PARFEX

Parents Finder in EXCEL™
Software Package for Parentage Allocation
Running in Microsoft® EXCEL™

User Guide

Masashi Sekino & Shigeho Kakehi

Tohoku National Fisheries Research Institute, Fisheries Research Agency of Japan
3-27-5 Shin-hama, Shiogama, Miyagi 985-0001, Japan

Contents

1. About PARFEX

1-1. What's PARFEX?

PARFEX (acronym of parents finder in EXCEL; current version is 1.0) is a software package for molecular parentage analysis. Many excellent computer programs for parentage allocation have been released (a recent list of software is available in Jones et al. 2010). The most favorable property of PARFEX is that it runs in Microsoft® EXCEL™ application. The well-known EXCEL is commonly used for genotype data storage. Since PARFEX consists of EXCEL macro programs (macros) written in VBA (Visual Basics for Applications) language, parentage testing proceeds in EXCEL after data transfer to a worksheet of PARFEX-bundled EXCEL workbook in a convenient copy-and-paste manner. Results are provided in other spreadsheets automatically created in the workbook. Thus, a series of parentage testing including the summarization of results completes in EXCEL.

1-2. Methods available in PARFEX

Among several methods of parentage allocation (reviewed in Jones & Ardren 2003; Jones et al. 2010), PARFEX performs exclusion and categorical likelihood methods (see section 2). In addition, PARFEX is furnished with several accessory macros, which we believe are useful for the analysis of parentage.

1-3. Types of DNA markers

PARFEX handles autosomal genotypes of allogamous diploid organisms. Co-dominant microsatellites and/or single nucleotide polymorphisms (SNPs) are allowed as DNA markers. For the likelihood-based method, markers should meet Hardy-Weinberg equilibrium (HWE) and gametic phase equilibrium (or no physical linkage).

1-4. Computing system

PARFEX works in the recent versions of EXCEL (ver. 2003 or later) on Windows XP platform. However, there are minor version-to-version differences in the way to launch macros. Please refer to version-specific EXCEL instruction. If user feels that PARFEX does not run properly, please email us. Upon request, we may recompile the source code so that PARFEX runs in EXCEL Macintosh.

1-5. Notes on computational speed

Overall, the computational speed of PARFEX is slow. It must be frustrating especially in simulation analyses. Although we will strive to resolve this problem, we would like to ask users to have patience for now. PARFEX may crash when another EXCEL workbook is running owing to some conflict between active workbooks. Thus, other EXCEL workbooks should be closed before running PARFEX.

1-6. Citation

MS designed PARFEX from the view of population genetics and the computer script was written by SK. We distribute PARFEX as free-share software; however, the copyright should be attributed to the authors. When users publish papers containing results obtained with PARFEX, please cite this:

Sekino M, Kakehi S (2012) PARFEX v1.0: an EXCEL™-based software package for parentage allocation. *Conservation Genetics Resources* 4:275–278

1-7. Contact information

PARFEX is built in an EXCEL workbook. It can be downloaded from http://cse.fra.affrc.go.jp/sekino/PARFEX/ anytime. We have detected so far no functional defect in the current version. In case that user finds some aberrant behavior of PARFEX, however, please let us know (email: MS, sekino@affrc.go.jp; SK, kakehi@affrc.go.jp). We would highly appreciate receiving requests and comments on PARFEX from users.

BE USED ON USERS' OWN RESPONSIBILITY. WE ACCEPT NO LIABILITY WHATSOEVER FOR ANY CLAIM RELATED TO THE USE OF THIS SOFTWARE.

2. Parentage testing in PARFEX

2-1. Exclusion

The exclusion method examines genotype incompatibilities between offspring and putative parents based on the rules of Mendelian inheritance (e.g., O'Reiily et al. 1998). Parent-offspring hypotheses are rejected when putative parents and offspring show genotype incompatibility at one or more markers. A robust parentage relationship is established if a single parent (or single parental pair) of offspring remains non-excluded from a parental pool (Fig. 2-1). The exclusion method can be used for populations to which classical population genetics assumptions do not hold (e.g., non-random mating). This

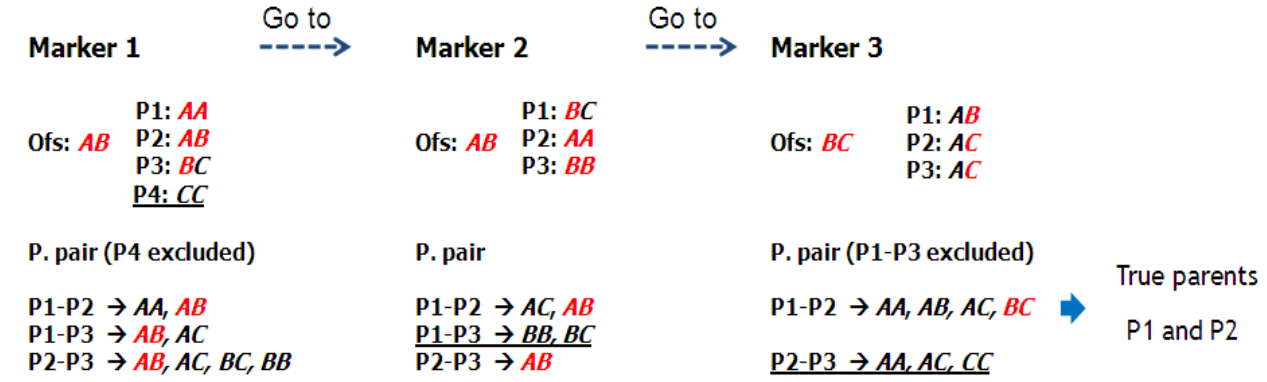


Fig. 2-1. Schematic representation of exclusion method (parental pair allocation)

Ofs: offspring; P1, P2, P3 and P4: all putative parents (non-sexed). Individual genotypes are given (e.g., *AB*: heterozygote of alleles *A* and *B*). Offspring genotypes expected from possible parental combinations are provided in 'P. pair'. Shared alleles between Ofs and putative parents as well as expected offspring genotypes (in P. pair) compatible with Ofs' are indicated in red. At Marker 1, P4 (underlined) should be excluded as she/he does not share any allele with Ofs. Next, parental pairs P1-P3 and P2-P3 should be rejected as neither of them should produce the genotype compatible with Ofs' at Marker 2 and 3, respectively. Consequently, a parental pair of P1 and P2 remains non-excluded; P1 and P2 are judged as the true parents of Ofs.

method, however, has several shortcomings (see Jones & Ardren 2003). For example, it may be required to screen a prohibitive number of markers to establish robust parentages for samples with a large number of candidate parents. Another problem is the presence of genotypic mismatches caused by human errors, PCR errors (e.g., microsat null alleles, see **Note 1)** and mutations, resulting in a false rejection of true parent-offspring hypotheses. PARFEX is designed to accommodate some genotypic mismatches (mismatched markers) in parent-offspring lines so as to deal with the latter issue.

2-2. Categorical likelihood-based method

The likelihood-based method available in PARFEX adopts the concept of Gerber et al (2000; 2003: FaMoz software) with modifications. The parentage inference relies on the difference in log-likelihood ratio (LOD) between related and unrelated relationships (Meagher and Thompson 1986; see glossaries). The strength of likelihood-based methods is that the quality of parentage allocation can be evaluated through simulations based on some population genetics assumptions. In addition, a true parent-offspring line may be identified according to LOD scores even if multiple putative parents remain non-excluded (Marshall et al. 1998; Jones et al. 2010). Furthermore, the method used in PARFEX requires a few assumptions therefore being easily understandable.

Both '*single parent search*' and '*parental pair search*' are available in PARFEX. The analysis consists of **1)** simulations to define a threshold LOD (LOD_C) to accept/reject possible parentage relationships and **2)** application of LOD_C to real genotype data to reconstruct parentages (Fig. 2-2):

1a) A pool of parental genotypes (POP_{m+f}; equal sex ratio) is produced according to allele frequencies obtained from real collected genotypes (or those provided by users). Offspring pool (POP_{ofs}) is created from POP_{m+f} assuming random mating between sexes. Random sampling of alleles from the allele frequency data yields a pool of random individuals (POP_{rand}) and they are assumed to have no parent in POP_{m+f}. Therefore, two alternative hypotheses are considered: individuals are related (POP_{m+f} *vs* POP_{ofs}) or unrelated (POP_{m+f} *vs* POP_{rand}). Genotypic errors are generated at a

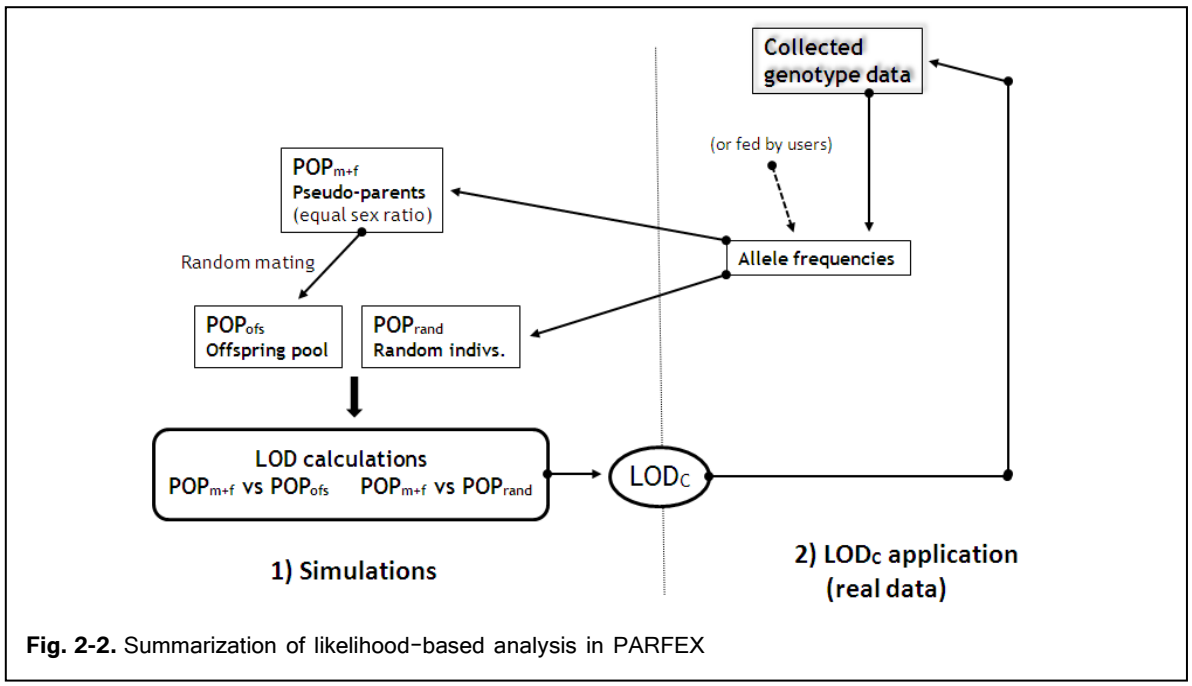


Fig. 2-2. Summarization of likelihood-based analysis in PARFEX

certain rate (e_{sim}) in the simulated genotypes following the random genotype replacement model (Marshall et al. 1998).

1b) LOD scores are calculated for POP_{m+f} *vs* POP_{ofs} and POP_{m+f} *vs* POP_{rand}. The way of calculating LOD score is different between *single parent* and *parental pair searches* (see glossaries). In *single parent search*, the first and second highest LOD scores are extracted for each member of POP_{ofs} and POP_{rand}. A putative parent who gives the first highest (or second highest) LOD for an offspring (POP_{ofs}) is the most likely parent of the offspring. In *parental pair search*, only the first highest LOD score is taken. A putative parental pair which gives the highest score for an offspring is the most likely parental pair of the offspring. For each offspring, the identity between true parent (or parental pair) and the most likely parent (parental pair) is checked.

1c) The extracted LOD scores are denoted here as L_{ofs} (POP_{ofs}) and L_{rand} (POP_{rand}). The L_{ofs} distribution is defined as the LOD distribution under the null hypothesis (H_0) that an individual has true parents in a population sample. The L_{rand} distribution represents the LOD distribution under the alternative hypothesis (no parent in the population sample). A threshold LOD (LOD_C) is determined according to the two LOD distributions (see later). In addition, type I and type II errors (type I, α: falsely rejecting H_0; type II, β: falsely accepting H_0) conditional upon the value of LOD_C are estimated.

1d) For simulated samples, parent-offspring hypotheses are examined based on LOD_C: pairs between putative parent (parental pair) and offspring having a LOD score smaller than the value of LOD_C are rejected. Based on the results, success rate of parentage allocation is estimated.

2) LOD scores between real genotypes of putative parents and offspring are calculated. Parentages are determined in the same way as described above (1d). The quality of parentage allocation can be measured by α, β and the success rate of parentage allocation obtained in the preceding simulations.

3. Macros bundled in PARFEX

PARFEX contains several macros. First, open the EXCEL workbook named 'PARFEX_v1_0.xlsm' downloaded from the PARFEX website (the workbook can be copied freely). EXCEL 2003 users better use 'PARFEX_v1_0_E03.xls' file. In case EXCEL gives a warning (Fig. 3-1), click 'Enable Content'. Go to '***View***' > '***Macros***' > '***View Macros***' tabs, or press keys '**Alt + F8**'. A window listing nine macros appears (Fig. 3-2). The intended purpose of each macro is as follows:

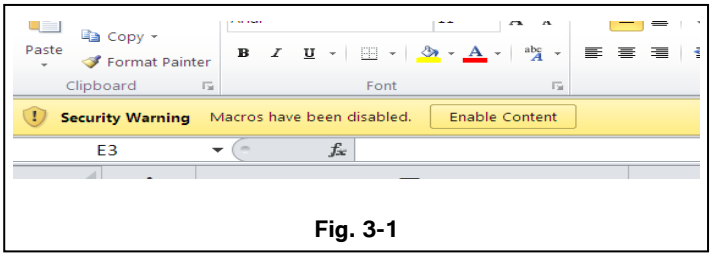
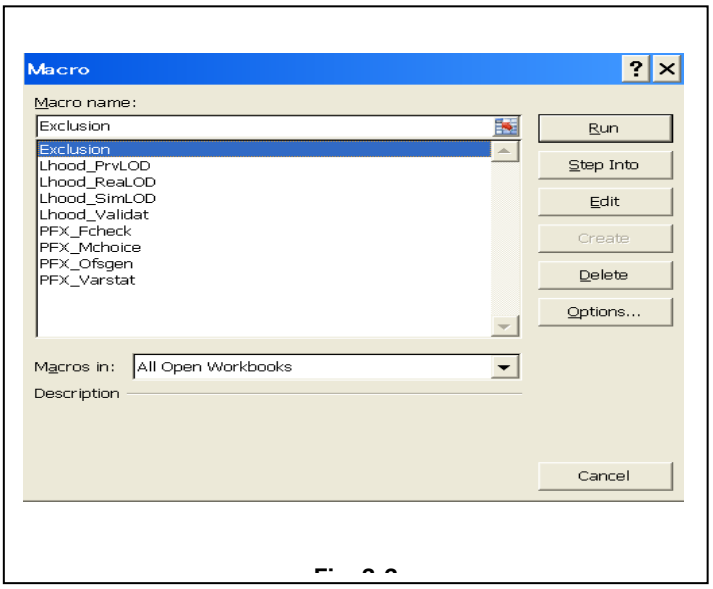


Fig. 3-1



►***Exclusion***: Implementation of exclusion method

►***Lhood_PrvLOD***:
►***Lhood_ReaLOD***:
►***Lhood_SimLOD***:
►***Lhood_Validat***:
→ Implementation of likelihood-based method

►***PFX_Fcheck***: Format error check for collected genotype data

►***PFX_Mchoice***: Identification of a set of markers for a high exclusion success

►***PFX_Ofsgen***: Generation of simulated offspring and random individuals

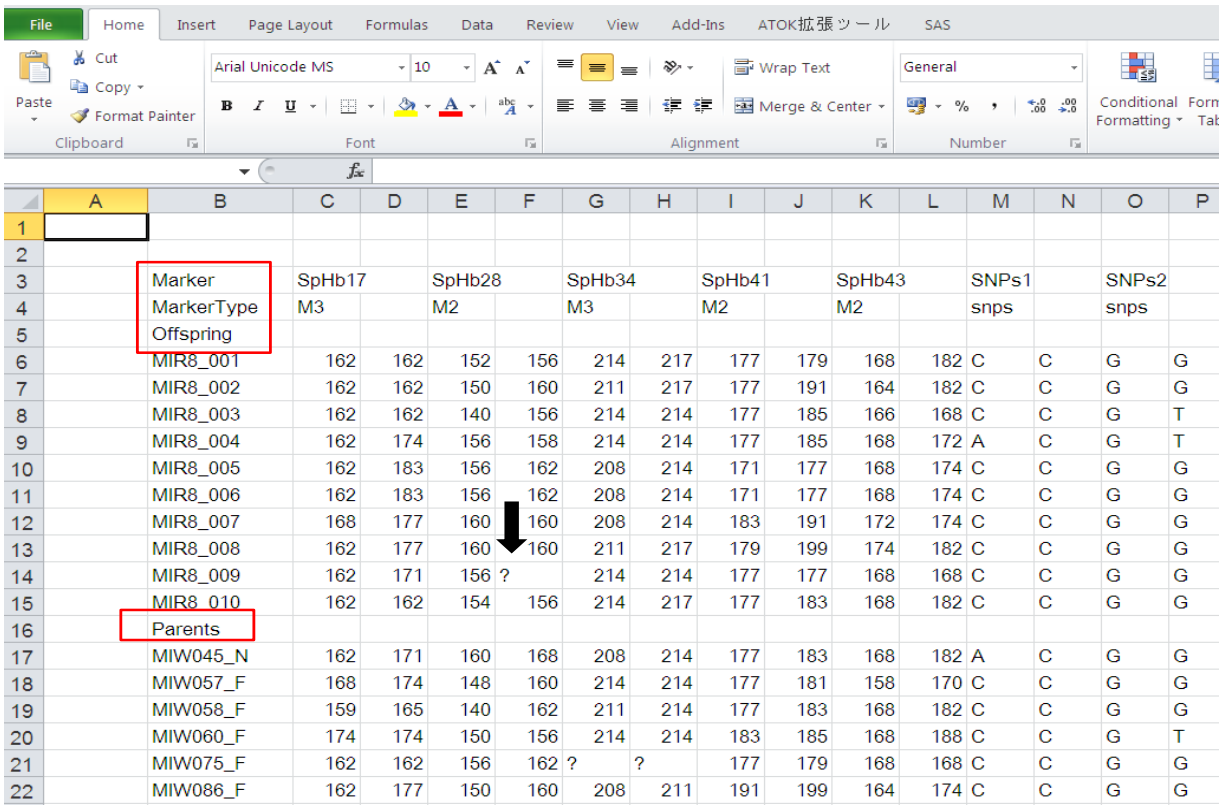
►***PFX_Varstat***: Calculation of polymorphisms statistics and allele frequencies

The macro ***Exclusion*** is used for the analysis of exclusion. Macros with the suffix '***Lhood***_' are for the likelihood method. Macros with the suffix '***PFX***_' are rather supplementary, but ***PFX_Varstat*** is necessary to calculate population allele frequencies used in the likelihood-based method.

4. Data preparation

4-1. Genotype data format

Genotype data should be prepared in a worksheet of PARFEX-bundled EXCEL workbook (Fig. 4-1). The worksheet can be named arbitrarily. We call it '**Data Genotype**' sheet throughout this documentation. Hereafter we use genotype data from a flatfish, the spotted halibut *Verasper variegatus*, which is distributed in the northwestern Pacific.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2																
3		Marker	SpHb17		SpHb28		SpHb34		SpHb41		SpHb43		SNPs1		SNPs2	
4		MarkerType	M3		M2		M3		M2		M2		snps		snps	
5		Offspring														
6		MIR8_001	162	162	152	156	214	217	177	179	168	182	C	C	G	G
7		MIR8_002	162	162	150	160	211	217	177	191	164	182	C	C	G	G
8		MIR8_003	162	162	140	156	214	214	177	185	166	168	C	C	G	T
9		MIR8_004	162	174	156	158	214	214	177	185	168	172	A	C	G	T
10		MIR8_005	162	183	156	162	208	214	171	177	168	174	C	C	G	G
11		MIR8_006	162	183	156	162	208	214	171	177	168	174	C	C	G	G
12		MIR8_007	168	177	160	160	208	214	183	191	172	174	C	C	G	G
13		MIR8_008	162	177	160	160	211	217	179	199	174	182	C	C	G	G
14		MIR8_009	162	171	156	?	214	214	177	177	168	168	C	C	G	G
15		MIR8_010	162	162	154	156	214	217	177	183	168	182	C	C	G	G
16		Parents														
17		MIW045_N	162	171	160	168	208	214	177	183	168	182	A	C	G	G
18		MIW057_F	168	174	148	160	214	214	177	181	158	170	C	C	G	G
19		MIW058_F	159	165	140	162	211	214	177	183	168	182	C	C	G	G
20		MIW060_F	174	174	150	156	214	214	183	185	168	188	C	C	G	T
21		MIW075_F	162	162	156	162	?	?	177	179	168	168	C	C	G	G
22		MIW086_F	162	177	150	160	208	211	191	199	164	174	C	C	G	G

Fig. 4-1. Genotype data sheet of PARFEX

The string of characters enclosed in red boxes (Fig. 4-1) should **NEVER** be changed (Marker, MarkerType, Offspring and Parents; **case-sensitive**). Nor is the order from upper to lower. No blank line within data is allowed. The characters serve as tokens letting PARFEX recognize your data.

- **Marker:** Marker names (here SpHb17 etc.). There is no limitation of the length of marker names (the same applies to offspring/parental ID). The right-adjacent cell of each marker name should be voided. Up to 120 markers can be used.
- **MarkerType:** For each microsat, place 'M' followed by the length of repeat-unit. For example, it should be 'M2' for dinucleotide-repeat microsats. This information is used in the subsequent data-format check. For SNPs, place 'snps' (case-*in*sensitive).

- **Offspring:** It declares the beginning of offspring genotype data from the next line. Genotype data should be preceded by offspring ID (here MIR8_001 etc.). Each offspring has one line of genotype data. One marker uses two cells each of which is occupied by one allele. Alleles should be provided in fragment size (bp) for microsats and four letters representing four nucleotides (A, C, G or T) for SNPs. Missing alleles should be denoted by '?'. Genotypes with one missing allele are allowed (arrow in Fig. 4-1) only in the exclusion method (macro ***Exclusion***) and generation of simulated individuals (***PFX_Ofsgen***). In other analyses, they are treated as missing genotypes. Up to 5×10^3 offspring are allowed.
- **Parents:** Parental genotype data should be prepared in the same way as described above. However, parental ID should end with the suffix '**_F**' for female and '**_M**' for male. Individuals with no suffix or the suffix '**_N**' are treated as non-sexed (Fig. 4-1). The maximum number of parental individuals is 5×10^3 for each sex, but the capacity reduces to a total of 5×10^3 if all are non-sexed. When sexed and non-sexed individuals are mixed, the maximum number follows this rule: $F + N \leq 5 \times 10^3$ **AND** $M + N \leq 5 \times 10^3$ (F: the number of females; M: males; N: non-sexed).

A simple way to prepare PARFEX genotype data is to copy genotypes stored in EXCEL as CONVERT format (data-conversion software; Glaubitz 2004) and paste (insert) them into the genotype data space (see **Note 2**).

Up to 20 lines above the token 'Marker' and/or 20 columns of the left side of the data can be used to put some comments, but duplication of any of the four tokens is not allowed.

4-2. Data format check (macro *PFX_Fcheck*)

Once 'Data Genotype' sheet is created, the data format should be checked. The macro ***PFX_Fcheck*** does it.

§ Show 'Data Genotype' sheet
§ Open the macro window, select ***PFX_Fcheck*** (Fig. 4-2) and click 'Run'.

First, ***PFX_Fcheck*** inspects if the four tokens as well as the marker information (name and type) are properly arranged. When there is something wrong, it gives a warning message. Next, it examines whether the genotype data contains such errors as **(1)** blank

cell, **(2)** duplicated individual ID, **(3)** anomalously short or long microsat allele-size (valid range: 50-400 bp), **(4)** microsat allele-size not explained by repeat-unit iterations and **(5)** SNPs alleles denoted by letters other than A, C, G or T. PARFEX does not run when any of the errors **(1)**, **(2)** and **(5)** is found, but tolerates potential errors **(3)** and **(4)**. Here is an example of error detection **(4)** (see **Note 3**):

- A warning message is given when an aberration is found (arrow in Fig. 4-3). It asks whether the potential error should be ignored.
- Click 'Yes ' if it is certain that the allele size is correct: the anomaly is ignored.
- Click 'No' if it is a true error: you are redirected to 'Data Genotype' sheet for correction.

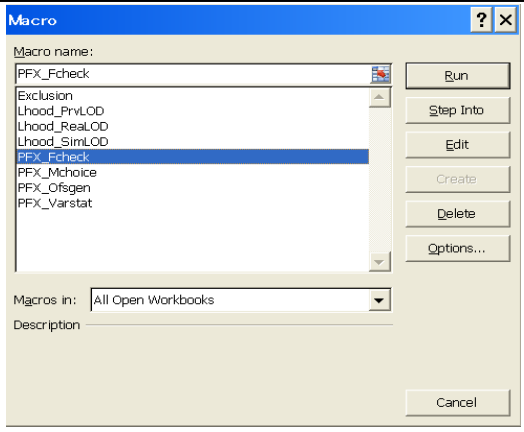


Fig. 4-2

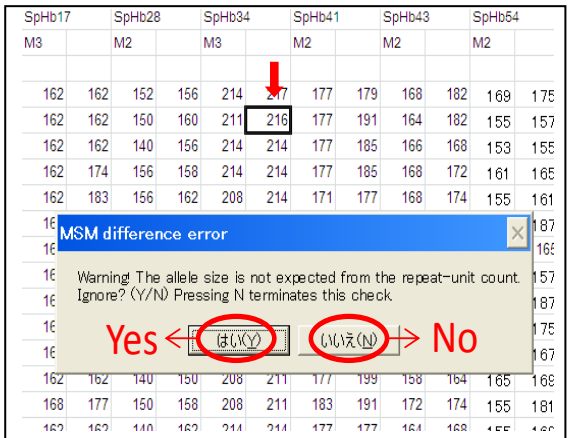


Fig. 4-3

5. PARFEX analyses

In PARFEX, each macro plays a particular role thereby producing a result sheet having a consistent macro-specific name. When reanalysis is done using the same macro in the same workbook (e.g., with different parameters), the existing results are replaced by new results in the same result sheet. This occurs because EXCEL does not allow the presence of two or more spreadsheets having an identical name in a workbook. Before reanalysis, therefore, the result sheet in which previous results are stored should be renamed. We should also note that several text files are created through PARFEX analyses, which **MUST** be kept until all the analyses complete. It would be better to use PARFEX in a specific folder in order to avoid scattering of text files.

5-1. Summary statistics of polymorphisms (macro *PFX_Varstat*)

PFX_Varstat calculates several statistics of polymorphisms and allele frequencies based on collected genotype data. It also performs an exact test for HWE. User can omit HWE analysis as the method used in PARFEX takes much time for computation for a large sample size.

§ Show 'Data Genotype' sheet.
§ Open the macro window, select and run ***PFX_Varstat***.
§ Decide whether or not HWE testing should be performed.
§ Computation status is given in the right cell of the token 'Offspring' in 'Data Genotype' sheet.

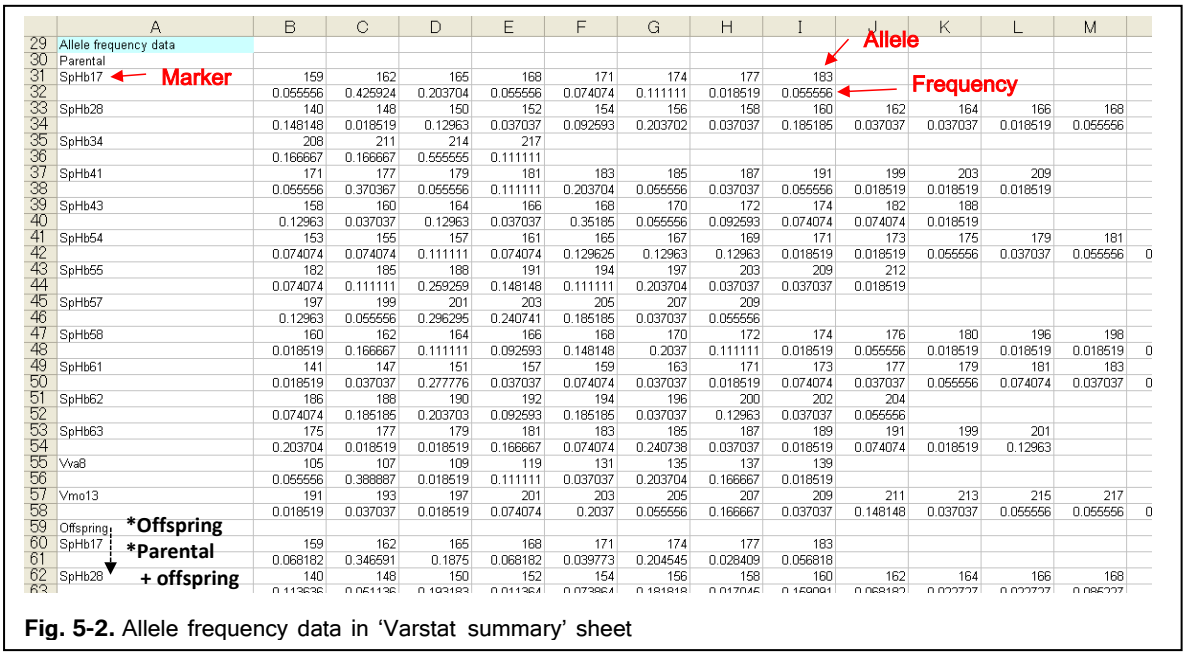
	A	B	C	D	E	F	G	H	I	J	K	L
1	Parental data											
2		SpHb17	SpHb28	SpHb34	SpHb41	SpHb43	SpHb54	SpHb55	SpHb57	SpHb58	SpHb61	SpHb6
3	Sample Size	27	27	27	27	27	27	27	27	27	27	
4	Aobs	8	12	4	11	10	17	9	7	13	17	
5	Auni	1	2	0	3	1	7	1	0	6	3	
6	Hobs	0.667	0.963	0.593	0.926	0.889	0.926	0.852	0.815	0.926	0.926	0.6
7	Hexp	0.764	0.884	0.635	0.809	0.832	0.928	0.852	0.811	0.887	0.901	0.8
8	PIC	0.722	0.854	0.579	0.773	0.799	0.904	0.816	0.766	0.857	0.877	0.8
9	Excl PP(Paternity)	0.555	0.736	0.385	0.623	0.659	0.82	0.676	0.602	0.74	0.781	0.7
0	Excl P1(One parent missing)	0.372	0.58	0.215	0.445	0.486	0.694	0.506	0.424	0.585	0.64	0.5
1	Excl P2(Both parents)	0.754	0.895	0.567	0.815	0.846	0.947	0.85	0.785	0.898	0.932	0.8
2	HWE P	0.014	0.886	0.482	0.187	0.952	0.816	0.193	0.112	0.022	0.723	0.0
3	HWE P_SE	0.00098	0.00243	0.00337	0.00326	0.00158	0.00312	0.00324	0.00234	0.00101	0.0034	0.001
4												
5	Offspring data											
6		SpHb17	SpHb28	SpHb34	SpHb41	SpHb43	SpHb54	SpHb55	SpHb57	SpHb58	SpHb61	SpHb6
7	Sample Size	100	100	100	100	100	100	100	100	100	100	1
8	Aobs	8	12	4	9	10	13	8	7	11	15	
9	Auni	1	0	0	0	0	1	0	0	1	2	
20	Hobs	0.77	0.83	0.65	0.83	0.92	0.95	0.7	0.74	0.89	0.91	0.
21	Hexp	0.722	0.811	0.656	0.807	0.826	0.882	0.689	0.733	0.829	0.849	0.8
22	PIC	0.692	0.782	0.6	0.783	0.8	0.866	0.658	0.685	0.803	0.829	0.8

Fig. 5-1. Results of *PFX_Varstat* analysis ('Varstat_summary' sheet)

Results are presented in '**Varstat_summary**' sheet (Fig. 5-1). The following indices are calculated for parental population: $\boldsymbol{A}_{\mathbf{obs}}$, number of different alleles; $\boldsymbol{A}_{\mathbf{uni}}$, number of unique alleles (alleles observed in just a single individual); $\boldsymbol{H}_{\mathbf{obs}}$ and $\boldsymbol{H}_{\mathbf{exp}}$, observed and unbiased estimate of expected heterozygosity; **PIC**, polymorphism information content; **ExclPP**, **ExclP1** and **ExclP2**, three types of exclusion probability; **HWE P**, exact *P* value of HWE analysis and its standard error (**HWE P_SE**) (17×10^3 Monte Carlo randomizations). Result space for HWE testing is left blank if user selected the 'No' option for this analysis.

The same statistics are calculated for offspring. However, there may be a case in which just one offspring is to be analyzed in the subsequent parentage testing. In such a case, the macro omits calculations for offspring and the result space for offspring is left empty.

Allele frequency data begins from Line 30 of 'Varstat_summary' sheet with the following order: parental, offspring and combined data (parental + offspring) (Fig. 5-2).



	A	B	C	D	E	F	G	H	I	J	K	L	M
29	Allele frequency data												
30	Parental												
31	SpHb17	159	162	165	168	171	174	177	183				
32		0.055556	0.425924	0.203704	0.055556	0.074074	0.111111	0.018519	0.055556				
33	SpHb28	140	148	150	152	154	156	158	160	162	164	166	168
34		0.148148	0.018519	0.12963	0.037037	0.092593	0.203702	0.037037	0.185185	0.037037	0.037037	0.018519	0.055556
35	SpHb34	208	211	214	217								
36		0.166667	0.166667	0.555555	0.111111								
37	SpHb41	171	177	179	181	183	185	187	191	199	203	209	
38		0.055556	0.370367	0.055556	0.111111	0.203704	0.055556	0.037037	0.055556	0.018519	0.018519	0.018519	
39	SpHb43	158	160	164	166	168	170	172	174	182	188		
40		0.12963	0.037037	0.12963	0.037037	0.35185	0.055556	0.092593	0.074074	0.074074	0.018519		
41	SpHb54	153	155	157	161	165	167	169	171	173	175	179	181
42		0.074074	0.074074	0.111111	0.074074	0.129625	0.12963	0.12963	0.018519	0.018519	0.055556	0.037037	0.055556
43	SpHb55	182	185	188	191	194	197	203	209	212			
44		0.074074	0.111111	0.259259	0.148148	0.111111	0.203704	0.037037	0.037037	0.018519			
45	SpHb57	197	199	201	203	205	207	209					
46		0.12963	0.055556	0.296295	0.240741	0.185185	0.037037	0.055556					
47	SpHb58	160	162	164	166	168	170	172	174	176	180	196	198
48		0.018519	0.166667	0.111111	0.092593	0.148148	0.2037	0.111111	0.018519	0.055556	0.018519	0.018519	0.018519
49	SpHb61	141	147	151	157	159	163	171	173	177	179	181	183
50		0.018519	0.037037	0.277776	0.037037	0.074074	0.037037	0.018519	0.074074	0.037037	0.055556	0.074074	0.037037
51	SpHb62	186	188	190	192	194	196	200	202	204			
52		0.074074	0.185185	0.203703	0.092593	0.185185	0.037037	0.12963	0.037037	0.055556			
53	SpHb63	175	177	179	181	183	185	187	189	191	199	201	
54		0.203704	0.018519	0.018519	0.166667	0.074074	0.240738	0.037037	0.018519	0.074074	0.018519	0.12963	
55	Vva8	105	107	109	119	131	135	137	139				
56		0.055556	0.388887	0.018519	0.111111	0.037037	0.203704	0.166667	0.018519				
57	Vmo13	191	193	197	201	203	205	207	209	211	213	215	217
58		0.018519	0.037037	0.018519	0.074074	0.2037	0.055556	0.166667	0.037037	0.148148	0.037037	0.055556	0.055556
59	Offspring												
60	SpHb17	159	162	165	168	171	174	177	183				
61		0.068182	0.346591	0.1875	0.068182	0.039773	0.204545	0.028409	0.056818				
62	SpHb28	140	148	150	152	154	156	158	160	162	164	166	168

Fig. 5-2. Allele frequency data in 'Varstat summary' sheet

5-2. Exclusion method

5-2-1. Marker selection (macro *PFX_Mchoice*; optional)

For closed captive-bred populations with known parental genotypes, *a priori* knowledge about a minimum set of markers which provides a high resolution of parentage allocation helps reduce the experimental cost and labor involved in the subsequent parentage testing. ***PFX_Mchoice*** proposes such a marker set through simulations:

1) Simulated offspring genotypes are generated from collected parental genotypes (random mating without selfing) (see **Note 4**).

2) Markers are ranked according to the extent of polymorphisms.

3) Simulated offspring and real parents are subjected to exclusion-based parentage testing based on the highest-ranked marker.

4) Parentage testing is continued with successive one-by-one addition of higher-ranked markers, from which the cumulative success rate of parentage allocation is obtained.

In ***PFX_Mchoice***, the success rate of parentage allocation is defined as the number of simulated offspring whose true parental pair is unambiguously identified divided by the total number of offspring.

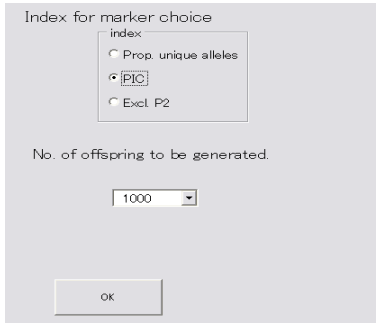


Fig. 5-3

§ Show 'Data Genotype' sheet.

§ Open the macro window, select and run ***PFX_Mchoice***.

§ Parameter setting window appears (Fig. 5-3).

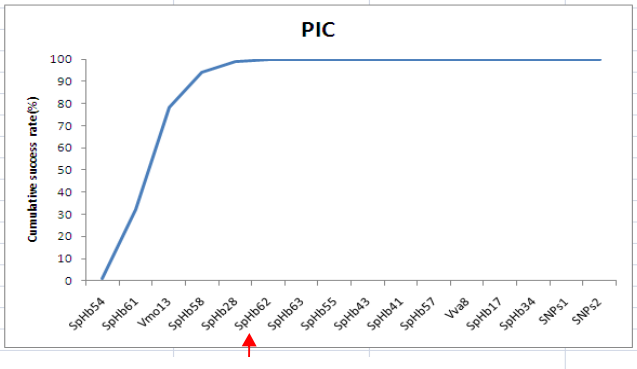
§ Select one of three statistics to rank markers: proportion of unique alleles (A_{uni}/A_{obs}: see **Note 5**), PIC and ExclP2.

§ Default number of simulated offspring is 10^3. Other numbers are available (100, 500, 2×10^3 and 5×10^3).

§ If there are some missing alleles ('?') in parental genotypes, one of the following options should be selected: 1) individual with missing data is removed or 2) missing allele is replaced by another allele (**Note 4**). In the case 1), the removed individual is not used in exclusion analysis. ***PFX_Mchoice*** does not ask so if the data has no missing allele.

Results are provided in '**Mchoice_summary**' sheet (Fig. 5-4). Here is an example of 27 putative parents (16 markers; index, PIC; 10^3 offspring). In 'Mchoice_summary' sheet, the cumulative success rate is plotted in a graph and the numerical details are given in the

	A	B	C	D	E
9	Index	Rank (High-Low)	Marker	Value	Success rate (single parental pair assigned, %)
20	PIC				
21		1	SpHb54	0.904	0.8
22		2	SpHb61	0.877	32.1
23		3	Vmo13	0.875	78.1
24		4	SpHb58	0.857	94.1
25		5	SpHb28	0.854	98.8
26		6	SpHb62	0.836	100
27		7	SpHb63	0.824	100
28		8	SpHb55	0.816	100
29		9	SpHb43	0.799	100
30		10	SpHb41	0.773	100
31		11	SpHb57	0.766	100
32		12	Vva8	0.731	100
33		13	SpHb17	0.722	100
34		14	SpHb34	0.579	100
35		15	SNPs1	0.221	100
36		16	SNPs2	0.2	100
37					
38	Excl. P2				
39		1	SpHb54	0.947	NA
40		2	SpHb61	0.932	NA
41		3	Vmo13	0.921	NA
42		4	SpHb58	0.898	NA
		5	SpHb28	0.895	NA



100% success rate achieved

NA: Not analyzed

Fig. 5-4 Results of *PFX_Mchoice* analysis ('Mchoice_summary' sheet)

left cells. In this case, we expect that the use of six markers will achieve a 100% success rate of parentage allocation. For a reference, ***PFX_Mchoice*** outputs the values of the other indices as well as the marker-ranking based on the indices.

5-2-2. Parentage allocation (macro *Exclusion*)

The macro ***Exclusion*** essentially searches for the parental pair of offspring. When no parental pair is found, it resorts to *single parent search* based on allele sharing between putative parents and offspring. Parental pairs within sexes are ruled out beforehand if all or some parents are sexed.

♣ **Mismatched marker**: An important parameter is 'Max Mismatch' ($MaxN_{MM}$). The brevity code N_{MM} stands for the number of mismatched markers and the $MaxN_{MM}$ is the maximum number of mismatched markers allowed by user. Let a $MaxN_{MM}$ be set at two. ***Exclusion*** performs parentage testing at N_{MM} of zero (strict exclusion: no mismatch is allowed across markers), one and two. A $MaxN_{MM}$ of zero means that the analysis to be done is strict exclusion. In the result sheet, probable parental pairs (or single parents) of offspring at each N_{MM} are shown with mismatched marker names.

♣ **Missing data**: Basically, markers with missing allele(s) are ignored and not counted in N_{MM}. They are shown in specified color in the result sheet independent of N_{MM} values (Fig. 5-6). An exceptional case is when one of the two alleles at a marker is missing (e.g., parental genotype: 156/?) and a parent-offspring relationship is reconstructed by exploiting the information of the scored allele (allele 156). In such a case, the marker name never appears in the result sheet despite the unavailability of one allele.

♣ **Null-allele segregation**: ***Exclusion*** asks if possible segregation of microsat null alleles (Note 1) should be tested. For the simple logic behind the test, see Note 6. This test is applicable to limited cases. Please click 'No' if user considers that this test is not useful.

§ Show 'Data Genotype' sheet.

§ Open the macro window, select and run ***Exclusion***.

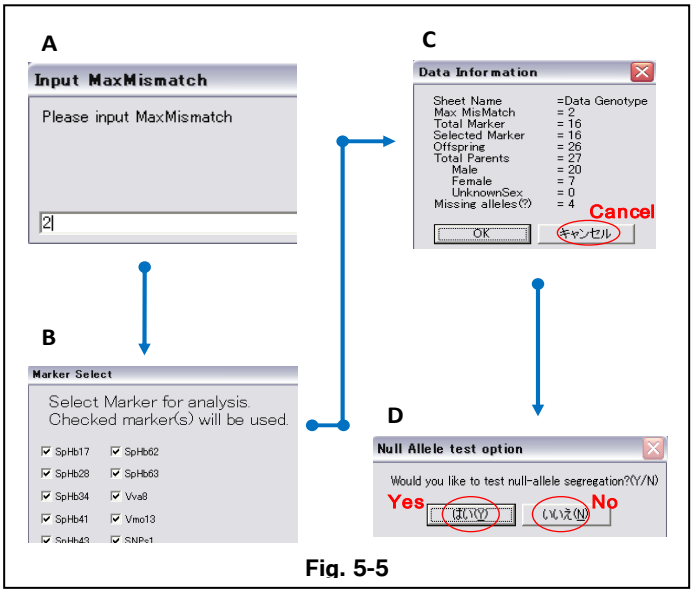


Fig. 5-5

§ Set MaxN_{MM} (Fig. 5-5A).

§ Select markers for analysis (Fig. 5-5B). Data confirmation window appears (Fig. 5-5C).

§ Notify null-allele segregation should be tested (Fig. 5-5D).

Results are given in '**Exclusion_summary**' sheet. An example is shown in Figure 5-6 (16 markers; MaxN_{MM} = 2; null-allele segregation test turned off).

• **Line 1** lists used markers; **Line 2** tells that markers with missing alleles were ignored (but see above); **Lines 3–6** show the categorization of markers (see below); **Lines 9–X** give the parentage results: column A, offspring ID; column B, N_{MM}; columns C and D; probable parent(s) identified at given N_{MM} (column C: male or non-sexed; col. D: female or non-sexed); from column E rightward, mismatched marker names (or markers with missing alleles).

Marker names are colored in order to deliver the following messages:

Blue: markers with missing allele(s) in offspring genotype.

Gray: markers with missing allele(s) in parental genotype.

Red: mismatched markers with suspected null-allele segregation (see **Note 6**).

Red: markers with mismatch caused by other than possible null-allele segregation.

The criterion of N_{MM} does not refer to the 'maximum' number of mismatched markers allowed by user and parentage relationships established at $N_{MM} = x$ are exclusive of those at $N_{MM} < x$. For example, when a MaxN_{MM} = 1 is set and a parental pair of offspring is identified at N_{MM} = 0, the parental pair is not shown in the result space for N_{MM} = 1.

In addition, as can be seen in Figure 5-6, for the large majority of offspring no result

	A	B	C	D	E	F
1	Markers employed (16/16):	SpHb17	SpHb28	SpHb34	SpHb41	SpHb43
2	Markers with missing data ('?') are NOT counted in the defined number of mismatch markers.					
3	Types of mismatch:	*Red*	=Other than null-allele segregation and missing alleles			
4		*Red*	=Suspected null-allele segregation			
5		*Blue*	=Missing allele in offspring			
6		*gray*	=Missing allele in parent			
7						
8	Offspring	No. Mismatch Marker	Parentage (Male or unknown)	Parentage (Female or unknown)	Incompatible Markers	
9	MIR8_001	0	MIW074_M	MIW075_F		
10		2	MIW014_M		*SpHb54*	*Vva8*
11	MIR8_002	0	MIW021_M	MIW086_F		
12	MIR8_021	0	MIW102_M	MIW057_F		
13	MIR8_022	0	MIW018_M	MIW086_F		
14	MIR8_023	0	MIW033_M	MIW075_F		
15	MIR8_024	0	MIW014_M	MIW060_F		
16	MIR8_025	0	MIW073_M	MIW086_F		
17	MIR8_026	0	MIW096_M	MIW075_F	*SNPs1*	
18	MIR8_042	0	MIW102_M	MIW057_F		
19	MIR8_043	0	MIW102_M	MIW057_F		
20	MIR8_044	0	MIW014_M	MIW060_F		
21	MIR8_082	0	MIW018_M	MIW045_F	*SpHb17*	
22	MIR8_083	0	MIW014_M	MIW060_F		
23	MIR8_084	0	MIW021_M	MIW086_F		
24		1		MIW075_F	*SpHb61*	
25	MIR8_085	0	MIW013_M	MIW086_F		
26	MIR8_086	0	MIW102_M	MIW057_F		
27	MIR8_087	0	MIW062_M	MIW075_F		
	MIR8_088	0	MIW073_M	MIW086_F		

Offspring ID | **N_{MM}** | **Parent** ♂ or non-sexed | **Parent** ♀ or non-sexed | **Mismatched marker**

Fig. 5-6. Results of *Exclusion* analysis ('Exclusion_summary' sheet)

is shown for N_{MM} = 1 and N_{MM} = 2 despite that a $MaxN_{MM}$ = 2 was set in this example. This means that that parentage testing was done at both N_{MM} = 1 and N_{MM} = 2 and no parent (parental pair) was found at the N_{MM}s. It does not mean that analyses were omitted at the N_{MM}s after the successful discovery of single parental pair of offspring at N_{MM} = 0.

When a parentage relationship is determined at a small value of N_{MM} (e.g., N_{MM} of one or two), the genotype data as well as electrophoretograms should be checked whether it contains genotyping errors. If it is certain that the data is error-free, the likelihood that the parentage relationship is false is high; nevertheless, the possibility that unverifiable allelic transmission errors, such as mutations, occurred in a true genealogical line cannot be rejected. This is a pitfall of exclusion method: it is usually difficult to accept/reject probable genealogical relationships having genotype incompatibilities at a few markers. This is problematic especially for samples with incomplete set of putative parents.

The paucity of DNA markers will result in non-exclusion of multiple putative parents (parental pairs), not allowing a resolution of parentages unless additional markers are used. This is another pitfall of exclusion method: for samples with a large parental pool, the number of markers required for complete exclusion may become prohibitively large.

The macro ***Exclusion*** cannot cope with these two limitations. Other methods may be used to resolve the problems (for a comprehensive review of currently available methods, see Jones & Ardren 2003; Jones et al. 2010). The likelihood method shown below is one alternative.

5-3. Likelihood-based method

To perform the likelihood-based method, user has to use three macros: ***Lhood_SimLOD***, ***Lhood_PrvLOD*** and ***LhoodReaLOD***. The ***PFX_Varstat*** also is used to calculate population allele frequencies if there is no pre-existing allele frequency data.

5-3-1. Preparation of allele frequency data (macro ***PFX_Varstat***)

§ Run ***PFX_Varstat*** (section 5-1).

§ Create a new spreadsheet and name it '**AlleleFreq**' (**case sensitive**). In the left-top cell, place a token '**Frequency**' (cell A1 in Fig. 5-7; **case sensitive**).

§ Copy the allele frequencies recorded in 'Varstat_summary' sheet and paste it onto the 'AlleleFreq' sheet.

No blank line from Line 1 to the end of data is allowed (Fig. 5-7). If another allele frequency data is available (e.g., data of a population from which parental samples were collected), change the format following this 'AlleleFreq' format. ***PFX_Varstat*** provides the

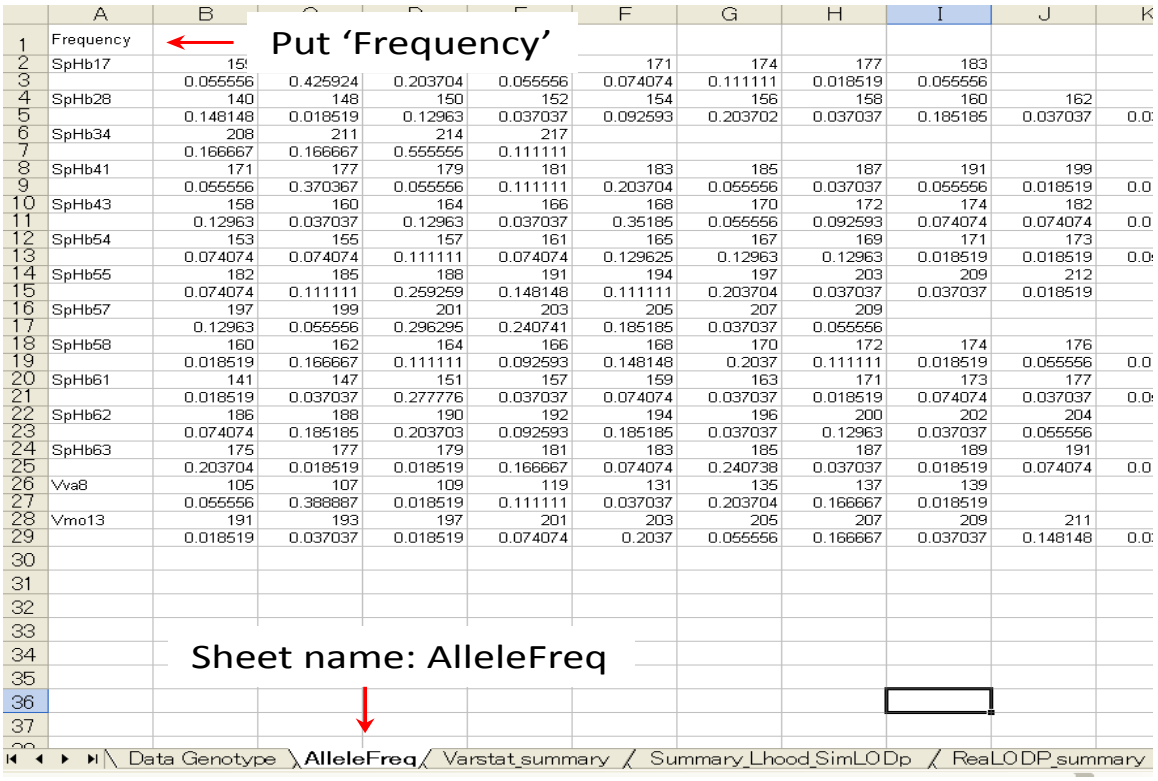


Fig. 5-7

allele frequencies of parental, offspring and combined (parents + offspring) samples separately. Typically, parental allele frequency data will be used (see **Note 7**).

5-3-2. Simulated LOD distributions (macro *Lhood_SimLOD*)

According to the allele frequency data, the macro ***Lhood_SimLOD*** carries out simulations to obtain a threshold LOD (LOD_C: see section 2-2 before going further).

§ Show 'AlleleFreq' sheet.

§ Open the macro window, select and run ***Lhood_SimLOD***.

§ Select markers in marker selection window.

§ Parameter setting window appears (Fig. 5-8; see below).

§ Click 'OK' to start calculations.

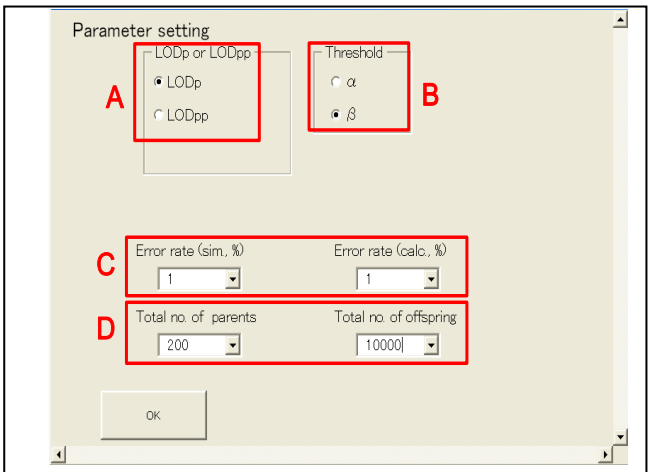


Fig. 5-8

Several essential parameters should be set here (Fig. 5-8). These include:

A) The type of LOD to be calculated. '**LOD_p**' is single parent LOD used for *single parent search.* '**LOD_{pp}**' is parental pair LOD for *parental pair search*.

B) In ***Lhood_SimLOD***, a LOD_C may be determined at the intersection between L_{ofs} and L_{rand} distributions (Fig. 5-10), where both types of errors (type I, α; type II, β) could be minimized (Gerber et al. 2000). In addition, user can select α or β to obtain LOD scores that may be used as a LOD_C. For example, by selecting β in the window, a LOD score as well as α corresponding to each pre-designated value of β (e.g., 0.05) is estimated based on the two LOD distributions. Of note, the power of test is 1 – β.

C) Set a genotypic error rate (%) for random replacement of simulated genotypes at each marker (e_{sim}; sim, %) and for LOD calculations (e_{calc}; calc, %). Error rate of 0.0, 0.1, 1.0 and 5.0% are available. User may set different values of e_{sim} and e_{calc}.

D) Set the number of parents (POP_{m+f}) created from allele frequencies. Default number is 200 (100 for each sex). Default number of offspring (POP_{ofs}) is 10^4 and the same number is applied to random individuals (POP_{rand}). The number of offspring can be chosen from 100, 400, 5×10^3, 10^4, 2×10^4 and 4×10^4. However, smaller numbers (100 and 400) should only be used to check the computational speed (see **Note 9**).

Results are given in '**Summary_Lhood_SimLODp**' sheet for LOD_p (**Summary_Lhood_SimLODpp**' sheet for LOD_{pp}). Here is an example of LOD_p calculations (Fig. 5-9).

Line 1: Types of LOD (LOD_p here).

Line 2: Markers used for LOD calculations.

Line 3: Parameters set by user.

Lines 4-7: In column B, α, β and LOD at the intersection are shown. In this example, we fixed β to obtain the corresponding LOD and α. These values are presented from columns C to F. For column G (yellow cells), see later**.**

Lines 9-12: Since the most likely parents (or parental pair) of offspring based on L_{ofs} may not be identical to the true parents (parental pair), the identities should be checked.

Single parent search: Percentage of offspring whose true parents (both parents) are identical to the two most likely parents is shown in 'Both sexes' category. 'Single sex' category provides percentage of offspring whose one parent is not identical to either the most likely parents. Percentage of offspring falling into neither these two categories is given in 'Failed' category.

Parental pair search: Percentage of offspring whose true parental pair is identical to

the most likely pair is shown in 'Both Sex' category. 'Single sex' category is omitted. Percentage of offspring whose true parental pair does not match with the most likely pair is shown in 'Failed' category.

Lines 20–X: L_{ofs} and L_{rand} distributions are numerically shown (percentage; POP_{m+f} *vs* POP_{ofs} and POP_{m+f} *vs* POP_{rand}, respectively).

Graph: L_{ofs} and L_{rand} are plotted in a graph (Fig. 5-10).

	A	B	C	D	E	F	G	H	I
1	LODp								
2	Markers employed (14/14):	SpHb17	SpHb28	SpHb34	SpHb41	SpHb43	SpHb54	SpHb55	SpHb57
3	Parents	200	Offspring	10000	e (sim) %	1	e (cal) %	1	
4		Intersection							
5	β	0.0369	0.001	0.01	0.05	0.1	User input threshold		
6	α	0.0263	0.23375	0.07105	0.0192	0.00835	input below		
7	Threshold LOD	3.72	9.54	5.93	3.09	1.38			
8									
9	Correct allocation (%)								
10	Both sexes	99.0							
11	Single sex	1.0							
12	Failed	0.0							
13									
14									
15									
16									
17									
18									
19									
20	Rank of LOD	POPm+f vs POP	POPm+f vs POPrand						
21	-18	0	0						
22	-17.2	0	0.01						
23	-16.4	0	0.04						
24	-15.6	0	0.1						
25	-14.8	0	0.18						
26	-14	0	0.12						
27	-13.2	0	0.31						
28	-12.4	0	0.74						

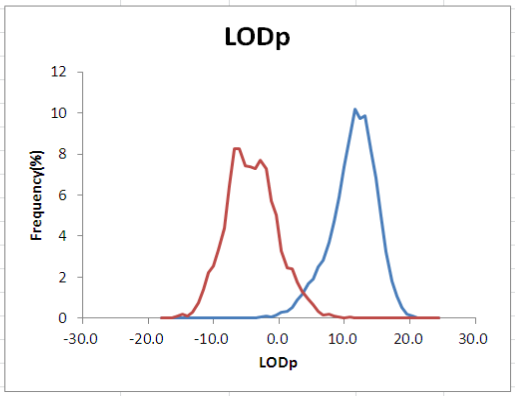


Fig. 5-9. Result sheet of *Lhood_SimLOD* (LOD_p) ('Summary_Lhood_SimLODp' sheet)

The aforementioned way to define LOD_C is possible only when the lower tail of L_{ofs} distribution and the upper tail of L_{rand} distribution overlap (Fig. 5-10). If the LOD distributions are segregated completely without intersection, the minimum value of L_{ofs} is suggested as a possible value of LOD_C (neither α nor β is shown) (see **Note 8**). In addition, L_{rand} is not estimated when e_{calc} is set at zero (see glossaries). In such a case, again, the minimum value of L_{ofs} is provided without showing both α and β.

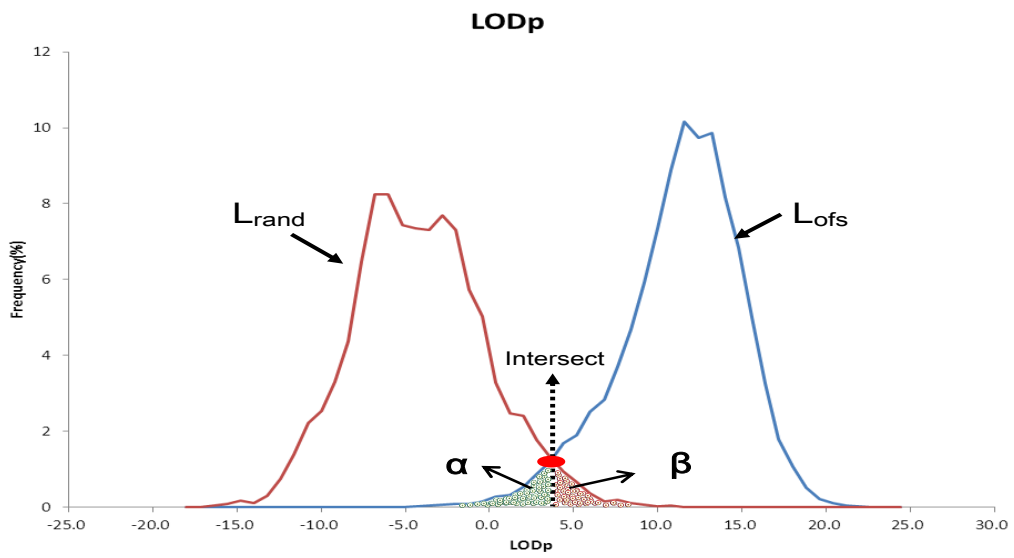


Fig. 5-10. Distributions of L_{ofs} and L_{rand} (LOD_p)

5-3-3. Validation of parentage allocation (macro *Lhood_PrvLOD*)

Success rate of parentage allocation conditional on LOD_C has to be evaluated using the macro ***Lhood_PrvLOD***. Success rate (%) is defined as the number of offspring with correct parentage allocation (at a LOD_C) divided by the total number of offspring.

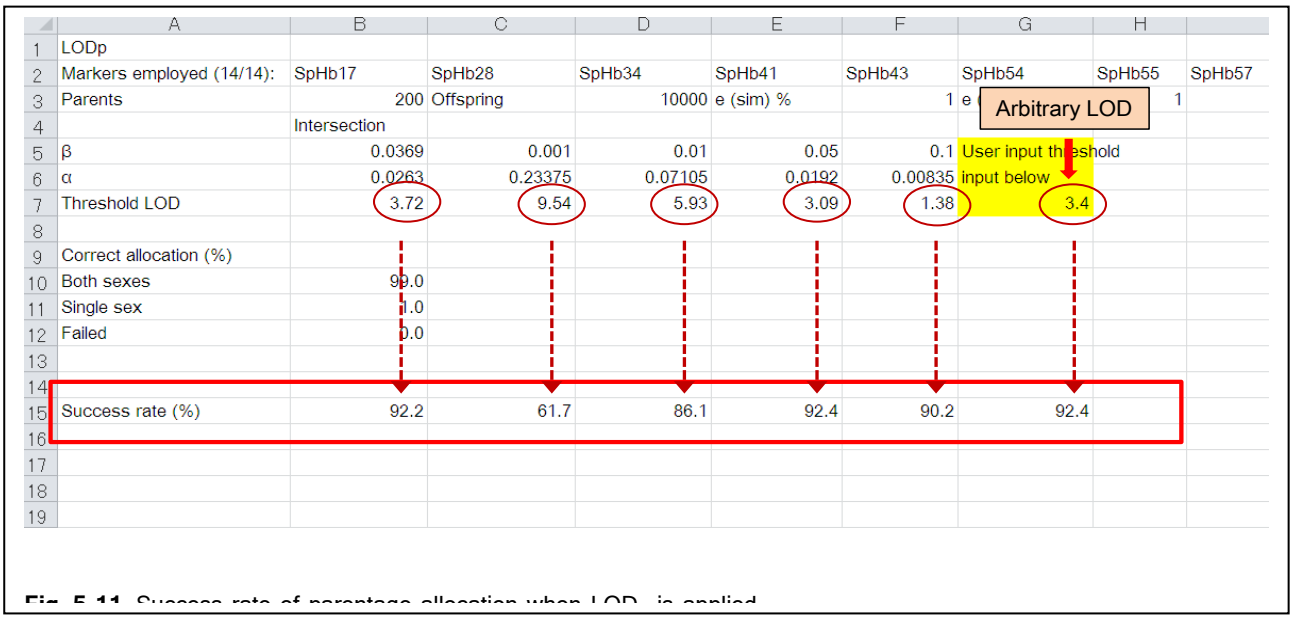
In *single parent search*, the term 'offspring with correct parentage allocation' is expressive of an offspring for whom none of putative parents excepting the true parents gives any LOD score exceeding the value of LOD_C. Thus, it is judged as incorrect allocation when the true parents plus additional putative parent(s) have a LOD score higher than the value of LOD_C. It is also considered as incorrect when either or both the true parents yield a LOD score smaller than the value of LOD_C. In *parental pair search*, a correct parentage allocation for offspring refers to the case where LOD score exceeding the value of LOD_C is exclusively obtained between the offspring and true parental pair (see **Note 10**).

§ Show 'Summary_Lhood_SimLODp' sheet (or _SimLODpp).

§ An arbitrary LOD_C value can be tested. Put it in the yellow cell G7 (Fig. 5-9 and 5-11).

§ Open the macro window, select and run ***Lhood_PrvLOD***'.

In ‘Summary_Lhood_SimLODp’ sheet, a result space comes up like Fig. 5-11 (red box). Check the values shown in Line 15, where the success rate of parentage allocation corresponding to each possible LOD_C is shown. The value of LOD_C determined at the



	A	B	C	D	E	F	G	H	
1	LODp								
2	Markers employed (14/14):	SpHb17	SpHb28	SpHb34	SpHb41	SpHb43	SpHb54	SpHb55	SpHb57
3	Parents	200	Offspring	10000	e (sim) %	1	e	1	
4		Intersection							
5	β	0.0369	0.001	0.01	0.05	0.1			
6	α	0.0263	0.23375	0.07105	0.0192	0.00835			
7	Threshold LOD	3.72	9.54	5.93	3.09	1.38	3.4		
8									
9	Correct allocation (%)								
10	Both sexes	99.0							
11	Single sex	1.0							
12	Failed	0.0							
13									
14									
15	Success rate (%)	92.2	61.7	86.1	92.4	90.2	92.4		
16									
17									
18									
19									

Fig. 5-11 Success rate of parentage allocation when LOD is applied

intersection of L_{ofs} and L_{rand} distributions is given in cell B7 and the corresponding success rate is found in the same column (column B). In this case, the success rate reaches *ca* 92% with the value of LOD_C from 3.1 to 3.7. In the next section, we apply a LOD_C of 3.7 (determined at the intersection) in parentage testing for real data.

5-3-4. Parentage reconstruction for real genotype data (macro *Lhood_ReaLOD*)

Based on the LOD_C estimated in the preceding simulations, the macro ***Lhood_ReaLOD*** conducts parentage testing for real data.

§ Show ‘Data Genotype’ sheet. Genotype with missing allele(s) is omitted from LOD calculations.

§ Open the macro window, select and run ***Lhood_ReaLOD***.

§ Parameter setting window appears (Fig. 5-12). Set the parameters. Note that markers listed in ‘Summary_Lhood_SimLODp’ sheet (‘_SimLODpp’ in *parental pair search*) are used (Fig. 5-9). Note also that the allele frequency data in ‘AlleFreq’ sheet is used in calculations. Therefore, **NEVER** change these sheet names.

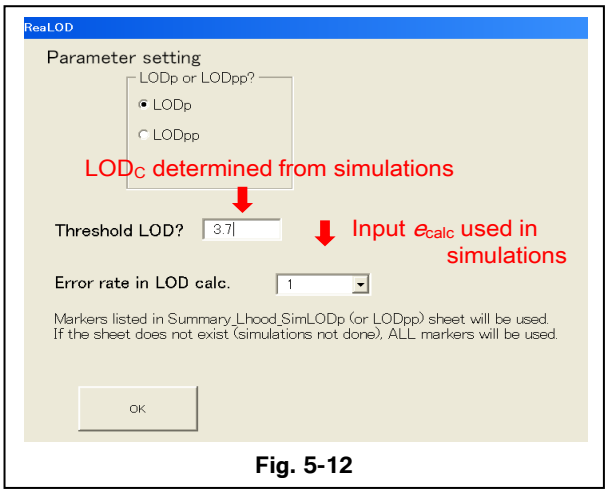


Fig. 5-12

Results are given in '**ReaLOD_summary**' sheet (Fig. 5-13). Markers used in LOD calculations are shown in Line 1. In Line 2, the parameters set by user are given. From

	A	B	C	D	E	
1	Markers employed (14/14):	SpHb17	SpHb28	SpHb34	SpHb41	SpHb43
2	Threshold	3.7	e (calc) %	1		
3	**Offspring**	**LOD**	**Parentage (Male or unknown)**	**Parentage (Female or unknown)**	**Incompatible Markers**	
4	MIR8_001	13.241	MIW074_M			
5		9.437		MIW075_F		
6	MIR8_002	12.042		MIW086_F		
7		11.407	MIW021_M			
8	MIR8_003	15.686	MIW079_M			
9		7.616		MIW075_F		
10	MIR8_004	11.056		MIW060_F		
11		10.595	MIW018_M			
12	MIR8_005	12.29	MIW096_M			
13		9.278		MIW075_F		
14	MIR8_006	13.374	MIW096_M			
15		9.485		MIW075_F		
16	MIR8_007	15.107		MIW086_F		
17		11.316	MIW018_M			
18	SimOfs1_MIW036_M-MIW045_F	12.908		MIW045_F		
19		11.455	MIW036_M			
20	SimOfs2_MIW073_M-MIW058_F	10.742		MIW058_F		
21		10.54	MIW073_M			
22	SimOfs3_MIW014_M-MIW097_F	10.321		MIW097_F		
23		7.356	MIW014_M			
24	Simrand_1					
25	Simrand_2					
26	Simrand_3					
27	Simrand_4					
28	Simrand_5					
29	Simrand_6					

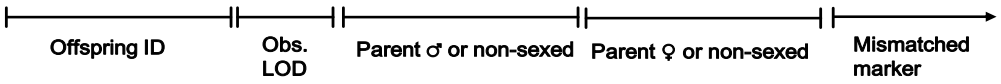


Fig. 5-13. Result sheet of *Lhood_ReaLOD* analysis ('ReaLood_summary' sheet)

Offspring with the suffix 'MIR8': real samples. Simulated offspring (SimOfs) and random individuals (Simrand) generated using ***PFX_Ofsgen*** also are included. Parental ID of each simulated offspring is found in offspring ID (e.g., in Line 18, SimOfs1_MIW036_M-MIW045_F: ID of the true male parent is MIW036_M and female parent MIW045_F)

Lines 4–X are the results of parentage allocation. Since we performed *single parent search*, the most likely parents of each offspring (columns C and D) are placed in separate lines. In *parental pair search*, one line should be given to respective parental pairs. Observed LOD scores are in column B. Putative parents whose LOD scores are smaller than the value of LOD_C do not appear in the results (Lines 24-29 in Fig. 5-13). Markers with genotype incompatibility between offspring and the most likely parent (parental pair) are presented from column E rightward, if any. When an e_{calc} of zero is set, LOD calculation is omitted for any combination between putative parent (parental pair) and offspring having one or more mismatched markers; such a relationship is rejected irrespective of the value of LOD_C (see glossaries).

!! Important !! A common result sheet name '**ReaLOD_summary**' is used to output the results of both LOD_p and LOD_{pp}. If both calculations are made in the same workbook, it needs to rename the existing result sheet (e.g., ReaLOD_summary → ReaLODp_summary and ReaLODpp_summary). Unless doing so the existing results will be lost.

It is not necessarily warranted that a LOD_C determined at the intersection of L_{ofs}

and L_{rand} distributions is most suitable to retrieve true genealogical relationships (see Note 11). Therefore, we recommend that several values of LOD_C should be tested. Success rate of parentage allocation, however, should be checked for every LOD_C (***Lhood_PrvLOD***). Corresponding α and β can be estimated by inspecting the LOD distributions numerically shown in the summary sheet (Fig. 5-9).

We emphasize that simulated offspring and random individuals should be analyzed (below, section 5-3-5) to assess *ad hoc* the validity of defined LOD_C. Simulated samples are easily obtained with the macro ***PFX_Ofsgen*** (below, section 5-4).

We validated the algorithm of likelihood method encoded in PARFEX using simulated data sets. A brief example is shown in Note 11.

5-3-5. Parentage success in simulated genotype data (macro *Lhood_Validat*)

Although manual checking for the correctness of parentage allocation for a number of simulated offspring and random individuals may be feasible, it is rather tedious work and time consuming. Thus, we offer the following easier way to calculate the success rate of parentage for simulated genotype data using the macro ***Lhood_Validat***.

§ Create new worksheet and name it arbitrarily (here we name it as '**Data Genotype_Sim**' sheet; Fig. 5-14).

§ Copy the contents of '**Data Genotype**' sheet (real genotype data) and paste it onto the newly created 'Data Genotype_Sim' sheet.

§ Replace the offspring data in the 'Data Genotype_Sim' sheet with simulated genotype data created by the macro ***PFX_Ofsgen*** (see section 5-4). The number of simulated offspring and random individuals is set by user (e.g., 1000 for each, 2000 in total). Be sure that the total number does not exceed the maximum number of offspring allowed by PARFEX (max. 5000: see section 4-1). And **NEVER** change the names of simulated offspring and random individuals.

§ Run the macro ***Lhood_ReaLOD*** on the 'Data Genotype_Sim' sheet. The same parameters for LOD calculation you used in the parentage allocation for real data (Fig. 5-12) should be set.

§ Parentage results are shown in '**ReaLOD_summary**' sheet. The name of this output sheet is identical to the one you got in the analysis for real data. Therefore, again, **the result sheets for real data should be renamed beforehand!** Otherwise, you will

completely lose the results for real data.

§ Rename the ‘ReaLOD_summary’ sheet arbitrarily. Here, ‘ValiSimLODp’ (Fig. 5-15). However, **NEVER** modify the contents of this sheet.

§ Run the macro ***Lhood_Validat*** on the ‘ValiSimLODp’ sheet.

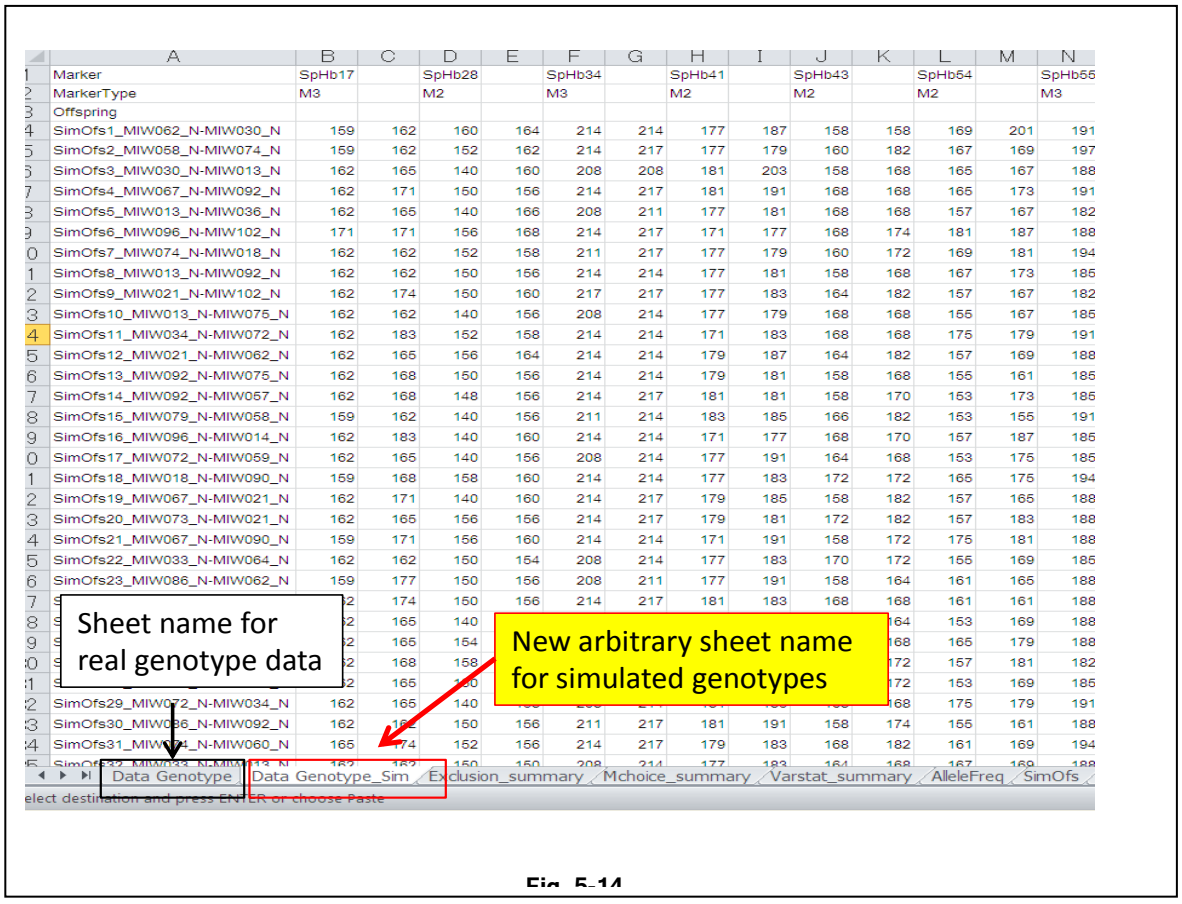


Fig. 5-14

The macro ***Lhood_Validat*** calculates the success rate of parentage allocation (**red box** in Fig. 5-15). For simulated offspring, the term ‘success rate of parentage’ is defined in section 5-3-3. Simulated offspring are categorized into ‘Correct allocation (successful allocation)’ and ‘Failed allocation’, and percentage of offspring fallen into each category is shown. Random individuals are assumed to have no parent in the parental pool. Thus, they are categorized into either ‘Correct rejection (no parent assigned)’ or ‘Failed rejection (one or more parents assigned)’ (for LOD_{pp}, please read by replacing ‘parent’ to ‘parental pair’). The macro ***Lhood_Validat*** automatically recognizes whether the parentage results are based on LOD_p or LOD_{pp} by looking at the cell ‘B3’ in the ‘ValiSimLODp’ sheet (**green box** in Fig. 5-15).

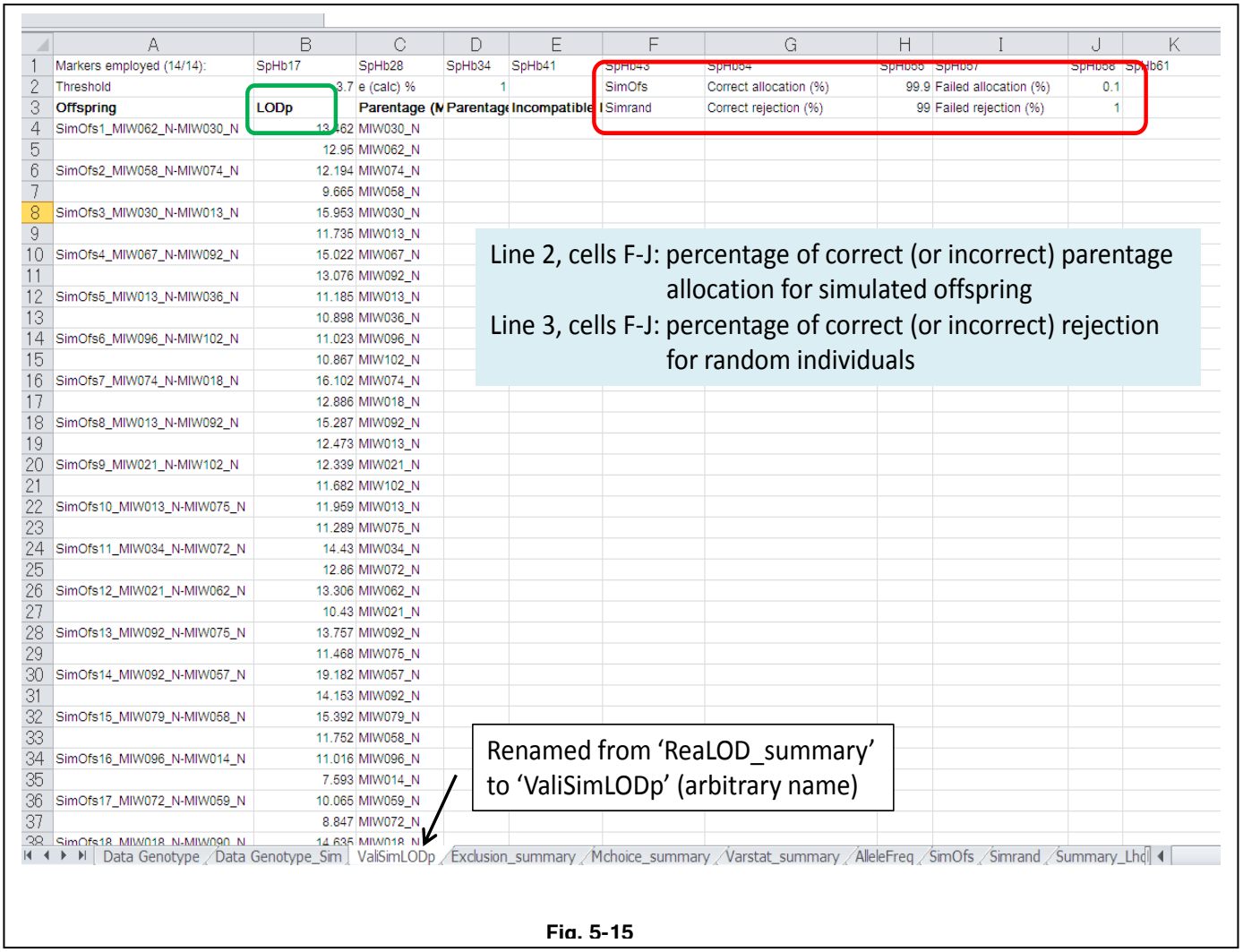


Fig. 5-15

5-4. Generation of simulated offspring (macro *PFX_Ofsgen*)

Assuming random mating, the macro ***PFX_Ofsgen*** produces simulated offspring from collected parental population. It also creates random individuals (genotypes) by random sampling of alleles according to parental allele frequencies.

§ Show 'Data Genotype' sheet.

§ Open the macro window and run ***PFX_Ofsgen***.

§ Notify whether parental sex information is used to create offspring (Fig. 5-16A, with known parental sexes). **Yes**: offspring are created assuming random mating between sexes (**Note 4**); **No**: parents are capable of mating with any other parents.

§ Set the number of offspring/random individuals (Fig.

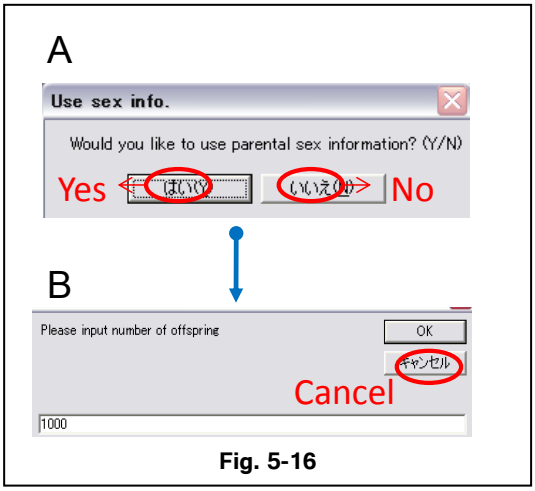


Fig. 5-16

5-16B) (max. 4×10^4 for each).

§ Notify how missing alleles should be handled (Fig. 5-17). Three options are available: 1) individuals having missing allele(s) are removed before generating offspring, 2) missing allele ('?') is treated as an existing allele and descended randomly to offspring and 3) missing allele is replaced by another allele randomly retrieved according to parental allele frequency data (See **Note 4**).

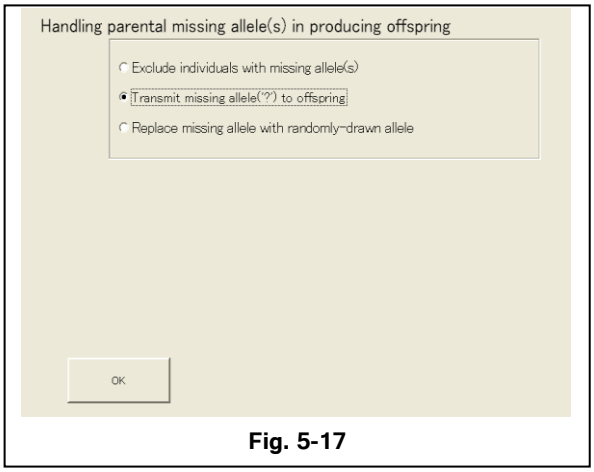


Fig. 5-17

Offspring and random individuals are provided in separate spreadsheets ('**SimOfs**' and '**Simrand**' sheet, respectively). Offspring genotypes are shown along with the true parental pair. Parental IDs of each offspring are incorporated in offspring ID (Fig. 5-18).

SimOfs

	A	B	C	D	E	F	G	H	I	J	K	L	M
1			Marker	SpHb17		SpHb28		SpHb34		SpHb41		SpHb43	
2	Male	Female	Offspring										
3	MIW036_M	MIW045_F	SimOfs1_MIW036_M-MIW045_F	162	162	140	168	211	214	177	183	168	1
4	MIW062_M	MIW057_F	SimOfs2_MIW062_M-MIW057_F	159	168	148	164	214	214	177	177	158	1
5	MIW018_M	MIW058_F	SimOfs3_MIW018_M-MIW058_F	162	165	140	158	211	214	177	183	172	1
6	MIW062_M	MIW060_F	SimOfs4_MIW062_M-MIW060_F	165	174	156	164	214	214	185	187	164	1
7	MIW014_M	MIW086_F	SimOfs5_MIW014_M-MIW086_F	162	162	156	160	208	214	183	199	170	1
8	MIW072_M	MIW045_F	SimOfs6_MIW072_M-MIW045_F	162	171	140	168	208	208	177	183	168	1
9	MIW034_M	MIW075_F	SimOfs7_MIW034_M-MIW075_F	162	183	154	156	214	214	177	181	160	1
10	MIW059_M	MIW097_F	SimOfs8_MIW059_M-MIW097_F	165	174	150	154	211	214	177	191	164	1

True parental pair Offspring + Parental ID Genotype

Simrand

	A	B	C	D	E	F	G	H	I	J	K	L	M
1			Marker	SpHb17		SpHb28		SpHb34		SpHb41		SpHb43	
2	Male	Female	Offspring										
3			Simrand_1	159	165	162	168	214	214	177	177	158	18
4			Simrand_2	162	165	154	156	208	214	177	181	168	17
5			Simrand_3	162	162	140	162	214	214	177	191	172	17
6			Simrand_4	162	165	150	154	211	214	177	183	168	17
7			Simrand_5	162	174	160	168	208	214	171	199	168	17
8			Simrand_6	162	171	140	156	208	214	177	181	164	16
9			Simrand_7	162	165	140	160	208	208	177	179	158	17
10			Simrand_8	165	174	150	160	211	214	185	187	174	18

Fig. 5-18. Simulated offspring and random individuals produced by *PFX_Ofsgen*

6. Notes

♠ **Note 1:** Mutations (insertions, deletions or substitutions) occurring in the nucleotide sequences of PCR primer binding sites prevent PCR amplification of either or both alleles at marker loci. Such non-amplifying alleles are called null alleles (e.g., Callen et al. 1993; Ede & Crawford 1995; Pemberton et al. 1995).

♠ **Note 2:** Alphabetical designation of alleles is not valid in CONVERT. Therefore, SNPs data is not compatible between CONVERT and PARFEX. However, it is easy to reciprocally transform the allelic codes using the 'REPLACE' function of EXCEL. Another software program for file conversion, CREATE (Coombs et al. 2008), uses similar genotype format.

♠ **Note 3:** ***PFX_Fcheck*** conducts the format check **(4)** using 'MarkerType' information, setting the first allele of the first offspring (Line 6 in Fig. 4-1) as the base allele at each marker. At *SpHb34* marker in Fig. 4-1, the allele *214* of offspring MIR8_001 (cell G6) is set as the base allele. The macro examines if the size of the other alleles at this marker is explained by the size of the base allele plus or minus $3 \times n$ ('3' comes from 'M3' and n is arbitrary integer). When the first allele of the first offspring is missing, the second allele is set as the base. When both alleles of the first offspring are missing, the first allele of the second offspring (MIR8_002 in Fig. 4-1,) is set as the base, and so on.

♠ **Note 4:** In simulations, mating within sexes is ruled out for sexed parents. Non-sexed parents are assumed to be capable of mating with any other parents. Sex information is also used in the exclusion method and the likelihood analysis for real data (***Lhood_ReaLOD***), as pairs within sexes are omitted before analysis. In producing offspring by simulation, user may opt to replace missing allele(s) by another allele randomly drawn according to parental allele frequencies. This procedure is applied to both cases where both alleles are missing (missing genotype) and one of the two alleles is missing. The replaced parental alleles are transmitted randomly to offspring following Mendelian inheritance of alleles.

♠ **Note 5:** When markers score a tied value of A_{uni}/A_{obs}, they are ranked based on the observed number of different alleles (A_{obs}): the larger the number, the higher the rank.

♠ **Note 6:** When one parent of offspring has a null allele at a marker, the offspring should have particular genotypes conditional on the genotype of the other parent (Table N6-1; it assumes complete amplification failure of null alleles). According to this table, six combinations of parental and offspring genotypes are expected to generate detectable mismatches (arrows in Table N6-1). The macro ***Exclusion*** hints a null-allele segregation

when it finds any of the six genotype configurations between offspring and its

Table N6-1. Expected genotype configuration of offspring with null allele segregation

	Genotype (P_2)					
	AA	***AØ*** (***AA***)	***AB***	***BB***	***BØ*** (***BB***)	***BC***
Genotype (P_1)	***AØ***† (***AA***)	***AØ***† (***AA***)	***AA***	***AB***	***AB***	***AB***
AØ (***AA***)	***AA***	***AA***	***AB***	***BØ*** (***BB***) ← 2	***AØ*** (***AA***) ← 3	***AC***
		ØØ‡	***AØ***† (***AA***)		***BØ*** (***BB***) ← 4	***BØ*** (***BB***) ← 5
			BØ (***BB***) ← 1		***ØØ***‡	***CØ*** (***CC***) ← 6

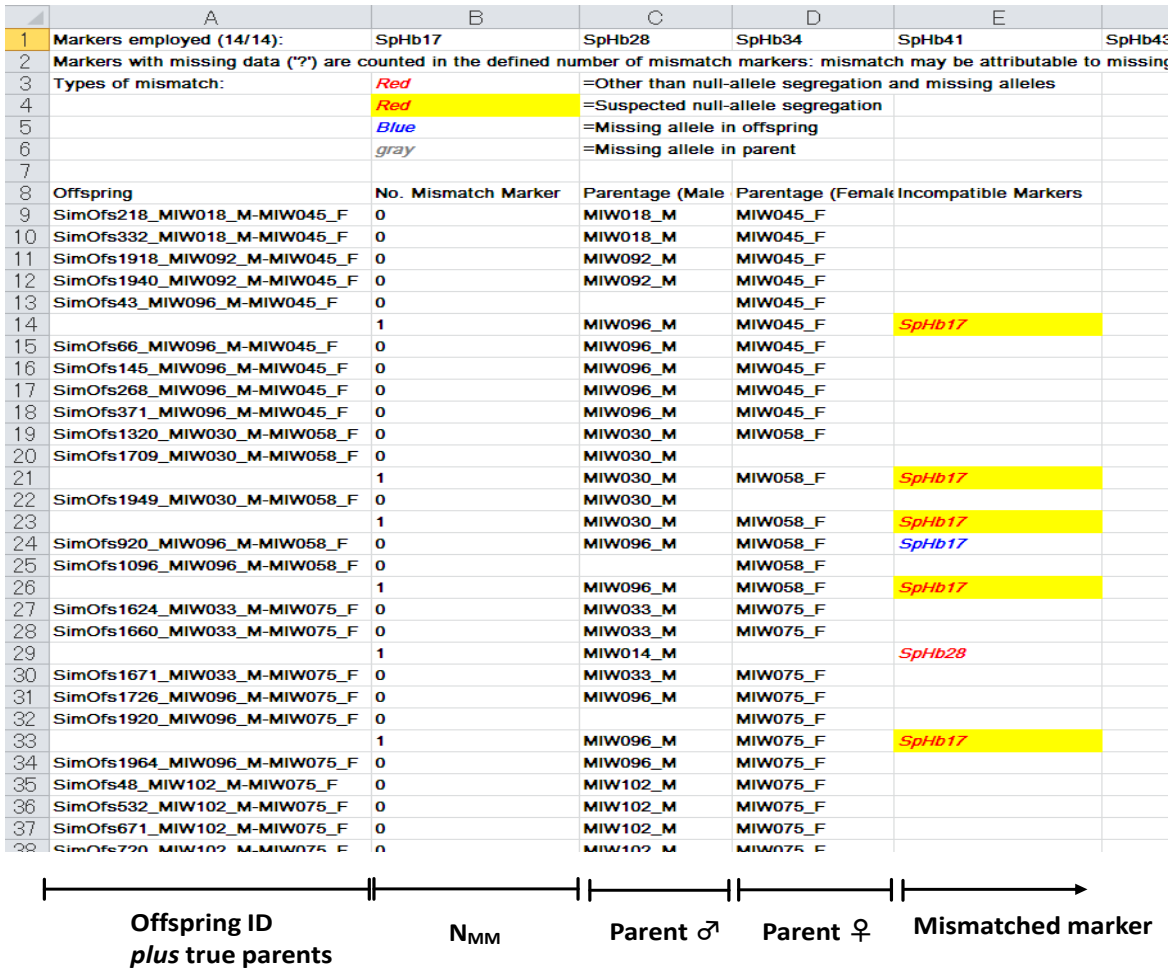
One parent P_1 carries a visible allele ***A*** and invisible null allele ***Ø***. For the other parent (P_2), all possible genotypes with or without null allele are shown. P_1's genotype ***AØ*** should be typed as an apparent homozygote ***AA*** (shown in parenthesis: the same applies to other genotypes with null allele). Homozygote of null allele (***ØØ***) that should be recorded as missing data is omitted from the genotypes of P_2. Offspring genotype with null allele(s) is shown in red. Arrows indicate detectable mismatches caused by null-allele segregation in parent-offspring genotypic triplets. †Offspring genotype with null allele indiscernible from true homozygote ***AA***. ‡Homozygote of null allele (***ØØ***) (missing genotype).

non-excluded parental pair. With an easy simulation, we show how this function works.

Genotype data (14 markers) was derived from a parental population of spotted halibut (27 fish: ♀ 7, ♂ 20). We simulated the transmission of null allele from parents to offspring at one marker (*SpHb17*), which showed a moderate variability in the parental population (Fig. 5-1). Among the parental fish, we randomly selected two fish (♀ ID: MIW058_F; ♂ ID: MIW096_M). For each, we replaced one of the two alleles at *SpHb17* by missing allele '?' (genotype: MIW058_F, 159/165→165/?; MIW096_M, 171/183→171/?), so that it could serve as a 'mock' null allele. We obtained simulated genotypes of offspring from the parental population using ***PFX_Ofsgen***, where the missing allele ('?') also was transmitted randomly to offspring (section 5-4). The true parents of each offspring were recorded. We assumed the genotype of MIW058_F and MIW096_M each as a homozygote of the unaltered allele (MIW058_F, 165/?→165/165; MIW096_M, 171/?→171/171). The same assumption was applied to offspring who received one missing allele. Offspring genotype with two missing alleles was treated as missing genotype. With the macro ***Exclusion***, we conducted parentage testing for this modified data set ($MaxN_{MM}$ = 1).

When the test for null-allele segregation available in the macro ***Exclusion*** is turned on, the result sheet 'Exclusion_summary' comes up like Figure N6-1. Markers with suspected null-allele segregation appear in red in yellow cells. A total of 2×10^3 simulated

offspring were generated, of which 1,606 were descended from neither the two parents having a mock null allele. Parental pairs of all the 1,606 offspring were determined at N_{MM} = 0, though a single false parent was allocated additionally to a few of the offspring at N_{MM} = 1. The remaining 394 were derived from either or both the null-allele-assigned parents (Table N6-2). In this case, all types of mismatch caused by quasi null-allele segregation were detected precisely.



	A	B	C	D	E	F
1	Markers employed (14/14):	SpHb17	SpHb28	SpHb34	SpHb41	SpHb43
2	Markers with missing data ('?') are counted in the defined number of mismatch markers: mismatch may be attributable to missing allele(s).					
3	Types of mismatch:	*Red*	=Other than null-allele segregation and missing alleles			
4		*Red*	=Suspected null-allele segregation			
5		*Blue*	=Missing allele in offspring			
6		*gray*	=Missing allele in parent			
7						
8	Offspring	No. Mismatch Marker	Parentage (Male	Parentage (Femal	Incompatible Markers	
9	SimOfs218_MIW018_M-MIW045_F	0	MIW018_M	MIW045_F		
10	SimOfs332_MIW018_M-MIW045_F	0	MIW018_M	MIW045_F		
11	SimOfs1918_MIW092_M-MIW045_F	0	MIW092_M	MIW045_F		
12	SimOfs1940_MIW092_M-MIW045_F	0	MIW092_M	MIW045_F		
13	SimOfs43_MIW096_M-MIW045_F	0		MIW045_F		
14		1	MIW096_M	MIW045_F	*SpHb17*	
15	SimOfs66_MIW096_M-MIW045_F	0	MIW096_M	MIW045_F		
16	SimOfs145_MIW096_M-MIW045_F	0	MIW096_M	MIW045_F		
17	SimOfs268_MIW096_M-MIW045_F	0	MIW096_M	MIW045_F		
18	SimOfs371_MIW096_M-MIW045_F	0	MIW096_M	MIW045_F		
19	SimOfs1320_MIW030_M-MIW058_F	0	MIW030_M	MIW058_F		
20	SimOfs1709_MIW030_M-MIW058_F	0	MIW030_M			
21		1	MIW030_M	MIW058_F	*SpHb17*	
22	SimOfs1949_MIW030_M-MIW058_F	0	MIW030_M			
23		1	MIW030_M	MIW058_F	*SpHb17*	
24	SimOfs920_MIW096_M-MIW058_F	0	MIW096_M	MIW058_F	*SpHb17*	
25	SimOfs1096_MIW096_M-MIW058_F	0		MIW058_F		
26		1	MIW096_M	MIW058_F	*SpHb17*	
27	SimOfs1624_MIW033_M-MIW075_F	0	MIW033_M	MIW075_F		
28	SimOfs1660_MIW033_M-MIW075_F	0	MIW033_M	MIW075_F		
29		1	MIW014_M		*SpHb28*	
30	SimOfs1671_MIW033_M-MIW075_F	0	MIW033_M	MIW075_F		
31	SimOfs1726_MIW096_M-MIW075_F	0	MIW096_M	MIW075_F		
32	SimOfs1920_MIW096_M-MIW075_F	0		MIW075_F		
33		1	MIW096_M	MIW075_F	*SpHb17*	
34	SimOfs1964_MIW096_M-MIW075_F	0	MIW096_M	MIW075_F		
35	SimOfs48_MIW102_M-MIW075_F	0	MIW102_M	MIW075_F		
36	SimOfs532_MIW102_M-MIW075_F	0	MIW102_M	MIW075_F		
37	SimOfs671_MIW102_M-MIW075_F	0	MIW102_M	MIW075_F		

Fig. N6-1. Result sheet of *Exclusion* analysis with test for null-allele segregation.

Parental IDs of simulated offspring are incorporated in offspring ID.

It is obvious that the inference of null-allele segregation will not always be correct. At markers with limited allelic richness, true mismatches, or possibly, other types of allelic transmission errors could frequently produce such null-allele-segregating genotype configurations. Moreover, it will be difficult, or rather, not possible to trace the transmission of null alleles unless the offspring pool in question contains a certain number of full- or half-sibs having null alleles. Therefore, we should emphasize that suggestions by ***Exclusion*** do NOT provide evidence of null-allele segregation. However,

the function could help quickly find null alleles at highly variable markers in captive-bred populations containing many sib families (e.g., from our interest, hatchery populations and aquaculture strains of fish and shellfish).

Table N6-2. Number of offspring descended from MIW058_F and MIW096_M

	Compatible	Type 1	Type 2	Types 5 & 6	Total
MIW058_F	181 (19)	47 (8)	44 (6)	23 (4)	295 (19)
MIW096_M	52 (6)	2 (1)	14 (2)	13 (3)	81 (6)
MIW058_F *vs* MIW096_M	Compatible	Types 3 & 4	Missing		
	3	10	5		18
					394

Offspring are categorized according to the genotype configurations of parent-offspring triplet (Table N6-1). For the mismatch types caused by null alleles (Type 1, Type 2 etc.), see Table N6-1. 'Compatible' indicates no genotype incompatibility between parental and offspring genotypes. The number of half-sib families originated from each parental fish is presented in parenthesis. The number of offspring produced from a parental pair of MIW058_F and MIW096_M, each of which has a mock null allele, is shown separately ('MIW058_F *vs* MIW096_M'). Offspring with two null alleles are classified into 'Missing' category.

♠ **Note 7:** If some of offspring's rare alleles are not found in parental data, the use of parental allele frequencies is problematic: it means a null frequency of the offspring genotypes in the parental population thereby resulting in an indefinable LOD (see the equations in glossaries). This is not matter in simulations to determine a LOD_C but poses a big problem in parentage testing for real data. In such a case, allele frequencies of 'parents + offspring' may be used. Another option is addition of very low frequency of offspring's rare alleles to the parental allele frequency data.

♠ **Note 8:** Applying the minimum LOD score as a LOD_C sometimes results in a lower success rate of parentage allocation due to the acceptance of false parentage relationships. In such a case, try a higher LOD score and check the success rate using ***Lhood_PrvLOD***.

♠ **Note 9:** Computational speed of simulations is especially slow. With the parameters shown in Figure 5-8 and 14 markers, it took *ca* 5 min to complete LOD_p calculation and more than four hrs for LOD_{pp} calculation on MS's PC (Intel Core^{TM} 2 Duo, CPU: 3.0 GHz; RAM: 3.24 GB). In LOD_{pp} calculation, you better start running the macro just before going

home. In the next morning, the mission will be accomplished (hopefully…)

♠ **Note 10:** In order to reduce computational burden, any combination between putative parent (parental pair) and offspring yielding a LOD score smaller than the minimum value of 'L_{rand}' is excluded from the analysis.

♠ **Note 11**: We show an example illustrating the accuracy of PARFEX likelihood-based parentage allocation as well as the validity of calculation codes thereof. Using ***PFX_Ofsgen***, we simulated 10^3 offspring from the 27 parental fish (section 5-3 and Note 6; 14 markers). We ignored the parental sex assuming that they could mate with any other parents, so that the number of possible parental combinations and the variations of offspring genotypes could be increased. To produce a pool of individuals who were assumed to be unrelated to the parental fish, we generated 10^3 individuals by random sampling of alleles using the allele frequency data. Based on this data set, we performed both exclusion ($MaxN_{MM}$ = 2) and likelihood-based methods under non-sexed condition. In simulations of likelihood-based method, we used a common parameter setting described above for all the analyses (no. of parents, 200; offspring and random individuals, 10^4; e_{sim} and e_{calc}, 1.0%). In *single parent search* (likelihood), we applied a LOD_C of 3.7 determined in section 5-3. In *parental pair search*, the highest success rate was obtained at a LOD_C of 16.0 (α = 0.01, β < 0.0001, success rate of 98.5%) rather than at a possible LOD_C determined at the intersection (LOD_C = 10.0, 92.5% success rate). Thus, we set the former value as an appropriate LOD_C.

The likelihood equations adopted in PARFEX are taken from Kalinowski et al. (2007), which are the revised version of old ones (these 'unrevised' are explicitly formulated in Marshall et al. 1998 for single parent LOD and Morrisey & Wilson 2005 for parental pair LOD). Of note, the frequencies of both offspring and parental genotypes are required to calculate LOD scores when the unrevised equations are used (c.f., revised equations in glossaries). For a comparative purpose, we also conducted parentage testing based on the 'unrevised' equations using the same data set. This was done by replacing the 'revised' equations encoded in the script of PARFEX with the 'unrevised' ones. We explored a LOD_C *de novo* through simulations for both *single parent* and *parental pair searches*. In *single parent search*, we obtained a LOD_C at the intersection (LOD_C = 4.9, α = 0.10, β = 0.10, success rate of 73.2%). The reliability of the LOD_C was low, but changing the value of LOD_C brought little benefit to improve it (e.g., success rate of 73.3% at a LOD_C of 3.5). In *parental pair search*, we applied a LOD_C of 14.5 since the highest success rate was found at this value (α = 0.05, β < 0.0001, success rate of 93.7%) rather than at a possible LOD_C of 11.7 determined at the intersection (66.5%

success rate).

The results of parentage allocation are summarized in Table N11. Almost all ambiguous allocations (offspring) and false acceptances (random individuals) remained in the exclusion analysis ($MaxN_{MM}$ = 2) disappeared in the likelihood method based on the 'revised' equations. Both *single parent* and *parental pair searches* accomplished a nearly perfect correct allocation/rejection with an overall success rate of more than 99%, strengthening the credibility of PARFEX likelihood analysis. On the other hand, the likelihood analysis using the 'unrevised' equations gave worse results, especially in *single parent search*. This reduced power was not unforeseeable given the lower reliability of LOD_C. In *single parent search* ('unrevised' equations), applying more relaxed LOD_C (3.5) did increase the percentage of correct allocation for offspring (95.7%) but decreased the percentage of correct rejection for random individuals (86.7%), resulting in little improvement in the overall success rate (91.5%).

Table N11. Success rate of parentage allocation (%) for simulated samples

	Exclusion	Likelihood_*single parent*		Likelihood_*parental pair*	
	$MaxN_{MM}$ = 2	Revised (LOD_C = 3.7)	Unrevised (LOD_C = 4.9)	Revised (LOD_C = 16.0)	Unrevised (LOD_C = 14.5)
Offspring (*N* = 1000)					
Correct allocation[a]	**94.1**	**99.9**	**83.9**	**99.9**	**98.8**
Ambiguous allocation[b]	**5.9**	**0.1**	**0.4**	**0.1**	**0.6**
Failed allocation[c]	**0.0**	**0.0**	**15.7 (0.1)**	**0.0**	**0.6**
Random individuals (*N* = 1000)					
Correct rejection[d]	**90.8**	**99.7**	**98.5**	**100.0**	**100.0**
False acceptance[e]	**9.2**	**0.3**	**1.5**	**0.0**	**0.0**
Overall success (*N*=2000)	**92.4**	**99.8**	**91.2**	**>99.9**	**99.4**

[a]Correct allocation in exclusion method is defined as the case where offspring for whom the true parents (both parents) were determined at N_{MM} = 0 with no allocation of false parent at any of the N_{MM}s. In likelihood analysis, it is defined as the case where offspring for whom no candidate parent (parental pair) other than the true parents (parental pair) was accepted at given threshold LOD (LOD_C).

[b]In exclusion, this category includes offspring for whom the true parents were determined at N_{MM} = 0 while one or more false parents were assigned at N_{MM} = 1 or 2. In likelihood analysis, offspring to whom the true parents (parental pair) plus false parents (parental pairs) were assigned are put in this category.

[c]This category includes offspring for whom the true parent-offspring relationship was rejected (rejection of either or both the true parents in exclusion and *single parent search*; rejection of true parental pair in *parental pair search*). In *single parent search*, percentage of offspring for whom both parents were rejected is shown in parenthesis.

[d]In exclusion, this category includes random individuals to whom no parent was assigned at any of the N_{MM}s. In likelihood analysis, it includes individuals to whom no parent (parental pair) was assigned.

[e]A category for the case where false parentage was accepted. In exclusion, all the cases occurred at N_{MM} =1 or 2 excepting one instance (at N_{MM} = 0).

7. Glossaries

Shown below are the mathematical formulas used in PARFEX (autosomal, co-dominant and unlinked markers are assumed). These are brief descriptions. For details, please refer to the literature cited.

Heterozygosity

In a population sample, the observed heterozygosity (H_{obs}) is simply obtained as the number of observed heterozygotes divided by the sample size (N). Unbiased estimate of expected heterozygosity (a.k.a. gene diversity) is calculated following Nei (1987):

$$H_{\mathrm{exp}} = \frac{2N}{2N-1}\left(1-\sum_{i=1}^{k} {p_i}^2\right),$$

where p_i is the frequency of i th allele and k is the number of different alleles.

Polymorphism information content (PIC)

The following expression comes from Hildebrand et al. (1992):

$$\mathrm{PIC} = 1-\sum_{i=1}^{k} {p_i}^2 - \left(\sum_{i=1}^{k} {p_i}^2\right)^2 + \sum_{i=1}^{k} {p_i}^4,$$

where p_i is the frequency of i th allele and k is the number of different alleles.

Exclusion probability

There are three types of exclusion probability: paternity exclusion (one parent exclusion; denoted by ExclPP *ad hoc* in PARFEX), exclusion with one parental genotype unknown (ExclP1) and exclusion for both parents (ExclP2). These can be calculated based on powers of population allele frequencies (Jamieson & Taylor 1997):

$$\mathrm{ExclPP} = 1-2\sum_{i=1}^{k} {p_i}^2 + \sum_{i=1}^{k} {p_i}^3 + 2\sum_{i=1}^{k} {p_i}^4 - 3\sum_{i=1}^{k} {p_i}^5 - 2\left(\sum_{i=1}^{k} {p_i}^2\right)^2 + 3\sum_{i=1}^{k} {p_i}^2 \sum_{i=1}^{k} {p_i}^3,$$

$$\mathrm{ExclP1} = 1-4\sum_{i=1}^{k} {p_i}^2 + 2\left(\sum_{i=1}^{k} {p_i}^2\right)^2 + 4\sum_{i=1}^{k} {p_i}^3 - 3\sum_{i=1}^{k} {p_i}^4,$$

$$\mathrm{ExclP2} = 1+4\sum_{i=1}^{k} {p_i}^4 - 4\sum_{i=1}^{k} {p_i}^5 - 3\sum_{i=1}^{k} {p_i}^6 - 8\left(\sum_{i=1}^{k} {p_i}^2\right)^2 + 8\sum_{i=1}^{k} {p_i}^2 \sum_{i=1}^{k} {p_i}^3 + 2\left(\sum_{i=1}^{k} {p_i}^3\right)^2,$$

where p_i is the frequency of i th allele and k is the number of different alleles.

For each type, combined exclusion probability over M markers is obtained by:

$$\mathrm{ExclP}_{\mathrm{Total}} = 1 - \prod_{m=1}^{M} (1 - P_m),$$

where P_m is the exclusion probability of mth marker (Jamieson & Taylor 1997). This calculation is not available in PARFEX, but it is easy to get the estimate using the 'PRODUCT' function of EXCEL.

Exact test for Hardy-Weinberg equilibrium (HWE)

In PARFEX, a conventional Monte Carlo exact test is used in HWE analysis (Guo & Thompson 1992).

In a sample with the size N taken from a population, we observe that a marker has k different alleles (a_1, a_2, ..., a_k) with the allele counts of (n_1, n_2, ..., n_k). By defining that g_{ij} represents the count of genotype $a_i a_j$ (a_i and a_j is i th and j th allele: $1 \le i \le j \le k$) and g = (g_{11}, g_{12}, g_{21}, g_{22}, ..., g_{kk}), the probability of observing g under HWE can be expressed as (e.g. Guo & Thompson 1992; Kalinowski 2006):

$$\mathrm{Prob}(g \mid n_1, n_2, \ldots, n_k) = \frac{N! \prod_{i=1}^{k} n_i! \left(2^{Het}\right)}{(2N)! \prod_{i \le j} g_{ij}!},$$

where $Het = \sum_{i<j} g_{ij}$ (i.e., total number of heterozygotes). The $\mathrm{Prob}(g)$ is compared against a probability distribution generated by random pairing of alleles with fixed marginal allele counts of (n_1, n_2, ..., n_k). Following the recommendation by Guo & Thompson (1992), a batching method is used (17×10^3 randomizations consisting of 100 batches and 170 randomizations per batch) so that P value as well as its standard error is calculated.

For a large sample size, the method is known to be very inefficient in terms of computational time (Guo & Thompson 1992), but it was out of our intention to dig into the efficiency of HWE testing. If users prefer, they can use more improved and efficient methods. Please consult other dedicated studies (e.g., Engels 2009 and references therein) or excellent software (e.g., GENEPOP or GENEPOP on the Web: Rousset 2008; ARLEQUIN: Excoffier et al. 2005).

LOD score (likelihood-based parentage allocation)

The following descriptions summarize excerptions from several articles (Meagher & Thompson 1986; Marshall et al. 1998; Gerber et al. 2000; Jones & Ardren 2003; Morrissey

& Wilson 2005; Kalinowski et al. 2007).

There are three individuals, O with the genotype g_{O}, A (g_{A}) and B (g_{B}), among which we consider three genealogical relationships:

1. The likelihood that the individuals are unrelated is expressed as

   $L(\mathrm{UR}) = P(g_{\mathrm{O}}) \cdot P(g_{\mathrm{A}}) \cdot P(g_{\mathrm{B}})$,

   where $P(g_i)$ is the frequency of genotype g_i in a random mating population.
2. The likelihood that A is the true parent of O but B is not is

   $L(\mathrm{P}) = T(g_{\mathrm{O}} | g_{\mathrm{A}}) \cdot P(g_{\mathrm{A}}) \cdot P(g_{\mathrm{B}})$,

   where T represents Mendelian transition probability of offspring genotype (g_{O}) given the parental genotype g_{A}.
3. The likelihood that A and B are the true parental pair of O is

   $L(\mathrm{PP}) = T(g_{\mathrm{O}} | g_{\mathrm{A}}, g_{\mathrm{B}}) \cdot P(g_{\mathrm{A}}) \cdot P(g_{\mathrm{B}})$

Therefore, the likelihood ratio of related relationship to unrelated relationship for single parent, $L(\mathrm{P})/L(\mathrm{UR})$, can be written as

$$\mathrm{L(single)} = \frac{T(g_{\mathrm{O}} \mid g_{\mathrm{A}})}{P(g_{\mathrm{O}})}$$

and for parental pair, $L(\mathrm{PP})/L(\mathrm{UR})$, as

$$\mathrm{L(pair)} = \frac{T(g_{\mathrm{O}} \mid g_{\mathrm{A}}, g_{\mathrm{B}})}{P(g_{\mathrm{O}})}$$

The likelihood ratio estimated for each marker is multiplied over markers. A LOD score over markers ($\mathrm{LOD_p}$, single parent LOD; $\mathrm{LOD_{pp}}$, parental pair LOD) is obtained by taking the natural logarithm (ln) of the multiplied product.

Mendelian transition probabilities between offspring and parental genotypes are concisely summarized in tables 1 and 2 of Marshall et al. (1998). The $\mathrm{LOD_p}$ adopted in PARFEX assumes that no information about the other parent is available. That is, it is different from that of typical paternity testing (see Marshall et al. 1998).

Marshall et al. (1998) incorporated genotypic error rate (e) into the likelihood equations based on the random genotype replacement model. However, those equations were reformulated later by Kalinowski et al. (2007) and its corrigendum, Kalinowski et al. (2010). PARFEX calculates LOD scores based on the following likelihood ratio derived from the revised likelihood equations (Kalinowski et al. 2007):

$$\mathrm{L(single)} = \frac{(1-e)^2 T(g_O \mid g_A) + 2e(1-e)\cdot P(g_O) + e^2 P(g_O)}{(1-e)^2 P(g_O) + 2e(1-e)P(g_O) + e^2 P(g_O)}$$

and

$$\mathrm{L(pair)} = \frac{(1-e)^3 T(g_O \mid g_A, g_B) + e(1-e)^2 [T(g_O \mid g_A) + T(g_O \mid g_B) + P(g_O)] + 3e^2(1-e)P(g_O) + e^3 P(g_O)}{(1-e)^3 P(g_O) + 3e(1-e)^2 P(g_O) + 3e^2(1-e)P(g_O) + e^3 P(g_O)}$$

A LOD score over markers is obtained in the same manner as described above. When e is assumed to be zero, LOD score can be calculated only for non-excluded relationships (see the likelihood ratios without error shown above). In PARFEX likelihood analyses with an e_{calc} = 0, therefore, LOD calculation is done only for non-excluded combinations between putative parent (parental pair) and offspring, which have no mismatched markers. For the same reason, LOD calculation for POP_{m+f} *vs* POP_{rand} is omitted in simulation analysis to determine a LOD_C as it is not possible to get a meaningful LOD distribution.

In LOD_p, when a putative parent gives a positive LOD score against offspring, the putative parent is more likely to have true genealogical relationship with the offspring than are other individuals randomly selected from the parental (panmictic) population. A LOD score of zero indicates that the putative parent and other randomly-drawn individuals are equally likely to be the true parent of the offspring. If the putative parent is less likely to be the true parent of the offspring compared with other randomly-drawn individuals, the LOD score should become negative. The same logic is applied to LOD_{pp}, but the focus of interest is of ‘parental pair’.

8. Literature cited

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* 32: 314−331

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC, Sutherland GR (1993) Incidence and origin of "null" alleles in the $(AC)_n$ microsatellite markers. *American Journal of Human Genetics* 52: 922−927

Coombs JA, Letcher BH, Nislow KH (2008) CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs. *Molecular Ecology Resources* 8: 578−580

Ede AJ, Crawford AM (1995) Mutations in the sequence flanking the microsatellite at the KAP8 locus prevent the amplification of some alleles. *Animal Genetics* 26: 43−44

Engels WR (2009) Exact test for Hardy-Weinberg Proportions. *Genetics* 183: 1431−1441

Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1: 47−50

Gerber S, Mariette S, Streiff R, Bodénès C, Kremer A (2000) Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. *Molecular Ecology* 9: 1037−1048

Gerber S, Chabrier P, Kremer A (2003) FaMoz: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. *Molecular Ecology Notes* 3: 479−481

Glaubitz JC (2004) CONVERT: a user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Molecular Ecology Notes* 4: 309−310

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48: 361−372

Hildebrand CE, Torney DC, Wagner R (1992) Informativeness of polymorphic DNA markers. *Los Alamos Science* 20: 100−102

Jamieson A, Taylor St.CS (1997) Comparisons of three probability formulae for parentage exclusion. *Animal Genetics* 28: 397−400

Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. *Molecular Ecology* 12: 2511−2523

Jones AG, Small CM, Paczolt KA. Ratterman NL (2010) A practical guide to methods of parentage analysis. *Molecular Ecology Resources* 10: 6−30

Kalinowski ST (2006) HW-QUICKCHECK: an easy-to-use computer program for checking genotypes for agreement with Hardy-Weinberg expectations. *Molecular Ecology Notes* 6: 974−979

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology* 16: 1099−1106

Kalinowski ST, Taper ML, Marshall TC (2010) Corrigendum. *Molecular Ecology* 19: 1512

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology 7: 639−655

Meagher TR, Thompson E (1986) The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology* 29: 87−106

Morrisey MB, Wilson AJ (2005) The potential costs of accounting for genotypic errors in molecular parentage analyses. *Molecular Ecology* 14: 4111−4121

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York, NY

Pemberton JM, Slate J, Bancroft DR, Barrett A (1995) Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. *Molecular Ecology* 4: 249−252

O'Reilly PT, Herbinger C, Wright JM (1998) Analysis of parentage determination in Atlantic salmon (*Salmo salar*) using microsatellites. *Animal Genetics* 29: 363−370

Rousset F (2008) Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources* 8: 103−106

Thompson EA (1976) Inference of genealogical structure. *Social Science Information* 15: 477−526