- 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

Fig. 4. Format of reference profiles.

Allele frequencies

- Two types of allele frequency exist: the allele count (Fig. 5) and allele probability (Fig. 6).
- If the allele frequencies are represented as allele counts, then the probabilities of each allele are determined by the expected values based on the Dirichlet distribution.
- If the allele frequencies are represented as probabilities, then these values are regarded as 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 (set in the “Calculation” tab) is used as the frequency of these alleles.
- Allele frequencies for Identifiler [2] and GlobalFiler [3] in the Japanese population can be downloaded from https://github.com/manabe0322/Kongoh/releases. The allele frequencies for the Identifiler are represented as probabilities. The allele frequencies for GlobalFiler are represented as allele counts.

Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA	D22S1045	D5S818	D13S317	D7S8
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	

Fig. 5. Format of allele frequencies, which are represented as allele counts.

Allele	D3S1358	vWA	D16S539	CSF1PO	TPOX	D8S1179	D21S11	D18S51	D2S441	D19S433	TH01	FGA	D22S1045	D5S818	D13S317	D7S8
5											0.001332					
6											0.220853					
7			0.000333	0.012325		0.000333				0.000333	0.270153			0.002665	0.001999	0.00
8			0.001666	0.000999	0.452032				0.002332		0.064624			0.005996	0.261825	0.13
8.1									0.000333							
9			0.358761	0.0503	0.117921	0.002665					0.396402			0.091606	0.13058	0.0
9.1									0.036642							0.00
9.2										0.000333						
9.3											0.037642					
10			0.200533	0.215856	0.033311	0.128914		0.001999	0.268155		0.008994			0.205197	0.112592	0.22
10.1									0.002332							0.00
10.2										0.000999						
10.3																0.00
11			0.187209	0.206529	0.357095	0.106262		0.004664	0.344437	0.003331			0.198201	0.288141	0.22485	0.3
11.1									0.000333							
11.2										0.000666						
11.3									0.022652							
12	0.001999		0.172219	0.422052	0.037975	0.123251		0.046636	0.17022	0.043971			0.001332	0.230513	0.202532	0.22
12.2										0.00433						
13	0.000999	0.000333	0.06962	0.069287	0.000999	0.225849		0.199201	0.048301	0.291139			0.001666	0.165556	0.052632	0.03
13.2										0.029647						
14	0.026316	0.195203	0.008661	0.017322	0.000666	0.208195		0.214191	0.099267	0.345769			0.002332	0.009327	0.012325	0.00
14.1																
14.2										0.083278						
15	0.397069	0.025983	0.000999	0.004664		0.136576		0.164224	0.004664	0.053298			0.311126	0.000999	0.000666	0.00

Fig. 6. Format of allele frequencies, which are represented as probabilities.

Parameters for Monte Carlo simulation

- The file format is shown in Fig. 7. The default files for Identifiler Plus (Parameter IDP.csv) and GlobalFiler (Parameter GF.csv) can be downloaded from https://github.com/manabe0322/Kongoh/releases.
- The file can be generated in the “Estimation of parameters” function. This function is explained in Section 4.

Factor	Marker	Repeat_length	Dye	Model	Locus_specific	Parameter1	Parameter2	Parameter3	Parameter4
AE	D3S1358	4	B	Log-normal	TRUE	0.11127412	37.53563063		
AE	vWA	4	B	Log-normal	TRUE	-0.043596988	28.14775483		
AE	D16S539	4	B	Log-normal	TRUE	-0.120129511	32.20360062		
AE	CSF1PO	4	B	Log-normal	TRUE	0.008902089	29.37678268		
AE	TPOX	4	B	Log-normal	TRUE	-0.376987711	28.42532211		
AE	D8S1179	4	G	Log-normal	TRUE	0.338328689	51.30888082		
AE	D21S11	4	G	Log-normal	TRUE	0.132878459	33.52475685		
AE	D18S51	4	G	Log-normal	TRUE	0.303146811	42.73801429		
AE	D2S441	4	Y	Log-normal	TRUE	-0.108962362	38.03856623		
AE	D19S433	4	Y	Log-normal	TRUE	-0.332618365	37.79491707		
AE	TH01	4	Y	Log-normal	TRUE	-0.69558261	53.51556212		
AE	FGA	4	Y	Log-normal	TRUE	-0.118678063	41.66196204		
AE	D22S1045	3	R	Log-normal	TRUE	-0.058400747	30.78565658		
AE	D5S818	4	R	Log-normal	TRUE	0.06070042	33.78133055		
AE	D13S317	4	R	Log-normal	TRUE	0.238492544	40.03948389		
AE	D7S820	4	R	Log-normal	TRUE	0.088359214	34.53825926		
AE	SE33	4	R	Log-normal	TRUE	0.20023827	39.53462544		
AE	D10S1248	4	P	Log-normal	TRUE	0.206132037	53.09754323		
AE	D1S1656	4	P	Log-normal	TRUE	0.076581236	38.88584337		
AE	D12S391	4	P	Log-normal	TRUE	-0.03005453	37.32104767		
AE	D2S1338	4	P	Log-normal	TRUE	-0.28969211	61.78089295		
Hb	D3S1358	4	B	Log-normal	TRUE	0.011330727	198.1780281		
Hb	vWA	4	B	Log-normal	TRUE	-0.005534235	147.6003219		
Hb	D16S539	4	B	Log-normal	TRUE	-0.025246045	122.646206		
Hb	CSF1PO	4	B	Log-normal	TRUE	0.028408787	128.1985617		
Hb	TPOX	4	B	Log-normal	TRUE	-0.032100031	139.1653515		
Hb	D8S1179	4	G	Log-normal	TRUE	-0.033464546	203.3382532		
Hb	D21S11	4	G	Log-normal	TRUE	0.021985612	156.6108448		
Hb	D18S51	4	G	Log-normal	TRUE	0.003766641	181.9947107		

Fig. 7. Format of parameters for Monte Carlo simulation.

- The first column is the “Factor” column. The names of each factor are as follows:
  - AE: locus-specific amplification efficiency
  - Hb: heterozygote balance
  - BSR: back stutter ratio
  - FSR: forward stutter ratio
  - DSR: double-back stutter ratio
  - M2SR: minus 2-nt stutter ratio
- In the “Model” column, the model names are entered per locus in each factor. The candidate model names are listed in Table 1.

Table 1 Candidates of model names in each factor

Model name	AE	Hb	BSR	FSR	DSR	M2SR
Log-normal	✓	✓				
Allele			✓	✓	✓	
LUS			✓	✓	✓	
Multi-seq			✓	✓	✓	
Uniform				✓	✓	✓
No*			✓	✓	✓	✓

* Not considered

- In the “Locus-specific” column, “TRUE” is entered when considering the locus specifically. “FALSE” is entered when considering multiple loci simultaneously. In loci with “FALSE”, the settings of the “Model” column and the “Parameter” columns must be identical.

- In the “Parameter 1–4” columns, the parameters for each model are entered. The meanings of parameters 1–4 for each parameter are shown in Table 2.

Table 2 Meanings of parameters 1–4 for each parameter

Factor	Model	Parameter1	Parameter2	Parameter3	Parameter4
AE	Log-normal	Mean [a]	Variance [a]		
Hb	Log-normal	Mean [a]	Variance [a]		
BSR	Allele	Slope	Intercept	Variance [a]	
	LUS	Slope	Intercept	Variance [a]	
	Multi-seq	Slope	Intercept	Variance [a]	x_l [b]
FSR	Allele	Slope	Intercept	Variance [a]	
	LUS	Slope	Intercept	Variance [a]	
	Multi-seq	Slope	Intercept	Variance [a]	x_l [b]
	Uniform	Mean [a]	Variance [a]		
DSR	Allele	Slope	Intercept	Variance [a]	
	LUS	Slope	Intercept	Variance [a]	
	Multi-seq	Slope	Intercept	Variance [a]	x_l [b]
	Uniform	Mean [a]	Variance [a]		
M2SR	Uniform	Mean [a]	Variance [a]		

[a] Logarithmic scale

[b] Number of repeats affecting stuttering in locus l.

Allele repeat correction

- The allele repeat correction file is necessary if the LUS model or multi-seq model is used as models of stutter ratios.
- The file format is shown in Fig. 8. Default files for Identifiler Plus (Allele repeat correction IDP.csv) and GlobalFiler (Allele repeat correction GF.csv) can be downloaded from https://github.com/manabe0322/Kongoh/releases.
- The file can be generated in the “estimation of parameters” function based on the experimental data prepared by users. This function is explained in Section 4.
- In the “LUS” column, the LUS values of each allele are entered.
- In columns “CA_BSR”, “CA_FSR”, and “CA_DSR”, corrected allele numbers in each allele ware entered based on the multi-seq model for the back stutter, forward stutter, and double-back stutter ratios, respectively.

Marker	Allele	LUS	CA_BSR	CA_FSR	CA_DSR
⋮					
CSF1PO	6	5.833333			
CSF1PO	7	6.916667			
CSF1PO	8	8			
CSF1PO	9	9			
CSF1PO	10	9.997921			
CSF1PO	11	10.98589			
CSF1PO	12	11.9888			
CSF1PO	13	13			
CSF1PO	14	14			
CSF1PO	15	15			
TPOX	5	5			
TPOX	6	6			
TPOX	7	7			
TPOX	8	8			
TPOX	9	9			
TPOX	10	10			
TPOX	11	10.99802			
TPOX	12	12			
TPOX	13	13			
D8S1179	8	8	0		
D8S1179	9	9	0		
D8S1179	10	10	0		

Fig. 8. Format of allele repeat correction.

3.3. Calculation of likelihood ratio

1. Set both the prosecutor (H_p) and defense (H_d) hypotheses. Verify the individuals to include them as contributors in each hypothesis. Fig. 9 shows an example of setting the hypotheses:

   H_p : victim + suspect (+ unknown contributor(s))

   H_d : victim (+ unknown contributor(s)).

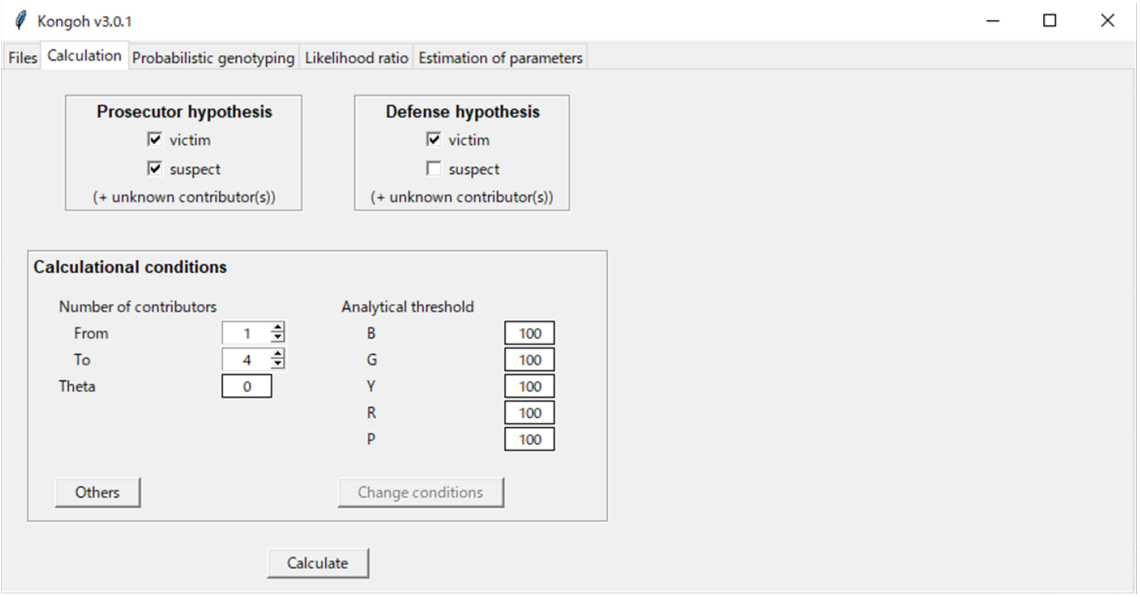

Fig. 9. "Calculation" tab where user can calculate likelihood ratios.

Note:

The number of unknown contributors need not be selected because *Kongoh* automatically calculates the likelihoods of all assumed numbers of contributors in both H_p and H_d.

2. Set the following calculational conditions:

   - Number of contributors

     Set the range of the assumed number of contributors. The likelihoods of all set numbers are calculated for both H_p and H_d.

   - Theta

     Set the theta correction to consider the subpopulation effect.

   - Analytical threshold

     Set the analytical thresholds for each dye. Peaks below the thresholds are not used in the calculation.

3. Other detailed conditions can be changed from the "Others" button. These conditions are explained in detail in Section 3.6.

4. Click the "Calculate" button, and the "Progress Bar" window will appear (Fig. 10).

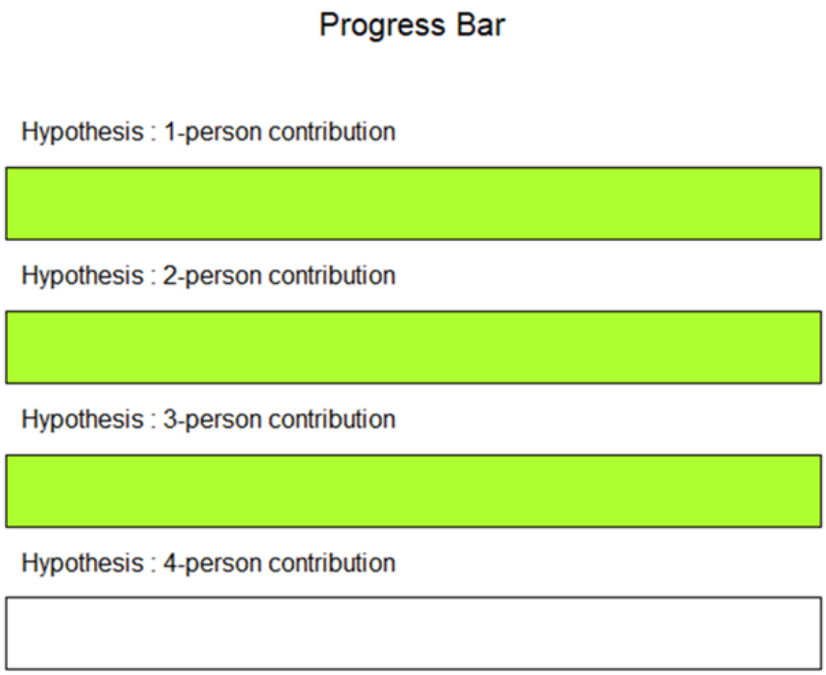

Fig. 10. "Progress Bar" window for calculation.

3.4. Results

1. After completing the likelihood calculation, the “Likelihood ratio” tab appears automatically, as shown in Fig. 11. A brief overview of the results is provided below.
   - Likelihood, estimated mixture ratios, and estimated degradation parameters of each assumed number of contributors in H_p and H_d
   - Likelihood ratios of each hypothesis
   - Ratio of maximum likelihood in H_p and H_d
2. This report can be exported to a .csv file by clicking the “Report” button.

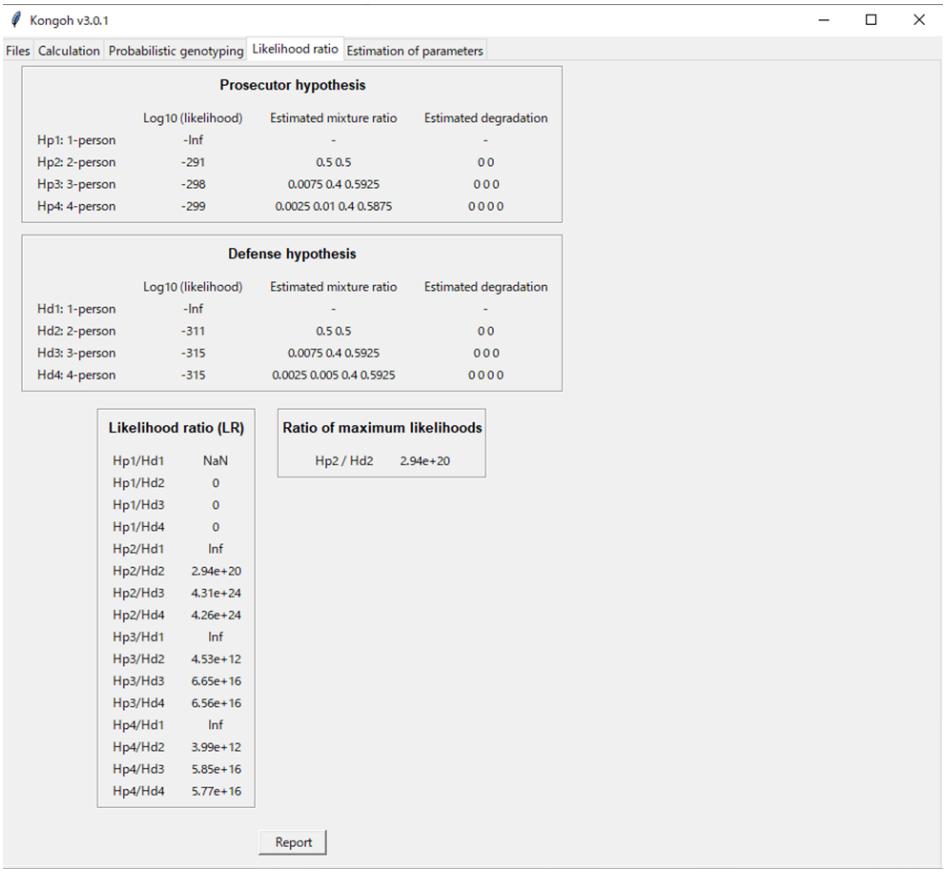

Fig. 11. Example of “Result” tab.

3.5. Verify results of probabilistic genotyping

1. Click the “Probabilistic genotyping” tab.
2. Set the conditions for “Hypothesis”, “Number of contributors”, and “Locus”.
3. Click the “Show” button.
4. The weight values of each genotype combination under the selected conditions are shown in Fig. 12.

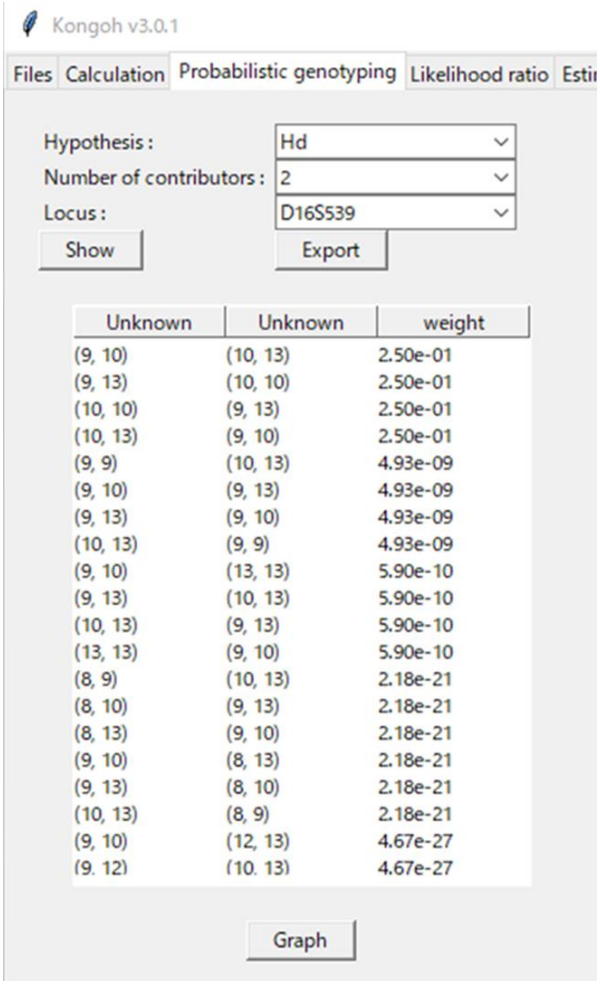

Fig. 12. Weight values of each genotype combination under selected conditions.

5. The observed peak heights and estimated gamma distributions can be compared by clicking the “Graph” button (Fig. 13).

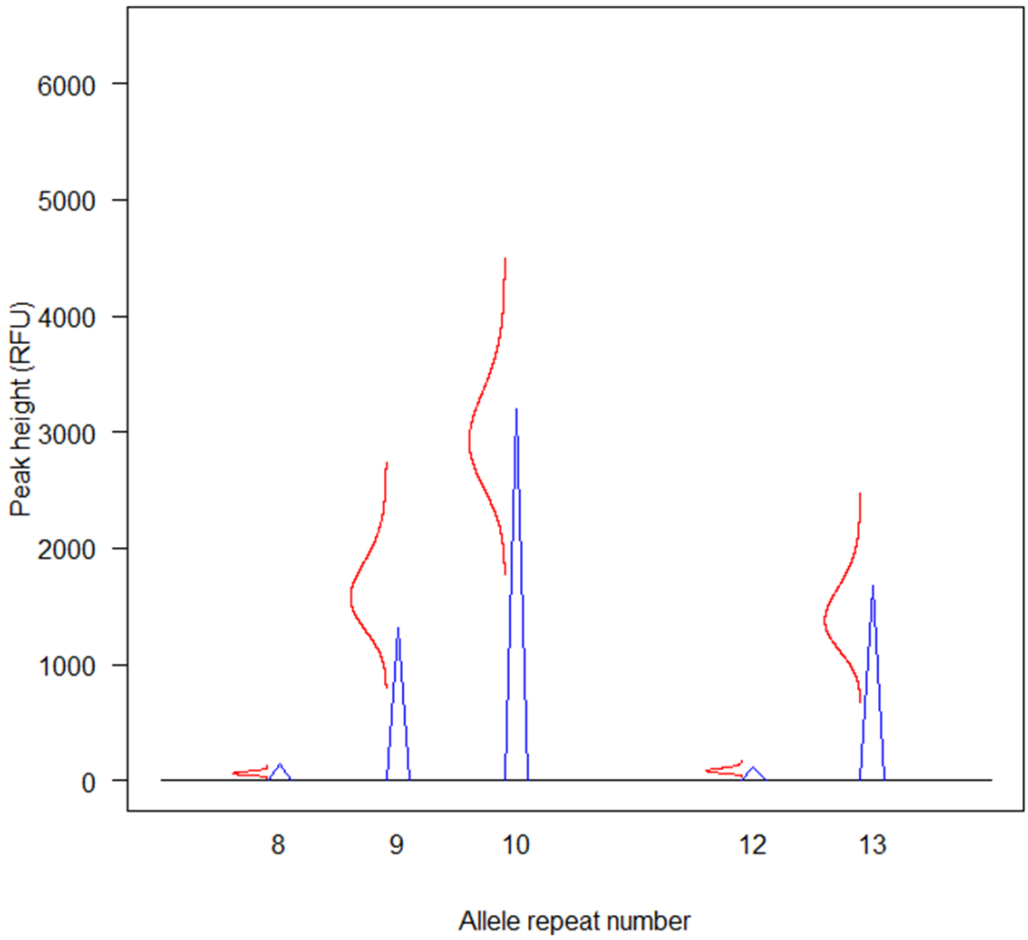

Fig. 13. Observed peak heights (blue) and estimated gamma distributions (red).

6. The weight values of all conditions can be exported by clicking the “Export” button.

3.6. Recalculation of likelihood ratio

When recalculating the likelihood ratio by changing H_p and H_d, click the "Calculation" tab, and then set the H_p and H_d. Fig. 14 shows an example of changing the hypotheses.

H_p : suspect (+ unknown contributor(s));

H_d : unknown contributor(s).

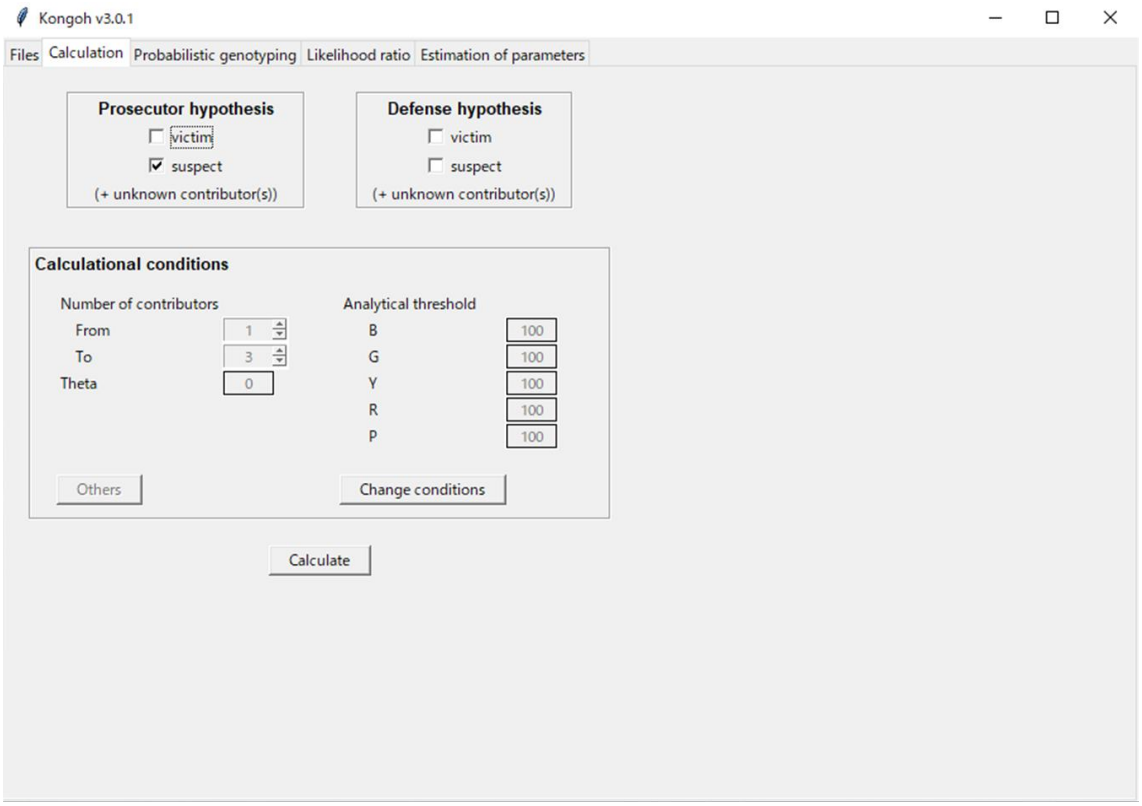

Fig. 14. "Calculation" tab for recalculating likelihood ratio.

After setting H_p and H_d, click the "Calculation" button. The calculation results are displayed immediately because the likelihoods of all assumed hypotheses, including the number of contributors and the combination of reference individuals, are already calculated in the first calculation.

When changing the calculational conditions, click the “Change conditions” button in the “Calculation” tab. The current calculation results will be deleted after the button is clicked.

3.7. Setting detailed conditions

After clicking the “Others” button in the “Calculation” tab, the “Setting other conditions” window is automatically opened (Fig. 15).

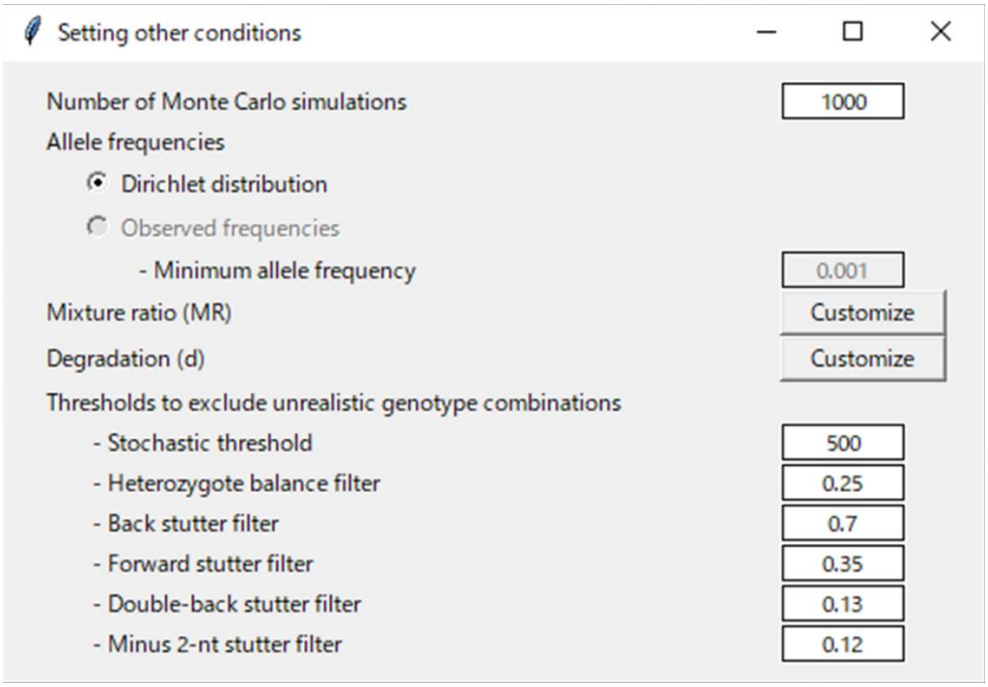

Fig. 15. “Setting other conditions” window.

Number of Monte Carlo simulations

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.

Allele frequencies

The setting depends on the input file of allele frequencies. If the allele frequencies are represented as allele counts, then the "Dirichlet distribution" is automatically selected. If the allele frequencies are represented as probabilities, then "Observed frequencies" is automatically selected and the value of "Minimum allele frequency" can be set.

Mixture ratio (MR)

By clicking the "Customize" button, the "Customize mixture ratio of one contributor" window (Fig. 16) is automatically opened. Users can manually add or delete the mixture ratio of one contributor.

The mixture ratio (MR_n) is the DNA proportion of contributor n ($n = 1, 2, \ldots, N$) in a crime stain profile, where N is the number of contributors. In *Kongoh*, MR_n must satisfy the condition $0 < MR_1 < MR_2 < \cdots < MR_{N-1} < MR_N$ ($N \geq 2$). In the "Customized mixture ratios" window, candidate values for MR_n ($n = 1, 2, \ldots, N-1$) can be set. MR_N is calculated as $1 - \sum_{n=1}^{N-1} MR_n$ in *Kongoh*.

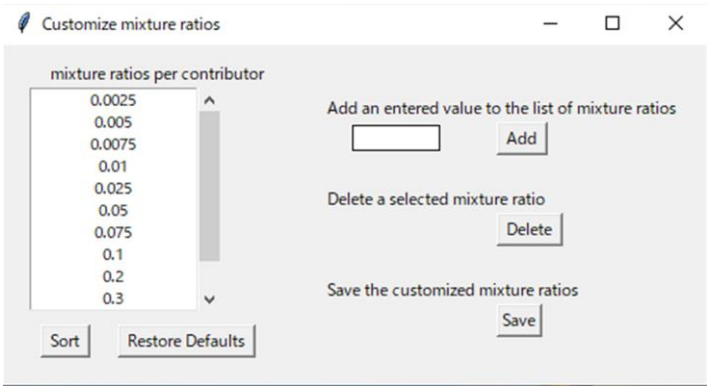

Fig. 16. Window for customizing mixture ratio of one contributor.

Degradation (d)

By clicking the “Customize” button, the “Customize degradation parameter” window (Fig. 17) is automatically opened. Users can manually add or delete the degradation parameters.

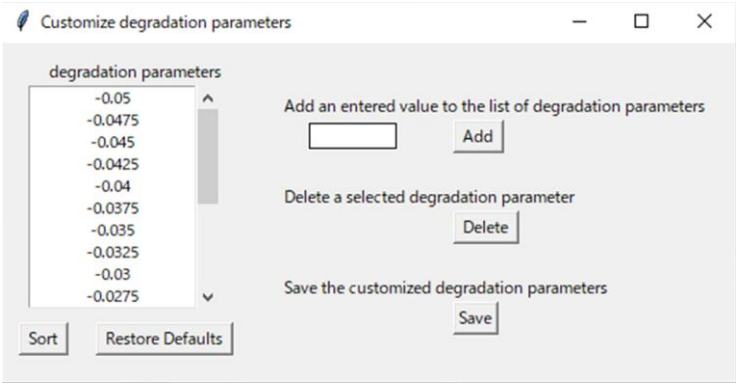

Fig. 17. Window for customizing degradation parameters.

Thresholds to exclude unrealistic genotype combinations

Stochastic threshold

If one of the allele peaks in a heterozygous genotype is above the threshold and the other allele peak is below the analytical threshold, then the genotype is regarded as unrealistic.

Heterozygote balance filter

If the heterozygote balance is less than the set value, the two peaks cannot be derived only from a single contributor. If one of the heterozygote allele peaks is below the stochastic threshold, then the filter will be ignored.

Back stutter filter, forward stutter filter, double-back stutter filter, and minus 2-nt stutter filter

If the stutter ratio is greater than the set value, then the peak of the stutter position cannot be derived from the stutter product.

4. Estimation of parameters

In *Kongoh* ver. 3.0.1, probability distribution models for biological parameters (i.e., locus-specific amplification efficiency, heterozygote balance, back stutter ratio, forward stutter ratio, double-back stutter ratio, and minus 2-nt stutter ratio) can be estimated using experimental data prepared by users. In this section, the preparation of the experimental data and procedure for the estimation of parameters are described.

4.1. Preparation of experimental data

The experimental data should satisfy the following conditions:

- The data are composed of single-source profiles.
- The genotypes of each profile are preliminary determined.
- 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.

The experimental data used for estimating default parameters for Identifiler Plus

(Parameter IDP.csv) and GlobalFiler (Parameter GF.csv) satisfy these conditions[4,5].

4.2. Input required files

1. Click the "Estimation of parameters" tab.
2. Click the "Open a new window", and the "parameter estimation" window is opened (Fig. 18).

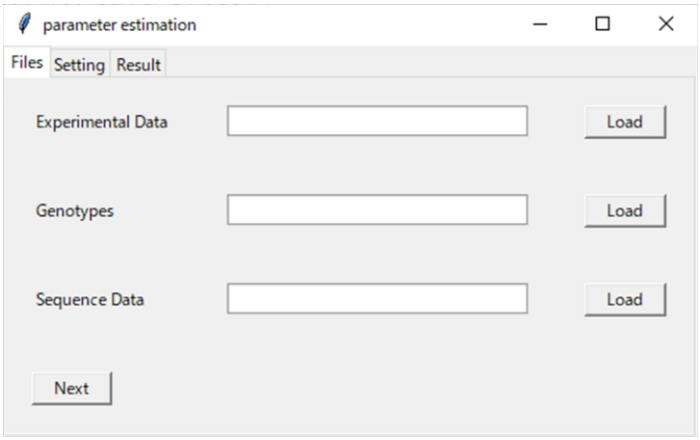

Fig. 18. "Parameter estimation" window.

3. Prepare the required files (.csv format) based on the "**Notes**" below.
4. Load three required files by clicking the "Load" button for each file.

[4] 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™ 3130xl Genetic Analyzer (Thermo Fisher Scientific) with injection parameters set at 3 kV for 10 s.

[5] 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/μL. 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™ 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).

5. Click the “Next” button.

Notes:

Experimental Data

- This file can be exported from the GeneMapper® ID-X software.
- This file must include information regarding the “Sample File”, “Sample Name”, “Marker”, “Dye”, “Allele”, “Size”, and “Height” as shown in Fig. 19.
- No filters are to be used to remove stutter peaks when analyzing sample electrophoretic data in software programs such as GeneMapper® ID-X software.
- Pull-up peaks and noises are not necessarily removed manually[6].
- Alleles and stutters that are not in the allelic ladders (i.e., off-ladder [OL]) must be renamed as appropriate repeat numbers. OL peaks other than alleles and stutters are not necessarily removed manually.

Genotypes

- This file must include information regarding the “Sample Name”, “Marker”, “Allele 1”, and “Allele 2” as shown in Fig. 20.
- Information regarding the “Sample Name” in the “Genotypes” file must be the same as that in the “Experimental Data” file.
- Two alleles in homozygotes must be entered in each column.

[6] *Kongoh* ver. 3.0.1 can automatically determine allele and stutter peaks based on the genotypes of each experimental data, as well as remove peaks located within ±1 base [4] of an allele peak(s) in different color channels to exclude the effect of spectral pull-up.

Sample File	Sample Name	Marker	Dye	Allele 1	Allele 2	Allele 3	...	Size 1	Size 2	Size 3	...	Height 1	Height 2	Height 3	...
Example1	Example1	D3S1358	B	OL	OL	12		98.47	106.22	109.39		13	14	14	
Example1	Example1	vWA	B	16	17	18		176.98	181.06	185.12		111	1668	1930	
Example1	Example1	D16S539	B	8	9	11		239.91	244.08	252.17		40	1345	71	
Example1	Example1	CSF1PO	B	OL	OL	8		278.51	282.58	291.09		28	26	37	
Example1	Example1	TPOX	B	7	8	10		346.68	350.81	358.91		18	929	17	
Example1	Example1	Yindel	G	1				81.4				2032			
Example1	Example1	AMEL	G	X	Y			98.79	104.91			1916	1630		
Example1	Example1	D8S1179	G	10	11	12		134.41	138.49	142.8		244	4546	35	
Example1	Example1	D21S11	G	28	29	30		199.56	203.65	207.57		137	2548	130	
Example1	Example1	D18S51	G	7	12	13		261.19	281.75	285.72		22	91	1811	
Example1	Example1	DYS391	G	9	10			373.37	377.42			34	945		
Example1	Example1	D2S441	Y	9	10	11		80.78	84.92	89.01		51	1315	1076	
Example1	Example1	D19S433	Y	OL	13	14		138.99	145.53	149.55		14	129	1980	
Example1	Example1	TH01	Y	6	8	9		187.24	195.47	199.38		858	16	538	
Example1	Example1	FGA	Y	20	21	21.2		251.62	255.59	257.57		115	1439	62	
Example1	Example1	D22S1045	R	10	11	12		94.4	97.35	100		31	1224	18	
Example1	Example1	D5S818	R	7	8	9		138.4	142.72	146.81		18	64	1655	
Example1	Example1	D13S317	R	7	8	OL		206.87	210.82	232.41		93	4335	20	
Example1	Example1	D7S820	R	9	10	11		274.69	278.6	282.58		92	2529	1967	
Example1	Example1	SE33	R	16	16.2	17		354.07	356.24	358.22		110	69	1749	
Example1	Example1	D10S1248	P	11	OL	12		97.91	99.28	101.97		28	28	125	
Example1	Example1	D1S1656	P	15	16	16.2		184.23	188.3	190.24		181	1887	34	
Example1	Example1	D12S391	P	OL	17	18		211.35	228.44	232.5		37	78	1299	
Example1	Example1	D2S1338	P	OL	18	19		279.25	309.13	313.13		28	93	1486	
Example2	Example2	D3S1358	B	9	OL	OL		97.27	98.55	104.56		16	17	16	
Example2	Example2	vWA	B	15	16	17		172.89	176.96	180.98		15	183	2900	
Example2	Example2	D16S539	B	8	9	10		239.91	244.12	248.14		112	2858	26	
Example2	Example2	CSF1PO	B	OL	OL	OL		278.68	282.69	285.48		33	36	19	
Example2	Example2	TPOX	B	7	8	10		346.82	350.87	359.11		46	3094	64	

Fig. 19. Format of experimental data.

Sample Name	Marker	Allele 1	Allele 2
Example1	D3S1358	15	15
Example1	vWA	17	18
Example1	D16S539	10	12
Example1	CSF1PO	9	11
Example1	TPOX	9	11
Example1	Yindel	1	1
Example1	AMEL	X	Y
Example1	D8S1179	12	12
Example1	D21S11	29	31
Example1	D18S51	13	15
Example1	DYS391	10	10
Example1	D2S441	11	11
Example1	D19S433	14	14
Example1	TH01	6	7
Example1	FGA	21	22
Example1	D22S1045	11	15
Example1	D5S818	9	11
Example1	D13S317	8	9
Example1	D7S820	10	11
Example1	SE33	17.2	27.2
Example1	D10S1248	13	15
Example1	D1S1656	15	17
Example1	D12S391	18	21
Example1	D2S1338	19	20
Example2	D3S1358	15	16
Example2	vWA	15	16
Example2	D16S539	9	9
Example2	CSF1PO	10	11
Example2	TPOX	8	8

Fig. 20. Format of genotypes.

Sequence Data

- This file must include information regarding the “Marker”, “Allele”, “Repeat Region”, and “Count” as shown in Fig. 21.
- “Count” implies the number of observations in sequence-based population data.

Marker	Allele	Repeat Region	Count
D1S1656	10	[TCTA]10	13
D1S1656	11	[TCTA]11	100
D1S1656	12	CCTA [TCTA]11	98
D1S1656	13	[TCTA]13	130
D1S1656	13	TCTA GCTA [TCTA]11	9
D1S1656	14	[TCTA]14	61
D1S1656	14.3	CCTA [TCTA]9 TCA [TCTA]4	5
D1S1656	15	CCTA [TCTA]14	291
D1S1656	15	CTTA [TCTA]14	1
D1S1656	15.3	CCTA [TCTA]11 TCA [TCTA]3	28
D1S1656	16	CCTA [TCTA]15	274
D1S1656	16	CTTA [TCTA]15	9
D1S1656	16.3	CCTA [TCTA]11 TCA [TCTA]4	138
D1S1656	17	CCTA [TCTA]16	81
D1S1656	17.3	CCTA [TCTA]12 TCA [TCTA]4	215
D1S1656	18	CCTA [TCTA]17	12
D1S1656	19.3	CCTA [TCTA]14 TCA [TCTA]4	19
TPOX	7	[AATG]7	15
TPOX	8	[AATG]8	448
TPOX	9	[AATG]9	285
TPOX	10	[AATG]10	124
TPOX	11	[AATG]11	505
TPOX	11	[AATG]10 TATG	1
TPOX	12	[AATG]12	106
D2S441	8	[TCTA]8	1
D2S441	9.1	A [TCTA]9	2
D2S441	10	[TCTA]8 TCTG TCTA	272
D2S441	10	[TCTA]10	149
D2S441	11	[TCTA]9 TCTG TCTA	57

Fig. 21. Format of sequence data.

4.3. Setting

1. By clicking the “Next” button in the “Files” tab, the “Setting” tab is automatically opened, as shown in Fig. 22.
2. Set conditions for estimating the parameters.

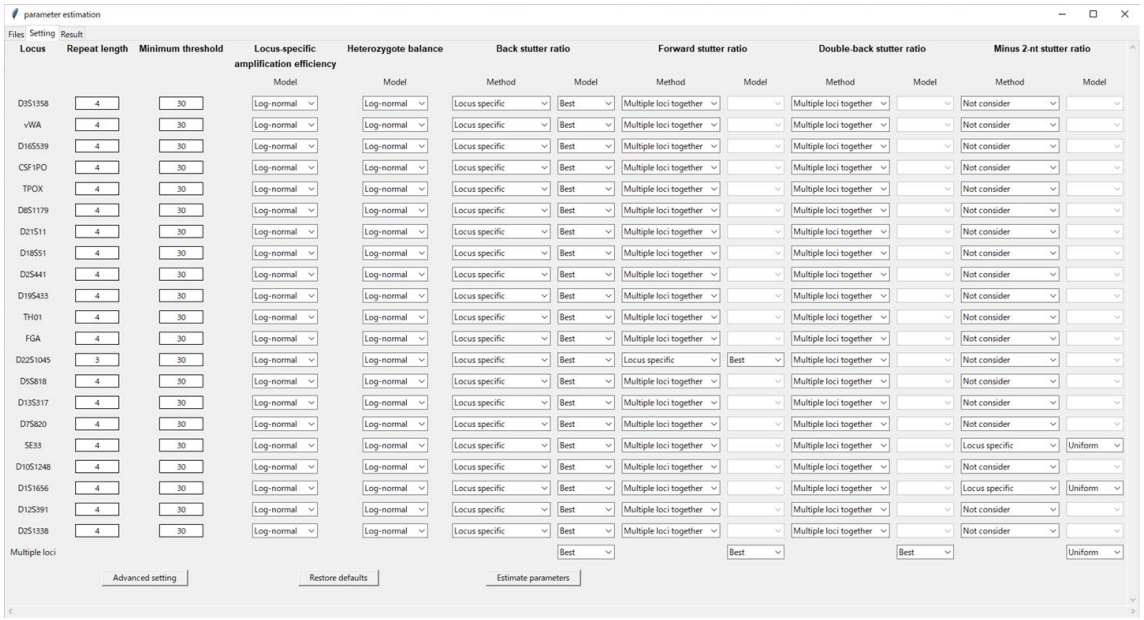

Fig. 22. “Setting” tab

Repeat length

Set the repeat lengths of each locus.

Minimum threshold

Set the minimum thresholds for each locus. Peaks below the thresholds are not used to estimate the parameters.

Locus-specific amplification efficiency

Model is fixed to “Log-normal” distribution in the current version of *Kongoh*.

Heterozygote balance

Model is fixed to “Log-normal” distribution in the current version of *Kongoh*.

Each type of stutter ratio

Select the “Method” and “Model” for each type of stutter ratio in each locus. Three options are available for the “Method,” as follows:

- **Locus specific**: The data of stutter ratios are considered separately from those in other loci.
- **Multiple loci together**: In the loci with the option, the data of stutter ratios are considered together.
- **Not consider**: Stutter ratios are not considered.

The options available for “Model” in each type of stutter ratio are shown in Table 3. Five options are available, as follows:

- **Best**: The best model for the stutter ratios is determined based on the AIC values of each model.
- **Allele**: Stutter ratios are positively correlated with allele numbers [5].
- **LUS**: Stutter ratios are positively correlated with the LUSs [6,7].
- **Multi-seq**: Stutter ratios are positively correlated with corrected allele numbers,

which consider not only the LUS, but also other uninterrupted stretches [8].

- **Uniform**: The mean value of the stutter ratios is uniform regardless of the allele number.

Table 3 Available options for "Model" for each stutter ratio type

Model	BSR	FSR	DSR	M2SR
Best	✓	✓	✓	
Allele	✓	✓	✓	
LUS	✓	✓	✓	
Multi-seq	✓	✓	✓	
Uniform		✓	✓	✓

When modeling stutter ratios based on the "Multiple loci together" method, the "Model" can be set in the bottom box of each "Model" column.

3. If necessary, click the “advanced setting” button to change the following conditions:
    - Minimum values (not logarithmic scale) of locus-specific amplification efficiency and heterozygote balance[7]
    - Maximum values (not logarithmic scale) of each stutter ratio type[8]
    - Initial, minimum, and maximum values of each parameter for maximum likelihood estimation[9]
4. Click the “Estimate parameters” button to estimate the parameters.

[7] The probability distributions of locus-specific amplification efficiency and heterozygote balance exhibit truncated log-normal distributions with the minimum and maximum parameter settings. The maximum parameter was set to 1/minimum.
[8] The probability distributions of each stutter ratio type exhibit truncated log-normal distributions with the maximum parameter setting.
[9] Initially, these conditions need not be changed. If the estimated parameters are not fitted to the experimental data, then these conditions should be changed.

4.4. Results

1. After the parameter estimation is completed, the "Result" tab appears automatically, as shown in Fig. 23. The models of all biological parameters at each locus are displayed.

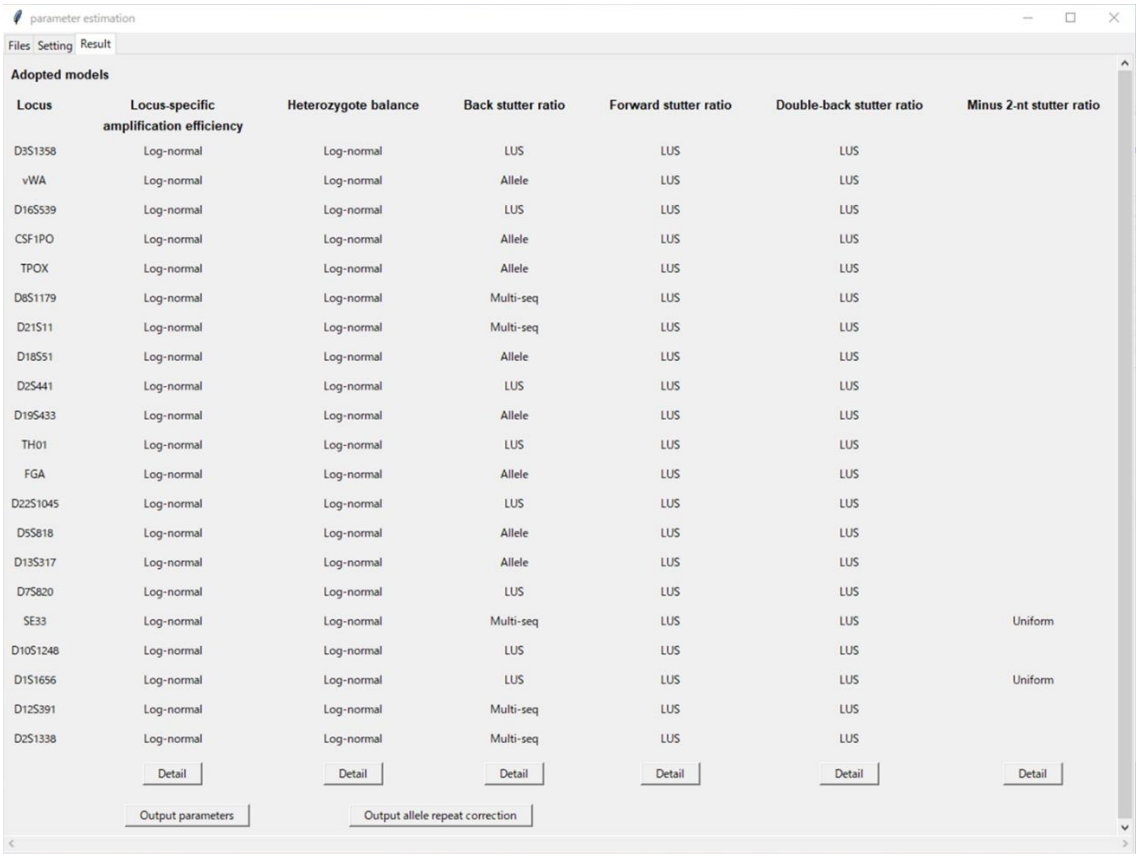

Fig. 23. Example of "Result" tab for parameter estimation.

2. The detailed results can be viewed by clicking the "Detail" button for each biological parameter. By clicking the button, a new window appears (Fig. 24).

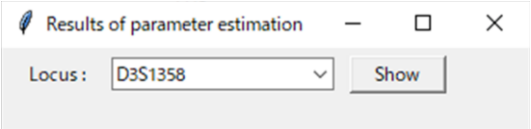

Fig. 24. Window showing detailed results.

3. Select a locus and click the “Show” button to display the detailed results of the selected locus (Fig. 25).

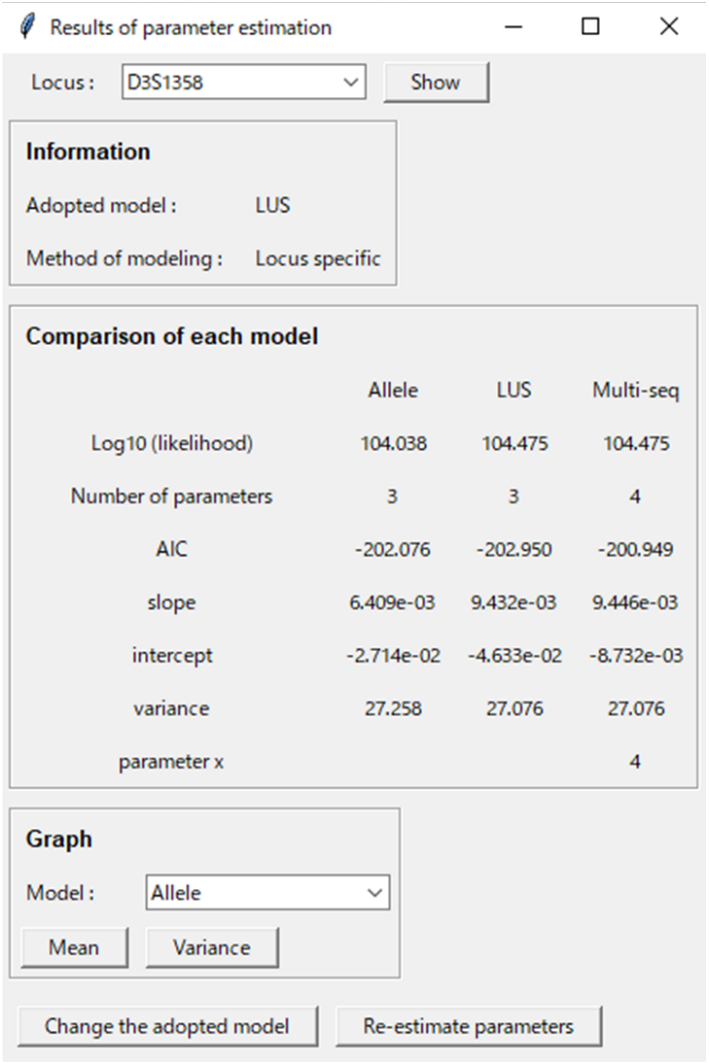

Fig. 25. Example of detailed results of back stutter ratios in D3S1358.

4. A brief overview of the results is displayed as follows:

Information

The adopted model[10] and modeling method are displayed.

Comparison of each model

The logarithmic likelihoods, number of parameters, AIC values, and each estimated parameter are displayed. When “Best” is selected in the “Model” of the “Setting” tab, the results of all compared models are displayed.

Graph

Users can verify whether the estimated parameters are fitted to the experimental data. The graph for variance can be described in terms of the locus-specific amplification efficiency and heterozygote balance. In each stutter ratio type, the graphs for mean and variance can be described by selecting a model arbitrarily. For example, the graphs for mean and variance of back stutter ratios can be described as shown in Figs. 26 and 27, respectively.

[10] The model featuring the smallest AIC was adopted.

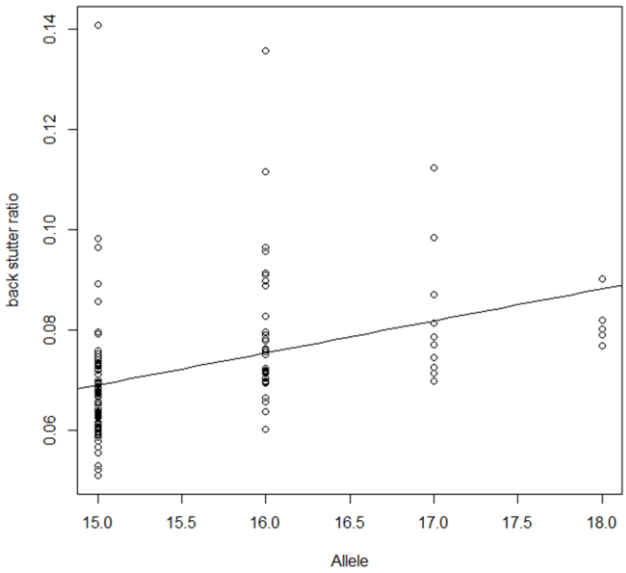

Fig. 26. Example of mean graph for back stutter ratios. Solid line indicates regression line of mean values.

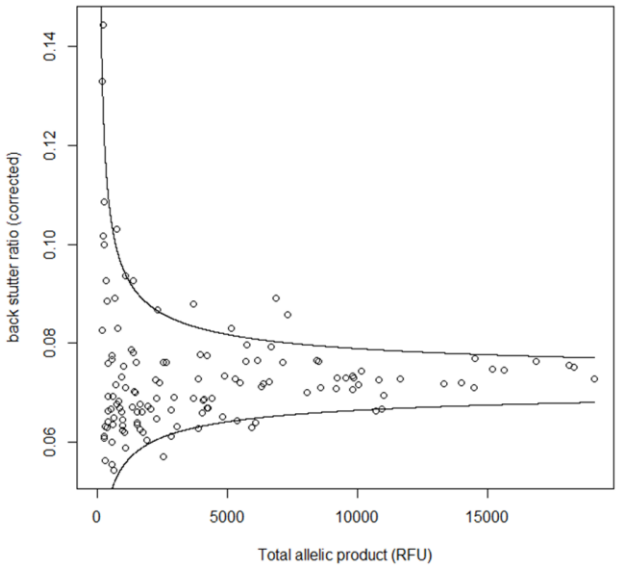

Fig. 27. Example of variance graph for back stutter ratios. Relationship between back stutter ratios and total allelic products is shown[11]. Upper and lower solid lines indicate 95% and 5% quantiles of back stutter ratios in each total allelic product, respectively. Back stutter ratios are corrected by eliminating difference in mean values in each allele.

[11] A total allelic product is the sum of peak heights of allelic and stutter products of an allele.

5. To change the adopted model for back stutter, forward stutter, and double-back stutter ratios, click the “Change the adopted model” button; subsequently, a new window will appear (Fig. 28).

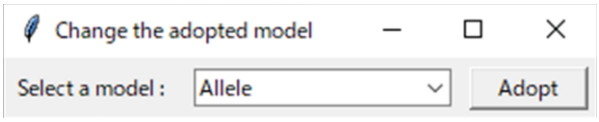

Fig. 28. “Change the adopted model” window.

6. Select the desired model, and then click the “Adopt” button. The user-selected model is marked with a dagger in the “Result” tab (Fig. 29).

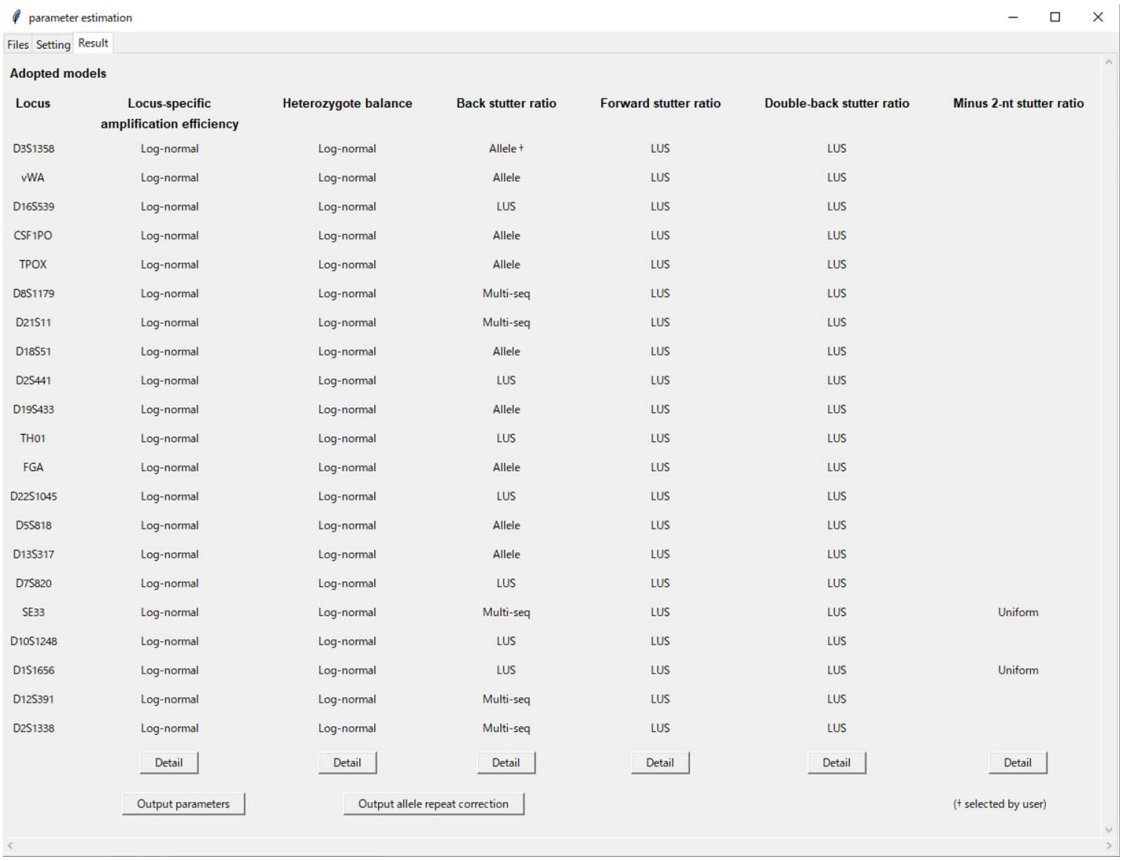

Fig. 29. Example of “Result” tab for parameter estimation. User-selected model is marked with a dagger (e.g., Allele model of back stutter ratio in D3S1358).

7. To re-estimate the parameters, click the “Re-estimate parameters” button; subsequently, a new window will appear (Fig. 30).

Fig. 30. A window of “Re-estimate parameters”.

8. Select the desired model for parameter re-estimation, and then click the “Set conditions” button.
9. Change the conditions of the initial, minimum, and maximum values of each parameter for the maximum likelihood estimation (Fig. 31).

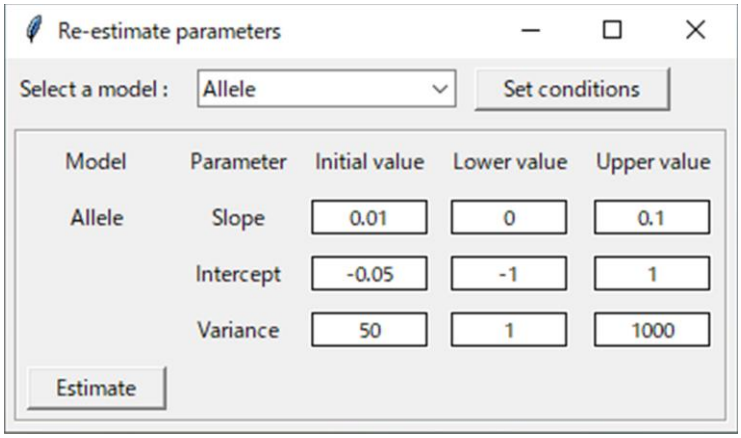

Fig. 31. Window for changing conditions of maximum likelihood estimation.

10. By clicking the “Estimate” button, the parameters are estimated. After completing the re-estimation, the “Result” tab is automatically opened.
11. The estimated parameters can be exported by clicking the “Output parameters” button. The export file can be used for the “Parameters for Monte Carlo simulation” file to analyze the crime stain profiles.
12. Information regarding the allele repeat correction can be exported by clicking the “Output allele repeat correction” button. The export file can be directly used in the “Allele repeat correction” file to analyze the crime stain profiles.

5. 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 (2017) e0188183.

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

[3] 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 (2015) 306–308.

[4] 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.

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

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

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

[8] D. Taylor, J.A. 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.