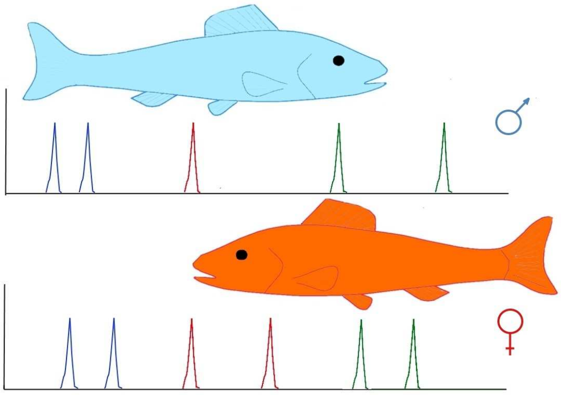
GENASSEMBLAGE 2.2



Maintaining genetic variation within broodstock is important for successful fish farming and successful conservation of hatchery-dependent species. Breeders, as well scientists involved in producing juveniles, should assemble spawning pairs using fish that are as genetically different as possible. Unfortunately, the genetic variation within hatchery dependent populations can decrease because of progressive elimination of allelic diversity from the genomes of individuals in the broodstocks (Koljonen et al. 1999, Koljonen et al.. 2002, Verspoor 2005). This decrease in genetic variation might result from obtaining a large group of juveniles from a few parental individuals, possible inbreeding events (Bryant et al. 1986, Kosowska and Nowicki 1999) or domestication of the broodstock, which may eliminate individuals with alleles present in the natural population (Kim et al. 1994). Molecular genetics offers a tool that has proven useful in assessment and maintenance of genetic diversity (Koljonen et al. 2002). On the basis of highly polymorphic fragments of DNA, such as microsatellites, individual genetic profiles of spawners can be prepared. These profiles can be used to identify possible spawners and assemble them in spawning pairs. Unfortunately, there is no computer-based tool available for identification of the best spawning pairs in large databases of genetic profiles of spawners. For this reason, Genassemblage software was constructed to help identify the best possible combinations of spawners within and between broodstocks and to manage genetic variation deposited in banks of cryopreserved gametes.

Table of contents

1. Program information

The creation of Genassemblage 2.2 software was supported by the National Science Centre in Poland as part of project No. 2014/15/B/NZ9/05240 for 2015-2019.

The author and holder of the license:

Dr. Dariusz Kaczmarczyk

Department of Environmental Biotechnology

University of Warmia and Mazury in Olsztyn

Słoneczna 45G 10-718 Olsztyn, Poland

Tel. (+48) 89 523-41-62

e-mail: d.kaczmarczyk@uwm.edu.pl

Ownership rights to Genassemblage are held by the

Inland Fisheries Institute in Olsztyn

M. Oczapowskiego 10

10-719 Olsztyn, Poland

The author of Genassemblage 2.2 grants consent to download the program free of charge from the author's website http://pracownicy.uwm.edu.pl/d.kaczmarczyk/main_page.htm and to use the program:

- for scientific research,
- as a tool for protection of populations and management of their genetic variation,
- as a tool in recreational or commercial (e.g. aquaculture) animal breeding,
- in education associated with biology, genetics, environmental protection or aquaculture.

Contact the author if you intend to use the program for other purposes than those mentioned above.

Whenever the program is used, please cite this paper:

Kaczmarczyk D**.** 2015. Genassemblage software, a tool for management of genetic diversity in human dependent populations. *Conservation Genetic Resources* 7:49-51
doi: 10.1007/s12686-014-0356-8

The author does not grant consent to:

- modify of any components of the program,
- put up the program for download on any other website than the author’s website.

The program was produced by:

Moonscape Software Grzegorz Janik
Mieszka I 11 / 21
26-617 Radom

2. Program description

Genassemblage 2.2 is an improved version of Geneassemblage 1.0 (Kaczmarczyk 2016). This software tool is designed to manage genetic variation in commercial stocks or human dependant populations. It enables selection of parent individuals based on their genetic characteristics and management of genetic variation deposited in banks of cryopreserved gametes. It allows analysis of diploid, tetraploid and partially tetraploid organisms. The program enables identification of the best variants of breeding pairs based on the predicted genetic variation of their offspring, and estimation of the effect of different variants of matched parental individuals based on the predicted genetic variation of the subsequent generation. Selection of parental individuals is optimized based on individual genetic characteristics (genetic profiles) determined for each individual. The profiles contain the name of the population or herd to which the individual belongs, the name or number of the individual and the list of alleles found within the analyzed genetic markers. The markers used in Genassemblage should have the following characteristics:

A. high polymorphism,

B. autosomal chromosome inheritance in accordance with Mendelian Laws, and

C. evolutionary neutrality (not subject to natural selection).

Individuals are selected for mating couples based on the following criteria and assumptions:

A. The sex (male or female) of the individuals is known, and each individual is fertile and ready for reproduction.

B. All individuals are marked and identifiable.

C. The probability of occurrence of two identical genotypes among unrelated individuals is negligible.

D. Individuals with a higher frequency of heterozygosity within the analyzed genetic markers are more likely to differ in loci that determine population viability and adaptability.

Samples stored in banks of cryopreserved gametes should meet the criteria given below:

A. The type of gametes, i.e. sperms or eggs, has been determined.

B. Genetic profiles of gamete donors are known, and follow the criteria for genetic markers given above.

C. All samples are marked and identifiable.

3. List of changes in Genassemblage version 2.2, compared to version 1.0.

1. A new, clearer and more user-friendly interface.

2. Easy to use design that consists of 4 modules.

3. A new module for selecting individuals for group spawning

4. A new module for management of genetic variation in a gamete bank

5. A redesigned tool for conversion of files to *.dat format

6. Introduction of the v index to the module for selecting the best breeding pair

7. Improved calculation and presentation of the values of heterozygosity, share of week heterozygotes and number of alleles in potential progeny of breeding pairs.

8. Improved mechanism for conversion of input files used in Genassemblage to *.arp files.

9. Compatibility with versions of MS Excel newer than 2013.

4. Application of the program

Genassemblage has been developed for species whose offspring is produced in controlled conditions (e.g. a hatchery). The software can be used to maintain the genetic diversity of broodstocks in commercial breeding and to conserve endangered species. It can also be used to reduce the probability of inbreeding or for educational purposes.

Genassemblage can be used for species with a diploid genome, for those with a tetraploid genome, (e.g. Acipenseridae fish), and these with a diploid genome containing tetraploid fragments, e.g. cyprinids, salmonids and some Acipenseridae fish.

The program can be used to convert *.xls, *.xlsx or *.dat files that contain genotyping data to to *.arp. files.

The program enables estimation of genetic variance indexes, such as heterozygosity (also the percentage of "weak heterozygotes" for tetrasomic loci) and the number of different alleles potentially inherited by offspring of selected breeding pairs. The term "weak heterozygote" denotes individuals with three identical alleles and a fourth allele that differs at a tetrasomic locus, e.g. AAAB (Kaczmarczyk and Fopp-Bayat 2012). The indices calculated for each individual are presented together in tables; it is possible to choose parental combinations whose offspring will have the best values of these indexes by using the *v* index.

A new functionality in Genassemblage 2.0 is that it can be used to identify a set of individuals that will be optimal for group breeding, resulting in the highest possible allelic diversity and heterozygosity in their progeny. Moreover, Genassemblage 2.0 can be used to manage genetic variation resources deposited in banks of cryopreserved gametes in order to obtain a target level of genetic variation in the next generation.

5. System requirements

1. The program operates in the Windows environment (it requires Microsoft NET Framework 4.7 or higher) with Microsoft Excel 2003 and later versions.
2. A file with input data should be created in MS Excel following the procedure presented in Chapter 8.
3. The upper limit of the size of the analyzed set of depends on version of MS Excel used. There is no limit on number of analyzed markers.
4. In order to calculate the share of individuals that are "weak heterozygotes", the program requires a set of input data containing the results of genotyping of at least one tetrasomic locus in the analyzed breeding couple. If the data contain tetrasomic loci, they can take any position in the sequence of analyzed loci and the ratio of tetrasomic to disomic loci can have any value.

We interested in improve a performance of our software. Please use the newest version of our Genassemblage software. If you find any bug we will be graceful for reporting it.

6. Program installation

The Genassemble installer in version .msi or .zip archive can be downloaded from the author's webpage http://pracownicy.uwm.edu.pl/d.kaczmarczyk/main_page.htm

The program requires Microsoft NET Framework ".NET", version 4.7 or later. If it has not been installed, please download it from the Microsoft website and install it.

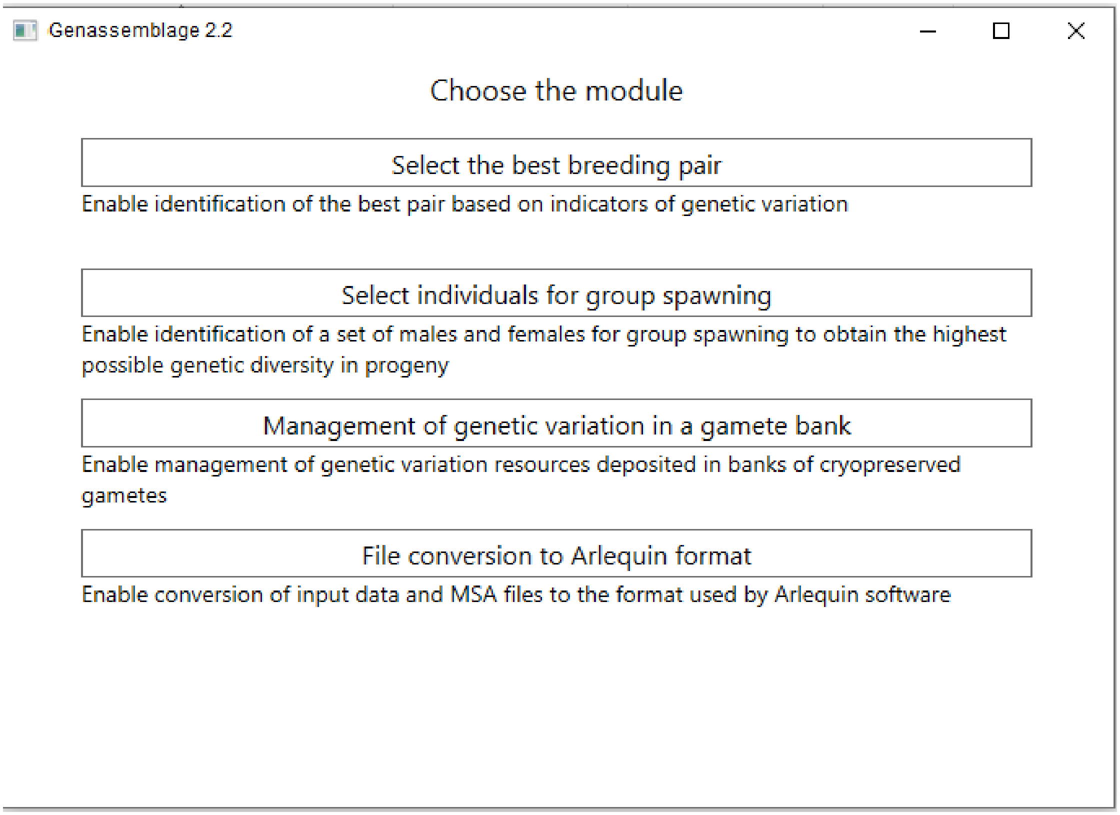
Before installing Genassemblage, MS Excel version 2003 or later must be installed on your computer.

After the Microsoft NET Framework package has been installed, click Genassemblage_2.0_installer.msi (for install 2.0 of the Genassemblage software). If you are going install version 2.01 you should first unzip zip archive and then cilck on Genassemblage_2.01_installer.msi or setup.exe. Next specify the target location of the program on your hard drive. After finishing the installation, the program is ready to use.

7. Interface

The Genassemblage 2.2 and its previous 2.02 version slightly differs in details of interface and result files. All pictures presented below are actual for 2.2 version.

The Genassemblage software can be launched by clicking Genassemblage.exe the following dialogue box (Figure 1).


Genassemblage 2.2
Choose the module
Select the best breeding pair
Select individuals for group spawning
Management of genetic variation in a gamete bank
File conversion to Arlequin format

Enable identification of the best pair based on indicators of genetic variation

Enable identification of a set of males and females for group spawning to obtain the highest possible genetic diversity in progeny

Enable management of genetic variation resources deposited in banks of cryopreserved gametes

Enable conversion of input data and MSA files to the format used by Arlequin software

Figure 1. Genassemblage 2.0 main menu.

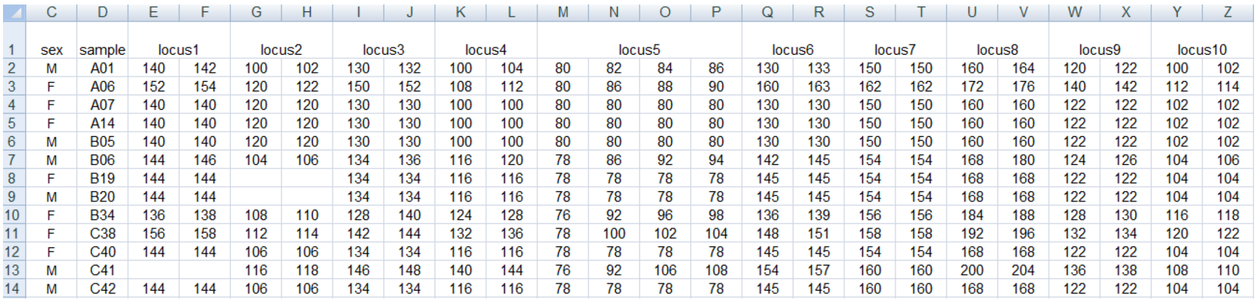
The four bars in the dialogue launch the specified modules, and a short description of each module is provided below the bar. Click on the respective bar to launch a module.

8. Input files

The input files to Genassemblage 2.0 differs across the modules. Their examples can be downloaded from authors' website and modified by users. If you use MS Excell 2007 or latter the input file extension .xlsx is recommended. If you use MS Excell 2003 please use .xls extension.

8.a. Input file for module 1 and 2.

To be used in module Select best breeding pairs and module Select individuals for group spawning an input file should be created in MS Excel in accordance with example (Figure 2) should be prepared.



	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	sex	sample	locus1		locus2		locus3		locus4		locus5				locus6		locus7		locus8		locus9		locus10	
2	M	A01	140	142	100	102	130	132	100	104	80	82	84	86	130	133	150	150	160	164	120	122	100	102
3	F	A06	152	154	120	122	150	152	108	112	80	86	88	90	160	163	162	162	172	176	140	142	112	114
4	F	A07	140	140	120	120	130	130	100	100	80	80	80	80	130	130	150	150	160	160	122	122	102	102
5	F	A14	140	140	120	120	130	130	100	100	80	80	80	80	130	130	150	150	160	160	122	122	102	102
6	M	B05	140	140	120	120	130	130	100	100	80	80	80	80	130	130	150	150	160	160	122	122	102	102
7	M	B06	144	146	104	106	134	136	116	120	78	86	92	94	142	145	154	154	168	180	124	126	104	106
8	F	B19	144	144			134	134	116	116	78	78	78	78	145	145	154	154	168	168	122	122	104	104
9	M	B20	144	144			134	134	116	116	78	78	78	78	145	145	154	154	168	168	122	122	104	104
10	F	B34	136	138	108	110	128	140	124	128	76	92	96	98	136	139	156	156	184	188	128	130	116	118
11	F	C38	156	158	112	114	142	144	132	136	78	100	102	104	148	151	158	158	192	196	132	134	120	122
12	F	C40	144	144	106	106	134	134	116	116	78	78	78	78	145	145	154	154	168	168	122	122	104	104
13	M	C41			116	118	146	148	140	144	76	92	106	108	154	157	160	160	200	204	136	138	108	110
14	M	C42	144	144	106	106	134	134	116	116	78	78	78	78	145	145	160	160	168	168	122	122	104	104

Figure 2. Format of an input file for module 1 and 2 of the Genassemblage 2.2

Columns and rows description:

Columns A and B – definition of terms: A population is a community of potentially interbreeding animals, such as all the members of one species of fish in a particular pond. A population grouping is chosen by the operator of Genassemblage and could be, for example, all the populations of a species of fish in a particular region or country.

Column A – the value here defines the group to which the population in column B belongs. If all the populations analysed in the input file belong to one group, then the value in this column = 1. If the analysed populations belong to a larger number of groups, then they should be assigned numbers according to where they belong, beginning 1, 2, 3, etc.,

Column B - the name of the population to which an individual belongs,

Column C - sex of the individuals (M - male, F - female),

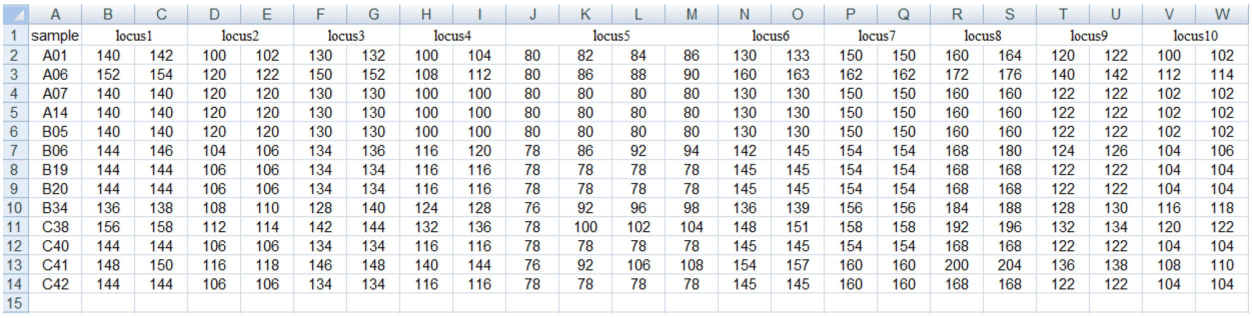
Column D - marking of the individuals, e.g. the tag numbers,

Columns E, F... etc. analyzed loci and their alleles in individual samples, row 1 (header) is the name of the locus. If the locus is tetrasomic, the header should include 4 consecutive columns in which alleles of the gene are situated. The program automatically detects whether the locus is disomic or tetrasomic; therefore their sequence in the input file and their proportions are not specified. If an individual has an incomplete set of data, blank cells should be left in the locus for which there are no data (like in the example, sample C41, locus 1).

8.b Input file for module 3.

The module Management the genetic variation in gamete bank require a input file formatted

as shown of the Figure 3.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	sample	locus1		locus2		locus3		locus4		locus5				locus6		locus7		locus8		locus9		locus10	
2	A01	140	142	100	102	130	132	100	104	80	82	84	86	130	133	150	150	160	164	120	122	100	102
3	A06	152	154	120	122	150	152	108	112	80	86	88	90	160	163	162	162	172	176	140	142	112	114
4	A07	140	140	120	120	130	130	100	100	80	80	80	80	130	130	150	150	160	160	122	122	102	102
5	A14	140	140	120	120	130	130	100	100	80	80	80	80	130	130	150	150	160	160	122	122	102	102
6	B05	140	140	120	120	130	130	100	100	80	80	80	80	130	130	150	150	160	160	122	122	102	102
7	B06	144	146	104	106	134	136	116	120	78	86	92	94	142	145	154	154	168	180	124	126	104	106
8	B19	144	144	106	106	134	134	116	116	78	78	78	78	145	145	154	154	168	168	122	122	104	104
9	B20	144	144	106	106	134	134	116	116	78	78	78	78	145	145	154	154	168	168	122	122	104	104
10	B34	136	138	108	110	128	140	124	128	76	92	96	98	136	139	156	156	184	188	128	130	116	118
11	C38	156	158	112	114	142	144	132	136	78	100	102	104	148	151	158	158	192	196	132	134	120	122
12	C40	144	144	106	106	134	134	116	116	78	78	78	78	145	145	154	154	168	168	122	122	104	104
13	C41	148	150	116	118	146	148	140	144	76	92	106	108	154	157	160	160	200	204	136	138	108	110
14	C42	144	144	106	106	134	134	116	116	78	78	78	78	145	145	160	160	168	168	122	122	104	104
15																							

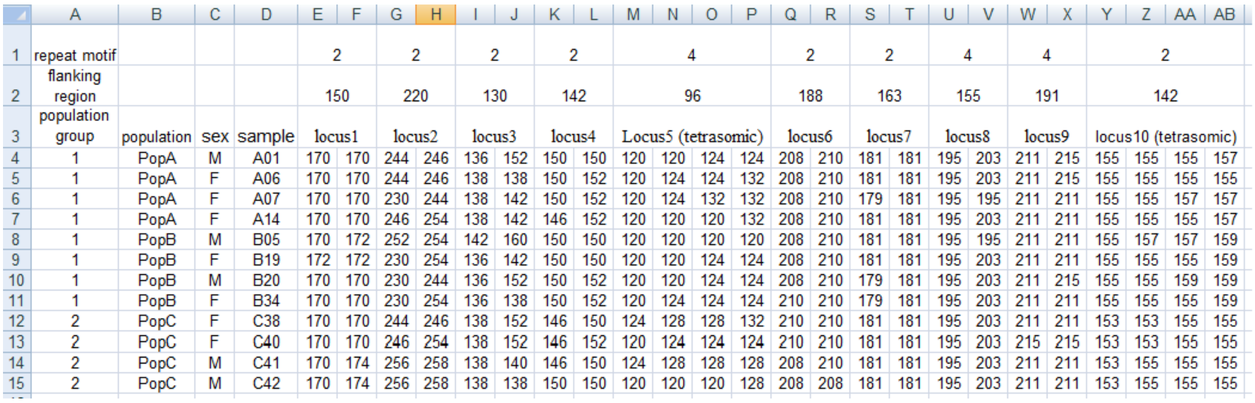
Figure 3. Format of an input file for module 3 of the Genassemblage 2.2

Column A (sample) - marking of the individuals, e.g. the tag numbers,

Columns B, C,... etc. analyzed loci and their alleles in individual samples, row 1 (header) is the name of the locus. If the locus is tetrasomic, the header should include 4 consecutive columns in which alleles of the gene are situated. The program automatically detects whether the locus is disomic or tetrasomic; therefore their sequence in the input file and their proportions are not specified.

8.c. Input files for conversion to the .arp format

The input file intended for conversion to the .arp format should not include tetrasomic markers; if it does the program divides them into two virtual disomic isoloci, which are inherited independently. The basis for this conversion is input data with two lines added. Line 1 include repeat motif of the microsatellite motif and line 2 length of the flanking region of the microsatellite DNA. Example of input file is given at Figure 4.

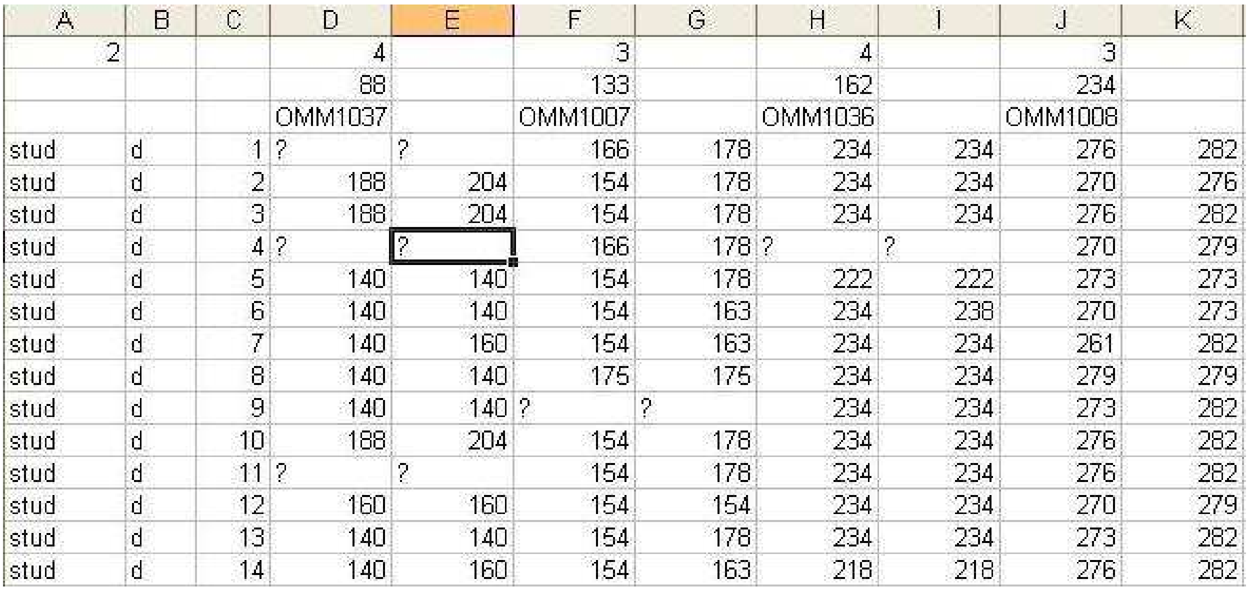


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	repeat motif				2		2		2		2		4				2		2		4		4		2			
2	flanking region				150		220		130		142		96				188		163		155		191		142			
3	population group	population	sex	sample	locus1		locus2		locus3		locus4		Locus5 (tetrasomic)				locus6		locus7		locus8		locus9		locus10 (tetrasomic)			
4	1	PopA	M	A01	170	170	244	246	136	152	150	150	120	120	124	124	208	210	181	181	195	203	211	215	155	155	155	157
5	1	PopA	F	A06	170	170	244	246	138	138	150	152	120	124	124	132	208	210	181	181	195	203	211	215	155	155	155	155
6	1	PopA	F	A07	170	170	230	244	138	142	150	152	120	124	132	132	208	210	179	181	195	195	211	211	155	155	157	157
7	1	PopA	F	A14	170	170	246	254	138	142	146	152	120	120	120	132	208	210	181	181	195	203	211	211	155	155	155	157
8	1	PopB	M	B05	170	172	252	254	142	160	150	150	120	120	120	120	208	210	181	181	195	195	211	211	155	157	157	159
9	1	PopB	F	B19	172	172	230	254	136	142	150	150	120	120	124	124	208	210	181	181	195	203	211	211	155	155	155	159
10	1	PopB	M	B20	170	170	230	244	136	152	150	152	120	120	124	124	208	210	179	181	195	203	211	215	155	155	159	159
11	1	PopB	F	B34	170	170	230	254	136	138	150	152	120	124	124	124	210	210	179	181	195	203	211	211	155	155	155	159
12	2	PopC	F	C38	170	170	244	246	138	152	146	150	124	128	128	132	210	210	181	181	195	203	211	211	153	153	155	155
13	2	PopC	F	C40	170	170	246	254	138	152	146	152	120	124	124	124	210	210	181	181	195	203	215	215	153	153	155	155
14	2	PopC	M	C41	170	174	256	258	138	140	146	150	124	128	128	128	208	210	181	181	195	203	211	211	153	155	155	155
15	2	PopC	M	C42	170	174	256	258	138	138	150	150	120	120	120	128	208	208	181	181	195	203	211	211	153	155	155	155

Figure 4. Example of input file for conversion from .xls and .xlsx file to the .arp file.

8.d. Input file for conversion of .dat file fto .arp format.

A .dat file intended for conversion should be formatted in accordance with the requirements specified in the instructions for the MSA (Dieringer and Schlötterer 2003) programe, there is an example of such a file called microsatellite-example.dat, in the Genassemblage 2.0 folder (saved as a text file with tab characters as separators). If a .dat file has a different structure than the input file is used for MSA, it should be rebuilt as described above and on Figure 5.



A	B	C	D	E	F	G	H	I	J	K
2			4		3		4		3	
			88		133		162		234	
			OMM1037		OMM1007		OMM1036		OMM1008	
stud	d	1	?	?	166	178	234	234	276	282
stud	d	2	188	204	154	178	234	234	270	276
stud	d	3	188	204	154	178	234	234	276	282
stud	d	4	?	?	166	178	?	?	270	279
stud	d	5	140	140	154	178	222	222	273	273
stud	d	6	140	140	154	163	234	238	270	273
stud	d	7	140	160	154	163	234	234	261	282
stud	d	8	140	140	175	175	234	234	279	279
stud	d	9	140	140	?	?	234	234	273	282
stud	d	10	188	204	154	178	234	234	276	282
stud	d	11	?	?	154	178	234	234	276	282
stud	d	12	160	160	154	154	234	234	270	279
stud	d	13	140	140	154	178	234	234	273	282
stud	d	14	140	160	154	163	218	218	276	282

Figure 5. An example of a .dat input file usable for conversion to the .arp format:

Column A – name of the population, column B - a disomic model of inheritance, column C – name, e.g. individual's tag number, Columns D-K the examined microsatellite sections, their properties and the allele's length in bp (row 4 and below). Columns D, F, H, J lengths of the motifes of basic microsatellite sequences (row 1), length of flanking regions of the microsatellite in bp (row 2), name of the locus (row 3). The file should be saved in the text file format, with tab characters as separators, and it should have a .dat extension.

9. Module 1. Select best breeding pairs

This module is designed to identify best variants of breeding pairs. After clicking a bar in main menu the window of this module (Figure 6) will appear.

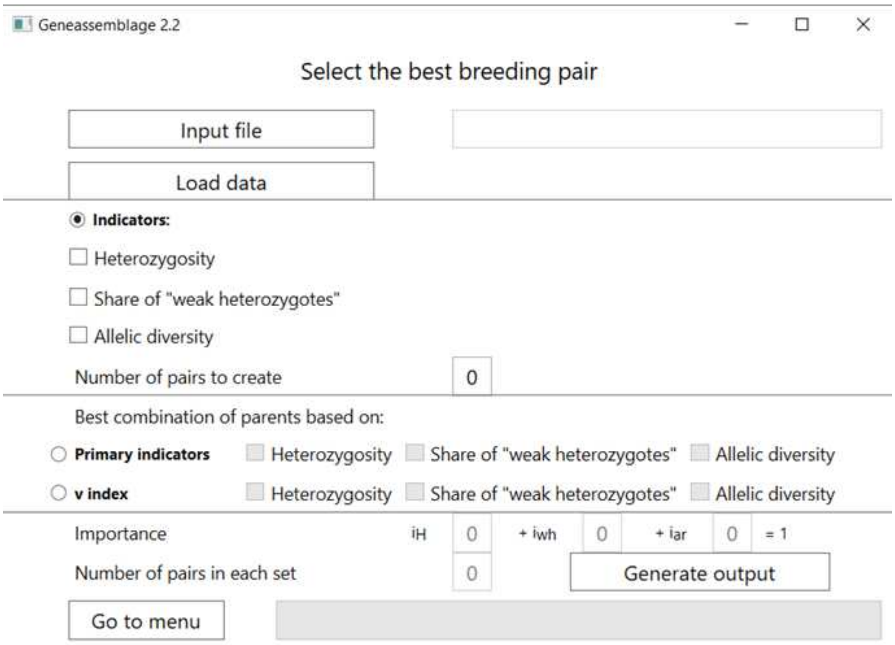

Geneassemblage 2.2
Select the best breeding pair
Input file
Load data
● Indicators:
□ Heterozygosity
□ Share of "weak heterozygotes"
□ Allelic diversity
Number of pairs to create 0
Best combination of parents based on:
○ Primary indicators □ Heterozygosity □ Share of "weak heterozygotes" □ Allelic diversity
○ v index □ Heterozygosity □ Share of "weak heterozygotes" □ Allelic diversity
Importance iH 0 + iwh 0 + iar 0 = 1
Number of pairs in each set 0
Generate output
Go to menu

Figure 6. Main menu of module 1.

They are three possible variants of calculations:

A. When the field **indicators** is marked and number of best values is chosen the program identify and show only best pairings across all data**.** The program will calculate the indicator such Heterozygosity share of weak heterozygotes and allelic diversity if those indicators will be marked by the user. The number of presented values will be equal the value introduced in the field **Numbers of pairs to create**.

B. When the field **primary indicators** is marked the program will calculate the indicatros chosen by the user for all variants of breeding pairs. **The user can chose what indicators will be included in the calculation.**

C. When section **v index** is chosen the program will calculate a the values of indicators chosen by the user and indicate a group of best pairs basing on the **v index.** To perform those calculations the **importance** of each component of the v index must be introduced by the user**.**

9.a. Mathematical methods

Calculation of expected heterozygosity of offspring

This value is calculated from disomic and tetrasomic markers for which there are complete genotyping results for both parental individuals. A locus for which there are data missing from one or both parental individuals is excluded from the calculations for this pair. The expected heterozygosity of the offspring (*H*) is calculated by dividing the sum of the expected shares of homozygous genotypes in the offspring of a specific breeding couple at the first (ph_1) second (ph_2), third (ph_3) and all subsequent (ph_n) loci, by the number of analysed loci (*nl*) then subtracting this value from 1 (Algorithm 1).

The values of probability (*ph*) are calculated assuming that gametes conjugate randomly, and the frequency of individual genotypes in offspring is not changed by natural selection.

[Algorithm 1]

$$H = 1 - \frac{(ph_1 + ph_2 + ph_3 + ...ph_n)}{nl}$$

Algorithm 1. Calculation of expected heterozygosity of offspring

Calculation of the expected percentage of "weak heterozygote" individuals

Individuals described as "weak heterozygotes" have three identical alleles and one different allele in their genotype, e.g. AAAB. The expected percentage of "weak heterozygote" individuals is calculated for tetrasomic markers only. The percentage of "weak heterozygotes" within the total offspring ranges from 0.000 to 0.667 and it is calculated by Algorithm 2, where *wh* is the expected percentage of "weak heterozygous" individuals. The values of probability (*pwh*) are calculated assuming that alleles of the tetrasomic fragment are located on four independently inherited homologous chromosomes, gametes conjugate randomly, and the frequency of individual genotypes in offspring is not changed by natural selection.

[Algorithm 2]

$$wh = \frac{(pwh_1 + pwh_2 + pwh_3 + ...pwh_n)}{nl}$$

Algorithm 2. Calculation of the expected percentage of "weak heterozygotes" Individuals.

A locus in which there are no data in one or both of the potential parents is excluded from calculations.

Calculation potential allelic diversity

To calculate the number of different alleles that potential offspring would inherit, all the alleles found in the markers (na_n) are summed (Algorithm 2). The calculations include all the alleles found within the di- and tetrasomic markers in a specific breeding couple.

[Algorithm 3]

$$ar = \sum (na_n)$$

Algorithm 3. Calculation of potential allelic diversity in the offspring of a specific couple.

Calculation of the *v* index

The *v* index shows a relative quality of given pair for conservation genetic variation. The higher value of the “v index” indicates the best parental combinations that will produce more genetically diverse and heterozygotic progeny and therefore is optimal for enhance purposes.

The user can chose what indicator will be used and weight of each of them. All the indicators cam be used in when tetrasomic loci are in the input file. When working with a diploid species, the share of “weak heterozygotes” is not taken into account in the calculation. The v index is a sum of the three components (*i*): heterozygosity (*i_H)*, allelic diversity *(i_a*) and share of “weak heterozygotes” (*i_{wh})* are their importance of each indicators, multipied by value of division in the brackets. If this was a diploid species, *i_{wh}* impotance would be 0. The *v* index was calculated by using Algorithm 4.

[Algorithm 4]

$$v = i_H \left[\frac{H_n}{H_{\max}} \right] + i_{wh} \left[\frac{wh_{\max} - wh_n}{wh_{\max}} \right] + i_{ar} \left[\frac{ar_n}{ar_{\max}} \right] \qquad \begin{array}{l} i_H + i_{wh} + i_{ar} = 1 \\ wh_{\max} > 0 \end{array}$$

Algorithm 4. Calculation of the *v* index.

In this algorithm, *H_n*, *wh_n*, and *ar_n*, are the values of the indicators of genetic variation expected in the progeny of each potential pairing, and *H_{max}*, *wh_{max}*, and *ar_{max}* are the maximal values of those indicators detected in all analysed pairings. After calculation a v index of each breeding pair the program finds a set of breeding pairs. The set is assembled in this way that same individual can't be present in more than one breeding. This will help to reduce inbreeding in future generations.

The χ^2 statistics describe a scale of differences in the value of the components used to calculate the v index between the group of all possible breeding pairs and set of best pairs chosen by the software. This statistic is based on (Algorithm 5).

[Algorithm 5]

$$\chi^2 = \left(\frac{H_{all} - H_{set}}{H_{all}}\right)^2 + \left(\frac{wh_{all} - wh_{set}}{wh_{all}}\right)^2 + \left(\frac{ar_{all} - ar_{set}}{ar_{all}}\right)^2$$

Algorithm 5. Calculation of the χ^2 statistic.

where:

H_{all} = average heterozygosity of the offspring of all possible breeding pairs

H_{set} = average heterozygosity of the offspring of the selected set of breeding pairs

wh_{all} = average percentage of weak heterozygotes in the offspring of all possible breeding pairs

wh_{set} = average percentage of weak heterozygotes in the offspring of the selected set of breeding pairs

ar_{all} = average diversity of allele offspring of all possible breeding pairs

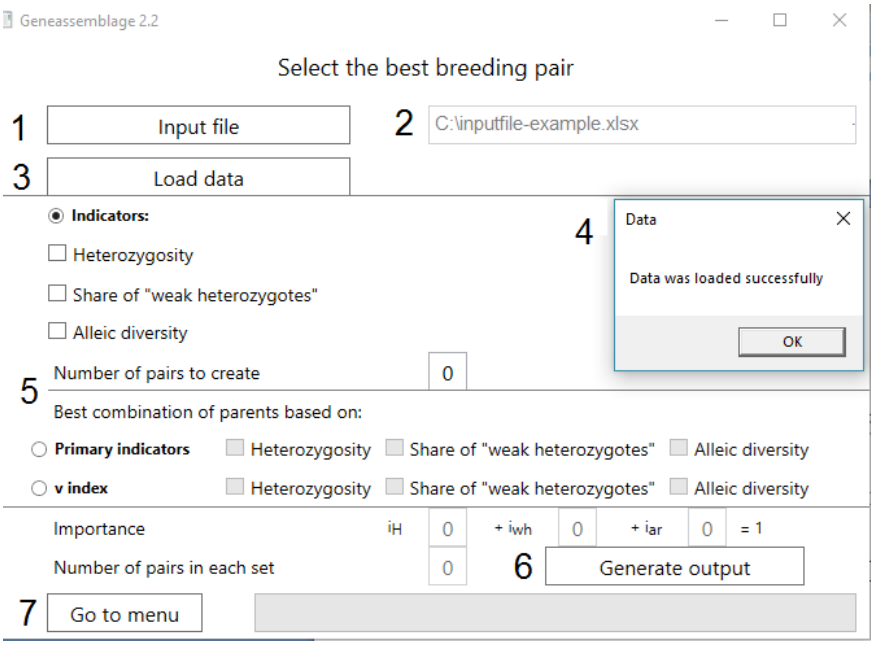
ar_{set} = average diversity of alleles of the offspring of the selected set of breeding couples

Algorithm 5. Method of calculating statistics χ^2.

9.b. Calculation setup

1. After launching the Genassemblage 2.0 click the bar **Select best breeding pair**

2. On the window (Figure 7):



Select the best breeding pair

C:\inputfile-example.xlsx

Data

Data was loaded successfully

● **Indicators:**

□ Heterozygosity

□ Share of "weak heterozygotes"

□ Alleic diversity

Number of pairs to create 0

Best combination of parents based on:

○ **Primary indicators** □ Heterozygosity □ Share of "weak heterozygotes" □ Alleic diversity

○ **v index** □ Heterozygosity □ Share of "weak heterozygotes" □ Alleic diversity

Importance iH 0 + iwh 0 + iar 0 = 1

Number of pairs in each set 0

Figure 7. Module 1, calculation setup.

1. Click the button **Input file**

2. Browse for **the input file**

3. Click the button **Load data**

4. The communicate **Data was loaded successfully** appears if input file is correct and was successfully loaded

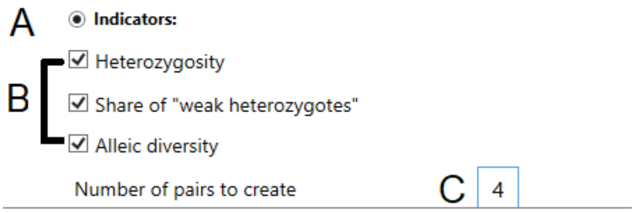
5. Chose and setup one among calculation **variants I-III** specified bellow.

6. Click **Generate output,** select **name** and **path of the output file** and click the button **Write.**

7. If you need to go back to **Genassemblage main menu** click **Go to menu.**

Calculation variants

Variant I. calculation with pointing best values of genetic variation indicators (Figure 8)



● **Indicators:**

☑ Heterozygosity

☑ Share of "weak heterozygotes"

☑ Alleic diversity

Number of pairs to create 4

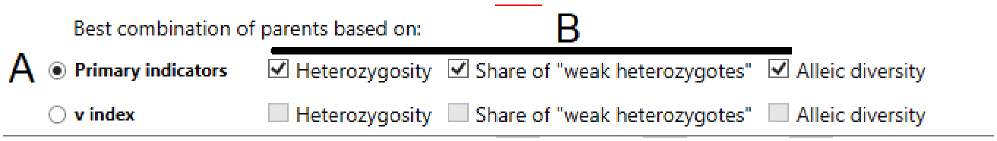
Figure 8. Setup of calculations tsht points to the best values of genetic variation indicators.

A. Mark the field **indicators**

B. Chose the **Indicators** (heterozygosity, share of "weak heterozygotes" or allelic diversity)

C. Introduce **Number of pairs to create**.

Variant II. calculation a values genetic variation indicators in all variants of breedingpairs (Figure 9)



Best combination of parents based on:

● Primary indicators ☑ Heterozygosity ☑ Share of "weak heterozygotes" ☑ Alleic diversity

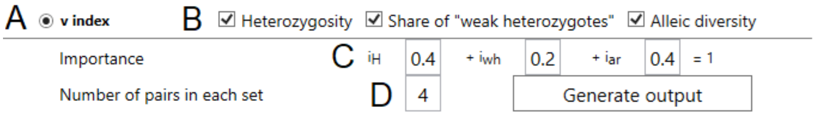
○ v index □ Heterozygosity □ Share of "weak heterozygotes" □ Alleic diversity

Figure 9. Setup of calculation a values genetic variation indicators.

A. Mark the field primary indicators

B. Chose the indicators (heterozygosity, share of "weak heterozygotes" or allelic diversity).

Variant III. Calculation a values genetic variation indicators in all variants of breeding pairs (Figure 10) and comparison by using v index



● v index ☑ Heterozygosity ☑ Share of "weak heterozygotes" ☑ Alleic diversity

Importance i_H 0.4 + i_{wh} 0.2 + i_{ar} 0.4 = 1

Number of pairs in each set 4

Figure 10. Setup of calculation and comparison of breeding pairs by use a v index.

A. Mark the field **v index**

B. Chose the **indicators**

C. Set a **importance coefficients** i_H, i_{wh} and i_{ar}

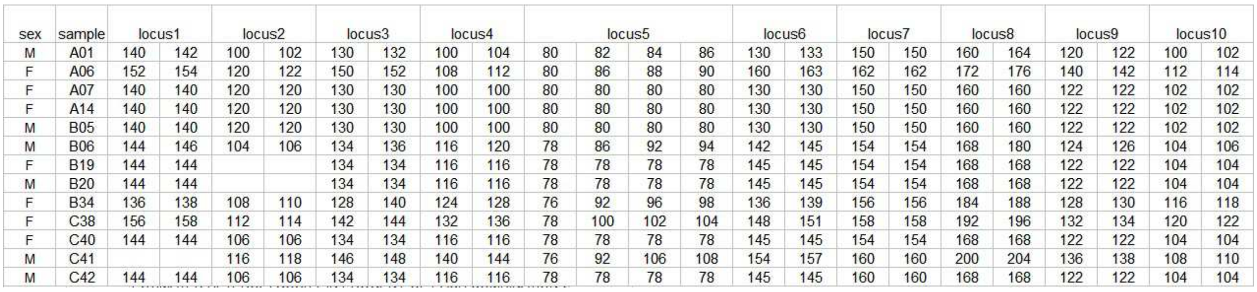
D. Set **number of pairs in each set** (number of sets should not exceed number of males or females in input data)

9.c. Example of use.

A genotyping data consisting of 13 genetic profiles of 6 males and 7 females was formatted according to requiremets of module 1. This data set included informationsuch such as: Individual sex, tag number and list of alleles of 10 microsatellite loci. One of them (locus 5) was tetrasomic.

In samples B19, B20 and C41 the data at one marker was missing. The aim of studies was to estimate a genetic variation in progeny of each possible breeding pair and find the best sets of pairs.

A Input file shown on the Figure 11 has been loaded.



sex	sample	locus1		locus2		locus3		locus4		locus5				locus6		locus7		locus8		locus9		locus10	
M	A01	140	142	100	102	130	132	100	104	80	82	84	86	130	133	150	150	160	164	120	122	100	102
F	A06	152	154	120	122	150	152	108	112	80	86	88	90	160	163	162	162	172	176	140	142	112	114
F	A07	140	140	120	120	130	130	100	100	80	80	80	80	130	130	150	150	160	160	122	122	102	102
F	A14	140	140	120	120	130	130	100	100	80	80	80	80	130	130	150	150	160	160	122	122	102	102
M	B05	140	140	120	120	130	130	100	100	80	80	80	80	130	130	150	150	160	160	122	122	102	102
M	B06	144	146	104	106	134	136	116	120	78	86	92	94	142	145	154	154	168	180	124	126	104	106
F	B19	144	144			134	134	116	116	78	78	78	78	145	145	154	154	168	168	122	122	104	104
M	B20	144	144			134	134	116	116	78	78	78	78	145	145	154	154	168	168	122	122	104	104
F	B34	136	138	108	110	128	140	124	128	76	92	96	98	136	139	156	156	184	188	128	130	116	118
F	C38	156	158	112	114	142	144	132	136	78	100	102	104	148	151	158	158	192	196	132	134	120	122
F	C40	144	144	106	106	134	134	116	116	78	78	78	78	145	145	154	154	168	168	122	122	104	104
M	C41			116	118	146	148	140	144	76	92	106	108	154	157	160	160	200	204	136	138	108	110
M	C42	144	144	106	106	134	134	116	116	78	78	78	78	145	145	160	160	168	168	122	122	104	104

Figure 11. Input file for example of use a module 1

Program setup

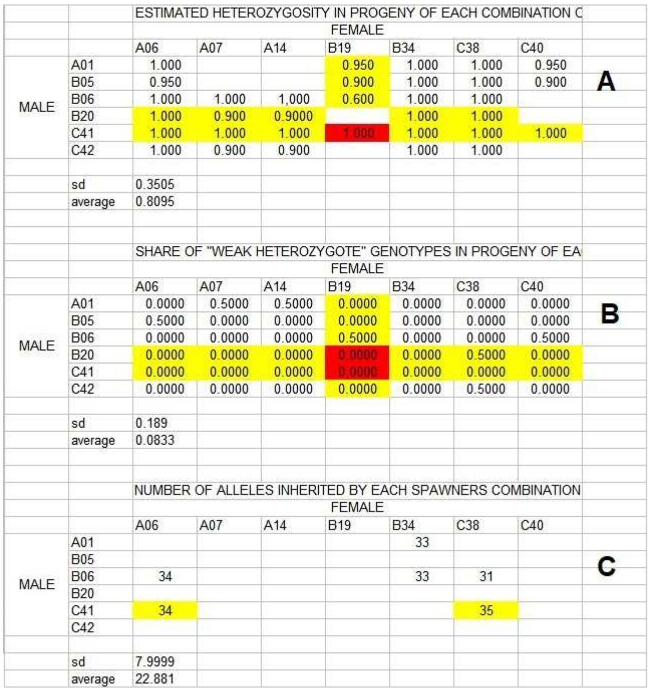
All variants of calculations specified above sa a variant I, II and III included 3 indicators (heterozygosity, share of weak heterozygotes and allelic diversity). In variant I a value number of pairs to create was 4. In variant III importance values were i_H = 0.4 i_{wh} = 0.2and i_{ar} = 0.4 number of pairs in each set was 4.

Results

Results are presented in the MS Excell file. Each result file include a sheets and their number is equal number of investigated indicators. The results of calculation a values of each indicator are given in separate sheets. The results file of variant III include additional sheet Summary where table of summarized v index index is added as well as χ^2 statistic.

Variant I

Results of calculations from sheets 1-3 were combined together and shown on Figure 9 (A, B, C).



ESTIMATED HETEROZYGOSITY IN PROGENY OF EACH COMBINATION C

		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	1.000			0.950	1.000	1.000	0.950
	B05	0.950			0.900	1.000	1.000	0.900
	B06	1.000	1.000	1,000	0.600	1.000	1.000	
	B20	1.000	0.900	0.9000		1.000	1.000	
	C41	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	C42	1.000	0.900	0.900		1.000	1.000	
	sd	0.3505						
	average	0.8095						

SHARE OF "WEAK HETEROZYGOTE" GENOTYPES IN PROGENY OF EA

		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	0.0000	0.5000	0.5000	0.0000	0.0000	0.0000	0.0000
	B05	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	B06	0.0000	0.0000	0.0000	0.5000	0.0000	0.0000	0.5000
	B20	0.0000	0.0000	0.0000	0.0000	0.0000	0.5000	0.0000
	C41	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	C42	0.0000	0.0000	0.0000	0.0000	0.0000	0.5000	0.0000
	sd	0.189						
	average	0.0833						

NUMBER OF ALLELES INHERITED BY EACH SPAWNERS COMBINATION

		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01					33		
	B05							
	B06	34				33	31	
	B20							
	C41	34					35	
	C42							
	sd	7.9999						
	average	22.881						

Figure 12. Results of calculations - variant 1

(A) heterozygosity

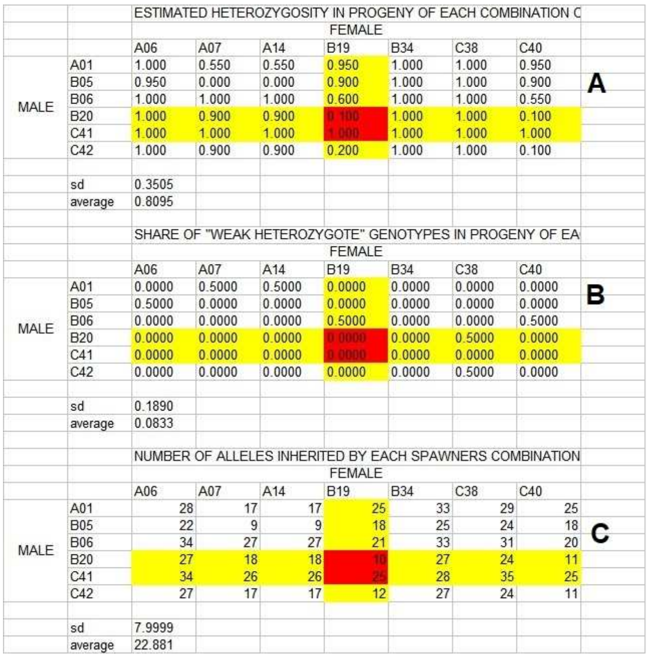
(B) share of weak heterozygotes

(C) number of alleles

In this version programe indicate pairs that have 4 best values of each indicators were presented. The values other that best 4 were not shown. If there is no data in one or more analyzed loci in one of a couple of individuals, the cell with the result of calculations is highlighted in yellow. If there is no data in one or more analyzed loci in both individuals, the cell with the result of calculations is highlighted in red. The average and standard deviation (SD) for values of given indicators calculated across progeny of all possible breeding pairs are shown just below each result table. Those values can be used for checking the differences in values shown in the tables of result comparing to average values for all variants.

Variant II. Calculation a values genetic variation indicators in all variants of breeding pairs

This variants of calculations presents a values of genetic variation indicators calculated for all breeding pairs. Its SD and average is given below table of results. The results of 3 sheets combined together are given at Figure 10 (A, B, C).



		ESTIMATED HETEROZYGOSITY IN PROGENY OF EACH COMBINATION C						
		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	1.000	0.550	0.550	0.950	1.000	1.000	0.950
	B05	0.950	0.000	0.000	0.900	1.000	1.000	0.900
	B06	1.000	1.000	1.000	0.600	1.000	1.000	0.550
	B20	1.000	0.900	0.900	0.100	1.000	1.000	0.100
	C41	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	C42	1.000	0.900	0.900	0.200	1.000	1.000	0.100
	sd	0.3505						
	average	0.8095						

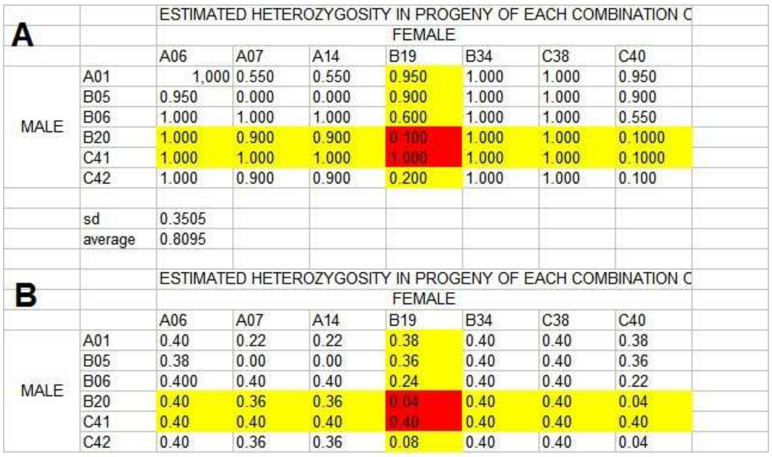
		SHARE OF "WEAK HETEROZYGOTE" GENOTYPES IN PROGENY OF EA						
		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	0.0000	0.5000	0.5000	0.0000	0.0000	0.0000	0.0000
	B05	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	B06	0.0000	0.0000	0.0000	0.5000	0.0000	0.0000	0.5000
	B20	0.0000	0.0000	0.0000	0.0000	0.0000	0.5000	0.0000
	C41	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	C42	0.0000	0.0000	0.0000	0.0000	0.0000	0.5000	0.0000
	sd	0.1890						
	average	0.0833						

		NUMBER OF ALLELES INHERITED BY EACH SPAWNERS COMBINATION						
		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	28	17	17	25	33	29	25
	B05	22	9	9	18	25	24	18
	B06	34	27	27	21	33	31	20
	B20	27	18	18	10	27	24	11
	C41	34	26	26	25	28	35	25
	C42	27	17	17	12	27	24	11
	sd	7.9999						
	average	22.881						

Figure 13. Results of calculations - Variant 2:
(A) heterozygosity
(B) share of weak heterozygotes
(C) number of alleles

Variant III. Calculation a values genetic variation indicators in all variants of breedingpairs and comparison by using *v* index

In this variant of calculations each sheet presents values genetic indicators calculated for all spawning pairs. Figure 14 (A). Below them the values of component v index is added (Figure 14B). This variant calculations have one additional sheet "summary" (Figure 15). On that sheet a values of summarized components of v index are given in the table (A) and recommended set of breeding pairs is marked bold. The coefficients of weight a components of v index are given in the table (B). The comparison of average values of genetic indicators and v index between all breeding pairs and set chosen by the Genassemblage software is given below results table (C). The scale of differences in indicators presents value of χ^2 statistic (D). Higher value indicate for bigger differences between selected set of spawning pairs and all.



A

ESTIMATED HETEROZYGOSITY IN PROGENY OF EACH COMBINATION C

		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	1,000	0.550	0.550	0.950	1.000	1.000	0.950
	B05	0.950	0.000	0.000	0.900	1.000	1.000	0.900
	B06	1.000	1.000	1.000	0.600	1.000	1.000	0.550
	B20	1.000	0.900	0.900	0.100	1.000	1.000	0.1000
	C41	1.000	1.000	1.000	1.000	1.000	1.000	0.1000
	C42	1.000	0.900	0.900	0.200	1.000	1.000	0.100
	sd	0.3505						
	average	0.8095						

B

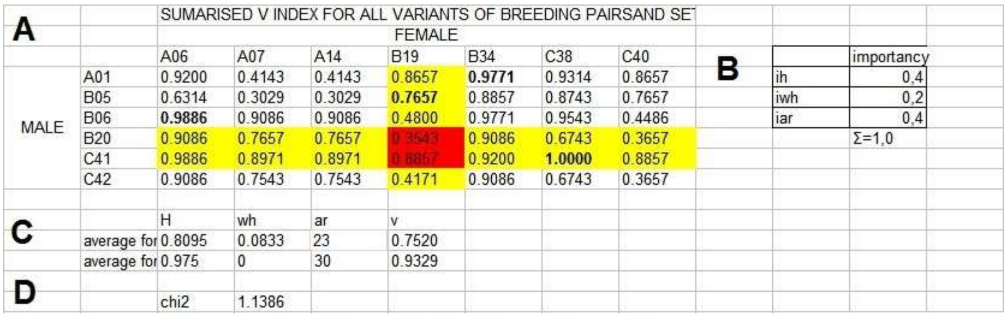
ESTIMATED HETEROZYGOSITY IN PROGENY OF EACH COMBINATION C

		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	0.40	0.22	0.22	0.38	0.40	0.40	0.38
	B05	0.38	0.00	0.00	0.36	0.40	0.40	0.36
	B06	0.400	0.40	0.40	0.24	0.40	0.40	0.22
	B20	0.40	0.36	0.36	0.04	0.40	0.40	0.04
	C41	0.40	0.40	0.40	0.40	0.40	0.40	0.40
	C42	0.40	0.36	0.36	0.08	0.40	0.40	0.04

Figure 14. Variant III, example of calculation results (heterozygosity) and values of i_h component quotient:

(A) heterozygosity

(B) i_h component of *v* index



A

SUMARISED V INDEX FOR ALL VARIANTS OF BREEDING PAIRSAND SET

		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	0.9200	0.4143	0.4143	0.8657	**0.9771**	0.9314	0.8657
	B05	0.6314	0.3029	0.3029	**0.7657**	0.8857	0.8743	0.7657
	B06	**0.9886**	0.9086	0.9086	0.4800	0.9771	0.9543	0.4486
	B20	0.9086	0.7657	0.7657	0.3543	0.9086	0.6743	0.3657
	C41	0.9886	0.8971	0.8971	0.8857	0.9200	**1.0000**	0.8857
	C42	0.9086	0.7543	0.7543	0.4171	0.9086	0.6743	0.3657

B

	importancy
ih	0,4
iwh	0,2
iar	0,4
	Σ=1,0

C

	H	wh	ar	v
average for	0.8095	0.0833	23	0.7520
average for	0.975	0	30	0.9329

D

chi2	1.1386

Figure 15. Summary of calculations based on v index:

(A) sumarised v index values

(B) weight of v index components

(C) comparison of average values of genetic variation indicators and v index between all breeding pairs and set chosen by program.

(D) χ^2 value

Summary

The optimal set of four spawning pairs would be male A1 and female B34, male B05 and female B19, male B06 and female A06 and male C41 and female C38 Figure 12. The "v index" for this set is 0.933, and they would produce a progeny cohort with very high average heterozygosity (0.975), and without weak heterozygous genotypes (0.00). Average allelic diversity inside this set is high (30 alleles) and equal 3 alleles per locus. Certain spawning pairs male B05 with female A06 or A07 should be avoided when constructing a set of spawning pairs because those individuals are the same genetic profile and this result in homozygosity in their progeny and low number of inherited alleles. Consequently those pairs have the lowest values of "v index" (0.3029) and therefore can be identified easily among all variants of breeding pairs. Moreover, those spawning pairs include male B05, thus increasing the risk of inbreeding in future generations.

10. Module 2. Selection a individuals for group breeding

This module enable to find an optimal composition a of group individuals that are intended to be used for group breeding. This feature is especially useful in conservation a human dependant fish species and group breeding in aquaculture conditions. Using this module the user introduce a number of males and females consisting to breeding group and specify what indicator of genetic variation is primary (heterozygosity or number of alleles).

10.a. Mathematical methods

The program starts calculations from finding in the input data a group of females that are genetically different from each other as much as possible. The size of this group is defined by the user. Those differences are measured as a sum of the different alleles (allelic diversity) across all loci included in their genetic profiles. In the case of more than one group of females have an identical allelic diversity, the program performs the same operations for each female group and saves the results as a separate files. example: group breeding (variant 1). xls. group breeding (variant 2) .xls, group breeding (variant 3) .xls .... After selecting the best group (or groups of females), the program begins to analyze the impact of adding individual males to allelic diversity or heterozygosity their progeny. Including or excluding a male from group is based on heterozygosity (H) or number of alleles (A) expected in progeny entire group. Adding a males to female group s continued until number of males reach a value specified by the user. The mechanism for calculating heterozygosity and the number of alleles in progeny is identical to specified in module 1. In module 2 a heterozygosity (*H*) or number of alleles (*A*) is calculated according to Algorithm 6. Finaly the values obtained for each locus are averaged Algoritm 7. Next the best compositions of breeding group are written to the output files.

[Algorithm 6]

$$H_1 = \frac{H_{x1y1} + H_{x1y2} + H_{x1yn} + H_{x2y1} + H_{x2y2} + H_{x2yn} + H_{xny1} + H_{xny2} + H_{xnyn}}{n_{xy}}$$

$$A_1 = \frac{A_{x1y1} + A_{x1y2} + A_{x1yn} + A_{x2y1} + A_{x2y2} + A_{x2yn} + A_{xny1} + A_{xny2} + A_{xnyn}}{n_{xy}}$$

Algorithm 6. Mechanism of calculation a heterozygosity (H_1) and number of alleles at given locus in module 2.

where:

H - heterozygosity in progeny of:

$x1, x2, xn$ - female nr 1, 2, n

$y1, x2, xn$ - male nr 1, 2, n

n_{xy} - number potential combinations males and females in the set

[Algorithm 7]

$$\overline{H} = \overline{x}(H_1, H_2, H_n) \quad \overline{A} = \overline{x}(A_1, A_2, A_n)$$

$\overline{x}$ - average

H_1, H_2, H_n - heterozygosity at locus 1, 2, n

A_1, A_2, A_n - number of alleles at locus 1, 2, n

Algorithm 7. Mechanism of calculation a overall heterozygosity ($\overline{H}$) and number of alleles $\overline{A}$ for set of spawners.

Input file

Input file is the same as described in module 1 and should be formatted as given in the chapter 8.4.

10.b. Calculation setup

A. After launching the Genassemblage 2.2 click the bar **Select individuals for group spawning**

B. On the window **(Figure 16)** please select

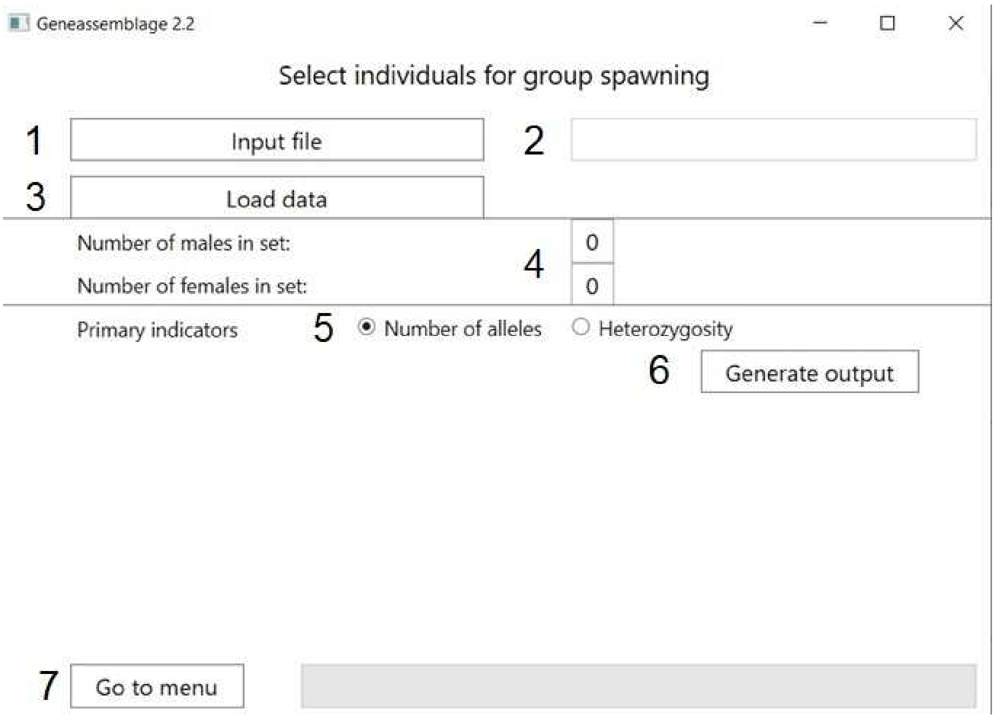


Figure 16. Window of module Select individuals for group spawning

1. click the button **Input file**

2. brose for **input file**

3. Click the button **Load data (**The communicate **Data was loaded successfully** appears if input file is correct and was successfully loaded)

4. Introduce a **number of males** and **females in set (D).**

5. Chose **primary indicator (number of alleles** or **heterozygosity)**

6. Click **Generate output, s**elect **name** and **path of the output file** and click the button **Write.**

7. If you need to go back to **Genassemblage main menu** click **Go to menu.**

10.c. Example of use

Data set given on Figure 8 was used. The aim of study is to find a set of 3 males and 3 females for group breeding that will contribute in highest heterozygosity of their progeny.

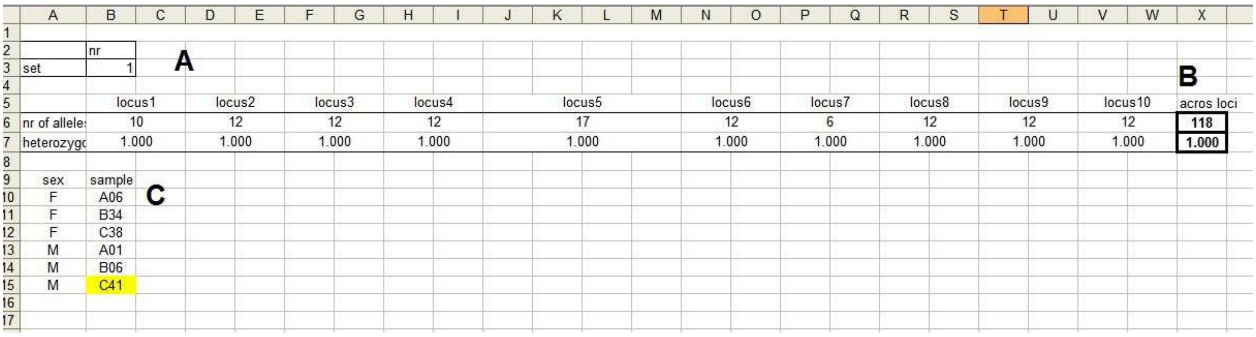
Calculation setup

Number of males that should be included in the set was 3 as well as number of females in set was 3. Heterozygosity has been marked as a primary indicator.

Results

The best set included females A06, B34, C38, and males A01, B06, C41.

This set enable to get highest possible heterozygosity in progeny (1.0) and inherit 118 alleles at 10 microsatellite loci (Figure 17). Male C41 was marked yellow because there is lacking data at one locus in this profile. Alternative set that have the same heterozygosity can be given below this marked as best.



	nr
set	1

	locus1	locus2	locus3	locus4	locus5	locus6	locus7	locus8	locus9	locus10	acros loci
nr of allele:	10	12	12	12	17	12	6	12	12	12	118
heterozygc	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

sex	sample
F	A06
F	B34
F	C38
M	A01
M	B06
M	C41

Figure 17. Results of identicication the best sets of individuals for group breeding. (A) number of set, (B) number of alleles and heterozygosity at given locuc and across all loci in progeny of this set, (C) tag numbers of individuals in this set.

Large datasets including dozens of males and females and sets including many results in huge number of potential sets of individuals that need to be assessed by program. Therefore it require some time to indicate optimal sets.

11. Module 3. Management genetic variation in gamete bank

This module was designed to help in management a genetic variation of species endangered extinction whose gametes were deposited in gamete bank. This management is based on genetic profiles of gamete donors. The program can identify a samples that differs as much as possible in alleles at genetic loci. Moreover, can find a group of samples that enable to reach a designed percentage of allelic diversity deposited in gamete bank by using a minimal number of samples. This module is especially useful in species that genetic differences between individuals is low. It enable to identify a samples that will contribute most to the genetic variation and reduction of number of samples that should be used in conservation or restoration of endangered populations.

11.a. Mathematical methods

In the input file disomic or tetrasomic loci are recognized by the software. The program automatically calculate a number of alleles at given locus and across all loci in the input file . The user introduce a percentage number of alleles identified in the input file that should be included in group of samples selected by program. After this the program calculate a target number of alleles that should be included in the proposed set of samples. If the user want to transfer a 90% of allelic diversity and 30 was identified in the input data, the program will find a samples that enable a transfer of 27 alleles. In the output file set of samples that meet two conditions is proposed. Those conditions are 1. the allelic diversity is identical target value 1. A number of samples included in the set is as small as possible. If target number of alleles is lower than number of alleles detected in input file the alleles with low frequency are preferred.

Searching optimal set of samples for transfer allelic variation is performed as follow:

Stage 1. When loading a batch file, the program calculates not only the number of alleles within all attempts and loci in the file (it gives it in the number of alleles in all samples field) but also their frequency. These values are loaded into the program memory and then used in the next step.

Stage 2. The program finds a sample in which there is the greatest diversity of alleles in its genotype. If more than one sample will have the same allele diversity, the one in which there are alleles with a lower frequency (the average calculated from the frequency of alleles identified in the genotype of this sample) is selected first.

Stage 3. Identifies which subsequent samples can contribute the most to an existing set of alleles.

Stage 4. He chooses the one that will contribute the most and if more than one sample contributes the same number of alleles he chooses the one in which the alleles occur less often (their average frequency in the examined set is lower)

Stage 5. Stages 3 and 4 are repeated until the number of user-defined alleles is reached.

In some cases is possible that more than one version of the sample set meet both conditions. Then the program randomly selects and show one variant of sample composition and in the results file add the Commnents line "There is more than one version of samples set at the same value of sample number and allelic diversity".

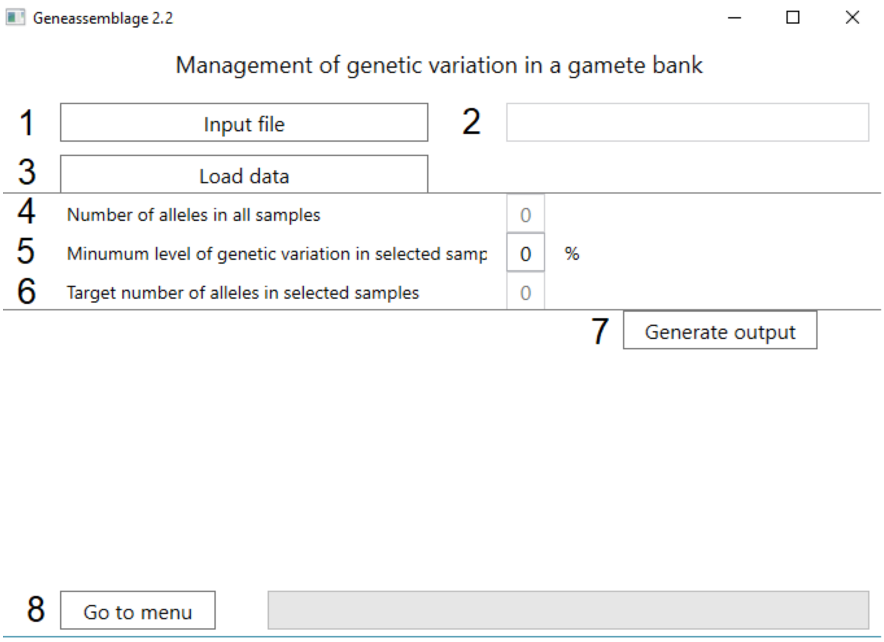
Input file

Please use a input file in format specified in the chapter 8.5. All profiles used in this module should have a complete set of data (alleles) in the input file.

11.b. Calculation setup

A. After launching the Genassemblage 2.0 click a **Management genetic variation in gamete bank**

B. On the module window (Figure 18) please:



Number of alleles in all samples 0

Minumum level of genetic variation in selected samp 0 %

Target number of alleles in selected samples 0

Figure 18. Window of module Management of genetic variation in a gamete bank.

1. click the button **Input file,**

2. brose for **input file,**

3. click the button **Load data (**The communicate **Data was loaded successfully** appears if input file is correct and was successfully loaded).

4. **Number alleles in all samples** is calculated automatically after loading an input file**.**

5. Please set a minimum level of genetic variation in % that should be transfered in the set of samples selected by a software.

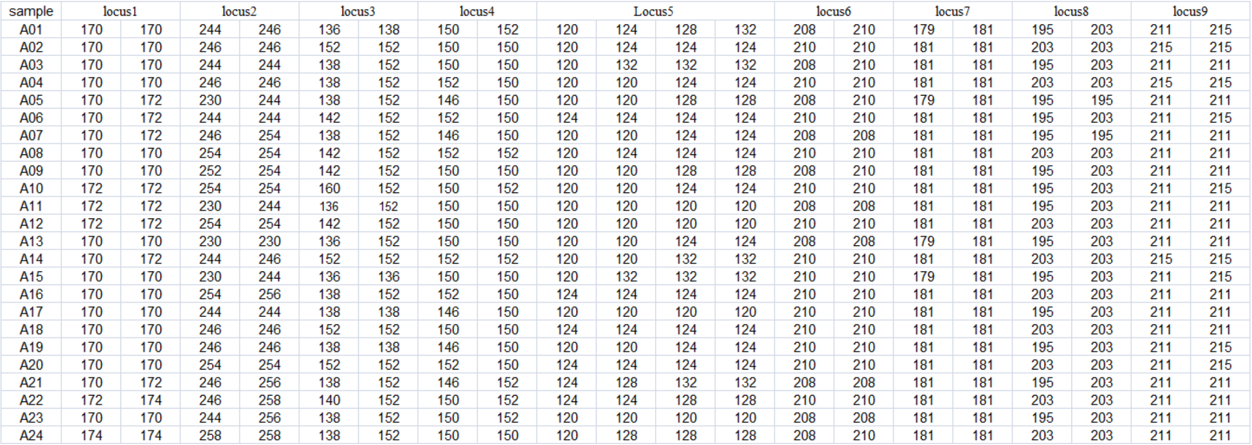
6. After this a software calculates target number of alleles that should be included in set of samples.

7. Click **Generate output,** select **name** and **path of the output file** and click the button **Write**

8. If you need to go back to **Genassemblage main menu** click **Go to menu.**

11. c. Example of use

A group 24 criopreserved sperm samples is stored in the gamete bank. Genetic profiles of their donors are included in the input file Figure 19. The aim of study was to identify a group of samples that enable 100% of allelic diversity identified in the samples deposited in the bank to the set of samples that will be used in conservation of this population.

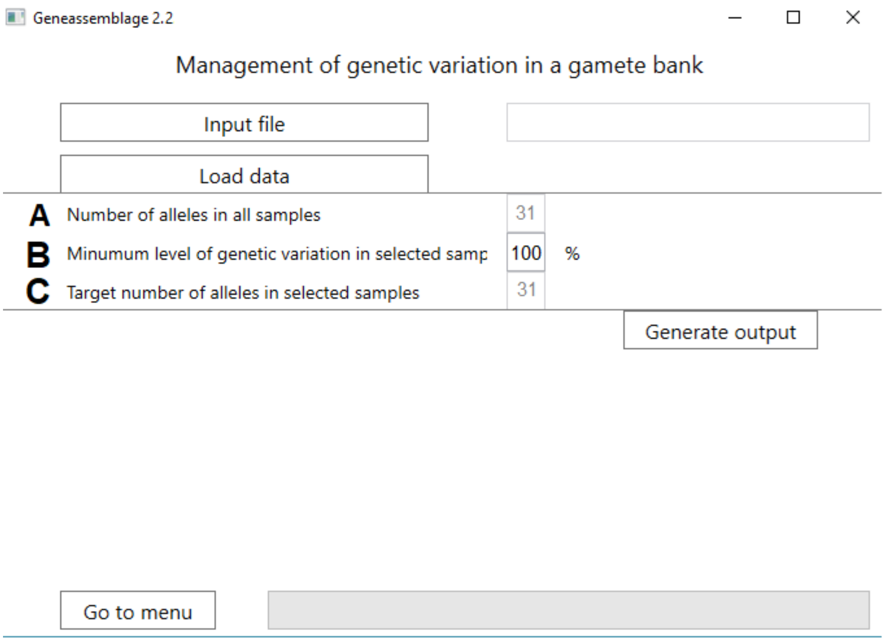


sample	locus1		locus2		locus3		locus4		Locus5				locus6		locus7		locus8		locus9	
A01	170	170	244	246	136	138	150	152	120	124	128	132	208	210	179	181	195	203	211	215
A02	170	170	246	246	152	152	150	150	120	124	124	124	210	210	181	181	203	203	215	215
A03	170	170	244	244	138	152	150	150	120	132	132	132	208	210	181	181	195	203	211	211
A04	170	170	246	246	138	152	152	150	120	120	124	124	210	210	181	181	203	203	215	215
A05	170	172	230	244	138	152	146	150	120	120	128	128	208	210	179	181	195	195	211	211
A06	170	172	244	244	142	152	152	150	124	124	124	124	210	210	181	181	195	203	211	215
A07	170	172	246	254	138	152	146	150	120	120	124	124	208	208	181	181	195	195	211	211
A08	170	170	254	254	142	152	152	152	120	124	124	124	210	210	181	181	203	203	211	211
A09	170	170	252	254	142	152	150	150	120	120	128	128	208	210	181	181	195	203	211	211
A10	172	172	254	254	160	152	150	152	120	120	124	124	210	210	181	181	195	203	211	215
A11	172	172	230	244	136	152	150	150	120	120	120	120	208	208	181	181	195	203	211	211
A12	172	172	254	254	142	152	150	150	120	120	120	120	210	210	181	181	203	203	211	211
A13	170	170	230	230	136	152	150	150	120	120	124	124	208	208	179	181	195	203	211	211
A14	170	172	244	246	152	152	152	152	120	120	132	132	210	210	181	181	203	203	215	215
A15	170	170	230	244	136	136	150	150	120	132	132	132	210	210	179	181	195	203	211	215
A16	170	170	254	256	138	152	152	150	124	124	124	124	210	210	181	181	203	203	211	211
A17	170	170	244	244	138	138	146	150	120	120	120	120	210	210	181	181	195	203	211	211
A18	170	170	246	246	152	152	150	150	124	124	124	124	210	210	181	181	203	203	211	211
A19	170	170	246	246	138	138	146	150	120	120	124	124	210	210	181	181	195	203	211	215
A20	170	170	254	254	152	152	152	150	124	124	124	124	210	210	181	181	203	203	211	215
A21	170	172	246	256	138	152	146	152	124	128	132	132	208	208	181	181	195	203	211	211
A22	172	174	246	258	140	152	150	152	124	124	128	128	210	210	181	181	203	203	211	211
A23	170	170	244	256	138	152	150	152	120	120	120	120	208	208	181	181	195	203	211	211
A24	174	174	258	258	138	152	150	150	120	128	128	128	208	210	181	181	203	203	211	211

Figure 19. Module 3, input file for example of calculations.

Calculation setup

After a input file given above was loaded a number of alleles across all loci (31), Figure 20 (A) was calculated. Minimum level of genetic variation in selected samples was set to 100% (B). Consequently, 31 alleles should be included in group of selected samples (C).



Management of genetic variation in a gamete bank

A Number of alleles in all samples 31

B Minumum level of genetic variation in selected samp 100 %

C Target number of alleles in selected samples 31

Figure 20. Setup a module management of genetic variation in a gamete bank program - example.

Results

In calculation results (Figure 21) a following data are presented:

A. The list of genetic profiles from the input data

B. List of alleles identified in the input data

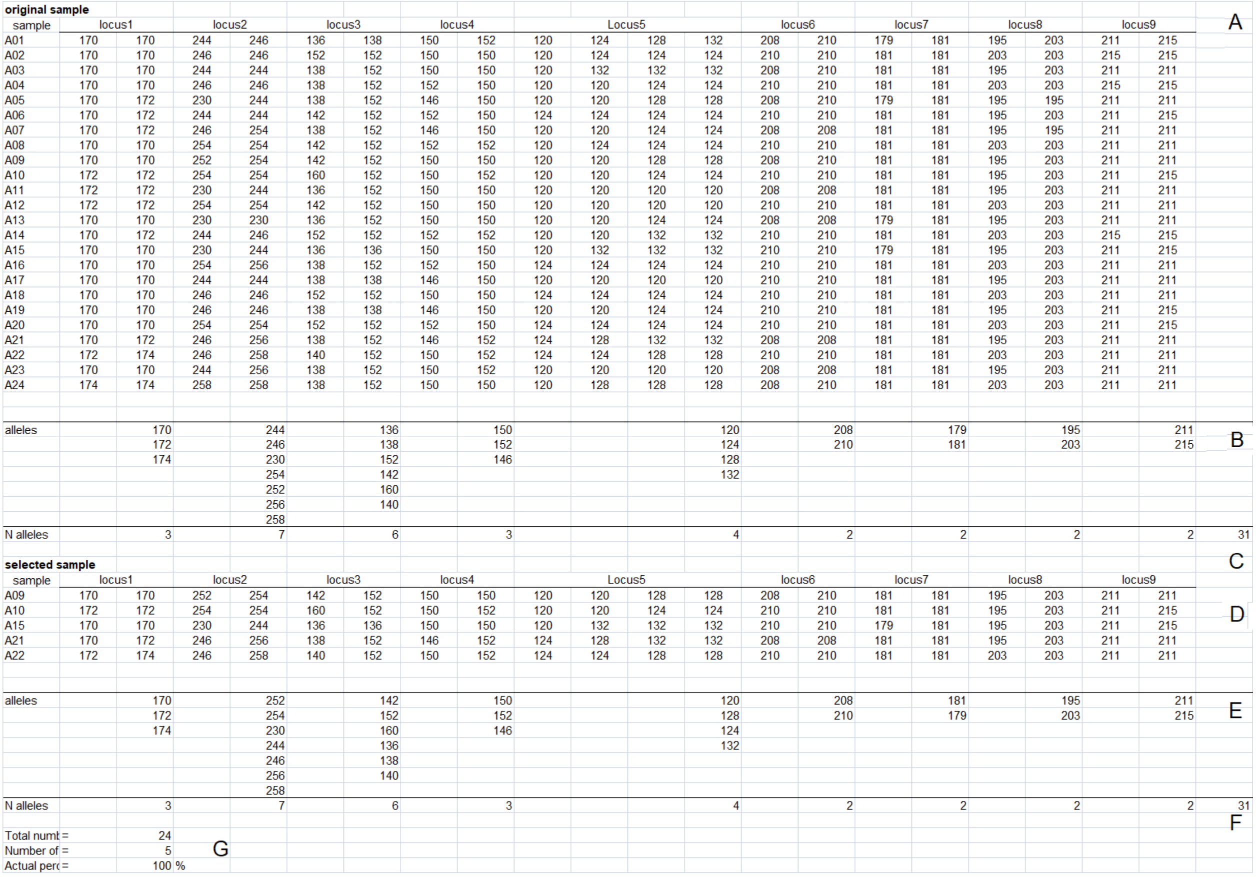
C. Number of alleles at given locus and across all investigated loci in the input data

D. Set of samples that meets a criteria of genetic variation that were set by a user.

E. List alleles that are included in the set of samples proposed by the software

F. Number of alleles at given locus and across all loci in the set of samples proposed by the software.

G. Summary of number all profiles in the input data, number of smples chosen by the software and percentage of total allelic diversity from the input data that were included in the set of samples.



original sample																					
sample	locus1		locus2		locus3		locus4		Locus5				locus6		locus7		locus8		locus9		A
A01	170	170	244	246	136	138	150	152	120	124	128	132	208	210	179	181	195	203	211	215	
A02	170	170	246	246	152	152	150	150	120	124	124	124	210	210	181	181	203	203	215	215	
A03	170	170	244	244	138	152	150	150	120	132	132	132	208	210	181	181	195	203	211	211	
A04	170	170	246	246	138	152	152	150	120	120	124	124	210	210	181	181	203	203	215	215	
A05	170	172	230	244	138	152	146	150	120	120	128	128	208	210	179	181	195	195	211	211	
A06	170	172	244	244	142	152	152	150	124	124	124	124	210	210	181	181	195	203	211	215	
A07	170	172	246	254	138	152	146	150	120	120	124	124	208	208	181	181	195	195	211	211	
A08	170	170	254	254	142	152	152	152	120	124	124	124	210	210	181	181	203	203	211	211	
A09	170	170	252	254	142	152	150	150	120	120	128	128	208	210	181	181	195	203	211	211	
A10	172	172	254	254	160	152	150	152	120	120	124	124	210	210	181	181	195	203	211	215	
A11	172	172	230	244	136	152	150	150	120	120	120	120	208	208	181	181	195	203	211	211	
A12	172	172	254	254	142	152	150	150	120	120	120	120	210	210	181	181	203	203	211	211	
A13	170	170	230	230	136	152	150	150	120	120	124	124	208	208	179	181	195	203	211	211	
A14	170	172	244	246	152	152	152	152	120	120	132	132	210	210	181	181	203	203	215	215	
A15	170	170	230	244	136	136	150	150	120	132	132	132	210	210	179	181	195	203	211	215	
A16	170	170	254	256	138	152	152	150	124	124	124	124	210	210	181	181	203	203	211	211	
A17	170	170	244	244	138	138	146	150	120	120	120	120	210	210	181	181	195	203	211	211	
A18	170	170	246	246	152	152	150	150	124	124	124	124	210	210	181	181	203	203	211	211	
A19	170	170	246	246	138	138	146	150	120	120	124	124	210	210	181	181	195	203	211	215	
A20	170	170	254	254	152	152	152	150	124	124	124	124	210	210	181	181	203	203	211	215	
A21	170	172	246	256	138	152	146	152	124	128	132	132	208	208	181	181	195	203	211	211	
A22	172	174	246	258	140	152	150	152	124	124	128	128	210	210	181	181	203	203	211	211	
A23	170	170	244	256	138	152	150	152	120	120	120	120	208	208	181	181	195	203	211	211	
A24	174	174	258	258	138	152	150	150	120	128	128	128	208	210	181	181	203	203	211	211	
alleles		170		244		136		150				120		208		179		195		211	
		172		246		138		152				124		210		181		203		215	B
		174		230		152		146				128									
				254		142						132									
				252		160															
				256		140															
				258																	
N alleles		3		7		6		3				4		2		2		2		2	31
selected sample																					C
sample	locus1		locus2		locus3		locus4		Locus5				locus6		locus7		locus8		locus9		
A09	170	170	252	254	142	152	150	150	120	120	128	128	208	210	181	181	195	203	211	211	
A10	172	172	254	254	160	152	150	152	120	120	124	124	210	210	181	181	195	203	211	215	D
A15	170	170	230	244	136	136	150	150	120	132	132	132	210	210	179	181	195	203	211	215	
A21	170	172	246	256	138	152	146	152	124	128	132	132	208	208	181	181	195	203	211	211	
A22	172	174	246	258	140	152	150	152	124	124	128	128	210	210	181	181	203	203	211	211	
alleles		170		252		142		150				120		208		181		195		211	
		172		254		152		152				128		210		179		203		215	E
		174		230		160		146				124									
				244		136						132									
				246		138															
				256		140															
				258																	
N alleles		3		7		6		3				4		2		2		2		2	31
																					F
Total numt	=	24																			
Number of	=	5	G																		
Actual perc	=	100	%																		

Figure 21. Example of calculation results performed by using a module Management of genetic variation.

A set of samples: A09, A10, A15, A21, A22 meets requirement of transfer to next generation a100% genetic variation identified across all samples in the bank. This set can be used for conservation this population.

12. Module 4. File conversion to the Arlequine format

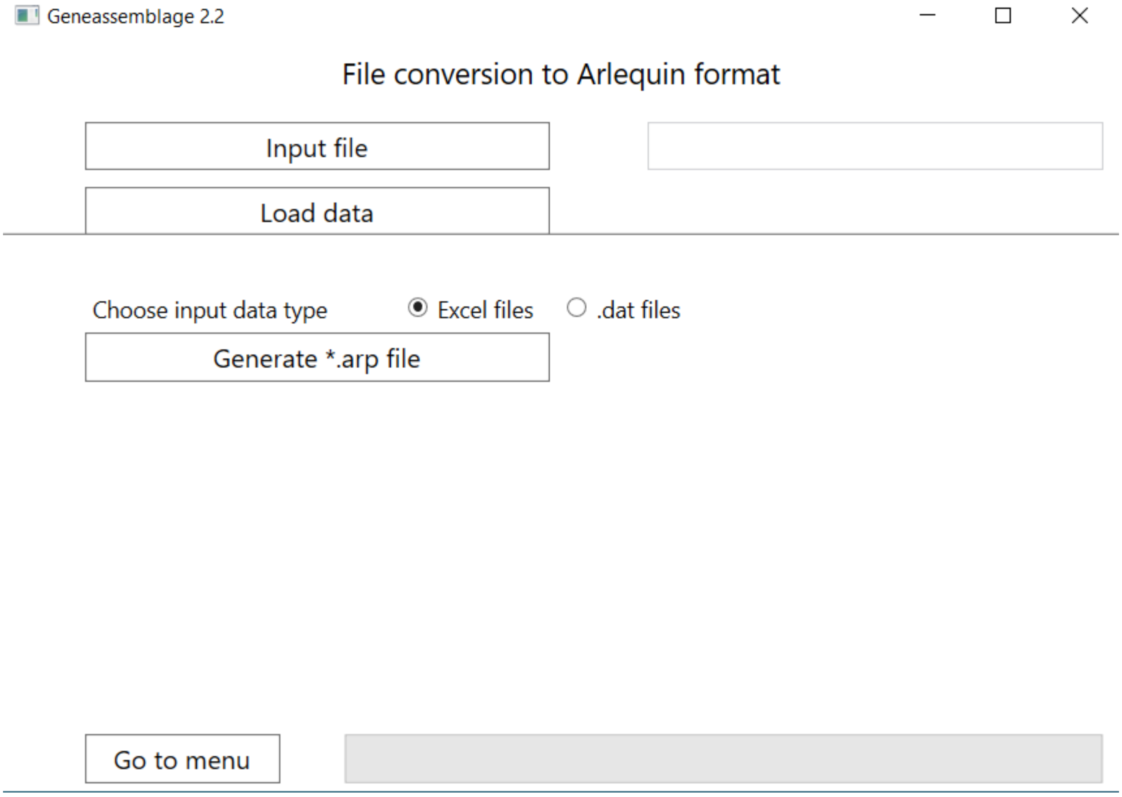
12.a. Conversion of Genasemblage files to Arlequin format

The Genassemblage enables direct conversion of a file consistent with its input format (.xls, and .xlsx) to the .arp format, i.e. used by Arlequin 3.0 and Arlequin 3.5 (Excoffier et al. 2005, Excoffier and Lischer 2010). The conversion process retains the structure of population groups contained in the “population group” column in the input data file and population names.

The input .xls or xlsx file should be formated as shown in the chapter 9.4 and Figure 4.

12.b. Conversion of MSA files to to Arlequin format

There is an embedded tool in Genassemblage which enables direct conversion of an input .dat file, used by Microsatellite Analyser (MSA) (Dieringer and Schlötterer 2003) to the .arp format, used by Arlequin 3.0 (Excoffier et al. 2005) and Arlequin 3.5. (Excoffier and Lisher 2010).


Geneassemblage 2.2

File conversion to Arlequin format

Choose input data type ● Excel files ○ .dat files

Figure 22. Window of module File conversion to Arlequin format.

12.c. Conversion setup

Conversion setup should be performed as described on the Figure 22 and text below.

After launching the Genassemblage 2.0 click the bar **File conversion to to Arlequin format**

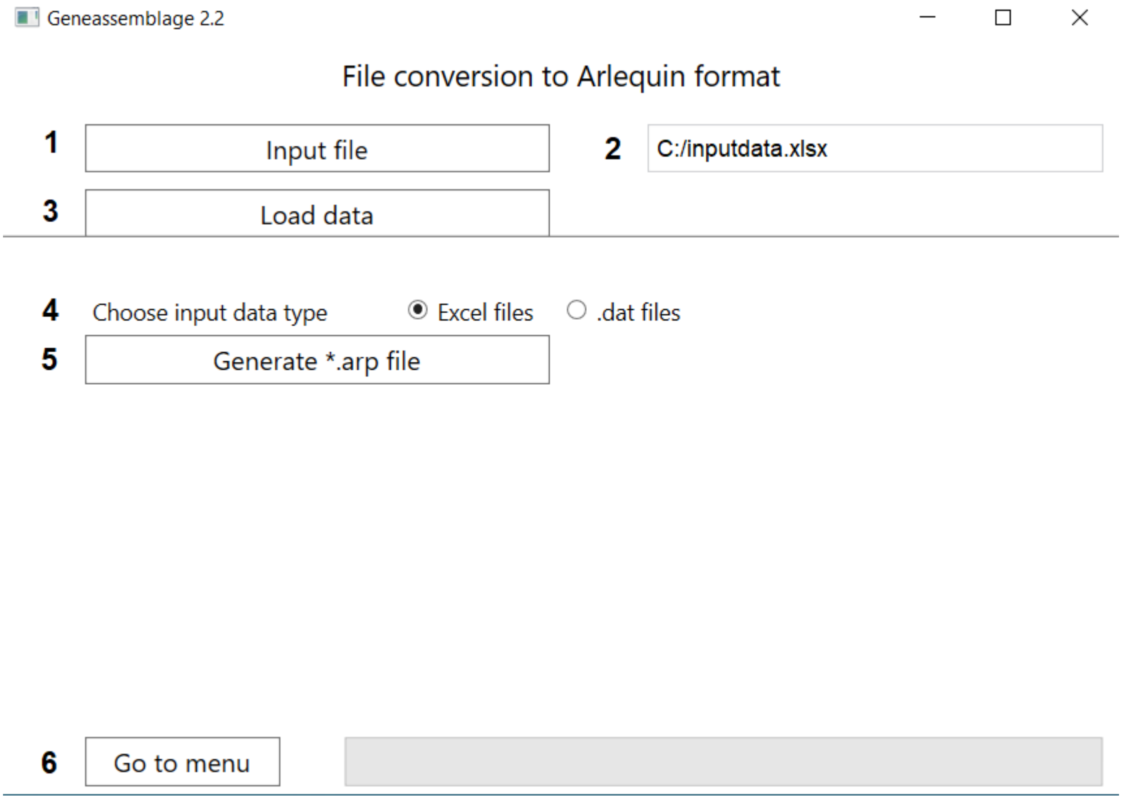


Figure 23. Conversion a input file to the Arlequin format (.arp).

1. click the button **Input file**

2. brose for **input file**

3. Click the button **Load data** (The communicate **Data was loaded successfully** appears if input file is correct and was successfully loaded)

4. Chose **input file (Excel file** or **.dat file)**

5. Click **Generate output,** select **name** and **path of the output file** and click the button **Write.**

6. If you need to go back to **Genassemblage main menu** click **Go to menu.**

The output file can be used as a input file for the Arlequine software.

Acknowledgments:

I would like to thank eng. Dominika Karyś for her help in development of this software and many very helpful suggestions that enabled improvement a interface of this software.

References

Dieringer D., Schlötterer C. (2003). Microsatellite analyzer (MSA): a platform-independent analysis tool for large microsatellite data sets. Ecology Notes, 167-169.

Excoffier L, Lischer H.E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows", *Molecular Ecology Resour*ces, 10:564-567

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

Kaczmarczyk D., Fopp-Bayat D. (2013). Assemblage of spawning pairs based on their individual genetic profiles – as tool for maintaining genetic variation within sturgeon populations. *Aquaculture Research* 44: 677–682.

Kaczmarczyk D. 2016. Genassemblage software, a tool for management of genetic diversity in human dependent population. *Conservation Genetic Resources*. 3: 49-51.