BACKGROUNDER: INTRODUCTION TO DNA ANALYSIS

By J.SCROGGIE and E. PORSZT

Fisheries and Oceans Canada

Acknowledgements

Special thanks to John Candy (Fisheries and Oceans Canada), Steve Latham (Pacific Salmon Commission) and Chuck Parken (Fisheries and Oceans Canada) for providing the authors with their expertise on this subject during the production of this document. Also, thanks to participants of the Joint Technical Working Group for providing useful feedback and edits.

Introduction

The Department of Fisheries and Oceans uses genetic and other techniques, such as coded-wire tags and thermal otolith marks, to provide managers, First Nations and stakeholders estimates of the contribution of various fish stocks to catches in mixed stock fisheries throughout the Pacific Region. Genetic techniques can provide information to manage fisheries to achieve rights-based, economic and conservation objectives for some stocks and species (e.g. Fraser sockeye) or for some fisheries (Northern BC troll fishery); whereas other species (e.g. Chinook) require both stock and age data, thus coded-wire tags are used coast-wide The purpose of this document is to provide the technical and non-technical reader with: 1) an overview of the genetic methods and techniques used in mixed-stock analysis; 2) a description of the way the results of these analyses are presented; and 3) a review of mixed-stock analysis results including key considerations to take into account during interpretation.

Method

Overview

It is possible to use genetic markers to estimate the regional stock group of origin for several salmon species using technology similar to the 'genetic fingerprinting' used to identify people from their DNA in forensics and criminal investigations. The ability to identify the stock of origin is particularly useful for managing multiple stocks in a mixed-stock fishery when management objectives vary by stock and the stock is abundant enough to detect in fishery samples.

The use of genetics to estimate stock of origin of mixed-stock samples relies primarily on prior identification of the genetic makeup of a large number of individuals for which stock of origin is known. These individuals, known to be from specific stocks, are called the baseline sample. The stocks of origin of fish used for the baseline sample are presumed known because these samples are collected during or close to the time of spawning in the natal stream, and these samples are assumed to not contain any stray fish from other rivers. Thanks to sampling across stocks throughout waterways draining into the north Pacific, there are very well defined baseline samples for Chinook and Sockeye salmon across their natural range, however there remain some regions that are represented by samples from a few locations only.

The differences in genetic composition among the stocks represented in the baseline allow the stock of origin to be estimated for an individual with unknown origin, based on the genetic composition of the individual. Because the genetic similarity of stocks varies with distance (generally, the closer two stocks are to one another geographically, the more similar they are genetically), it is often difficult to resolve exactly to which baseline stock an individual fish of unknown origin belongs. For this reason, computer programs are used to mathematically estimate the probability associated with assignments of unknown samples to stock of origin. However, even with the use of high powered computers, the computer is unable to estimate the probability that a fish originated from a river that is not in the baseline.

Details

The genetic makeup of an individual fish is characterized by measuring a particular suite of genetic characteristics (called genetic "loci"). It is the specific combination of measurement across loci that provide an individual's unique genetic signature. The same loci are read for each fish when performing a genetic analysis.

Genetic stock identification estimates both the relative contributions of distinct stocks in a mixture of individuals (i.e. the percentage each of a number of stocks comprises of the total sample), as well as stock-specific identification of individuals (i.e. the probability that one individual belongs to a particular stock). The first step in performing mixture analyses is to assemble a baseline of genotypic characteristics (specific genetic makeup) for all stocks that potentially contribute to the catch. A mixture of individuals (unknown stock of origin) is randomly sampled from the area and time of interest and the same suite of genetic characteristics (loci) are measured for each individual in the sample.

CBAYES is a computer program which uses Bayesian statistical methods for mixed-stock DNA analysis. Bayesian methods utilize prior knowledge of the genetic makeup of potentially contributing stocks (baseline stocks), combined with the likelihood of the genetic makeup from the mixture given different stock compositions of the mixture, to develop a probability (posterior distribution) of both the genetic makeup of baseline stocks and the percent each baseline stock (or region) contributes to the mixture. The program runs with three files, a baseline file of genotypes from known stocks, a mixture file of genotypes of unknown origin, and a control file of parameters required for the assignment of mixture genotypes to the baseline stocks. The baseline file contains information about genetic loci sampled from each of the potentially contributing stocks (known stock of origin). The mixture file contains genetic loci information for individuals in the mixtures (unknown stock of origin).

Because of the large number of calculations that need to be performed with Bayesian methods, a sampling approach called "Monte Carlo Markov Chains" (MCMC) is used, which reduces computational requirements. This technique draws an assigned number of samples from the posterior (probability) distribution of stock composition (or region composition) using a prescribed number of sampling "chains".

After an initial "burn-in" period, the MCMC chains should converge, so that the variance in the samples drawn within each chain is similar to the variance in the samples across chains. The diagnostic used to determine the degree of convergence across MCMC chains is called the Gelman-Rubin diagnostic. The estimated posterior (probability) distribution of the percent each stock/region contributes to the mixture is then based on summarizing the results of the samples taken once the "burn-in" samples are discarded.

Detailed descriptions of the methods used for mixed-stock analysis of salmon, specifically Chinook and Sockeye, are provided by these references:

Beacham, T.D., J.R. Candy, B. McIntosh, C. MacConnachie, A.Tabata, K. Kaukinen, L. Deng, K. M. Miller, R. E. Withler, and N. V. Varnavskaya. 2005. Estimation of stock composition and individual identification of sockeye salmon on a Pacific Rim basis using microsatellite and major histocompatibility complex variation. Transactions of the American Fisheries Society 134: 1124-1146.

Beacham, T. D., J. R. Candy, K. L. Jonsen, J. Supernault, M. Wetklo, L. Deng, K. M. Miller, and R. E. Withler. 2006. Estimation of stock composition and individual identification of Chinook salmon across the Pacific Rim using microsatellite variation. Transactions of the American Fisheries Society 135: 861-888.

Neaves, P. I., Wallace, C. G., Candy, J. R., and Beacham, T. D. 2004-2008. CBayes: Computer program for mixed stock analysis of allelic data. Version v5.01.

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

Interpretation of results

General

Each mixture consists of individual fish samples grouped together for analysis, based on the question at hand. After mixtures have been run through CBAYES two main types of reports are produced: stock composition reports and individual stock assignment reports. Both reports are produced at the stock and regional level (i.e., the composition of individual stocks in the mixture, and the composition of geographic regions). Another report summarizes the results of the Gelman-Rubin convergence diagnostic.

Headers with the following information (parameters defined by the analyst prior to the run) are included on each worksheet in the results file:

	A	B	C	D	E	F	G	H	I	J	K
1	Species = chinook Number of populations = 14 Baseline Description = Coastwide062607 Number of loci = 12 Max missing loci = 5										
2	Number of chains = 10 Number of Reps = 20000 Reps Kept = 1000										

Species = The species being analyzed.

Number of populations = The number of baseline stocks used in the analysis (baseline stocks consist of individuals known to be from specific stocks with estimates of their genetic makeup).

Baseline Description = A description of the baseline used in the analysis, usually consisting of information relating to the geographical area covered by baseline stocks (e.g., coast-wide or limited to the Fraser River stocks) and the date that the baseline file was created.

Number of loci = The number of genetic loci assayed in the baseline stocks.

Max missing loci = Maximum number of missing loci for a mixture individual. If a mixture individual exceeds the maximum number of missing loci (due to a poor tissue sample) it is excluded from the CBAYES analysis and will not be assigned to a stock of origin.

Number of chains = The number of Monte Carlo Markov Chains (MCMC) that ran. Each chain draws an assigned number of samples from the posterior (probability) distribution of stock/region composition.

Number of Reps = The number of samples drawn from the posterior distribution.

Reps Kept = An assigned number of samples from the posterior (probability) distribution of stock/region composition are kept and these samples make up the estimated posterior (probability) distribution of composition of each stock/region in the mixture (a pre-determined number of burn-in samples are discarded prior to summarizing the results).

Inventory worksheet

Provides the following information for each mixture:

- Sample number
- Vial ID (how vials are labeled in the field)
- Gear used, area and date of sampling
- Sample size (number of individual fish grouped together for analysis)
- Number of samples excluded from the analysis because the number of genetic loci with missing data was too large (i.e., exceeds the maximum missing loci value in the header of each worksheet, which is a parameter defined by the analyst prior to the run).

	A	B	C	D	E	F	G	H	I	J	K
1	Species = chinook Number of populations = 14 Baseline Description = Coastwide062607 Number of loci = 12 Max missing loci = 5										
2	Number of chains = 10 Number of Reps = 20000 Reps Kept = 1000										
3											
4	Sample	Vial ID	Year	Gear	Area	Mix Date	N	Excluded			
5	1		2003	Gillnet	Skeena Test	June8-14	36	5			
6	2		2003	Gillnet	Skeena Test	June15-21	82	7			
7	3		2003	Gillnet	Skeena Test	June22-30	78	12			
8	4		2003	Gillnet	Skeena Test	July1-7	45	18			
9	5		2003	Gillnet	Skeena Test	Jul8-14	89	21			
10	6		2003	Gillnet	Skeena Test	Jul15-21	46	10			
11	7		2003	Gillnet	Skeena Test	Jul22-28	41	5			
12	8		2003	Gillnet	Skeena Test	Jul29-Aug4	14	3			
13	9		2003	Gillnet	Skeena Test	Aug5-11	3	2			
14	10		2003	Gillnet	Skeena Test	Aug12-18	3	0			
15	11		2003	Gillnet	Skeena Test	Aug19-25	2	0			
16	12		2003	Gillnet	Skeena Test	June8-Aug2	439	83			
17	13		2003	Sport	Lower Skeena		63	1			
18											

Stock composition reports

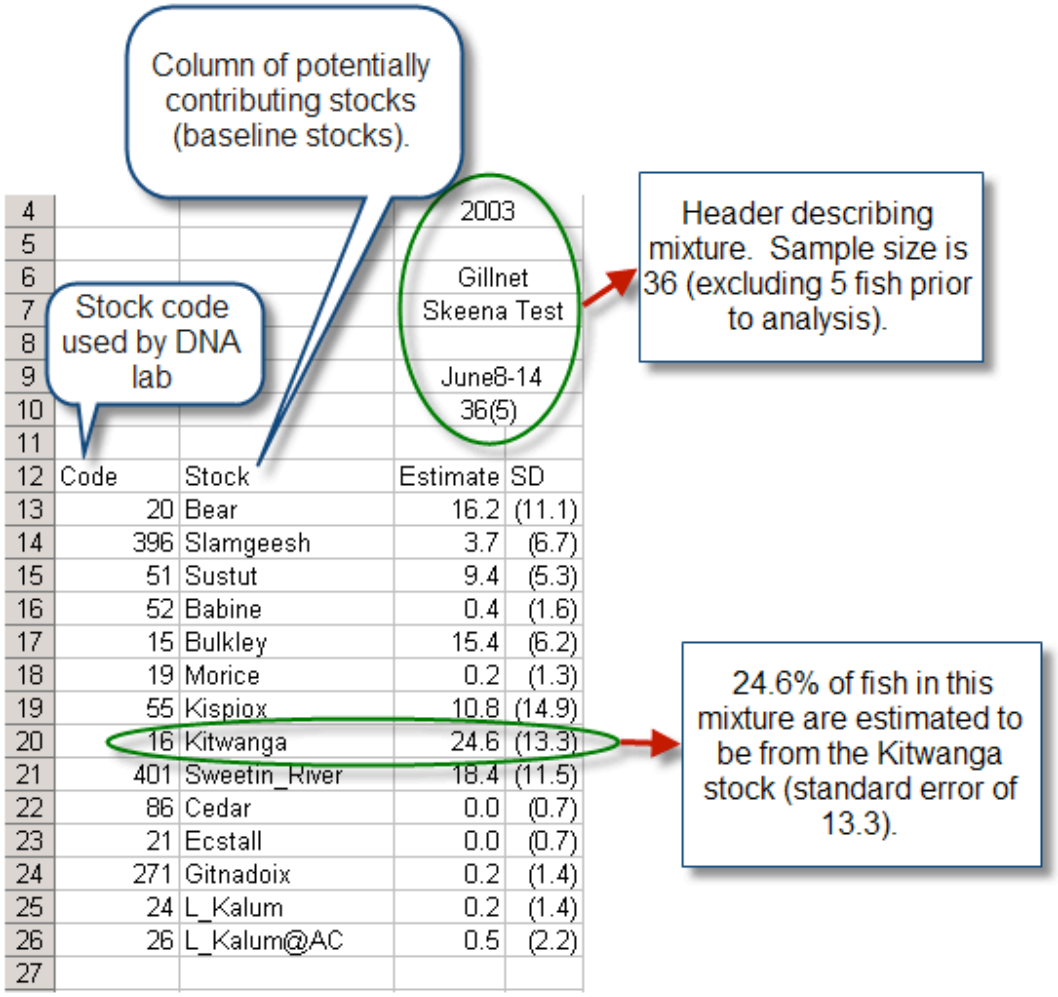
Each mixture is described by a header (see figure below). The header provides information about the mixture such as area and date sampled, gear type, and sample size (number of individual fish grouped together for analysis). The sample size is the last row of the mixture header and includes: the number of samples (N) that amplified with fewer than the maximum number of missing loci and in brackets the number of samples which were excluded from the analyses because more than the maximum number of loci were missing. It is important to note that samples that did not amplify (could not be analyzed) at all are not included in either number. Two columns report the estimated percent that each baseline stock/region contributes to that mixture (*Estimate* described below) and the standard error of this estimate (*SD* described below).

Stock proportion worksheet

Estimate: For each stock in the baseline, the roll-up of the probabilities of all individual assignments to that stock.

- Each individual fish has a probability of being assigned to each stock.
- For each stock, the probabilities of being assigned to it are summed across all individual fish in the mixture
- For each stock, the summed probability is divided by the total probability across all stocks
- This results in estimates of stock composition for each mixture that will sum to 100% across all stocks
- **Estimate is the estimated percent contribution of fish from that stock to the mixture**
- The estimated percent contribution of each stock is reported, even if a given stock is not estimated to contribute to the mixture at all (percent = 0.0)
- The contributing stocks will depend on the biology of the stocks, the time and area of sampling, and bias associated with the collection of the sample

SD: Standard error of the mean of the estimated posterior (probability) distribution of that stock's contribution to the mixture.



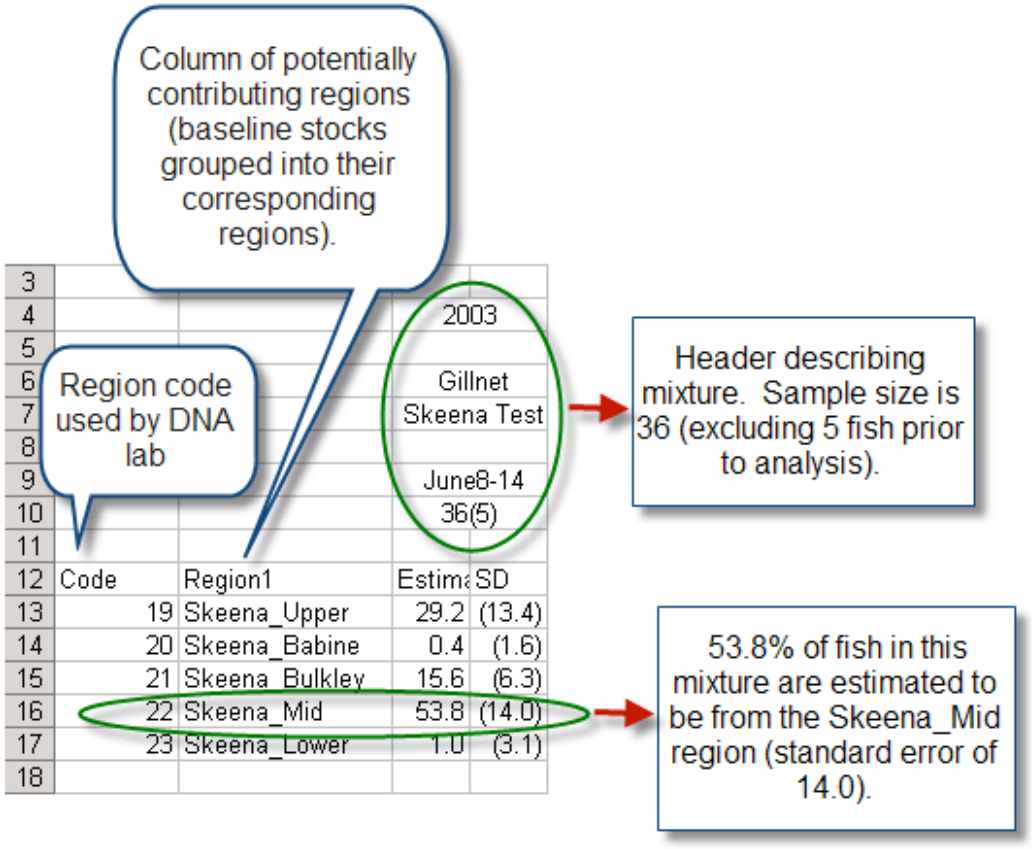
			2003	
			Gillnet	
			Skeena Test	
			June8-14	
			36(5)	
Code		Stock	Estimate	SD
	20	Bear	16.2	(11.1)
	396	Slamgeesh	3.7	(6.7)
	51	Sustut	9.4	(5.3)
	52	Babine	0.4	(1.6)
	15	Bulkley	15.4	(6.2)
	19	Morice	0.2	(1.3)
	55	Kispiox	10.8	(14.9)
	16	Kitwanga	24.6	(13.3)
	401	Sweetin_River	18.4	(11.5)
	86	Cedar	0.0	(0.7)
	21	Ecstall	0.0	(0.7)
	271	Gitnadoix	0.2	(1.4)
	24	L_Kalum	0.2	(1.4)
	26	L_Kalum@AC	0.5	(2.2)

Regional worksheet

Estimate: For each regional grouping in the baseline, the roll-up of the probabilities of all individual assignments to that region.

- Each individual fish has a probability of being assigned to each region
- For each region, the probability of being assigned to it is summed across all individual fish in the mixture
- For each region, the summed probability is divided by the total probability across all regions
- This results in estimates of region composition for each mixture that will sum to 100% across all regions
- **Estimate is the estimated percent contribution of fish from that region to the mixture**
- The estimated percent contribution of each region is reported, even if a given region is not estimated to contribute to the mixture at all (percent = 0.0)

SD: Standard error of the mean of the estimated posterior (probability) distribution of that region's contribution to the mixture

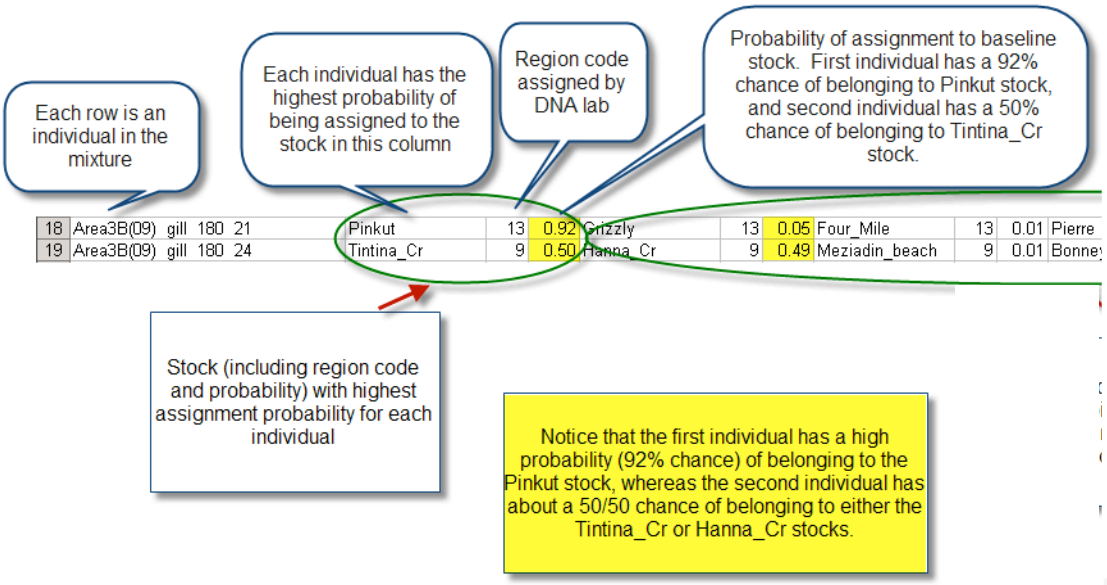


Individual stock assignment reports

Individual IDs worksheet

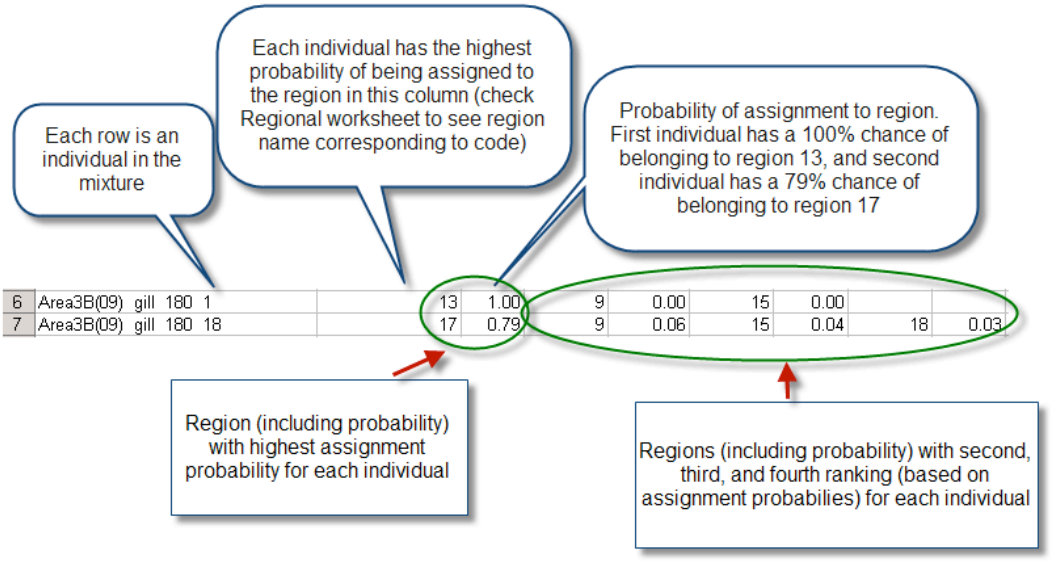
At each MCMC chain iteration (i.e., each time a sample is drawn from the posterior probability distribution) every individual fish in a mixture is assigned to one stock. These MCMC samples of stock assignments are summarized with the proportion of times each mixture individual was assigned to each stock. This is the probability of that individual being assigned to a particular baseline stock. For each mixture individual the top five most likely stocks are reported, consisting of the name of the assigned stock, the stock group or regional identification number, and the probability of assignment.

Although an assignment probability of each individual fish to each potentially contributing stock is calculated, only the top five stocks that each individual is assigned to (based on assignment probabilities) is reported in the results. If less than five stocks are reported it is because that individual has no probability of assignment to any other stocks. There will be individuals with a high assignment probability to one stock, and low assignment probabilities to all other stocks. However, there will also be individuals with similar assignment probabilities to several stocks. These stocks are most likely genetically similar and thus difficult to discriminate. In this case using Regional results may be a preferred approach due to the uncertainty of the Individual ID stock assignments.



Individual Region IDs worksheet

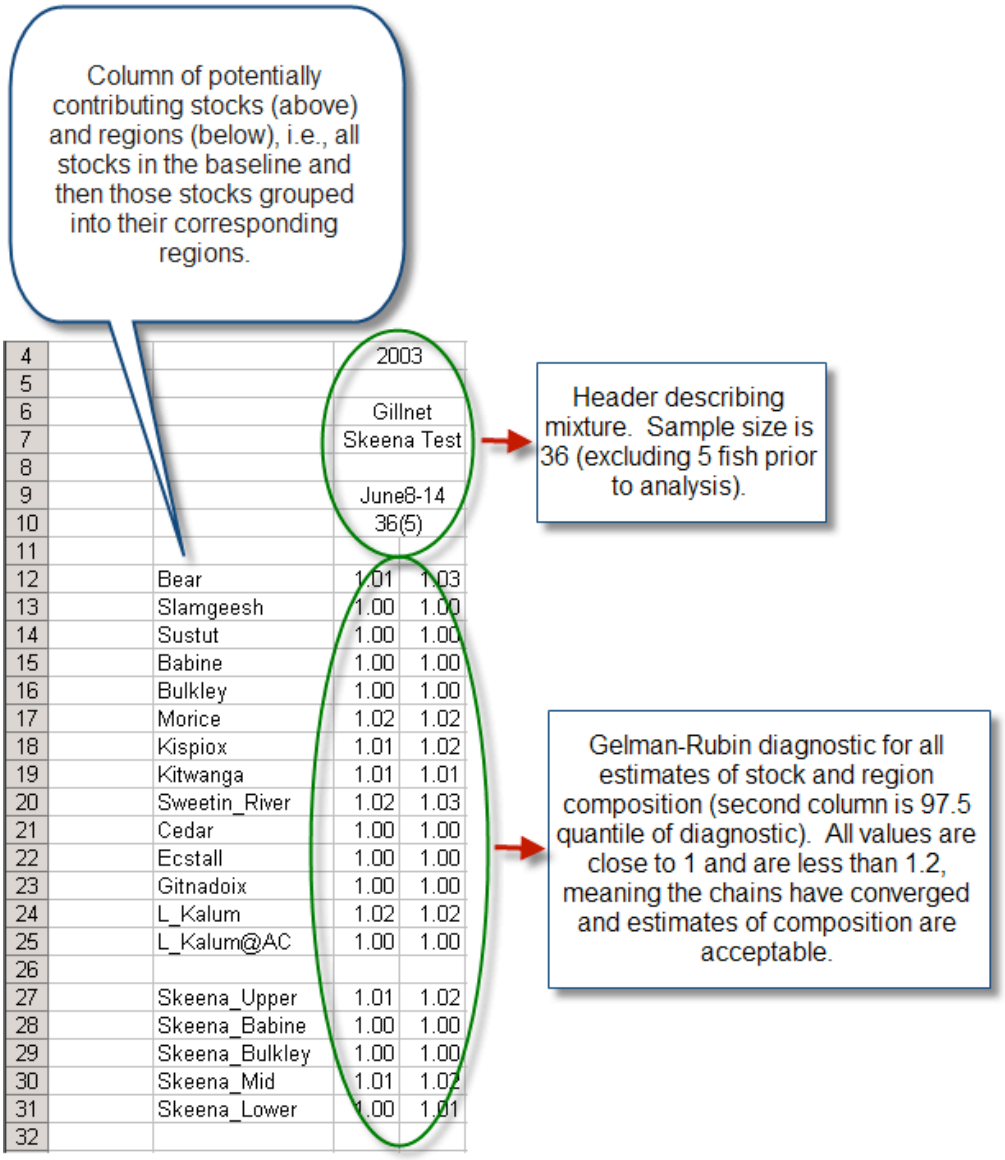
Same as Individual IDs worksheet except individuals are assigned to regions instead of stocks.



Gelman-Rubin worksheet

The Gelman-Rubin (G-R) convergence diagnostic is an estimate of convergence among MCMC chains. The G-R diagnostic is the ratio of the variability of the pooled chains to the average variability within chains. The value should be close to 1, indicating that the variance within the chains is the same as the variance across chains, and values less than 1.2 are acceptable. A value greater than 1.2 is an indicator of poorly converged chains, and suggests that a greater number of samples should be taken.

G-R diagnostics are provided for each stock and region composition estimate within a mixture (along with 97.5 quantile), in order to provide information on the chain convergence for each stock/region composition posterior (probability) distribution.



			2003	
4				
5				
6			Gillnet	
7			Skeena Test	
8				
9			June8-14	
10			36(5)	
11				
12		Bear	1.01	1.03
13		Slamgeesh	1.00	1.00
14		Sustut	1.00	1.00
15		Babine	1.00	1.00
16		Bulkley	1.00	1.00
17		Morice	1.02	1.02
18		Kispiox	1.01	1.02
19		Kitwanga	1.01	1.01
20		Sweetin_River	1.02	1.03
21		Cedar	1.00	1.00
22		Ecstall	1.00	1.00
23		Gitnadoix	1.00	1.00
24		L_Kalum	1.02	1.02
25		L_Kalum@AC	1.00	1.00
26				
27		Skeena_Upper	1.01	1.02
28		Skeena_Babine	1.00	1.00
29		Skeena_Bulkley	1.00	1.00
30		Skeena_Mid	1.01	1.02
31		Skeena_Lower	1.00	1.01
32				

Key Considerations

Interpretation and use of DNA mixture estimates

It is important to note that DNA mixture estimates, in the form presented in this document, by themselves do not provide:

- exploitation rates
- survival rates
- maturation rates
- stock-specific details required for forecasting
- age-specific details for population dynamics statistics, such as brood year contribution
- known stock identification of mixture individuals (only baseline samples are known with certainty, mixture individuals have a probability of assignment to all baseline stocks)

Currently, for Chinook salmon a coast wide assessment program based on Coded Wire Tag analyses aims to assess fishery specific exploitation rates and survival rates by cohort. The following report briefly discusses some limitations of DNA based assessment in more detail:

> *Expert Panel on the Future of the Coded Wire Tag Program for Pacific Salmon. 2005. Report. Pacific Salmon Comm. Tech. Rep. No. 18: 230 p.*

Assuming a representative sampling design, including minimum sampling targets per strata, DNA mixture results can be used to estimate stock composition of landed catch in a particular time/area fishery. Keep in mind large sample sizes will be necessary to generate reliable estimates of fishery contributions for small stocks, and results will be sensitive to small assignment errors for large stocks and ages.

Sampling design and data collection

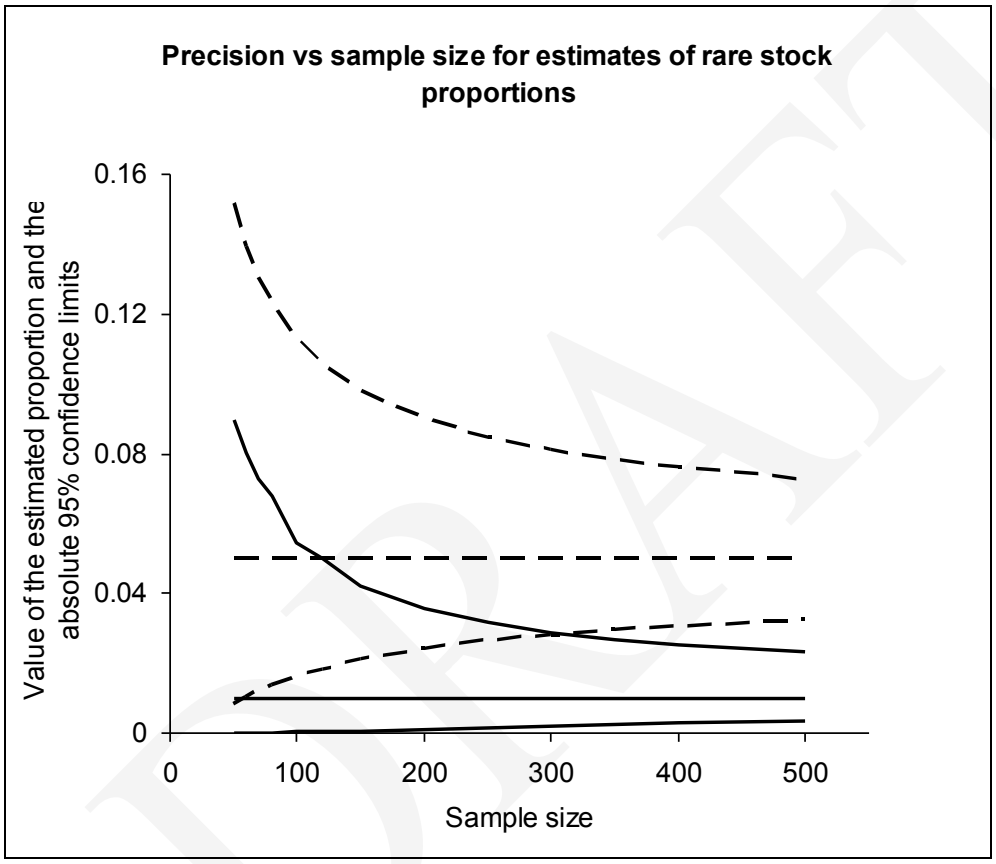
Important Assumption: The fish samples used in the mixture analysis are assumed to be independent, random and to be representative samples of the actual available population (the fish returning to an area during any given sampling period). The molecular genetics lab does not evaluate the study design used to collect the samples or the statistical population that inferences are made for: that responsibility lies with those collecting the samples, funding the analysis, and summarizing inferences. Sampling design should ensure that individual fish of unknown stock of origin are sampled in a manner that ensures this assumption is met.

The percent composition of stocks captured in a fishery may not represent the true percent composition of stocks in the actual available population for reasons including gear selectivity, various fisheries management measures (e.g., time/area closures, size/sex/species restrictions) or biases inherent in the sampling design (e.g. samples from boat-based anglers may not represent the stock composition for the catch of shore-based anglers in ocean fisheries).

The activity of random sampling is critical to make inferences beyond the sample. For example, if one aims to make an inference about the stock composition of catch, and a non-random sample was collected from small fish only, then inferences beyond the sample only represent small fish that were caught; no inferences would be accurate for medium and large fish.

Thought should be given to sampling design in order to have a sufficient number of samples within each mixture. Mixtures are set up to include the individuals at a grouping level (area/time) of relevance for stock/region composition. Mixtures usually consist of individual fish samples from a given management unit (area and time period). For example, if the fishery is managed by area and month each mixture will consist of all individual fish sampled within a given area and month, and stock composition and individual IDs will be provided for this area and month. Sampling designs should include enough samples to represent each genetic group with a reasonable level of precision. However, some genetic groups are

rare or less frequent in catches and high sampling rates are important to have achieve precision and confidence in stock composition estimates that are small (e.g. <5%). In some cases, for rare stocks reasonable relative precision will be difficult to achieve at practical sample sizes, and even large increases will not necessarily achieve very good relative precision. The figure below describes the 95% absolute confidence limits (CL) versus sample size, for estimated proportions of 5% (dashed curves) and 1% (solid curves); the matching horizontal lines indicate the estimated proportion, using a simple random sample. From the examples in the figure below, if 400 samples were taken randomly from the population the mean absolute 95% CLs are 5% +/-2.3% (relative mean 95% CL+/- 45%) and 1% +/- 1.1% (relative mean 95% CL +/- 113%). Note: The values in the figure are for proportions estimated using the binomial method and a simple random sample for an infinite population. The confidence limits are non-symmetrical and do not go below zero.



Complete and accurate data recording during sampling periods helps ensure that few samples will be eliminated from genetic analyses. Maintaining sufficient sample sizes over time and space is of key importance and a reduction of DNA samples within the analysis due to incomplete or inaccurate recording should be avoided. Care should also be taken to avoid sample duplication (tissue from one fish collected and interpreted as two different fish during collection and analyses). When duplication is not detected this procedure can skew results significantly.

Main considerations:

- Independent, random and representative sampling of the population
- Sufficient sample size within each management unit over time strata to achieve precision targets
- Effort should be given to ensure complete and accurate data collection and labelling and consideration should be given to avoid sample duplication during sample collection

Probabilities of assignment to stock of origin

A high probability of assignment to an individual stock usually occurs when the baseline stock has a very distinct genotype. It also implies that there are enough samples from that baseline stock to provide sufficient genetic information to be able to discriminate easily between stocks.

When an individual has a similar probability of assignment to numerous baseline stocks, this means that the genetic information provided by the individual was not distinct enough to clearly discriminate between potential stocks (these potential stocks are generally genetically similar). Genetically similar stocks are usually found in close proximity to each other, and there is most likely a high probability of assignment of that individual to the region containing these baseline stocks.

Lower assignment probabilities could also result when there are not enough samples from certain baseline stocks to provide sufficient genetic information to be able to discriminate between stocks.

Baseline dependence of results

Stock composition estimates and individual assignments to stock of origin depend on the stocks in the baseline. Results from mixed-stock analysis provide the most likely composition of baseline stocks found in the mixture. However, individuals can only be assigned to stocks in the baseline, and the true stock of origin for individuals may not be included in the baseline. Individuals will then be assigned to the most likely stock in the baseline, usually a stock in the same region as the true stock of origin. It is important to have good baseline sampling in the area of interest (enough samples from enough stocks to be useful).

Mixture dependence of results

Individual assignments to stock of origin are also mixture dependent. Individual assignments are the probability of an individual being assigned to a certain stock given the genetic makeup of baseline stocks and the genetic makeup of all individuals in the mixture. Therefore, the assignment of an individual may be different if it is included in different mixtures. This mainly applies to individuals with similar probabilities of assignment to numerous stocks. Those individuals with high probabilities of assignment to a particular stock will likely have the same stock assignment regardless of the mixture of which they are included.

Individual assignments may change on each run of the model but the stock composition will generally not change between runs. Individual assignments are ranked based on the highest probability of assignment to a stock of origin, but for individuals with very similar probabilities of assignment to several baseline stocks the rank order of stock assignments may change between model runs. Stock composition is a roll-up of all probabilities of individual assignments to baseline stocks, regardless of which individuals are assigned to which stock, and are thus more robust to replication.