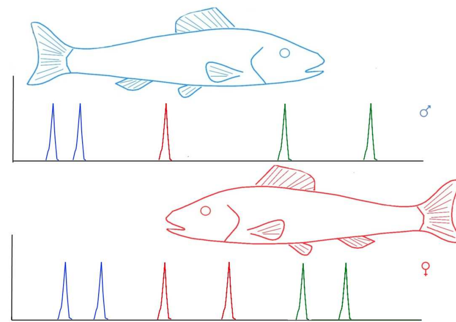
GENASSEMBLAGE 1.0



The maintaining of a genetic variation within broodstock is important for successful fish farming and successful conservation of the hatchery-dependant species. Breeders, as well scientists involved in producing juveniles should assemble the spawning pairs from fish that are as genetically different as possible. Unfortunately, the genetic variation within hatchery dependant populations may be decreased because of progressive elimination of allelic diversity from genomes of individuals within to the broodstocks (Koljonen et al. 1999, Koljonen et al.. 2002, Verspoor 2005). This decrease of genetic variation might be a result of obtaining a large group of juveniles from a few parental individuals, possible inbreeding events (Bryant et al. 1986, Kosowska and Nowicki 1999) as well as domestication of the broodstock when may eliminate individuals with alleles specific for natural population (Kim et al. 1994). Molecular genetics offers a tool which has proved useful in assessment and maintenance of genetic diversity (Koljonen et al. 2002). The highly polymorphic fragments of the DNA such microsatellites can be utilized in preparing individual genetic profiles of spawners. They might be used in identification of different possible spawners and assembling them into spawning pairs. Unfortunately, there is no computer-based tool available for identification of the best spawning pairs among the large databases of genetic profiles of spawners. The Genassemblage software was constructed to help in identification of the best possible combinations of spawners within and between broodstocks.

Table of contents

1. Program information
2. Program description
3. Application of the program
4. System requirements
5. Mathematical methods
    a. Calculation of expected heterozygosity of offspring
    b. Calculation of the percentage of “weak heterozygotes”
    c. Calculation of the number of alleles potentially inherited by offspring
6. Tools for file conversion
    a. Conversion of input files to the .arp format
    b. Conversion of .dat files to the .arp format
6. Installation
7. Interface description
8. Input files
9. Calculations
10. Calculation results
11. Conversion of files to the .arp format

1. Program information

The creation of Genassemblage program was supported by Polish Ministry of Science (project number N N304 400338)

The author and holder of the licence and all the ownership rights to Genassemblage is:

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

The author of Genassemblage 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 quote the work:

Kaczmarczyk D. 2015. Genassemblage software, a tool for management of genetic diversity in human dependent population. *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:

Astcon
al. Zwycięstwa 96/98
81-451 Gdynia Poland

2. Program description

Genassemblage is a computer tool designed to manage selection of parent individuals based on their genetic characteristics. It is unique in that it allows analysis of diploid, tetraploid and partially tetraploid organisms. The program enables identification of the best variants of matching reproductive couples based on genetic variation of their offspring, and estimation of the effect of different variants of matching parental individuals on the genetic variation of the subsequent generation. Selection of parental individuals is optimised based on individual genetic traits (genetic profiles) developed 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 analysed genetic markers. The markers usable in Genassemblage should have the following characteristics:

A. A high level of polymorphism,

B. autosomal chromosome inheritance in a accordance with the Mendelian Laws

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, 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. The individuals with a higher frequency of heterozygotes within the analysed genetic markers are more likely to differ in loci that determine population viability and adaptability.

118

119 **3. Application of the program**

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

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

127 The program can be used to convert *.xls or *.dat files to *.arp. files.

128 The program enables estimation of genetic variance indexes, such as heterozygosity (also the
129 percentage of “weak heterozygotes” for tetrasomic loci) and the number of different alleles
130 potentially inherited by offspring of selected breeding couples. The term “weak heterozygote” is
131 understood to denote individuals with three identical alleles and a different fourth one that differs
132 at a tetrasomic locus, e.g. AAAB (Kaczmarczyk and Fopp-Bayat 2012). The indices calculated
133 for each individual are presented together in tables; it is possible to chose parental combinations
134 whose offspring will show the best values of these indexes.

135

136 **4. System requirements**

137 1. The program operates in the Windows environment (it requires Microsoft NET Framework 3.0
138 or higher) and in Microsoft Excel 2013 or earlier.

139 2. A file with input data should be created in MS Excel by following the procedure presented in
140 Chapter 7.

141 3. The upper limit of the size of the analysed set of individuals is 254 females and 65533 males.
142 There is no limit on number of analysed markers.

143 4. In order to calculate the share of individuals which are “weak heterozygotes”, the program
144 requires a set of input data containing the results of genotyping of at least one tetrasomic
145 locus in the analysed breeding couple. If the data contain tetrasomic loci, they can take any
146 position in the sequence of analysed loci and their ratio to disomic loci can have any value.

147

148 **4. Mathematical methods**

149 **a. Calculation of expected heterozygosity of offspring**

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

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

159 [Algorithm 1]

160 $$H = 1 - \frac{(ph_1 + ph_2 + ph_3 + ...ph_n)}{nl}$$

161 Algorithm 1. Calculation of expected heterozygosity of offspring

162

163 **b. Calculation of the expected percentage of “weak heterozygote” individuals**

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

172 [Algorithm 2]

173 $$wh = \frac{(pwh_1 + pwh_2 + pwh_3 + ...pwh_n)}{nl}$$

174 Algorithm 2. Calculation of the expected percentage of “weak heterozygotes” Individuals.

175

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

178

179

180 **c. Calculation potential allelic diversity**

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

184 [Algorithm 3]

185 $$ar = \sum(na_n)$$

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

187

188 **6. Tools for file conversion**

189 **A. Conversion of Genasemblage files to Arlequin format**

190 There is an embedded tool in Genassemblage which enables direct conversion of a file from its
191 input format (.xls) to the .arp format, used by Arlequin 3.0 (Excoffier et al. 2005), and Arlequin
192 3.5 (Excoffier and Lischer 2010). The conversion process retains the structure of the population
193 groups contained in the “population group” column in the input data file and the population
194 names. The input file intended for conversion to .arp format should not include tetrasomic
195 markers; if it does, such markers should be deleted from the input data file. If there is a
196 tetrasomic marker in the input file, the program divides them into two virtual disomic isoloci,
197 which are inherited independently.

198

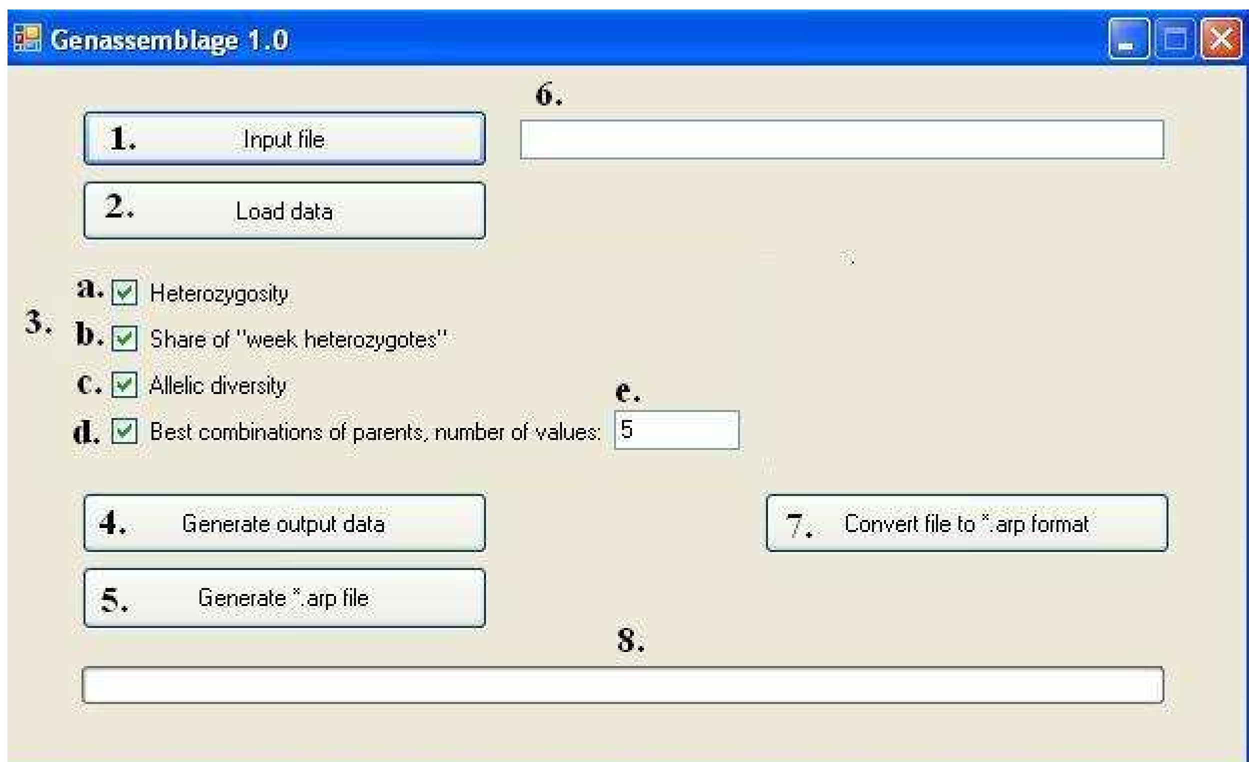
199 **B. Conversion of MSA files to to Arlequin format**

200 There is an embedded tool in Genassemblage which enables direct conversion of an input .dat
201 file, used by Microsatellite Analyser (MSA) (Dieringer and Schlötterer 2003) to the .arp format,
202 used by Arlequin 3.0 (Excoffier et al. 2005) and Arlequin 3.5. (Excoffier and Lisher 2010) A .dat
203 file intended for conversion should be formatted in accordance with the requirements specified in
204 the instructions for the MSA program, there is an example of such a file called microsatellite-
205 example.dat, in the Genassemblage folder (saved as a text file with tab characters as separators).
206 If a .dat file has a different structure than the input file is used for MSA, it should be rebuilt as
207 described in Chapter 9B.

208

209 **7. Program installation**

210 The Genassemble installer can be download from the author's webpage
211 http://pracownicy.uwm.edu.pl/d.kaczmarczyk/main_page.htm
212 The program requires Microsoft NET Framework ".NET", version 3.0 or later. If it has not been
213 installed, please download it from the Microsoft website: http://www.microsoft.com/pl-
214 pl/download/details.aspx?id=31 and install it.
215 Before installing the Genassemblage program MS Excel in version 2007 or earlier must be
216 installed on your computer.
217 After the Microsoft NET Framework package has been installed, click
218 Genassemblage_Installer.exe and specify the target location of the program on your hard drive.
219 After finishing the instalation, the program is ready to use.
220
221 **8. Interface**
222 The following dialogue box (Figure 1) appears after clicking Genassemblage.exe.



223
224 **Figure 1.** Genassemblage interface
225 where:
226 1. button for selection an input file,
227 2. button for loading data from the input file to the program memory,
228 3. program settings (for chosing the type of calculations done by the program):

229 a. heterozygosity,

230 b. percentage of “weak heterozygotes”,

231 c. number of inherited alleles,

232 d. listing of the optimum breeding pairs,

233 e. the number of the optimum breeding pairs that the program will indicate,

234 4. button for generating an output file,

235 5. button for conversion of an input file (.xls format) to the .arp format (for Arlequin 3.0 and

236 3.5)

237 6. the location of the input file on the hard drive,

238 7. button for converting of a .dat file to .arp format Arlequin 3.0

239 8. progress bar for calculation

240

241 **9. Input files**

242 An input file should be created in MS Excel in accordance with example (Figure 2):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	population group	population	sex	sample	name of locus1		name of locus2 (tetrasomic)				name of locus3		locus4 (tetrasomic)			
2	1	PopA	M	A01	170	170	120	120	124	124	244	246	155	155	155	157
3	1	PopA	F	A06	170	170	120	124	124	132	244	246	155	155	155	155
4	1	PopA	F	A07	170	170	120	124	132	132	230	244	155	155	157	157
5	1	PopA	F	A14	170	170	120	120	120	132	246	254	155	155	155	157
6	1	PopB	M	B05	170	172	120	120	120	120	252	254	155	157	157	159
7	1	PopB	F	B19	172	172	120	120	124	124	230	254	155	155	155	159
8	1	PopB	M	B20	170	170	120	120	124	124	230	244	155	155	159	159
9	1	PopB	F	B34	170	170	120	124	124	124	230	254	155	155	155	159
10	2	PopC	F	C38	170	170	124	128	128	132	244	246	153	153	155	155
11	2	PopC	F	C40			120	124	124	124	246	254	153	153	155	155
12	2	PopC	M	C41	170	174					256	258	153	155	155	155
13	2	PopC	M	C42	170	174	120	120	120	128	256	258	153	155	155	155

243

244 **Figure 2.** Format of an input file for Genassemblage where:

245 Columns A and B – definition of terms: A population is a community of potentially

246 interbreeding animals, such as all the members of one species of fish in a particular pond. A

247 population grouping is chosen by the operator of Genassemblage and could be, for example,

248 all the populations of a species of fish in a particular region or country.

249 Column A – the value here defines the group to which the population in column B belongs. If

250 all the populations analysed in the input file belong to one group, then the value in this

251 column = 1. If the analysed populations belong to a larger number of groups, then they

252 should be assigned numbers according to where they belong, beginning 1, 2, 3, etc.,

253 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 C40, locus 1).

9A. Input files for conversion to the .arp format

The Genassemblage enables direct conversion of a file consistent with its input format (.xls) 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 file intended for conversion to the .arp format should not include tetrasomic markers; if it does, such markers should be deleted from the input data file. If there is a tetrasomic marker in the input file, the program divides them into two virtual disomic isoloci, which are inherited independently.

8B. Example conversion of .dat file fto .arp format.

If a .dat file with a different structure than the input file is used for MSA, it should be rebuilt as shown in Figure 3 before it is converted to .arp format.

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 3. 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.

10. Calculations

1. Clicking the input file selection button [1] will open the file selection window; indicate the input file for Genassemblage, e.g. input file - example.exe, this program should be located in the program folder.
2. Load the data from the input file to the program memory by clicking button [2]; after the data has been loaded, the "data loaded" message will appear.
3. Select the type of calculations [3], bearing in mind that the percentage of "weak heterozygotes" can be calculated only for the data which contain at least one tetrasomic locus. By default, all types of calculations are done; if, for example, only one index is needed, unselect the other ones. If it is necessary to identify spawning couples which will produce the most heterozygous offspring, with the smallest percentage of "weak heterozygotes" and the most diverse alleles, such information can be listed in an Excel

file generated by leaving the field in the “best combination of parents” selected and by entering in the field next to it the number of the best values of each index presented.

4. After selecting the indexes to be calculated, click the “generate output data” button; calculation progress will be shown in the bar [8].
5. After the calculations are completed, the program will open a window which allows the choice of location on the hard drive and the name of output file (please enter the file name and indicate the disc and folder where it will be saved). Subsequently, a file will be generated listing the offspring of breeding couples with the best values of all the indexes. After the file is generated, the program will open a window where you should indicate the name of the file with the listing and location where it will be saved.

10A. Conversion of input files to the .arp format

1. Clicking the input file selection button [1] will open the file selection window; indicate the input file for Genassemblage, consistent with the requirement specified in Chapter 8, e.g. input file - example.xls.
2. Load the data from the input file to the program memory by clicking button [2]; after the data has been loaded, the “data loaded” message will appear.
3. After the data has been loaded, click the “generate .arp file” button [5].
4. A window will open in which you should indicate the disc and folder in which the converted file is to be saved.

10.B Conversion of .dat files to the .arp format

1. Please select a file for conversion, consistent with the format described in section **8A. Input dat file to be converted into the arp format.** To do this, click the “convert to arp. file” button [7] and select the .dat file you want converted.
2. The resulting .arp file will be automatically saved in the same folder as the input file used for conversion. It can be used for calculations done with the Arlequin program.

10. Calculation results

The results include one or two MS Excel files. One of them contains the results of calculations done for all the matching variants (the combination of spawners). Each sheet in the file is described in accordance with the name of the index whose values are shown in it.

332 The values of an index are presented as a table whose row 2 is the name of the index, row 4 -
333 numbers of females, column 2 - numbers of males. In intersections of rows and columns there
334 are values of expected indexes in offspring of individual breeding couples. Figure (4 , 5, 6). If
335 there is no data in one or more analysed loci in one of a couple of individuals, the cell with
336 the result of calculations is highlighted in yellow. If there is no data in one or more analysed
337 loci in both individuals, the cell with the result of calculations is highlighted in orange.

338

	ESTIMATED HETEROZYGOSITY IN PROGENY OF GIVEN COMBINATION OF SPAWNERS							
		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	0.4931	0.6667	0.6042	0.9236	0.6667	0.6042	0.8611
	B05	0.8750	0.8681	0.6875	0.7708	0.8125	0.8750	0.9167
	B20	0.6389	0.6181	0.7083	0.9028	0.6458	0.6806	0.9630
	C41	0.6667	0.8056	0.7500	0.9167	0.7500	0.8056	0.9583
	C42	0.7500	0.8542	0.7500	0.9167	0.8125	0.8542	0.9722

339

340 **Figure 4.** Expected heterozygosity among the offspring of the analysed couples. Yellow, the
341 pair with incomplete data for one parent. Orange the pair with incomplete data for both
342 parents.

343

	The share of "weak heterozygotes" genetypes within progeny of given spawners combination (all = 1)							
		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	0.4097	0.3403	0.4583	0.4722	0.5000	0.2778	0.4861
	B05	0.5000	0.4028	0.4792	0.5000	0.3750	0.0417	0.2917
	B20	0.4931	0.1944	0.4167	0.4722	0.5000	0.1667	0.3750
	C41	0.5000	0.4722	0.5833	0.5833	0.5833	0.5139	0.5139
	C42	0.3750	0.3611	0.5417	0.5139	0.4167	0.2986	0.3819

344

345 **Figure 5.** Expected percentage of individuals which are “weak heterozygotes” among the
346 offspring of the analysed couples.

	The number of alleles inherited by progeny of given spawners combination							
		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01	8	9	9	11	10	10	9
	B05	12	12	10	10	10	14	11
	B20	9	9	11	9	8	11	10
	C41	8	9	9	10	9	8	8
	C42	12	13	12	13	12	12	11

347

348

Figure 6. Diversity of alleles of the analysed markers inherited by the offspring of the analysed breeding couples.

If the option has been selected of presentation of couples whose offspring would have the best values of individual indexes, Genassemblage will create a separate output file in which only the matching variants are presented with the best values of a specific index (Figure 7, 8 and 9). The method of highlighting and formatting of the file used in it is the same as in the output file with calculation results for all possible matching variants.

		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01				0.9236			
	B05							0.9167
	B20							0.963
	C41				0.9167			0.9583
	C42				0.9167			0.9722

Figure 7. Best couples in terms of the value of parameter "estimated offspring heterozygosity"

		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01						0.2778	
	B05						0.0417	0.2917
	B20		0.1944				0.1667	
	C41							
	C42							

Figure 8. Best couples in terms of the parameter "percentage of offspring – "weak heterozygotes".

		FEMALE						
		A06	A07	A14	B19	B34	C38	C40
MALE	A01				11	10	10	
	B05	12	12	10	10	10	14	11
	B20			11			11	10
	C41				10			
	C42	12	13	12	13	12	12	11

Figure 9. Best couples in terms of the parameter "diversity of inherited alleles".

Among possible variants of spawning pairs the best seems to be male B05 with female C38, because their progeny are distinguished by: highest allelic diversity, lowest share of "weak heterozygotic" genotypes as well as high heterozygosity.

368

369 **11. Example of application the Genassemblage software**

370 The broodstock consist of 11 tagged individuals. Their genetic profiles, including 4
371 microssatelite loci were prepared. The species is partially tetrasomic i.e. rainbow trout. The
372 breeder would like to identify which combinations of parents are best for maintaining genetic
373 variation of this broodstock and will result in .high heterozygosity, low share of "weak
374 heterozygotic" genotypes, and high allelic richness in the next generation. The genetic
375 profiles of those spawners were given in Figure 2

376

377

378 **An example of the use of Genassemblage**

379

380 **Background**

381 A breeder needs to construct a set of four spawning pairs that will be best for maintaining the
382 genetic variation of his broodstock. The genetic profiles of the spawners are given in Figure 2.

383

384 **Setup**

385 The genetic profiles of the spawners include tetrasomic loci, so all three indicators of genetic
386 variation can be included in the calculation. Thus, the analysis will include a calculation of
387 heterozygosity, the share of "weak heterozygotes", and the number of different of alleles
388 potentially inherited by the progeny. These functions must be checked in the interface (Figure 1);
389 however, the box labelled "Best combination of parents, number of values" should be unchecked
390 – checking this box enables the user to select the number of breeding pairs that the program will
391 indicate, but in this example the breeder is going to find pairs that have the optimal balance of the
392 above three characteristics. A step-by-step guide of how to setup and perform calculations with
393 the Genassemblage interface is provided in the Genassemblage manual.

394

395 **Analysis of output data**

396 The data from spreadsheets entitled "Heterozygosity", "Weak heterozygotes" and "Alleles
397 inherited" was analysed directly in an output file using MS Excel. The average and standard
398 deviation (SD) were calculated by using Excel.

399

400 The values for heterozygosity, the share of "weak heterozygotes" and allelic diversity that could
401 be expected in groups of progeny from a potential breeding pair were used to calculate a *v* index.
402 For each potential pairing, this index = 0 if there are no genetic differences between all
403 individuals in the broodstock; the index ≈ 0 if both parents have the same genetic profile and are
404 homozygotic at all investigated loci; and the index = 1 if the parents give the highest values of
405 heterozygosity, allelic diversity and the lowest share of "weak heterozygtic genotypem" of all
406 possible pairs. The values of the *v* index increase in proportion to the differences between the
407 genetic profiles of the parental organisms .

408

409 The calculation of the *v* index takes into account the relative importance that the breeder places
410 on the indicators of genetic variation. In addition, when working with a diploid species, the share
411 of "weak heterozygotes" is not taken into account in the calculation. Here, the breeder considered
412 allelic diversity and heterozygosity more important than a low share of "weak heterozygotes".
413 The relative importance of the indicators was expressed in importance coefficients (i); the
414 importance coefficient of heterozygosity (i_H) was 0.40; of allelic diversity (i_a), 0.40; and of the
415 share of "weak heterozygotes" (i_{wh}), 0.20; if this was a diploid species, i_{wh} would be 0. The *v*
416 index was calculated by using Algorithm 4.

417 $$v = i_H \left[\frac{H_n}{H_{\max}} \right] + i_{wh} \left[\frac{wh_{\max} - wh_n}{wh_{\max}} \right] + i_{ar} \left[\frac{ar_n}{ar_{\max}} \right] \qquad \begin{matrix} i_H + i_{wh} + i_{ar} = 1 \\ wh_{\max} > 0 \end{matrix}$$

418 Algorithm 4. Calculation of the *v* index.

419 In this algorithm, H_n, wh_n, and ar_n, are the values of the indicators of genetic variation expected
420 in the progeny of each potential pairing, and H_{max}, wh_{max}, and ar_{max} are the maximal values of
421 those indicators detected in all analysed pairings.

422

423 To test the significance of the differences in the average values of genetic variation indicators
424 between the set of four pairs chosen as optimal and the average values for all possible pairings, a
425 one tailed Mann-Whitney test was used, with significance set at p = 0.05. The calculations were
426 performed using STATISTICA 10.0 software (Statsoft USA).

427

428 **Results**

429 Figure 3 shows the expected heterozygosity of the progeny of each potential pair of spawners, as
430 calculated by Genassemblage. This value ranges from 0.49 to 0.97, and the best 4 values of this
431 indicator are calculated for the following pairs: female C40 x male C42, female C40 x male B20,
432 female C40 x male C41, and female B19 x male A01. The average heterozygosity was 0.78 with
433 a standard deviation of 0.12.

434

435 Figure 4 shows the expected share of "weak heterozygous genotypes" for each potential pairing.
436 The values range from 0.04 to 0.58 and the average is rather high (0.42, SD +/- 0.12). The best 4
437 values of this indicator were calculated for female C38 x male B05, female C38 x male B20,
438 female A07 x male B20, and female C38 x male A01.

439

440 The number of inherited alleles at investigated loci ranged from 8 to 14 (Figure 5). The average
441 number of alleles in all possible groups of progeny iss 10. Keeping in mind that a single
442 individual should not be bred twice to avoid inbreeding, the following are the best breeding pairs
443 in terms of allelic richness: In the progeny of female C38 x male B05, the number of inherited
444 alleles is calculated to be 14; in the progeny of female A07 x male C42, it is 13; with both female
445 A14 x male B20 and female B19 x male A01, it was 11.

446

447 The calculation of the *v* index – which in this case, give twice as much weight to heterozygosity
448 (i_H) and allelic diversity (i_a) as to the share of „weak heterozygotes" (i_{wh}) – show that some
449 pairings are better for maintaining genetic variation than others (Figure 6). Once again, each
450 individual was only bred once to increase genetic variation and reduce the risk of inbreeding. The
451 average value of this index for all possible pairs is 0.68 with a SD of 0.10. The minimal value of
452 the index is 0.49 and the maximal value was 0.95. The highest genetic diversity is expected in the
453 progeny of female C38 and male B05 (0.95). In addition, other pairs indicated as appropriate for
454 maintaining genetic variation are female A07 x male C42 (v = 0.80), female C40 x male B20 (v
455 = 0.75) and female B19 x male A01 (0.73).

456

457 The four pairs selected by using the *v* index show better values for all of the above indicators; the
458 significance of the difference is assessed by using a Mann-Whitney test. The expected
459 heterozygosity was 0.90), as opposed to 0.78 ($p < 0.05$); the share of "weak heterozygous
460 genotypem" is 0.31, as opposed to 0.42, but this difference is only significant at a p value of 0.10
461 because of the high standard deviation; the number of inherited alleles is 12, as opposed to 10 (p
462 < 0.05); and a *v* index calculated just for these 4 pairs is 0.80, as opposed to 0.68 ($p < 0.05$).

463

464 **Discussion**

465 As shown in the above example of the use of Genassemblage, the calculations performed by the
466 software facilitate the identification of an optimal set of pairs for maintaining genetic variation in
467 a breeding stock. In this example, the potential offspring of the four pairings would have better
468 than average indicators of genetic variation. Although it would have been possible to create sets
469 of pairs with a higher average *v* index, average heterozygosity and average number of alleles, this
470 would have required the use of one individual in two or more pairings. It is better to use different
471 individuals in each pairing to increase genetic variation and reduce the risk of inbreeding.
472
473 When using data from Genassemblage to calculate a *v* index, the relative weight given to the
474 coefficients of heterozygosity (i_H), of allelic diversity (i_{ar}), and of the share of “weak
475 heterozygotes” (i_{wh}) can be adjusted. Because the heterozygosity represented by (i_H)and allelic
476 diversity (i_{ar}) are the most important parameters used in quantification of the genetic variability
477 (Nei 1973), and evolutional potential of populations (James 1971, Hill and Rasbash 1986) , they
478 had been weigh twice over (i_{wh}). For example, when working with a diploid species, (i_{wh}) would
479 be 0.
480

481 In addition to the identification of optimal sets of breeding pairs, Genassemblage can also be
482 used for other tasks related to the management of genetic variation in breeding stocks or human-
483 dependant populations, for example, in maintaining the genetic variability in progeny in
484 hatcheries. The simplicity of the software and the fact that it presents data as tables in a
485 Microsoft Excel spreadsheet enables easy presentation and ranking of results, the use of filters
486 for selecting desirable values of indicators, and further analysis of data with a wide range of
487 mathematical and statistical tools.
488
489
490

491 **Acknowledgments:** I would like thank to: Dr Pawel Brzuzan and Dr. Mirosław Luczynski for
492 their help in beta testing of this software, and all helpful suggestions applied in preparation of
493 this manual.
494

495

496 **References**

497 **Bryant E.H., McCommas S.A., Combs L.M.** (1986). The effect of an experimental
498 bottleneck upon quantitative genetic variation in the housefly. *Genetics,* 114: 1191-
499 1211.
500 **Dieringer D., Schlötterer C.** (2003). Microsatellite analyzer (MSA): a platform-
501 independent analysis tool for large microsatellite data sets. Ecology Notes*,* 167-169.
502 **Excoffier L, Lischer H.E.** (2010). Arlequin suite ver 3.5: a new series of programs to
503 perform population genetics analyses under Linux and Windows", *Molecular Ecology*
504 *Resour*ces, 10:564-567
505 **Excoffier L., Laval G., Schneider S.** (2005). Arlequin ver. 3.0: An integrated software
506 package for population genetics data analysis. Bioinformatics Online*,* 47-50.
507 **Hill, W. G., Rasbash J.** (1986). Models of long term artificial selection in finite
508 populations. *Genetic Resources* 48: 41–50.
509 **James J. W.** (1971). The founder effect and response to artificial. *Genetic Resources*
510 12: 249–266.
511 **Kaczmarczyk D., Fopp-Bayat D**. (2013). Assemblage of spawning pairs based on their
512 individual genetic profiles – as tool for maintaining genetic variation within sturgeon
513 populations. *Aquaculture Research* 44: 677–682.
514 **Kim J.E., Withler R.E., Ritland C., Cheng K.M.** (2004). Genetic variation within and
515 between domesticated chinook salmon, *Oncorhynchus tshawytscha*, strains and
516 their progenitor populations. *Developments in Environmental Biology of Fishes*, 23:
517 371-378.
518 **Koljonen M-L., Jansson H., Paaver T., Vasin O., Koskiniemi. J.** (1999).
519 Phylogeographic lineages and differentiation pattern of Atlantic salmon in the Baltic
520 Sea with management implications. *Canadian Journal of Fisheries and Aquatic*
521 *Sciences*, 56**:** 1766–1780.
522 **Koljonen M-L., Tähtinen J., Säisä M., Koskiniemi J.** (2002). Maintenance of genetic
523 diversity of Atlantic salmon (*Salmo salar*) by captive breeding programs and the
524 geographic distribution of microsatellite variation. *Aquaculture*, 212: 69-92.

525 **Koljonen M-L., Tähtinen J., Säisä M., Koskiniemi J.** (2002). Maintenance of genetic
526 diversity of Atlantic salmon (*Salmo salar*) by captive breeding programs and the
527 geographic distribution of microsatellite variation. *Aquaculture*, 212: 69-92.
528 **Kosowska B., Nowicki B.** (1999). Genetic similarity. Pedigree. W: Genetyka
529 weterynaryjna p. 237-246 *Wydawnictwo Lekarskie PZWL*, Warszawa 312 pp. (in
530 Polish)
531 **Nei M.** (1973). Analysis of gene diversity in subdivided populations. *Proceedings of*
532 *National Academy of Science*. USA 70: 3321–3323.
533 **Verspoor E., Beardmore J.A., Consuegra S., García de Léaniz C., Hindar K.,**
534 **Jordan W.C., Koljonen M-L., Mahkrov A.A., Paaver T., Sánchez J.A., Skaala**
535 **Ø., Titov S., Cross T.F.** (2005). Population structure in the Atlantic salmon:
536 insights from 40 years of research into genetic protein variation. *Journal of Fish*
537 *Biology,* 67:Suppl A, 3–55.
538
539