INEST 2.3

The user manual

written by Igor J. Chybicki

e-mail: igorchy@ukw.edu.pl

homepage: www.ukw.edu.pl/pracownicy/strona/igor_chybicki/

Department of Genetics, Kazimierz Wielki University, Chodkiewicza 30, Bydgoszcz, Poland

(Last update: 26/04/2024)

1. Introduction

The purpose of the INEST software is to provide unbiased multilocus estimates of inbreeding coefficients within a population, which are robust to a presence of null alleles. The methods described in Chybicki and Burczyk (2009) have been originally coded in the version 1.x, starting from 2007. At the beginning, a small number of routines allowed me writing the software in the single unit. However, after 6 years (in 2013), due to a large number of additional lines (majority of routines still in a single unit) and especially the obscure structure of some parts, I have found the code difficult to follow and to develop further. That's why I decided to implement the next version completely from scratch. The profit is twofold. First, I have split the code into parts related with different methods (statistical and others). Second, and probably much more important, I revised the routines responsible for the estimation. In effect, the second major release of the software differs slightly in estimation algorithms (compared with 1.x) and may return slightly different results. As users of the former version can easily note, as compared with the version 1.x, the present release has changed the interface. Because the Bayesian approach (IIM) has better statistical properties as compared with the maximum likelihood approach (PIM) (Chybicki and Burczyk 2009), I decided to organise the interface to make IIM the method of choice. However, PIM is still available using the menu 'Miscellanea'.

Version 2.0 was coded and maintained entirely under Delphi 7, Borland environment. However, because D7 (released in 2002) shows some compatibility problems with Windows 10, since version 2.1, INEST has been successfully ported to LAZARUS (www.lazarus-ide.org) and, later, to Turbo Delphi, which appeared to be a more stable programming environment. Because INEST 2.1 is compiled using a different compiler, there may be some changes in run-time of procedures.

Since version 2.2, INEST is maintained under Embarcadero Delphi 10.2, providing the support for high-definition screens and system re-scaling. Also, INEST 2.2 provides a new functionality in terms of empirical Bayes estimates of observed and expected heterozygosity corrected for null alleles (the functionality inspired by Javier Morente).

Since version 2.3, INEST maintenance is again under LAZARUS.

2. Data file

INEST reads only simple text files (no Excel, no Word, and all these types of rich file formats). Data have to be organised in the specific order, which will be best explained based on the following example:

```
5       3      1
MarkerName1
MarkerName2
MarkerName3
10      12.3   56.2   10     9      5      5      1      1
20      28.1   15.0   9      3      0      0      10     3
44      12.8   25.9   9      10     6      2      1      5
61      24.9   34.8   10     8      5      3      10     3
12      36.0   44.6   -1     -1     5      4      10     10
```

Data file begins with the header line, which has 3 integer numbers: *Ni* – the number of individuals, *Nl* – the number of loci, *Ca* – the binary indicator for spatial coordinates. *Ca* = 1 if coordinates are available and *Ca* = 0 otherwise. Starting from the second line (and up to *Nl*+1) a file must contain marker names. Then, a file must contain a table of data for individuals (*Ni* lines are expected). Each line starts from a number corresponding to the numerical indicator of an individual (Numerical ID). Then, if *Ca* = 1, X and Y coordinates are expected. They are followed by a genotype. Alleles must be separated by space or tab. The example without coordinates looks like this:

```
5       3      0
MarkerName1
MarkerName2
MarkerName3
10      10     9      5      5      1      1
20      9      3      0      0      10     3
44      9      10     6      2      1      5
61      10     8      5      3      10     3
12      -1     -1     5      4      10     10
```

Missing genotypes are represented by 0 0 or a pair negative integers. Unlike the version 1.x, in the version 2.x missing data are always utilised in the estimation. Therefore, there is no difference between 0 and -1, when coding alleles. Data are loaded through the standard open dialog launched by pressing the button […] located right to the ‘Data file name:’ field in the ‘Input/Output’ box. Once data are correctly parsed, some summary stats are written to the screen.

3. Estimation of inbreeding coefficients

The main purpose of the software is to estimate inbreeding coefficients. As previously, INEST 2.0 offers two approaches: PIM (or maximum likelihood) and IIM (or Bayesian). However, because IIM behaves apparently better than PIM (especially for small samples), now it is available readily through the main window.

3.1. Estimation based on IIM model:

IIM is implemented as a Bayesian approach. Therefore, it requires additional assumption concerning prior distributions for parameters of the model (the likelihood function). Briefly, inbreeding coefficients are assumed to follow beta distribution while allele frequencies at a given locus follow Dirichlet distribution.

In order to use IIM model, the number of cycles, the burnin and thinning must be chosen. Number of cycles refers to the number of MCMC iterations (or updates of the parameters) and generally the more the better. Default values are okay as for a trial analysis, but for final estimates longer chains must be used (200,000 cycles or more). Thinning (i.e. keep every n-th update) parameter is used to avoid strong autocorrelation between updates and to avoid extremely large output files. It is recommended to keep at least 1,000 to get a good approximation of posterior distributions. However, I would say that less than 10,000 stored updates would be a good option to avoid extremely large text files. Burnin is the number of disregarded updates, counting from the first one. It is used to estimate some properties of the model, including Deviance Information Criterion (DIC) used for model comparison. Because IIM uses the Gibbs sampler, 10% of the total number of updates is enough for burnin. Therefore, for 500,000 cycles in total, burnin can be set to 50,000 cycles.

The sampler is run pressing [Start] button. At this stage, after pressing [Start], a model must be specified. The model is defined in terms of the parameters involved. There are three types of parameters: n, f and b. n is null alleles, f is inbreeding coefficients and b is genotyping failures. Based on these three letters one is able to construct the model of choice. For example, nf means that the model includes null alleles and inbreeding (but ignores a possibility of genotyping failures). Thus, nfb refers to the full model, with all the parameters included. In order to define the null model, when all parameters equal zero, the special word ‘null’ must be used.

Once the model is chosen, the sampler starts running. The progress is shown at the bottom, in the status bar. During the analysis successive updates are saved to tab-delimited text files. Output files are located in the directory specified in the ‘Output file name’ field (in the ‘Input/Output’ group box). In order to obtain final estimates, some post-processing is needed, as described below.

After the specified number of cycles the sampler stops with no successive action. Then, quantiles of the marginal posterior distributions for all the parameters can be extracted using the inbuilt post-processing routine. Alternatively, any other software can be used (probably R package is the most convenient one, but even MS Excel can be used). There are six output files generated for a given model: *.bj, *.dic, *.fi, *.hyp, *.pjk and *.het (where * is the specified name of the output file). All these files can be loaded pressing […] button located right to the 'Posterior distrib.' field.

Contents of output files

File type (extension)	Content
***.bj**	Samples from the marginal posterior distribution of genotyping failure rates
***.dic**	Summary values to be used for model comparison, including effective number of parameters (pD) and Deviance Information Criterion (DIC)
***.fi**	Samples from the marginal posterior distribution of individual inbreeding coefficients
***.hyp**	Samples from the marginal posterior distribution of hyper-parameters, including (theoretical) inbreeding level (MeanF) and (empirical) average inbreeding coefficient (Avg(Fi))
***.pjk**	Samples from the marginal posterior distribution of allele frequencies, including null alleles (denoted by zero)
***.het**	Samples from the posterior distribution of observed (cHo) and expected (cHe) heterozygosity per locus corrected for null alleles

As for a prior distribution, a beta distribution was chosen for inbreeding. Consequently, there is no possibility to directly verify whether inbreeding is larger from zero, because formally F cannot be equal zero under the beta prior. However, still there is a possibility to perform alternative analysis based on a model without inbreeding (i.e. assuming that F = 0). Then, model comparison approach can be used to verify which model better fits to data. For this purpose INEST 2.x (unlike INEST 1.x) computes Deviance Information Criterion for each model. Generally, the model with the lowest DIC outperforms the others.

The example data analysed with the full model (nfb) gave the following results. After loading the file *.hyp, one can see the table:

```
X(i)        HSMode    Mean      Q(2.5%)   Q(5.0%)   Q(25.0%)  Q(50.0%)  Q(75.0%)  Q(95.0%)  Q(97.5%)  HPDl(95%) HPDh(95%)
LogL        -1516.0475          -1515.6548          -1531.3280          -1529.0300          -1520.5840          -1515.3850
  -1510.2550          -1503.7330          -1501.4280          -1529.9600          -1501.0970
af          0.3370    0.5173    0.1040    0.1380    0.2640    0.3810    0.6060    1.2810    1.5680    0.0410    1.2810
bf          2.1920    3.3580    0.9460    1.1010    1.7670    2.4620    3.7070    7.5550    10.4030   0.7000    7.5900
MeanF       0.1443    0.1377    0.0572    0.0723    0.1093    0.1362    0.1645    0.2141    0.2259    0.0626    0.2281
Avg(Fi)     0.1453    0.1358    0.0695    0.0817    0.1130    0.1366    0.1582    0.1934    0.2017    0.0682    0.1994
ab          0.0010    0.0567    0.0010    0.0020    0.0120    0.0300    0.0720    0.2090    0.2590    0.0000    0.2080
bb          0.0180    6.8773    0.0230    0.0550    0.3740    1.0130    2.8030    17.5830   39.5850   0.0030    17.1010
an          0.3460    0.5102    0.0870    0.1250    0.2860    0.4520    0.6860    1.0830    1.2000    0.0080    1.0740
```

Avg(Fi) refers to the sample mean inbreeding coefficient. The posterior mean is 0.1358, while 95% highest posterior density interval is from 0.0682 to 0.1994. We can see that there is strong support for a presence of inbreeding here. To perform Bayesian procedure of model comparison, we need to perform the analysis setting the model to 'nb'. Then, we can load *.dic files to see, which model has lower DIC value. For 'nfb' model, DIC is

```
Model:           nfb
Avg(logl(X)):    -1515.723
logl(Avg(X)):    -1473.742
DBar:            3031.446
Dhat:            2947.484
pD:              83.962
DIC:             3115.408
```

while for 'nb' model it is

```
Model:           nb
Avg(logl(X)):    -1534.916
logl(Avg(X)):    -1500.505
DBar:            3069.832
Dhat:            3001.009
pD:              68.823
DIC:             3138.654
```

So, DIC for nb is approximately 3139 while that for 'nfb' is 3115, supporting the inbreeding model as the better one. In other words, one can conclude that inbreeding is the significant component of the model. However, because DIC has no clear relationship with the probability of the model, the word "significant" has more qualitative (like "important") than quantitative (like "statistically significant at the level of…") meaning.

The list of parameters in the output files

Parameter	Output file	Description
LogL	*.hyp	Not a parameter but log-likelihood function of the model
af	*.hyp	The hyper-parameter (alpha) of the beta distribution used as a prior for inbreeding
bf	*.hyp	The hyper-parameter (beta) of the beta distribution used as a prior for inbreeding
MeanF	*.hyp	The average of the prior distribution; the theoretical inbreeding level of the population, from which the sample was taken
Avg(Fi)	*.hyp	The sample mean inbreeding coefficient
ab	*.hyp	The hyper-parameter (alpha) of the beta distribution used as a prior for genotyping failure rate
bb	*.hyp	The hyper-parameter (beta) of the beta distribution used as a prior for genotyping failure rate
MeanB	*.hyp	The average of the prior distribution; the theoretical genotypic failure rate
an	*.hyp	The only variant hyper-parameter of the Dirichlet distribution used as a prior for allele frequencies; it is used to assess the overall probability for null alleles across loci (the remaining alleles have uniform prior distribution)
b[j]	*.bj	The parameter of the likelihood function; the rate of random genotyping failure at the j-th locus
f[i]	*.fi	The parameter of the likelihood function; the individual

		inbreeding coefficient (of the i-th individual)
p[j,k]	*.pjk	The parameter of the likelihood function; the frequency of the k-th allele at the j-th locus; p[j,0] denotes null allele frequency at the j-th locus
Avg(logl(X))	*.dic	Average of log-likelihood function across iterations
logl(Avg(X))	*.dic	The log-likelihood function for the posterior means
DBar	*.dic	The mean deviance estimated across iterations
Dhat	*.dic	The deviance estimated for the posterior means
pD	*.dic	The effective number of parameters
DIC	*.dic	The deviance information criterion for the model
cHo[j]	*.het	The observed heterozygosity for the j-th locus corrected for null alleles
cHe[j]	*.het	The expected heterozygosity for the j-th locus corrected for null alleles

Generally, the post-processing procedure returns: posterior mode (estimated using the half-sample algorithm), posterior mean (Mean), quantiles of the posterior distribution (Q(…%)), and upper and lower limit of the highest density posterior interval (HPDl, HPDh).

3.2. Estimation based on PIM model:

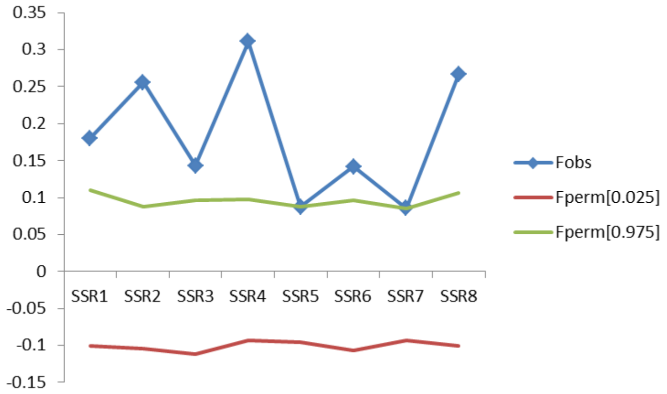
PIM is implemented as the expectation-maximization (EM) algorithm. In the current version, the algorithm is implemented in such a way that there is no need to set-up any parameter in order to perform the analysis. So, the EM procedure is launched through ‘Miscelenea|INEST (PIM)’ command in the main menu. There is only a need to agree (or not) on performing the jackknife procedure across loci. The jackknifing is used to estimate standard errors (SE) of estimates.

Please note that, once SEs are computed, significance of the parameters can be assessed using the Z-test based on the normal distribution theory. In order to test whether an estimate is different from zero, one needs to compute Z as Estimate_value / SE and compare it to the quantile of the normal N(0,1) distribution. For example, if one wishes to test if FIS is significantly larger from zero that Z = f_coef / SE and if Z > 1.96 then the decision is ‘yes, it is significantly larger’ and ‘no, FIS does not differ significantly from zero’ otherwise.

4. Additional functionality

4.1. Permutation test for heterozygosity excess

The permutation test can be used to test for the heterozygosity excess. The test is based on the conventional inbreeding coefficient, estimated as F = 1 – Ho/He, where Ho and He refers to the observed and expected heterozygosity, respectively. Assuming random allele pairing (panmixia), any deviation of F from zero is only due to statistical sampling. In order to verify whether the observed deviation is a signature of genetic sampling (e.g. inbreeding, outreeding) or null alleles, INEST reconstructs the empiric null distribution of F under panmixia based on Np permutations of alleles among genotypes. (By default Np is set to 100,000, but I would suggest 1,000,000 as for the final estimates.) The procedure produces the 95% confidence interval of the null distribution of F. If the observed F falls within the confidence interval, heterozygosity excess does not depart from zero (at the significance level 0.05).



The result of the permutation procedure obtained for the attached example data.

4.2. Spatial genetic structure

If coordinates are provided, INEST 2.x can be used to analyse spatial genetic structure using spatial autocorrelogram. Five measures of genetic similarity are available: Nason's F (Loiselle et al. 1995), Ritland's ρ (Rotland 1996), Moran's I (estimated as in Streiff et al. 1998; eqn. 4), Queller and Goodnight's r (Queller and Goodnight 1989) and semivariance (eqn. 3 in Wagner et al. 2005; in fact equal to the unweighted genetic distance introduced in Smouse and Peakall 1999). Nason and Ritland represent kinship (Malecot's coancestry) coefficient, while Moran and Q&G represent relatedness coefficient (see e.g. Hardy and Vekemans 1999 for definitions). The user can also choose whether intervals are to be determined based on the "equal number of pairs" or "equal width" criterion. The former will produce correlogram, in which every distance class has the same number of points (pairs of individuals). The latter will produce correlogram, in which every distance class will be equally long. In consequence, "equal number of pairs" and "equal width" will result in variable width and number of points, respectively.

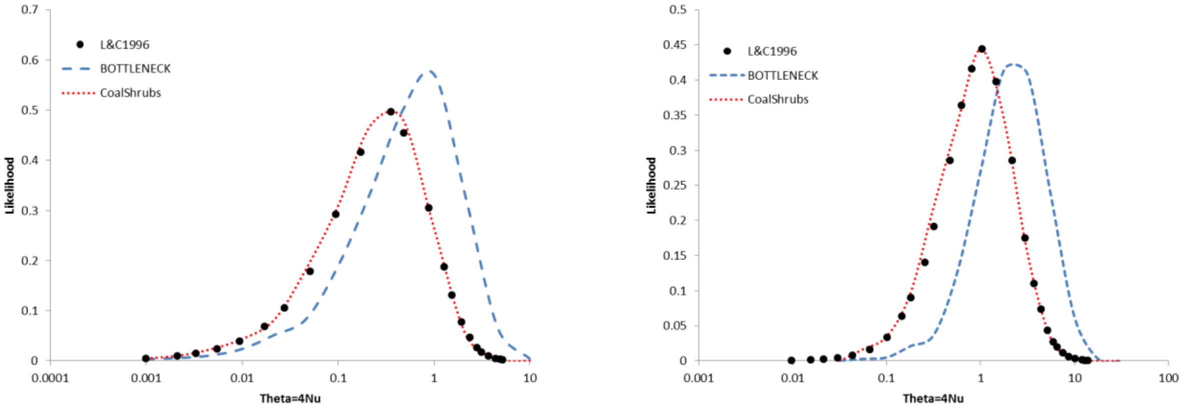
Significance of the correlogram is determined based on permutations of spatial positions of individuals. The number of permutations can be set in the main window (by default it is set to 999). After permutations, the null distribution is produced, of which some percentiles are shown. If the observed (Obs) value at a given distance class lays outside the bounds [0.025] and [0.975], the spatial autocorrelation (i.e. Obs) is statistically significant at the significance level of 0.05.

If a kinship or relatedness measure is chosen, INEST computes the slope of the log-linear regression function, which can be used to quantify the overall relationship between genetic similarity and distance. Under the isolation by distance (Rousset 2000), the slope should be negative, what means that relatedness decreases together with a distance between individuals. So, if the slope (Obs) of the correlogram lays outside the bounds [0.025] and [0.975], then there is the overall spatial genetic structure (of form expected for IBD).

The menu 'Export' in the window menu allows to extract some data, including 'Genetic similarity matrix', 'Distance matrix', 'Allele frequencies' used in the estimation and 'Distance to Nearest Neighbours'. The first two matrices can be copy/pasted to a file and used for other purposes (e.g. Mantel test, Principle Coordinate Analysis). Allele frequencies may be worth looking at, if one is interested in the frequencies computed without accounting for null alleles and missing data. In fact, these allele frequencies are used to compute genetic similarity measures in the SGS unit. The last option extracts distances to the nearest neighbour for every individual.

4.3. Bottleneck tests

This option was implemented in order to perform a test for demographic bottleneck. The main test is based on the phenomenon known as the excess of heterozygosity (or genetic diversity) in respect to a given number of alleles, typically generated in consequence of bottleneck as compared to a demographically stable population (i.e. a population of a constant size). The test was developed by Luikart and Cornuet (1996; Genetics). Although a widely-cited BOTTLENEK software (Piry et al. 1999) implements this test, I found BOTTLENECK to be probably buggy in respect to procedures for the Stepwise Mutation Model (SMM) and the Two-Phase Model (TPM). For example, I was unable to reproduce the distribution of conditional likelihood for a mutation rate (theta; Fig. 2 in Luikart and Cornuet 1996) as well as the original results for the case study in Luikart and Cornuet (1996) for the SMM model. In fact, the likelihood of the mutation rate (theta) estimated with BOTTLENECK, required to perform the test under SMM and TPM model of mutation, is shifted right (see figure below). As a result, BOTTLENECK overestimates the reference equilibrium heterozygosity. This effect is especially strong for low-to-moderate number of alleles at a locus, where little change in a mutation rate leads to visible differences in genetic diversity. In consequence, p-value of the test for heterozygosity excess (under SMM or TPM) is over-estimated, leading too often to false negatives. This finding led me to implement my own version of the algorithm, which performs correctly under both models.



The likelihood of theta for K=2 (left) and 3 (right) and 50 haploid chromosomes in a sample. L&C1996 reproduces the original plot in Luikart and Cornuet 1996, BOTTLENECK – the likelihood estimated with BOTTLENECK software, CoalShrubs – The likelihood estimated with my own implementation (used in INEST). To obtain L&C1996, I have digitalized the original figure and interpolated a series of discrete points along the distribution.

The interface is very simple. ‘Average multi-step mutation size’ equals to δ_g parameter of the two-sided geometric distribution used to model multi-step mutations (see Eq. 1 in Williamson-Natesan 2005, Conserv Genet). The equation for variance in multi-step mutation size, equal to $\sigma_g^2 = 2\delta_g - 3\delta_g + 2$, can be used to translate the parameter used in BOTTLENECK software (σ_g^2) into δ_g and vice versa. ‘Proportion of multi-step mutations’ corresponds to p_g parameter in the model (see Williamson-Natesan 2005). Note that BOTTLENECK uses ($1 - p_g$). The default settings are likely okay for most data, assuming they are microsatellite-based genotypes. Parameters of the Two-Phase Model were set according to recommendations by Peery et al. (2012; Mol Ecol). However, different values may be used to verify robustness of the bottleneck test in respect to assumptions of a mutation model. Alternatively, INEST implements a procedure where either p_g or d_g or both vary randomly during simulations. When p_g is set to a negative value then in each iteration a random value of p_g is drawn from a beta distribution with parameters $\mu=\alpha/(\alpha+\beta)= 0.292$ and $\gamma=1/(\alpha+\beta+1)= 0.207$ (where α and β are standard parameters of beta distribution). If d_g is set to a negative value then in each iteration d_g **=** $2 + X$, where X is drawn from a log-normal distribution with parameters $\mu=-0.246$ and $\sigma=0.770$. Note, that p_g and d_g can be set to vary independently. The values of parameters for a beta distribution and a log-normal distribution were determined based on data in Peery et al. (2012; Table 4). These values can be changed, however, because INEST requests to confirm them at the beginning of the analysis. Setting mutation parameters to vary during coalescent simulations is expected to increase a variance of equilibrium parameter values. So, it may have larger effect on the tests based on combined Z-values (Test 2 in Luikart & Cornuet 1996) than the Wilcoxon test (see below). It should be stressed that such an approach is not an original solution, and it was suggested earlier by Williamson-Natesan (2005) who, instead of using fixed p_g, proposed to allow p_g to vary during coalescent simulations (p_g was drawn from a uniform distribution in the range 0-0.2).

INEST implements two statistical tests: the Z-test based on combined Z scores for particular loci and the Wilcoxon signed-rank test. In the case of the Wilcoxon test p-values are determined both based on the assumption of normality (number of informative loci should be >20) and based on 1,000,000 permutations to approximate the exact value (normality not assumed here). However, as permutations may raise issues when a huge number of markers is used (e.g. thousands of SNPs), this option can be disabled using 'Options|Use permutation for p-value' (when unchecked).

In addition to equilibrium heterozygosity values, INEST returns also values of M-Ratio values (Garza and Williamson 2001). However, in order to perform properly the test for the deficiency in M-Ratio (treated as a signal of bottleneck), one needs to provide repeat motif lengths for microsatellite loci (using 'Options|Set motif lengths' menu). The idea behind the test for M-Ratio is analogous to that for the excess of heterozygosity (Williamson-Natesan 2005). However, due to specificity of a mutation process, the analysis is meaningful only under SMM and TPM models.

4.4. Convert FSTAT to INEST files

This tool can be used to prepare input files readable for INEST 2.x based on the FSTAT (Goudet 1995) input files. Each population in the FSTAT file will be written to the separate input file.

5. Closing remarks

This is an uncommercial software and may contain some bugs, although the author have done his best to check for errors in the routines related with the estimation. Therefore, the user is requested to inform about any erratic behaviour. Also, although no regular assistance can be assured, the user can count on the author's help, if needed. In this case he/she is requested to use the header "INEST" in their correspondence. E-mails will be replied as soon as possible. Good luck!

6. Bug fixes

10. April, 26, 2024. Symptom: computational issues when a huge number of markers is used for bottleneck testing. Fixed.

9. July, 8, 2020. Symptom: problems with output file post-processing due to repeated analysis under the same output file name. Fixed.

8. June, 8, 2017. Symptom: 'Invalid floating-point operation' error raised when 'PIM' estimator is used for data with no variation in genotypes at some loci. Fixed.

7. January, 30, 2015. Symptom: data on marker(s) with totally missing genotypes (across all

individuals) cause crush. Fixed.

6. January, 29, 2015. Symptom: expected heterozygosities differ slightly compared with other genetic software. In fact, INEST offered a very simple, yet approximately unbiased estimator developed with regard to uninbred populations. Now, a bit more sophisticated estimator is implemented (see E. 7 in Shete 2003; doi: 10.1093/jhered/esg078). Fixed.

5. April, 2, 2014. Symptom: If the system decimal separator is set to comma (French SI), the post-processing routine causes fatal error. The expected decimal separator for this routine is dot (English SI), while INEST output files are generated using the system setting. Fixed.

4. March 31, 2014. Symptom: If missing genotypes are present then SGS analysis with the Ritland kinship coefficient can cause fatal error. Fixed.

3. March 28, 2014. Symptom: If missing genotypes are present then 'all' keyword used to specify the model for the Bayesian analysis causes fatal error. Fixed.

2. March 12, 2014. Symptom: The main edit window does not scroll down automatically after loading an input file. (It may look like the main summary table does not update.) Fixed.

1. March 12, 2014. Symptom: Once an incorrect input file is tried to be loaded, the program does not allow to load any other file (correct or not) anymore, unless restarted. Fixed.