

AutoBin v 0.9

AutoBin is Excel Macro written in Microsoft Visual Basic (VBA) and has been developed by Franck Salin (INRA Pierroton-UMR BIOGECO). It automatically analyzes raw data generated from microsatellites genotyping software such as STRand (Toonen *et al.*, 2001) or GeneMapper (Applied Biosystems, USA). It can deal with unlimited number of loci and samples, with no consideration of the type of SSRs motifs. AutoBin helps the user to bin his data with visual alerts and format automatically the data for downstream analysis.

Input format:

Input data are in the form of four columns with fixed fields. As a consequence, data must always follow the example below:

	A	B	C	D	
1	Sample	Marker	Size1	Size2	
2	Sample_1	Locus1	211.04	216.92	
3	Sample_2	Locus1	211.05	211.05	
4	Sample_3	Locus1	211.23	212.16	
5	Sample_4	Locus1	212.03	212.03	
6	Sample_1	Locus2	106.24	109.96	
7	Sample_2	Locus2	106.24	109.91	
8	Sample_3	Locus2	106.28	109.84	
9	Sample_4	Locus2	106.29	109.89	
10	Sample_1	Locus3	100.84	105.63	
11	Sample_2	Locus3	102.51	105.57	
12	Sample_3	Locus3	102.56	105.66	
13	Sample_4	Locus3	102.58	105.69	
14	etc...	etc...	etc...	etc...	
15					

Column A: Sample name

Column B: Marker name

Column C: Allele 1 in raw size

Column D: Allele 2 in raw size

Column headings must not be changed by the user. In some software, homozygous genotypes are missing one allele, so AutoBin automatically detects this problem and duplicates the first allele. Note that there are no limitations for the number of samples and loci, except the intrinsic limits of Excel. Names used for samples or loci are free.

Step by step execution of AutoBin

Step 1:

By clicking the 'Format' button on sheet 'data', the number of samples and loci is automatically detected and genotypes for each locus are copied and pasted to a new sheet, with the locus name.

Step 2:

By clicking the 'Binning' button on sheet 'data', for each locus, data are formatted from two to one column format (two lines per samples, on column by locus). Ascending sort of alleles by raw size is done automatically.

Step 3:

By clicking the 'Validation' button on sheet 'data', for each locus, size difference between contiguous alleles is calculated. At this step, the user chooses two thresholds for gap detection. In red will appear strong gaps, in orange will appear moderate gaps for further checking. A conditional formatting of allele in raw sizes is done and the user can easily visualize the putative gaps (defined above). An arbitrary binning is proposed for each locus in column E. The user must check the binning with caution and if necessary modify it in column F (at least one value for each bin, which is copied automatically to other alleles).

Step 4:

By clicking the 'Graph' button on sheet 'data', for each locus, a graph plot of cumulative length distribution is done, with automatic adjustment of vertical scale to omit missing data and optimize the interpretation. This step can be useful to visualize ambiguous samples.

Step 5:

By clicking the "Format" button on sheet 'data', three output files are generated, for further analysis with STRUCTURE (Pritchard *et al.*, 2000) or GENEPOP (Raymond *et al.*, 1995):

- STRUCTURE format (two lines by sample – one column by locus)
- STRUCTURE format (one line by sample – two columns by locus)
- GENEPOP format (one line by sample – two columns by locus)

From GENEPOP format, the user can reformat his data for commonly used population genetic software using CONVERT (Glaubitz, 2004).

IMPORTANT

At any time during the procedure, data can be erased by clicking the 'Erase Data' button.

At any time during the procedure, graph can be erased by clicking the 'Erase Graph button.

References:

- Glaubitz JC (2004) CONVERT: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Molecular Ecology Notes* **4**, 309-310.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
- Raymond M, Rousset F (1995) GENEPOP (Version 1.2) - Population-genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**, 248-249.
- Toonen, Robert J, Hughes, Shayne (2001) *Increased throughput for fragment analysis on an ABI PRISM 377 automated sequencer using a membrane comb and STRand software*
Eaton, Natick, MA, ETATS-UNIS.