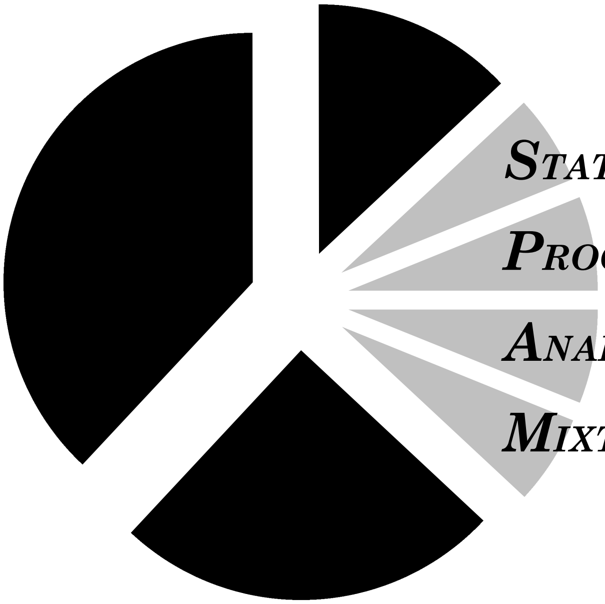
SPAM

Version 3.7

Addendum II to User's Guide for Version 3.2



STATISTICS
PROGRAM FOR
ANALYZING
MIXTURES

Special Publication No. 15

Alaska Department of Fish and Game
Division of Commercial Fisheries
Gene Conservation Laboratory
333 Raspberry Road
Anchorage, Alaska 99518

25 June 2003

(this page intentionally blank)

SPAM

(Statistics Program for Analyzing Mixtures)

Version 3.7

Addendum II to User's Guide for Version 3.2

Addendum II to Special Publication No. 15

Alaska Department of Fish and Game
Division of Commercial Fisheries
Gene Conservation Laboratory
333 Raspberry Road
Anchorage, Alaska 99518

25 June 2003

Citing the Software:

SPAM 3.2:

Debevec, E. M., R. B. Gates, M. Masuda, J. Pella, J. Reynolds, L. W. Seeb. 2000. SPAM (Version 3.2): Statistics Program for Analyzing Mixtures. Journal of Heredity 91 (6):509-510.

SPAM 3.5:

Alaska Department of Fish and Game. 2001. SPAM Version 3.5: Statistics Program for Analyzing Mixtures. Alaska Department of Fish and Game, Commercial Fisheries Division, Gene Conservation Lab. Available for download from http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm.

Citing the Manual:

SPAM 3.2:

Alaska Department of Fish and Game. 2000. SPAM Version 3.2: User's Guide. Special Publication 15, Alaska Department of Fish and Game, Commercial Fisheries Division, Gene Conservation Lab, 333 Raspberry Road, Anchorage, Alaska, 99518. 61 pages. Available for download from http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm.

SPAM 3.5:

Reynolds, J. H. 2001. SPAM Version 3.5: User's Guide Addendum. Addendum to Special Publication 15, Alaska Department of Fish and Game, Commercial Fisheries Division, Gene Conservation Lab, 333 Raspberry Road, Anchorage, Alaska, 99518. 63 pages. Available for download from http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm.

Table of Contents

(this page intentionally blank)

Overview of Enhancements

This addendum describes the new features of SPAM version 3.7. For a full description of the program, including input and output files, please see the SPAM version 3.2 User's Guide and the SPAM version 3.5 Addendum to the User's Guide, both available online[1]. The following descriptions assume a familiarity with material in the User's Guide and Addendum, hereafter referred to as UG:3.2.

The enhancements in SPAM 3.7 accommodate the estimation of baseline allele frequencies for loci with many low-frequency alleles. Under the conditional maximum likelihood (CML) scheme, a stock that has a sampled frequency of zero for an allele is assumed to be an impossible source for a mixture individual with that allele. This assumption may be unrealistic and cause bias and/or imprecision in stock-composition estimates. SPAM 3.7 offers a solution to this problem by allowing Bayesian modeling of baseline allele frequency distributions within the maximum likelihood scheme. That is, maximum likelihood is still used to estimate stock composition. SPAM 3.7 offers two Bayesian models of baseline allele frequency distributions: 1) Rannala-Mountain (Rannala and Mountain 1997) and 2) Pella-Masuda baseline posteriors (Pella and Masuda 2001). Both baseline posteriors are Dirichlet distributions. Rannala and Mountain (1997) use an equal-probability prior distribution (their Eq. 1) for the alleles at a locus with mean frequency equal to one over the number of distinct alleles. That is, all alleles at a locus are assumed to be equally abundant for all stocks before the baseline samples become available. The mean of the Rannala and Mountain baseline posterior is an unweighted average of the observed allele relative frequencies and the prior mean (Eq. 6 of Rannala and Mountain 1997). Pella and Masuda (2001) use a pseudo-Bayes method to determine the baseline posterior distribution for the alleles at a locus. The baseline center or unweighted arithmetic mean of the allele frequencies among stocks at a locus is used as the mean of the prior distribution. The mean of the Pella-Masuda baseline posterior is a weighted average of the observed allele relative frequencies and the baseline center with weights determined by an objective risk criterion (Eq. 4 of Pella and Masuda 2001). The user may still perform with SPAM 3.7 a traditional CML analysis, using the maximum likelihood estimates of baseline allele frequencies (i.e. no Bayesian modeling of baseline allele frequency distributions) if baseline sampling of zeros for alleles is not of concern.

Assuming the Rannala-Mountain (or Pella-Masuda) model of baseline allele frequency distributions, the user may perform estimation (with or without bootstrapping) or simulation (with baseline resampling). If the analysis involves no baseline resampling (for bootstrapping or simulation), then baseline allele frequencies are estimated by the mean of the Dirichlet posterior distributions. If the analysis involves baseline resampling

---

[1] http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm

(for bootstrapping or simulation), then allele frequencies are repeatedly drawn from the Dirichlet posterior distributions. If simulations are done, mixture genotypes are generated by sampling alleles with fixed relative frequencies equal to the posterior mean.

New Control File Options

For full documentation regarding required input files, including the general options available in the control file, see the SPAM version 3.2 User's Guide and the SPAM version 3.5 Addendum to the User's Guide.

* Estimation/Simulation

No new options.

* Options

See Appendix for example control file.

```
* options selected for optimization
  use IRLS algorithm in optimal search       : t
  print mixture file                         : t
  print baseline relative frequencies        : t
  print conditional genotype probabilities   : t
  print conditional population probabilities : t
  print bootstrap estimates                  : t
  print iterations                           : t
  print likelihoods of simulations/resamples : t
  compute likelihood confidence intervals    : t
  compute infinitesimal jackknife std. dev.  : t
  compute studentized conf. intervals        : t
  compute likelihood at external estimate    : t
  resample mixture frequencies               : t
  resample baseline                          : t
  use rannala-mountain model                 : t
  use pella-masuda model                     : f
```

This section is used to select performance and output options. The keywords for the various options are listed here.

Keyword(s)	*Default*	*Description*
`IRLS`	F	Use IRLS algorithm in optimal search
`PRIN`		Print…
`BASE`	F	baseline relative frequencies
`MIXT`	F	mixture file
`GENO`	F	conditional genotype probabilities
`POPU` or `STOC`	F	conditional population (stock) probabilities

	BOOT	F	estimates from each bootstrap resample
	ITER	F	MLE search iterations
	LIKE	F	likelihood for each simulation or resample
COMP			Compute…
	CONF	F	likelihood confidence intervals
	JACK	F	infinitesimal jackknife standard deviations
	STUD	F	studentized bootstrap conf. intervals
	EXTE	F	use externally provided mle in likelihood ratio
RESA			Resample…
	MIXT	F	mixture frequencies
	BASE	F	baseline
RANN		**F**	**Use Rannala-Mountain model of baseline allele frequency distributions**
PELL		**F**	**Use Pella-Masuda model of baseline allele frequency distributions**

RANNALA-MOUNTAIN

The keyword `RANNALA` is used for modeling baseline allele frequency distributions with the Rannala-Mountain baseline posterior (Rannala and Mountain 1997). If the analysis involves no baseline resampling, then baseline allele frequencies are estimated by the mean of the Dirichlet posterior distributions. If the analysis involves baseline resampling, then allele frequencies are repeatedly drawn from the Dirichlet posterior distributions.

PELLA-MASUDA

The keyword `PELLA` is used for modeling baseline allele frequency distributions with the Pella-Masuda baseline posterior (Pella and Masuda 2001). If the analysis involves no baseline resampling, then baseline allele frequencies are estimated by the mean of the Dirichlet posterior distributions. If the analysis involves baseline resampling, then allele frequencies are repeatedly drawn from the Dirichlet posterior distributions.

If both keywords `RANNALA` and `PELLA` are set to true, then the program will default to the Rannala-Mountain model.

* Parameters

See Appendix for example control file.

The control parameters specify the number of populations and characters in the analysis, upper limit parameters, tolerances to control the optimization search, and features to synchronize and partition simulated mixture samples across runs of SPAM 3.5 (for use in Monte Carlo simulation of likelihood ratios for testing mixture equality).

```
* control parameters
  number of populations in analysis  :  14
  number of characters in analysis   :   9
  maximum number of genotypes        : 200
  maximum number of classes          :  20
  maximum # of iterations            : 300
  maximum number of missing loci     :   4
  estimate tolerance                 :  .1E-3
  likelihood tolerance               :   1.0E-10
  genotype tolerance                 :   1.0E-6
  algorithm switch tolerance         :   0.01
  GPA                                :  90
  number of resamplings              : 100
  simulation sample size             : 100
  number of null observations after  :  0
  confidence intervals               :  90
  random seed (negative)             :  -718805
  second random seed (positive)      :  99733654
```

The keywords for the commands are listed here:

Keyword(s)		***Default***	***Description***
`NUMB`			Number of…
	`POPU` or `STOC`	-	populations (stocks) in the analysis
	`CHAR`	-	characters in the analysis
	`RESA`	100	bootstrap resamplings
	`BEFO`	0	observations to simulate, but not use, before simulating mixture sample of interest
	`AFTE`	0	observations to simulate, but not use, after simulating mixture sample of interest
`MAXI`			Maximum number of…
	`GENO`	100	genotypes
	`CLAS`	1	classes
	`ITER`	100	iterations
	`MISS`	0	missing (unscored) loci in mixture
`TOLE`			Tolerances for…
	`ESTI`	10^{-4}	estimates

`LIKE` or `FUNC`	10^{-10}	likelihood (function)
`GENO`	10^{-10}	genotype probability
`ALGO` or `SWIT`	10^{-2}	algorithm switch (CG to IRLS)
`GUAR`, `PERC`, or `GPA`	`90`	Guaranteed percent achievement of the maximal likelihood (GPA)
`CONF`	`90`	Confidence interval size (percent)
`SIZE`	`100`	Simulation sample size
`SEED`	*From CPU clock*	Random number generator seed
`SECO`	***From CPU clock***	**Random number generator seed**

SECOND

The keyword `SECOND` allows the user to specify a seed for a second random number generator for reproducible results. The seed takes a positive value between 0 and 2,147,483,648. If a negative seed is given, it will be changed to positive. Therefore, the seed –12,345 will give the same sequence of numbers as 12,345. The seed will be printed in the bootstrap and/or simulation output files.

If a random seed is not declared in the control file, SPAM will generate a seed based on the current CPU time. The seed will be printed in the bootstrap and/or simulation output files for reference.

* Characters

No changes. See UG:3.2 for full description.

* Populations

No changes. See UG:3.2 for full description.

* Regions

No changes. See UG:3.2 for full description.

* Files

No changes. See UG:3.2 for full description.

* Run

No changes. See UG:3.2 for full description.

Input Files

No changes. See UG:3.2 for full description.

Output Files

All results from a SPAM analysis are printed to a collection of ASCII text files that can be viewed through the SPAM environment or separately with any text editor. The set of files created depends on the analysis requested in the control file. All files, except the resampled estimate files, are formatted for convenient viewing and printing. Every SPAM analysis will produce a log file (`*.log`) and either an estimation (`*.est`) or a simulation file (`*.sim`), depending on the type of analysis run. Only changes in content or new files are discussed below. See UG:3.2 for a full description of the other output files created by SPAM 3.7.

Estimation (`*.est`)

See UG:3.2 for a full description.

The file will indicate if either the Rannala-Mountain or Pella-Masuda model of baseline allele frequency distributions is used.

If the Pella-Masuda model is selected, then values for the baseline posterior parameters are determined from the pseudo-Bayes method described in Pella and Masuda (2001). The posterior mean of allele or type relative frequencies is computed as a weighted average of the observed and prior mean relative frequencies. Weights for prior means are output to the `*.est` file for each character and stock. Weights for observed means are simply one minus weights for prior means. Large weights for prior means may indicate that there is little variation in the character among stocks. The character's value in discriminating stocks should be examined.

Simulation (`*.sim`)

See UG:3.2 for a full description.

The file will indicate if either the Rannala-Mountain or Pella-Masuda model of baseline allele frequency distributions is used in the resampling of the baseline. The value of the "`second random seed (positive)`" is reported here.

If the Pella-Masuda model is selected, then values for the baseline posterior parameters are determined from the pseudo-Bayes method described in Pella and Masuda (2001). The posterior mean of allele or type relative frequencies is computed as a weighted average of the observed and prior mean relative frequencies. The posterior mean is the underlying fixed allele or type relative frequencies of the simulated populations. Weights for prior means are output to the `*.sim` file for each character and stock. Weights for observed means are simply one minus weights for prior means. Large weights for prior means may indicate that there is little variation in the character among stocks. The character's value in discriminating stocks should be examined.

Bootstrap (*`.bot`)

See UG:3.2 for a full description.

The file will indicate if either the Rannala-Mountain or Pella-Masuda model of baseline allele frequency distributions is used in the resampling of the baseline. The value of the "`second random seed (positive)`" is reported here.

Baseline (*`.bsl`)

See UG:3.2 for a full description.

If either the Rannala-Mountain or Pella-Masuda model is selected, the means of the Dirichlet posterior distributions of allele frequencies are printed below the observed baseline allele frequencies.

Conditional Genotype Probabilities (*`.gen`)

See UG:3.2 for a full description.

If either the Rannala-Mountain or Pella-Masuda model is selected, conditional probabilities of types found in the mixture are computed from baseline allele frequencies estimated by the mean of the Dirichlet posterior distributions.

Correspondence

We welcome correspondence regarding SPAM. If you would like to be included on the mailing list and receive notifications of updates, please contact us at the address below. Please report any bugs as soon as possible so we can assess the problem and make any necessary corrections to the program.

Bill Templin – Bill_Templin@fishgame.state.ak.us
Lisa Seeb – Lisa_Seeb@fishgame.state.ak.us

Alaska Department of Fish and Game
Division of Commercial Fisheries
Gene Conservation Laboratory
333 Raspberry Road
Anchorage, Alaska 99518
USA

Limited Warranty and Disclaimer

This software and accompanying written materials (including instructions for use) are provided “as is” without warranty of any kind. Further, Alaska Department of Fish and Game (ADF&G) does not warrant, guarantee, or make any representations regarding the use, or the results of use, of the software or written materials in terms of correctness, accuracy, reliability, currentness, or otherwise. The entire risk as to the results and performance of the software is assumed by you. If the software or written materials are defective, you, and not ADF&G or its employees, assume the entire cost of all necessary servicing, repair, or correction.

The above is the only warranty of any kind, either express or implied, including but not limited to the implied warranty of fitness for a particular purpose, that is made by ADF&G. No oral or written information or advice given by ADF&G or its employees shall create a warranty or in any way increase the scope of this warranty and you may not rely on any such information or advice.

Neither ADF&G nor anyone else who has been involved in the creation, production or delivery of this product shall be liable for any direct, indirect, consequential or incidental damages (including damages for loss of business profits, business interruption, loss of business information, and the like) arising out of the use or inability to use such product even if ADF&G has been advised of the possibility of such damages.

Use of this product for any period of time constitutes your acceptance of this agreement and subjects you to its contents.

Literature Cited

Pella, J. and Masuda, M. 2001. Bayesian methods for analysis of stock mixtures from genetic characters. Fishery Bulletin. 99:151-167.

Rannala, B. and Mountain, J. L. 1997. Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences of the United States of America. 94:9197-9201.

Appendix

Example Control File for Modeling Baseline Allele Frequency Distributions.

```
* estimation: Columbia Example

* options selected for optimization
  use IRLS algorithm in optimal search       :  f
  print mixture file                         :  t
  print baseline relative frequencies        :  t
  print conditional genotype probabilities   :  t
  print conditional population probabilities :  t
  print bootstrap estimates                  :  t
  print iterations                           :  t
  compute likelihood confidence intervals    :  t
  compute infinitesimal jackknife std. dev.  :  t
  compute studentized bootstrap intervals    :  t
  resample mixture frequencies               :  f
  resample baseline                          :  f
  use rannala-mountain model                 :  t
  use pella-masuda model                     :  f

* control parameters
  number of populations in analysis  :  14
  number of characters in analysis   :   9
  maximum number of genotypes        : 200
  maximum number of classes          :  20
  maximum # of iterations            : 300
  maximum number of missing loci     :  4
  estimate tolerance                 : .1E-3
  likelihood tolerance               :  1.0e-10
  genotype tolerance                 :  1.0e-6
  algorithm switch tolerance         :  0.01
  GPA                                :  90
  number of resamplings              :  25
  confidence intervals               :  90
  random seed                        :  -718805
  second random seed (positive)      : 85728283
...
```