ONCOR

A computer program for genetic stock identification

Steven T Kalinowski

Kezia R Manlove

Mark L Taper

Department of Ecology

310 Lewis Hall

Montana State University

Acknowledgements

Funding for the development of this program has been provided by the Northwest Fisheries Science Center and the United States Chinook Technical Committee.

Table of Contents

1. Introduction

2. Data files

    2a. Baseline files

    2b. Reporting group files

    2c. Mixture files

    2d. Fishery files

    2e. Sample size files

3. Mixture analysis menu

    3a. Estimation

    3b. Simulate a single mixture sample.

    3c. 100% Simulations

    3d. Realistic fishery simulations

    3e. Three-way error decomposition

4. Individual assignment menu

    4a. Estimation

    4b. Leave-one-out test

    4b. Realistic fishery simulation

1. Introduction

ONCOR is a computer program that uses genetic data to estimate the population of origin of individuals. This program was specifically written to estimate the composition of mixed stock fisheries of Pacific salmon (genus *Oncorhynchus*), and this manual is written in that context, but the program can be used for many other applications. This manual assumes that the user is familiar with mixture analysis and assignment tests. Its main purpose is to describe which statistical tests the program implements and what the user needs to do in order to perform the calculations.

ONCOR performs data analysis and simulations for mixture analysis and assignment tests. In brief, a mixture analysis estimates the proportion of fish in a fishery that belong to different stocks that may have contributed to the fishery. Assignment tests estimate the origin of each individual fish. In addition to estimating mixture proportions and conducting assignment tests, ONCOR performs several types of simulations that evaluate how accurate mixture analysis or assignment testing is likely to be.

2. Input files

ONCOR uses five types of data files—baseline, reporting groups, mixture, fishery, and sample size. A baseline file is needed for all types of analysis, the others are used for specific applications.

2a.Baseline files. Baseline files contain genotypes from populations that may have contributed fish to a fishery. ONCOR reads baseline data in standard GENEPOP format or a slight modification of GENEPOP format. If the user is not familiar with the GENEPOP format, please see the documentation at the GENEPOP website (http://genepop.curtin.edu.au/). ONCOR reads genotypes that are either four or six digits long. An example of a standard GENEPOP file is shown below. Note that the first line contains the title for the data set, the second line lists the names of loci included in the data, and that missing data is indicated by zeros.

```
GAPS 2.1 - Stikine & Taku populations
Ogo2, Ogo4, Oki100, Omm1080, Ots201b, Ots211, Ots212, Ots213, Ots3M, Ots9
POP
AndrewCr, 220226 132158 224236 282282 214226 170202 240272 139195 278298 148154 105107
AndrewCr, 220222 136158 244349 242258 153173 182222 232268 139139 294306 144148 105107
AndrewCr, 216222 000000 236305 194246 190214 186190 228232 139171 262334 144144 105105
AndrewCr, 200222 136162 236256 166290 178214 190218 240256 143147 278310 144150 105107
AndrewCr, 220222 132160 248298 166330 190238 186206 256280 143171 282290 144144 105105
POP
KowatuaCr, 216222 136152 208244 282298 173214 190242 208284 143175 266298 144146 103107
KowatuaCr, 220222 160162 244264 294298 186242 190230 228232 143155 274290 144146 105105
KowatuaCr, 222226 160160 232244 282282 202230 174294 236292 143179 294322 142144 105107
KowatuaCr, 216216 138160 264329 242290 182214 162170 244300 147175 290310 144146 105107
KowatuaCr, 220222 132136 236248 270278 169262 174186 240244 139143 258290 146148 105107
KowatuaCr, 222234 136136 244244 222294 190218 162166 228312 135155 282310 138146 103105
KowatuaCr, 222222 132136 260275 218250 186258 174226 232236 143163 254258 146146 107107
```

The GENEPOP format does not include a convenient place for listing sample names, so ONCOR supports a slight modification of standard GENEPOP format. ONCOR will look for populations names immediately following (and on the same line as) the “POP” indicator in a

GENEPOP file. ONCOR will read GENEPOP files that have spaces in the population names, but fishery and sample size files cannot have spaces in population names, so the user may want to remove spaces from all population names.

2b. Reporting group files. Reporting groups are collections of populations, usually in the same geographic region, that are managed together (e.g., populations of Chinook salmon in the Puget Sound comprise a reporting group). Users of ONCOR do not need to define reporting groups, but such grouping can be useful when baseline populations are genetically similar. Identifying the exact population of origin of a fish is difficult when baseline populations are genetically similar, but if reporting groups are composed of populations that are genetically related, estimating the origin of fish to reporting group should be more accurate.

An example of a reporting group file is shown below.

```
GAPS 2.1 Reporting groups: Stinkine/Taku
AndrewCr       Stikine
KowatuaCr      Taku
LTahltanR      Stikine
NakinaR        Taku
TatsatuaCr     Taku
UNahlinR       Taku
```

In this example, there are six baseline populations (AndrewCr, KowatuaCr, LTahltanR, NakinaR, TatsatuaCr, UNahlinR) that belong to one of two reporting groups (Stikine, Taku). Reporting group files must the following requirements:

1. The first line of the file must contain a title describing the file.
2. Each subsequent line of the file lists a baseline population and the reporting group that it belongs to.
3. There can be no spaces in the names of baseline populations.
4. Baseline populations do not have to be listed in the same order as in the baseline file.
5. Baseline population names must be spelled the same as in the baseline file.
6. All baseline populations must be assigned to a reporting group.

2c. Mixture files. A mixture file contains genotypes of fish sampled from a fishery. Mixture files must be in GENEPOP format, and must contain only one GENEPOP "POP" designation. The mixture file must have the same loci as the baseline file, and the loci must be listed in the same

order. Mixture files are required for any sort of estimation (i.e. mixture analysis or assignment tests), but are not needed for simulations.

2d. Fishery files. A fishery file is used during computer simulations that estimate how accurate GSI is likely to be. ONCOR reads two types of fishery files: proportion format and count format. Both describe the stock proportions in the fishery to be sampled. In a proportion format file, the parametric proportions of each stock in the fishery are listed in the fishery file. If this type of file is used, ONCOR will simulate a mixture sample by drawing fish with replacement from the proportions provided. For example, in the file show below, there is a 90% chance that a fish in the mixture sample will be from AndrewCr. In a count format fishery file, the exact number of fish in the mixture sample from each stock is listed. In the example file show below, each simulated sample will include exactly 90 fish from AndrewCr.

Proportion format:

```
Test mixture for Stikine/Taku populations: 90% AndrewCr
AndrewCr       0.90
KowatuaCr      0.02
LTahltanR      0.02
NakinaR        0.02
TatsatuaCr     0.02
UNahlinR       0.02
```

Count format:

```
Test mixture for Stikine/Taku populations: 90% AndrewCr
AndrewCr       90
KowatuaCr      2
LTahltanR      2
NakinaR        2
TatsatuaCr     2
UNahlinR       2
```

2e. Sample size files. Sample size files are used in simulations to specify how large the simulated baseline should be. This is useful if the user wants to examine how increasing baseline sample sizes will affect the accuracy of genetic stock identification. The format of sample size files is similar to reporting group and fishery files. The first line of the file contains a title. After that, each line lists a baseline population and the sample size (in diploid individuals) to use for it. An example of a sample size file is show below.

```
Stikine 500 / Taku same
AndrewCr       500
KowatuaCr      SAME
LTahltanR      500
NakinaR        Same
TatsatuaCr     Same
UNahlinR       Same
```

In this example, simulated baseline data sets will have 500 *individuals* for the AndrewCr and LTahltanR populations, and use the same exact sample sizes for the other populations as in the actual baseline. If the "Same" identifier is specified, ONCOR will use the same number of genotypes per population *per locus* as in the baseline that is being analyzed. Missing data is accounted for when keeping track of sample sizes, so the sample size for a population may be different for each locus.

3. MIXTURE ANALYSIS

3a. Estimation. A mixture analysis uses baseline genetic data to estimate the stock composition of a sample from a mixed stock fishery.

Methods. Follow the following procedure to estimate mixture proportions.

1. Open a baseline file.
2. Open a reporting groups file (optional).
3. Select [Estimate Mixture Proportions] from the [Mixture Analysis] menu, and indicate whether you want to perform bootstrapping to estimate 95% confidence intervals (this will be slower, especially if the baseline is large).

ONCOR uses conditional maximum likelihood to estimate mixture proportions (Millar 1987). The EM algorithm is used to estimate mixture proportions, and iteration is continued until the total change of mixture proportions from one iteration to the next (summed across all stocks) is less than 10^{-6}. Genotype probabilities are calculated using the method of Rannala and Mountain (1997). If the bootstrapping option is selected, both baselines and mixtures are bootstrapped. Mixture samples are bootstrapped by resampling individuals with replacement from the individuals in the mixture file. Baseline genotypes are bootstrapped by resampling alleles from

baseline samples using the method of Rannala and Mountain (Equations 23-25 of Rannala & Mountain, 1997).

Results. An example of the output for this option is show below.

```
ONCOR Output
10/20/2007 8:52:20 AM

Baseline ID      GAPS 2.1 - Stikine & Taku populations
Baseline file    StikineTaku BASELINE.txt
Mixture ID       Simulated mixture data from StikineTaku BASELINE.txt
Mixture file     StikinTaku - MIXTURE - Simulated example.txt
NMixture         100
NBootstraps      1000

POPULATION ESTIMATES w/ 95% CI
AndrewCr       0.8752  (0.677, 0.909)
KowatuaCr      0.0128  (0.000, 0.097)
LTahltanR      0.0572  (0.000, 0.178)
NakinaR        0.0041  (0.000, 0.095)
TatsatuaCr     0.0194  (0.000, 0.092)
UNahlinR       0.0314  (0.000, 0.124)
```

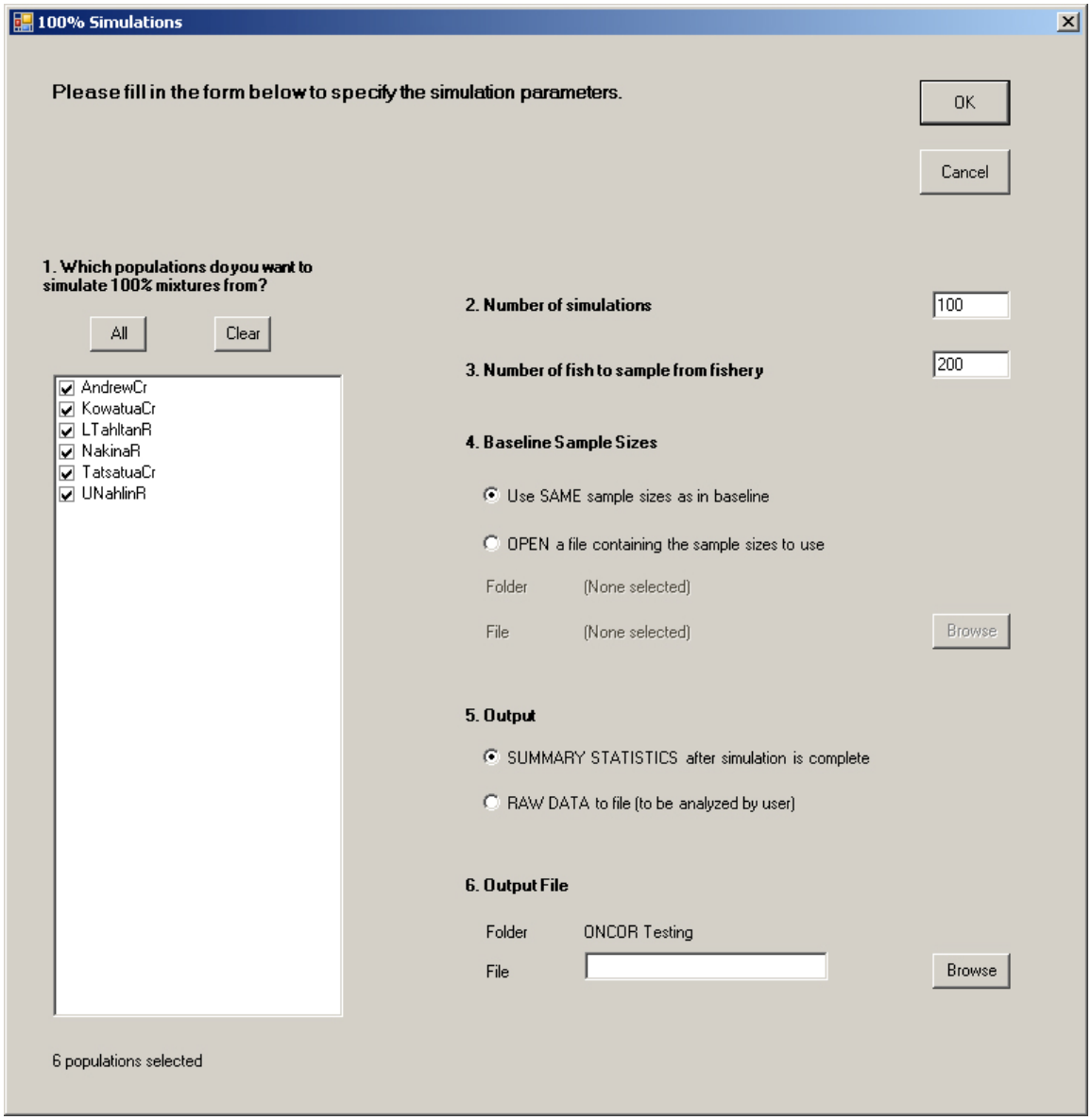
Discussion. The conditional maximum likelihood method used by ONCOR to estimate mixture proportions is the traditional method for estimating mixture proportions from genetic data. These estimates are usually biased towards $1/k$ (where k is the number of populations in the baseline). The bias is strongest when populations are genetically similar (i.e. low F_{ST} for baseline populations). The bootstrapping method used by ONCOR to estimate confidence intervals is traditional, but the method has not been tested to verify that its intervals contain parametric mixture proportions 95% of the time. The recent work of Anderson et al. (2007) suggests that this sort of bootstrapping probably introduces some sort of bias, but the affect of this on confidence intervals has not yet been explored.

3b. Simulate a single mixture. On occasion, it is useful to simulate a sample of individuals from a mixed stock fishery. ONCOR has a function for doing this in the [Mixture Analysis] menu. If this function is used, the user will be asked to open a fishery file, which will be used to determine which baseline populations the fish in fishery came from. Genotypes for the fishery sample are simulated by drawing alleles from reduced variance estimates of allele frequencies in the baseline

populations (Kalinowski et al. 2007). The genotypes simulated by this routine are output to a text file in GENEPOP format.

3c. 100% Simulations. ONCOR performs a few different types of simulation to examine how accurate mixture analysis is likely to be. The 100% simulation is one of the standard tools used by the Pacific salmon community to evaluate the accuracy of GSI. In this type of simulation, a fishery sample is simulated in which all of the individuals are from the same population.

Methods. After opening a baseline file (and a reporting groups file, if wanted), select [100% Simulations] from the [Mixture analysis] menu. The following dialog box will appear.



Each of the six parts of this form must be completed before a 100% simulation can be run.

1. Which populations do you want to simulate 100% mixtures from? This set of checkboxes let the user select a subset of populations to simulate data from.

2. Number of simulations. This is the number of mixture samples to generate for each stock. If, for example, there are six baseline populations selected for analysis, and 100 mixtures are simulated for each (as shown on the example form above), 600 mixture samples will be generated and analyzed.

3. Number of fish to sample from the fishery. This is mixture sample size.

4. Baseline sample sizes. The user has the option of using the existing baseline in these simulations or of providing alternative baseline sample sizes to explore how it might affect the accuracy of the estimates. If the user wants to use different sample sizes than the actual baseline, a sample size file must be opened. To do this, click on the "OPEN a file containing sample sizes to use" radio button and then click on the "Browse" button to select a file.

5. Output. There are two choices for output for this simulation. The default value is "SUMMARY STATISTICS". If the user chooses this option, averages and standard deviations will be output to ONCOR's output window after the calculations have been performed. Alternatively, all mixture estimates can be written to file and the user can analyze them with a statistics package. This option may be useful when using large baselines. Mixture analysis is slower when baselines are large, and it can be inconvenient to have to wait for the entire program to run to obtain results.

6. *Output file*. This specifies the file to output results to. The default location for this file is the same folder as the baseline file. Press the <Browse> button to select a different folder.

Methods. ONCOR uses one of two new methods for running these simulations. If the baseline sample size is set to "Use SAME sample sizes as baseline," ONCOR uses the method of Anderson et al. (2007) to simulate mixture genotypes and to estimate their probability of occurrence in baseline populations. If the user selects "OPEN a file containing sample sizes to use," ONCOR assumes that the parametric allele frequencies in baseline populations are equal to the reduced-variance estimates of Kalinowski et al. (2007), and then simulates baselines and mixture samples from these allele frequencies.

Results. An example of the output from this program is show below.

```
ONCOR Output – 100% Simulations
10/16/2007 9:57:17 AM

Baseline ID     GAPS 2.1 - Stikine & Taku populations
Baseline file   StikineTaku BASELINE.txt
NMixture        200
NSimulations    1000
Sample sizes    Same as empirical baseline


                   POPULATION ESTIMATES

               AVG    ST DEV  (95 PERCENT INT)
-------------------------------------------
AndrewCr       0.9082  0.0261  (0.8558, 0.9582)
KowatuaCr      0.4936  0.0512  (0.3940, 0.5940)
LTahltanR      0.5664  0.0502  (0.4650, 0.6666)
NakinaR        0.4272  0.0499  (0.3208, 0.5311)
TatsatuaCr     0.5584  0.0496  (0.4587, 0.6517)
UNahlinR       0.5968  0.0506  (0.4994, 0.6921)
-------------------------------------------

Run duration: 2 minutes, 45 seconds
```

In the output above, the average estimate for AndrewCr was 0.9082 (when the correct value was 1.0). The standard deviation across estimates was 0.0261, and 95% of the estimates fell in the interval (0.8558, 0.9582).

The data above shows the output if the "SUMMARY STATITSICS" option is chosen. If the user selects the "RAW DATA" option, all the estimates for all of the stocks will be output to a tab delimited text file. An example of this output is shown below.

Run_ID	100_Percent_Stock	Stock_Estimate	Estimate
10/20/2007	AndrewCr	AndrewCr	0.88508061
10/20/2007	AndrewCr	KowatuaCr	0.01653885
10/20/2007	AndrewCr	LTahltanR	0.07251415
10/20/2007	AndrewCr	NakinaR	0.02586348
10/20/2007	AndrewCr	TatsatuaCr	0.00000000
10/20/2007	AndrewCr	UNahlinR	0.00000292
10/20/2007	KowatuaCr	AndrewCr	0.01377819
10/20/2007	KowatuaCr	KowatuaCr	0.52343579
10/20/2007	KowatuaCr	LTahltanR	0.01428265
10/20/2007	KowatuaCr	NakinaR	0.11739555
10/20/2007	KowatuaCr	TatsatuaCr	0.28614100
10/20/2007	KowatuaCr	UNahlinR	0.04496681

In the example output above, the first six lines of output (i.e. lines 2-7) show mixture estimates for a sample in which fish from AndrewCr comprised 100% of the fishery

sample. The next six lines show estimates when fish from KowatuaCr comprise 100% of the fishery, etc. In each case, the column "100_Percent_Stock" lists the stock that has a frequency of 1.0 in the fishery. The "Stock_Estimate" column lists the stock for which the next column shows the estimate for. The data is output in tab delimited format so that it can easily be imported to a spreadsheet, database, or statistics program. If the data is copied and pasted into a spreadsheet, a pivot table can easily be used to analyze the data.

Discussion. The method of Anderson et al. (2007) is currently the preferred method for examining the likely accuracy of GSI. It is more accurate than previous methods that resampled alleles with replacement from the allele frequencies in baseline samples. One disadvantage of this method is that it cannot simulate baselines that are larger than the available baseline. The method Kalinowski et al. (2007) does not have this limitation, but this new method has not been tested extensively, and therefore, should be used with some caution. Testing to date shows that this method produces slightly optimistic results (i.e. smaller errors than it should), but is substantially more accurate than bootstrapping from allele frequencies in baseline samples.

3d. Realistic fishery simulations. The 100% simulation described above serves as a convenient benchmark for assessing the accuracy of GSI, but has the disadvantage of being unrealistic. Real fisheries are likely to have complex mixtures of baseline stocks, and it is natural to wonder how accurate GSI is likely to be when fisheries realistic proportions. ONCOR provides a function to perform such simulations. The program is very similar to the 100% simulations described above, with the single exception that the user must open a fishery file that specifies the parametric stock proportions in the fishery.

Methods. The method used to simulate data for this analysis depends on whether the user selects "Use SAME sample size as baseline" or "OPEN a file containing sample sizes to use." If baseline sample sizes are the same as the actual baseline, ONCOR uses the resampling method of Anderson et al. (2007) to simulate mixture genotypes and to estimate their probability of occurring in baseline populations. If the user selects "OPEN a file containing sample sizes to use," ONCOR assumes that the parametric allele frequencies in baseline populations are equal to the reduced-variance estimates of

Kalinowski et al. (2007), and then simulates baselines and mixture samples from these allele frequencies.

Results. Here is an example of the output from a realistic fishery simulation.

```
ONCOR Output – Realistic fishery simulations
10/16/2007 9:33:58 AM

Baseline ID     GAPS 2.1 - Stikine & Taku populations
Baseline file   StikineTaku BASELINE.txt
NMixture        200
NSimulations    100
Sample sizes    Same as empirical baseline
```

		ESTIMATES		
	ACTUAL VALUE	AVG	ST DEV	(95 PERCENT INT)
AndrewCr	0.9000	0.8050	0.0330	(0.7395, 0.8699)
KowatuaCr	0.0200	0.0258	0.0177	(0.0000, 0.0624)
LTahltanR	0.0200	0.0718	0.0274	(0.0257, 0.1311)
NakinaR	0.0200	0.0383	0.0233	(0.0000, 0.0860)
TatsatuaCr	0.0200	0.0240	0.0161	(0.0000, 0.0571)
UNahlinR	0.0200	0.0350	0.0206	(0.0000, 0.0781)

Each line in the table above shows the actual proportion of a stock in the simulated fishery and a few summary statistics for the estimates of that proportion. For example, 90% of the fish in the fishery were from AndrewCr, but the average estimate was 0.8050 (which shows that the estimates for this stock are biased low). The standard deviation of estimates for AndrewCr was 0.0330 and 95% of the estimates fell in the interval (0.7395, 0.8699). Note that the average estimate for all of the stocks is biased towards $1/k$ (where k is the number of stocks in the baseline, which is six in this example).

If the user selects the "RAW DATA" output option, the output will look like this:

STOCK	PARAMETRIC	ESTIMATE
AndrewCr	0.85	0.83348815
KowatuaCr	0.01	0.00000000
LTahltanR	0.02	0.00656216
NakinaR	0.03	0.07702571
TatsatuaCr	0.04	0.05498693
UNahlinR	0.05	0.02793705

```
AndrewCr        0.85            0.67118058
KowatuaCr       0.01            0.00000024
LTahltanR       0.02            0.05537179
NakinaR         0.03            0.08855978
TatsatuaCr      0.04            0.02936097
UNahlinR        0.05            0.15552664
AndrewCr        0.85            0.80853497
KowatuaCr       0.01            0.04654310
```

Each row lists a stock, the parametric frequency of that stock in the mixture and the estimated proportion. The first six lines of data show results for the first simulated mixture; the second six lines show results for the second simulated mixture etc. The data is in tab delimited format so that it can easily be imported to spreadsheet, database, or statistics program.

Discussion. The method of Anderson et al. (2007) is currently the preferred method for performing GSI simulations. It is more accurate than previous methods that resampled alleles with replacement from the allele frequencies in baseline samples. One disadvantage of this method is that it cannot bootstrap baselines that are larger than the available baseline. The method Kalinowski et al. (2007) does not have this limitation, but has not been tested extensively, and therefore, should be used with some caution. Testing to date shows that this method produces slightly optimistic results (i.e. less error than it should), but is substantially more accurate than bootstrapping from allele frequencies in baseline samples.

When running "realistic" fishery simulations, the parametric stock proportions should not be set to $1/k$ (where k is the number of stocks in the baseline, which is six in this example). Mixture analysis is biased towards $1/k$, so if $1/k$ is used for parametric stock proportions, the estimates will be closer to this true value than if some other (probably more realistic) fishery proportions were used.

3e. Three-way error decomposition. Several sources of error cause mixture estimates to differ from the parametric stock proportions in a mixed stock fishery. ONCOR performs a simulation based analysis that estimates the relative magnitude of three types of error—fishery sampling error, genotypic sampling error, and baseline sampling error. Fishery sampling error is error introduced by sampling a finite number of fish from fishery. If, for example, only ten fish are sampled, mixture estimates will be inaccurate no matter how many loci are genotyped. Genotypic sampling error is the estimation error that is caused by genotyping a finite number of loci. If for

example, only one locus is genotyped, mixture estimates are likely to be inaccurate no matter how large baseline sample sizes are. Genotypic error is estimates by assuming that the allele frequencies in baseline populations are known without error. Baseline sample sampling error is the error caused by not knowing exactly the allele frequencies in baseline populations.

Methods. ONCOR uses the method of Kalinowski et al. (2007) to perform the error decomposition. The program outputs the percentage of the total error that is attributable to each type of error. It is important to realize that these numbers are percentages. If the total error is small, each source of error may be small. ONCOR performs an error decomposition for each stock and an overall summary of the results. The summary results are obtained by summing mean squared errors over all stocks and then determining which proportion of the total MSE is from fishery, genotypic, and baseline sampling error.

Results. An example of the output for an error decomposition is shown below.

```
ONCOR Output - Error Decomposition
10/20/2007 9:56:01 AM

Baseline ID     GAPS 2.1 - Stikine & Taku populations
Baseline file  StikineTaku BASELINE.txt
NMixture        100
NSimulations    100
Sample sizes    Same as empirical baseline

SUMMARY RESULTS FOR ALL POPULATIONS
Fishery        12.4%
Genotypic      10.0%
Baseline       77.6%

RESULTS BY POPULATION
               FISH    GENO    BASE
AndrewCr       10.0%   2.3%    87.7%
KowatuaCr      29.6%   47.0%   23.4%
LTahltanR      9.1%    14.7%   76.2%
NakinaR        11.5%   13.5%   75.0%
TatsatuaCr     35.4%   41.0%   23.6%
UNahlinR       25.2%   28.4%   46.4%

Run time: 3 seconds
```

Note that baseline sampling error is the largest source of error for most of the stocks, but for a couple stocks, genotypic sampling error is substantial.

Discussion. The following points may assist interpretation of the error decomposition.

i. If the fishery error is large compared to the genotypic and baseline errors, this means that the genetic data in the baseline is accurately estimating the mixture proportions in the mixture samples. From a geneticist's point of view, this is a desirable result.

ii. If the genotypic error is substantial, more loci will have to be genotyped to improve the accuracy of GSI.

iii. If the baseline error is large (as in the example below), increasing baseline sample sizes should increase the accuracy of mixture estimates. However, increasing the number of loci genotyped would also be an option. That last point is worth repeating. Increasing baseline sample sizes is not the only way to reduce baseline sampling error. The results produced by ONCOR are for a given set of loci. If additional loci are genotyped, GSI is likely to be more accurate and the error caused by baseline sampling is likely to decrease.

4. ASSIGNMENT TESTS

Assignment tests estimate the origin of each fish in a sample from a fishery. ONCOR performs assignment tests and performs two types of simulation to assess the accuracy of assignment. All of these functions are available in the [Individual Assignment] menu.

4a. Individual assignment

Methods. Assignment tests can be performed by selecting the [Assign individuals to baseline populations] option.

ONCOR assigns individuals in a mixture sample to the baseline population that would have the highest probability of producing the given genotype *in the mixture*. Emphasis is placed on the phrase "in the mixture" because ONCOR uses both genotype frequencies and mixture proportions when estimating the origin of individuals. ONCOR performs these calculations as

follows. Let P_{ij} represent the probability that individual i (of unknown origin) belongs to population j in the baseline. P_{ij} can be estimated from estimates of the genotype frequencies in each baseline population and an estimate of the stock composition of the fishery. Let f_{ij} represent frequency of the i^{th} fish's genotype in the j^{th} population of the baseline. ONCOR uses the method of Rannala and Mountain (1997) to estimate this probability. Let m_j represent the estimated stock composition of the sample. ONCOR uses the methods described above for mixture estimation to estimate m_j. Given these estimates, the probability that the i^{th} individual belongs to the j^{th} baseline population, P_{ij}, is equal to

$$P_{ij} = \frac{m_j f_{ij}}{\sum_j m_j f_{ij}}$$

Results. ONCOR outputs the results for assignment tests in three ways. Output to the screen lists each individual and the population that that would most likely have that produced that individual's genotype in the mixture sample and the probability of that genotype coming from the population indicated. A sample of screen output is shown below.

```
ONCOR Output
10/20/2007 10:50:57 AM

Baseline ID     GAPS 2.1 - Stikine & Taku populations
Baseline file   StikineTaku BASELINE.txt
Mixture ID      Simulated data from StikineTaku BASELINE.txt
Mixture file    Simulated example.txt
N Mixture       100


RESULTS BY POPULATION

IND             ORIGIN          PROBABILITY
AndrewCr-1      AndrewCr        0.5118
AndrewCr-2      AndrewCr        0.9995
AndrewCr-4      AndrewCr        0.9986
AndrewCr-5      AndrewCr        0.9955
AndrewCr-6      AndrewCr        1.0000
AndrewCr-7      AndrewCr        0.9839
AndrewCr-8      AndrewCr        1.0000
AndrewCr-9      AndrewCr        0.9998
AndrewCr-10     AndrewCr        0.9859
AndrewCr-11     AndrewCr        0.9995
AndrewCr-12     AndrewCr        0.9454
```

The output file contains the same data, but also lists second, third, and fourth (as necessary) most likely populations and their probabilities. An example of this type of output is show below.

IND	Best Estimate	Probability	2nd Best Estimate	Probability	3rd Best Estimate, etc.	
AndrCr-1	AndCr	0.5118	LTahltanR	0.2387	UNahlinR	0.1814
AndCr-2	AndCr	0.9995				
AndCr-4	AndCr	0.9986				
AndCr-5	AndCr	0.9955				
AndCr-6	AndCr	1.0000				
AndCr-7	AndCr	0.9839	KowatuaCr	0.0104		
AndCr-8	AndCr	1.0000				
AndCr-9	AndCr	0.9998				
AndCr-10	AndCr	0.9859	LTahltanR	0.0132		
AndCr-11	AndCr	0.9995				
AndCr-12	AndCr	0.9454	LTahltanR	0.0532		
AndCr-13	AndCr	1.0000				
AndCr-14	AndCr	0.9967				
AndCr-15	AndCr	0.9050	LTahltanR	0.0372	TatsatuaCr	0.0280
AndCr-16	AndCr	0.6727	LTahltanR	0.3078	NakinaR	0.0173

Possible origins of each individual are given until the sum of all probabilities is greater than 0.99.

In addition, the output file contains probabilities for all of the fish in the mixture for all of the baseline populations. The output for this looks something like:

PROBABILITY OF EACH INDIVIDUAL IN THE MIXTURE BELONGING TO EACH BASELINE POPULATION

	Andr	Kowa	LTah	Naki	Tats	UNah
AndrewCr-1	0.5117	0.0243	0.2387	0.0096	0.0341	0.1814
AndrewCr-2	0.9995	0.0000	0.0004	0.0000	0.0000	0.0000
AndrewCr-3	0.4167	0.0101	0.5575	0.0066	0.0031	0.0058
AndrewCr-4	0.9985	0.0000	0.0010	0.0000	0.0000	0.0002

4b. Leave one out test. The leave one out test evaluates how well fish can be assigned to their population of origin. The test is conducted by removing fish from baseline populations (one at a time) and then estimating their origin.

Methods. The leave one out test is found in the [Individual assignment] menu. During the test, each fish in each baseline population is sequentially removed from the baseline, and its origin is estimated using the rest of the baseline. If a fish's multilocus genotype is incomplete (i.e. there is missing data), it will be dropped from analysis (but will remain in the baseline in order to estimate the origin of other fish). Once every fish has been tested, ONCOR records the fraction of

assignments for each population that were correct and the population to which individuals were most often incorrectly assigned to.

Results. An example of output from a leave one out test is shown below.

```
ONCOR Output
10/20/2007 11:29:09 AM
Baseline ID = GAPS 2.1 - Stikine & Taku populations
Baseline file = StikineTaku BASELINE.txt


PROPORTION OF BASELINE INDIVIDUALS CORRECTLY ASSIGNED TO POPULATION

               N      % Correct      Largest Misidentification

AndrewCr       103    78.6%          LTahltanR      7.8%
KowatuaCr      121    41.3%          TatsatuaCr     29.8%
LTahltanR      110    33.6%          UNahlinR       20.0%
NakinaR        116    26.7%          KowatuaCr      19.8%
TatsatuaCr     103    41.7%          KowatuaCr      31.1%
UNahlinR       124    35.5%          NakinaR        18.5%
```

The summary results shown above are output to screen. ONCOR also writes more complete results to file. An example is shown below.

```
ROWS list where individuals were FROM.
COLUMNS list where individuals were assigned TO.

               Andr    Kowa    LTah    Naki    Tats    UNah
AndrewCr       81      2       8       4       2       6
KowatuaCr      4       50      10      11      36      10
LTahltanR      11      6       37      19      15      22
NakinaR 4      23      23      31      12      23
TatsatuaCr     7       32      8       10      43      3
UNahlinR       3       18      22      23      14      44
```

4c. Realistic fishery simulations. This option simulates samples from a fishery and tests how well the baseline data can identify the origin of each individual.

Methods. ONCOR simulates data for realistic fishery assignment tests in the same manner as for for mixture analysis. As indicated above, ONCOR simulates both fishery samples and a baseline to

analyze them. If the user specifies that baseline sample sizes are the same as the actual baseline, ONCOR uses the resampling method of Anderson et al. (2007) to simulate mixture genotypes and to estimate their probability of occurring in baseline populations. If other sample sizes are used, ONCOR assumes that the parametric allele frequencies in baseline populations are equal to the reduced-variance estimates of Kalinowski et al. (2007), and then simulates baselines and mixture samples from these allele frequencies.

Results. When this type of simulation is run, ONCOR output a brief summary of the results to screen, and more comprehensive results to file. An example of screen output is shown below.

```
ONCOR Output - Assignment test simulation
10/20/2007 11:52:57 AM

Baseline ID     GAPS 2.1 - Stikine & Taku populations
Baseline file   StikineTaku BASELINE.txt
NMixture        200
NSimulations    1000
Sample sizes    Same as empirical baseline


Proportion of individuals correctly assigned to POPULATION (in all simulations)

                N       % Correct

AndrewCr        180070  90.8%
KowatuaCr       4058    27.0%
LTahltanR       3968    40.7%
NakinaR         3969    23.5%
TatsatuaCr      3913    35.7%
UNahlinR        4022    34.7%


Additional results have been written to: C:\Documents and
Settings\skalinowski\Desktop\GSI\ONCOR Testing\assign test.txt

Run time: 28 seconds
```

In the example data above, 1000 samples from a fishery were simulated. In these samples, there was a total of 18,070 fish from AndrewCr. Of these 18,070 fish, 90.8% were correctly assigned to AndrewCr.

The output file contains the raw data from the simulations as well as a more comprehensive summary of the data. Selections from an output file are shown below.

```
Where were individuals assigned TO?

        Row labels indicate the ACTUAL origin of individuals.
        Column labels indicate the ESTIMATED origin of individuals.

                    Andr    Kowa    LTah    Naki    Tats    UNah
        AndrewCr    0.9084  0.0091  0.0419  0.0157  0.0083  0.0167
        KowatuaCr   0.1451  0.2698  0.1390  0.1400  0.2287  0.0774
        LTahltanR   0.2177  0.0572  0.4073  0.1573  0.0552  0.1053
        NakinaR     0.1842  0.0980  0.2414  0.2348  0.0816  0.1600
        TatsatuaCr  0.1142  0.2088  0.1357  0.1145  0.3568  0.0700
        UNahlinR    0.1872  0.0547  0.1850  0.1790  0.0475  0.3466


Where did individuals come FROM?

        Row labels below indicate the ESTIMATED origin of individuals.
        Column labels below indicate the ACTUAL origin of individuals.

                    Andr    Kowa    LTah    Naki    Tats    UNah
        AndrewCr    0.9797  0.0035  0.0052  0.0044  0.0027  0.0045
        KowatuaCr   0.3743  0.2493  0.0517  0.0886  0.1860  0.0501
        LTahltanR   0.6308  0.0472  0.1352  0.0801  0.0444  0.0622
        NakinaR     0.4613  0.0929  0.1021  0.1525  0.0733  0.1178
        TatsatuaCr  0.3282  0.2039  0.0481  0.0712  0.3067  0.0420
        UNahlinR    0.4970  0.0520  0.0693  0.1052  0.0454  0.2310


RAW DATA

        Row labels indicate the ACTUAL origin of individuals.
        Column labels indicate the ESTIMATED origin of individuals.

                    Andr    Kowa    LTah    Naki    Tats    UNah
        AndrewCr    163574  1644    7540    2819    1494    2999
        KowatuaCr   589     1095    564     568     928     314
        LTahltanR   864     227     1616    624     219     418
        NakinaR     731     389     958     932     324     635
        TatsatuaCr  447     817     531     448     1396    274
        UNahlinR    753     220     744     720     191     1394
```

5. LITERATURE CITED

Anderson EC, RS Waples, ST KALINOWSKI (2007). An improved method for estimating the accuracy of genetic stock identification. Canadian Journal of Fisheries and Aquatic Sciences (Accepted pending revision July 24, 2007).

Kalinowski ST, ML Taper, ST, KR Manlove, WD Templin, EC Anderson. (2007) Estimating how different types of sampling affect the accuracy of genetic stock identification. In preparation.

Millar, R.B. 1987. Maximum likelihood estimation of mixed stock fishery composition. Canadian Journal of Fisheries and Aquatic Sciences 44: 583–590.

Rannala, B., and Mountain, J.L. 1997. Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA. 94: 9197–9201.