Kongoh version 2.0.1 User Manual

7th February 2019

Sho Manabe

Department of Forensic Medicine, Kyoto University Graduate School of Medicine

manabe@fp.med.kyoto-u.ac.jp

Contents

1. What is *Kongoh*? 3

2. Changes in ver.2.0.1 4

3. Tutorial 5

3.1. Getting started 5

3.2. Input a crime stain profile 7

3.3. Input reference profiles 8

3.4. Input allele frequencies 9

3.5. Calculation of likelihood ratios 10

3.6. Results 13

3.7. Setting 14

4. Appendix 16

4.1. Improving the allelic drop-out model 16

4.2. Changing the model of locus specific amplification efficiency 17

4.3. Changing the method of estimating the parameters of the gamma distribution 19

4.4. Alteration of the likelihood ratios by changing the calculation model 20

5. Reference 22

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]. *Kongoh* performs a Monte Carlo simulation based on probability distributions of the biological parameters determined by means of empirical data. Subsequently, the peak heights generated by the simulation are approximated by gamma distributions. 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).

The profile, typed by AmpFℓSTR® Identifiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA), can be interpreted using *Kongoh* in its current version. The Identifiler Plus system runs for 28 amplification cycles according to the manufacturer’s protocol. The PCR products are analyzed using an ABI 3130xl Genetic Analyzer (Thermo Fisher Scientific) with a 10 s injection time.

In *Kongoh*, there is no requirement to designate a peak, located at the position of the -1 backward stutter peak, as an allele or stutter because the derivation of the peak in the stutter position can be determined probabilistically. Thus, we can remove the stutter filters of all loci. Moreover, *Kongoh* considers the allelic drop-out, which is the event of a peak under the analytical threshold. In contrast, the drop-in is not considered, however, spontaneous drop-in peaks could be explained by additional unknown contributors.

The likelihood ratios are calculated by the ratio of maximum likelihoods in prosecution and defense hypotheses. The likelihoods of 1–4 contributors are automatically calculated; therefore, the number of contributors does not need to be determined manually prior to the analysis.

2. Changes in ver. 2.0.1

- The equation for the probability density of observing an allelic drop-out was changed as follows:

  $f(O_{al}|G_{l,i}, MR_n, d) = \frac{\sum_{Z=1}^{\mathrm{AT}-1} f(Z|G_{l,i}, MR_n, d)}{\mathrm{AT}-1}$ (ver.1.0.1),

  $f(O_{al}|G_{l,i}, MR_n, d) = \sum_{Z=1}^{\mathrm{AT}-1} f(Z|G_{l,i}, MR_n, d)$ (ver.2.0.1).

- The model of locus specific amplification efficiency was changed from a normal distribution to a log-normal distribution. Therefore, the R-package “truncnorm” is no longer required.
- The shape parameter and scale parameter of the gamma distribution are calculated from expected peak heights, which are generated by Monte Carlo simulations. The R-package “MASS” is no longer required.
- The variance parameters of the stutter ratio were modified.
- The “Calculation” tab was simplified by merging two calculation buttons together. The range of the assumed contributor numbers can be changed in the “Setting” tab.
- The progress bars for the calculations were improved.
- A button for the output weight values was added in the “Result” tab.
- All required packages are automatically installed when *Kongoh* is launched.
- An open source license (GNU GPLv3) was added.

(Further information can be found in the Appendix)

3. Tutorial

3.1. Getting started

First, ensure that the R software has been installed. It is available from the R Development Core Team website (http://www.R-project.org).

The *Kongoh* program is freely available at GitHub (https://github.com/manabe0322/Kongoh/releases) and can be accessed by clicking the file named “Kongoh v2.0.1.RData”.

After the installation of the R software and the *Kongoh* program, start an R session. Subsequently, load the “Kongoh v2.0.1.RData” file from “Load Workspace” in the “File” tab (Fig. 1).

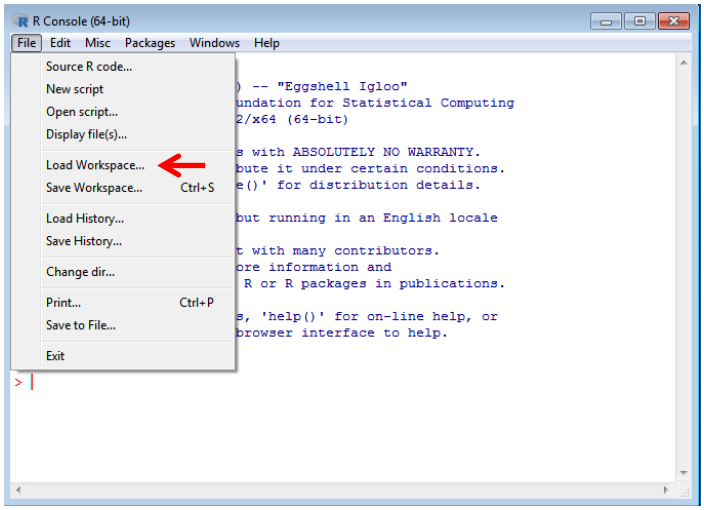

Fig. 1. “Load Workspace” in the “File” tab.

After loading the file, the *Kongoh* software is launched by the following command:

```
Kongoh()
```

All required packages used in *Kongoh* (tcltk, tcltk2, gtools, and snow) are automatically installed.

After all packages are loaded, the "Files" tab opens as shown in Fig. 2.

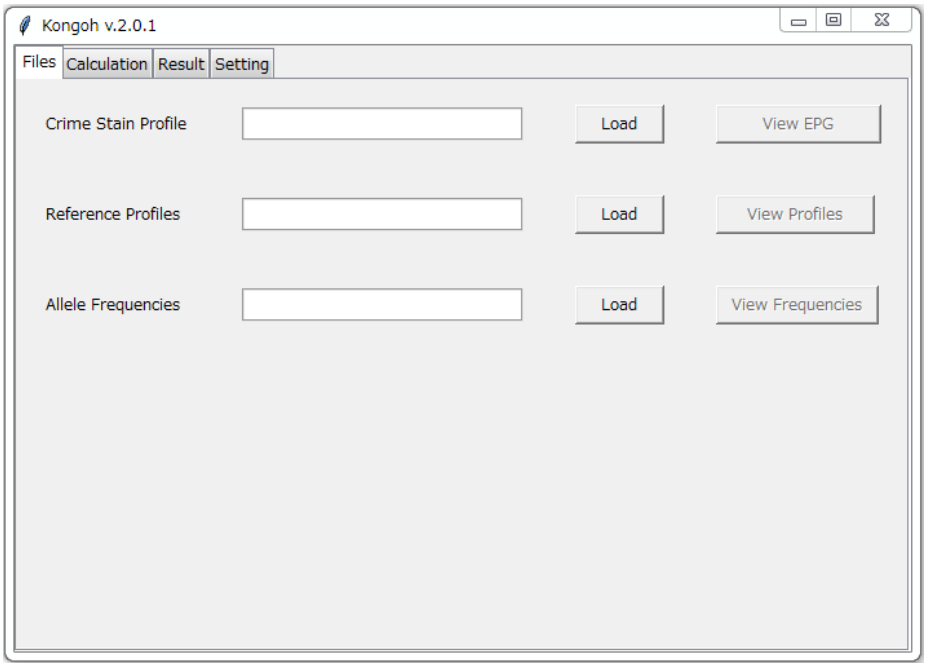

Fig. 2. The "Files" tab, where the user can import a crime stain profile, reference profiles, and allele frequencies.

3.2. Input a crime stain profile

After the “Files” tab is opened as shown in Fig. 2, press the “Load” button for the crime stain profile. The profile must be typed by the Identifiler Plus Kit and analyzed using an ABI 3130xl Genetic Analyzer according to the manufacturer’s protocol. The peak, located at the position of the −1 backward stutter, does not need to be designated as an allele or stutter because the derivation of the peak can be determined probabilistically. Forward stutters, −2 backward stutters, and pull-up peaks need to be removed manually.

The input file of the crime stain profile should be in .csv format as shown in Fig. 3. The file must include information of the “Sample File,” “Marker,” “Allele,” “Size,” and “Height.” This file could also be exported from the GeneMapper® software. After loading the file for a crime stain profile, the electropherogram can be saved into a PDF file by pressing the “View EPG” button.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Sample File	Marker	Allele 1	Allele 2	Allele 3	Allele 4	Size 1	Size 2	Size 3	Size 4	Height 1	Height 2	Height 3	Height 4
2	Mixture	D8S1179	10	12	13	15	130.02	138.02	142.17	150.5	64	106	1423	112
3	Mixture	D21S11	29	30	31.2	32.2	202.76	206.71	212.64	216.73	52	695	41	550
4	Mixture	D7S820	10	11	12	13	270.19	274.16	278.23	282.26	465	60	133	367
5	Mixture	CSF1PO	10	11	12		318.68	322.82	326.82		466	552	87	
6	Mixture	D3S1358	15	16	17		122.55	126.63	130.59		442	1083	257	
7	Mixture	TH01	6	7	9		168.67	172.72	180.86		494	178	647	
8	Mixture	D13S317	11	12			227.45	231.43			153	1032		
9	Mixture	D16S539	9	10	12	13	267.28	271.33	279.29	283.22	161	586	46	514
10	Mixture	D2S1338	16	17	22	23	308.81	312.83	333.3	337.24	129	468	39	571
11	Mixture	D19S433	12	13	14		112.22	116.29	120.24		39	634	343	
12	Mixture	vWA	15	16	17		168.24	172.3	176.36		55	535	460	
13	Mixture	TPOX	8	9	11		228.83	232.9	240.98		522	358	62	
14	Mixture	D18S51	13	14	15	19	285.22	289.22	293.22	309.17	36	586	546	71
15	Mixture	AMEL	X	Y			105.07	110.92			535	528		
16	Mixture	D5S818	9	11	12		140.75	148.98	153.03		418	71	273	
17	Mixture	FGA	20	21	22	23	225.19	229.35	233.42	237.49	30	526	362	124

Fig. 3. The format of the crime stain profile.

3.3. Input reference profiles

You can input reference profiles by pressing the respective “Load” button. The profiles must include 15 short tandem repeat (STR) loci in the Identifiler system. The input file of the reference profiles should be in .csv format as shown in Fig. 4. This file must include information of the “Marker” and the name of each profile (e.g., victim and suspect). Consequently, you can view the profiles by pressing the “View Profiles” button.

	A	B	C	D	E
1	Marker	Victim	Victim	Suspect	Suspect
2	D8S1179	13	13	10	15
3	D21S11	30	32.2	30	30
4	D7S820	10	13	11	12
5	CSF1PO	10	11	11	12
6	D3S1358	16	16	15	17
7	TH01	6	9	7	7
8	D13S317	12	12	11	12
9	D16S539	10	13	9	10
10	D2S1338	17	23	16	23
11	D19S433	13	14	13	13
12	vWA	16	17	16	17
13	TPOX	8	9	8	11
14	D18S51	14	15	14	19
15	AMEL	X	Y	X	Y
16	D5S818	9	12	9	11
17	FGA	21	22	23	23

Fig. 4. The format of the reference profiles.

3.4. Input allele frequencies

You can input allele frequencies by pressing the respective “Load” button. The input file of the allele frequencies should be in .csv format as shown in Fig. 5. The file must include information of the “Allele” and the name of each locus. Consequently, you can view the frequencies by pressing the “View Frequencies” button. To calculate the frequencies of rare alleles, which are not observed in the population database, you should set the minimum allele frequency in the “Setting” tab as described in section 3.7.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
2	5						0.001852									
3	6						0.223621									
4	7	0.001852		0.003332	0.010367		0.266938	0.001852	0.001852				0.001852		0.002962	
5	8			0.12699	0.001852		0.066642	0.267308	0.002221				0.454646		0.007034	
6	9	0.001852		0.045909	0.04702		0.398741	0.129211	0.353943				0.114402		0.086264	
7	9.1			0.001852												
8	9.2										0.001852					
9	9.3						0.035172									
10	10	0.132914		0.219178	0.223621		0.009256	0.115143	0.196964				0.363199	0.002592	0.201037	
11	10.1			0.001852												
12	10.2										0.001852					
13	10.3			0.001852												
14	11	0.109219		0.328767	0.208071			0.221399	0.187338		0.004073		0.363199	0.004813	0.292484	
15	11.2										0.001852					
16	12	0.122917		0.235098	0.418734	0.002221		0.202518	0.178823		0.040726		0.035913	0.04813	0.235468	
17	12.2										0.005553					
18	13	0.225102		0.035172	0.069234	0.001852		0.051462	0.072936		0.287671	0.001852	0.001852	0.199556	0.166975	
19	13.2										0.030359					
20	14	0.205109		0.006294	0.018141	0.029248		0.013328	0.008515		0.34987	0.194372	0.001852	0.22251	0.009256	
21	14.2										0.088486					
22	15	0.134765		0.001852	0.005553	0.39615		0.001852	0.001852		0.051092	0.027027		0.168456	0.001852	
23	15.2										0.115143					
24	16	0.064421			0.001852	0.306553				0.008886	0.005553	0.184376		0.125879		
25	16.2										0.019622					
26	17	0.006664				0.199926				0.097742		0.282858		0.081822		0.003702
27	17.1													0.001852		
28	17.2										0.003332					
29	18	0.001852				0.06368				0.1592	0.001852	0.225842		0.048501		0.021844
30	19					0.003332				0.209182		0.074047		0.036653		0.067382
31	20									0.105887		0.010367		0.022214		0.089226
32	21					0.001852				0.01518		0.002962		0.01592		0.131063
33	22									0.050722		0.001852		0.013328		0.201777
34	22.2															0.002221
35	23									0.146983				0.007775		0.205479
36	23.2															0.005183
37	24									0.106108				0.004073		0.157349
38	24.2															0.001852
39	25									0.061829				0.001852		0.073676
40	25.2															0.002221
41	26									0.029248				0.001852		0.03221
42	27		0.001852							0.008886				0.001852		0.008145
43	28		0.042577							0.002592						0.002592
44	28.2		0.005183													
45	29		0.246946													

Fig. 5. The format of the allele frequencies.

3.5. Calculation of likelihood ratio

After loading the three required files, press the “Calculation” tab, Fig. 6.

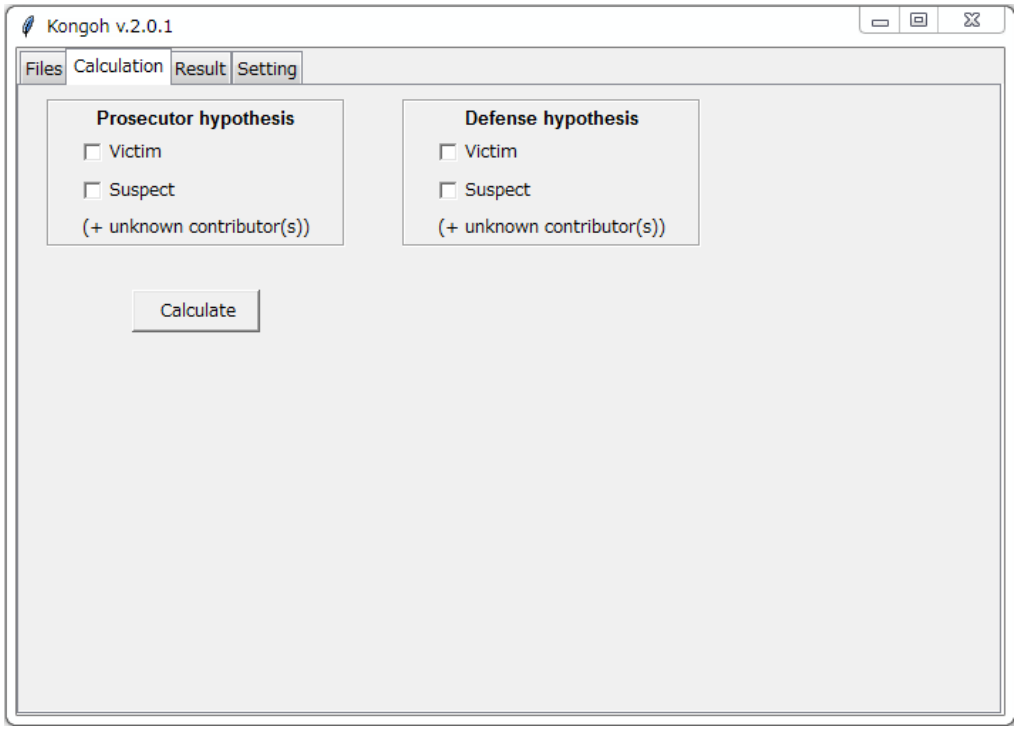

Fig. 6. The “Calculation” tab where the user can calculate the likelihood ratios.

You can calculate the likelihood values by setting both prosecutor (H_p) and defense (H_d) hypotheses. Check the individuals, to include them as contributors in each hypothesis. Fig. 7 shows an example of setting the hypotheses:

H_p : victim + suspect (+ unknown contributors)

H_d : victim (+ unknown contributors).

The number of unknown contributors does not have to be selected because *Kongoh* automatically calculates the likelihoods of 1–4 contributors in both H_p and H_d. The range of the assumed numbers of contributors can be changed in the "Setting" tab. Each parameter value can be changed before calculation by pressing the "Setting button". After each hypothesis has been set, the likelihood values can be calculated by pressing the "Calculation" button.

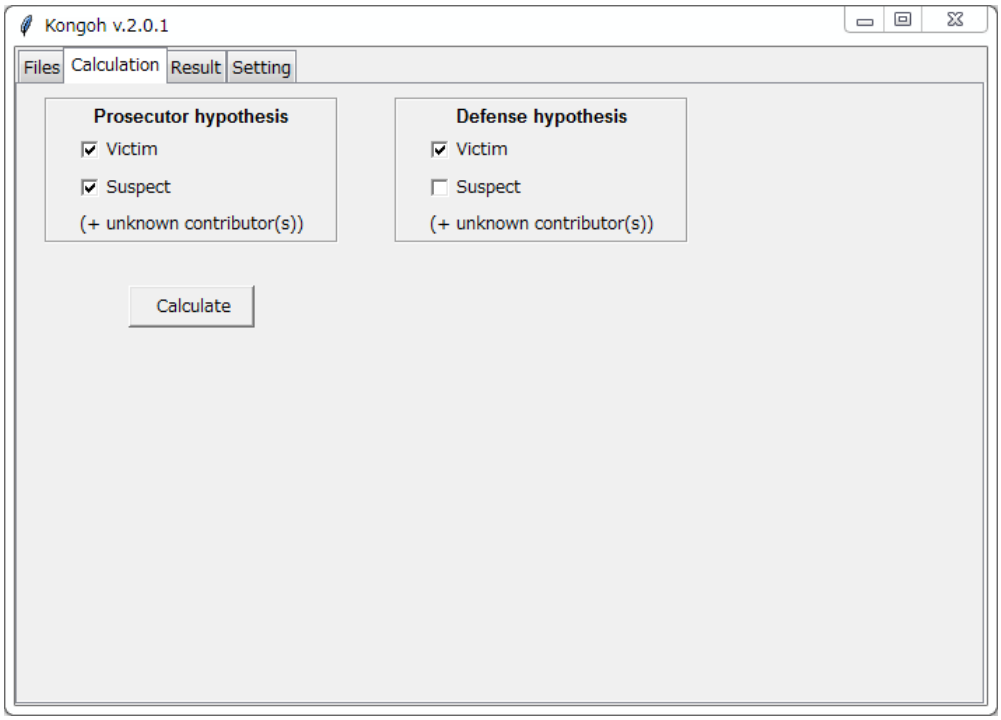

Fig. 7. An example of setting the hypotheses.

After pressing the “Calculate” button, a “Progress Bar” window appears, Fig. 8. Initially, *Kongoh* calculates the weight values of all possible genotype combinations for the 1–4 contributors. Subsequently, the likelihood values in both H_p and H_d are calculated by changing the number of contributors from one to four.

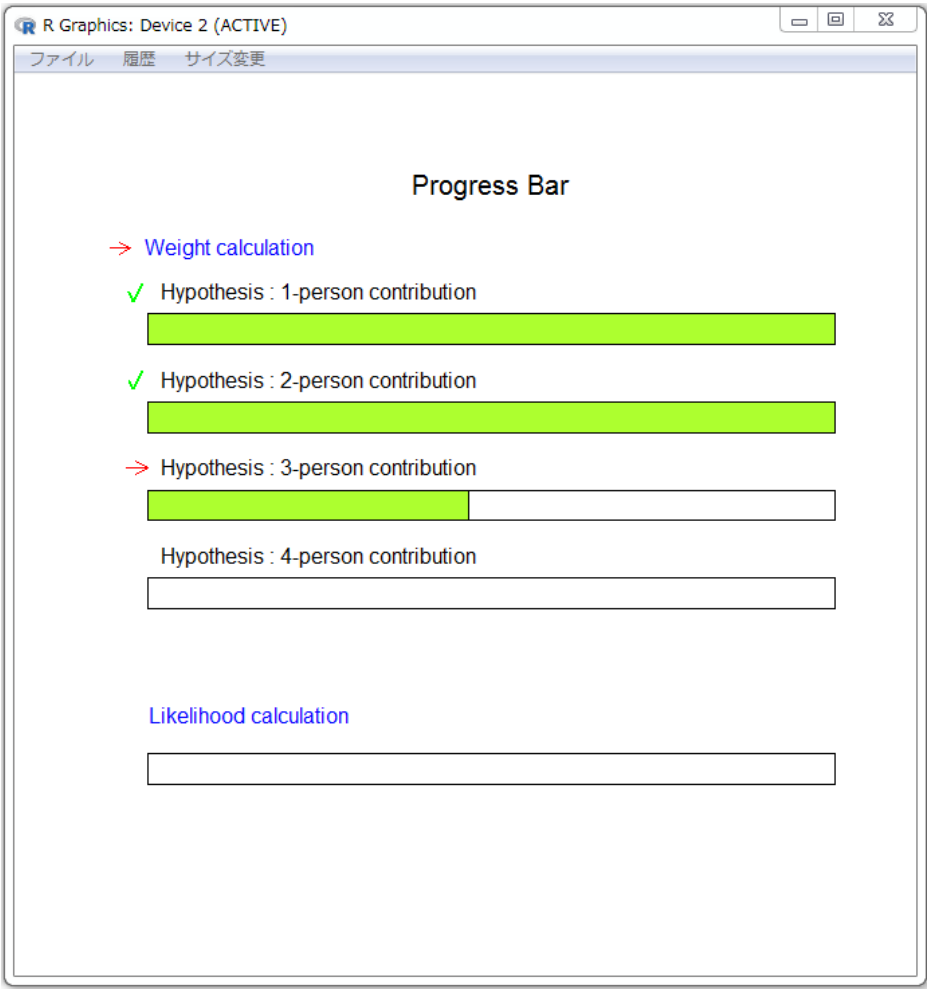

Fig. 8. A “Progress Bar” window for the calculation steps.

3.6. Results

After finishing the likelihood calculation, the "Result" tab automatically appears as is shown in Fig. 9. A brief overview of the results is displayed (i.e., likelihoods and estimated parameters in H_p and H_d, likelihood ratios, and the ratio of maximum likelihood in H_p and H_d). The report can be exported into a .csv file by pressing the "Report" button. The report includes detailed information such as the set parameter values, hypotheses, likelihoods in each locus, and estimated mixture ratios including information of each contributor. The weight values of each locus can be exported as well into .csv files by pressing the "Output weights". These files are automatically exported in the working directory.

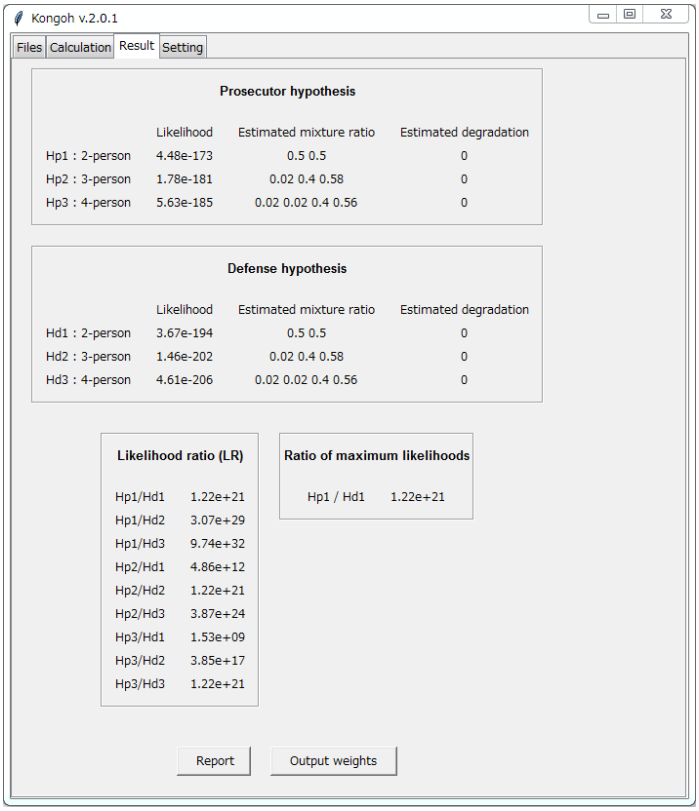

Fig. 9. An example of the "Result" tab.

3.7. Setting

Parameters can be changed from the "Setting" tab before each calculation, Fig. 10.

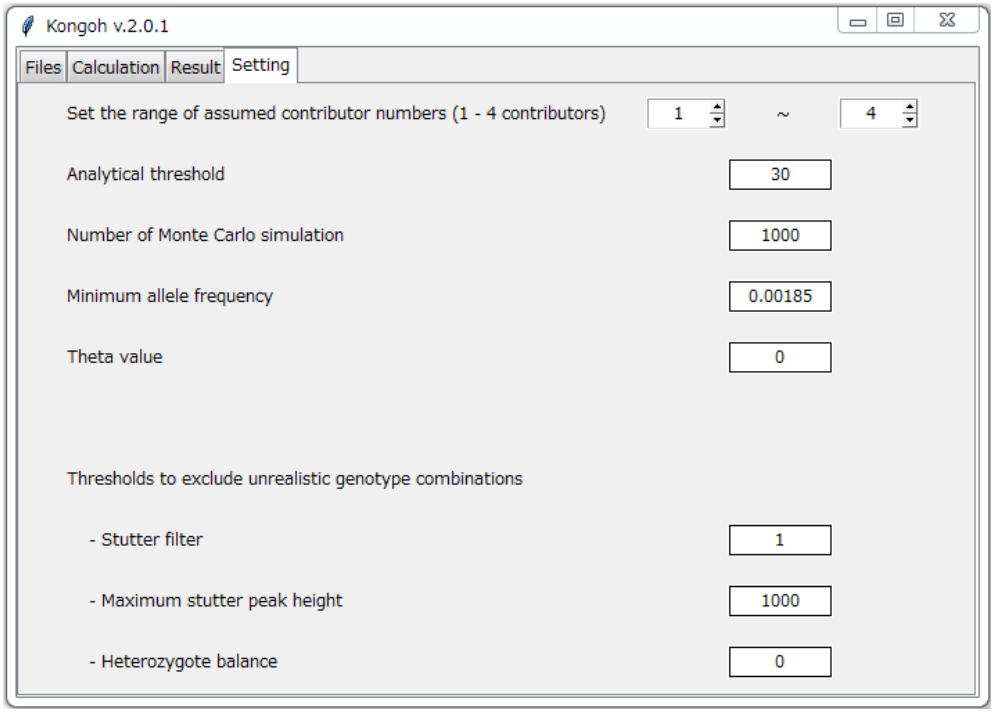

Fig. 10. The "Setting" tab. Default parameter values are already entered in each entry box.

Range of assumed contributor number:

You can select 1 to 4 contributors. The likelihoods of all set numbers are calculated in both H_p and H_d.

Analytical threshold:

Peaks below the analytical threshold are not used for the calculation. *Kongoh* ver. 2.0.1 was validated using 30 relative fluorescence units (RFU).

Number of Monte Carlo simulations:

The default value is 1,000, which is enough for the simulation. If the number of simulations is increased, the result becomes more robust; however, the runtime is also increased.

Minimum allele frequency:

The frequency is used for rare alleles, which are not observed in the population database. The default value is calculated by the 5 / 2N rule [2] in the Japanese population including 1,350 individuals [3] (i.e., $5 / (2 \times 1{,}350) \approx 0.00185$). This value should be changed according to the population data used.

Theta value:

This value is used to take into consideration the population substructure.

Thresholds to exclude unrealistic genotype combinations:

Stutter filter:

If a stutter ratio is greater than the set value, it is not possible to derive the stutter position's peak only from the stutter product.

Maximum stutter peak height:

If a stutter position's peak is greater than the set value, it is not possible to derive the peak only from the stutter product.

Heterozygote balance:

If a heterozygote balance is less than the set value, the two peaks cannot be derived only from a single contributor.

4. Appendix

4.1. Improving the allelic drop-out model

The weight values ($w_{l,i}$), which represent how good the observed peak heights fit to the genotype combination ($G_{l,i}$), are calculated by comparing the observed peak heights (O_{al}) with the expected peak heights under the condition of $G_{l,i}$, mixture ratio (MR_n), and DNA degradation (d). The equation is as follows:

$$w_{l,i} = \prod_a f(O_{al}|G_{l,i}, MR_n, d).$$

When the O_{al} is less than the analytical threshold (AT), the allele is regarded as drop-out.

In *Kongoh* ver. 1.0.1, the O_{al} value is regarded as one of the integers between 1 to AT – 1, each of which are assumed equally probable. Therefore, $f(O_{al}|G_{l,i}, MR_n, d)$ for the drop-out peak is calculated by the "mean" of $f(Z|G_{l,i}, MR_n, d)$ for each Z (i.e., integers between 1 to AT – 1). The equation is as follows:

$$f\left(O_{al}\middle|G_{l,i}, MR_n, d\right) = \frac{\sum_{Z=1}^{\mathrm{AT}-1} f\left(Z\middle|G_{l,i}, MR_n, d\right)}{\mathrm{AT}-1}.$$

In *Kongoh* ver. 2.0.1, the allelic drop-out is regarded as the event in which the allele peak is below AT. Therefore, the $f(O_{al}|G_{l,i}, MR_n, d)$ for the drop-out peak is calculated by the "sum" of $f(Z|G_{l,i}, MR_n, d)$ for each Z. The equation is as follows:

$f(O_{al}|G_{l,i}, MR_n, d) = \sum_{Z=1}^{AT-1} f(Z|G_{l,i}, MR_n, d)$.

By changing the model of the allelic drop-out, the estimated probability of the allelic drop-out is higher in *Kongoh* ver. 2.0.1 than in *Kongoh* ver. 1.0.1.

4.2. Changing the model of locus specific amplification efficiency

The locus-specific amplification efficiency (AE_l) is the relative amplification level of each locus. The variability of AE_l typically increases because of the stochastic effects when amplifying the low levels of a DNA template. Original data of 234 single-source profiles suggested that AE_l followed a normal distribution in each locus [4]. However, these data do not include low-template DNA samples.

Therefore, the data of low-template DNA samples were obtained from the publicly available data sets in the Project Research Openness for Validation with Empirical Data (PROVEDIt: https://lftdi.camden.rutgers.edu/provedit/files/). The data is 392 single-source profiles without DNA degradation, and include low-template DNA samples (0.0078–0.73 ng). The model of AE_l was reconsidered by using these data.

Fig. 11 shows the relationship between AE_l and the sum of peak heights in D3S1358 locus. The data demonstrated that AE_l tended to vary widely as the sum of peak heights decreased (i.e., DNA amount decreased). In addition, the distribution of AE_l was upper tailed in each peak height. This result seems to suggest that a log-normal distribution more appropriately represents the variability of AE_l rather than a normal distribution. Other loci have the same tendency as the D3S1358 locus.

Therefore, the model of locus specific amplification efficiency was changed from a

normal to a log-normal distribution in *Kongoh* ver. 2.0.1. In addition, the variance parameters of stutter ratio (SR_{al}) were also corrected by using the above-mentioned PROVEDIt dataset.

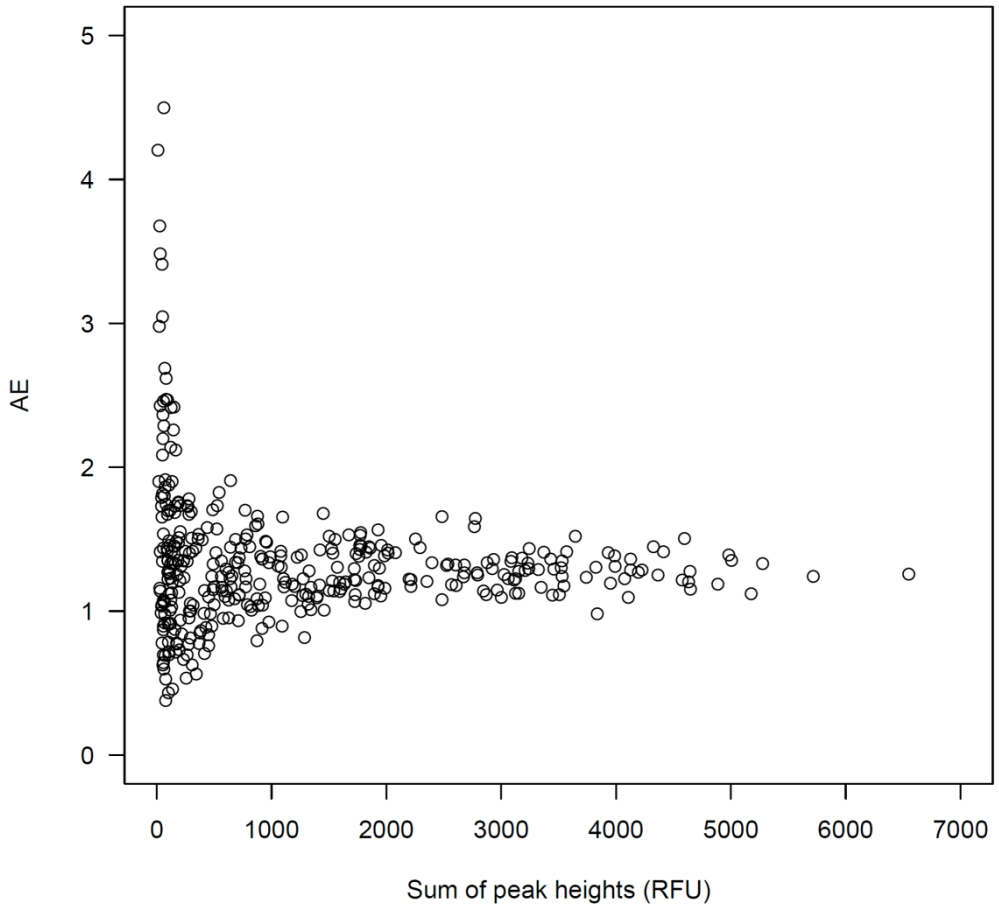

Fig. 11. The relationship between AE_l and the sum of peak heights in the D3S1358 locus.

4.3. Changing the method of estimating the parameters of the gamma distribution

The variability of the expected peak heights (E_{al}) in each allele or stutter generated by Monte Carlo simulation is approximated by gamma distributions in *Kongoh*. The gamma distribution of E_{al} is expressed as follows:

$$f(E_{al}|k,\theta) = \frac{(E_{al})^{k-1}}{\Gamma(k)\theta^{k}} \exp\left(-\frac{E_{al}}{\theta}\right).$$

In *Kongoh* ver. 1.0.1, the shape parameter (k) and scale parameter (θ) are determined by fitdistr function of R-package MASS, which can perform the maximum likelihood estimation. However, it is sometimes difficult to fit a gamma distribution to the variability of E_{al} values especially when extremely small amount of DNA samples are analyzed.

Therefore, in *Kongoh* ver. 2.0.1, the k and θ are calculated from expected peak heights. The mean and the variance of the gamma distribution are defined as follows:

$$E(E_{al}) = k\theta$$
$$V(E_{al}) = k\theta^2$$

where $E(E_{al})$ and $V(E_{al})$ are the mean and the variance of E_{al}, respectively. The k and θ are calculated using the following equations:

$$k = \frac{\left(E(E_{al})\right)^2}{V(E_{al})}$$

$$\theta = \frac{V(E_{al})}{E(E_{al})}$$

4.4. Alteration of the likelihood ratios by changing the calculation model

The LR values of *Kongoh* ver. 2.0.1 were compared with those of *Kongoh* ver. 1.0.1 using experimentally prepared mixtures. Fig. 12 shows the LR values in a two-person mixture with a mixture ratio of 9:1. The DNA amount of the mixture was 0.25 ng, and there were 11 drop-out alleles of the minor contributor. If the minor contributor is a person of interest (POI), the LR for the POI was strongly supportive for defense hypothesis (i.e., the POI is not a contributor) in ver. 1.0.1. Conversely, the LR was 2.92×10^3 in ver. 2.0.1, and similar to other quantitative continuous model software (i.e., *EuroForMix* ver. 1.9 [5] and *likeLTD* ver. 6.3.0 [6]). If the major contributor is a POI, the LR values for the POI were similar in each version of *Kongoh* because there was no drop-out allele of the major contributor. The same tendency was also observed in a three-person mixture with the mixture ratio of 3:2:1, Fig. 13.

The differences of the LR values between each version of *Kongoh* were mainly due to improving the allelic drop-out model. Changing the model of locus specific amplification efficiency, and changing the method of estimating the shape and scale parameters of the gamma distribution had little effect on the differences.

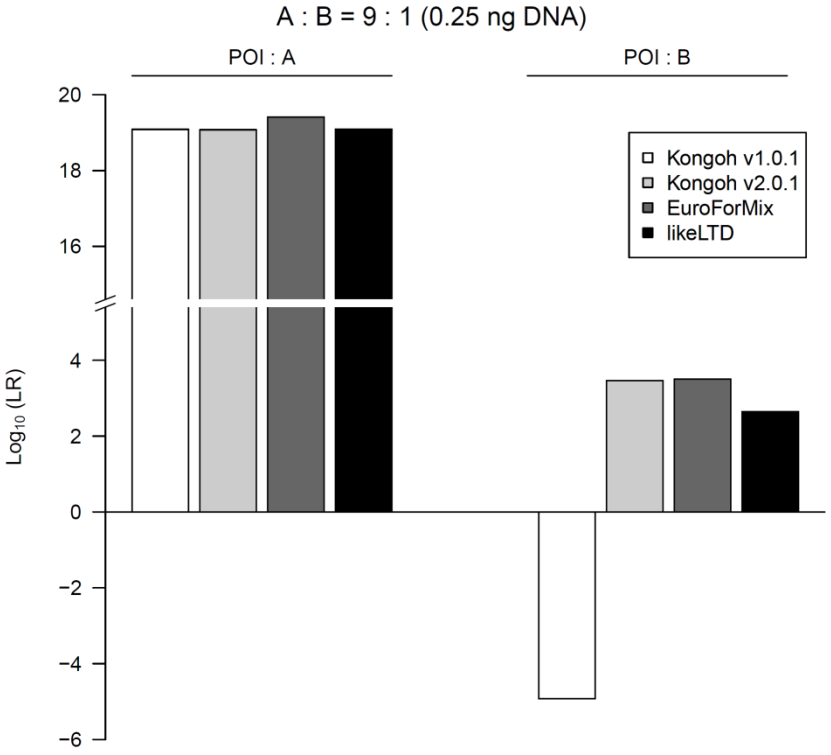

Fig. 12. LR values of each version of *Kongoh* and two more software in a 9:1 mixture.

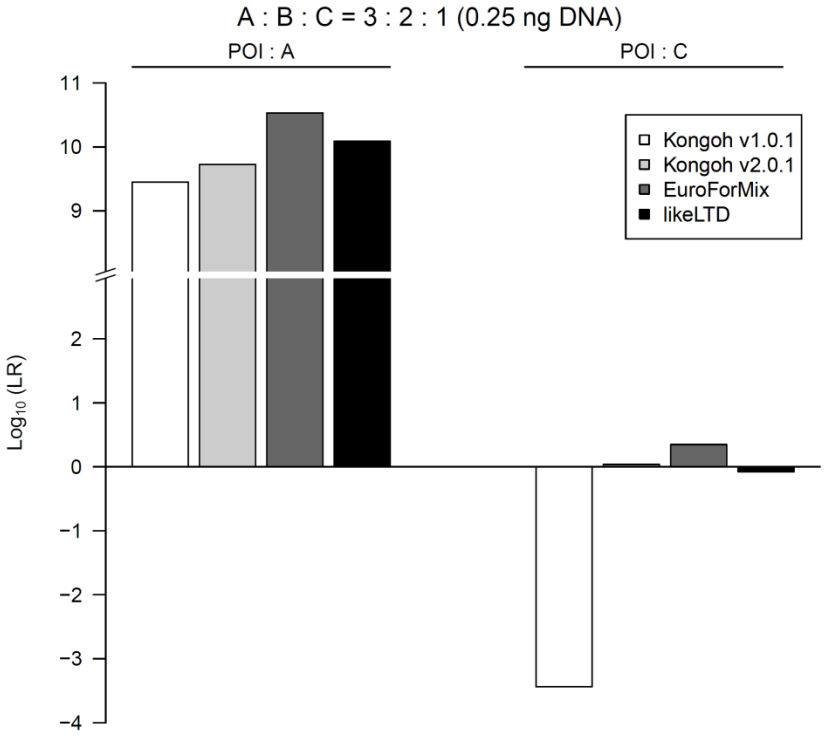

Fig. 13. LR values of each version of *Kongoh* and two more software in a 3:2:1 mixture.

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] National Research Council (NRC) Committee on DNA Forensic Science. The Evaluation of Forensic DNA evidence, Washington, DC: National Academy Press. (1996).

[3] 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.

[4] S. Manabe, Y. Hamano, C. Kawai, C. Morimoto, K. Tamaki, Development of new peak-height models for a continuous method of mixture interpretation, Forensic Sci Int Genet Suppl Ser. 5 (2015) e104-e106.

[5] Ø. Bleka, G. Storvik, P. Gill, EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci Int Genet. 21 (2016) 35-44.

[6] C.D. Steele, M. Greenhalgh, D.J. Balding, Evaluation of low-template DNA profiles using peak heights, Stat Appl Genet Mol Biol. 15 (2016) 431-445.