Patrick Kück & Carola Greve

AMARE v1.3beta - Manual

Contents

List of Figures

List of Tables

1 Introduction

AMARE is designed to optimize the signal-to-noise ratio in AFLP data sets (single or multiple primer combinations). It identifies potentially erroneous AFLP genotyping errors of GeneMapper (Applied Biosystems) or CEQTM System Fragment Analysis v.9.0.25 (Beckman Coulter) as the number of unreproducibly scored markers between replicated AFLP profiles of single individuals. As a measure of overall data quality, it simultaneously estimates general replicate error rates. Strength and accuracy of the approach depend on the number of replicates and whether they are representative for the whole data set.

As input file, AMARE reads exported AFLP binary character matrices in either GeneMapper (Applied Biosystems) or CEQ^{TM} System Fragment Analysis v. 9.0.25 (Beckman Coulter) table format. A mixture of both formats is not possible. AMARE concatenates multiple AFLP matrices and then analyzes the concatenated supermatrix in one process. The concatenated, unreduced supermatrix is printed out as separate text (.txt) file. Missing sample data in single matrices are replaced by "?", which are not further considered in the analyses. For each individual threshold set, AMARE generates i) a single log file reporting the masking of bins and replicates and ii) a character matrix in text (.txt) and nexus (.nex) format, if error rate conditions and minimum number of remaining bins are met. A summary of all threshold sets and corresponding error rates are stored in the main log file. Nexus files can be directly executed in other programs like `PAUP`. Additionally, AMARE plots a graphical overview of the original and each masked replicate matrix.

AMARE was written on Linux and works on WindowsPCs, Mac OS and Linux running systems. Input files originating from Windows, CRFL line feeds should be converted into Unix (LF) line feeds in advance, especially, if the user changes the operating system. This can be done in several editors like e.g. `Bioedit`, `Notepad++` or `Scite`. AMARE usually replaces them, but might not succeed in every instance.

2 Start AMARE

To run AMARE, open the terminal of your running system. Navigate through your directory path to the folder where AMARE and input files are placed. Notice that *all input files* have to be located in the AMARE including folder. To execute AMARE, a Perl interpreter must be installed on the current system. Linux and Mac OS systems do not need a subsequent installation because the Perl interpreter is usually included as a standard tool. Unfortunately, Windows users have to install a Perl interpreter *ex post*. We recommend the `ActivePerl` interpreter which can be downloaded for free under:

- http://activeperl.softonic.de/

2.1 Command Line

AMARE has to be started directly via command line. Therefore, commands and chosen options have to be typed in one row in the terminal. Starting AMARE via command line simplifies the implementation of AMARE into complex process pipelines. Navigate through your directory path to the folder where AMARE and your files are located and type the name of the AMARE version, followed by a blank and the demanded setup options with a minus (-) sign in front of each. Then press <enter>. Make sure you write

the input setup correctly, for example '-i' and not '- i'. Otherwise AMARE will not start working. Instead AMARE will open a synopsis menu (figure 1) with a short command description and a few examples. The command order is arbitrary. For example:

- C:\AMARE_Folder> perl AMARE_v1.3beta.pl **-help** <enter>
- C:\AMARE_Folder> perl AMARE_v1.3beta.pl **-i** infile.txt **-d** 0.25 **-l** 88.0 **-e** j **-t** a **-r** 0<enter>

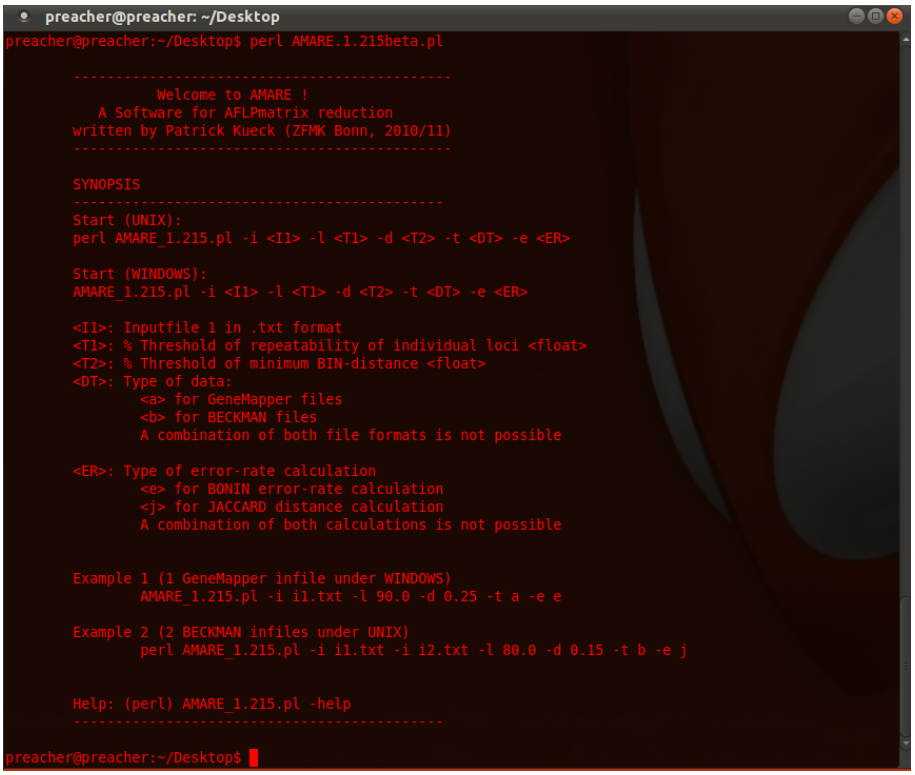

Figure 1: AMARE SYNOPSIS

3 Options

Except of the '-r' command, AMARE needs all setup options described above in one command line. Except of the '-i' command, multiple commands of the same type are not allowed. AMARE skips to the synopsis menu and ends with a specified error prompt if unknown or multiple command options are used or if a required option lack. Usable options are listed in table 1.

Table 1: AMARE setup options

General options	**Command**	
Help menu	-help	
Preface	-a	
Synopsis	-s	
Setup options	**Command**	**Suffix**
Defined Input File	-i	string.txt
Minimum Bin Distance	-d	floats
BIN Reliability	-l	floats
Type of Error Rate	-e	e/j
Reject 0|0 BIN States	-r	0/1
Datatype	-t	a/b

3.1 Infile (-i Option)

AMARE is able to handle two different infile datatypes (.txt), which have to be in either GeneMapper or Beckman Coulter table format. The software can handle one or multiple infiles in one single process run using the '-i' command multiple times (for each input file once). Infiles have to be given to AMARE with the corresponding TEXT associated suffix '.txt' (e.g. -i infilename_1.txt -i infilename_2.txt). Sample names have to consist of only alphanumeric signs and underscores (_). AMARE will issue an error prompt and abort if any non-alphanumeric sign is encountered within sample names.

3.1.1 Concatenation of Multiple Table Infiles

The datatype of multiple infiles have to be either in GeneMapper or Beckman Coulter table format. AMARE checks the sample names of each infile and recognizes missing samples in single files by name comparison. Therefore sequence names of equal samples within different infiles have to be typed completely unique (upper and lower case included). Be aware of presents and absents of double quotes in sample names. The number of replicates can be different within single infiles, but have to be a multiple of two in total. AMARE will abort with an error prompt if not. Note, that missing sample data will be replaced by question marks in the concatenated matrix and not further considered by AMARE.

3.1.2 GeneMapper Table Format

GeneMapper table formats have to be saved as tabstop delimited TEXT files. The first line of each GeneMapper table format has to be started with a space and a tabstop, followed by single tabstop delimited BIN numbers. The second line has to be started with 'Sample Name' followed by single user specified, tabstop delimited sample names. The third line will not be taken into account by AMARE. The only conditions for the third lines is, that it has to start with the word 'Dye'. The following lines comprise the real BIN matrix with 'Allele' and 'Size' labeled lines followed by numbers either as single

numbers or enclosed by brackets. Table 2 shows an example of GeneMapper table format which will be recongnized by AMARE.

Table 2: Example of readable GeneMapper table format

Column 1		**Column 2**		**Column 3**			**Column n**
¡Space¿	<TAB>	1	<TAB>	2	...	<TAB>	n+1
Sample Name	<TAB>	Sample_1	<TAB>	Sample_2	...	<TAB>	Sample_n+1
Dye	<TAB>	B	<TAB>	B	...	<TAB>	B
Allele 1	<TAB>	50	<TAB>	57	...	<TAB>	70
Allele 2	<TAB>	55(2)	<TAB>	60(2)	...	<TAB>	72(2)
⋮	⋮	⋮	⋮	⋮		⋮	
Allele n	<TAB>	86(1)	<TAB>	90(1)	...	<TAB>	94(1)
Size 1	<TAB>	50	<TAB>	57	...	<TAB>	60
Size 2	<TAB>	55(2)	<TAB>	60(2)	...	<TAB>	65(2)
⋮	⋮	⋮	⋮	⋮		⋮	⋮
Size n	<TAB>	86(1)	<TAB>	90(1)	...	<TAB>	93(1)

3.1.3 Beckman Coulter Table Format

Single columns within each line must be saved as comma separated TEXT file. Except of the first column and the second line ('Dye'), the Beckman Coulter table format consists of only numerical signs. Single BIN states are represented as single integers, most suitable 0 or 1. Table 3 shows an example of typical Beckman Coulter table format recognizable by AMARE. The existence of double quotes is not mandatory.

Table 3: Example of readable Beckman Coulter table format

Column 1		**Column 2**		**Column 3**			**Column n**
"BIN"	,	55	,	56	, ... ,		83
"Dye"	,	"D3"	,	"D3"	, ... ,		"D3"
"Samples"	,	59	,	61	, ... ,		117
Fragments"	,	59	,	61	, ... ,		117
"XMin"	,	55.01	,	56.01	, ... ,		65.91
"XMax"	,	55.75	,	56.75	, ... ,		66.62
"XMean"	,	55.21	,	56.21	, ... ,		66.29
"XVar"	,	0.03	,	0.02	, ... ,		0.02
"YMean"	,	1343	,	1353	, ... ,		1830
"BIN_1"	,	0	,	1	, ... ,		1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
"BIN_n"	,	1	,	0	, ... ,		1

Table 4: A good declaration

Sample Names	**Defined as...**
aSample_1	Replicate 1
aSample_2	Replicate 1
aNextPair_1	Replicate 2
aNextPair_2	Replicate 2
anotherPair_1	Replicate 3
anotherPair_2	Replicate 3
aLast_one_1	Replicate 4
aLast_one_2	Replicate 4
ASingleSample	Single 1
⋮	⋮
A_notherSample	Single 2

Table 5: A bad declaration

Sample Names	**Defined as...**
aSample_1	Replicate 1
aNextPair_1	Replicate 2
aNextPair	Replicate 3
aSample_2	Replicate 1
anotherPair_a	Replicate 4
ALast_one_1	Single Sample 1
anotherPair_b	Replicate 4
aLast_one_2	Replicate 5
aSingleSample	Replicate 6
⋮	⋮
a_notherSample	Replicate 5

3.1.4 Replicates

Except of their sample specific suffix, all replicate names have to be named identically and to be labeled by a lowercased '**a**' prefix. Replicates which belong together should be arranged successively in each inputfile. AMARE identifies replicates by their lowercased '**a**' prefix and assigns them together by unique name identity. Samples which are not represented as replicate are not allowed to begin with a lowercase '**a**' prefix. The software aborts with an error prompt if AMARE identifies an impair number of replicates. Table 4 gives an example of a good replicate declaration whereas table 5 shows an example of a bad replicate declaration. The respective sample declaration identified by AMARE for each of the two sample definitions is given in the right column of each example.

3.2 Minimum BIN Distance (-d Option)

The Minimum Bin Distance is a user specified threshold which determines the minimum acceptable BIN distance between differently sized BINs. BINs are removed if the inferred BIN distance is at or below the specified Minimum BIN Distance. The Minimum BIN Distance has to be specified as float number by using the '-d' option (e.g. -d 0.25).

3.3 Minimum BIN Reliability Threshold (-l Option)

The BIN Reliability defines the relative number of reproducible BIN states over all replicates which can range between 0 and 1. The user specified BIN Reliability Threshold sets the acceptance value of the minimum number of reproducible (0,0) and (1,1) BIN states. BIN states at or below the predefined threshold are removed.

3.4 Error Rate (-e Option)

The Error Rate reflects the quality of the replicates. The higher the Error Rate, the higher the proportion of unreproducible markers among replicates. Two different Error Rates can be specified. Error Rates are printed in the main logfile if $r_{error} < 0.1$. If

$r_{error} > 0.1$, a reduced matrix will not be generated. Replicates which include missing data ('?') are not considered into Error Rate calculations.

3.4.1 Mismatch Error Rate r_{BONIN} (-e e)

The replicate Mismatch Error Rate r_{BONIN} is defined as the relative number of unreproducible N (0,1) and N (1,0) summed over all n replicates.

$$r_{BONIN} = \frac{N_x(0,1) + N_x(1,0)}{N_x(0,0) + N_x(0,1) + N_x(1,0) + N_x(1,1)} \quad (1)$$

3.4.2 Mismatch Error Rate $r_{JACCARD}$ (-e j)

The average JACCARD Mismatch Error Rate $r_{JACCARD}$ divides the number of unreproducible N (0,1) and N (1,0) markers by the sum of reproducible N (1,1), unreproducible N (0,1) and N (1,0) markers.

$$r_{JACCARD} = \frac{N_x(0,1) + N_x(1,0)}{N_x(0,1) + N_x(1,0) + N_x(1,1)} \quad (2)$$

3.5 Reject 0|0 BIN States (-r Option)

AMARE removes all BINs without any (1,1) pairs among replicates to avoid spurious background noise in the data. If BINs of only (0,0) pairs among replicates should not be exclusively removed by default use the '-r 0' option.

3.6 Datatype (-t Option)

The input table format has to be specified by using the '-t' option. The '-t a' command defines GeneMapper matrices while '-t b' defines Beckman Coulter matrices. AMARE does not allow handling of both kinds of input matrices in one single process run.

4 Output Files

Output files are placed in three AMARE generated subfolders. All file names are declared by the prefix 'AMARE' continued by a content name (matrix, log or svgplot) and the associated parameter combination of Minimum BIN Reliability and Replicate Reliability Threshold (e.g. AMARE_matrix_90_80.txt). Additionally, SVG-plots include the suffix reduced or unreduced depending on which matrix is concerned (e.g. AMARE_svgplot_90_80_reduced.svg).

4.1 AFLP Matrices

Reduced AFLP matrices of all parameter settings are given as separate TEXT (.txt) and NEXUS (.nex) files if the obtained Error Rate is below 0.1 and the number of remaining BINs higher as 5. All printed TEXT matrices are Beckman Coulter table formatted. The NEXUS output can be used for further analyses like e.g. distance analyses in *PAUP*. Both file types are printed in the new subfolder 'AMARE_matrices/'. If multiple

infiles are used, this subfolder contains also the concatenated, unreduced supermatrix 'AMARE_supermatrix.txt'.

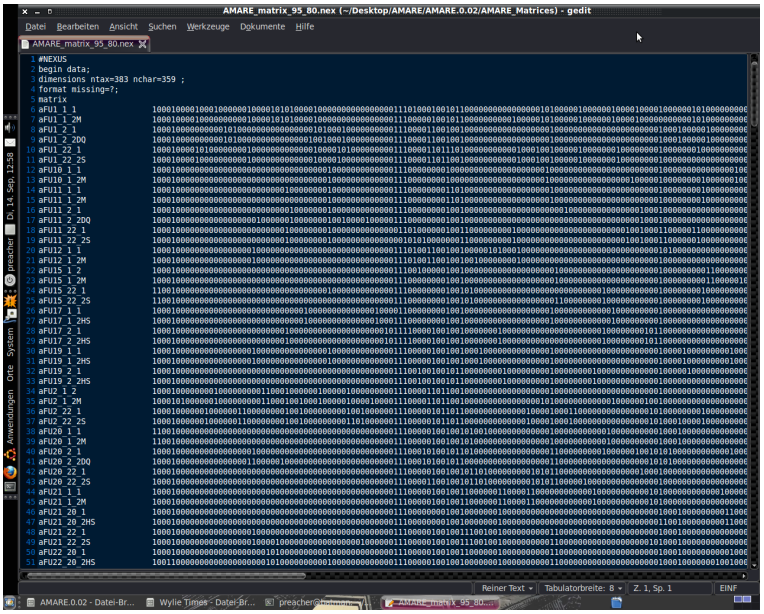

Figure 2: Cutout of a NEXUS matrix

4.2 AMARE Logfiles

The new subfolder 'AMARE_logfiles/' includes a logfile for each analysed parameter setting. Each logfile contains BIN and replicate information about each reduction step, and a main logfile 'AMARE_main_log.txt'. The main logfile gives an overview about each parameter setting and its corresponding results. Associated with each reduced matrix the main logfile lists also the number of remaining BINs and replicates, the total number of remaining characters, the Error Rate value, and the filename of the associated matrix.

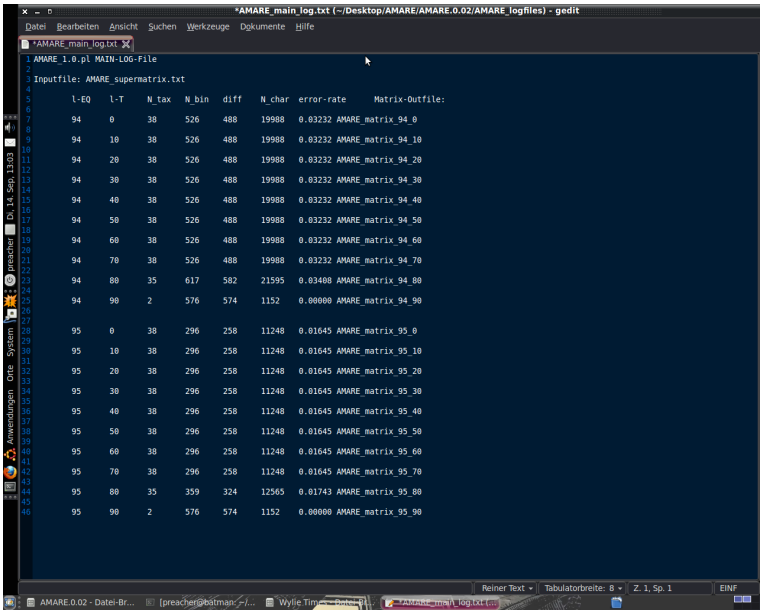

Figure 3: Cutout of a main logfile content

4.3 AMARE Vector Plots

Graphical plots of each reduced and unreduced matrix of a given Minimum BIN Reliability threshold are printed in the new subfolder 'AMARE_svg/'. Incongruent states of replicates are coloured red, equal (0,0) states are shown in light blue, equal (1,1) states in dark blue, and states of missing data in white. Replicates are listed vertically, BINs horizontally. Figure 4 and 5 give cutout examples of an unreduced and reduced matrix svg plot.

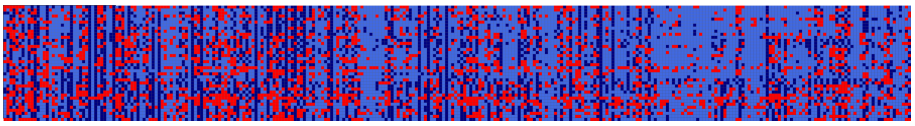

Figure 4: Vector plot of an unreduced matrix

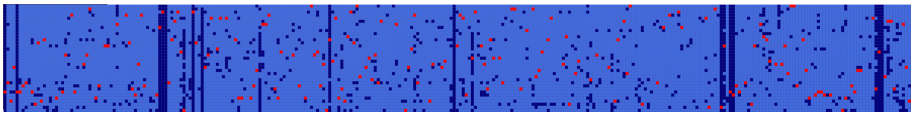

Figure 5: Vector plot of a reduced matrix

5 Error Reports

Each error allocates AMARE to stop all running processes; AMARE will abort with an specified error message.

5.1 List of possible COMMAND-ERRORS

After starting the program, AMARE checks the command line syntax. If something is wrong, AMARE opens the synopsis menu and aborts with a COMMAND-ERROR prompt. This subsection gives a short explanation about possible COMMAND-ERROR prompts.

5.1.1 COMMAND-ERROR: Missing parameter for option: *-option* !

Parameter values have to be specified for the '-d', '-l', '-i', '-e' option.

5.1.2 COMMAND-ERROR: Multiple usage of option: *-option* !

Except of the '-i' option, all options can only be used once.

5.1.3 COMMAND-ERROR: Unknown command: *command* !

The listed command can not be identified by AMARE. For further explanation see subsection 'Command Line' and table 1.

5.1.4 COMMAND-ERROR: Unknown type of specified table format: *command* !

Two different input table formats can be specified. Either GeneMapper formatted (-t a) or Beckman Coulter formatted (-t b) tables. See subsection 'Datatype (-t Option)'.

5.1.5 COMMAND-ERROR: Unknown type of specified ERROR RATE: *command* !

Two different Error Rates can be specified. Either r_{BONIN} (-e e) or $r_{JACCARD}$ (-e j). See subsection 'Error Rate (-e Option)'.

5.1.6 COMMAND-ERROR: Unknown type of specified BIN state handling: *command* !

BINs of only (0,0) pairs among replicates are exclusively excluded by default. To disable this option type '-r 0'. See subsection 'Reject 0|0 BINs (-r Option)'.

5.1.7 COMMAND-ERROR: Unknown option: *command* !

See section 'Options'.

5.1.8 COMMAND-ERROR: Wrong Minimum BIN Distance Threshold: *command* !

The Minimum BIN Distance Threshold has to be given as float (e.g. '-d 0.25'). See subsection 'Minimum Bin Distance (-d Option)'.

5.1.9 COMMAND-ERROR: Wrong Minimum BIN Reliability Threshold: *command* !

The Minimum BIN Reliability Threshold has to be given as float (e.g. '-l 80.0'). See subsection 'Minimum BIN Reliability Threshold (-l Option)'.

5.2 List of possible FILE-ERRORS

AMARE checks each input table of correct format and forbidden characters. This subsection gives a short explanation about possible FILE-ERROR prompts.

5.2.1 FILE-ERROR: Cannot open *infile* !

Possible reasons for that error prompt could be a wrong written filename (Be aware of case-sensitivity), a forgotten '.txt' suffix or absence of the infile in the AMARE-software folder.

5.2.2 FILE-ERROR: Cannot find predefines replicates !

The AMARE analysis depends on replicates. Replicates have to be defined by a lower case 'a' at the beginning of the replicate name (e.g. aSamplename or a_Samplename). The AMARE algorithm cannot start without predefined replicates. See subsection 'Replicate' and table 4 and 5.

5.2.3 FILE-ERROR: Impair number of predefined replicates !

If using a single infile, replicates have to be double samples. If multiple infiles are used, AMARE checks all infiles for predefined replicates. Maybe one or more replicates are not predefined with an 'a' as name prefix (be aware of lower case) or maybe single samples have an 'a' as name prefix and are therefore recognized as replicate (single samples are not allowed to start with a lower case 'a'). See subsection 'Replicate' and table 4 and 5. Note: AMARE prints a list of predefined replicates for which no other predefined replicate can be assigned ('FILE-ERROR: Unpaired replicates: *replicate name* !').

5.2.4 FILE-ERROR: *infile* is empty !

The named infile has no content.

5.2.5 FILE-ERROR: Line 1 of table *infile* is not in GeneMapper table format !

If GeneMapper formatted tables, the first line has to be startet with a space and a 'tab' followed by tab delimited integers. At the end of the line can either be a 'tab' or an integer. Only numbers and tabstops are allowed. See subsection 'GeneMapper Table Format' and table 2.

5.2.6 FILE-ERROR: Line 2 of table *infile* is not in GeneMapper format !

In GeneMapper formatted tables, the second line has to be startet with 'Sample Name' followed by tab delimited words composed of alphanumeric signs. At the end of the line can either be a 'tab' or a word. See subsection 'GeneMapper Table Format' and table 2.

5.2.7 FILE-ERROR: Line 3 of table *infile* is not in GeneMapper format !

In GeneMapper formatted tables, the third line has to start with the word 'Dye' followed by a tabstop. Following signs in line three are not recognized by AMARE. See subsection 'GeneMapper Table Format' and table 2.

5.2.8 FILE-ERROR: Unequal number of columns in line *linenumber* of GeneMapper table *infile* !

In GeneMapper formatted tables, the number of columns in each line has to be constant. See subsection 'GeneMapper Table Format' and table 2.

5.2.9 FILE-ERROR: Forbidden signs in line *linenumber* of GeneMapper table *infile* !

This error prompt refers to an specific allele or size line. Each of such lines has to start with the word 'Allele' or 'Size', followed by a continuous numbering of tabstop delimited allele or size values (parenthesis allowed). See subsection 'GeneMapper Table Format' and table 2.

5.2.10 FILE-ERROR: Sample *samplename* appears more than once in table *infile* !

AMARE does not allow equal sample names within the same infile. If multiple infiles are used be aware that sequences from same samples share identic names in all infiles (be aware of case-sensitive). AMARE can not concatenate multiple sequences of identic samples if the sample names are different. See subsections 'Infile (-i Option)', 'Concatenation of Multiple Table Infiles', and 'Replicates'.

5.2.11 FILE-ERROR: Wrong format of Bin-line in Beckman Coulter table *infile* !

The first line (BIN-line) of Beckman Coulter formatted tables has to start with the word 'BIN', followed by comma separated integers. Double quotes are not recognized (e.g. BIN,1,2,3...). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.12 FILE-ERROR: Wrong format of Dye-line in Beckman Coulter table *infile* !

The second line (Dye-line) of Beckman Coulter formatted tables has to start with the word 'Dye', followed by comma separated alphanumeric signs. Double quotes are not recognized (e.g. Dye,D3,D3,D3...). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.13 FILE-ERROR: Wrong format of Samples-line in Beckman Coulter table *infile* !

The third line (Samples-line) of Beckman Coulter formatted tables has to start with the word 'Samples', followed by comma separated integers. Double quotes are not recognized (e.g. Samples,1,2,3...). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.14 FILE-ERROR: Wrong format of Fragments-line in Beckman Coulter table *infile* !

The fourth line (Fragments-line) of Beckman Coulter formatted tables has to start with the word 'Fragments', followed by comma separated integers. Double quotes are not recognized (e.g. Fragments,1,2,3...). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.15 FILE-ERROR: Wrong format of XMin-line in Beckman Coulter table *infile* !

The fifth line (XMin-line) of Beckman Coulter formatted tables has to start with the word 'XMin', followed by comma separated integers or floats. Double quotes are not recognized (e.g. XMin,1,1.3,1.76...). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.16 FILE-ERROR: Wrong format of XMax-line in Beckman Coulter table *infile* !

The sixth line (XMax-line) of Beckman Coulter formatted tables has to start with the word 'XMax', followed by comma separated integers or floats. Double quotes are not recognized (e.g. XMax,1.1,1.5,2...). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.17 FILE-ERROR: Wrong format of XMean-line in Beckman Coulter table *infile* !

The seventh line (XMean-line) of Beckman Coulter formatted tables has to start with the word 'XMean', followed by comma separated integers or floats. Double quotes are not recognized (e.g. XMean,1.2,1.8,2...). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.18 FILE-ERROR: Wrong format of XVar-line in table *infile* !

The eighth line (XVar-line) of Beckman Coulter formatted infiles has to start with the word 'XVar', followed by comma separated integers or floats. Double quotes are not recognized (e.g. XVar,1,2.2,3...). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.19 FILE-ERROR: Wrong format of YMean-line in Beckman Coulter table *infile* !

The ninth line (YMean-line) of Beckman Coulter formatted tables has to start with the word 'YMean', followed by comma separated integers or floats. Double quotes are not recognized (e.g. YMean,1,1.8,2...). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.20 FILE-ERROR: Wrong format of sample-line *number* in Beckman Coulter table *infile* !

Sample lines of Beckman Coulter formatted tables have to start with a sample name of alphanumeric signs, followed by comma separated integers (BINs). Double quotes are not recognized in sample names. BINs have to be without double quotes (e.g. Sample_1,50,60,80...). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.21 FILE-ERROR: Unequal number of columns in line *linenumber* of Beckman Coulter table *infile* !

In Beckman Coulter tables, the number of columns in each line has to be constant (either comma separated or tabstop delimited). See subsection 'Beckman Coulter Table Format' and table 3.

5.2.22 FILE-ERROR: Wrong format of BIN states in Beckman Coulter table *infile* !

In Beckman Coulter tables, BIN lines have to start with the sample name, followed by comma separated BIN values which have to be either integers or '?' (e.g. taxon,1,?,3...). See subsection 'Beckman Coulter Table Format' and table 3.

6 License/Help-Desk/Citation

AMARE v1.3beta was developed by Patrick Kück, Carola Greve, France Gimnich, and Bernhard Misof. The software was written in PERL by Patrick Kück in 2010. It can be distributed and/or modified under the terms of the GNU General Public License as published by the Free Software Foundation; either 2 of the license, or (at your option) any later version.
This program is distributed with the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details. You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 675 Mass Ave, Cambridge, MA 02139, USA.

If you have any problems, error-reports or other questions about AMARE, feel free to write an email to patrick_kueck@web.de which is the official help desk email account for AMARE.

For further free downloadable programs from our institute visit:
http://software.zfmk.de.

If you use AMARE please cite:

Kück P., Greve C., AMARE, Version 1.0, Zool. Forschungsmuseum A. Koenig, Germany, 2010

7 Copyright